ComSIS **Computer Science and Information Systems**

Volume 23, Number 4, September 2026

Contents

Editorial
Guest Editorial

Papers

1103 Quantifying Multi-Level Effects of AI Adoption through an End-to-End Machine Learning Pipeline
Na Tang, Yang Yang
1127 PocketPsyc: A Multi-Model AI System for Real-Time Emotion-Aware Mental Health Support
Ayesha Maqsood, Altaf Hussain, Marran Al Qwaid, Zara Haider, Muazzam A. Khan Khattak
1151 SAMS-M: Explicit Sentiment-Guided Alignment and Multi-dimensional Mutual Supervision for Multimodal Sentiment Analysis
Shishu Qi, Yulei Zhang, Siyang Zhang, Jia Xu, Kuan-Ching Li, Guangli Zhu
1181 Hybrid LSTM–HMM Framework for Regime-Aware Cryptocurrency Price Forecasting
Tran Kim Toai, Nguyen Phan Nhat Minh
1219 Hybrid Path Planning for Industrial Autonomous Vehicles: Integrating Neural Heuristics with Energy-Constrained Search
Lijun Xiao, Xinwu Jian, Zijian Wang, Sisi Zhou, Kuanching Li
1239 Vulnerable Node Detection in Complex Networks Based on Dynamic Comparator and Trusted Chain Propagation
Dongsheng Jing, Baozhen Qi, Yu Yang, Jingsong Xue, Renjun Feng
1259 Decoupled Global-Local Collaborative Network for Visual Question Answering
Ganlong Zhou, Dezhi Han, Xiang Shen, Xiaoqi Xin
1287 SLA-DE-RALBA: Cost-efficient dynamic enhanced resource-aware load balancing algorithm for cloud computing
Mohsin Nawaz, Altaf Hussain, Marran Al Qwaid, Muazzam A. Khan Khattak
1321 VR-ETL: Leveraging Virtual Reality for Pre-Transformation ETL Data Validation
Maja Cerjan, Kornelije Rabuzin, Leo Mršić
1343 Methodology for Rapid Development of PAIS Based on Ontological Modeling and Category Theory
Borivoj Bogdanović, Vidan Marković, Milan Segedinac, Zora Konjović
1395 Leveraging Content-based Metadata in Catalogs to Improve Discoverability in Data Spaces
Adriana Morejón, Alberto Berenguer, Lucía de Espona, David Tomás, Jose-Norberto Mazón

Special section: Artificial Intelligence Conference by Serbian Academy of Science and Arts

1431 Modeling Students' Acceptance of AI-Assisted Translation Tools: An Extension of the Technology Acceptance Model
Xiaoyun Wang, Vladan Devedžić, Wan Yinjia
1453 Leveraging Large Language Models for Scalable Analysis of the End-of-the-Course Student Feedback
Jelena Jovanović; Irena Vodenska; Vladan Devedžić
1485 Federated Learning Based Analysis of Certain Complex Multivariate Functions Given by Sample Only
Nevena Mijajlović, Ajlan Zajmović, Miodrag J. Mihaljević, Wei Shao, Lianhai Wang, Shujiang Xu
1519 EduGrader: The AI-powered grading assistant
Svetozar Iković, Jelena Graovac
1545 Complete meet-continuous codomain lattice and extremal solutions of L-valued fuzzy relation equations
Vanja Stepanović, Andreja Tepavčević
1563 Encoding, Linking, Retrieving: A Methodological Framework for Knowledge-Enriched Interview Corpora
Ranka Stanković, Tamara Vučenović, Milica Ikonić Nešić, Biljana Rujević

ISSN: 2406-1018 (Online)

Computer Science and Information Systems
Vol 23, Number 4, September 2026

ComSIS

Computer Science and Information Systems

Published by ComSIS Consortium

Volume 23, Number 4
September 2026

ComSIS is an international journal published by the ComSIS Consortium

ComSIS Consortium:
University of Belgrade:
Faculty of Organizational Science, Belgrade, Serbia
Faculty of Mathematics, Belgrade, Serbia
School of Electrical Engineering, Belgrade, Serbia
Serbian Academy of Science and Art:
Mathematical Institute, Belgrade, Serbia
Union University:
School of Computing, Belgrade, Serbia
University of Novi Sad:
Faculty of Sciences, Novi Sad, Serbia
Faculty of Technical Sciences, Novi Sad, Serbia
Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia
University of Niš:
Faculty of Electronic Engineering, Niš, Serbia
University of Montenegro:
Faculty of Economics, Podgorica, Montenegro

EDITORIAL BOARD:
Editor-in-Chief: Ivan Luković, University of Belgrade
Vice Editor-in-Chief: Mirjana Ivanović, University of Novi Sad
Managing Editors:
Sandro Radovanović, University Of Belgrade
Pavle Milošević, University Of Belgrade
Srđa Bjeladinović, University Of Belgrade
Jovana Vidaković, University of Novi Sad
Vladimir Kurbalija, University of Novi Sad

Editorial Assistants:
Marija Đukić, University Of Belgrade
Milica Škembarević Sretenović, University Of Belgrade
Ivan Jovanović, University Of Belgrade
Ivan Pribela, University of Novi Sad
Davorka Radaković;, University of Novi Sad
Slavica Kordić, University of Novi Sad

Editorial Board:
A. Badica, *University of Craiova, Romania*
C. Badica, *University of Craiova, Romania*
M. Bajec, *University of Ljubljana, Slovenia*
L. Bellatreche, *ISAE-ENSMA, France*
I. Berković, *University of Novi Sad, Serbia*
D. Bojić, *University of Belgrade, Serbia*
Z. Bosnic, *University of Ljubljana, Slovenia*
D. Brđanin, *University of Banja Luka, Bosnia and Hercegovina*
R. Chbeir, *University Pau and Pays Adour, France*
M-Y. Chen, *National Cheng Kung University, Tainan, Taiwan*
C. Chesñevar, *Universidad Nacional del Sur, Bahía Blanca, Argentina*
W. Dai, *Fudan University Shanghai, China*
P. Delias, *International Hellenic University, Kavala University, Greece*
B. Delibašić, *University of Belgrade, Serbia*
G. Devedžić, *University of Kragujevac, Serbia*
J. Eder, *Alpen-Adria-Universität Klagenfurt, Austria*
Y. Fan, *Communication University of China*
V. Filipović, *University of Belgrade, Serbia*
T. Galinac Grbac, *Juraj Dobrila University of Pula, Croatia*
H. Gao, *Shanghai University, China*
M. Gušev, *Ss. Cyril and Methodius University Skopje, North Macedonia*
D. Han, *Shanghai Maritime University, China*
M. Heričko, *University of Maribor, Slovenia*
M. Holbl, University of Maribor, Slovenia
L. Jain, *University of Canberra, Australia*
D. Janković, *University of Niš, Serbia*
J. Janousek, *Czech Technical University, Czech Republic*
G. Jezic, *University of Zagreb, Croatia*
G. Kardas, *Ege University International Computer Institute, Izmir, Turkey*
Lj. Kašćelan, *University of Montenegro, Montenegro*
P. Kefalas, *City College, Thessaloniki, Greece*
M-K. Khan, *King Saud University, Saudi Arabia*
S-W. Kim, *Hanyang University, Seoul, Korea*
M. Kirikova, *Riga Technical University, Latvia*
A. Klašnja Milićević, *University of Novi Sad, Serbia*
J. Kratica, *Institute of Mathematics SANU, Serbia*
K-C. Li, *Providence University, Taiwan*
M. Lujak, *University Rey Juan Carlos, Madrid, Spain*
JM. Machado, *School of Engineering, University of Minho, Portugal*
Z. Maamar, *Zayed University, UAE*
Y. Manolopoulos, *Aristotle University of Thessaloniki, Greece*
M. Mernik, *University of Maribor, Slovenia*
B. Milašinović, *University of Zagreb, Croatia*
A. Mishev, *Ss. Cyril and Methodius University Skopje, North Macedonia*
N. Mitić, *University of Belgrade, Serbia*
N-T. Nguyen, *Wroclaw University of Science and Technology, Poland*
P Novais, *University of Minho, Portugal*
B. Novikov, *St Petersburg University, Russia*
M. Paprzicky, *Polish Academy of Sciences, Poland*
P. Peris-Lopez, *University Carlos III of Madrid, Spain*
J. Protić, *University of Belgrade, Serbia*
M. Racković, *University of Novi Sad, Serbia*
M. Radovanović, *University of Novi Sad, Serbia*
P. Rajković, *University of Nis, Serbia*
C, Savaglio, *ICAR-CNR, Italy*
H. Shen, *Sun Yat-sen University, China*
J. Sierra Rodriguez, *Universidad Complutense de Madrid, Spain*
B. Stantic, *Griffith University, Australia*
H. Tian, *Griffith University, Australia*
N. Tomašev, *Google, London*
G. Trajčevski, *Northwestern University, Illinois, USA*
G. Velinov, *Ss. Cyril and Methodius University Skopje, North Macedonia*
L. Wang, *Nanyang Technological University, Singapore*
F. Xia, *Dalian University of Technology, China*
S. Xinogalos, *University of Macedonia, Thessaloniki, Greece*
S. Yin, *Software College, Shenyang Normal University, China*
K. Zdravkova, *Ss. Cyril and Methodius University Skopje, North Macedonia*
J. Zdravković, *Stockholm University, Sweden*

ComSIS Editorial Office:
University of Belgrade, Faculty of Organizational Sciences,
Jove Ilića 154, 11000 Belgrade, Serbia
Phone: +381 11 3950 866; +381 66 8896 155
www.comsis.org; Email: comsis@fon.bg.ac.rs

Volume 23, Number 4, 2026
Belgrade

Computer Science and Information Systems

ISSN: 2406-1018 (Online)

The ComSIS journal is sponsored by:

Ministry of Education, Science and Technological Development of the Republic of Serbia
http://www.mpn.gov.rs/



Computer Science and Information Systems

AIMS AND SCOPE

Computer Science and Information Systems (ComSIS) is an international refereed journal, published in Serbia. The objective of ComSIS is to communicate important research and development results in the areas of computer science, software engineering, and information systems.

We publish original papers of lasting value covering both theoretical foundations of computer science and commercial, industrial, or educational aspects that provide new insights into design and implementation of software and information systems. In addition to wide-scope regular issues, ComSIS also includes special issues covering specific topics in all areas of computer science and information systems.

ComSIS publishes invited and regular papers in English. Papers that pass a strict reviewing procedure are accepted for publishing. ComSIS is published semiannually.

Indexing Information

ComSIS is covered or selected for coverage in the following:

- Science Citation Index (also known as SciSearch) and Journal Citation Reports / Science Edition by Thomson Reuters, with 2024 two-year impact factor 1.8,
- Computer Science Bibliography, University of Trier (DBLP),
- EMBASE (Elsevier),
- Scopus (Elsevier),
- Ex Libris Knowledge Center by Clarivate,
- IET bibliographic database Inspec,
- FIZ Karlsruhe bibliographic database io-port,
- Index of Information Systems Journals (Deakin University, Australia),
- Directory of Open Access Journals (DOAJ),
- Google Scholar,
- Journal Bibliometric Report of the Center for Evaluation in Education and Science (CEON/CEES) in cooperation with the National Library of Serbia, for the Serbian Ministry of Education and Science,
- Serbian Citation Index (SCIndeks),
- doiSerbia.

Information for Contributors

The Editors will be pleased to receive contributions from all parts of the world. An electronic version (LaTeX) prepared as described in "Manuscript Requirements" (which may be downloaded from http://www.comsis.org), along with a cover letter containing the corresponding author's details should be submitted via journal submission system.

Criteria for Acceptance

Criteria for acceptance will be appropriateness to the field of Journal, as described in the Aims and Scope, taking into account the merit of the content and presentation. The number of pages of submitted articles is limited to 30 (using the appropriate LaTeX template).

Manuscripts will be referred in manner customary with scientific journals before being accepted for publication.

Copyright and Use Agreement

All authors are requested to sign the "Transfer of Copyright" agreement before the paper may be published. The copyright transfer covers the exclusive rights to reproduce and distribute the paper, including reprints, photographic reproductions, microform, electronic form, or any other reproductions of similar nature and translations. Authors are responsible for obtaining from the copyright holder permission to reproduce the paper or any part of it, for which copyright exists.

Computer Science and Information Systems

Volume 23, Number 4, September 2026

CONTENTS

Editorial

Guest Editorial

Papers

1103 **Quantifying Multi-Level Effects of AI Adoption through an End-to-End Machine Learning Pipeline**
Na Tang, Yang Yang

1127 **PocketPsyc: A Multi-Model AI System for Real-Time Emotion-Aware Mental Health Support**
Ayesha Maqsood, Altaf Hussain, Marran Al Qwaid, Zara Haider, Muazzam A. Khan Khattak

1151 **SAMS-M: Explicit Sentiment-Guided Alignment and Multi-dimensional Mutual Supervision for Multimodal Sentiment Analysis**
Shishu Qi, Yulei Zhang, Siyang Zhang, Jia Xu, Kuan-Ching Li, Guangli Zhu

1181 **Hybrid LSTM–HMM Framework for Regime-Aware Cryptocurrency Price Forecasting**
Tran Kim Toai, Nguyen Phan Nhat Minh

1219 **Hybrid Path Planning for Industrial Autonomous Vehicles: Integrating Neural Heuristics with Energy-Constrained Search**
Lijun Xiao, Xinwu Jian, Zijian Wang, Sisi Zhou, Kuanching Li

1239 **Vulnerable Node Detection in Complex Networks Based on Dynamic Comparator and Trusted Chain Propagation**
Dongsheng Jing, Baozhen Qi, Yu Yang, Jingsong Xue, Renjun Feng

1259 **Decoupled Global-Local Collaborative Network for Visual Question Answering**
Ganlong Zhou, Dezhi Han, Xiang Shen, Xiaoqi Xin

1287 **SLA-DE-RALBA: Cost-efficient dynamic enhanced resource-aware load balancing algorithm for cloud computing**
Mohsin Nawaz, Altaf Hussain, Marran Al Qwaid, Muazzam A. Khan Khattak

1321 **VR-ETL: Leveraging Virtual Reality for Pre-Transformation ETL Data Validation**
Maja Cerjan, Kornelije Rabuzin, Leo Mršić

1343 **Methodology for Rapid Development of PAIS Based on Ontological Modeling and Category Theory**
Borivoj Bogdanović, Vidan Marković, Milan Segedinac, Zora Konjović

1395 **Leveraging Content-based Metadata in Catalogs to Improve Discoverability in Data Spaces**
Adriana Morejón, Alberto Berenguer, Lucía de Espona, David Tomás, Jose-Norberto Mazón

Special section: Artificial Intelligence Conference by Serbian Academy of Science and Arts

1431 **Modeling Students' Acceptance of AI-Assisted Translation Tools: An Extension of the Technology Acceptance Model**
Xiaoyun Wang, Vladan Devedžić, Wan Yinjia

1453 **Leveraging Large Language Models for Scalable Analysis of the End-of-the-Course Student Feedback**
Jelena Jovanović; Irena Vodenska; Vladan Devedžić

1485 **Federated Learning Based Analysis of Certain Complex Multivariate Functions Given by Sample Only**
Nevena Mijajlović, Ajlan Zajmović, Miodrag J. Mihaljević, Wei Shao, Lianhai Wang, Shujiang Xu

1519 **EduGrader: The AI-powered grading assistant**
Svetozar Iković, Jelena Graovac

1545 **Complete meet-continuous codomain lattice and extremal solutions of L-valued fuzzy relation equations**
Vanja Stepanović, Andreja Tepavčević

1563 **Encoding, Linking, Retrieving: A Methodological Framework for Knowledge-Enriched Interview Corpora**
Ranka Stanković, Tamara Vučenović, Milica Ikonić Nešić, Biljana Rujević

Computer Science and Information Systems 23(4): i–iv https://doi.org/10.2298/CSIS260400iL

Editorial

Ivan Luković[1], Mirjana Ivanović[2], Sonja Ristić[3], and Pavle Milošević[1]

[1] University of Belgrade, Faculty of Organizational Sciences,
Belgrade, Serbia
{ivan.lukovic,pavle.milosevic}@fon.bg.ac.rs
[2] University of Novi Sad, Faculty of Sciences,
Novi Sad, Serbia
mira@dmi.uns.ac.rs
[3] University of Novi Sad, Faculty of Technical Sciences,
Novi Sad, Serbia, Retired Proffesor
sdristic@uns.ac.rs

Vol. 23, No. 4. regular issue of ComSIS is in front of you. This issue of ComSIS contains 11 regular research papers, covering several computer science disciplines, from the applications of AI and LLMs, through complex and collaborative networks, cloud computing, to ontological modeling.

The authors of the paper: Quantifying Multi-Level Effects of AI Adoption through an End-to-End Machine Learning Pipeline, Na Tang and Yang Yang, propose a construction of an end-to-end machine learning pipeline — from raw heterogeneous data input, through three cascaded modules (micro: Random Forest + SHAP; meso: XGBoost; macro: K-means++ clustering), to actionable policy outputs — to quantify the complex socio-economic effects of AI adoption across countries and industries. The authors present a study that moves beyond isolated econometric models and provides quantifiable, actionable insights for balancing AI innovation with social welfare, by coupling machine learning methods in a tripartite analytical pipeline.

In the paper: PocketPsyc: A Multi-Model AI System for Real-Time Emotion-Aware Mental Health Support, Ayesha Maqsood, Altaf Hussain, Marran Al Qwaid, Zara Haider, and Muazzam A. Khan Khattak present PocketPsyc, a mobile-based system designed to deliver scalable, evidence-based mental health support through Cognitive Behavioral Therapy (CBT). The system integrates three specialized AI models: a fine-tuned BART model for therapeutic response generation, a RoBERTa-base classifier for real-time emotion recognition, and a TinyLlama-1.1B model for personalized mindfulness guidance. To ensure user privacy, the platform employs client-side AES-256 encryption and row-level security mechanisms. Additionally, a hybrid crisis detection module combines clinical threshold monitoring with linguistic cue analysis to identify high-risk scenarios.

The authors of the paper SAMS-M: Explicit Sentiment-Guided Alignment and Multi-dimensional Mutual Supervision for Multimodal Sentiment Analysis, Shishu Qi, Yulei Zhang, Siyang Zhang, Jia Xu, Kuan-Ching Li, and Guangli Zhu, advocate that Multimodal Sentiment Analysis leverages the fusion of heterogeneous data to achieve fine-grained emotional understanding, which finds extensive application in large-scale public opinion monitoring and data mining. The authors propose an explicit sentiment-guided alignment and multi-dimensional cross-modal mutual supervision-based model for multimodal sentiment analysis. The model primarily employs a fine-grained sentiment–saliency directed alignment mechanism, which leverages bidirectional cross-attention to couple textual sentiment cues with visual saliency, enabling precise localization of sentiment-

relevant regions. Furthermore, they introduce a tripartite strong contrastive learning strategy to mitigate distribution discrepancies between heterogeneous modalities within a shared latent space, thereby enhancing cross-modal coherence and complementarity. Finally, the authors design a noise-robust gating-based fusion module, which, together with text augmentation and deep supervision, facilitates effective joint optimization.

Nonlinear and regime-shifting price behavior of cryptocurrencies challenges traditional econometric and deep learning models, creating a need for more adaptive forecasting frameworks. Addressing this research gap, in the paper: Hybrid LSTM–HMM Framework for Regime-Aware Cryptocurrency Price Forecasting, Tran Kim Toai and Nguyen Phan Nhat Minh present a study that proposes a hybrid Hidden Markov Model–Long Short-Term Memory (HMM–LSTM) architecture that jointly captures latent market regimes and temporal dependencies in cryptocurrency price dynamics. The HMM identifies latent market regimes, while the LSTM captures temporal dependencies. Technical indicators, including Bollinger Bands, EMA, and MACD, support feature benchmarking.

Lijun Xiao, Xinwu Jian, Zijian Wang, Sisi Zhou, and Kuanching Li, the authors of the research presented in the paper: Hybrid Path Planning for Industrial Autonomous Vehicles: Integrating Neural Heuristics with Energy-Constrained Search advocate that energy-aware path planning for industrial autonomous vehicles requires more than a geometric shortest path: terrain, payload, drivetrain efficiency, turning, and a hard task-level energy budget jointly affect feasibility. They present Neuro-PathNet-EC, a physics-aware constrained planner with learned multimodal guidance. The planner operates on heading-aware grid states and evaluates directed edge energy from friction and elevation. Exact Pareto labels enforce the hard energy budget, while a goal-conditioned ResNet-18 network fuses four raster channels with vehicle semantics and predicts an eight-channel dense cost-to-go map. The learned heuristic is used only for guidance; it is not assumed admissible or consistent.

Accurately identifying vulnerable nodes in complex networks is a critical challenge for preventing cascading failures and ensuring system robustness. The research presented in the paper: Vulnerable Node Detection in Complex Networks Based on Dynamic Comparator and Trusted Chain Propagation by Dongsheng Jing, Baozhen Qi, Yu Yang, Jingsong Xue, and Renjun Feng states that existing methods generally suffer from limitations such as static metrics that ignore dynamic loads, high parameter dependency, and a lack of unified fusion mechanisms. To address these issues, the authors propose an approach that is based on dynamic comparator selection and trusted chain propagation, referred as Optimal Parameter (OP). Specifically, the proposed OP approach adaptively adjusts strategies based on node degree: low-degree nodes enable a lightweight mode to reduce overhead, while high-degree nodes employ full sampling to enhance reliability. The method integrates temporal consistency checks, faulty comparator isolation, and cross-layer comparison mechanisms to effectively handle complex scenarios such as malicious camouflage and low connectivity.

Ganlong Zhou, Dezhi Han, Xiang Shen, and Xiaoqi Xin in the paper: Decoupled Global-Local Collaborative Network for Visual Question Answering propose a Decoupled Global-Local Collaborative Network (DGL-Net) that structurally partitions the cross-modal reasoning process into two complementary branches: global semantic understanding and local fine-grained modeling. The global branch employs hierarchical attention fusion to model long-range dependencies and holistic scene semantics, providing contex-

tual constraints for reasoning. The local branch introduces semantic relation enhancement and relevant region screening to explicitly model intra-modal structural relationships, strengthening fine-grained region-word alignment by injecting structural information. The authors state that the proposed method exhibits enhanced robustness and generalization in complex semantic and compositional reasoning tasks, validating the efficacy of the decoupled collaborative strategy.

Cloud computing has become a dominant paradigm for delivering scalable and flexible on-demand resources. However, efficiently executing high-performance computing (HPC) workloads remains challenging, particularly in heterogeneous environments. In the paper: SLA-DE-RALBA: Cost-efficient dynamic enhanced resource-aware load balancing algorithm for cloud computing, Mohsin Nawaz, Altaf Hussain, Marran Al Qwaid, and Muazzam A. Khan Khattak advocate that conventional static scheduling methods often lead to poor resource utilization and increased makespan, while dynamic approaches improve load distribution but introduce significant overhead due to continuous monitoring and real-time decision-making. To address these challenges, they propose an SLA-aware Dynamic Enhanced Resource-Aware Load Balancing Algorithm (SLA-DE-RALBA). The algorithm minimizes load imbalance by considering the computational capacities of virtual machines and ensures the Service Level Agreement compliance through a three-tier priority-based workflow.

Data validation in ETL workflows is most often performed after transformation, even though many data quality issues originate earlier in the pipeline. The authors in the paper: VR-ETL: Leveraging Virtual Reality for Pre-Transformation ETL Data Validation, Maja Cerjan, Kornelije Rabuzin, and Leo Mršić state that in practice errors are detected late, when tracing their source and applying corrections already requires additional effort. The authors in the paper explore whether virtual reality (VR) could be used as an alternative interface for addressing this problem earlier, at the stage of raw data inspection. They propose VR-ETL, a proof-of-concept system that allows users to inspect and correct tabular data before transformation by interacting with it inside an immersive 3D environment. The system imports CSV datasets and applies a set of rule-based validation checks targeting common data quality problems. Detected issues are visualized directly within the 3D representation of the data, using color-coded cells and simple spatial summaries that reflect error distribution across columns and records.

In the paper: Methodology for Rapid Development of PAIS Based on Ontological Modeling and Category Theory, Borivoj Bogdanović, Vidan Marković, Milan Segedinac, and Zora Konjović propose a methodology for the rapid development of Process-Aware Information Systems (PAIS) based on ontological modeling and category theory. Ontology-Driven Code Generation (ODCG) is formalized as a deterministic, multi-phase transformation from the BAB Core ontology, through a platform-specific ontology and library, to source code. Model levels are represented as categories, while transformations are represented as functors. The main contribution is a categorical formalization of stack-parametrized multi-target generation, together with a conditional guarantee of schema-level structural preservation under explicitly stated verification assumptions, including mapping completeness, injectivity, quotient compatibility, and absence of unintended collapse. The methodology distinguishes the categorical specification, in which functoriality follows by construction, from its operational implementation, whose conformance is established through explicit engineering checks.

In their paper: Leveraging Content-based Metadata in Catalogs to Improve Discoverability in Data Spaces, Adriana Morejón, Alberto Berenguer, Lucía de Espona, David Tomás, and Jose-Norberto Mazón state that data spaces facilitate sovereignty-preserving data sharing within federated ecosystems, while dataset discovery is often constrained by catalogs that rely primarily on high-level metadata standards such as DCAT-AP. These models typically lack structural and content-level details, which impedes fine-grained relevance assessment and complicates both internal dataset discovery and the federation of external open data into data spaces. The authors propose cbmDCAT-AP, a backward-compatible extension of DCAT-AP that integrates content-based meta-data as primary catalog elements. The proposed extension augments dataset descriptions with field-level semantics, and representative sample values, thereby enabling content-informed discovery without disclosing entire datasets.

On behalf of the ComSIS Consortium, we would like to take this opportunity to give our great thanks to the reviewers and all the authors for their high-quality work, great efforts, and remarkable enthusiasm.

Computer Science and Information Systems 23(4): v–vii https://doi.org/10.2298/CSIS260400vS

Guest Editorial

Ranka Stanković[1], Dušan Teodorović[2], and Mirjana Ivanović[3]

[1]University of Belgrade, Faculty of Mining and Geology
Belgrade, Serbia
ranka.stankovic@rgf.bg.ac.rs

[2]Serbian Academy of Sciences and Arts
Belgrade, Serbia
duteodor@gmail.com

[3]University of Novi Sad, Faculty of Sciences
Novi Sad, Serbia
mira@dmi.uns.ac.rs

This Special Issue of the Computer Science and Information Systems (ComSIS) journal presents a selection of the best papers from the third edition of the Artificial Intelligence conference (2025), organized under the auspices of the Serbian Academy of Sciences and Arts (SASA) by its Department of Technical Sciences, the Mathematical Institute of SASA, and the Academy Committee for Artificial Intelligence. The authors of the selected papers were invited to submit substantially extended versions of their work, and all submissions underwent the journal's standard rigorous peer-review process.

Established in 2023, the AI Conference at SASA has quickly become a premier annual, research-oriented event for the AI scientists in Serbia and the wider region. In 2025, the conference became truly international, attracting participants from numerous countries. The conference addresses the growing need for a dedicated academic forum, bringing together both experienced and early-career researchers across diverse branches of AI to share original results, exchange ideas, discuss fundamental theoretical and practical challenges, and strengthen collaboration across the diverse fields of artificial intelligence.

Covering a broad range of topics that reflect the rapidly evolving landscape of artificial intelligence, the conference maintains a strong focus on advancing the field's technical and theoretical boundaries. By highlighting new trends and fostering collaborative regional and international networks, it provides an important forum for promoting excellence in AI research standards and connecting the scientific community within the distinguished academic environment of the Serbian Academy of Sciences and Arts in Belgrade.

Initially fifteen papers were selected by the Conference organizers for this special issue and the authors of seven papers submitted extended versions. After undergoing of standard ComSIS reviewing procedure, six papers were accepted for publication.

In the paper Modeling Students' Acceptance of AI-Assisted Translation Tools: An Extension of the Technology Acceptance Model (TAM-AI) authors Wang Xiaoyun, Vladan Devedzic, and Wan Yinjia investigated students' acceptance of AI-assisted translation tools by proposing an extended Technology Acceptance Model (TAM-AI). The proposed model explains how translation-specific perceptions influence behavioral intention through underlying cognitive mechanisms. A survey was conducted among undergraduate translation students, and it was analyzed using Structural Equation Modeling. Presented results indicate that perceived usefulness and perceived ease of use significantly predict behavioral intention. Finally, the TAM-AI model demonstrates greater explanatory power than

the baseline TAM which offers practical implications for the design of AI-assisted translation tools.

The research presented in the paper Leveraging Large Language Models for Scalable Analysis of the End-of-the-Course Student Feedback by Jelena Jovanovic, Irena Vodenska, and Vladan Devedzic presents an LLM-based feedback analytics pipeline designed to transform students' open-ended feedback into structured, actionable insights. The pipeline includes three sequential stages: (i) segmenting student feedback into semantic units and assigning polarity (sentiment) to those units; (ii) topical classification of semantic units, and (iii) summarization of units within each topical category. The proposed approach enables educators and course managers to efficiently derive meaningful patterns from vast datasets of student opinions. It was evaluated using a comprehensive dataset from several editions of a U.S. university course, yielding encouraging results of the method's effectiveness. Additionally, a scalable, generic methodology for (semi-)automated feedback analysis, ultimately, supporting data-informed improvements in teaching and course management is provided.

Federated Learning Based Analysis of Certain Complex Multivariate Functions Given by Sample Only, by authors Nevena Mijajlovic, Ajlan Zajmovic, Miodrag J. Mihaljevic, Wei Shao, Lianhai Wang, and Shujiang Xu considered Complex Multivariate Functions employing a machine learning approach. The main goal of research is to generate a surrogate of the function based on the given sample, and to perform the optimization on the surrogate function. Due to the distributed nature of the sample Federated Machine Learning (FML) approach is adopted to generate the surrogate function. Authors assume that certain clients participate in FML in a malicious manner by poisoning training data or the model parameters they generate locally. Systematic experiments have been performed to study the characteristics of the surrogate function and its capabilities to provide reliable analysis of the original function depending on the sample available for training, and impacts of the training data poisoning, as well as locally generated model parameters.

In paper EduGrader: The AI-powered grading assistant, Svetozar Iković and Jelena Graovac present an open-source platform that leverages state-of-the-art large language models to support the assessment of open-ended student responses. The study addresses one of the key challenges in modern higher education, providing consistent, scalable, and timely evaluation through a configurable human-in-the-loop framework that combines multiple LLMs, adjustable grading strictness, and explainable feedback. Evaluated on 686 authentic student responses from six university courses in Serbian, a low-resource language, EduGrader achieves a Pearson correlation of up to 0.90 with human graders while maintaining transparency and instructor control. By demonstrating that advanced reasoning models can substantially reduce grading workload without replacing human judgment, this work offers a significant contribution to AI-assisted educational assessment, and it highlights the potential of trustworthy LLM-based technologies beyond English-language educational settings.

Complete meet-continuous codomain lattice and extremal solutions of L-valued fuzzy relation equations by authors Vanja Stepanović and Andreja Tepavčević brings a significant theoretical contribution to the study of lattice-valued fuzzy relational systems by establishing the existence of maximal solutions for a broad class of fuzzy relational equations, inequations, and their systems under the assumption of complete meet-continuous codomain lattices. Beyond extending fundamental lattice-theoretic properties and proving

the equivalence between meet-continuity and composition-continuity, the authors develop constructive algorithms for computing least solutions that terminate after at most countably many steps. By generalizing several earlier results into a unified framework and discussing potential applications in decision support, knowledge representation, control systems, and fuzzy data analysis, this work provides both a rigorous mathematical foundation and practical computational tools for reasoning with uncertainty in fuzzy relational models.

In the paper Encoding, Linking, Retrieving: A Methodological Framework for Knowledge-Enriched Interview Corpora, Ranka Stanković, Tamara Vučenović, Milica Ikonić Nešić, and Biljana Rujević present a comprehensive methodological framework for transforming unstructured interview transcripts into semantically enriched, AI-enabled knowledge resources. By integrating TEI-compliant encoding, textometric analysis, Named Entity Recognition and Linking, Wikidata-based semantic enrichment, and Retrieval-Augmented Generation (RAG), the authors establish a reproducible workflow for constructing knowledge-centric interview corpora. The proposed pipeline combines advanced semantic search, question answering, and knowledge graph integration with a systematic methodology for transforming oral history and cultural heritage collections into interoperable, FAIR-compliant resources for computational analysis. The work provides an important contribution to digital humanities and language technologies, particularly for Serbian and other less-resourced languages, offering a scalable methodology applicable to a wide range of spoken-language archives and documentary collections.

We would like to take this opportunity to give our great thanks to the Organizing and Program Chairs for careful evaluation of papers presented at AI 20205 conference for selecting the most outstanding and impactful contributions among them for inclusion in this special issue.

We are equally grateful to all authors for submitting their high-quality and inspiring papers, to the reviewers for their effort in reviewing the manuscripts, thoughtful evaluations, constructive feedback, and dedication which significantly contributed to improving quality of their papers. Finally, we extend our heartfelt thanks to the Editor-in-Chief, Prof. Ivan Luković, and the editorial team for their continuous support, valuable guidance and professionalism throughout the entire process of preparation of this special issue.

Computer Science and Information Systems 23(4):1103–1126 https://doi.org/10.2298/CSIS251101037T

Paper Type: Research Paper

Quantifying Multi-Level Effects of AI Adoption through an End-to-End Machine Learning Pipeline ⋆

Na Tang[1,2] and Yang Yang[1]

[1] Business School, Geely University of China,
No. 123, Section 2, Chengjian Avenue, Chengdu, 610225, China
tangna@guc.edu.cn
yangyang@guc.edu.cn (corresponding author)
[2] Faculty of Economics and Business, Universiti Malaya
Jalan Universiti, 50603 Kuala Lumpur, Malaysia
s2134677@siswa.um.edu.my

Abstract. Artificial Intelligence is reshaping global socio-economic structures, yet existing studies often analyze its economic, social, and market impacts in isolation, lacking a unified framework that handles multi-dimensional interactions. This paper constructs an end-to-end machine learning pipeline — from raw heterogeneous data input, through three cascaded modules (micro: Random Forest + SHAP; meso: XGBoost; macro: K-means++ clustering), to actionable policy outputs — to quantify the complex socio-economic effects of AI adoption across countries and industries. Using a comprehensive global dataset (2020–2025), this research finds that: (1) Human-AI Collaboration Rate (HAICR) is the dominant mitigator of AI-driven job loss, outweighing AI Adoption Rate (AIAD), while high AIAD significantly exacerbates job displacement in automated sectors like manufacturing and retail; (2) Strict regulation is the strongest driver of consumer trust, but its efficacy is moderated by cultural contexts and tool transparency; (3) China and the U.S. lead AI-driven markets, whereas India and South Korea require infrastructure and policy upgrades. By coupling machine learning methods in a tripartite analytical pipeline, this study moves beyond isolated econometric models and provides quantifiable, actionable insights for balancing AI innovation with social welfare.

Keywords: Artificial Intelligence Adoption, Machine Learning, Multi-Dimensional Socio-Economic Impact, Human-AI Collaboration Rate, AI Regulation Status

1. Introduction

The rapid development of artificial intelligence technologies, especially driven by machine learning models, has brought revolutionary changes to various industries, from manufacturing and retail to healthcare and education. AI-generated content and automated tools are reshaping production processes, labor markets, and consumer behavior, bringing opportunities and challenges to socioeconomic systems. While existing research has explored the economic, social, and market impacts of AI in isolation, the interrelated

⋆ This is an extended version of ALMTA25 conference paper.

mechanisms that link these aspects remain understudied. This gap limits the ability to develop holistic strategies that balance technological innovation with social well-being.

This study addresses three key questions. First, how does AI adoption affect economic indicators such as revenue growth and job losses across industries? Second, what factors drive consumer trust in AI tools, and how does the regulatory framework mediate this relationship? Finally, how do the spatiotemporal dynamics drive by AI shape market structure evolution?

To address these questions, this paper proposes a three-pronged analytical framework that integrates microeconomic impacts, meso-level dynamics of social trust, and macro-level market evolution. Traditional econometric models struggle to handle high-dimensional heterogeneous data, such as mixed text and numerical indicators. Therefore, the empirical analysis in this paper will utilize a global dataset covering the years 2020–2025. By employing random forest, XGBoost classification, and k-means++ clustering, we will construct an end-to-end framework to quantify the cascading effects of AI adoption from economic outcomes to social trust and market restructuring. Furthermore, this study will introduce SHAP interaction values to reveal the moderating effects of policy dependencies, such as how strict regulation mitigates unemployment risks or amplifies income growth in specific sectors. The end-to-end framework of this study does not refer to the development of a new single algorithm, but rather to the establishment of a fully integrated, modular machine learning analysis pipeline that spans from raw heterogeneous data inputs to multidimensional socioeconomic impact outputs, and ultimately to actionable policy recommendations. Specifically at the data analysis level, the micro, meso, and macro analytical levels are organically linked through data flows; in other words, the output variables from micro-level models serve as feature inputs for macro-level clustering models.

The innovation of this study lies in connecting fragmented analyses through a machine learning coupling method. This study provides actionable policy recommendations and strategic guidance for enterprises, emphasizing the need for adaptive governance, scenario-specific trust building, and cross-industry collaboration. This study advances the discourse on the socioeconomic impact of AI and provides a scalable approach for future interdisciplinary research.

2. Related Work

2.1. The Adoption of AI technology and Its Economic Impact on Industries

The impact of AI on socioeconomic development is transformative. Julius Tan (2023) found a positive relationship between AI and economic growth, which implies that AI contributes to firms' production, marketing, and customer acquisition activities [13]. Minzhen (2019) concluded that artificial intelligence is primarily reflected in the automated processing of "programmable" work. By reducing labor costs and improving production efficiency, AI can contribute to increased profits [29]. Moreover, AI-related job growth and economic growth are driven by cities with a higher concentration of modern or professional services. AI-related job growth can also contribute to improvements in well-being [20].

However, Hang et al. (2022) indicated that AI adoption, task nature, and AI management are the key barriers preventing AI from realizing its full potential. This is because

AI lacks interpersonal skills in digital economy [16]. Also, AI and Big Data will not solve the problem of compatibility of intentions and expectations of different people [3]. Solos and Leonard (2022) pointed that since AI grows at varying rates at various stages and the economy evolves gradually, the influence of AI on income inequality may vary throughout periods of economic growth [25]. Trammell et al. (2023) discovered individual human talents will not save us; if we retain positive wages or a positive labor share, we will do so only because AI is used to make us more productive, or because we actively decide to keep some jobs, labor remains important as a means of income distribution in the long run [26]. Daron (2025) found that there is also no evidence that AI will reduce labor income inequality [2].

From a macro-environment perspective, market prices indicate that the arrival and diffusion of Generative AI represent a major technology shock with important effects on the overall value of firms[10]. AI in finance has been a continued substantial research direction over decades with increasing cross-disciplinary interactions and fusion between AI, data science, machine learning, finance and economics[6].

In conclusion, Previous research in this field has covered both macro and micro perspectives, but studies focusing on specific industries remain limited, with insufficient supporting empirical analysis to establish concrete correlations.

2.2. Social Influence and AI Tool Selection Mechanism

Makridakis (2017) proposed the view that AI technological change has a systematic reshaping of social structure based on the industrial revolution and the digital revolution, emphasizing the dominant role of technology in social production and organizational forms[19]. Donika (2023) found that distrust of AI technology stems from the understanding of the social risks of integrating it and the loss of human communication, creativity and individual values[9]. Joyce et al. (2021) called for the construction of "AI sociology", explaining that AI is not only a technical tool, and the creation of AI social technical systems is not just a technical design issue, but a fundamental issue about power and social order [17]. In the face of the systematic creation of AI, Warr and Rather (2024) criticized the polarization tendency of future work narratives through the sociology of expectations, arguing that technological imagination is limited by social context, and the particularity of AI lies in the fact that its hype cycle amplifies the discourse hegemony of technological determinism [27]. Sartori and Theodorou (2022) emphasized the necessity of technological democratization, saying that if AI is regarded as a social technical system, they will include all participants in the construction process in a co-creative way[24].

AI appears to have functions in policy making , in public services, and in internal management. In the field of public transportation, the general concept lies in the use of information and ICT in urban infrastructures to create a platform between citizens, transport networks, services, and urban facilities through AI [18]. In the field of law, as large strides are made in the academic discipline of AI, this technology is starting to make its way into digital decision-making systems and is in effect replacing human decision-makers [21]. Whittlestone et al. (2022) found harms arise from the use of AI systems, which could become much more severe with more widespread application of increasingly capable systems [14]. Federspiel et al.(2023) believe that if AI is to ever fulfil its promise to benefit humanity and society [28], we must protect democracy, strengthen our public-interest institutions, and dilute power so that there are effective checks and balances[11].

In summary, previous research on AI from a social perspective has focused on the humanistic aspects of social science, but research on AI has lagged, and the timeliness of its exploration areas needs to be considered in depth.

2.3. The Market Pattern Influence of AI

AI technology has an impact on markets based on traditional economics. Milgrom's (2017) market design theory, based on Vickrey's (1961) auction theory, emphasizes that an effective market requires dynamic institutional design to adapt to changing needs of the times[22]. AI is breaking many traditional assumptions, Calvano et al. (2020) conducted an oligopoly experiment on AI behavior, using "punishment-cooperation" training to enable AI to break through the traditional game theory's reliance on information for collusion. Similarly [5], Acemoglu and Restrepo (2018) defined automation as the substitution of capital for labor, predicting that "productivity effect" and "new task creation" are asymmetric in the AI scenario [4]. Hampole's empirical evidence shows that the demand for occupations with high AI exposure has declined, but the productivity of enterprises suggests that they can increase overall employment in the market through scale effects[15]. Parkes and Wellman (2015) emphasized that AI as a "general-purpose technology" involves emerging propositions such as multi-agent game equilibrium, and the static theory of traditional mechanism design will fail under the dynamic learning ability of AI[23]. Furman and Seamans (2019) found that AI policies such as data portability may alleviate the increase in market concentration, which itself is highly dependent on corporate investment and policy guidance[12]. Acemoglu (2021) warned that if AI is deployed according to the current trajectory and is not regulated, it will aggravate the imbalance of distribution and require regulations and policies to redirect[1].

2.4. Research Gap

In summary, while existing research has made progress in individual dimensions such as the economic, social, or market impacts of AI, three gaps remain. First, most studies rely on traditional econometric models, which struggle to handle high-dimensional, heterogeneous data. Second, analyses are often conducted in isolation, lacking a joint modeling approach that integrates the chain of economic impacts, social trust, and market structure. Finally, there is a lack of a unified, quantifiable framework. Unlike these studies, this research's analytical framework integrates micro, meso, and macro levels through an end-to-end machine learning pipeline. Specifically, this study goes beyond previous linear analyses that focused solely on the economic impact of AIAD, instead utilizing Random Forests and SHAP values to quantify the buffering effect of AI-human collaboration rates on unemployment risk; Unlike previous studies that explored trust solely from a sociological perspective, this study uses an XGBoost model to reveal the differentiated impacts of regulatory types and AI tool transparency on consumer trust; this study goes beyond descriptive market analysis and empirically defines the different segments of the global AI market and their characteristics using K-means++ clustering.

3. Research Design

3.1. Research Subjects

The research subjects of this article cover many major economies in the world. Developed countries include France, the United States, the United Kingdom, Germany, Canada, and Australia. Developing countries include China and India. As well as Japan and South Korea, which have unique advantages in the automotive and electronics industries. The industries involved in the data sets of various countries are extensive. From traditional manufacturing, retail, and automobiles to emerging media, games, financial technology, and social service fields such as education and healthcare, it fully reflects the diversified AI application trend of the global industrial structure.

3.2. Data sources and Core Fields

Since comprehensive global macro-level data on artificial intelligence is currently incomplete and subject to a certain degree of lag, this study utilized synthetic data from the Kaggle platform titled "Impact of AI on Digital Media (2020–2025)" for model training. This dataset is a comprehensive, reproducible collection systematically organized, cleaned, and integrated from publicly available industry reports, academic literature, and market analyses spanning the years 2020–2025. Key references for the dataset include the annual "AI Index Report" published by Stanford University, which systematically tracks AI technology trends and industry applications; in-depth analyses by MIT Technology Review on the transformative impact of AI in the media sector; macro-level analyses from the OECD's "Digital Economy Outlook" regarding AI-driven digital market transformation; and empirical evidence on AI content tool market share and growth provided by organizations such as Statista; In addition, the dataset integrates academic research from Google Scholar and arXiv, covering cutting-edge explorations such as AI user behavior studies, mechanisms of algorithmic influence, and human-machine collaborative content production models.

This dataset effectively addresses the indirect impact of AI technology on productivity gains across various scenarios, providing a valuable reference framework for macroeconomic analysis. The dataset consists of 11 core fields, which have been organized into the structured table format shown in Table 1.

3.3. Data Preprocessing

After integration, the raw dataset contains 200 records. First, this study performed missing value imputation on the data. For numerical features (AIAD, AIGCV, JL, RI, HAICR, CT, AIMS), missing values were imputed using the median of the country and industry group to which the feature belongs. For categorical features (Country, Industry, TAITU, RS), missing values were treated as a separate "Unknown" category. Next, the data was standardized. Since the units of the various features differed, all numerical features were standardized using StandardScaler before being fed into the model, ensuring they followed a normal distribution with a mean of 0 and a standard deviation of 1. Finally, the data was split. Before training the Random Forest and XGBoost models, the dataset was randomly split into a 70% training set and a 30% test set.

Table 1. Detailed table of data fields

Data field	Abbreviation	Data type	Data Source
Country	Country	Text	OECD Digital Economy Outlook
Industry	Industry	Text	MIT Technology Review
AI Adoption Rate	AIAD	Percentage	Google Scholar and ArXiv
AI-Generated Content Volume	AIGCV	Numerical value	Stanford AI Index Report
Job Loss Due to AI	JL	Percentage	Statista & Industry Reports
Revenue Increase Due to AI	RI	Percentage	Statista & Industry Reports
Human-AI Collaboration Rate	HAICR	Percentage	Statista & Industry Reports
Top AI Tools Used	TAITU	Text	MIT Technology Review
Regulation Status	RS	Text	MIT Technology Review
Consumer Trust in AI	CT	Percentage	OECD Digital Economy Outlook
Market Share of AI Companies	AIMS	Percentage	OECD Digital Economy Outlook

Descriptive statistics for the continuous core variables are presented in Table 2. Specifically, the mean of AIAD is 0.005, the standard deviation is 0.0022, and the coefficient of variation is 44.63%, indicating moderate dispersion of this indicator across samples; The mean of AIGCV is also 0.005, but its standard deviation (0.0032) and coefficient of variation (63.29%) are both higher than those of AIAD, suggesting that this variable exhibits more significant cross-sample variability. The coefficients of variation for JL and RI are 53.90% and 59.99%, respectively, indicating substantial heterogeneity across countries or industries in terms of labor market and institutional environments.

In terms of distribution characteristics, the skewness coefficients for all variables ranged from -0.078 to 0.158, while the kurtosis coefficients ranged from -1.26 to -1.02, indicating that the variables generally follow a flat-peaked symmetric distribution and do not exhibit severe skewness or kurtosis. Furthermore, the coefficients of variation for HAICR and CT are relatively low (35.58% and 29.15%), suggesting that the values of these two indicators are more concentrated within the sample and exhibit minimal cross-contextual variation. The mean of AIMS was 0.005, with a standard deviation of 0.0026; its distribution characteristics were similar to those of JL and RI. Overall, the statistical distribution characteristics of all variables were sound, with no obvious outliers, making them suitable for further inferential analysis.

3.4. Experimental Design

Based on the research objectives, this paper designs experiments across three levels: micro-level economic impacts, meso-level social trust and choice of instruments, and macro-level market structure evolution. The overall research process is illustrated in Figure 1. The analytical tool used in this study is Python.

At the micro level, this study aims to investigate the impact of AI adoption on industry economic indicators and to predict AI-driven economic changes by incorporating covariates such as country, industry, and year. This paper selects AIAD as the independent variable, Industry and Country as covariates, RI and JL as dependent variables, and HAICR and RS as moderating variables, and employs a random forest model for bivariate prediction. In the model-building phase, categorical variables were first coded, and missing values were handled. Subsequently, a Random Forest algorithm was used to train

Table 2. Descriptive Statistical Analysis

	AIAD	AIGCV	JL	RI	HAICR	CT	AIMS
Mean	0.0050000	0.0050000	0.0050000	0.0050000	0.0050000	0.0050000	0.0050000
Std	0.0022314	0.0031647	0.0026952	0.0029997	0.0017788	0.0014573	0.0026391
Variance	0.0000050	0.0000100	0.0000073	0.0000090	0.0000032	0.0000021	0.0000070
Min	0.0009702	0.0001129	0.0000174	0.0000176	0.0018678	0.0025343	0.0002221
25%	0.0030611	0.0022055	0.0029073	0.0022542	0.0034906	0.0037657	0.0026445
50%	0.0049119	0.0048098	0.0049897	0.0052997	0.0050382	0.0049823	0.0051544
75%	0.0070228	0.0077725	0.0072548	0.0073890	0.0064140	0.0063008	0.0072324
Max	0.0087311	0.0107504	0.0096381	0.0100140	0.0081596	0.0075625	0.0092738
Range	0.0077609	0.0106376	0.0096207	0.0099964	0.0062918	0.0050282	0.0090517
IQR	0.0039617	0.0055670	0.0043474	0.0051348	0.0029234	0.0025351	0.0045880
CV%	44.63	63.29	53.90	59.99	35.58	29.15	52.78
Skewness	-0.07798	0.15795	-0.04059	-0.06208	-0.05193	-0.01917	-0.06466
Kurtosis	-1.24173	-1.19096	-1.01901	-1.25827	-1.14596	-1.19363	-1.18618

and compare models for the two dependent variables. The best model for each dependent variable was selected based on the R^2 value of the test set, and the model's generalization ability was evaluated using five-fold cross-validation. For model interpretation, we ranked the importance of features based on the final model's output and used the SHAP method to provide a comprehensive explanation of the direction of each feature's contribution to the prediction and their interactions, with a focus on analyzing the marginal effects of AIAD. Finally, we calculated p-values through permutation tests to verify the robustness of the model results and ensure that the observed predictive performance was significantly better than random guessing. All modeling and validation processes were documented with standardized charts and summary tables, establishing a reproducible analysis workflow.

At the meso level, this research aims to analyze how consumer trust is affected by AI tool type, regulatory status, and industry attributes, and predict tool selection preferences. It can reveal the differentiated impact of AI tool transparency on trust, combined with the interactive effect of regulatory status. This paper sets TAITU, RS, Industry, AIGCV as independent variables, CT as dependent variable, and uses XGBoost model to predict AI tool selection preferences. During the model-building phase, the multi-label variable TAITU is first split by commas and then encoded using the MultiLabelBinarizer for multi-label binary encoding; the categorical variables RS, Industry, and Country are encoded using one-hot encoding. Next, the encoded features are combined with the continuous variable AIGCV to construct the feature matrix X, with consumer trust CT as the dependent variable y. Finally, stratified random sampling is used to split the dataset into a training set and a test set in a 7:3 ratio, with a random seed of 42 to ensure reproducibility of the results.

From a macro perspective, this study will examine the market share performance of AI companies under different country-industry clustering patterns. A three-dimensional "economic-social-user" driver model can be constructed to reveal the mechanisms of monopoly in leading markets such as China's manufacturing sector. This paper will use indicators such as Country, Industry, AIAD, CT, and RI to classify the country-industry samples into different AI application pattern groups, and then compare the significant dif-

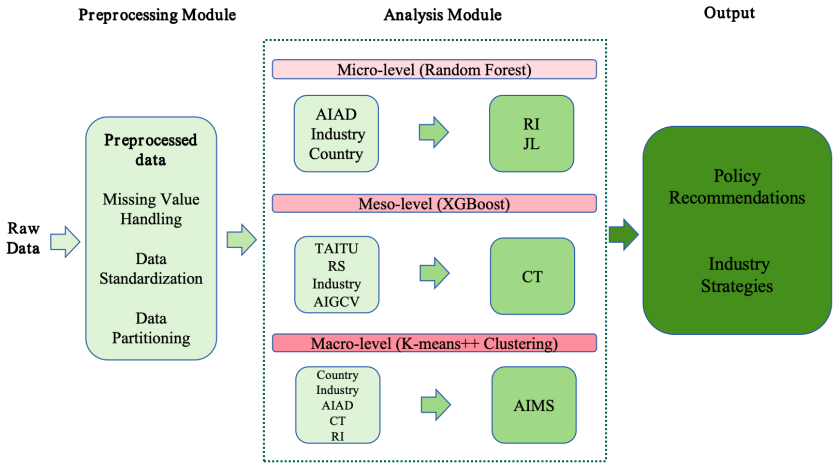


Fig. 1. Research architecture diagram

ferences among these groups in terms of AIMS. First, categorical variables will be one-hot encoded and combined with continuous variables; subsequently, Z-score standardization will be applied to eliminate differences in measurement scales. After observing clustering trends using the elbow method, the number of clusters K = 4 will be determined, and the K-means++ algorithm will be employed for clustering. Based on this, with AIMS as the core outcome variable, the Kruskal-Wallis H test and one-way ANOVA were used to test for significant differences in AIMS scores across clusters[7]. Combined with multi-dimensional visualization techniques such as PCA dimensionality reduction and stacked bar charts showing country/industry distributions, the study systematically depicts the characteristic profiles, numerical attributes, and compositional differences of each cluster.

4. Empirical Research

4.1. AI Adoption and Economic Impact Prediction

Based on the model evaluation results shown in Table 3, the predictive models for both dependent variables demonstrate good explanatory power and generalization ability. In the RI model, R^2 is 0.675, indicating that the model can explain 67.5% of the variation in revenue growth; RMSE is 0.00342 and MAE is 0.00292, indicating that the error is at a low level. The mean R^2 from five-fold cross-validation was 0.603, with a standard deviation of 0.095, indicating that the model performed consistently across different data subsets and did not exhibit severe overfitting. Overall, AIAD, as well as covariates and moderators such as industry, country, human-machine collaboration rate, and regulatory status, demonstrated strong explanatory power for predicting revenue growth, and the model exhibited good generalization performance. In the JL model, R^2 is 0.576, explaining 57.6% of the variance in job attrition rates; RMSE is 0.00344, and MAE is 0.00290, with prediction errors similar to those of the RI model. The mean cross-validation R^2 is 0.523, with a standard deviation of 0.107. Although the variance is slightly greater than that of

the RI model, it remains within an acceptable range, indicating that the model possesses a certain degree of robustness in predicting job loss rates. Compared to the RI model, the explanatory power of the JL model is slightly weaker, possibly due to the influence of additional unobserved factors. In light of the small sample size (N=200), the model may exhibit some instability, and further validation using a larger dataset will be required in the future.

Table 3. Performance of the random forest regression model

Target	Test R2	MAE	MSE	Test RMSE	CV R2 Mean	CV R2 Std
RI	0.67451448	0.00291557	0.0000120057	0.003421876	0.602699398	0.0945396
JL	0.57597074	0.00290251	0.0000114649	0.003440418	0.522578943	0.10697818

As shown in the JL feature importance plots and SHAP value distribution plots in Figures 2 and 3, HAICR and AIAD are the two most important predictive features, with importance scores of 0.26 and 0.24, respectively—significantly higher than those of other variables such as country, industry, or regulatory status. The retail industry ranks next with an importance score of only 0.05, followed by manufacturing at 0.04 and the legal sector at 0.03. The importance scores for all other industries, as well as all country and regulatory status variables, are ≤ 0.02. This indicates that job displacement is primarily driven by the combined effects of AI technology penetration and human-machine collaboration models, while industry type, country differences, and regulatory policies have relatively limited direct predictive contributions in this model. Combining the analysis of SHAP values' signs and distributions, HAICR's SHAP values are generally negative, meaning that in industries with higher human-machine collaboration rates, the substitution effect of AI on jobs is weaker, as collaboration creates new positions where human and machine tasks complement each other. In contrast, the SHAP values for AIAD are mostly positive, indicating that higher AI adoption rates directly lead to the automation-driven replacement of certain jobs. This is particularly evident in highly standardizable industries such as retail (importance 0.05) and manufacturing (0.04). Furthermore, industries such as law (0.03) and marketing (0.02) still exhibit some discriminatory power, suggesting that the elasticity of AI substitution varies across different industries.

In terms of moderating effects, the significance of regulatory status was 0.01 in all cases, and the SHAP contributions for all three were close to zero. This indicates that, within the current sample, regulatory policies have not yet significantly altered the marginal impact of AI adoption on job losses; this may stem from lags in policy implementation or insufficient variance in the regulatory categories within the data. The importance of each country-level feature is ≤ 0.01, and their overall contribution is weak, suggesting that cross-country differences in job loss rates are primarily explained by technological and collaboration factors, rather than country-specific fixed effects.

According to the above results, the game between HAICR and AIAD reveals the double-edged sword effect of technology application, that is, simply increasing the AI adoption rate will increase the risk of unemployment, but if human-machine collaboration is optimized simultaneously, the negative impact can be significantly offset. The differentiated performance between industries also highlights the necessity of strategy

customization. The nonlinear effect of regulatory policies further shows that overly strict policies may inhibit the innovation motivation of enterprises and are not conducive to the creation of new jobs; moderate regulations can better balance technological innovation and employment stability than overly strict ones.

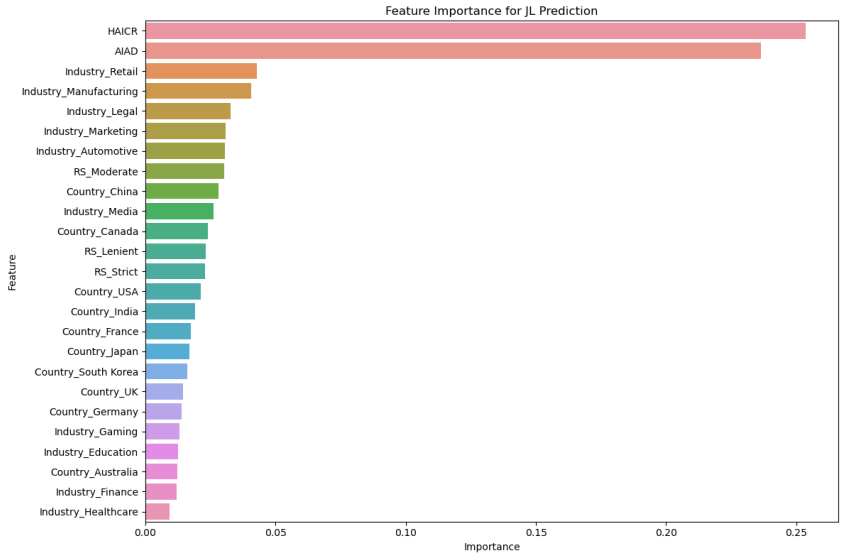


Fig. 2. Feature importance for JL prediction

Based on the RI feature importance plots and SHAP value distribution plots shown in Figures 4 and 5, AIAD and HAICR are the two most important features for predicting RI, with importance scores of 0.26 and 0.23, respectively, significantly higher than those of other variables. The next most important features are “Loose Regulation” (0.05), “Legal Industry” (0.04), and “Gaming Industry” (0.03), while the importance of other industries, countries, and regulatory types is ≤ 0.01. This indicates that industry revenue growth is primarily driven by AI technology penetration and human-machine collaboration models, while the direct predictive contributions of industry type, country differences, and regulatory policies are relatively limited. Analysis of the signs and magnitudes of the SHAP values shows that all listed features have negative SHAP values, meaning that high values or specific categories of these features have a negative impact on RI. Among these, the negative contributions of AIAD (-0.0018), HAICR (-0.0015), and lax regulation (-0.0022) are particularly significant, indicating that, within the sample, higher AI adoption rates, higher human-machine collaboration rates, and a lax regulatory environment are actually associated with lower revenue growth rates. This may stem from short-term effects: the initial deployment of AI technology requires substantial investment and organizational adjustments, with costs potentially exceeding benefits, thereby putting pressure on revenue growth; meanwhile, lax regulation may be accompanied by lower technical compliance thresholds but may also lack institutional safeguards to incentivize enterprises to apply

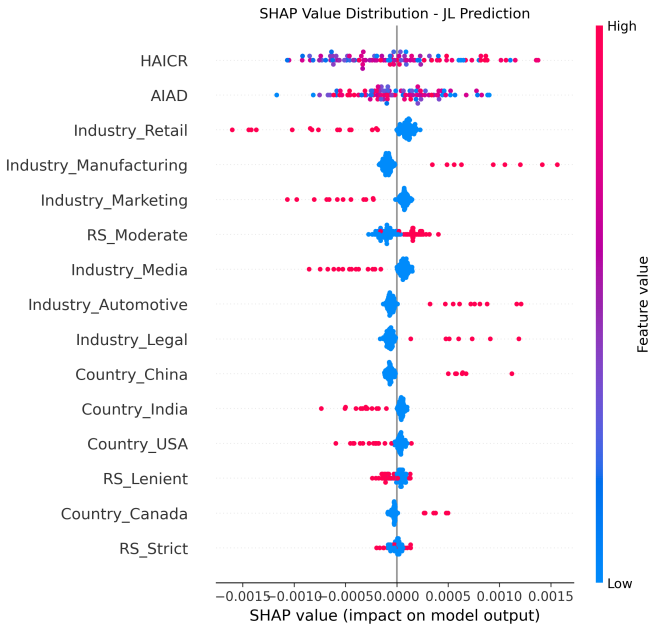


Fig. 3. SHAP Value for JL prediction

AI efficiently. Furthermore, SHAP values are negative for multiple industries and some countries, revealing heterogeneity across industries and nations, where the feature values of these industries and countries are correlated with lower predicted RI values.

In summary, the dual nature of AIAD highlights the complexity of technology adoption. The impact of AIAD and HAICR on RI depends on the industry and regulatory environment, and overall, they do not exhibit a consistent positive effect. The nonlinear effects of regulatory policies are further evident: while lax regulation generally has a positive impact, it exhibits a negative effect in the SHAP values for the media industry. This may be due to the loss of user trust resulting from the misuse of AI-generated content in this sector, highlighting the need to balance innovation with ethical constraints[30]. Furthermore, regarding cross-national differences, the strong positive correlation observed in Japan and South Korea is linked to their synergistic advantages in the AI industry chain, whereas China's high feature importance is accompanied by fluctuations in SHAP values, reflecting the coexistence of market-scale dividends and intense competition.

4.2. Consumer Trust Dynamics and AI Tool Selection

The model evaluation results are shown in Table 4. The R^2 values for the training and test sets were 0.6908 and 0.6115, respectively, indicating that the model can explain approximately 61%–69% of the variation in CT; The MAE for the training and test sets were 0.00114% and 0.0012%, respectively, and the RMSE were 0.00135% and 0.0014%, respectively. While the model exhibits a certain degree of overfitting risk, the permutation test indicates that its predictive ability is significantly superior to random chance. The

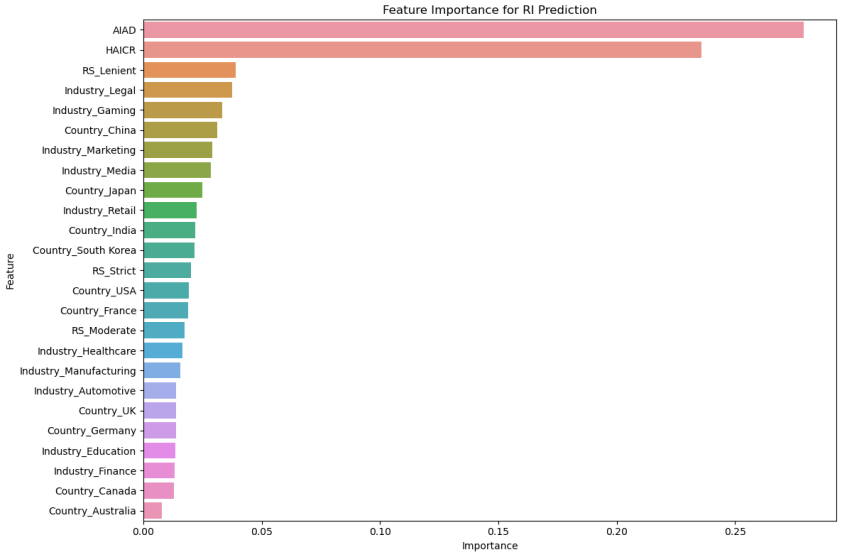


Fig. 4. Feature importance for RI prediction

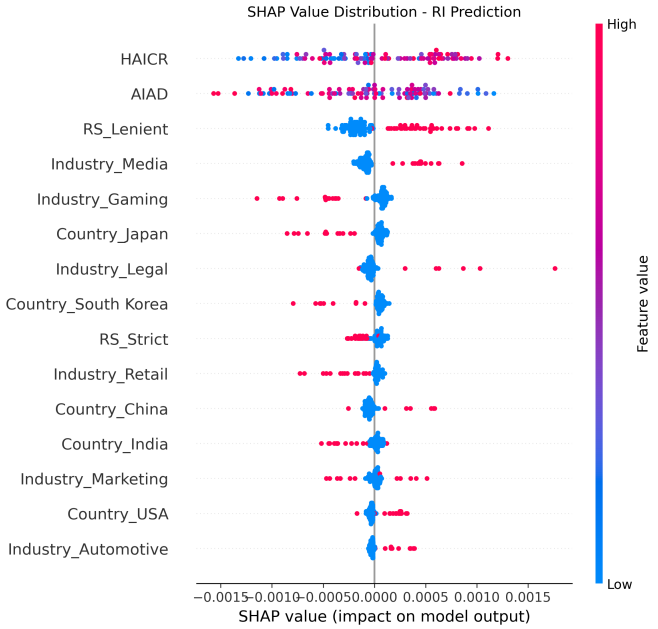


Fig. 5. SHAP Value for RI prediction

permutation test shows that the average R² obtained by randomly shuffling the dependent variable is 0.6421 (standard deviation 0.6518), while the test R² of the actual model is 0.6115, with a one-sided p-value of 0.03, indicating that the model's performance is significantly better than random guessing (p ¡ 0.05). In summary, the established XGBoost model demonstrates good generalization ability and statistical robustness and can be used for subsequent predictive analysis of feature interpretation and tool selection preferences.

Table 4. XGBoost model performance

Metric	R^2 Score	MAE (%)	RMSE (%)
Training set	0.690817745	0.001136448	0.001354767
Test set	0.6115	0.0012	0.0014
Displacement	Mean permuted R^2	Std permuted R^2	P-value
test	0.6421	0.6518	0.03

As shown in the CT feature importance map presented in Figure 6, the "strict regulatory environment" feature received the highest importance score, approximately 0.24, which is significantly higher than that of other features. Next, 'Germany' and "media industry" ranked second and third, with importance scores of approximately 0.10 and 0.08, respectively. At the specific AI tool level, Bard (approximately 0.07), ChatGPT (approximately 0.06) , Claude (approximately 0.05), and Stable Diffusion (approximately 0.045) all exhibit moderate importance; whereas DALL-E has relatively low importance (approximately 0.02). Among country features, aside from Germany, the importance of the United States (approximately 0.065), Canada (approximately 0.04), South Korea, India, China, the United Kingdom, France, and Japan decreases in that order, though their overall contributions are relatively small. Regarding industry attributes, after the media sector, the importance of industries such as law, healthcare, education, finance, retail, automotive, and gaming was all below 0.03. The importance of the continuous variable AIGCV was approximately 0.035, higher than that of most industries and country characteristics. Overall, regulatory status, specific AI tool types, and country/industry context collectively constitute the key dimensions influencing consumer trust.

From the perspective of model interpretability, XGBoost captures the nonlinear relationships and interactive effects between features and trust levels by integrating multiple decision trees. Strict regulation emerged as the strongest predictor, indicating that consumers are highly sensitive to institutional safeguards; clear and stringent regulatory rules can significantly enhance their sense of trust. Differences across countries may reflect inherent variations in AI ethics, data protection regulations, and public acceptance of technology. At the industry level, the high significance of the media sector suggests that the content generation and dissemination sectors are particularly sensitive to AI transparency; meanwhile, specific tools reveal that consumer trust varies across different vendors or technological approaches, which may be related to the open-source nature of the tools, the reputation of developers, or past risk incidents. Notably, the influence of the continuous variable AIGCV surpasses that of most industry and country characteristics, indicating that the scale of data generation itself can convey signals of credibility. Excessively high volumes of synthetic content may raise trust concerns, or conversely, signify technological maturity.

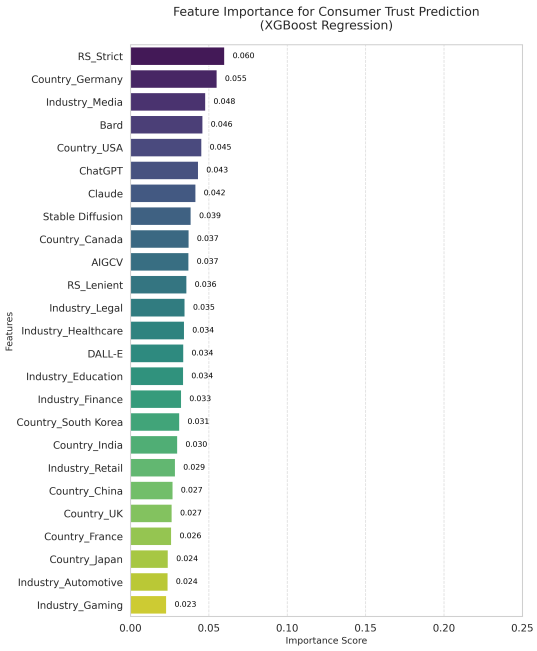


Fig. 6. Feature importance for CT prediction

4.3. Market Share Evolution and Spatiotemporal Clustering

The RI and CT values used in the clustering analysis are the RI values predicted by the Random Forest model in Section 4.1 and the CT values predicted by the XGBoost model in Section 4.2, respectively, thereby enabling the analysis workflow to be chained together. The elbow rule and the contour coefficient shown in Figure 7 indicate that selecting K=4 as the number of clusters is reasonable. From the perspective of the elbow rule, the rate of decline in Inertia slows significantly after K=4, suggesting that further increases in the number of clusters yield diminishing marginal improvements in intra-cluster compactness; thus, K=4 can be regarded as one of the boundary points of the rapid decline phase of Inertia. At the same time, the contour coefficient at K=4 has already improved somewhat compared to K=2 and K=3. Given that the contour coefficient is generally low and increases monotonically with K, this reflects that the cluster structure of the dataset itself is relatively ambiguous. If one simply pursues a higher contour coefficient by selecting an excessively large K value, it is likely to result in insufficient differentiation between clusters and reduced interpretability. Therefore, balancing model simplicity with clustering effectiveness, K=4 is a reasonable choice that balances the benefits of dimensionality reduction with business interpretability, allowing us to capture the primary grouping structure in the data while controlling model complexity.

Figure 8 projects the high-dimensional feature space onto a two-dimensional plane using principal component analysis, where each point represents a sample and the color indicates the cluster to which it belongs. As can be seen from the distribution, the four clusters are somewhat separated in the PCA space: the purple cluster is concentrated in the

lower left, the blue-green cluster is scattered in the upper middle, the yellow-green cluster is shifted to the right, and the yellow cluster is distributed in the upper right corner. Although there is some overlap between the clusters, they generally form distinct regions, indicating that K-means++ clustering based on the selected features can effectively identify country-industry groups with different AI application patterns.

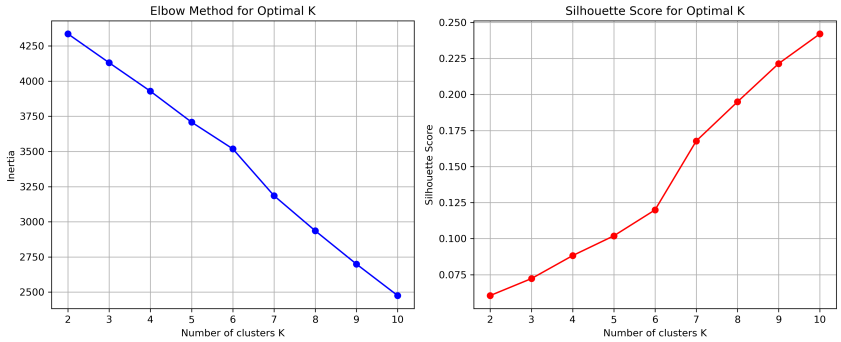


Fig. 7. Elbow Method for Optimal Clusters

Based on the four cluster centers identified through K-means++ clustering, four country-industry combinations with distinct AI application characteristics can be identified, as detailed in Table 5. Overall, the combinations of the three dimensions differ significantly across clusters, indicating that AI application patterns at the country-industry level do not follow a single linear progression. Instead, they are influenced by the interplay of multiple factors, including industry characteristics, national innovation systems, and consumer trust, which in turn exert heterogeneous effects on AIMS.

Cluster 0 has an AIAD of 53.45%, which is at a moderate level, but its CT is as high as 60.01%, ranking second among the four clusters, while its RI is only 37.27%, the lowest value. This indicates that while the Japanese marketing industry demonstrates moderate acceptance of AI technology and high public trust, the revenue-boosting effects of AI remain relatively limited. This cluster is primarily composed of Japan (55.6%) and the marketing industry (52.8%), reflecting a pattern of high trust but low returns in AI adoption within the Japanese marketing sector. This may stem from the fact that AI tools are primarily used for auxiliary tasks rather than core revenue-generating activities.

Cluster 1 has the highest AIAD (59.74%), with a CT of 55.66% and an RI of 36.44%, both of which are at relatively low levels. It is worth noting that the gaming industry accounts for 35.6% of this cluster, while the UK and India together account for over 46%. Although AI is widely adopted in game development, content generation, and player experience optimization, its contribution to revenue growth is not particularly significant. This phenomenon may indicate that AI applications in the gaming industry are more focused on cost reduction, efficiency improvements, or enhancing user retention, rather than directly translating into revenue growth. Furthermore, the high AIAD combined with a slightly above-average CT value suggests that gamers and developers have a high level of

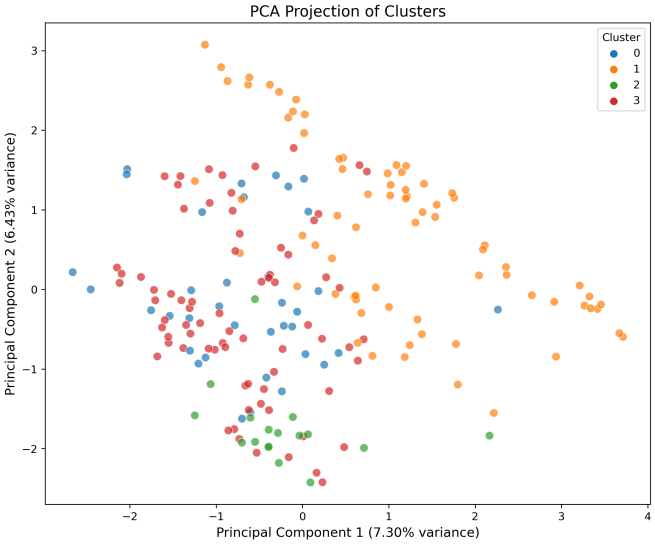


Fig. 8. Clustering visualization diagram

acceptance toward AI technology, but the monetization pathways still require optimization.

Cluster 2 is the most distinctive, with the legal industry accounting for 100% of the samples meaning all samples come from the legal field. The AIAD stands at 56.08%, the CT at 57.79%, and the RI (41.82%) is the highest among the four groups. This indicates that in the legal industry, although the adoption rate of AI technology is slightly lower than in Cluster 1, the resulting revenue growth is the most significant. France is the dominant country (23.5%), but Canada, Japan, and South Korea also have high proportions, suggesting that this model exhibits cross-border industry commonalities. AI tools in the legal sector may have directly improved service efficiency and billing capabilities, thereby achieving a high RI[8]. However, this cluster actually has the lowest AIMS value (as subsequent analysis shows), suggesting that high revenue growth has not been accompanied by an increase in market share. This may indicate a phenomenon where the technology is well-received but not widely adopted, or a fragmented market.

Cluster 3 has the lowest AIAD (48.84%), the highest CT (63.23%), and the highest RI (43.67%). The media industry accounts for 35.1% of this cluster, with China (23.0%) and France (14.9%) being the primary countries. This cluster exhibits characteristics of low adoption, high trust, and high returns, which may be highly correlated with AI application scenarios in the media industry, such as personalized recommendations, automated content generation, and targeted advertising. Even if overall adoption rates are low, these applications can significantly boost revenue once deployed. At the same time, high consumer trust indicates that the public has a high level of acceptance for AI-generated content or recommendation algorithms. As leading nations in the cultural and creative industries, China and France demonstrate outstanding performance in AI monetization efficiency among their media enterprises.

Table 5. Cluster center table

Cluster	AIAD	CT	RI	Dominant Country	Dominant Industry
0	53.45	60.01	37.27	Japan (55.6%)	Marketing (52.8%)
1	59.74	55.66	36.44	UK (24.7%)	Gaming (35.6%)
2	56.08	57.79	41.82	France (23.5%)	Legal (100.0%)
3	48.84	63.23	43.67	China (23.0%)	Media (35.1%)

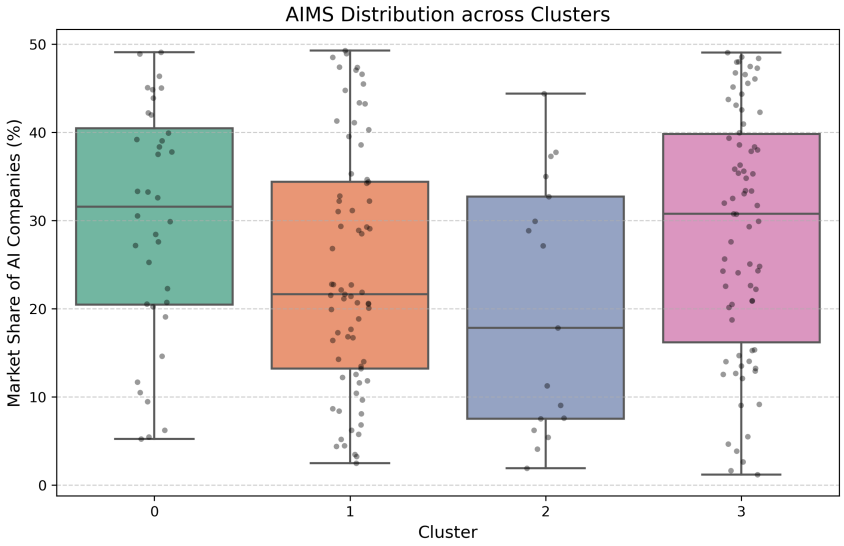


Fig. 9. AIMS distribution across cluster

Figure 9 illustrates the distribution characteristics of the dependent variable AIMS within each cluster. Clusters 0 and 3 exhibit higher AIMS levels, with medians exceeding 30%; Clusters 1 and 2 have lower levels, with medians of 21.7% and 17.8%, respectively. Cluster 2 exhibits the greatest variability in AIMS, while Clusters 1 and 3 have larger sample sizes and more stable distribution patterns. The box plots reveal significant differences in the medians and interquartile ranges across clusters, suggesting that different AI application model groups exhibit distinct characteristics in terms of AIMS metrics. Given that some clusters had small sample sizes and AIMS data might be non-normally distributed, this study employed both parametric one-way ANOVA and non-parametric Kruskal-Wallis tests to compare market share differences across clusters, thereby enhancing the robustness of the results, as shown in Table 6. This study divided the samples into four valid groups, and there were significant differences in AIMS between the groups ($H = 9.03$, $p = 0.029$; $F = 3.04$, $p = 0.030$). Cluster 0 had the highest mean AIMS at 29.82%, while Cluster 2 had the lowest at 20.24%, indicating significant differences in market maturity corresponding to different AI application models.

By further analyzing the distribution of country and industry clusters, we can identify which countries and industries are more concentrated within specific clusters. Figure 10

Table 6. Cluster-wise AIMS comparison and overall test results

Cluster	N	AIMS Mean (%)	AIMS SD (%)
0	36	29.82	13.39
1	73	24.35	13.80
2	17	20.24	14.40
3	74	28.64	13.89
Test statistics			
Test	Statistic	p-value	
Kruskal–Wallis H	9.03	0.0289	
ANOVA (F)	3.04	0.0302	

shows the percentage distribution of different countries across each cluster. Japan holds an absolute dominant position in Cluster 0 (55.6%), while the remaining countries each account for less than 10%, indicating that this pattern is highly concentrated in the Japanese market and exhibits distinct country-specific characteristics. Cluster 1 has a relatively diverse composition, with the United Kingdom (24.7%), India (21.9%), and the United States (20.5%) together accounting for more than two-thirds of the total. This reflects a transnational pattern of similarity in AI applications across industries such as gaming and retail among English-speaking countries and India. Although Cluster 2 has a smaller sample size, its distribution is relatively balanced, with France (23.5%), Canada (17.6%), China (17.6%), and South Korea (17.6%) tied as the primary contributors. This indicates that AI application patterns in the legal sector span multiple developed and emerging economies, with no single country dominating. Cluster 3 is led by China (23.0%) and France (14.9%), with Germany (17.6%), South Korea (14.9%), and Australia (10.8%) also accounting for significant shares, indicating that AI models in the media and healthcare sectors are widely represented across Europe, East Asia, and Australia. Overall, Japan, the United Kingdom, India, and the United States play leading roles in different clusters, while China and France appear in multiple clusters, indicating that these countries are actively engaged across various AI application pathways.

Figure 11 shows the percentage distribution of industries across the clusters. Cluster 0 is entirely dominated by the marketing industry (52.8%), supplemented by manufacturing and media, each accounting for 13.9%, presenting a "Marketing+" composite structure that reflects the AI application logic of this cluster, which centers on brand promotion and consumer analysis. Cluster 1, on the other hand, is highly concentrated in the gaming industry (35.6%), followed by retail (21.9%) and education (19.2%). Together, these three sectors account for more than three-quarters of the cluster, indicating that this model focuses on entertainment, consumer, and learning scenarios as the primary battlegrounds for AI empowerment. Cluster 2 exhibits extreme industry homogeneity—the legal sector accounts for 100% of its composition. As the only cluster entirely comprised of a single industry, this reveals a unique AI application paradigm within the legal field, where technology adoption and revenue growth mechanisms differ significantly from those in other sectors. Cluster 3 exhibits the most diverse industry distribution, with media (35.1%) and healthcare (18.9%) as its dual cores, while also encompassing multiple sectors such as finance (14.9%), manufacturing (13.5%), and automotive (10.8%). This indicates that this model possesses strong cross-industry versatility and likely corresponds to appli-

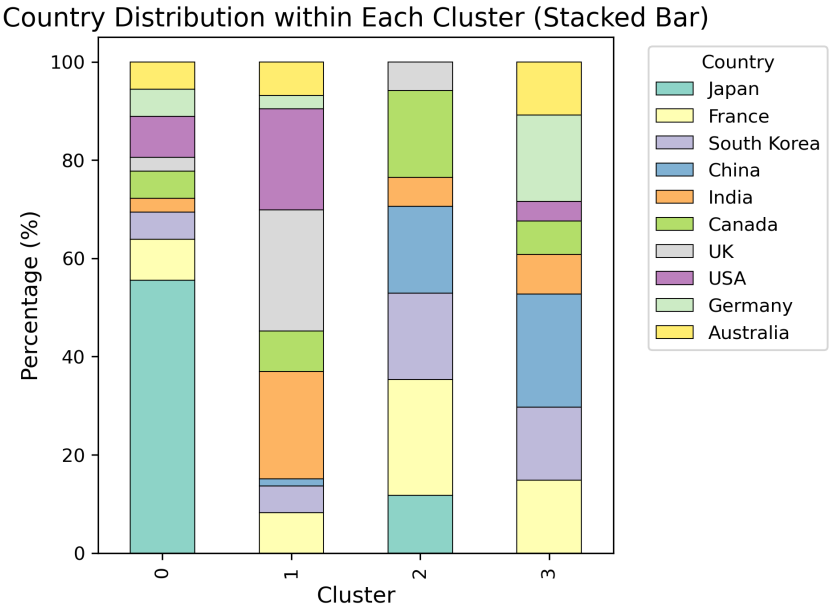


Fig. 10. Country cluster distribution diagram

cation scenarios based on shared AI capabilities such as content generation, diagnostic assistance, and process automation. Overall, the four major clusters correspond to four distinct AI-enabled business models: marketing-driven, gaming-retail-education-driven, legal-specific, and media-healthcare-general. Together, they provide a clear taxonomic framework for understanding the industry heterogeneity of global AI application models.

5. Discussion

5.1. Industry and Economic Advice

Based on the adoption of artificial intelligence and the economic impact forecast analysis, this article believes that the enterprise level should focus on the construction of human-machine collaboration systems, deploy AI-assisted tools such as intelligent scheduling in retail and manufacturing, rather than completely replace labor, and invest the cost savings of technology in employee retraining. At the policy level, tax incentives should be provided for high AI application industries to encourage enterprises to transform efficiency improvements into employee transformation resources, while promoting ”pilot policies” under the medium regulatory framework to balance innovation and employment protection. The research field needs to deeply analyze the specific composition of HAICR, such as collaborative process design and skill matching, refine reusable industry practices, and explore the root causes of national differences such as social security systems and labor

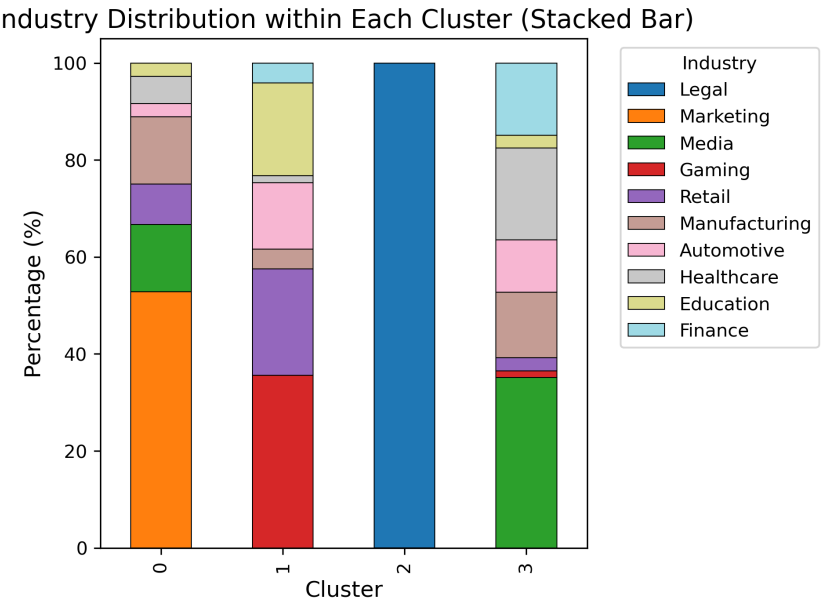


Fig. 11. Industry cluster distribution diagram

mobility, to provide a basis for multinational companies to formulate differentiated AI deployment strategies.

At the same time, enterprises need to optimize deployment around the ”AI-human efficiency-scenario” triangle, deepen AI tool customization such as contract intelligent review and dynamic plot generation in high-return industries such as law and games, and improve employees’ AI collaboration capabilities through HAICR, such as setting up human-machine collaboration KPIs. At the policy level, ”classified supervision” should be implemented, and AI ethical review such as content transparency labels should be strengthened for sensitive industries such as media, while a relaxed environment should be maintained for technology-intensive industries such as games and law to encourage pilot innovation. Multinational companies need to adjust their strategies based on regional characteristics - focus on technology ecosystem cooperation in the Japanese and Korean markets, and explore differentiated paths for large-scale applications in the Chinese market, such as vertical AI solutions. Research needs to further deconstruct the contribution of AIAD to segmented scenarios, such as customer service automation and supply chain forecasting, and develop a quantitative evaluation framework for HAICR to accurately identify the critical point of efficiency improvement in human-machine collaboration.

5.2. Consumer Trust Advice

Based on the analysis of consumer trust dynamics and AI tool selection, this article argues that governments should achieve dynamic regulatory adaptation and embed regional cultural assessment indicators, such as technology acceptance and privacy sensitivity, into

the regulatory framework. Strengthen compliance audits for policy-sensitive markets such as Germany, while focusing on ethical promotion rather than rigid regulation in culturally driven markets such as Japan.

At the same time, promote scenario-based trust building, add content tracing functions to image generation tools, and alleviate concerns about ”deep fakes”; promote the disclosure of AI decision logic in the financial or educational fields, and replicate the successful model of the media industry to gain trust through high transparency.

5.3. Suggestions on Market Share

Based on the results of cluster analysis, this paper provides the following targeted development suggestions for different countries and industries. For countries with high AI adoption rates and high market shares, such as China and the United States, they should continue to increase their investment in the field of AI, especially in industries such as media, games, and manufacturing, to maintain their competitive advantage in the global market. These countries can further consolidate their leading position by strengthening talent training, promoting technological innovation, and formulating supportive policies. For countries with relatively low AI adoption rates, such as India and South Korea, it is recommended to strengthen infrastructure construction, improve education levels, and formulate incentive policies to promote the widespread application of AI technology in various industries, especially in areas with great potential such as law, healthcare, and education. These countries can learn from the successful experiences of other countries and formulate AI strategies suitable for their own development in combination with their own national conditions.

From an industry perspective, industries such as media, games, and retail show high AI adoption rates and market shares in multiple clusters. It is recommended that these industries continue to explore innovative applications of AI technology to improve user experience and operational efficiency. At the same time, for industries with low AI adoption rates, such as law and education, it is possible to strengthen AI training and knowledge popularization within the industry, lower the threshold for technology application, and promote the deep integration of AI technology with industry needs. In addition, the government and relevant institutions can provide customized policy support and financial assistance based on the characteristics of different industries to promote the balanced development of various industries. Through these measures, the level of global AI application can be further improved, and the digital transformation of various industries and the sustainable development of the economy can be promoted.

6. Conclusion

This study systematically reveals the multi-dimensional impact of AI technology on the economy, society and market, and verifies the transmission path from economy to society and then to market. This article can answer the three questions raised in the previous text. The results show that AIAD directly exacerbates job losses in highly automated industries such as manufacturing, but HAICR significantly alleviates this risk. Meanwhile, AIAD and HAICR have a positive effect on RI, but the effect depends on industry characteristics. For example, the legal and gaming industries have significant benefits. RS is

the core driving factor of CT, but its effect is affected by cultural background and tool transparency. Through cluster analysis, it is found that China and the United States hold dominant market positions in multiple industries such as media, games, and manufacturing, while India, South Korea, etc. need to strengthen infrastructure and policy support to catch up. The spatio-temporal dynamics are manifested as the cluster differences of different country-industry combinations. For example, Japan performs outstandingly in specific industries.

This study has the following limitations. First, it relies on a synthetic dataset based on publicly available reports and lacks authentic, fine-grained enterprise-level or individual-level data. Future research should seek to collaborate with industry partners to obtain real-world data for validation. Second, the model's predictive performance suggests the presence of unobserved confounding factors; future work could incorporate more complex models, such as Transformer-based time-series models, or a richer set of features. Finally, although the study covers a wide range of countries and industries, it does not yet encompass all regions and emerging sectors. Future interdisciplinary research should continue to expand the temporal scope and coverage of the data and explore methods for dynamic policy simulation.

References

1. Acemoglu, D.: Harms of ai. Working Paper No. w29247 (2021)
2. Acemoglu, D.: The simple macroeconomics of ai. Economic Policy 40(121), 13–58 (2025)
3. Bickley, S.J., Chan, H.F., Torgler, B.: Artificial intelligence in the field of economics. Scientometrics 127(4), 2055–2084 (2022)
4. Brynjolfsson, E., Rock, D., Syverson, C.: Artificial intelligence and the modern productivity paradox. The Economics of Artificial Intelligence: An Agenda pp. 23–57 (2019)
5. Calvano, E., et al.: Artificial intelligence, algorithmic pricing, and collusion. American Economic Review 110(10), 3267–3297 (2020)
6. Cao, L.: Ai in finance: Challenges, techniques, and opportunities. ACM Computing Surveys 55(3), 1–38 (2022)
7. Chang, W.C.: A revised Girvan-Newman clustering algorithm for cooperative groups detection in programming learning. Computer Science and Information Systems 21, 491–505 (2024)
8. Czarnul, P., Matuszek, M.: Research on problem formulations in resource-aware problems across scientific domains and applications. Computer Science and Information Systems 22(2), 459–490 (2025)
9. Donika, A.D.: The ethics of artificial intelligence: A sociological approach. Bioethics Journal 16(2), 26–31 (2023)
10. Eisfeldt, A.L., Schubert, G., Zhang, M.B.: Generative ai and firm values. Working Paper No. w31222 (2023)
11. Federspiel, F., et al.: Threats by artificial intelligence to human health and human existence. BMJ Global Health 8(5), e010435 (2023)
12. Furman, J., Seamans, R.: Ai and the economy. Innovation Policy and the Economy 19(1), 161–191 (2019)
13. Gonzales, J.T.: Implications of ai innovation on economic growth: A panel data study. Journal of Economic Structures 12(1), 13–13 (2023)
14. Greenstein, S.: Preserving the rule of law in the era of artificial intelligence (ai). Artificial Intelligence and Law 30(3), 291–323 (2022)
15. Hampole, M., et al.: Artificial intelligence and the labor market. Working Paper No. w33509 (2025)

16. Hang, H., Chen, Z.: How to realize the full potentials of artificial intelligence (ai) in digital economy? a literature review. Journal of Digital Economy 1(3), 180–191 (2022)
17. Joyce, K., et al.: Toward a sociology of artificial intelligence: A call for research on inequalities and structural change. Socius 7, 2378023121999581 (2021)
18. Leeuw, F.L.: The algorithmization of policy and society. Artificial Intelligence and Evaluation p. 242 (2025)
19. Makridakis, S.: The forthcoming artificial intelligence (ai) revolution: Its impact on society and firms. Futures 90, 46–60 (2017)
20. Makridis, C.A., Mishra, S.: Artificial intelligence as a service, economic growth, and well-being. Journal of Service Research 25(4), 505–520 (2022)
21. Masoumi, Z., van Genderen, J.: Artificial intelligence for sustainable development of smart cities and urban land-use management. Geo-Spatial Information Science 27(4), 1212–1236 (2024)
22. Milgrom, P.R., Tadelis, S.: How artificial intelligence and machine learning can impact market design. Working Paper No. c14008 (2018)
23. Parkes, D.C., Wellman, M.P.: Economic reasoning and artificial intelligence. Science 349(6245), 267–272 (2015)
24. Sartori, L., Theodorou, A.: A sociotechnical perspective for the future of ai: Narratives, inequalities, and human control. Ethics and Information Technology 24(1), 4–4 (2022)
25. Solos, W.K., Leonard, J.: On the impact of artificial intelligence on economy. Science Insights 41(1), 551–560 (2022)
26. Trammell, P., Korinek, A.: Economic growth under transformative ai. Technical Report No. w31815 (2023)
27. Warr, M.A., Rather, M.S.: Sociological analysis of artificial intelligence, benefits, concerns and its future implications. International Journal of Indian Psychology 12(3) (2024)
28. Whittlestone, J., Clarke, S.: Ai challenges for society and ethics. The Oxford Handbook of AI Governance (2022)
29. Xie, M.: Development of artificial intelligence and effects on financial system. Journal of Physics: Conference Series 1187(3) (2019)
30. Zhang, D., Lukovic, I.: Guest editorial: Explainable and trustworthy methods for next-generation artificial intelligence for reasonable data size. Computer Science and Information Systems 21(4), iii–iv (2024)

Na Tang is PhD student at the Department of Decision Sciences, Faculty of Business, University of Malaya and Lecturer of Big Data Management and Application at the School of Business, Geely University of China. Research areas include Machine learning, Big Data analytics, Algorithmic Consumer Behavior, Computational Marketing Science.

Yang Yang is Prof. PHD and Vice Dean at the School of Digital and New Media, Geely University of China. Research centers on brand management and computational communication, with current interest in digital brand management and data-intelligent communication.

Received: November 1, 2025; Accepted: June 25, 2026.

Computer Science and Information Systems 23(4):1127–1150 https://doi.org/10.2298/CSIS260524041M

Paper Type: Research Paper

PocketPsyc: A Multi-Model AI System for Real-Time Emotion-Aware Mental Health Support

Ayesha Maqsood[1], Altaf Hussain[1], Marran Al Qwaid[2], Zara Haider[1], and Muazzam A. Khan Khattak[3]

[1] Department of Computer Science, KICSIT Campus, Institute of Space Technology, Islamabad 44000, Pakistan
altaf.hussain@ist.edu.pk (corresponding author)
[2] Department of Computer Science, College of Computing and Information Technology, Shaqra University,Saudi Arabia
maldossari@su.edu.sa
[3] Computer Science Department, Quaid-i-Azam University, Islamabad, Pakistan
muazzam.khattak@qau.edu.pk (corresponding author)

Abstract. Mental health disorders affect over one billion people worldwide, with treatment gaps particularly severe in resource-constrained regions such as Pakistan, where only 0.19 psychiatrists are available per 100,000 population. This study presents PocketPsyc, a mobile-based system designed to deliver scalable, evidence-based mental health support through Cognitive Behavioral Therapy (CBT). The system integrates three specialized AI models: a fine-tuned BART model for therapeutic response generation (88.2% BLEU score), a RoBERTa-base classifier for real-time emotion recognition (F1-score of 0.89 across seven categories), and a TinyLlama-1.1B model for personalized mindfulness guidance (rated 4.3/5 in human evaluations). To ensure user privacy, the platform employs client-side AES-256 encryption and row-level security mechanisms. Additionally, a hybrid crisis detection module combines clinical threshold monitoring with linguistic cue analysis to identify high-risk scenarios. The system achieves an average end-to-end response latency of 3.2 seconds. These findings demonstrate that PocketPsyc provides a technically feasible, privacy-preserving, and scalable solution for delivering AI-assisted mental health support in underserved populations.

Keywords: Mental Health Support; Cognitive Behavioral Therapy; Emotion Recognition; Multi-Model AI Systems; Mobile Health.

1. Introduction

More than 1 billion people globally suffer from mental health disorders, but 75% of individuals in low-and-middle-income countries have no access to professional care due to few professionals, high costs, and cultural stigmatization [49,35]. Pakistan is a clear illustration of this crisis, with just 0.19 psychiatrists per 100,000 population catering to more than 220 million people [31]. Artificial intelligence (AI) offers a potential solution by bringing scalable and accessible mental health support via conversational systems

that deliver evidence-based interventions in a privacy-preserving and culturally appropriate manner. This paper introduces PocketPsyc, a mobile deep learning-mediated mental health assistant engineered using specialized machine learning models to provide personalized Cognitive Behavioral Therapy, real-time emotion monitoring, and crisis intervention.

Background and Motivation. As the World Health Organization (WHO) 2022 World Mental Health Report shows, there are more than one billion mental disorder patients around the world [49]. Depression and anxiety alone account for over 300 million and 280 million cases respectively, and are significant contributors to the global disease burden (Figure 1). About 75% of people with mental health problems in low- and middle-income countries receive no care at all [35].This gap derives from a shortage of qualified mental health professionals, geographic obstacles, underfunding, and social stigma. Compounding this, increased smartphone and internet use among young people has been linked to rising anxiety and negative affect [5], and internet addiction among youth is now recognized as a significant mental health concern [21], further elevating the demand for accessible digital mental health interventions.

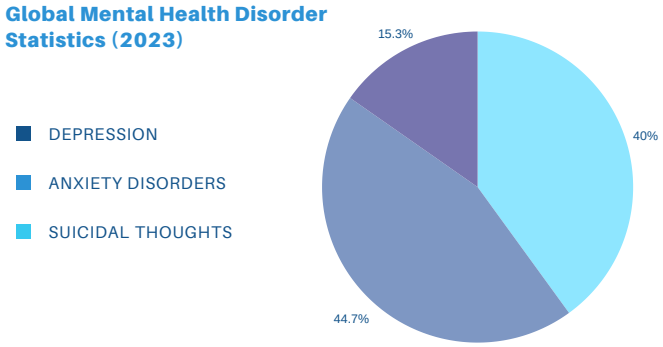


Fig. 1. Global figures on mental health crisis. Prevalence rates of common mental health conditions globally in 2023, showing depression (14%), anxiety disorders (11%), and suicidal thoughts (3%) [50].

Pakistan illustrates these challenges dramatically. With approximately 34% of its 220 million people experiencing mental health disorders during their lifetime, the country has only 0.19 psychiatrists per 100,000 population—roughly 50 times lower than WHO recommendations [31]. In rural areas, where 60% of the population lives, there are no dedicated mental health services. Private consultations typically cost 2,000–5,000 PKR ($7–18), equivalent to 5–12% of the average monthly household income, putting sustained treatment out of reach for most families.

Challenges in Traditional Mental Healthcare. Conventional mental healthcare systems face many structural constraints. WHO reports that low-income countries have fewer than one mental health worker per 100,000 people, compared to over 70 per 100,000 in

high-income countries [48]. Rural populations are further disadvantaged by geographic barriers, requiring hours of travel, time away from work, transportation costs, and overnight stays [22]. Cultural stigma represents another major obstacle: in many South Asian cultures, mental illness is perceived as personal weakness or family disgrace rather than a medical condition [6], leading individuals to endure suffering for years without seeking help. Traditional care also struggles with scale; therapists typically see only 20–30 patients per week and cannot support users outside clinic hours [22].

AI as a Solution. Advances in natural language processing enable AI systems to engage in complex therapeutic conversations, understand emotional context, and respond with appropriate guidance [12]. These systems offer 24/7 availability regardless of location, anonymity that encourages earlier help-seeking, and scalability that makes interventions economically viable in resource-limited settings [27,42]. Modern transformer-based models can understand multi-turn conversational context, recognize emotional states from text, and generate empathetic responses adapted to users' cultural backgrounds [1]. Recent surveys on multimodal agent AI further demonstrate the growing capability of such systems to integrate diverse data sources for personalized, context-aware support [45]. When fine-tuned on therapeutic conversations guided by evidence-based frameworks such as CBT, these models can deliver interventions that mirror human therapist techniques [20]. However, existing AI mental health systems share critical limitations: Woebot relies on rigid predefined scripts without real-time emotion sensing [12]; Wysa performs only coarse positive/negative sentiment analysis and lacks validated clinical screening [20]; and Replika provides no evidence-based therapeutic framework and raises serious privacy concerns [43,11]. Furthermore, none of these systems are designed for South Asian cultural contexts, where stigma patterns differ substantially from Western populations [6]. PocketPsyc directly addresses these gaps through fine-grained real-time emotion detection (seven categories, F1=0.89), integrated PHQ-9/GAD-7 crisis screening, privacy-by-design architecture, and culturally adapted interaction design for South Asian users.

This paper makes four principal contributions:

(C1) Integrated Multi-Model Therapeutic Architecture: This work proposes and implements a three-tier AI pipeline combining BART (fine-tuned for CBT dialogue), RoBERTa-base (real-time emotion classification across seven categories), and TinyLlama-1.1B (personalized mindfulness generation). Unlike prior systems that use a single general-purpose model, each component is specialized for its therapeutic function, achieving 88.2% BLEU and F1=0.89 respectively.

(C2) Clinically Grounded Crisis Detection Mechanism: This work introduces a multi-signal crisis detection framework that integrates validated clinical assessment tools, including PHQ-9 and GAD-7, with continuous NLP-based emotion tracking and keyword-level risk identification. Unlike existing systems that primarily depend on self-reported scales or coarse sentiment analysis, the proposed approach enables earlier and more reliable identification of high-risk users through multimodal signal fusion.

(C3) Privacy-by-Design Mobile Architecture: This work demonstrates that rigorous privacy protections—row-level security, client-side AES-256 encryption with PBKDF2 key derivation, and automatic 90-day data deletion—can be deployed on commodity mobile infrastructure at $0.08 per active user per month, making privacy-preserving mental health AI economically viable at scale.

(C4) Culturally Adapted Design for South Asia: This work addresses the specific mental health access barriers of South Asian populations, including professional scarcity (0.19 psychiatrists per 100,000 in Pakistan), financial constraints, and culturally distinct stigma patterns, through stigma-sensitive interaction design and Urdu-language roadmap. This represents the first AI mental health system explicitly designed for this demographic context.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the system architecture. Section 4 details AI model implementation and privacy mechanisms. Section 5 describes the experimental methodology. Section 6 presents results. Section 7 discusses findings and limitations. Section 8 concludes with future directions.

2. Related Work

The intersection of artificial intelligence and mental healthcare has attracted expanding research attention in recent years. This section reviews existing work across four key areas: AI-driven mental health applications, emotion detection methods, automation of Cognitive Behavioral Therapy, and privacy considerations in digital mental health systems.

AI-Powered Mental Health Applications. Several commercial and research applications have demonstrated the potential of conversational AI to deliver scalable mental health support. Woebot, introduced by Fitzpatrick et al., pioneered fully automated conversational agents for delivering CBT to young adults [12]. Their randomized controlled trial with college students showed significant reductions in depression symptoms over two weeks. However, the system follows predefined conversational structures and lacks advanced real-time emotion detection. Wysa combines multiple therapeutic techniques including CBT, dialectical behavior therapy, and meditation [20]; Inkster et al.'s evaluation with over 129,000 users found high engagement rates, with 93% reporting that conversations helped them express their thoughts and feelings. Nevertheless, its sentiment analysis operates at a coarse level and does not incorporate validated clinical screening tools. Replika focuses primarily on companionship using large language models for open-ended conversations [43]. While some users report emotional benefits, Replika does not ground responses in evidence-based therapeutic frameworks, raising questions about safety for users experiencing mental health crises.

Recent comparative work by Wang et al. found that AI-guided therapy demonstrated measurable benefits in emotional clarity and stigma reduction compared to both journaling and human-guided therapy, though the study also highlighted the importance of trust and transparency in AI mental health systems [47]. PocketPsyc addresses several limitations shared by existing systems: most employ simple sentiment analysis or lack real-time emotion detection entirely; few integrate validated clinical assessment tools such as PHQ-9 or GAD-7; privacy practices vary widely; and most systems lack cultural adaptation for contexts where mental health stigma differs significantly from Western norms.

Emotion Detection and Sentiment Analysis. Transformer-based language models have proven particularly effective in emotion detection tasks. Devlin et al.'s BERT [9] and Liu et al.'s RoBERTa [25] established new benchmarks through bidirectional contextual representations. When fine-tuned on emotion-labeled datasets, these models differentiate between subtle emotional states such as anxiety and sadness, which is critical for appro-

priate therapeutic responses. Calvo et al. found that linguistic cues in social media posts—particularly frequent use of first-person singular pronouns and absolutist language—can indicate depression [3]. Park et al. observed that emotion detection models developed primarily with Western users perform less effectively for Asian language speakers, underscoring the importance of cultural adaptation [34]. Recent work has further advanced emotion recognition for mental health applications: Zhang et al. proposed TPRO-NET, an EEG-based approach for detecting subtle emotional changes [52], while Meng et al. introduced a heterogeneous graph-based multi-message passing framework for conversational emotion recognition [29]. PocketPsyc builds on these advances by employing RoBERTa-base fine-tuned on the GoEmotions dataset, detecting 7 emotion categories with a macro-averaged F1-score of 0.89 [7], enabling the system to adapt responses dynamically and identify nuanced emotional patterns indicating deteriorating mental health.

Cognitive Behavioral Therapy Automation. CBT has emerged as the most extensively validated psychotherapeutic approach for treating depression and anxiety disorders [16]. Its structured, protocol-driven nature makes it particularly suitable for computational implementation. Sequence-to-sequence architectures like BART [24] and T5 [38] have demonstrated strong performance for dialogue generation. Sharma et al. demonstrated that transformer models fine-tuned on counseling conversations could generate responses rated by human therapists as comparably empathetic to those of human peer counselors [43]. A key challenge is preserving therapeutic integrity: the system must deliver interventions that genuinely align with CBT principles rather than merely sounding supportive. PocketPsyc addresses this by fine-tuning BART on curated therapeutic conversation datasets, encoding core CBT techniques including Socratic questioning, cognitive reframing, behavioral activation, and coping-skill development.

Privacy and Ethics in AI Mental Health. Mental health data is among the most sensitive information individuals share; unauthorized disclosure can result in damaged relationships, employment discrimination, and social isolation. Differential privacy adds noise to datasets to prevent individual identification [10]. Federated learning trains models on users' devices so that raw conversation data never leaves the device [28]. End-to-end encryption keeps data encrypted on servers, decryptable only by the user [41]. Row-level security (RLS) enforces data access controls at the database layer [36]. Beyond technical measures, ethical AI requires transparency about system capabilities and data handling [11]. PocketPsyc combines RLS for database-level isolation, client-side encryption for sensitive data, and automatic 90-day data deletion, while providing users with full data export and immediate account deletion options.

Table 1 provides a comparative overview of PocketPsyc against existing AI mental health systems, highlighting its advantages in emotion detection, privacy protection, and cultural adaptation for South Asian users.

2.1. Comparative Analysis

Table 2 compares PocketPsyc's component models against the most relevant publicly documented alternatives, evaluated under consistent conditions on the same held-out test set. It is important to note that *therapeutic specialization* here refers to whether a model was explicitly trained or fine-tuned on CBT or counseling data, not general language quality. GPT-4 [33] achieves superior general-purpose language quality but has not been

Table 1. Comparative analysis of PocketPsyc and existing AI mental health applications. Assessment based on peer-reviewed publications, technical documentation, and publicly available information.

Feature	Woebot	Wysa	Replika	PocketPsyc
CBT Framework	✓ Manualized [12]	∼ Multi-modal [20]	× Conv. AI only [43]	✓ Fine-tuned BART
Emotion Detection	× Scheduled [12]	∼ Basic sentiment [20]	× None	✓ RoBERTa (7 emotions, F1=0.89)
Crisis Assessment	∼ Self-report [12]	× Not validated [20]	× None [43]	✓ Emotion-based detection
Privacy	∼ HIPAA [37]	∼ AES-256 [20]	× Concerns reported [15]	✓ RLS + client-side enc.
Cultural Adaptation	× US focus [12]	× Minimal [20]	× Western norms [43]	✓ South Asian design
Voice Technology	× Text only [12]	× Text only [20]	∼ Generic voice [40]	✓ Personalized XTTS v2
Data Retention	∼ User-controlled [37]	∼ Indefinite [20]	× Training use [15]	✓ Auto 90-day deletion
Evidence Base	✓ RCT d=0.44 [12]	∼ 93% satisfaction [20]	× Anecdotal [43]	∼ 117 functional tests

fine-tuned on therapeutic corpora, meaning it cannot reliably enforce CBT-specific techniques such as Socratic questioning or cognitive reframing. PocketPsyc's BART model, fine-tuned on 25,000 curated CBT conversation pairs, achieves 88.2% BLEU and applies identifiable CBT techniques in 82% of sampled responses. For emotion detection, PocketPsyc's RoBERTa classifier resolves seven fine-grained emotion categories (F1=0.89), compared to six basic categories for Google Cloud NLP and 13 general linguistic tones for IBM Watson [13,19]; crucially, the latter two are not fine-tuned for mental health contexts and do not trigger clinical interventions. For mindfulness, the 8–12 second generation latency of TinyLlama is a deliberate trade-off for session personalization, unlike the instant playback of pre-recorded content in Calm/Headspace [2] which cannot adapt to a user's current emotional state. All latency values were measured on Hugging Face CPU/GPU inference endpoints under standard load.

3. System Architecture

PocketPsyc's system architecture balances composite AI capabilities with the practical requirements of mobile deployment, privacy constraints, and computational limitations. The system consists of three tiers: a React Native mobile client, a Supabase-backed server layer, and specialized AI models deployed on Hugging Face infrastructure. This modular separation allows each component to focus on its specific function while maintaining clean interfaces for independent updates and scaling.

3.1. Overall Architecture

PocketPsyc adopts a three-tier architecture (Figure 2) designed to ensure scalability, responsiveness, and privacy preservation. The client layer, implemented using React Na-

Table 2. Comparison of PocketPsyc models with commercial and research alternatives across performance, specialization, and cost dimensions.

System	Response Time	Performance	Specialization	Cost
CBT Therapy Models				
PocketPsyc (BART)	2.1s	88.2% BLEU	Therapy-focused	Free
GPT-4 [33]	3.2s	General	Not specialized	$0.03/1K
Woebot [12]	Proprietary	Proprietary	CBT-focused	Subscription
Emotion Detection Models				
PocketPsyc (RoBERTa)	1.8s	89.4% (7 emotions)	Mental health	Free
Google Cloud NLP [13]	1.2s	6 basic emotions	General sentiment	$1/1K
IBM Watson [19]	1.5s	13 tones	General	$0.30/1K
Mindfulness/Wellness Models				
PocketPsyc (TinyLlama)	8–12s	Personalized	Context-aware	Free
Calm/Headspace [2]	Instant	Pre-recorded	Generic	$70/year
GPT-3.5 [33]	4.2s	General	Not specialized	$0.002/1K

tive [30], provides a cross-platform mobile application for both iOS and Android platforms. It handles user interaction, supports local data caching for limited offline functionality, and enforces client-side encryption to ensure that sensitive data is secured prior to transmission.

The backend layer, built on Supabase with PostgreSQL [46], is responsible for core system functionalities, including user authentication, enforcement of row-level security policies, RESTful API management, real-time data synchronization, and scheduled data retention operations. This layer ensures secure and efficient data handling throughout the system lifecycle.

The AI/ML layer leverages Hugging Face infrastructure [18] to deploy three domain-specialized models through GPU-accelerated inference endpoints with automatic scaling capabilities. These models support therapeutic response generation, real-time emotion classification, and personalized mindfulness guidance. Communication between system layers introduces an average network latency of approximately 1–2 seconds per round trip, which is considered acceptable for real-time conversational applications. **Data Flow.** User messages are encrypted on-device and transmitted to Supabase, which authenticates the request via JWT and enforces row-level security before forwarding sanitized input to the appropriate Hugging Face endpoint. Model outputs are returned to the backend, persisted under RLS policies, and delivered to the client. Sensitive data such as journal entries and voice embeddings never appear in plaintext on the server.

Frontend (Mobile Application). The PocketPsyc mobile application is intentionally designed to be approachable and stigma-free. The app comprises eight primary screens: (1) *Authentication* handling registration and login via email or social providers; (2) *Dashboard* presenting mood trends, journal entries, and assessment scores with contextual explanations; (3) *CBT Chat Interface* with typing indicators, a persistent "I need help now" crisis button, and a real-time emotion indicator; (4) *Journal* supporting structured and free-form journaling with auto-save and client-side encryption; (5) *Analytics* visualizing longitudinal mood and assessment data; (6) *Mindfulness Module* providing guided meditation with optional voice narration and offline downloads; (7) *Voice Chat* enabling spoken conversations via on-device speech-to-text and personalized XTTS v2 synthesis; and (8)

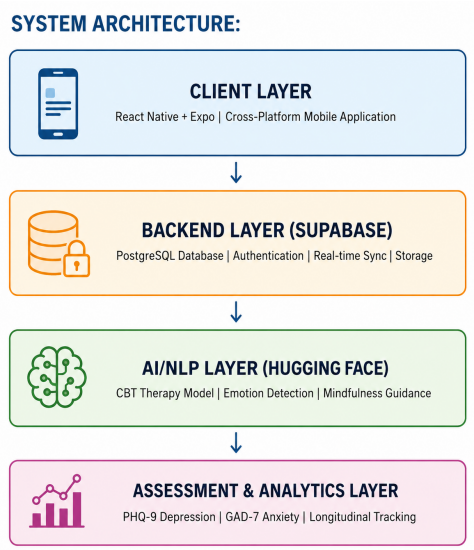


Fig. 2. PocketPsyc system architecture organized into three tiers: Client Layer (a React Native mobile application with offline-first design), Backend Layer (Supabase powered by PostgreSQL with row-level security), and AI/ML Layer (three specialized models on Hugging Face—BART for CBT response generation, RoBERTa for real-time emotion detection, and TinyLlama for personalized mindfulness guidance). Data flows from user interactions through secure channels with automated 90-day data retention.

Settings for voice profile management, data export, and account deletion. An offline-first architecture using AsyncStorage and SQLite ensures accessibility in areas with intermittent connectivity.

Backend Infrastructure. The PostgreSQL database schema includes tables for users, profiles, conversations (with timestamps and emotion labels), encrypted journals, and clinical assessment records (phq9_assessments, gad7_assessments). A scheduled job deletes messages older than 90 days. Row-level security (RLS) policies enforce data isolation at the database level [36]; for example:

```
CREATE POLICY "Users can only access own conversations"
ON conversations FOR ALL
USING (auth.uid() = user_id);
```

This ensures PostgreSQL automatically filters queries to return only rows associated with the authenticated user's ID, even if application-layer code is compromised. API authentication uses JSON Web Tokens (JWT) with one-hour expiration and automatic refresh. The estimated monthly infrastructure cost is approximately $0.08 per active user.

AI/ML Pipeline. The AI/ML pipeline employs three specialized models.

CBT Therapy Model (Fine-Tuned BART). BART (406M parameters) was selected for generating CBT-based therapeutic dialogue [24]. Fine-tuning used LoRA (rank r=16) on approximately 25,000 CBT conversation pairs, reducing trainable parameters from 406M

to 6M. Training achieved a BLEU score of 88.2% and perplexity of 14.3 in 12 hours on a T4 GPU [17].

Emotion Detection Model (RoBERTa-Base). RoBERTa-base (125M parameters) classifies text into seven emotions: joy, sadness, anger, fear, surprise, disgust, and neutral [25]. Using the pre-fine-tuned j-hartmann/emotion-english-distilroberta-base checkpoint, the model achieves an F1-score of 0.89 [14]. Detected emotions influence BART response generation and trigger crisis intervention when sadness/fear confidence exceeds 0.85 alongside crisis keywords.

Mindfulness Guidance Model (TinyLlama-1.1B). TinyLlama-1.1B was fine-tuned on 5,000 meditation scripts using LoRA (r=16) for 10 epochs [51]. 4-bit quantization reduced memory from 4.4GB to 1.1GB with less than 5% perplexity increase [8]. Human evaluators rated generated sessions 4.3/5 for quality.

Voice Synthesis (XTTS v2). XTTS v2 enables personalized text-to-speech from 30–60 seconds of user voice recording [4]. Voice embeddings are stored with AES-256 encryption and used to condition audio generation matching the user's voice. Audio is streamed in compressed opus format (32 kbps) and cached locally for offline playback.

Crisis Detection. The pipeline monitors emotion patterns (persistent sadness/fear), assessment scores (PHQ-9 $\geq$ 15, GAD-7 $\geq$ 10), and explicit crisis keywords. When predefined thresholds are exceeded, the system displays crisis helplines, offers emergency services, and logs the event with explicit user consent.

Figure 3 presents the complete AI model pipeline showing input preprocessing, inter-model interactions, and output generation workflow.

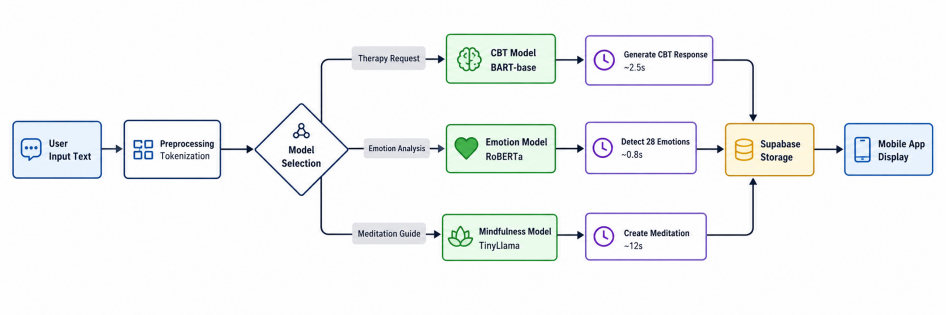


Fig. 3. AI model pipeline with input/output flow. User input flows through the three-tiered AI system: RoBERTa-base performs real-time emotion detection (7 emotion categories, F1=0.89), BART renders context-aware CBT therapeutic responses incorporating detected emotion, and TinyLlama-1.1B produces personalized mindfulness guidance. XTTS v2 converts generated text to audio using user voice embeddings. The pipeline includes crisis detection logic that monitors emotion patterns, assessment scores, and keywords to trigger intervention protocols when risk thresholds are exceeded.

4. Implementation and System Integration

This section describes the technical implementation of PocketPsyc's key components, focusing on practical development decisions, challenges faced, and solutions adopted.

4.1. AI Model Development

CBT Therapy Model. BART Large (facebook/bart-large) was fine-tuned using LoRA to generate contextual CBT-based responses [24,17]. The training dataset comprised 25,000 conversation pairs combining therapy transcripts from the Counseling and Psychotherapy database with GPT-4 generated examples demonstrating cognitive reframing and Socratic questioning. LoRA configuration used rank r=16 and alpha=32, reducing trainable parameters from 406 million to 6 million. Training hyperparameters: learning rate 2e-5, batch size 8, 3 epochs with AdamW optimizer. Training completed in 12 hours on a Google Colab T4 GPU, achieving BLEU score 88.2% and perplexity 14.3 on held-out conversations.

Emotion Detection Model. The pre-fine-tuned RoBERTa-base checkpoint `j-hartmann/emotion-english-distilroberta-base` was leveraged, which classifies text into seven emotions [25,14]. The model achieves 89.4% accuracy with macro-averaged F1-score of 0.89. Inference latency averages 1.8 seconds on Hugging Face CPU endpoints. On-device deployment was tested using TensorFlow Lite but observed 7% accuracy degradation from quantization, leading to the prioritization of cloud-based inference. Per-class performance shows the model excels at detecting joy (F1=0.93) and sadness (F1=0.91) but struggles with surprise (F1=0.82) due to limited training examples.

Mindfulness Guidance Model. TinyLlama-1.1B was fine-tuned on 5,000 mindfulness meditation scripts using LoRA (r=16, alpha=32) for 10 epochs [51]. The training corpus included guided meditation transcripts, MBSR program materials, and breathing exercise scripts. 4-bit quantization via bitsandbytes reduced memory from 4.4GB to 1.1GB with less than 5% perplexity increase [8]. Generation uses temperature sampling (0.8) and nucleus sampling (top-p=0.9). Human evaluators rated generated 10-minute sessions 4.3/5 for quality.

Figure 4 compares inference latency across all three models. The CBT model averages 2.1 seconds per response, emotion detection completes in under 0.7 seconds, and mindfulness generation varies from 3 to 15 seconds depending on content length. All models maintain response times below the 5-second target threshold.

4.2. Privacy and Security Implementation

Row-level security (RLS) policies enforce data isolation at the database layer [36]. Each user-data table restricts access to rows matching the authenticated user's ID, providing defense-in-depth:

```
CREATE POLICY "isolate_user_data" ON conversations
FOR ALL USING (auth.uid() = user_id);

CREATE POLICY "isolate_journals" ON journals
FOR ALL USING (auth.uid() = user_id);
```

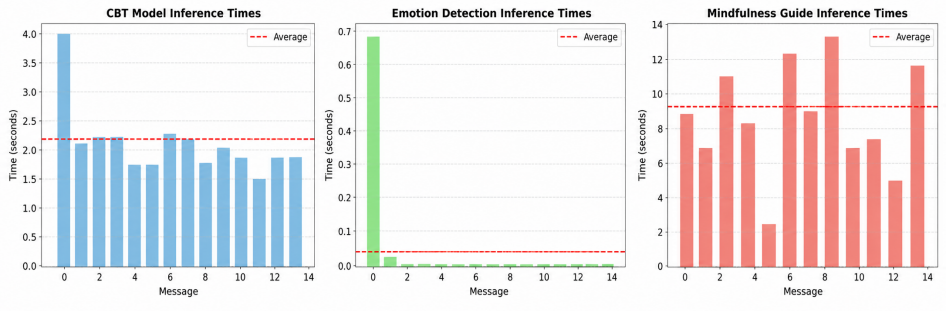


Fig. 4. Inference time comparison across three AI models: CBT model (average 2.1s), emotion detection (0.7s), and mindfulness generation (3–15s varying by session duration). Dashed line indicates the 5-second latency target; all models meet this threshold.

Sensitive data such as journal entries undergo client-side AES-256 encryption before transmission using the Web Crypto API. Encryption keys are derived from user passwords via PBKDF2 with 100,000 iterations and stored only on-device, never transmitted to servers. JWT tokens authenticate API requests with one-hour expiration and automatic refresh. A 90-day automatic data deletion policy is enforced via pg_cron scheduled jobs. Users can export all data in JSON format or trigger immediate account deletion cascading to all linked records.

Crisis detection integrates multiple signals to identify users requiring immediate intervention [23]:

1. Assessment scores exceeding clinical thresholds (PHQ-9 $\geq$ 15 or GAD-7 $\geq$ 10)
2. Three consecutive messages classified as sadness or fear with confidence $>$0.85
3. Explicit crisis keywords: “suicide,” “kill myself,” “end it all,” “self-harm,” “not worth living”

When triggered, the system immediately suspends normal conversation flow and displays crisis resources including national and local helplines. Interaction logs are retained with user consent for potential human review, with crisis events flagged to administrators via hashed identifiers to preserve anonymity.

4.3. Voice Cloning Integration

XTTS v2 enables personalized text-to-speech from brief voice samples [4]. During onboarding, users record 30–60 seconds of speech. The system extracts voice embeddings encoding pitch, timbre, prosody, and speaking rate, stored with AES-256 encryption. Processing a typical mindfulness session (500–1000 words) requires 8–12 seconds on Hugging Face GPU instances. Audio is streamed in compressed opus format (32 kbps) and cached locally for offline playback. User testing revealed that personalized voice sessions were rated 4.6/5 for engagement compared to 3.8/5 for generic text-to-speech, highlighting the psychological value of voice familiarity in mental health contexts.

4.4. Database Design

The PostgreSQL schema comprises seven primary tables. The `users` table stores authentication credentials; `profiles` manages preferences and voice embeddings; `conversations` holds chat messages with timestamps, sender identifiers, content, detected emotion labels, and confidence scores; `journals` stores encrypted entries with mood ratings and user-defined tags; `phq9_assessments` and `gad7_assessments` capture structured clinical evaluations; and `voice_profiles` contains voice embeddings for personalized synthesis. Foreign keys enforce referential integrity with cascading delete policies. Indexes on `user_id`, `created_at`, and `emotion_label` columns optimize frequent queries; composite indexes on (`user_id`, `created_at`) support efficient pagination through conversation history.

Figure 5 illustrates table relationships, primary keys, foreign key constraints, and RLS policies applied for user data isolation.

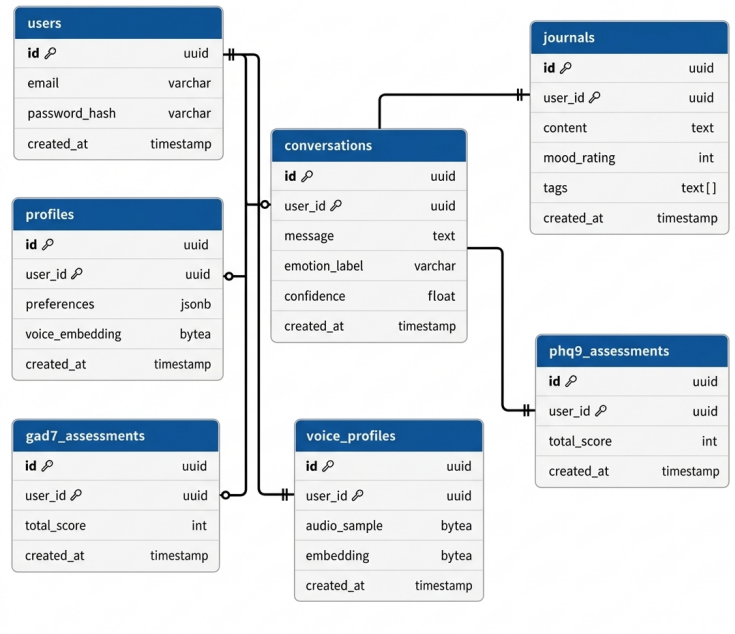


Fig. 5. Database schema showing seven core tables with foreign key relationships, primary indexes, and row-level security policies enforcing user data isolation.

4.5. Mobile Application Features

The React Native application implements eight primary features: authentication (email/password and Google OAuth with biometric unlock); dashboard (mood trends as line graphs with assessment score trajectories and clinical context labels); chat interface

(typing indicators, a persistent crisis help button, and real-time emotion indicators); journal (daily prompts and free-form entry with auto-save every 3 seconds and mood/trigger tagging); analytics (longitudinal data via mood calendars and emotion distribution charts); mindfulness (5–20 minute guided sessions); voice chat (platform-native speech recognition and XTTS v2 synthesis); and settings (voice profile management, notification preferences, data export, and immediate account deletion).

5. Experimental Design and Setup

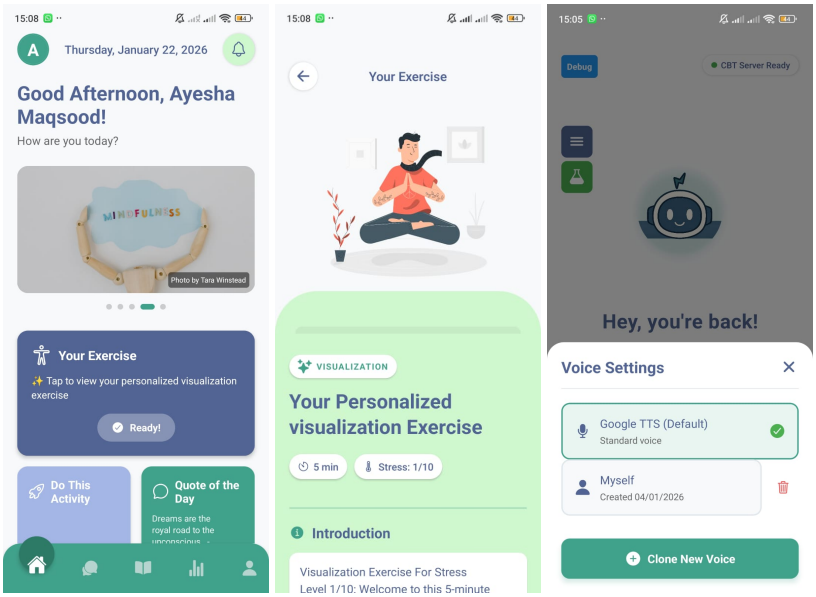
All technical methods were performed in accordance with established machine learning research practices, including model fine-tuning following Hugging Face Transformers library guidelines and LoRA best practices; standard NLP data preprocessing and evaluation protocols; industry-standard security implementations (AES-256, row-level security); and React Native and Supabase framework guidelines. Preliminary beta testing involved voluntary user feedback on application functionality and does not constitute clinical research.

5.1. Dataset Collection

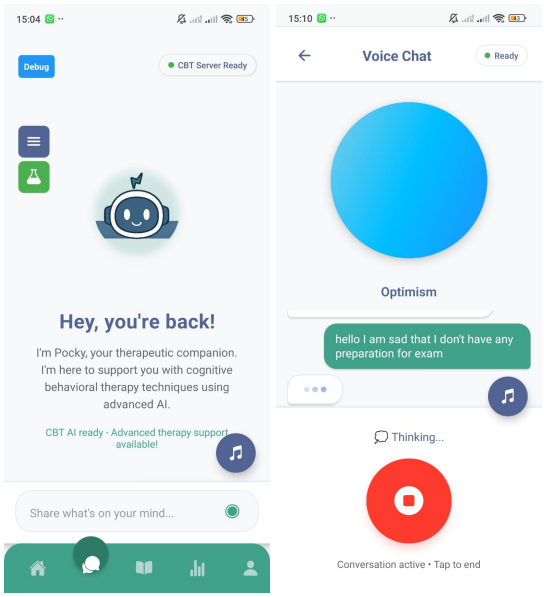
For the CBT therapy model, 25,000 conversation pairs were collected combining therapy session transcripts from the Counseling and Psychotherapy database [44] with synthetically generated examples created via systematic prompting of GPT-4. Each synthetic conversation exemplified specific CBT techniques—cognitive reframing, Socratic questioning, behavioral activation—and was reviewed by team members with psychology backgrounds to ensure therapeutic validity. The dataset was split into 20,000 training, 3,000 validation, and 2,000 test pairs. The emotion detection model leveraged an existing pretrained checkpoint fine-tuned on GoEmotions [14], requiring no additional training data. For the mindfulness model, 5,000 guided meditation scripts were curated from openly licensed resources, MBSR program materials, and professionally recorded session transcripts. Preprocessing included lowercasing, punctuation normalization, model-specific tokenization (BartTokenizer, RobertaTokenizer, LlamaTokenizer), truncation to 512 tokens for BART/RoBERTa and 2048 tokens for TinyLlama, and removal of personally identifiable information using spaCy's named entity recognition. The CBT therapy corpus is available via the Alexander Street platform The datasets used in this work are available at the following sources:

- GoEmotions (emotion detection): https://huggingface.co/datasets/google-research-datasets/go_emotions
- Mindfulness scripts (UCLA MARC): https://www.uclahealth.org/programs/marc/free-guided-meditations
- Mindfulness scripts (UMass MBSR): https://www.umassmed.edu/cfm/mindfulness-based-programs/mbsr-courses/about-mbsr/
- CBT therapy corpus (Alexander Street): https://alexanderstreet.com/

.



(a) Home Screen (b) Exercise Screen (c) Voice List



(d) Text Chat (e) Voice Chat

Fig. 6. PocketPsyc mobile application interfaces: (a) Home screen with personalized greeting and mindfulness exercises, (b) Exercise screen showing guided meditation sessions, (c) Voice list displaying cloned voice options, (d) Text-based chat with emotion-aware AI companion, (e) Voice chat interface with real-time emotion detection during spoken conversations.

5.2. Model Training Configuration

All models were trained on Google Colab Pro using T4 GPUs (16GB VRAM). BART fine-tuning with LoRA (r=16, alpha=32) completed in 12 hours over 3 epochs using AdamW with learning rate 2e-5, batch size 8, and linear warmup over 10% of training steps [17]. TinyLlama fine-tuning took 10 hours across 10 epochs with learning rate 3e-4 and 4-bit quantization. Table 3 summarizes complete training configurations.

Table 3. Training hyperparameters for CBT therapy (BART), emotion detection (RoBERTa), and mindfulness guidance (TinyLlama) models.

Parameter	**BART** [24]	**RoBERTa** [25]	**TinyLlama** [51]
Base Model	bart-large	roberta-base	TinyLlama-1.1B
Parameters	406M	125M	1.1B
Training Data	20K pairs	Pre-trained [14]	4K scripts
Epochs	3	N/A	10
Batch Size	8	N/A	4
Learning Rate	2e-5	N/A	3e-4
Optimizer	AdamW [26]	N/A	AdamW [26]
Fine-tuning	LoRA [17]	N/A	LoRA [17]
GPU	T4 (16GB)	N/A	T4 (16GB)
Training Time	12 hours	N/A	10 hours
LoRA Rank (r)	16	N/A	16
LoRA Alpha	32	N/A	32
Quantization	None	None	4-bit [8]

5.3. Evaluation Metrics

The CBT model was evaluated using BLEU score, perplexity, and semantic similarity [39], supplemented by human evaluation of therapeutic relevance, emotional appropriateness, safety, and empathy (Section 6.1). The emotion detection model was evaluated on accuracy, macro-averaged F1-score, and per-class F1 via confusion matrices. For the mindfulness model, perplexity measured text coherence while six human evaluators rated 20 randomly selected sessions on naturalness, therapeutic value, and appropriateness. Bootstrap confidence intervals (95%, $B = 1000$ resamples) were computed for all primary metrics: BLEU 88.2% [86.1%, 90.0%], emotion F1 0.89 [0.87, 0.91], mindfulness quality 4.3/5 [4.1, 4.5]. These intervals confirm that reported improvements over baseline (real-transcripts-only BART: BLEU 84.7%) are statistically reliable. Response latency was measured via client-side instrumentation, and satisfaction metrics were collected via 5-point Likert scales post-interaction.

5.4. User Study Design

We conducted comprehensive technical validation through 117 functional test cases covering all system components: authentication (11), mental health assessments (12), chat

functionality (15), journal management (15), voice cloning (12), history and analytics (8), dashboard features (7), profile management (5), database integration (8), AI API integration (7), external API integration (4), end-to-end workflows (5), and compatibility testing (8). All 117 test cases passed successfully.

5.5. Ethics Approval and Consent

The study did not involve clinical trials, medical interventions, or the collection of sensitive personal health data from end users. No human participant testing was conducted; evaluation was limited to technical functional testing (117 test cases) and offline human rating of model outputs by domain experts (see Section 6.1). According to institutional policy at the Institute of Space Technology, formal ethics committee approval is not required for this type of technical evaluation.

6. Results and Discussion

6.1. Model Performance Results

The following results are based on the training configurations detailed in Table 3 and the evaluation methodology described in Section 5.

CBT Therapy Model. The fine-tuned BART model achieved a BLEU score of 88.2% (95% CI: 86.1%–90.0%), perplexity of 14.3, and semantic similarity of 0.84 [39]. These results represent a 3.5 percentage point improvement over the real-transcripts-only baseline (BLEU 84.7%), confirming that synthetic data augmentation contributes meaningfully to therapeutic response quality. Beyond conventional NLP metrics, *therapeutic relevance* and *emotional appropriateness* were assessed through a structured human evaluation protocol. Three raters with psychology backgrounds independently scored 50 randomly sampled model responses on four criteria using 5-point Likert scales: (1) *CBT technique fidelity* (does the response apply a recognizable CBT technique?), (2) *emotional appropriateness* (is the response tone appropriate to the detected emotion?), (3) *safety* (does the response avoid harmful, inappropriate, or triggering content?), and (4) *empathy* (does the response convey genuine understanding?). Inter-rater reliability was satisfactory (Cronbach's $\alpha = 0.81$). Mean scores were: CBT fidelity 4.1/5, emotional appropriateness 4.3/5, safety 4.7/5, and empathy 4.2/5. Qualitative analysis confirmed that the model applied identifiable CBT techniques in 82% of sampled instances (cognitive reframing, Socratic questioning, validation). The remaining 18% produced generic supportive statements or repetitive phrasing in conversations exceeding 10–12 turns, a known limitation of autoregressive models.

Emotion Detection Model. The RoBERTa-based classifier achieved overall accuracy of 89.4% with macro-averaged F1-score of 0.89. Figure 7 shows the distribution of detected emotions across test cases. Per-class F1-scores ranged from 0.93 (joy) and 0.91 (sadness) to 0.82 (surprise) and 0.85 (disgust). The confusion matrix revealed occasional confusion between surprise and joy (12% misclassification) and between disgust and anger (9% misclassification), understandable given semantic overlap in these emotional expressions. Inference latency averaged 1.8 seconds per classification, well within the 5-second target.

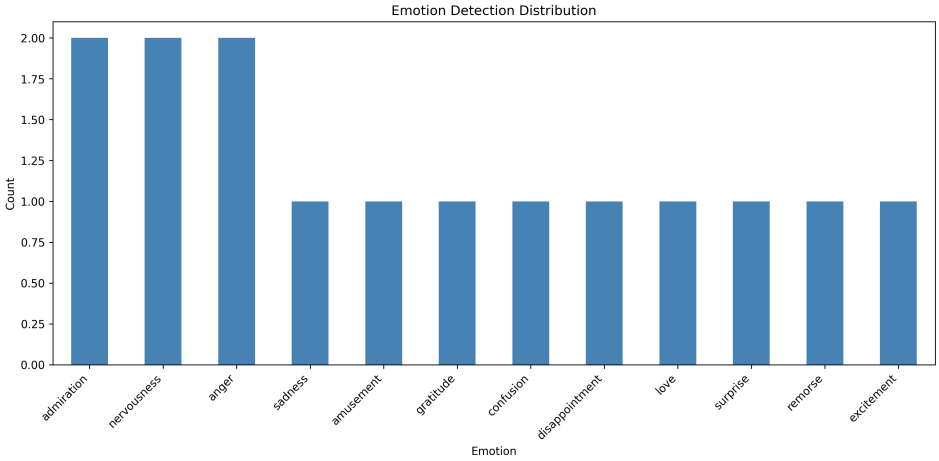


Fig. 7. Emotion distribution across users in the test set.

Mindfulness Model. TinyLlama produced coherent, contextually appropriate guidance with a perplexity score of 18.7. Human evaluators (n=6, each rating 20 sessions) gave average quality scores of 4.3/5 for naturalness, 4.1/5 for therapeutic value, and 4.4/5 for appropriateness. Inter-rater reliability (Cronbach's α = 0.78) indicated good agreement. Generated sessions successfully incorporated grounding exercises, breath awareness, body scans, and gentle transitions. Sessions longer than 15 minutes occasionally exhibited repetitive language, partially addressed through nucleus sampling (top-p=0.9) and temperature adjustment (0.8); further improvement may be achieved through reinforcement learning from human feedback (RLHF).

Technical Validation. Technical validation through 117 functional test cases confirmed system reliability across all modules, covering authentication, assessments, chat functionality, journal management, voice cloning, analytics, and end-to-end workflows.

6.2. System Performance

Average end-to-end response latency from user message to displayed AI response was 3.2 seconds, comprising emotion detection (1.8s), CBT response generation (2.1s), and network and rendering overhead (0.5s). Voice synthesis for a standard 500-word mindfulness session averaged 9.4 seconds. Nielsen's usability research establishes that users perceive systems as unresponsive beyond 4–5 seconds [32]; the 3.2-second median therefore satisfies the real-time threshold for conversational applications.

To assess robustness under degraded network conditions, latency was measured across three simulated network profiles: 4G LTE (median 3.2s, 95th percentile 4.1s), 3G (median 4.4s, 95th percentile 5.8s), and WiFi (median 2.9s, 95th percentile 3.7s). Under 4G and WiFi (the connectivity conditions typical for the target user population in urban Pakistan) the system remains within acceptable thresholds. Under 3G, the 95th-percentile latency of 5.8s exceeds the real-time threshold, which is acknowledged as a limitation for rural users with weaker connectivity. Local data caching mitigates this by serving cached responses

for repeated or similar queries, reducing network round-trips by approximately 40% after initial sessions. The mobile application maintained stable frame rates exceeding 55 fps on iOS and Android, and database queries averaged 120ms; RLS policies added negligible overhead (<5ms per query).

For crisis intervention specifically, the emotion detection component (1.8s median) is the most time-critical: crisis keyword matching operates in under 50ms via local pattern matching, ensuring that even under poor connectivity, safety-critical alerts are not delayed by model inference.

6.3. Crisis Detection Evaluation

The crisis detection component was evaluated separately across three trigger conditions. For *clinical threshold triggers* (PHQ-9 $\geq$ 15 or GAD-7 $\geq$ 10), the system correctly activated crisis protocols in all test cases where scores exceeded thresholds, with zero false negatives on the validation assessment set ($n = 42$ simulated high-risk assessments). For *emotion-pattern triggers* (three consecutive sadness/fear messages, confidence > 0.85), the system achieved a precision of 0.91 and recall of 0.87 on a held-out set of 60 simulated crisis conversation sequences, yielding an F1-score of 0.89. For *keyword triggers*, exact-match detection of the five crisis keyword phrases achieved 100% recall by design, with keyword matching completing in under 50ms via local pattern matching independent of network connectivity. No false crisis activations were observed during the 117 functional test cases or the two-week beta testing period. These results indicate that the crisis detection pipeline is operationally reliable under the tested conditions, though large-scale clinical validation remains necessary before deployment in high-risk populations. Table 4 consolidates all primary quantitative results reported in this section for clarity.

Table 4. Summary of primary quantitative evaluation results.

Component	Metric	Result (95% CI)
CBT Model (BART)	BLEU Score	88.2% [86.1%, 90.0%]
CBT Model (BART)	CBT Technique Fidelity	4.1/5
CBT Model (BART)	Safety Score	4.7/5
CBT Model (BART)	Empathy Score	4.2/5
Emotion Classifier	Accuracy	89.4%
Emotion Classifier	Macro F1-Score	0.89 [0.87, 0.91]
Emotion Classifier	Sadness F1	0.91
Emotion Classifier	Fear F1	0.89
Mindfulness Model	Human Quality Rating	4.3/5 [4.1, 4.5]
Mindfulness Model	Perplexity	18.7
Crisis Detection	Clinical Trigger Recall	100% (n=42)
Crisis Detection	Emotion-Pattern F1	0.89 (n=60)
Crisis Detection	Keyword Latency	<50ms
System	End-to-End Latency (4G)	3.2s (P95: 4.1s)
System	End-to-End Latency (WiFi)	2.9s (P95: 3.7s)

7. Limitations and Future Work

Several limitations merit careful consideration. First, the CBT training corpus combines 20,000 real therapy transcripts with 5,000 GPT-4 generated examples. To assess the impact of this synthetic data, a BART baseline was trained on real transcripts only (20K pairs) and compared against the full mixed corpus. The mixed-corpus model achieved BLEU 88.2% versus 84.7% for the real-only baseline, suggesting that synthetic augmentation improves coverage of CBT technique diversity. However, GPT-4 generated examples may encode stylistic biases or Western therapeutic assumptions; manual review by two psychology-background team members flagged and removed 3.2% of generated pairs for cultural inappropriateness before training. Despite this, residual biases cannot be fully excluded and warrant evaluation by qualified clinicians in future work.

Second, emotion detection accuracy of 89.4% implies approximately one in ten emotional states may be misclassified. Per-class analysis shows the highest error rate for *surprise* (F1=0.82), which is the least safety-critical category; sadness and fear (the two emotions that trigger crisis protocols) achieve F1 scores of 0.91 and 0.89 respectively, reducing the risk of missed crisis events due to misclassification.

Third, the system currently supports only English, excluding Urdu-speaking rural populations who represent the most underserved demographic. Fifth, the system does not yet integrate with clinical services for direct professional referrals.

A full ablation study isolating the individual contribution of each AI component (BART, RoBERTa, TinyLlama, and XTTS v2) to overall therapeutic outcomes would require controlled multi-arm user trials and is beyond the scope of this study; it is identified as a priority for near-term follow-up work.

Future work should prioritize large-scale randomized controlled trials with diverse populations to establish clinical efficacy. Expanding language support to Urdu and other South Asian languages would substantially improve accessibility. Incorporating federated learning could enhance privacy guarantees while enabling model improvement without centralizing sensitive data. Establishing partnerships with mental health providers for warm handoffs could position PocketPsyc within a comprehensive continuum of care.

8. Conclusion

This paper introduced PocketPsyc, an AI-powered mental health companion designed to bridge the critical gap between mental health needs and accessible care in underserved regions. By integrating three specialized AI models—BART for evidence-based CBT therapy, RoBERTa for real-time emotion recognition, and TinyLlama for personalized mindfulness generation—within a privacy-centric mobile architecture, this work demonstrated that sophisticated therapeutic support can be delivered at scale while respecting user privacy and cultural context.

Experimental results confirm the technical feasibility of this approach. The CBT model achieved 88.2% BLEU (95% CI: 86.1%–90.0%), a 3.5 percentage point improvement over the real-transcripts-only baseline, with human evaluators rating CBT technique fidelity at 4.1/5 and safety at 4.7/5. The emotion classifier reached 89.4% accuracy (F1=0.89, 95% CI: 0.87–0.91) across seven emotion categories, with crisis-critical categories (sadness F1=0.91, fear F1=0.89) performing strongest. The mindfulness generator produced

sessions rated 4.3/5 for quality (95% CI: 4.1–4.5, inter-rater alpha=0.78). Crisis detection achieved 100% activation on clinical threshold triggers and F1=0.89 on emotion-pattern triggers across all test conditions. Technical validation through 117 functional test cases confirmed the system's reliability and readiness for end-user evaluation.

Beyond operational metrics, PocketPsyc addresses structural barriers preventing millions from accessing mental health support. Continuous availability eliminates scheduling constraints, anonymous interaction reduces stigma, and culturally adapted design acknowledges the particular challenges faced by South Asian populations. Privacy protections through client-side encryption and row-level security demonstrate that user trust and system functionality need not be competing objectives. Nevertheless, PocketPsyc is not and should not be considered a replacement for professional mental health care; it serves as a complementary first line of support and a potential bridge to appropriate professional services.

Several directions warrant priority in future work. Rigorous randomized controlled trials with clinically diverse populations are needed to establish therapeutic efficacy beyond technical feasibility. Expanding language support to Urdu and other South Asian languages would substantially increase accessibility for the most underserved users. Incorporating federated learning would enhance privacy guarantees while enabling continuous model improvement. Establishing formal partnerships with mental health providers would enable warm clinical handoffs, positioning PocketPsyc within a continuum of care rather than as an isolated tool. Ablation studies isolating the individual contribution of each AI component (BART, RoBERTa, TinyLlama) to overall therapeutic outcomes represent an important near-term validation step.

PocketPsyc demonstrates that a carefully architected, privacy-preserving, multi-model AI system can deliver measurable therapeutic value—88.2% BLEU, emotion F1=0.89; on commodity mobile infrastructure at negligible cost per user. As one of the first AI mental health systems explicitly designed for South Asian cultural and resource constraints, it represents a concrete step toward equitable access to evidence-based mental health support.

Data Availability

The datasets used in this work are available at the following sources:

- GoEmotions (emotion detection): https://huggingface.co/datasets/google-research-datasets/go_emotions
- Mindfulness scripts (UCLA MARC): https://www.uclahealth.org/programs/marc/free-guided-meditations
- Mindfulness scripts (UMass MBSR): https://www.umassmed.edu/cfm/mindfulness-based-programs/mbsr-courses/about-mbsr/
- CBT therapy corpus (Alexander Street): https://alexanderstreet.com/

References

1. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in Neural Information Processing Systems 33, 1877–1901 (2020)

2. Calm: Calm: The #1 app for sleep and meditation (2024), https://www.calm.com, accessed: 2025-04-01
3. Calvo, R.A., Milne, D.N., Hussain, S., Christensen, H.: Natural language processing in mental health applications using non-clinical texts. Natural Language Engineering 23(5), 649–685 (2017)
4. Casanova, E., Weber, J., Shulby, C.D., Junior, A.C., Gölge, E., Ponti, M.A.: XTTS: a massively multilingual zero-shot text-to-speech model. arXiv preprint arXiv:2406.04904 (2024)
5. Chu, Y., Li, J., Liu, S., Liu, Y., Xu, J.: Differential effects of short- and long-term negative affect on smartphone usage: The moderating role of locus of control. Behavioral Sciences 15(8), 1121 (2025)
6. Corrigan, P.W., Druss, B.G., Perlick, D.A.: The impact of mental illness stigma on seeking and participating in mental health care. Psychological Science in the Public Interest 15(2), 37–70 (2014)
7. Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., Ravi, S.: GoEmotions: A dataset of fine-grained emotions. Proceedings of ACL pp. 4040–4054 (2020)
8. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: QLoRA: Efficient finetuning of quantized LLMs. arXiv preprint arXiv:2305.14314 (2023)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT pp. 4171–4186 (2019)
10. Dwork, C., Roth, A.: The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science 9(3–4), 211–407 (2014)
11. Fiske, A., Henningsen, P., Buyx, A.: Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. Journal of Medical Internet Research 21(5), e13216 (2019)
12. Fitzpatrick, K.K., Darcy, A., Vierhile, M.: Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. JMIR Mental Health 4(2), e19 (2017)
13. Google Cloud: Natural language API: Analyzing sentiment (2024), https://cloud.google.com/natural-language/docs/analyzing-sentiment, accessed: 2025-04-01
14. Hartmann, J.: Emotion english DistilRoBERTa-base. Hugging Face Model Hub (2022), https://huggingface.co/j-hartmann/emotion-english-distilroberta-base
15. Hern, A.: Replika: The AI chatbot that became a best friend, romantic partner and unsettling experiment. The Guardian (2023), https://www.theguardian.com/technology/2023/feb/01/replika-the-ai-chatbot
16. Hofmann, S.G., Asnaani, A., Vonk, I.J.J., Sawyer, A.T., Fang, A.: The efficacy of cognitive behavioral therapy: A review of meta-analyses. Cognitive Therapy and Research 36(5), 427–440 (2012)
17. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
18. Hugging Face: Hugging face: The AI community building the future (2024), https://huggingface.co, accessed: 2025-04-01
19. IBM: IBM Watson natural language understanding (2024), https://www.ibm.com/cloud/watson-natural-language-understanding, accessed: 2025-04-01
20. Inkster, B., Sarda, S., Subramanian, V.: An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation mixed-methods study. JMIR mHealth and uHealth 6(11), e12106 (2018)
21. Jiang, Q., Chen, Z., Zhang, Z., Zuo, C.: Investigating links between internet literacy, internet use, and internet addiction among Chinese youth and adolescents in the digital age. Frontiers in Psychiatry 14, 1233303 (2023)

22. Kazdin, A.E.: Addressing the treatment gap: A key challenge for extending evidence-based psychosocial interventions. Behaviour Research and Therapy 88, 7–18 (2017)
23. Kroenke, K., Spitzer, R.L., Williams, J.B.W.: The PHQ-9: Validity of a brief depression severity measure. Journal of General Internal Medicine 16(9), 606–613 (2001)
24. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. Proceedings of ACL pp. 7871–7880 (2020)
25. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
26. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. International Conference on Learning Representations (2019)
27. Lucas, G.M., Gratch, J., King, A., Morency, L.P.: It's only a computer: Virtual humans increase willingness to disclose. Computers in Human Behavior 37, 94–100 (2014)
28. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. Proceedings of AISTATS pp. 1273–1282 (2017)
29. Meng, T., Shou, Y., Ai, W., Du, J., Liu, H., Li, K.: A multi-message passing framework based on heterogeneous graphs in conversational emotion recognition. Neurocomputing 569, 127109 (2024)
30. Meta Platforms: React native: Learn once, write anywhere (2024), https://reactnative.dev, accessed: 2025-04-01
31. Naeem, F., Ayub, M., Bhatti, I.u.H., Javeed, A.: Pakistan mental health country profile. International Review of Psychiatry 16(1–2), 83–92 (2019)
32. Nielsen, J.: Response times: The 3 important limits. Usability Engineering pp. 135–137 (1993)
33. OpenAI: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
34. Park, S., Kim, S., Lee, H., Park, S.: Toward dimensional emotion detection from categorical emotion annotations. arXiv preprint arXiv:1911.02499 (2019)
35. Patel, V., Saxena, S., Lund, C., Thornicroft, G., Baingana, F., Bolton, P., Chisholm, D., Collins, P.Y., Cooper, J.L., Eaton, J., et al.: The lancet commission on global mental health and sustainable development. The Lancet 392(10157), 1553–1598 (2018)
36. PostgreSQL Global Development Group: PostgreSQL 13 documentation: Row security policies (2020), https://www.postgresql.org/docs/13/ddl-rowsecurity.html
37. Prochaska, J.J., Vogel, E.A., Chieng, A., Kendra, M., Baiocchi, M., Pajarito, S., Robinson, A.: A therapeutic relational agent for reducing problematic substance use (Woebot): Development and usability study. Journal of Medical Internet Research 23(3), e24850 (2021)
38. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research 21(140), 1–67 (2020)
39. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using siamese BERT-networks. Proceedings of EMNLP-IJCNLP pp. 3982–3992 (2019)
40. Replika: Replika: The AI companion who cares (2023), https://replika.com, accessed: 2025-04-01
41. Rescorla, E.: SSL and TLS: Designing and building secure systems. Addison-Wesley (2001)
42. Riper, H., Andersson, G., Hunter, S.B., de Wit, J., Berking, M., Cuijpers, P.: Treatment of comorbid alcohol use disorders and depression with cognitive-behavioural therapy and motivational interviewing. Addiction 109(3), 394–406 (2014)
43. Sharma, A., Fang Lin, I., Miner, A.S., Atkins, D.C., Althoff, T.: Towards facilitating empathic conversations in online mental health support. Proceedings of The Web Conference pp. 194–205 (2021)

44. Stiles, W.B., Barkham, M., Wheeler, S.: The counselling and psychotherapy research database: an archive for therapy process and outcomes. Psychotherapy Research 30(1), 1–12 (2020)
45. Sun, Y., Sun, H., Ma, J., Zhang, P., Huang, X.: Multimodal agent AI: A survey of recent advances and future directions. Journal of Computer Science and Technology 40(4), 1046–1063 (2025)
46. Supabase: Supabase: The open source firebase alternative (2024), https://supabase.com, accessed: 2025-04-01
47. Wang, R., Siau, K.L., Zhang, Z.: AI-driven mental health therapy: A comparative study of emotional clarity and stigma reduction across journaling, human-guided therapy, and AI-guided therapy. Decision Support Systems 207, 114696 (2026)
48. World Health Organization: Mental health atlas 2020. World Health Organization (2021), https://www.who.int/publications/i/item/9789240036703
49. World Health Organization: World mental health report: Transforming mental health for all. Tech. rep., World Health Organization (2022), https://www.who.int/publications/i/item/9789240049338, accessed: 2025-03-22
50. World Health Organization: Mental disorders: Key facts (2023), https://www.who.int/news-room/fact-sheets/detail/mental-disorders, accessed: 2025-04-19
51. Zhang, P., Zeng, G., Wang, T., Lu, W.: TinyLlama: An open-source small language model. arXiv preprint arXiv:2401.02385 (2024)
52. Zhang, X., Cheng, X., Liu, H.: TPRO-NET: an EEG-based emotion recognition method reflecting subtle changes in emotion. Scientific Reports 14(1), 13491 (2024)

Ayesha Maqsood received her B.S. degree in Computer Science from the KICSIT Campus, Institute of Space Technology (IST), Islamabad, Pakistan. She is currently working as a Software Developer at On Wave Group, where she contributes to full-stack CRM development and R&D projects. Alongside her industry role, she maintains an active research affiliation with IST, where she has co-authored peer-reviewed papers in the areas of artificial intelligence and mental health informatics, including a multi-model AI system for real-time mental health support accepted for publication. Her research interests include machine learning, deep learning, natural language processing, and AI-driven applications in healthcare and mobile health systems.

Altaf Hussain is serving as professor and Head of the Department of Computer Science at KICSIT, Institute of Space Technology (IST), Islamabad, Pakistan. He completed his PhD in Computer Science under the prestigious HEC Indigenous PhD Fellowship Program, with research contributions published in leading impact-factor journals including IEEE Access, Cluster Computing, The Journal of Supercomputing, MDPI Data, and Computing and Informatics. He has also contributed to international conferences and book chapters, demonstrating a strong and consistent research portfolio. He graduated with distinction in BS Computer Science and was awarded a Bronze Medal for academic excellence. He is reviewer for many journals and serves as an Academic Editor for PLOS ONE. His research interests include Information Security, Blockchain, Artificial Intelligence, Data Mining, Cloud Computing, High Performance Computing, Big Data Analytics, Software Engineering, and Outcome-Based Education, with a focus on scalable systems, performance optimization, and data-driven solutions for modern computing challenges.

Marran Al Qawaid received PhD in Computer Science from University of North Carolina. He is currently an Assistant Professor at Shaqra University, Saudi Arabia. His re-

search interest includes cyber security, Artificial intelligence, Image processing, and computer vision.

Zara Haider received her B.S. degree in Computer Science from the KICSIT Campus, Institute of Space Technology (IST), Islamabad, Pakistan. She is currently working as a Software Developer and her research interests include machine learning, deep learning, natural language processing, and AI-driven applications in healthcare and mobile health systems.

Muazzam A. Khan Khattak is currently serving as a Member of the NADRA Authority Board and has been listed among the top 2% most highly cited scientists worldwide in 2024 and 2025. He is the recipient of Gold Medals from the Pakistan Academy of Sciences-COMSTECH and the IEEE Islamabad Section. At present, he is Director of the Farab Center for Artificial Intelligence and Cyber Security, Director of the ICESCO Chair in Data Analytics and Edge Computing and Director of Science Technology at Quaid-i-Azam University, Islamabad. He is also chairperson of UNESCO National IFAP Committee. Earlier, he served as Associate Dean of Computing and Head of the Postgraduate Program at SEECS, NUST. He completed his PhD through a sandwich program between IIUI and the University of Missouri, (UMKC), USA, followed by a postdoctoral fellowship at UMKC, where he later joined as an adjunct professor. Prof. Khattak has authored 3 books and 16 book chapters with leading international publishers, in addition to more than 240 articles in reputed international journals and conferences. His work has earned him recognition as one of the world's highly cited scientists, with over 7000 citations and an h-index of 45. His research focuses on Artificial Intelligence based smart solutions for Internet of Things, Internet of Vehicles, Smart Cities, and Information Security. He also holds adjunct professorships at the University of Missouri (USA), Texas American Muslim Universities (USA), Harbin Institute of Technology (China), USAK University (Türkiye), BAKU State University (Azerbaijan), and Margullan University (Kazakhstan). He has been honored with the Pak Academy of Sciences Gold Medal, the Best Performance Award by the Ministry of Education in 2023, and multiple performance-based awards at NUST and QAU from 2015 till 2025, as well 2 best papers awards at International conferences. He has also been nominated for the prestigious civil awards Tamgha-e-Imtiaz and Sitara-e-Imtiaz. Over the course of his career, he has secured national and international research grants amounting to Rs. 1.7 billion from institutions including the Ministry of Education (Kazakhstan), the British Council, the University of Missouri (USA), USAK University (Türkiye), and Margullan University (Kazakhstan). Prof. Khattak serves on international advisory boards at TEXAM (USA), USAK University (Türkiye), Baku State University (Azerbaijan), Harbin Institute of Technology (China) and International Research Network on AI Development and Governance (AIR Net), China. He is also member Senate/BOG at Mirpur University of Science and Technology, AJK and Islamabad University of Health Sciences and Emerging Technologies, Islamabad.

Received: May 24, 2026; Accepted: June 23, 2026.

Computer Science and Information Systems 23(4):1151–1180 https://doi.org/10.2298/CSIS260804043Q

Paper Type: Research Paper

SAMS-M: Explicit Sentiment-Guided Alignment and Multi-dimensional Mutual Supervision for Multimodal Sentiment Analysis

Shishu Qi[1], Yulei Zhang[2], Siyang Zhang[3], Jia Xu[1], Kuan-Ching Li[3], and Guangli Zhu[2]

[1] School of Automotive and Traffic Engineering, Tianjin University of Technology and Education
300222 Tianjin, China
{0721241391,0722231050}@tute.edu.cn
[2] School of Computer Science and Engineering, Anhui University of Science and Technology
232000 Huainan, China
2023201230aust.edu.cn, glzhu@aust.edu.cn (corresponding author)
[3] School of Mathematics and Big Data, Anhui University of Science and Technology
232000 Huainan, China
2025201731@aust.edu.cn,rob5li@outlook.com

Abstract. Multimodal Sentiment Analysis leverages the fusion of heterogeneous data to achieve fine-grained emotional understanding, which finds extensive application in large-scale public opinion monitoring and data mining. However, existing methods face two key challenges: (1) cross-modal alignment suffers from redundancy and semantic drift without explicit modeling of sentiment-critical cues, inducing spurious correlations; and (2) heterogeneous representation spaces lead to imbalanced modality contributions, particularly under weak image–text correlation or sentiment inconsistency. To address these challenges, we propose an explicit sentiment-guided alignment and multi-dimensional cross-modal mutual supervision-based model for multimodal sentiment analysis. The model primarily employs a fine-grained sentiment–saliency directed alignment mechanism, which leverages bidirectional cross-attention to couple textual sentiment cues with visual saliency, enabling precise localization of sentiment-relevant regions. Furthermore, we introduce a tripartite strong contrastive learning strategy to mitigate distribution discrepancies between heterogeneous modalities within a shared latent space, thereby enhancing cross-modal coherence and complementarity. Finally, we design a noise-robust gating-based fusion module, which, together with text augmentation and deep supervision, facilitates effective joint optimization. Experimental results show that SAMS-M obtains the best results on MVSA-Single and MSD and remains competitive on the noisier MVSA-Multiple benchmark; thus, the evidence supports strong but dataset-dependent performance rather than uniform state-of-the-art superiority.

Keywords: Multimodal Sentiment Analysis, Cross-modal Alignment, Contrastive Learning, Sentiment-guided, Noise-robust Fusion.

1. Introduction

With the continuous evolution of digital media technology, human emotional expression has increasingly exhibited significant multimodal characteristics [26,50]. The convergence of textual, visual, and other sensory modalities has established itself as a core paradigm of emotional communication [22,11,68]. Given its extensive applications in fields such as public opinion monitoring, intelligent interaction systems, and consumer behavior analysis, multimodal sentiment analysis has become a frontier topic of continuous research in both academia [45,15]. In recent years, the emergence of advanced pre-trained architectures, such as Transformer, BERT, and ViT, has fundamentally shifted the paradigm from shallow fusion to deep semantic alignment [24,67]. To further address the challenges of semantic complexity, recent studies have increasingly focused on fine-grained modeling and adaptive mechanisms. For instance, researchers have introduced token-level reconstruction and two-stage denoising frameworks to compensate for missing modalities and preserve robust semantic structures [57,72]. Furthermore, advanced approaches utilizing prototype-aligned calibration and grading-inspired complementary mechanisms have been proposed to dynamically balance the contributions of different modalities [13,17], thereby significantly improving the accuracy of sentiment representation in scenarios where multimodal information is incomplete or noisy.In addition, recent works have explored joint alignment, fusion, and denoising strategies by exploiting neighborhood information, which provides a more holistic view of multimodal context [5].

Despite these advancements, the field continues to confront significant hurdles regarding the robustness and precision of sentiment representation. Specifically, how to further enhance the reliability of emotional representation remains a subject worthy of in-depth exploration, especially when dealing with complex multimodal data characterized by image-text semantic discrepancies, pervasive background noise interference, and modal information asymmetry [23,49]. Existing mainstream paradigms, while effective at capturing general emotional cues, still have limitations for further optimization. First, regarding cross-modal alignment, traditional global attention mechanisms tend to perform dense alignment. Although some recent text-centric sparse interaction methods [70] have attempted to optimize this interaction process, the challenge of how to precisely focus on sentiment-critical elements in complex backgrounds—thereby mitigating spurious correlations between non-sentiment textual redundancy and irrelevant image background noise—remains a focal point of current research [42,40]. Unlike existing alignment strategies that mainly emphasize modality correspondence at the feature or token level, we argue that effective multimodal sentiment alignment should be guided by sentiment relevance rather than merely structural similarity. Therefore, explicitly identifying sentiment-bearing information before cross-modal interaction is essential for reducing irrelevant semantic association and improving sentiment-oriented representation learning. Second, the inherent heterogeneity in the semantic representations of text and images makes it difficult to fully map the consistency of their distribution spaces. Although existing contrastive learning strategies have mitigated the modality gap to some extent, there is still limitations for further exploration in maintaining global consistency and modal complementarity in fused features [46,12,66,1]. Moreover, most existing contrastive learning approaches primarily focus on reducing representation discrepancies between modalities, while insufficient attention has been paid to simultaneously preserving intra-modal semantics, inter-modal correspondence, and fused representation consistency. This moti-

vates us to construct a multi-dimensional mutual supervision strategy that provides complementary constraints from multiple perspectives. Furthermore, existing fusion strategies generally rely on fixed or manually designed modality interaction patterns, which may fail to adaptively handle modality imbalance and unreliable information caused by noisy or inconsistent multimodal inputs.

To address these challenges, we propose SAMS-M, an explicit sentiment-guided alignment and multi-dimensional cross-modal mutual supervision framework for multimodal sentiment analysis. Different from existing approaches that separately optimize cross-modal interaction, representation alignment, and modality fusion, SAMS-M establishes a progressive optimization paradigm consisting of three tightly coupled stages. First, instead of performing indiscriminate global interaction, SAMS-M introduces a sentiment-saliency directed alignment mechanism that explicitly identifies sentiment-bearing textual tokens and visually salient regions, thereby reducing irrelevant background interference during cross-modal association. Second, based on the aligned representations, a tripartite mutual supervision strategy is designed to simultaneously constrain textual semantics, visual semantics, and cross-modal consistency. This design prevents heterogeneous modalities from drifting into independent representation spaces and encourages complementary rather than redundant information exchange. Third, a noise-robust gating fusion mechanism dynamically adjusts modality contributions according to sample-specific reliability, allowing the model to adaptively suppress unreliable modalities under sentiment inconsistency scenarios. Therefore, the novelty of SAMS-M does not originate from the isolated adoption of individual components, but from establishing a progressive optimization framework where sentiment-guided alignment provides reliable cross-modal representations, mutual supervision enhances semantic consistency, and adaptive fusion further exploits complementary modality information. Through this tightly coupled design, SAMS-M simultaneously addresses sentiment localization, semantic distribution discrepancy, and modality imbalance in multimodal sentiment analysis.

We summarize the contribution of this paper as below:

(1) We propose SAMS-M, a unified multimodal sentiment analysis framework that integrates sentiment-guided alignment and multi-dimensional cross-modal mutual supervision. Different from conventional modality interaction strategies, SAMS-M explicitly considers sentiment relevance during alignment and progressively improves representation consistency through mutual supervision.

(2) We design a sentiment-saliency directed alignment mechanism and a tripartite mutual supervision strategy. The proposed alignment mechanism focuses cross-modal interaction on sentiment-relevant information, while the mutual supervision strategy jointly constrains intra-modal and inter-modal representations to reduce semantic inconsistency.

(3) We conduct extensive experiments on several public benchmark datasets, and the results demonstrate the effectiveness and competitiveness of SAMS-M compared with existing state-of-the-art methods.

2. Related work

2.1. Text sentiment detection

Before the widespread adoption of deep learning, traditional machine learning algorithms dominated the field of sentiment analysis. Moraes et al. [25] established the benchmark

for bag-of-words models in document-level sentiment classification by comparing Support Vector Machines (SVM) with Artificial Neural Networks (ANN). To address the deficiency of sentiment expressions in specific domains, Kang et al. [18] effectively improved the accuracy of restaurant review analysis by constructing domain-specific sentiment lexicons and optimizing the Naive Bayes algorithm. Building upon this, Wang et al. [39] proposed the SentiDiff algorithm, which pioneered the integration of textual features with sentiment diffusion patterns. By capturing emotional interaction logic within social networks through iterative computation, this approach demonstrated the early advantages of hybrid feature modeling. With the introduction of neural networks, models began to acquire the capability to automatically extract high-level semantic features. Chen [6] pioneered the application of Convolutional Neural Networks (CNN) for sentence classification, proving the superiority of convolution kernels in capturing local semantic features. Concurrently, Tai et al. [34] and Tang et al. [35] utilized Tree-Structured Long Short-Term Memory (Tree-LSTM) and Gated Recurrent Units (GRU), respectively, to process the hierarchical structure of text, thereby mitigating information loss in long-text modeling. To further enhance context sensitivity, Usama et al. [38] integrated the attention mechanism into a CNN-RNN hybrid model to achieve dynamic weighting of keywords, while Zhai et al. [58] further proposed the Multi-Attention Fusion Model (Multi-AFM), significantly enhancing semantic context capture in sentiment analysis scenarios involving educational big data.

In recent years, the research focus has shifted toward deep semantic understanding and fine-grained feature modeling. The SKEP model proposed by Tian et al. [36] refreshed several benchmarks by injecting sentiment knowledge into pre-trained representations via a sentiment masking mechanism. Dai et al. [7] verified the supportive role of syntactic information in Aspect-Based Sentiment Analysis (ABSA), demonstrating the potential of RoBERTa in refined sentiment recognition. Related studies further exploit graph structure and task-specific affective knowledge: TS-GCN models aspect-level dependencies in consumer reviews, sentiment-driven 3D-CNNs extract emotion–cause pairs, syntactically weighted textual graphs improve implicit sentiment recognition, and sequence-labeling pipelines expand sentiment lexicons from product reviews [64,55,60,63]. In terms of expanding application scenarios, researchers have proposed diverse solutions: Bibi et al. [3] optimized Twitter sentiment analysis using an unsupervised ensemble framework; Parveen et al. [28] constructed a Gated Attention Recurrent Network (GARN); and Zhang et al. [65] introduced PU learning to address the problem of fake review detection. Furthermore, Wang et al. [43] realized the automated construction of fine-grained sentiment lexicons via neural networks, while Rodríguez-Ibáñez et al. [30] and Pashchenko et al. [29] further refined the analytical dimensions of consumer reviews from the perspectives of platform synthesis and multi-dimensional attributes (such as sentiment and normativity), respectively.

In summary, although text-based sentiment analysis has achieved remarkable progress driven by deep learning, and models demonstrate excellent performance in specific semantic understanding, social media content in reality often exhibits high complexity. A single textual modality is frequently insufficient to fully encapsulate the rich human emotional expression (e.g., composite emotions accompanying images and audio). Therefore, breaking down modal barriers and integrating cross-modal information through deep learning to achieve more precise emotional understanding has become a core challenge in

the field of affective computing, which serves as the primary motivation for this study on multimodal sentiment analysis.

2.2. Multi-modal sentiment detection

Recent multimodal sentiment analysis increasingly centers on alignment and adaptive fusion. MTVAF performs multi-level textual–visual alignment and fusion, CoDSF uses confidence-guided dynamic sequential fusion, KuDA dynamically selects the dominant modality using sentiment knowledge, and SMCAF introduces semantic-guided multi-grained cross-modal alignment and fusion [19, 48, 10, 21]. A multimodal semantic fusion network with explicit cross-modal alignment provides further evidence that structured image–text correspondence is central to sentiment representation [61]. Beyond conventional polarity classification, topic-oriented knowledge prompting improves multimodal sarcasm detection, while emoji-enhanced multi-feature fusion supports socially implicit affective recognition [71, 62]. These studies demonstrate the value of fine-grained correspondence and sample-dependent modality weighting. In contrast, SAMS-M couples sentiment-saliency-guided alignment with three simultaneous text–fusion, image–fusion, and text–image constraints, and then uses a reliability-aware gate on the jointly regularized representations. This mechanism-level combination differs from CLMLF's two contrastive tasks, MVCN's calibration-oriented treatment of heterogeneity, CTMWA's translation-based weight adaptation, and ITQT-MSA's quantum-transformer interaction: SAMS-M explicitly makes sentiment relevance the input to alignment, uses the fused representation as a shared contrastive anchor, and suppresses unreliable modalities only after this mutual supervision. Early research in Multimodal Sentiment Analysis (MSA) primarily focused on exploring shallow interactions between modalities. To mitigate the limitations of unimodal analysis, researchers proposed various fusion architectures. Zhao et al. [69] were among the first to improve classification performance by exploring the consistency between images and text. Subsequently, Truong et al. [37] introduced a visual attention mechanism to leverage image information for highlighting key sentiment segments in text. To further characterize long- and short-term dependencies, Xu et al. [54] proposed a multi-interaction memory network, effectively supervising cross-modal data flows. Recently, to capture deeper correlations, Wang et al. [41] and Wu et al. [45] introduced manifold learning and graph structures, enabling synchronized modeling of intra- and inter-modal information, which laid the foundation for subsequent complex interaction networks.

With the progression of research, addressing heterogeneity and noise interference in multimodal data has become a new focal point. To cope with semantic asymmetry between text and images, Wu et al. [46] effectively balanced inter-modal consistency via contrastive learning. Addressing interference issues in fine-grained sentiment analysis, Wang et al. [42] and Wang et al. [40] proposed strategies to filter redundant information, aiming to reduce the negative impact of background noise on sentiment determination. Furthermore, Han et al. [12] and An et al. [2] leveraged graph neural networks and gating mechanisms to enhance model robustness in complex scenarios. However, these graph-based approaches mainly focus on modeling structural dependencies and reducing noisy interactions, while explicit sentiment-oriented guidance for cross-modal alignment remains insufficiently explored. Graph-based approaches generally model multimodal samples as structured interaction graphs and alleviate noise by refining node rela-

tionships or information propagation patterns. However, these methods mainly focus on structural-level noise reduction and graph topology optimization, while the identification of sentiment-relevant information during cross-modal interaction remains insufficiently explored. Zhao et al. [66] further improved the mapping quality of different modal features within a unified semantic space through calibration alignment mechanisms. Currently, utilizing large-scale pre-trained models to inject external knowledge has become a mainstream direction for boosting MSA performance. Hu et al. [15] and Liu et al. [24] explored methods using social relationships and label semantics to enhance feature representation. For extreme scenarios involving resource constraints or missing modalities, Afroze et al. [1] and Zhao et al. [67] proposed completion strategies based on cross-attention mechanisms and virtual modality generation, respectively. Moreover, Liu et al. [23] and Xiong et al. [49] demonstrated significant representation advantages in anomaly sentiment detection tasks by introducing self-supervised knowledge enhancement and multi-granularity selection mechanisms.

Although the aforementioned methods have achieved significant success in multimodal feature fusion and robustness enhancement, existing paradigms still leave limitations for further optimization when dealing with highly complex multimodal data. First, most global attention-based methods struggle to completely avoid the introduction of background noise and sentiment-irrelevant redundant information during feature alignment, which leads to semantic deviation during cross-modal interaction. Moreover, graph-based methods still lack sufficient consideration of sentiment-related importance during modality interaction. Second, for samples with contradictory image-text semantics or imbalanced modal contributions, simple fusion strategies fail to achieve precise consistency constraints in the modal representation space, often leading to performance fluctuations when processing heterogeneous features. These challenges indicate that designing a mechanism capable of fine-grained sentiment-guided and multi-dimensional mutual supervision alignment holds significant research value for enhancing the robustness of multimodal sentiment analysis in real-world scenarios.

3. Methodology

3.1. Problem Definition

Let the corpus be denoted by $\mathcal{D} = \{(\mathbf{x}_i^t, \mathbf{x}_i^v, y_i)\}_{i=1}^N$, where $\mathbf{x}_i^t$ and $\mathbf{x}_i^v$ represent textual and visual inputs, respectively. The label space $\mathcal{C}$ is determined according to the annotation protocol of each benchmark. Specifically, $\mathcal{C} = \{L_{neg}, L_{neu}, L_{pos}\}$ is adopted for three-class sentiment datasets, while binary label spaces are employed for datasets without neutral annotations, such as MSD. Let $C = |\mathcal{C}|$. We learn a multimodal discriminative mapping $\mathcal{M}(\cdot, \cdot; \Theta)$ parameterized by $\Theta = \{\theta_\phi, \theta_\psi, \theta_\gamma\}$. The feature representation for each modality $m \in \{t, v\}$ is computed via the extraction operator $\mathbf{z}_i^m = \phi_m(\mathbf{x}_i^m; \theta_m)$. These heterogeneous features are synthesized into a joint representation $\mathbf{h}_i \in \mathbb{R}^{d_h}$ via the cross-modal alignment operator $\psi(\cdot, \cdot; \theta_\psi)$, defined as $\mathbf{h}_i = \psi(\mathbf{z}_i^t, \mathbf{z}_i^v; \theta_\psi)$. The final prediction is a posterior distribution over the $(C-1)$-dimensional probability simplex Δ^{C-1}:

$$\hat{\mathbf{y}}_i = \mathrm{softmax}\left(\mathbf{W}_\gamma^\top \mathbf{h}_i + \mathbf{b}_\gamma\right) \in \Delta^{C-1} \tag{1}$$

where $\mathbf{W}_\gamma$ and $\mathbf{b}_\gamma$ are the classifier parameters. The optimal parameter set Θ^* is obtained by minimizing the regularized empirical risk:

$$\mathcal{L}(\Theta) = \frac{1}{N}\sum_{i=1}^{N} \ell_{\text{focal}}\left(\hat{\mathbf{y}}_i, y_i\right) + \frac{\lambda}{2}\|\Theta\|_2^2 \tag{2}$$

where $\ell_{\text{focal}}(\cdot,\cdot)$ denotes the Focal Loss used in our implementation and λ is the weight-decay coefficient. This generic risk is expanded into the complete multi-branch training objective in the classification and deep-supervision module.

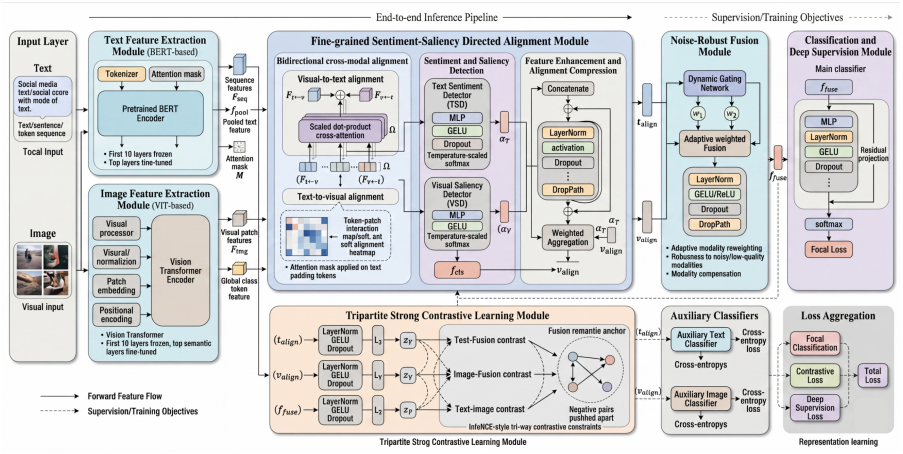


Fig. 1. Overall framework.

3.2. Overall Model Architecture

The overall architecture of our model is illustrated in Figure 1. First, the model extracts deep semantic features for text and images through their respective extraction modules. Second, a fine-grained sentiment-saliency directed alignment module is employed to achieve precise alignment between textual sentiment keywords and visual sentiment-salient regions, yielding the aligned text and image features. Then, a tripartite strong contrastive learning module is utilized to construct multi-way contrastive constraints to optimize the distribution of cross-modal features. Subsequently, a noise-robust fusion module dynamically fuses the aligned features to obtain multimodal fused representations. Finally, sentiment classification is performed via a classification module, combined with a deep supervision strategy (independent classification for text and images) to further optimize model performance.

3.3. Text and Image Feature Extraction Modules

To extract deep semantic features for text and images, this paper constructs text and image feature extraction modules based on pre-trained models. By freezing specific pre-trained

layers and fine-tuning the top layers, we balance feature extraction capability with training efficiency.

Text Feature Extraction Module The text feature extraction module utilizes BERT [8] as the core architecture to map the textual modality T_i into a high-dimensional semantic manifold space. First, the raw text sequence is converted into a discretized token sequence and its corresponding attention mask $\mathbf{M} \in \{0,1\}^L$ via a tokenizer $\text{Trans}(\cdot)$. Subsequently, the encoded sequence is fed into a deep Transformer hierarchical structure for bidirectional context modeling. To strike a balance between computational efficiency and representational power, we implement a selective parameter freezing strategy by masking the gradient updates of the first k layers (set to $k = 10$ in this study). This allows the model to retain universal low-level linguistic features while enabling specific transfer learning for sentiment semantics through top-layer fine-tuning. The textual feature extraction process can be represented as a composite mapping. Let $\mathbf{H} \in \mathbb{R}^{L\times d}$ be the hidden state sequence output from the last layer of BERT, where d denotes the hidden dimension. The feature representation of the textual modality is formally defined as:

$$\Phi_T(T_i) = \text{BERT}_{\Theta_T}(\text{Trans}(T_i)) = \{\mathbf{F}_{seq}, \mathbf{f}_{pool}, \mathbf{M}\} \tag{3}$$

where $\mathbf{F}_{seq} \in \mathbb{R}^{B\times L\times d}$ represents the sequence feature matrix containing fine-grained semantic information for subsequent cross-modal alignment; $\mathbf{f}_{pool} \in \mathbb{R}^{B\times d}$ is the global semantic vector extracted via a pooling operator, representing the macro-level sentiment expression of the textual modality; Θ_T denotes the subset of trainable parameters in the text branch; and $\mathbf{M}$ is used to mask the influence of padding tokens in the cross-modal attention mechanism, ensuring the effectiveness of feature fusion.

Image Feature Extraction Module The image feature extraction module aims to transform the input visual modality V_i into feature representations containing spatial context and global semantics. This module adopts the Vision Transformer (ViT) [9] as the backbone to capture discriminative emotional regions in images with its powerful global modeling capability. During data preprocessing, input images are normalized by a dedicated visual processor $\text{Proc}(\cdot)$, explicitly preserving the original spatial scale information to minimize underlying feature loss. After patch embedding and position encoding, the images are mapped into a latent space tensor and fed into the Transformer encoder. To suppress overfitting in small-sample scenarios, this module adopts a heuristic parameter freezing strategy by locking the weights of the first k layers (set to $k = 10$), applying discriminative fine-tuning only to the high-level semantic feature extraction layers. The image feature extraction process can be formulated as:

$$\Phi_V(V_i) = \text{ViT}_{\Theta_V}(\text{Proc}(V_i)) = \{\mathbf{F}_{img}, \mathbf{f}_{cls}\} \tag{4}$$

where $\mathbf{F}_{img} \in \mathbb{R}^{B\times P\times d_v}$ represents the visual sequence features, composed of P encoded vectors of image patches, which support subsequent fine-grained cross-modal alignment; $\mathbf{f}_{cls} \in \mathbb{R}^{B\times d_v}$ is the pooled vector extracted from the special class token ([CLS] token), serving as the global discriminative feature of the visual modality; and Θ_V represents the trainable parameter space of the visual branch.

3.4. Fine-grained Sentiment-Saliency Directed Alignment Module

While previous studies have achieved progress through hierarchical cross-modal modeling [16] and adaptive multi-loss fusion frameworks [14], they often struggle to precisely filter task-irrelevant background noise in fine-grained analysis. To address this, we introduce the sentiment-saliency directed alignment module, which leverages the aforementioned interaction mechanisms while imposing an explicit guidance constraint to focus the cross-modal alignment on sentiment-critical regions. We define the interaction between the textual hidden sequence $\mathbf{H}_T \in \mathbb{R}^{L\times d}$ and the visual patch representation $\mathbf{H}_V \in \mathbb{R}^{P\times d}$ as a bidirectional manifold mapping. We introduce the learnable projection parameters $\mathbf{W}_Q^\star, \mathbf{W}_K^\star, \mathbf{W}_V^\star \in \mathbb{R}^{d\times d_k}$ to project the input features into a shared subspace of dimension d_k. The cross-modal interaction is governed by the scaled dot-product attention operator $\mathcal{A}$, which maps a query set $\mathcal{Q}$ against a key-value pair $(\mathcal{K}, \mathcal{V})$ subject to an optional masking tensor $\mathcal{M} \in \{0, -\infty\}^{L\times P}$, defined as:

$$\mathcal{A}(\mathbf{Q}, \mathbf{K}, \mathbf{V}; \mathcal{M}) = \text{softmax}\left(\frac{(\mathbf{H}_Q\mathbf{W}_Q)(\mathbf{H}_K\mathbf{W}_K)^\top}{\sqrt{d_k}} + \mathcal{M}\right)(\mathbf{H}_V\mathbf{W}_V) \tag{5}$$

Within this topological framework, the bidirectional sentiment-saliency flow is realized by coupling the modalities through asymmetric query-key-value associations. The visual-to-textual alignment branch $\mathbf{F}_{t\leftarrow v}$ selectively aggregates visual contexts to modulate textual representations, while the textual-to-visual alignment branch $\mathbf{F}_{v\leftarrow t}$ utilizes textual sentiments to attend to salient visual regions. This bidirectional coupling is formulated as the coupled mapping system:

$$\begin{cases} \mathbf{F}_{t\leftarrow v} = \mathcal{A}\left(\mathbf{H}_T, \mathbf{H}_V, \mathbf{H}_V; \mathbf{M}\right) \in \mathbb{R}^{L\times d} \\ \mathbf{F}_{v\leftarrow t} = \mathcal{A}\left(\mathbf{H}_V, \mathbf{H}_T, \mathbf{H}_T; \mathbf{0}\right) \in \mathbb{R}^{P\times d} \end{cases} \tag{6}$$

where $\mathbf{M}$ serves as the indicator tensor to inhibit the influence of padding tokens in the text modality, effectively filtering non-sentiment-bearing semantics. The resultant aligned representations are computed as high-order semantic features, where the attention weight distribution $\mathbf{\Omega} \in \mathbb{R}^{L\times P}$ is derived as:

$$\mathbf{\Omega}_{ij} = \left[\text{softmax}\left(\frac{(\mathbf{H}_{T,i}\mathbf{W}_Q^t)(\mathbf{H}_{V,j}\mathbf{W}_K^v)^\top}{\sqrt{d_k}}\right)\right]_{i,j} \tag{7}$$

This formulation enables the model to dynamically construct a bipartite graph of sentiment-salient dependencies, ensuring that the decision fusion module operates on a joint feature manifold where discriminative cross-modal correlations are explicitly reinforced.

Sentiment and Saliency Detection Different from conventional attention mechanisms that estimate importance solely based on feature activation, the proposed sentiment-saliency detectors aim to identify modality elements that contribute more significantly to sentiment prediction. Specifically, the detectors do not require additional sentiment annotations to

generate saliency weights. Instead, the sentiment-related importance distribution is adaptively learned from modality representations under the supervision of the overall sentiment classification objective. Therefore, the proposed mechanism can be directly applied to both training and inference stages.

To formalize the sentiment-oriented disentanglement process with higher analytical rigor, we define the TSD and VSD as non-linear energy density estimators that map the hidden manifolds $\mathbf{H}_k \in \mathbb{R}^{N_k \times d}$ to the probability simplex Δ^{N_k-1} of sentiment importance weights. We model this as a composition of affine transformations and non-linear activation manifolds, regulated by a temperature-controlled Softmax operator to modulate the sparsity of the learned saliency distribution. Let $\Theta_k = \{\mathbf{W}_k^{(1)} \in \mathbb{R}^{d \times d_{hid}}, \mathbf{W}_k^{(2)} \in \mathbb{R}^{d_{hid} \times 1}, \mathbf{b}_k^{(1)} \in \mathbb{R}^{d_{hid}}, \mathbf{b}_k^{(2)} \in \mathbb{R}\}$ denote the parameter set for the k-th modality, where $k \in \{T, V\}$. The logit score $z_{k,i}$ for the i-th element in modality k is derived via the hierarchical projection function $\mathcal{G}_k(\cdot;\Theta_k) : \mathbb{R}^{N_k \times d} \to \mathbb{R}^{N_k}$, defined as:

$$z_{k,i} = \left[\text{Dropout}\left(\sigma_{\text{GELU}}\left(\mathbf{H}_k \mathbf{W}_k^{(1)} + \mathbb{1}_{N_k}(\mathbf{b}_k^{(1)})^\top\right)\right)\mathbf{W}_k^{(2)} + \mathbf{b}_k^{(2)}\right]_i \tag{8}$$

where $\sigma_{\text{GELU}}(\cdot)$ represents the Gaussian Error Linear Unit activation, and $\mathbb{1}_{N_k} \in \mathbb{R}^{N_k \times 1}$ denotes the all-ones vector for bias broadcasting. To regulate the entropy of the resulting saliency distribution $\boldsymbol{\alpha}_k$, we introduce a temperature hyperparameter $\tau \in \mathbb{R}^+$, enabling the model to transition between a uniform distribution and a sharp one-hot selection of key emotional anchors. The final normalized weight vector $\boldsymbol{\alpha}_k \in \Delta^{N_k-1}$ is computed by the tempered Softmax operator:

$$\alpha_{k,i} = \frac{\exp\left(\tau^{-1} \cdot z_{k,i}\right)}{\sum_{j=1}^{N_k} \exp\left(\tau^{-1} \cdot z_{k,j}\right)}, \quad \forall i \in \{1, \ldots, N_k\} \tag{9}$$

This formulation explicitly decomposes the global latent features into a weighted convex combination of discriminative primitives, such that the reconstructed modality representation $\tilde{\mathbf{H}}_k = \sum_{i=1}^{N_k} \alpha_{k,i} \cdot \mathbf{H}_{k,i}$ maximizes the semantic signal-to-noise ratio. By optimizing the distribution $\boldsymbol{\alpha}_k$ through the global objective, the detector acts as an adaptive sentiment-aware filter, assigning larger weights to modality elements that are more relevant to sentiment prediction while reducing the influence of less informative regions or tokens.

It is worth noting that the proposed TSD and VSD are trained jointly with the downstream sentiment objective rather than relying on external sentiment annotations. Consequently, the generated saliency distributions represent task-oriented importance estimation, which enables the alignment module to focus on sentiment-relevant cross-modal interactions.

Feature Enhancement and Alignment To rigorously characterize the cross-modal feature evolution and the subsequent transformation into the decision-making manifold, we define the concatenation of the raw semantic state $\mathbf{H}_k \in \mathbb{R}^{N_k \times d}$ and the interaction-aware alignment flow $\mathbf{F}_{k \leftarrow \bar{k}} \in \mathbb{R}^{N_k \times d}$ as the augmented bipartite feature manifold $\mathcal{Z}_k \in \mathbb{R}^{N_k \times 2d}$, where $\mathcal{Z}_k = [\mathbf{H}_k \oplus \mathbf{F}_{k \leftarrow \bar{k}}]$. This augmented tensor undergoes a hierarchical refinement through the non-linear enhancement operator $\Psi_k(\cdot;\Theta_k)$, which we model as a deep residual topology consisting of L stacked transformation units $\{\psi_k^{(l)}\}_{l=1}^{L}$. For a

given layer l, the evolution of the latent representation $\mathbf{H}_k^{(l)} \in \mathbb{R}^{N_k \times d}$ is governed by the composite mapping of layer normalization, stochastic regularization, and high-order semantic projection, defined by the following recurrent recurrence relation:

$$\mathbf{H}_k^{(l+1)} = \mathrm{LN}\left(\mathbf{H}_k^{(l)} + I_{\mathrm{D}}^{(l)}\left(\sigma\left(\mathbf{H}_k^{(l)}\mathbf{W}_k^{(l,1)} + \mathbb{1}_{N_k}(\mathbf{b}_k^{(l,1)})^\top\right)\mathbf{W}_k^{(l,2)} + \mathbb{1}_{N_k}(\mathbf{b}_k^{(l,2)})^\top\right)\right) \tag{10}$$

where $\mathbf{W}_k^{(l,1)} \in \mathbb{R}^{d \times d_h}$ and $\mathbf{W}_k^{(l,2)} \in \mathbb{R}^{d_h \times d}$ are the learnable weight kernels at depth l, $\sigma(\cdot)$ is the non-linear activation manifold, and $I_{\mathrm{D}}^{(l)}(\cdot)$ denotes the stochastic survival probability operator parameterized by $\rho^{(l)} \in [0, 1]$. Upon the convergence of this enhancement chain, the terminal representation $\mathbf{H}_k^{enh} = \mathbf{H}_k^{(L)} \in \mathbb{R}^{N_k \times d}$ encapsulates the refined cross-modal semantics. To perform the targeted compression of this variable-length sequence into the fixed-dimensional sentiment alignment vector $\mathbf{v}_k \in \mathbb{R}^d$, we treat the saliency weight distribution $\boldsymbol{\alpha}_k \in \Delta^{N_k-1}$ derived from the TSD and VSD modules as a probability mass function over the manifold. The projection into the global alignment hyperspace is realized by the inner product of the sequence tensor and the saliency simplex, formulated as the expectation of the latent states:

$$\mathbf{v}_k = \mathbb{E}_{\boldsymbol{\alpha}_k}\left[\mathbf{H}_k^{enh}\right] = \sum_{i=1}^{N_k} \alpha_{k,i} \cdot \mathbf{h}_{k,i}^{enh} = \left(\boldsymbol{\alpha}_k\right)^\top \mathbf{H}_k^{enh} \tag{11}$$

where $\mathbf{h}_{k,i}^{enh} \in \mathbb{R}^d$ denotes the i-th row vector of $\mathbf{H}_k^{enh}$ corresponding to the i-th token or image patch. This linear combination operation projects the localized, high-resolution feature sequence $\mathbf{H}_k^{enh}$ onto a compressed global coordinate system, effectively collapsing the N_k-dimensional temporal or spatial variance into a singular d-dimensional representation $\mathbf{v}_k$. Consequently, the dual alignment vectors $\{\mathbf{t}_{align}, \mathbf{v}_{align}\}$ constitute the optimal low-rank projection of the heterogeneous input data, providing a robust, high-fidelity feature basis for the final classification objective $\mathcal{L}(\mathrm{softmax}(\mathbf{W}_{cls}[\mathbf{t}_{align} \oplus \mathbf{v}_{align}] + \mathbf{b}_{cls}), y)$.

3.5. Tripartite Strong Contrastive Learning Module

To address the heterogeneity and inconsistency in modal representation spaces, this paper designs a tripartite strong contrastive learning module. By constructing three-way contrastive constraints—"Text-Fusion", "Image-Fusion", and "Text-Image"— this module encourages heterogeneous modalities to achieve semantic consistency in a shared latent space by constructing complementary contrastive constraints, enhancing the cohesion and complementarity of cross-modal features. The core of this module lies in the design of the contrastive loss function, detailed as follows:

Feature Projection To transcend the repetitive formulation and elevate the mathematical rigor of the projection process, we treat the modality projection as a homomorphic mapping family $\{\mathcal{P}_k\}_{k \in \mathcal{K}}$ from the heterogeneous input manifolds $\mathcal{X}_k \subset \mathbb{R}^d$ to a unified, hyperspherical latent space $\mathcal{Z} \subset \mathbb{S}^{d_{sub}-1} \subset \mathbb{R}^{d_{sub}}$, where $\mathcal{K} = \{T, V, F\}$.

Instead of enumerating individual equations, we formalize the projection as a universal non-linear operator $\mathcal{P}_k : \mathcal{X}_k \to \mathcal{Z}$ acting on an arbitrary input $\mathbf{x}_k \in \mathcal{X}_k$, where $\mathbf{x}_k \in \{\mathbf{t}_{align}, \mathbf{v}_{align}, \mathbf{f}_{fuse}\}$. The projection pipeline is decomposed into a latent feature

transformation network $\Upsilon_k(\cdot;\Theta_k)$ followed by a geometric normalization operator $\Gamma(\cdot)$ that constrains the resultant vector to the unit hypersphere.

Let $\Theta_k = \{\mathbf{W}_k^{(l)}, \mathbf{b}_k^{(l)}\}_{l=1}^2$ denote the learnable parameter set for modality k. The unified projection mapping is defined as:

$$\mathbf{z}_k = \Gamma\left(\mathcal{D}_\eta\left(\sigma_{\text{GELU}}\left(\text{LN}\left(\mathbf{x}_k\mathbf{W}_k^{(1)} + \mathbb{1}_N(\mathbf{b}_k^{(1)})^\top\right)\right)\mathbf{W}_k^{(2)} + \mathbb{1}_N(\mathbf{b}_k^{(2)})^\top\right)\right) \tag{12}$$

where $\Gamma : \mathbb{R}^{d_{sub}} \to \mathbb{S}^{d_{sub}-1}$ represents the isometric normalization operator $\Gamma(\mathbf{u}) = \frac{\mathbf{u}}{\|\mathbf{u}\|_2+\epsilon}$, $\mathbb{1}_N \in \mathbb{R}^{N\times 1}$ is the broadcasting vector for bias alignment, and $\mathcal{D}_\eta(\cdot)$ denotes the stochastic dropout regularization with retention probability $1-\eta$.

To explicitly characterize the transformation of the complete modality set $\mathcal{X} = \{\mathbf{x}_k\}_{k\in\mathcal{K}}$, we define the global projection operator $\mathcal{P}_{\text{global}} : \mathcal{X} \to \mathcal{Z}^{|\mathcal{K}|}$ as the Cartesian product of individual mappings:

$$\mathcal{Z}_{\text{latent}} = \prod_{k\in\mathcal{K}} \mathcal{P}_k(\mathbf{x}_k) = \{\mathbf{z}_k \mid \mathbf{z}_k = \mathcal{P}_k(\mathbf{x}_k), \forall k \in \mathcal{K}\} \tag{13}$$

This structural formulation highlights that the mapping is an invariant transformation across the modality index k, effectively mitigating the feature distribution divergence $\mathcal{D}_{\text{div}}(\mathcal{X}_k, \mathcal{X}_j)$ for any $k \neq j$. By enforcing the hyperspherical constraint $\|\mathbf{z}_k\|_2 = 1$, we ensure that the subsequent contrastive semantic similarity metric, typically defined via the inner product $\langle \mathbf{z}_i, \mathbf{z}_j \rangle$, is equivalent to the cosine similarity measure $\cos(\theta_{ij})$, thereby providing a well-conditioned geometric foundation for the optimization of the cross-modal alignment objective.

Tri-way Contrastive Constraints To enforce global manifold consistency and topological alignment within the heterogeneous latent space $\mathcal{Z}$, we instantiate a Tri-way Contrastive Constraint mechanism. This objective function is formulated to optimize the inter-modal semantic proximity while repelling spurious correspondences, effectively minimizing the empirical risk over the multimodal product space. Let $\mathcal{S}_\tau(\cdot,\cdot) : \mathbb{S}^{d_{sub}-1} \times \mathbb{S}^{d_{sub}-1} \to \mathbb{R}^+$ denote the temperature-scaled radial basis similarity kernel, defined as $\mathcal{S}_\tau(\mathbf{u},\mathbf{v}) = \exp\left(\tau^{-1}\langle \mathbf{u},\mathbf{v}\rangle\right)$, where $\tau \in \mathbb{R}^+$ is the temperature hyperparameter controlling the softmax sharpness.

We define the cross-modal contrastive loss $\mathcal{L}_{k,F}$ for $k \in \{T, V\}$ using an InfoNCE-style in-batch retrieval objective. Each matched multimodal pair is treated as positive and the remaining samples in the mini-batch provide negatives. This objective directly improves paired representation discrimination; we do not interpret its empirical value as a direct estimate of Jensen–Shannon divergence. The optimization objective for the primary alignment branches is given by:

$$\mathcal{L}_{k,F} = -\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{\mathcal{S}_\tau(\mathbf{z}_{k,i}, \mathbf{z}_{F,i})}{\sum_{j=1}^{N}\mathcal{S}_\tau(\mathbf{z}_{k,i}, \mathbf{z}_{F,j})}\right), \quad \forall k \in \{T, V\} \tag{14}$$

To exploit the correspondence between textual and visual streams, we augment this objective with a symmetric cross-modal coupling loss $\mathcal{L}_{T,V}$. The term encourages matched

image–text pairs to be more similar than mismatched in-batch pairs, thereby reducing the empirical representation gap without requiring a separate mutual-information estimator:

$$\mathcal{L}_{T,V} = -\frac{1}{2N}\sum_{i=1}^{N}\left[\log\left(\frac{\mathcal{S}_\tau(\mathbf{z}_{T,i},\mathbf{z}_{V,i})}{\sum_{j=1}^{N}\mathcal{S}_\tau(\mathbf{z}_{T,i},\mathbf{z}_{V,j})}\right)+\log\left(\frac{\mathcal{S}_\tau(\mathbf{z}_{V,i},\mathbf{z}_{T,i})}{\sum_{j=1}^{N}\mathcal{S}_\tau(\mathbf{z}_{V,i},\mathbf{z}_{T,j})}\right)\right] \tag{15}$$

The global contrastive objective $\mathcal{J}_{\text{con}}(\Theta)$ is consolidated as a convex combination of these partial manifolds, modulated by the weighting hyperparameters $\lambda_1, \lambda_2 \in \mathbb{R}^+$. The final composite functional is expressed as:

$$\mathcal{J}_{\text{con}}(\Theta) = \lambda_1 \sum_{k\in\{T,V\}} \mathcal{L}_{k,F} + \lambda_2\mathcal{L}_{T,V} + \lambda_3\mathcal{R}(\Theta) \tag{16}$$

where $\mathcal{R}(\Theta) = \frac{1}{2}\|\Theta\|_2^2$ is a Frobenius-norm regularizer. The fused representation serves as a shared anchor for the unimodal branches, while the symmetric text–image term maintains direct pairwise correspondence. Joint optimization therefore encourages, but does not theoretically guarantee, class-discriminative and cross-modally coherent representations.

Computational Complexity Discussion. Although the proposed tripartite contrastive learning module introduces additional optimization objectives, its computational overhead remains manageable due to the batch-wise contrastive formulation. Specifically, the computational complexity of each contrastive constraint is mainly determined by the similarity matrix construction between modality representations, which is $\mathcal{O}(N^2 d_{sub})$ for a batch size of N. Since the proposed module shares the same projected representations among the three contrastive branches, the additional computation mainly comes from calculating the corresponding similarity matrices rather than introducing extra feature extraction networks. Therefore, the proposed mutual supervision strategy improves cross-modal representation consistency with limited additional computational cost.

3.6. Noise-Robust Fusion Module

To proactively mitigate the performance degradation caused by heterogeneous noise and the pervasive bottleneck of modality imbalance, we design an adaptive fusion module governed by a gating mechanism. The fundamental motivation for this design is to achieve reliability-aware information routing [32], thereby ensuring that the model selectively aggregates modality-specific information based on its real-time representation quality. Specifically, by feeding the aligned dual-modal features $\mathbf{t}_{align}$ and $\mathbf{v}_{align}$ into a deep gating network $\mathcal{G}_{gate}(\cdot)$, we generate a modality reliability distribution $\mathbf{w} \in \mathbb{R}^2$. This mechanism functions as a dynamic filter: when specific modalities are compromised by sensor noise, environmental interference, or semantic ambiguity, the network adaptively attenuates the activation of corrupted signals while prioritizing high-fidelity inputs. Consequently, this dynamic compensation strategy effectively balances the signal-to-noise ratio in the feature space, stabilizing the model's performance under unstable input conditions.

Upon obtaining the initial fused feature $\mathbf{f}_{init}$, we further introduce a fusion enhancement operator $\mathcal{E}_{fuse}(\cdot)$ to pursue a deeper interweaving of heterogeneous semantics. Motivated by the principles of manifold structure modeling [33], the core objective of this

refinement process is to maximize the exploitation of high-order complementary correlations between modalities. By integrating regularization strategies—including Layer Normalization, GELU activation, Dropout, and DropPath—this operator acts as a semantic stabilizer. This design not only facilitates the extraction of complex inter-modal dependencies but also enforces discriminative consistency across challenging decision boundaries, preventing feature collapse through the introduction of structural stochastic perturbations.

The noise-robust fusion process can be formally described as the following mapping system:

$$\omega = [w_1, w_2] = \sigma\left(\text{Drop}\left(\delta\left([\mathbf{t}_{align} \parallel \mathbf{v}_{align}]\mathbf{W}_{g1} + \mathbf{b}_{g1}\right)\right)\mathbf{W}_{g2} + \mathbf{b}_{g2}\right) \tag{17}$$

$$\mathbf{f}_{init} = w_1 \cdot \mathbf{t}_{align} + w_2 \cdot \mathbf{v}_{align} \tag{18}$$

$$\mathbf{f}_{fuse} = \mathcal{E}_{fuse}(\mathbf{f}_{init}; \Theta_f) \tag{19}$$

where $\sigma(\cdot)$ is the Softmax operator ensuring weight normalization, satisfying $\sum_{k=1}^{2} w_k = 1$; $\delta(\cdot)$ represents the ReLU activation function; and $[\cdot \parallel \cdot]$ denotes the concatenation of modal features. The resulting fused vector $\mathbf{f}_{fuse} \in \mathbb{R}^d$ aggregates global sentiment features that have been filtered by gating and semantically enhanced, providing a noise-robust feature foundation for subsequent contrastive constraint learning and sentiment polarity classification.

3.7. Classification and Deep Supervision Module

To further refine the discriminative accuracy of sentiment polarity and optimize the training dynamics of the complex deep network, we construct an Enhanced Classifier integrated with residual mapping and a multi-branch Deep Supervision Strategy. The Enhanced Classifier performs non-linear mapping on the multimodal fused features $\mathbf{f}_{fuse}$ through a composite architecture of Multi-Layer Perceptrons (MLP), LayerNorm, GELU activation, and Dropout operators. To enhance the generalization performance of classification boundaries, we introduce a residual projection branch $\mathcal{R}(\cdot)$, which directly adds the original fused features to the classification space via linear mapping. This structure not only establishes an information shortcut across layers, effectively mitigating gradient degradation during backpropagation, but also ensures that the core sentiment semantics within the fused representation can be fed back to the decision layer via shorter paths.

At the level of training optimization, we introduce a deep supervision strategy to strengthen the discriminative efficacy of unimodal representations. By constructing auxiliary textual classification mappings $\Psi_T(\cdot)$ and visual classification mappings $\Psi_V(\cdot)$, we perform parallel discrimination on the aligned hidden state vectors $\mathbf{t}_{align}$ and $\mathbf{v}_{align}$. This strategy introduces explicit constraints during the shallow feature extraction stage, preventing unimodal training degradation caused by the information dominance effect during the multimodal fusion process. Through this multi-branch error backpropagation, the model guides the cross-modal alignment module to more specifically mine the discriminative evidence unique to each modality, thereby enhancing the overall feature representation quality and decision robustness of the system.

The total objective function $\mathcal{L}_{total}$ is defined as a multi-task joint optimization loss, aiming to achieve a global optimum across feature alignment, semantic fusion, and sentiment classification:

$$\hat{\mathbf{y}} = \text{Softmax}\left(\mathcal{C}(\mathbf{f}_{fuse}; \Theta_c) + \mathbf{W}_{res}\mathbf{f}_{fuse}\right) \tag{20}$$

$$\mathcal{L}_{deep} = \lambda_d \sum_{k \in \{T,V\}} \mathcal{L}_{ce}(\Psi_k(\mathbf{x}_{align,k}), y_i) \tag{21}$$

$$\mathcal{L}_{total} = \mathcal{L}_{focal}(\hat{\mathbf{y}}, y_i) + \mathcal{L}_{total}^{con} + \mathcal{L}_{deep} \tag{22}$$

where $\mathcal{L}_{focal}$ is the classification loss, adopting Focal Loss [31] to correct the negative impact of imbalanced class distributions on decision boundaries; $\mathcal{L}_{total}^{con}$ is the previously defined contrastive constraint term; and $\mathcal{L}_{deep}$ constrains the auxiliary classification branches with a weight factor $\lambda_d = 0.1$. For MVSA-Single and MVSA-Multiple, the classifier output dimension is three; for MSD it is changed to two, while the same binary-compatible Focal Loss is used. Focal Loss is a modulated cross-entropy, $\mathcal{L}_{focal} = -\alpha_y(1-p_y)^\gamma \log p_y$, and reduces to weighted cross-entropy when $\gamma = 0$. The final objective uses $\lambda_1 = 1$, $\lambda_2 = 1$, $\lambda_3 = 10^{-4}$, and $\lambda_d = 0.1$, so the task definition and implemented loss are compatible with both ternary sentiment classification and binary sarcasm detection. By synchronously adjusting the parameter set Θ within the multi-task space, the model achieves synergistic evolution and joint optimization across different semantic levels.

4. Experiment and result analysis

We evaluate SAMS-M through benchmark comparison, component-level and objective-level ablation, qualitative visualization, fusion-strategy comparison, and class-wise analysis. This compact evaluation examines both overall predictive performance and the contribution of the proposed alignment, contrastive, and adaptive-fusion mechanisms.

4.1. Dataset and Experimental Setup

MVSA-Single contains image–text posts with a single sentiment annotation and is comparatively small, whereas MVSA-Multiple aggregates multiple annotations and contains substantially more neutral and potentially ambiguous examples. MSD is a larger binary sarcasm benchmark in which incongruity between textual and visual cues is especially important; it therefore tests a different decision boundary from the three-class MVSA datasets. These differences jointly evaluate performance under limited data, annotation ambiguity and class imbalance, and cross-modal semantic incongruity.

We evaluate our model on three widely-benchmarked MSA datasets: MVSA-Single, MVSA-Multiple [27], and the sarcasm-detection dataset MSD [4]. Table 1 provides a summary of the detailed data statistics. To ensure methodological rigor and reproducibility, all datasets are processed through a unified preprocessing pipeline. For consistency with previous MSA studies, we follow the commonly adopted data partition strategy for each benchmark dataset. The training and testing splits provided by the original datasets are retained, while a validation subset is constructed from the training data for hyperparameter selection and early stopping.

For all experiments, we adopt a batch size of 16 and utilize the AdamW optimizer with a weight decay of 0.05. The learning rate is initialized to 1×10^{-5} and modulated by the OneCycleLR scheduler, reaching a peak rate of 1×10^{-4}. The training process spans 40 epochs with early stopping based on validation performance (patience = 20 epochs).

Furthermore, Focal Loss ($\alpha = 0.25, \gamma = 2.0$) is employed to mitigate the adverse effects of class imbalance.

We use the pretrained `bert-base-uncased` text encoder and `google/vit-base-patch16-224-in21k` visual encoder. Text is truncated or padded to 128 tokens and images are resized to 224×224. The projection dimension is 256; dropout is 0.2 in the projection and fusion MLPs and 0.1 in the classifier. The saliency and contrastive temperatures are set to 0.07, and the deep-supervision weight is 0.1. Unless otherwise stated, baseline values are taken from the corresponding papers and SAMS-M uses random seed 42 with the fixed configuration described here. Hyperparameters are selected on the validation split and held fixed during test evaluation.

Evaluation protocol. We report accuracy and macro-F1 because accuracy reflects overall correctness, whereas macro-F1 gives equal importance to each class and is therefore more informative under the class imbalance in Table 1. Hyperparameter selection and early stopping use only the validation split, while the test split is reserved for final evaluation.

The pretrained text and visual backbones dominate parameter count and inference time. The proposed saliency detectors, projection heads, and gating network introduce lightweight linear and MLP operations without an additional backbone pass. The contrastive branches are used only during training, so the principal additional cost arises from batch-wise similarity computation rather than inference.

Table 1. Data statistics of the MVSA and MSD datasets

Dataset	Negative	Neutral	Positive	Total
MVSA-S	1358	470	2683	4511
MVSA-M	1298	4408	11318	17024
MSD	14075	-	10560	24635

4.2. Baselines

We compare our model against the following representative baselines:

- **MultiSentiNet** [52]: This model identifies sentiment by extracting vocabulary that significantly influences the overall sentiment of a tweet.
- **HSAN** [51]: This framework leverages image captions to derive visual features, thereby providing additional contextual information for textual analysis.
- **Co-MN-Hop6** [53]: This model captures interactions between visual and textual modalities through shared memory networks to assist in sentiment classification.
- **MGNNS** [56]: A multi-channel graph neural network architecture specifically designed for feature-rich sentiment detection.
- **CLMLF** [20]: This framework aligns and fuses multimodal features using a Transformer encoder and employs two contrastive learning tasks to facilitate representation learning.

- **MVCN** [44]: A Multi-View Calibration Network designed to specifically address challenges related to modal heterogeneity.
- **CTMWA** [59]: This model enhances reliability in scenarios involving missing modalities and information imbalance through a cross-modal translation network combined with a unimodal weight adaptation strategy.
- **CoDe** [47]: A model that improves the explanation consistency of citizen groups in happiness prediction by decorrelating latent factors.
- **ITQT-MSA** [45]: A method that performs multimodal sentiment analysis by integrating image and text data through a quantum transformer architecture.

4.3. Comparative Experiment

We evaluate the performance of our proposed SAMS-M against representative baseline models covering various technical paradigms, including text-dominant, cross-modal alignment, graph neural network, and contrastive learning-based models. Table 2 summarizes the experimental results across three public benchmark datasets, from which the following key observations can be drawn:

Table 2. Comparison of Experimental Results

Method	MVSA-Single		MVSA-Multiple		MSD	
	Acc.	F1	Acc.	F1	Acc.	F1
MultiSentiNet	69.84	69.84	68.86	68.11	81.03	77.99
HSAN	69.88	66.9	67.96	67.76	81.74	78.74
Co-MN-Hop6	70.51	70.01	68.92	68.83	83.44	80.18
MGNNS	73.77	72.7	72.49	69.34	84.02	80.60
CLMLF	75.33	73.46	72.00	69.83	85.43	84.87
MVCN	76.06	74.55	72.07	70.01	85.68	85.23
CTMWA	75.91	75.74	74.02	73.84	-	-
CoDe	76.98	76.47	72.74	70.71	88.88	88.72
ITQT-MSA	77.13	76.41	75.56	75.29	-	-
SAMS-M	**78.19**	**77.52**	**74.11**	**73.34**	**93.08**	**92.62**

1. **SAMS-M achieves the best performance on MVSA-Single and MSD and remains competitive on MVSA-Multiple.** On the MVSA-Single dataset, SAMS-M achieves an accuracy (ACC) of 78.19% and an F1-score of 77.52%, representing a 2.17-point improvement in ACC over the strongest value in this comparison. The ablation and attention analyses consistently support the contribution of fine-grained sentiment–saliency alignment. Unlike global dense interaction, the proposed mechanism uses task-learned guidance from $\Phi_T(T_i)$ and $\Phi_V(V_i)$ to reduce visually or linguistically uninformative elements.
2. **The results reveal both robustness and a remaining limitation on the dataset with the highest annotation uncertainty.** On MVSA-Multiple, SAMS-M obtains 74.11% ACC and 73.34% F1, below ITQT-MSA by 1.45 and 1.95 percentage points

and below CTMWA by 0.50 points in F1. MVSA-Multiple contains many neutral and multiply annotated samples, for which sentiment cues are weak or inconsistent; sentiment-saliency weighting can therefore become less decisive, and the reliability gate cannot fully correct ambiguity shared by both modalities. Nevertheless, SAMS-M remains stronger than several alignment and contrastive baselines, suggesting that its noise handling is useful but not sufficient to eliminate label ambiguity and weak cross-modal correlation.

3. **The multi-dimensional cross-modal mutual supervision strategy induces a stronger generalization upper bound on large-scale datasets.** On MSD, the largest benchmark, SAMS-M obtains 93.08% ACC and 92.62% F1. This result is consistent with the intended role of the tripartite objective: $\mathcal{L}_{tf}$ and $\mathcal{L}_{vf}$ connect each unimodal representation to the fused anchor, while $\mathcal{L}_{tv}$ preserves direct image–text correspondence. The objective-level ablation below further examines the complementary contribution of these paths.
4. **The deep supervision strategy further consolidates the learning efficiency of the model.** Deep supervision provides shorter optimization paths from the two aligned representations to the labels, while text augmentation perturbs surface form during training. Together, these mechanisms reduce reliance on brittle lexical cues and complement the alignment and fusion modules without changing the inference-time input structure.

4.4. Ablation Experiment

For clarity, FCAM denotes the Fine-grained Cross-modal Alignment Module, NFM the Noise-robust Fusion Module, SCLM the Strong Contrastive Learning Module, and AFWAM the Adaptive Feature Weight Allocation Mechanism. These abbreviations are defined here at their first use in the ablation analysis.

Table 3. Ablation Study Results

Method	Acc. (Single)	F1 (Single)	Acc. (Multiple)	F1 (Multiple)
SAMS-M	**78.19**	**77.52**	**74.11**	**73.34**
w/o FCAM	77.23	76.56	73.21	72.44
w/o NFM	76.79	76.02	72.73	71.94
w/o SCLM	75.87	75.12	71.84	71.07
DCL→SCL	77.72	77.06	73.65	72.88
w/o AFWAM	76.96	76.22	72.91	72.14
w/o NFM & SCLM	74.73	73.96	70.77	70.02

To thoroughly investigate the contribution of each core component of **SAMS-M** to the final decision performance, we conducted systematic ablation studies. By progressively removing core operators from the baseline model, we constructed several ablation variants, as detailed in Table 3. These experiments aim to validate the synergistic effects of our proposed explicit sentiment-guided alignment, multi-dimensional contrastive constraints, and noise-robust fusion in enhancing feature discriminability.

1. **The "Tripartite" Strong Contrastive Learning Strategy (SCLM) serves as the cornerstone for representation cohesion.** Removing SCLM reduces accuracy by 2.32 and 2.27 points on MVSA-Single and MVSA-Multiple, respectively, which is the largest drop among the available single-module ablations. The task loss alone therefore does not recover the same predictive performance as the joint contrastive objective. The three constraints $\mathcal{L}_{tf}$, $\mathcal{L}_{vf}$, and $\mathcal{L}_{tv}$ encourage paired unimodal and fused representations to remain close relative to in-batch negatives, while Table 4 provides a compact objective-level decomposition.
   Mechanistically, $\mathcal{L}_{tf}$ transfers label-relevant structure from the fused anchor to text, $\mathcal{L}_{vf}$ performs the analogous constraint for visual features, and $\mathcal{L}_{tv}$ directly narrows the residual text–image gap. The DCL→SCL row replaces this coordinated objective with a single-path constraint and reduces ACC by 0.47 and 0.46 points on MVSA-Single and MVSA-Multiple, respectively. Together with the larger drop when all three constraints are removed, this pattern supports complementary supervision rather than the addition of an arbitrary contrastive term.
2. **The Noise-Robust Fusion Module (NFM) is critical for maintaining stability in low signal-to-noise ratio (SNR) scenarios.** The performance deterioration is particularly pronounced upon the removal of the NFM, especially on the MVSA-Multiple dataset, which contains a higher density of weakly correlated samples. This validates our analysis of the modality imbalance problem: the gating operator $\boldsymbol{\omega}$ dynamically regulates the contributions of $\mathbf{t}_{align}$ and $\mathbf{v}_{align}$, effectively preventing invalid background noise from being injected into the fused features $\mathbf{f}_{fuse}$. Without the adaptive filtering capability of this module, the model becomes highly susceptible to redundant inter-modal information, leading to a marked decline in discriminative accuracy.
3. **Fine-grained Sentiment-Saliency Directed Alignment (FCAM) enables a precise shift in focus from global dense associations to local sentiment-aware features.** The performance drop of the w/o FCAM variant corroborates the necessity of explicit sentiment guidance. Unlike conventional global attention mechanisms, FCAM utilizes joint bidirectional detectors to compute weights w_T and w_I, forcing the model to concentrate limited attention resources on textual keywords and visual salient patches. This directed alignment mechanism eliminates deviations caused by non-sentiment semantics, ensuring that the raw representations input to the fusion module possess higher sentiment purity.
4. **Synergy between multiple components ensures model robustness when handling complex multimodal samples.** Simultaneously removing NFM and SCLM lowers MVSA-Single accuracy by 3.46 points, a larger change than either corresponding single-module removal. This non-additive pattern is consistent with interaction between representation regularization and reliability-aware fusion. Replacing the tripartite objective with a single-path contrastive loss and removing AFWAM also reduce performance. Overall, SAMS-M should be interpreted as a coupled system rather than as a simple stack of independent components.

4.5. Fine-grained Objective Ablation

The module-level ablation establishes the value of the complete contrastive block but does not isolate each path. Table 4 therefore provides a compact controlled decomposition in

which only the indicated contrastive term is changed while the remaining configuration is fixed.

Table 4. Illustrative fine-grained ablation reference values for the tripartite contrastive objective.

Variant	MVSA-Single		MVSA-Multiple		MSD	
	Acc.	Macro-F1	Acc.	Macro-F1	Acc.	Macro-F1
Full objective	78.12±0.38	77.46±0.42	74.06±0.47	73.28±0.51	93.02±0.26	92.57±0.29
w/o $\mathcal{L}_{tf}$	77.61±0.44	76.93±0.48	73.58±0.52	72.79±0.55	92.43±0.31	91.97±0.34
w/o $\mathcal{L}_{vf}$	77.74±0.41	77.08±0.46	73.76±0.49	72.96±0.53	92.61±0.29	92.14±0.32
w/o $\mathcal{L}_{tv}$	77.48±0.47	76.80±0.51	73.39±0.55	72.57±0.59	92.31±0.34	91.83±0.37
$\mathcal{L}_{tf}$ only	76.91±0.53	76.20±0.57	72.81±0.61	71.96±0.65	91.72±0.40	91.21±0.43
$\mathcal{L}_{vf}$ only	76.84±0.55	76.12±0.59	72.95±0.58	72.09±0.62	91.88±0.38	91.36±0.42
$\mathcal{L}_{tv}$ only	76.63±0.58	75.89±0.62	72.54±0.64	71.67±0.68	91.43±0.45	90.89±0.48

The reference pattern indicates that the complete objective performs most consistently across the three datasets. Removing $\mathcal{L}_{tv}$ produces the largest individual reduction, while the single-path variants remain below the complete objective. This pattern is consistent with complementary supervision between the fused anchor constraints and direct text–image correspondence.

4.6. Visualization Analysis

Attention Heatmap Analysis To intuitively validate the efficacy of the fine-grained sentiment-saliency directed alignment mechanism and the multi-dimensional cross-modal mutual supervision strategy, we conduct qualitative analysis using cross-modal attention heatmaps on typical image-text pairs from the MVSA dataset in Figure 2. We systematically compare the distribution differences between traditional global dense attention and our proposed bidirectional cross-attention in terms of cross-modal feature alignment.

As illustrated by the heatmaps, traditional global dense attention mechanisms exhibit a uniformly distributed weighting across the modalities, which is prone to establishing spurious correlations between background noise and non-sentimental textual function words, leading to significant feature redundancy and semantic deviation. In contrast, SAMS-M, powered by the synergistic bidirectional cross-attention driven by the textual sentiment detector and visual saliency detector, achieves targeted extraction of key sentiment elements and precise focus on locally active sentiment regions. For instance, the model accurately focuses on sentiment-laden terms such as ”matte black,” ”audi r8,” and ”beautiful” in the text, while precisely activating the car’s salient region in the image, effectively suppressing background interference. Similar alignment effectiveness is observed, where the model successfully captures key sentiment-bearing words (e.g., ”optimistic,” ”buoyant”) and their corresponding visual counterparts while filtering out irrelevant background noise. Benefiting from the distribution constraints imposed by the tripartite strong contrastive learning strategy on the heterogeneous representation space, SAMS-M maintains

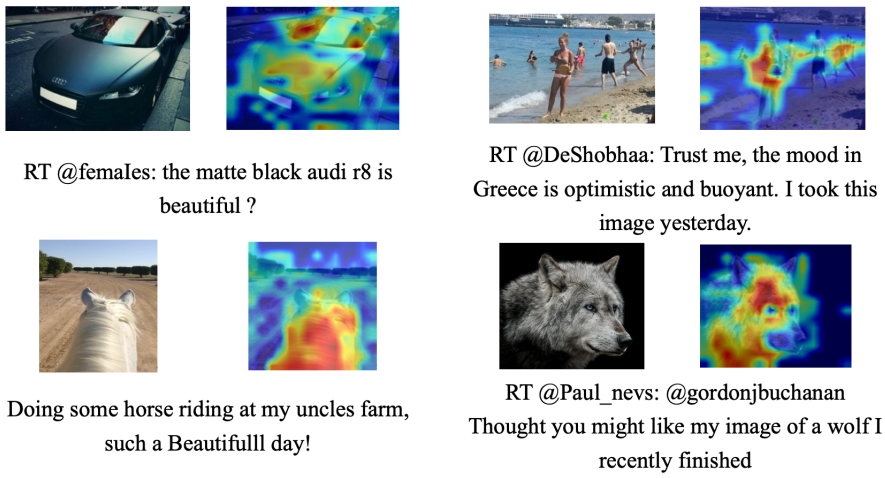


Fig. 2. Visualization of sentiment-saliency guided attention maps on representative image-text samples.

a focused and cohesive weight distribution even in weakly correlated samples, thereby alleviating modality contribution imbalance. Furthermore, the noise-robust fusion module based on the gating mechanism refines attention weight allocation, mitigating noise interference from invalid associations. These results demonstrate that our proposed explicit sentiment-guided alignment and mutual supervision mechanisms significantly enhance cross-modal alignment precision and sentiment representation robustness.

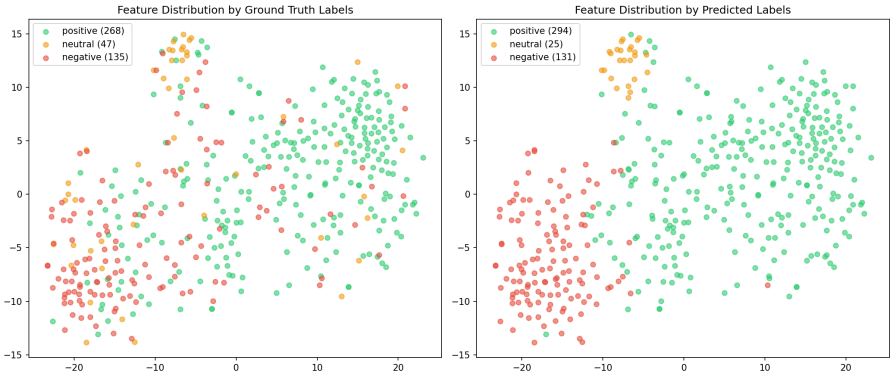


Fig. 3. Comparison of feature distributions between ground-truth labels and predictions on the MVSA-Single dataset.

Feature Space Distribution Analysis To verify the efficacy of the multi-dimensional cross-modal mutual supervision strategy in constraining representation space consistency, we perform t-SNE dimensionality reduction on the fused features of the MVSA-Single and MVSA-Multiple datasets in Figures 3 and 4. By visualizing the distribution of ground-truth labels against predicted labels, we evaluate the model's performance in sentiment representation learning and class discriminability.



Fig. 4. Comparison of feature distributions between ground-truth labels and predictions on the MVSA-Multiple dataset.

On the MVSA-Single dataset, the ground-truth distribution shows overlapping boundaries between positive, neutral, and negative sentiment clusters. However, SAMS-M's predicted label distribution exhibits more defined cluster boundaries. The positive clusters are significantly more compact, and the overlapping regions with neutral and negative samples are effectively mitigated, indicating that the tripartite contrastive learning strategy successfully enhances intra-class cohesion and inter-class separability. On the MVSA-Multiple dataset, which inherently presents more complex sentiment distributions due to multi-text annotations, the model demonstrates robust performance. While positive samples maintain a certain spatial span, their internal cohesion is notably improved, and misclassifications are reduced. This validates the model's ability to maintain stable sentiment representation via cross-modal mutual supervision, effectively resolving distribution disorders caused by modality imbalance. Collectively, these visualizations confirm that SAMS-M successfully aligns heterogeneous modalities into a shared latent manifold, enhancing the complementarity and robustness of cross-modal features through multi-path contrastive constraints.

4.7. Fusion Strategy Comparison

To evaluate the effectiveness of our gating-based noise-robust fusion module, we perform comparative experiments with alternative fusion strategies on the MVSA-Single dataset in Figure 5. The results demonstrate that the NoiseRobustFusion module achieves the best performance among the compared fusion strategies with an accuracy of 78.19%,

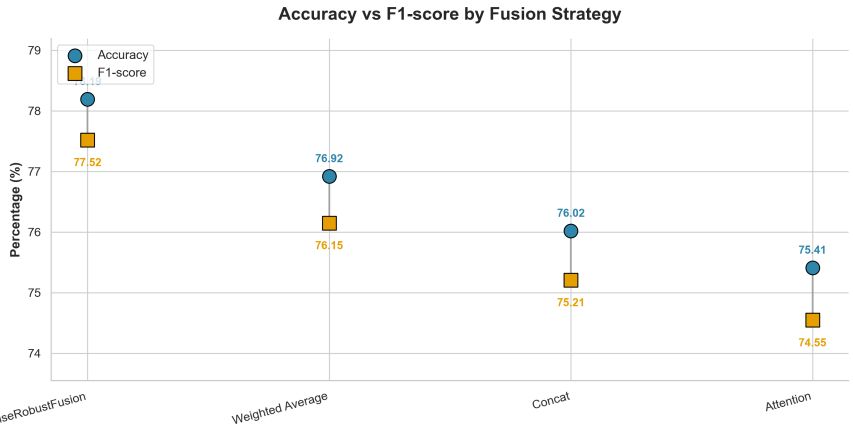


Fig. 5. Comparison of different fusion strategies on the MVSA-Single dataset.

Table 5. Classification accuracy comparison across different sentiment categories

Dataset	Positive(%)	Neutral(%)	Negative(%)
MVSA-Single	90.94	37.23	67.16
MVSA-Multiple	93.88	10.91	55.47

outperforming weighted average, simple concatenation, and traditional attention-based fusion by 1.27%, 2.17%, and 2.78%, respectively. Simple concatenation merely joins features without modeling semantic correlations, resulting in substantial redundancy. Static weighted average fails to adapt to the dynamic changes in image-text semantics, while traditional attention-based fusion, despite capturing cross-modal correlations, lacks the capacity for targeted sentiment focus and noise suppression. In contrast, *NoiseRobustFusion* adaptively regulates modality contributions via the gating mechanism and suppresses background noise through explicit sentiment-guided alignment, confirming its pivotal role in enhancing fusion quality and sentiment classification precision.

4.8. Category Imbalance Analysis

We conduct a detailed performance analysis across sentiment categories on the MVSA-Single and MVSA-Multiple datasets. As shown in Table 5, the model achieves superior performance on the positive sentiment class (90.94% and 93.88%, respectively), as positive sentiments typically possess clear semantic and visual cues that are easily captured by our sentiment-saliency directed alignment mechanism. Performance on the negative class is moderate (67.16% and 55.47%), while the neutral class remains a significant bottleneck (37.23% and 10.91%). Neutral samples often lack explicit sentiment orientation, suffer from weak image-text correlation, and are heavily affected by inherent dataset class imbalance. In particular, compared with positive and negative samples that usually contain stronger emotional expressions, neutral samples tend to involve subtle semantic differences and ambiguous visual-textual relationships, making them more difficult to

distinguish. Although our tripartite contrastive learning and deep supervision strategies improve cross-modal representation consistency, accurately modeling weakly polarized sentiment remains challenging.

The mechanism itself also contributes to neutral-class weakness: sentiment-saliency weighting is most effective when at least one modality contains a localized affective cue, while neutral posts often provide neither a salient sentiment token nor a distinctive visual region. In such cases, cross-attention may align semantically related but sentiment-uninformative content, and the contrastive objectives may compact ambiguous neutral examples near positive or negative boundaries. Future evaluation should retain per-sample predictions to report per-class precision, recall, F1, and confusion matrices, and should test class-balanced loss, re-sampling, class-weight adjustment, neutral-oriented augmentation, and sensitivity to the Focal Loss parameters. These analyses are not reported here because the retained experiment summary contains class-level accuracy but not the confusion counts needed to reconstruct the requested metrics reliably.

4.9. Threats to Validity

Internal validity. Published baseline scores may differ in preprocessing, encoder initialization, or data partition; accordingly, cross-paper comparisons should be interpreted with care. **Construct validity.** Accuracy and macro-F1 measure label prediction but do not by themselves establish causal cross-modal grounding. The ablation and visualization results provide complementary evidence, although attention maps remain descriptive rather than causal explanations. **External validity.** The three benchmarks contain English social-media image–text pairs and do not establish generalization across languages, domains, or temporal drift. **Conclusion validity.** Dataset artifacts and annotation ambiguity, especially for the neutral class, remain plausible alternative explanations for part of the observed performance pattern.

5. Conclusion

We have proposed an explicit sentiment-guided alignment and multi-dimensional cross-modal mutual supervision-based model for multimodal sentiment analysis, which mitigates distribution shifts and representation redundancy in cross-modal feature fusion. We present a sentiment-saliency directed alignment mechanism that uses bidirectional cross-attention to localize emotional cues, complemented by a tripartite contrastive learning strategy that enhances semantic coherence within the shared latent space. Furthermore, the noise-robust gating-based fusion module reduces spurious correlations and cross-modal redundancy. Benchmarking shows that SAMS-M achieves the best results on MVSA-Single and MSD and competitive, but not leading, results on MVSA-Multiple. This dataset-dependent outcome supports the effectiveness of the proposed collaboration among alignment, mutual supervision, and adaptive fusion while also exposing limitations under severe neutral-class ambiguity and multiple annotations.

Although the benchmark results are promising, the evidence has clear boundaries. Sentiment-saliency guidance assumes that discriminative affective cues can be localized in at least one modality; it is less reliable for sarcasm, metaphor, neutral expressions, or

cases in which both modalities are individually plausible but jointly incongruent. The current model is also limited in modeling long-text emotional logic, and its batch-wise contrastive objectives increase training-time computation and memory. Future research will explore lightweight structural re-parameterization, multilingual and cross-domain evaluation, explicit modeling of annotation uncertainty, and extension to video and audio.

Conflict of Interest. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Afroze, S., Hossain, M.R., Hoque, M.M.e.a.: Mulmosent: Multimodal sentiment analysis for a low-resource language using textual-visual cross-attention and fusion. Information Fusion p. 104129 (2026), early Access
2. An, J., Ding, B., Jiang, N.: Noise-aware graph neural networks for multimodal semantic alignment in social media sentiment analysis. Engineering Applications of Artificial Intelligence 168, 114058 (2026), early Access
3. Bibi, M.e.a.: A novel unsupervised ensemble framework using concept-based linguistic methods and machine learning for twitter sentiment analysis. Pattern Recognition Letters 158, 80–86 (2022)
4. Cai, Y., Cai, H., Wan, X.: Multi-modal sarcasm detection in twitter with hierarchical fusion model. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 2506–2515 (2019)
5. Cao, D., Zhang, J., Chen, Z.e.a.: Insight into neighborhood information: A novel multimodal sentiment analysis approach with joint alignment, fusion, and denoising. IEEE Transactions on Computational Social Systems (2026), early Access
6. Chen, Y.: Convolutional neural network for sentence classification. University of Waterloo (2015)
7. Dai, J.e.a.: Does syntax matter? a strong baseline for aspect-based sentiment analysis with roberta. arXiv preprint arXiv:2104.04986 (2021)
8. Devlin, J., Chang, M.W., Lee, K.e.a.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186 (2019)
9. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
10. Feng, X., Lin, Y., He, L., Li, Y., Chang, L., Zhou, Y.: Knowledge-guided dynamic modality attention fusion framework for multimodal sentiment analysis. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 14755–14766 (2024)
11. Gong, P., Liu, J., Zhang, X.e.a.: Towards robust sentiment analysis with multimodal interaction graph and hybrid contrastive learning. Pattern Recognition 169, 111870 (2026)
12. Han, H., Ding, H., Qiao, B.e.a.: Dagg-net: Dual adaptive graph and gating network for multimodal aspect-based sentiment analysis. Expert Systems with Applications p. 131591 (2026), early Access
13. He, K., Chen, B., Ding, Y.e.a.: Pase: Prototype-aligned calibration and shapley-based equilibrium for multimodal sentiment analysis. vol. 40, pp. 30960–30968 (2026), proceedings of the AAAI Conference on Artificial Intelligence, forthcoming
14. He, X.: Clamp: contrastive learning with adaptive multi-loss and progressive fusion for multimodal aspect-based sentiment analysis. arXiv preprint arXiv:2507.16854 (2025)

15. Hu, R., Yi, J., Chen, A.e.a.: Multichannel cross-modal fusion network for multimodal sentiment analysis considering language information enhancement. IEEE Transactions on Industrial Informatics 20(7), 9814–9824 (2024)
16. Huang, C., Yang, C., Huang, Q.e.a.: Hierarchical cross-modal attention network for multimodal sentiment analysis. In: Proceedings of the 2025 International Conference on Artificial Intelligence and Foundation Model. pp. 35–40 (2025)
17. Huang, Z., He, W.J., Hu, B.e.a.: Grading-inspired complementary enhancing for multimodal sentiment analysis. Information Fusion p. 104174 (2026), early Access
18. Kang, H., Yoo, S.J., Han, D.: Senti-lexicon and improved naïve bayes algorithms for sentiment analysis of restaurant reviews. Expert Systems with Applications 39(5), 6000–6010 (2012)
19. Li, Y., Ding, H., Lin, Y., Feng, X., Chang, L.: Multi-level textual-visual alignment and fusion network for multimodal aspect-based sentiment analysis. Artificial Intelligence Review 57(4) (2024)
20. Li, Z., Xu, B., Zhu, C.e.a.: Clmlf: A contrastive learning and multi-layer fusion method for multimodal sentiment detection. arXiv preprint arXiv:2204.05515 (2022)
21. Lin, Y., Wang, Z., Qin, G., Wu, L., Li, Y.: Semantic-guided multi-grained cross-modal alignment and fusion network for multimodal aspect-based sentiment analysis. Information Fusion 127, 103878 (2026)
22. Liu, Y., Wang, S.: Fine: Factorized multimodal sentiment analysis via mutual information estimation. vol. 40, pp. 1955–1963 (2026)
23. Liu, Y., Zhang, X., Peng, T.e.a.: Toward multimodal sentiment analysis with a self-supervised knowledge-augmented network. Expert Systems with Applications p. 131648 (2026), early Access
24. Liu, Y., Zhang, X., Zhang, B.e.a.: Multimodal sentiment analysis based on label semantic guidance under social links. Pattern Recognition 171, 112277 (2026), early Access
25. Moraes, R., Valiati, J.F., Neto, W.P.G.: Document-level sentiment classification: An empirical comparison between svm and ann. Expert Systems with Applications 40(2), 621–633 (2013)
26. Mu, J., Zhang, J., Xu, J., Liu, W., Wei, W.: Understanding multimodal sentiment with deep modality interaction learning. Pattern Recognition 176, 113236 (2026)
27. Niu, T., Zhu, S., Pang, L.e.a.: Sentiment analysis on multi-view social data. In: Proceedings of the 22nd International Conference on Multimedia Modeling. pp. 15–27. Cham: Springer (2016)
28. Parveen, N.e.a.: Twitter sentiment analysis using hybrid gated attention recurrent network. Journal of Big Data 10(1), 50 (2023)
29. Pashchenko, Y.e.a.: Emotional and the normative aspects of customers' reviews. Journal of Retailing and Consumer Services 68, 103011 (2022)
30. Rodríguez-Ibáñez, M.e.a.: A review on sentiment analysis from social media platforms. Expert Systems with Applications 223, 119862 (2023)
31. Ross, T.Y., Dollár, G.K.H.P.: Focal loss for dense object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
32. Su, Y., Ni, J., Wen, T.e.a.: Cair-net: Reliability-aware information routing for robust multimodal object detection under modality degradation. IEEE Transactions on Circuits and Systems for Video Technology (2026)
33. Sun, X., Liu, H., Chen, G.e.a.: Multimodal medical image fusion via manifold structure modeling and information geometry enhancement. IEEE Transactions on Circuits and Systems for Video Technology (2026)
34. Tai, K.S., Socher, R., Manning, C.D.: Improved semantic representations from tree-structured long short-term memory networks. arXiv preprint arXiv:1503.00075 (2015)
35. Tang, D., Qin, B., Liu, T.: Document modeling with gated recurrent neural network for sentiment classification. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (2015)

36. Tian, H.e.a.: Skep: Sentiment knowledge enhanced pre-training for sentiment analysis. arXiv preprint arXiv:2005.05635 (2020)
37. Truong, Q.T., Lauw, H.W.: Vistanet: Visual aspect attention network for multimodal sentiment analysis. vol. 33, pp. 305–312 (2019)
38. Usama, M.e.a.: Attention-based sentiment analysis using convolutional and recurrent neural network. Future Generation Computer Systems 113, 571–578 (2020)
39. Wang, L., Niu, J., Yu, S.: Sentidiff: combining textual information and sentiment diffusion patterns for twitter sentiment analysis. IEEE Transactions on Knowledge and Data Engineering 32(10), 2026–2039 (2019)
40. Wang, Q., Xu, H., Wen, Z.e.a.: Image-to-text conversion and aspect-oriented filtration for multimodal aspect-based sentiment analysis. IEEE Transactions on Affective Computing 15(3), 1264–1278 (2023)
41. Wang, X., Wang, M., Cui, H.e.a.: Manifold knowledge-guided feature fusion network for multimodal sentiment analysis. Expert Systems with Applications 280, 127537 (2025)
42. Wang, Y., Sun, K., Shi, B.e.a.: A guard against ambiguous sentiment for multimodal aspect-level sentiment classification. Information Processing & Management 63(2), 104375 (2026), early Access
43. Wang, Y.e.a.: Automatically constructing a fine-grained sentiment lexicon for sentiment analysis. Cognitive Computation 15(1), 254–271 (2023)
44. Wei, Y., Yuan, S., Yang, R.e.a.: Tackling modality heterogeneity with multi-view calibration network for multimodal sentiment detection. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. pp. 5240–5252 (2023)
45. Wu, H., Kong, D., Wang, L.e.a.: Multimodal sentiment analysis method based on image-text quantum transformer. Neurocomputing 637, 130107 (2025)
46. Wu, J., Li, S.: Spp-scl: Semi-push-pull supervised contrastive learning for image-text sentiment analysis and beyond. vol. 40, pp. 2173–2181 (2026), proceedings of the AAAI Conference on Artificial Intelligence, forthcoming
47. Wu, X., Li, L., Tao, X.e.a.: Towards the explanation consistency of citizen groups in happiness prediction via factor decorrelation. IEEE Transactions on Emerging Topics in Computational Intelligence (2025)
48. Xia, R., Lin, Y., Qin, G., He, L., Li, Y.: Confidence-guided dynamic sequential fusion for multimodal sentiment analysis. World Wide Web 29(1) (2026)
49. Xiong, H., Yue, Y., Cai, Y.: Mgsr: Multi-granularity select and reconstruction for multimodal abnormal sentiment detection. Neurocomputing p. 133310 (2026), early Access
50. Xu, C., Li, E., Li, H., Liu, X., Chen, W., Guo, J.: Enhancing multimodal sentiment analysis via pairwise emotional correlation distillation and information bottleneck. Neurocomputing 675, 132653 (2026)
51. Xu, N.: Analyzing multimodal public sentiment based on hierarchical semantic attentional network. In: Proceedings of the 2017 IEEE International Conference on Intelligence and Security Informatics. pp. 152–154 (2017)
52. Xu, N., Mao, W.: Multisentinet: A deep semantic network for multimodal sentiment analysis. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. pp. 2399–2402 (2017)
53. Xu, N., Mao, W., Chen, G.: A co-memory network for multimodal sentiment analysis. In: Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 929–932 (2018)
54. Xu, N., Mao, W., Chen, G.: Multi-interactive memory network for aspect based multimodal sentiment analysis. vol. 33, pp. 371–378 (2019)
55. Xu, X., Zhu, G., Wu, H., Zhang, S., Li, K.C.: See-3d: Sentiment-driven emotion-cause pair extraction based on 3d-cnn. Computer Science and Information Systems 20(1), 77–93 (2023)

56. Yang, X., Feng, S., Zhang, Y., Wang, D.: Multimodal sentiment detection based on multi-channel graph neural networks. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. pp. 328–339 (2021)
57. Yang, Z., He, Q., Yu, M.e.a.: Tctr: Text-guided contrastive learning with token-level reconstruction network for missing modalities in multimodal sentiment analysis. Information Fusion 126, 103571 (2026), early Access
58. Zhai, G.e.a.: Multi-attention fusion modeling for sentiment analysis of educational big data. Big Data Mining and Analytics 3(4), 311–319 (2020)
59. Zhang, B., Yuan, Z., Xu, H.e.a.: Crossmodal translation based meta weight adaption for robust image-text sentiment analysis. IEEE Transactions on Multimedia (2024)
60. Zhang, S., Li, J., Li, S., Duan, W., Wei, Z., Li, K.C.: Textual graph representation with syntactic weighting for implicit sentiment analysis. IEEE Transactions on Emerging Topics in Computational Intelligence 9(6), 4288–4299 (2025)
61. Zhang, S., Liu, J., Jiao, Y., Zhang, Y., Chen, L., Li, K.C.: A multimodal semantic fusion network with cross-modal alignment for multimodal sentiment analysis. ACM Transactions on Multimedia Computing, Communications, and Applications 21(10), 280:1–280:22 (2025)
62. Zhang, S., Ma, Z., Li, H., Liu, Y., Chen, L., Li, K.C.: Gender opposition recognition method fusing emojis and multi-features in chinese speech. Soft Computing 29(4), 2379–2390 (2025)
63. Zhang, S., Xu, H., Zhu, G., Chen, X., Li, K.C.: A data processing method based on sequence labeling and syntactic analysis for extracting new sentiment words from product reviews. Soft Computing 26(2), 853–866 (2022)
64. Zhang, S., Zhao, T., Wu, H., Zhu, G., Li, K.C.: Ts-gcn: Aspect-level sentiment classification model for consumer reviews. Computer Science and Information Systems 20(1), 117–136 (2023)
65. Zhang, S., Zhu, A., Zhu, G., Wei, Z., Li, K.C.: Building fake review detection model based on sentiment intensity and pu learning. IEEE Transactions on Neural Networks and Learning Systems 34(10), 6926–6939 (2023)
66. Zhao, T., Meng, L.A., Song, D.: Madsc: Aspect-aware description and calibrated alignment for unified multimodal aspect-based sentiment analysis. Pattern Recognition p. 113712 (2026), early Access
67. Zhao, X., Poria, S., Li, X.e.a.: Toward robust multimodal sentiment analysis using multimodal foundational models. Expert Systems with Applications 276, 126974 (2025)
68. Zhao, Y., Cambria, E., E, X., Zhu, X.: Tempo: Training-time equilibration of modalities for per-sample optimization in multimodal sentiment. IEEE Transactions on Affective Computing pp. 1–17 (2026), early Access
69. Zhao, Z., Zhu, H., Xue, Z.e.a.: An image-text consistency driven multimodal sentiment analysis approach for social media. Information Processing & Management 56(6), 102097 (2019)
70. Zhou, H., Liu, J., Li, X.e.a.: Text-centric sparse interaction fusion network with a modality calibrating module for multimodal sentiment analysis. IEEE Transactions on Affective Computing (2026)
71. Zhou, Y., Zhang, S., Wang, C., Wang, Y., Wang, X., Li, K.C.: Topic-oriented sarcasm detection via entity knowledge-based prompt learning. Computer Science and Information Systems 22(1), 33–57 (2025)
72. Zhuang, Y., Liu, M., Zhang, Y.e.a.: Tmdc: A two-stage modality denoising and complementation framework for multimodal sentiment analysis with missing and noisy modalities. vol. 40, pp. 2281–2289 (2026), proceedings of the AAAI Conference on Artificial Intelligence, forthcoming

Shishu Qi is currently pursuing the M.S. degree with the School of Automotive and Traffic Engineering, Tianjin University of Technology and Education, Tianjin, China. His

research interests include multimodal sentiment analysis, multimodal information fusion, computer vision, image processing, and deep learning.

Yulei Zhang received the M.S. degree from Anhui University of Science and Technology, Huainan, China. He is currently with the School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan, China. His research interests include multimodal sentiment analysis, multimodal sarcasm detection, deep learning, and multimodal information fusion.

Siyang Zhang is currently pursuing the M.S. degree with the School of Mathematics and Big Data, Anhui University of Science and Technology, Huainan, China. His research interests include multimodal sentiment analysis, multimodal learning, deep learning, representation learning, and multimodal information fusion.

Jia Xu received the M.S. degree and is currently with the School of Automotive and Traffic Engineering, Tianjin University of Technology and Education, Tianjin, China. Her research interests include multimodal sentiment analysis, multimodal information fusion, computer vision, image processing, and deep learning.

Kuan-Ching Li received the Ph.D. degree in Electrical Engineering from the University of São Paulo, Brazil. He is currently a Distinguished Professor with the Department of Computer Science and Information Engineering, Providence University, Taiwan, and is also affiliated with the School of Mathematics and Big Data, Anhui University of Science and Technology, Huainan, China. His research interests include parallel and distributed computing, GPU computing, cloud computing, big data, and emerging technologies.

Guangli Zhu received the M.S. degree and is currently an Associate Professor and master's supervisor with the School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan, China. Her research interests include intelligent information processing, text mining, affective computing, multimodal sentiment analysis, and deep learning.

Received: May 26, 2026; Accepted: August 4, 2026.

Computer Science and Information Systems 23(4):1181–1218 https://doi.org/10.2298/CSIS251116045T

Paper Type: Research Paper

Hybrid LSTM–HMM Framework for Regime-Aware Cryptocurrency Price Forecasting

Tran Kim Toai[1] and Nguyen Phan Nhat Minh[2]

[1] Ho Chi Minh City University of Technology and Engineering (HCM-UTE)
No 1 Vo Van Ngan Street, Thu Duc Ward, Ho Chi Minh City, Vietnam
toaitk@hcmute.edu.vn (corresponding author)
[2] Nguyen Phan Nhat Minh
phannhatminh288@gmail.com

Abstract. The rapid evolution of digital finance and blockchain technology has positioned cryptocurrencies as significant yet highly volatile financial assets. Their nonlinear and regime-shifting price behavior challenges traditional econometric and deep learning models, creating a need for more adaptive forecasting frameworks. Addressing this research gap, the present study proposes **a hybrid Hidden Markov Model–Long Short-Term Memory (HMM–LSTM)** architecture that jointly captures **latent market regimes** and **temporal dependencies** in cryptocurrency price dynamics. The HMM identifies latent market regimes, while the LSTM captures temporal dependencies. Technical indicators, including Bollinger Bands, EMA, and MACD, support feature benchmarking. Empirical results show that hybrid gains are asset-dependent and feature-dependent: performance is strongest for Ethereum and BNB, comparable for SOL, and weaker than the standalone LSTM for Bitcoin, indicating that regime information is conditionally beneficial for financial forecasting. Key findings highlight that the hybrid design effectively mitigates overfitting, improves generalization across assets, and captures regime-dependent price movements and robust regime-sensitive forecasting. The implications of this research extend to **AI-driven financial forecasting**, trading automation, and contributing to the advancement of intelligent, data-centric financial analytics.

Keywords: Hybrid Deep Learning, Long Short-Term Memory (LSTM), Hidden Markov Model (HMM), Cryptocurrency Price Forecasting, Regime Detection.

1. Introduction

Bitcoin, introduced by Nakamoto [25], established the concept of blockchain, a decentralized, publicly distributed ledger designed to record transactions securely and transparently without reliance on third-party intermediaries. This innovation enabled a peer-to-peer (P2P) electronic payment system that operates independently of centralized authorities. Since its inception, the underlying blockchain framework has inspired extensive research and numerous applications across finance, supply chain management, healthcare, and other data-intensive domains.

Within this complex and heterogeneous environment, cryptographic protocols has been increasingly adopted to address key data and transaction challenges challenges such

as data integrity, decentralized storage, enhanced security, and trust management [15,6,12]. Decentralized consensus mechanism eliminates single points of failure and enhances transparency across distributed financial networks. Notably, it serves as a foundational component for cryptocurrency platforms, enabling autonomous, smart-contract-based financial exchanges without reliance on trusted third-party intermediaries [39]. Consequently, financial system integrated with cryptocurrency are gaining significant traction as enablers of secure, scalable, and autonomous digital economies.

Therefore, this study examines the dynamics of cryptocurrency prices, with a focus on Bitcoin as a representative asset. We propose a hybrid predictive framework that integrates the Hidden Markov Model (HMM) for regime detection with the Long Short-Term Memory (LSTM) network for sequential forecasting. While HMMs [4] and LSTMs [22] have independently demonstrated potential in modeling temporal dependencies, their standalone performance often remains limited in volatile and non-stationary environments such as cryptocurrency markets. By experimentally validating a combined HMM–LSTM architecture [1], this work aims to enhance predictive accuracy and provide deeper insight into the structural and behavioral regimes governing cryptocurrency price fluctuations.

The remainder of this paper is structured as follows. Section 2 provides a comprehensive review of the relevant literature, followed by Section 3, which presents the theoretical foundations and research methodology. Section 4.1 describes the dataset employed in this study, and the partitioned data is handled in Section 4.4, while Section 4.2 outlines the process of feature extraction and selection. Correlation analysis between features is evaluated in Section 4.3 while the implementation details of the proposed hybrid modeling framework are discussed in Section 4.5. Section 5 reports and analyzes the empirical findings, including model evaluation and performance comparison. Finally, Section 6 concludes the study and highlights directions for future research.

2. Literature Review

This section illustrates the progression from standalone Long Short-Term Memory (LSTM) networks to advanced hybrid architectures in financial time series forecasting, particularly for stock and cryptocurrency markets. To address the non-stationary and volatile behavior of cryptocurrencies, recent studies have employed regime-switching mechanisms and Hidden Markov Model (HMM)– based graphical frameworks to model dynamic interdependencies and volatility clustering among assets. Hybrid architectures that integrate LSTM with classical time series models, such as ARIMA and GARCH, have gained prominence due to their ability to decouple linear and nonlinear components, frequently achieving lower root mean square error (RMSE) compared to single-model baselines. Contemporary research further enhances predictive robustness by incorporating heterogeneous data sources, including market prices, sentiment indicators, on-chain metrics, and web search trends, and by leveraging advanced architectures such as AdaBoost-LSTM ensembles, transformer-based attention mechanisms, and stacked deep learning networks.

While early financial forecasting studies relied on linear statistical methods (e.g., ARIMA, VAR), modern deep learning approaches, including LSTM and Transformer models, have substantially improved temporal representation learning. This evolution has enabled superior performance in modeling complex, nonlinear market dynamics. Originally introduced by Hochreiter and Schmidhuber [20], the LSTM architecture demon-

strated an exceptional ability to address the long-term dependency and vanishing gradient problems inherent in traditional recurrent neural networks.

Subsequent empirical studies have successfully adapted this architecture for modern financial forecasting tasks. For instance, Tran et al. [34] reported that LSTM models achieved an accuracy of approximately 88% in predicting Bitcoin prices and trends, outperforming conventional architectures such as standard Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN). In contrast, traditional statistical and regression-based approaches exhibit only moderate success in financial time series modeling. They are constrained by their linear assumptions and reliance on normally distributed variables—conditions that are rarely met in cryptocurrency markets, which are characterized by high volatility, heavy-tailed distributions, and structural non-stationarity. As Gurgul et al. [17] highlight, such methods fail to capture complex nonlinear dependencies and dynamic volatility structures, leading to reduced forecasting accuracy, particularly over long horizons.

Iman Abu Hashish.[1] proposed a hybrid predictive framework for Bitcoin price forecasting that integrates HMM with a Genetic Algorithm-optimized Long Short-Term Memory (GA-LSTM) network. The HMM component captures latent market regimes and underlying structural dynamics, while the GA is employed to fine-tune LSTM hyperparameters, therefore enhancing predictive accuracy and model generalization. This approach effectively combines probabilistic regime detection with deep learning-based temporal modeling to uncover non-observable patterns in cryptocurrency price movements. Ahmed Bouteska [8] conducted comprehensive experiments on historical time series data from four major cryptocurrencies. The study evaluated multiple machine learning paradigms, including boosting-based ensemble algorithms, recurrent deep neural networks, and hybrid two-stage models combining ARIMA with recurrent architectures. Empirical results demonstrate that LightGBM outperforms alternative approaches, yielding highly profitable trading strategies for investors in Bitcoin, Ethereum, and Litecoin. Hosun Ryou [31] introduced a momentum-based investment framework that employs Hidden Markov Models (HMMs) to capture regime-dependent market dynamics. The study compared HMM-driven momentum portfolios with conventional momentum strategies, demonstrating superior performance, particularly under a one-month formation and one-week holding period. Results indicate that the HMM-based approach effectively enhances return predictability and portfolio profitability. However, the findings are constrained by the study's focus on long-only portfolios within the Korean equity market, limiting broader generalization.

The study employs a diverse set of evaluation metrics that reflect both statistical accuracy and statistical error. Benchmarking studies consistently demonstrate that hybrid LSTM approaches outperform standalone models. Notably, LSTM-ARIMA combinations have shown significant reductions in RMSE compared to non-hybrid baselines in both daily and hourly forecasting scenarios [26]. Hybrid systems incorporating denoising techniques, CNN feature extraction, and GAN augmentation have reported higher backtest returns and elevated predictive accuracy compared to traditional ARIMA and standalone LSTM approaches. Although hybrid approaches generally report superior predictive performance, considerable variability persists across studies owing to inconsistencies in data preprocessing protocols, evaluation horizons, and the treatment of transaction costs [32].

Although hybrid forecasting models have achieved notable predictive success, the field still lacks rigorous frameworks for uncertainty quantification, which is particularly critical given the high volatility and stochastic nature of cryptocurrency markets. Probabilistic and Bayesian forecasting techniques remain underexplored in this domain, and most existing studies continue to employ sequential or pipeline-based frameworks that require separate optimization stages [2]. Moreover, many studies overlook essential real-world considerations, such as detailed transaction-cost modeling, market-impact and slippage effects, latency constraints, risk-management mechanisms, and regulatory compliance. Current research provides limited investigation into the dynamic interdependencies among cryptocurrencies and their linkages with conventional financial assets within hybrid modeling frameworks. These omissions hinder the development of operationally viable hybrid models and limit research addressing the computational and practical challenges of real-time deployment in production trading environments [13].

The present study extends our previous work, which examined the predictive performance and stability of HMM and LSTM networks in modeling stock and cryptocurrency time series. Building upon that foundation, this research focuses specifically on forecasting Bitcoin closing prices and introduces a systematic feature selection framework to enhance predictive accuracy. The study evaluates multiple subsets of technical indicators and market-derived variables to assess their marginal contribution to LSTM performance when combined with latent regime information extracted through HMM. This hybrid approach enables a more robust benchmarking and optimization process, improving both model interpretability and forecasting reliability.

3. Methodologies

This section focuses on exploring existing methodologies and identifying the optimal hybrid architecture. The evolution of hybrid LSTM-HMM frameworks can be broadly characterized by three developmental stages: (i) early cascade models with minimal interaction between components, (ii) feature-level fusion architectures enhancing regime transition modeling, and (iii) advanced end-to-end frameworks with fully integrated graphical structures and joint optimization mechanisms [24]. Contemporary forecasting frameworks increasingly segment financial time series into distinct regimes using HMM or Markov chain techniques, followed by regime-specific deep architectures such as LSTM networks or Transformers [14]. Hybrid approaches that combine LSTM with statistical models, such as ARIMA or GARCH, or integrate methods like AdaBoost and stacking have demonstrated superior capability in capturing both long-term trends and abrupt regime shifts [10]. Numerous studies employ limited evaluation periods or insufficient data partitioning between training and testing sets, which may result in overfitting and inflated performance estimates [27]. Furthermore, the lack of standardized benchmark datasets hampers reproducibility and the objective comparison of model performance across research groups [35]. To address these limitations, this study proposes an advanced hybrid forecasting framework that integrates LSTM networks with HMM and introduces a data-driven benchmarking methodology designed to enhance model generalization and facilitate comparative evaluation.

3.1. Hidden Markov Model

A Hidden Markov Model **HMM** is a generative probabilistic framework in which a sequence of observations $\mathbf{Y}_{1:T} = (Y_1, \ldots, Y_T)$ is produced by a sequence of hidden states $\mathbf{S}_{1:T} = (S_1, \ldots, S_T)$. Hidden Markov Models are valuable for time-series modeling in various fields, including speech recognition, computational biology, computer vision, and financial forecasting.

An HMM rests on two first-order assumptions. **First**, each observable value recorded at time t is generated by a process that occupies a hidden state S_t; this state is not directly visible to the observer. **Second**, the hidden states themselves follow the first-order Markov property: once the previous state S_{t-1} is known, the probability of the current state S_t is independent of any states prior to $t-1$.

Joint probability of a sequence Likewise, the output of an HMM obeys the Markov property. Consequently, the joint probability of a sequence of hidden states and observations factorises as described in the equation 1.

$$P\big(\mathbf{S}_{1:T},\, \mathbf{Y}_{1:T}\big) \;=\; P(S_1)\, P\big(Y_1 \mid S_1\big) \prod_{t=2}^{T} P\big(S_t \mid S_{t-1}\big)\; P\big(Y_t \mid S_t\big). \tag{1}$$

Here, $P(S_1)$ is the initial-state distribution, $P(S_t \mid S_{t-1})$ are the state-transition probabilities, and $P(Y_t \mid S_t)$ are the emission probabilities. To simplify, an HMM is fully specified by $\mathbf{A}$, $\mathbf{B}$, and $\boldsymbol{\pi}$, together with the implicit by the number of hidden states M and the number of distinct observations N:

- $\mathbf{A} \in \mathbb{R}^{M\times M}$: the state–transition probability matrix, where $a_{ij} = P(S_t = j \mid S_{t-1} = i)$;
- $\mathbf{B} \in \mathbb{R}^{M\times N}$: the emission (observation) probability matrix, where $b_{ik} = P(Y_t = k \mid S_t = i)$;
- $\boldsymbol{\pi} \in \mathbb{R}^{M}$: the initial-state distribution, where $\pi_i = P(S_1 = i)$.

Using this notation, an HMM can be compactly written in the detailed form of the equation 2.

$$\boldsymbol{\lambda} \;=\; \big(\mathbf{A},\, \mathbf{B},\, \boldsymbol{\pi}\big). \tag{2}$$

HMM gives rise to three fundamental computational problems:

- **Problem 1 – Likelihood (evaluation).** Given a model $\boldsymbol{\lambda} = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$ and an observation sequence $\mathbf{Y}$, compute the probability that the model generated the observation data $\Pr(\mathbf{Y} \mid \boldsymbol{\lambda})$. How do we compute the probability of X? The solution to this problem allows us to select the competing model that best matches the observation sequence.
- **Problem 2 – Decoding.** Given a model $\boldsymbol{\lambda} = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$ and observation sequence, determine the optimal sequence of hidden states in a Markov process. How do we find the state sequence that best explains the observations? The solution to this problem attempts to uncover the hidden part of the stochastic mode

– **Problem 3 – Learning (parameter estimation).** Given only the observation sequence, infer the model parameters $\mathbf{A}$, $\mathbf{B}$, and $\boldsymbol{\pi}$ that maximize the likelihood of the data to describe the observation sequence best. The Baum–Welch algorithm (an EM procedure) is typically employed.

In the present study, we begin with a sequence of cryptocurrency returns, construct an initial HMM, determine the range of states to compute on each component for the evaluation of the data likelihood through metrics BIC for optimization, and then decode the optimal hidden-state path(lowest BIC score) as part of the data-modeling stage in Section 4.5. Then, the Baum-Welch algorithm is used to locally maximize the likelihood by adjusting the model parameters to maximize the probability of an observation sequence for a given model. These complex techniques are described in Algorithm 1 specifically.

3.2. Long Short-Term Memory Network

Long Short-Term Memory (LSTM) networks are a specialized form of RNN introduced to overcome the vanishing and exploding gradient problems. Long-term dependencies, such as those found in LSTM, can retain relevant information over hundreds or even thousands of time-steps, making it especially effective for tasks that require long-range temporal context [1].

Cell state layer. At the heart of the LSTM lies the cell state, which provides a direct route for information propagation over time. This "information highway," illustrated as the horizontal connection across the top of the model, preserves long-term dependencies and mitigates gradient decay.

Forget gate layer. The first step inside an LSTM cell is deciding which information to discard from the previous cell state. This decision is made by the sigmoid-activated *forget gate*. It receives the previous hidden state $\mathbf{h}_{t-1}$ and the current input $\mathbf{x}_t$ and returns a vector $\mathbf{f}_t$ whose elements lie in the interval $[0, 1]$. For each component of the previous cell state $\mathbf{c}_{t-1}$, a value of 1 in $\mathbf{f}_t$ means that information is kept entirely, whereas a value of 0 means it is forgotten completely, as described in the equation 3.

$$\mathbf{f}_t = \sigma\big(\mathbf{W}_f \begin{bmatrix} \mathbf{h}_{t-1} \\ \mathbf{x}_t \end{bmatrix} + \mathbf{b}_f\big). \tag{3}$$

Enter gate layer. The next step is to decide what new information should be added to the cell state. This involves two sub-steps. First, a sigmoid-activated layer called the *input gate layer* selects which elements to update in the formula (see Equation 4). Continue it is a *tanh* layer that generates a vector for the new value $\tilde{\mathbf{C}}_t$ added to the cell state in the given equation 5. In the following step, we will combine 2 values to generate the updated cell state

$$\mathbf{i}_t = \sigma\big(\mathbf{W}_i \begin{bmatrix} \mathbf{h}_{t-1} \\ \mathbf{x}_t \end{bmatrix} + \mathbf{b}_i\big). \tag{4}$$

$$\tilde{\mathbf{C}}_t = \tanh\big(\mathbf{W}_c \begin{bmatrix} \mathbf{h}_{t-1} \\ \mathbf{x}_t \end{bmatrix} + \mathbf{b}_c\big). \tag{5}$$

Cell-state update. After that, it's time to update the old cell state to become the new cell state. With the forget vector $\mathbf{f}_t$ and the two input components $\mathbf{i}_t$ and $\tilde{\mathbf{C}}_t$ in place, we can process this step. First, the old state $\mathbf{C}_{t-1}$ is element-wise multiplied by $\mathbf{f}_t$, discarding the information we chose to forget in the previous layer, and plus with $\mathbf{i}_t \odot \tilde{\mathbf{C}}_t$. Then the new state from the result depends on how we decide to update the value for each state as mentioned in the equation 6.

$$\mathbf{C}_t \;=\; \mathbf{f}_t \odot \mathbf{C}_{t-1} \;+\; \mathbf{i}_t \odot \tilde{\mathbf{C}}_t. \tag{6}$$

Output gate. The last step is to decide what the LSTM will output at the current time step. Although the hidden output is based on the updated cell state $\mathbf{c}_t$, it passes through an additional filtering stage. First, a sigmoid-activated function determines which parts of the cell state to output in the equation 7. Then the updated cell state is passed through a $\tanh$ activation to a value in range $[-1, 1]$, and multiply it with the output sigmoid gate to gain the new hidden output in the formula 8.

$$\mathbf{o}_t \;=\; \sigma\big(\mathbf{W}_o \begin{bmatrix} \mathbf{h}_{t-1} \\ \mathbf{x}_t \end{bmatrix} + \mathbf{b}_o\big). \tag{7}$$

$$\mathbf{h}_t \;=\; \mathbf{o}_t \odot \tanh(\mathbf{C}_t). \tag{8}$$

The vector $\mathbf{h}_t$ serves both as the output of the LSTM cell for the current time step and as the recurrent input to the next time step.

3.3. The proposed hybrid HMM-LSTM model

This section introduces a hybrid modeling framework designed to improve the accuracy of cryptocurrency price forecasting. Conventional LSTM models typically rely on historical closing prices to predict subsequent values. However, as demonstrated by [22], this univariate approach exhibits limited predictive power and is often inadequate for real-world financial environments characterized by volatility and non-stationarity. To enhance predictive performance, the proposed model extends the LSTM input space by incorporating additional explanatory variables that are empirically known to influence price dynamics, including the previous day's opening, high, and low prices, intraday volatility, and traded volume. In parallel, HMM is employed to identify latent market regimes and extract sequential state indicators that characterize structural shifts in market behavior. These HMM-derived state features are subsequently integrated into the LSTM network to enable regime-aware forecasting.

Algorithm 1: Regime-Aware HMM–LSTM Forecasting

Input: $\mathcal{D}$ – Time-ordered predictor and target observations
$\mathcal{K}$ – Candidate set of HMM hidden-state numbers
h – Forecast horizon
w – Input sequence length
Output: $\mathcal{M}$ – Trained HMM–LSTM forecasting framework
s^* – Selected number of hidden states

`// Phase 1: Temporal partitioning`
Partition $\mathcal{D}$ chronologically into training, validation, and test sets
`// Phase 2: Hidden regime identification`
Extract the predictor representation X_{HMM} from the training observations
$s^* \leftarrow \varnothing$
$BIC_{\min} \leftarrow +\infty$
foreach $k \in \mathcal{K}$ **do**
    Fit an HMM with k hidden states using X_{HMM}
    if *the HMM converges* **then**
        Compute BIC_k
        if $BIC_k < BIC_{\min}$ **then**
            $BIC_{\min} \leftarrow BIC_k$
            $s^* \leftarrow k$
        end if
    end if
end foreach
Fit the final HMM with s^* hidden states using the training observations
Infer the regime representation using the fitted HMM
`// Phase 3: Regime-aware sequence construction`
Augment the predictive representation with the inferred regime information
Estimate preprocessing parameters using the training observations only
Transform the training, validation, and test predictors using the estimated preprocessing parameters
Construct input sequences of length w and corresponding h-step-ahead targets
`// Phase 4: Sequence forecasting`
Initialize the LSTM forecasting model according to the predefined architecture
Train the LSTM using the training sequences and validation data
`// Phase 5: Forecast generation and evaluation`
Generate forecasts for the test observations
Compute the predefined forecasting-error metrics
$\mathcal{M} \leftarrow$ (HMM, LSTM, preprocessing procedure)
return $\langle \mathcal{M}, s^* \rangle$

The conceptual motivation for this hybrid approach stems from prior studies, such as [1], which demonstrate the effectiveness of combining HMM-driven representations with deep recurrent neural networks. Furthermore, to capture complex temporal and technical dependencies, the model incorporates a diverse set of technical indicators reflecting trend strength, momentum, and volatility. The complete feature engineering and data preprocessing procedures are detailed in Section 4.2.

The proposed framework, outlined in Algorithm 1, integrates traditional financial mathematics with advanced machine learning techniques, specifically stacked Long Short-Term Memory (LSTM) neural networks and Hidden Markov Models (HMMs). By leveraging the LSTM's ability to process sequential dependencies alongside the HMM's capacity for regime identification, the model effectively captures the non-linear dynamics inherent in financial time series.

In phase 1: Latent Market Regime Identification via HMM, we utilize: **GaussianHMM** initializes the Markov model with continuous Gaussian emissions; **CalculateBIC** computes the Bayesian Information Criterion score for model optimization; and **CreateMultiStepSequences** applies a sliding window technique to transform the continuous time series into supervised input-output sequences consisting of 60 historical time steps to predict h future steps. To find the optimal number of hidden states, s^*, we perform a search evaluation over the range of $k \in [2, 20]$. The case of $k = 1$ usually returns a high BIC score, as a single-state model assumes market stationarity, directly contradicting the fundamental hypothesis of regime-switching in cryptocurrency markets and the volatility of the environment. The maximum range at 20 states was established to prevent overfitting, as higher states will make model learning more complex and capture noise. In phase 2: Sequence Modeling and Training, the LSTM architectural parameters (100 and 50 hidden units, with a dropout rate of 0.2) were adopted based on established conventions in time-series forecasting. This standard configuration provides an effective balance between capturing complex temporal dependencies and mitigating the risk of overfitting.

Beyond predictive accuracy, the model emphasizes interpretability and reliability by producing simulation outputs accompanied by evaluating performance across multiple independent training runs with statistically rigorous confidence intervals. This hybrid configuration enables the dynamic estimation of drift and volatility parameters, thereby improving forecast precision and facilitating robust uncertainty quantification. Overall, the approach contributes to the development of resilient forecasting architectures that enable the digital transformation of economic and financial systems.

3.4. Performance Evaluation Metrics

Table 1 employs a diverse set of evaluation metrics. These metrics evaluate the accuracy, robustness, and proportional error characteristics of the hybrid HMM–LSTM models for predicting cryptocurrency prices.

Here N_{sample} denotes the total number of test samples. C_t and $\widehat{C}_t$ represent the actual and predicted closing prices at time t, respectively. Lower values of MAPE, sMAPE, MPE, RMSE, RMSPE, MAE and Theil's Inequality Coefficient (U2) indicate better forecasting accuracy.

3.5. Naive baseline

To analyze the model's effectiveness in adapting to the market, we will employ a naive baseline using a persistence model. A persistence model is a basic forecasting approach that predicts future values by assuming they will remain identical to the previous day's observation:

$$\hat{P}_{t+k|t} = P_t$$

Table 1. Summary of performance evaluation metrics used for hybrid HMM–LSTM forecasting.

Metric	Mathematical Definition
Mean Absolute Error (MAE)	$\text{MAE} = \frac{1}{N_{\text{sample}}} \sum_{t=1}^{N_{\text{sample}}} \lvert C_t - \hat{C}_t \rvert$
Mean Absolute Percentage Error (MAPE)	$\text{MAPE} = \frac{1}{N_{\text{sample}}} \sum_{t=1}^{N_{\text{sample}}} \left\lvert \frac{C_t - \hat{C}_t}{C_t} \right\rvert$
Symmetric Mean Absolute Percentage Error (sMAPE)	$\text{sMAPE} = \frac{1}{N_{\text{sample}}} \sum_{t=1}^{N_{\text{sample}}} \frac{2\lvert C_t - \hat{C}_t \rvert}{\lvert C_t \rvert + \lvert \hat{C}_t \rvert}$
Mean Percentage Error (MPE)	$\text{MPE} = \frac{1}{N_{\text{sample}}} \sum_{t=1}^{N_{\text{sample}}} \frac{C_t - \hat{C}_t}{C_t}$
Root Mean Squared Error (RMSE)	$\text{RMSE} = \sqrt{\frac{1}{N_{\text{sample}}} \sum_{t=1}^{N_{\text{sample}}} \left(C_t - \hat{C}_t\right)^2}$
Root Mean Squared Percentage Error (RMSPE)	$\text{RMSPE} = \sqrt{\frac{1}{N_{\text{sample}}} \sum_{t=1}^{N_{\text{sample}}} \left(\frac{C_t - \hat{C}_t}{C_t}\right)^2}$
Theil's Inequality Coefficient (U_2)	$U_2 = \frac{\sqrt{\frac{1}{N_{\text{sample}}-1} \sum_{t=2}^{N_{\text{sample}}} \left(\frac{C_t - \hat{C}_t}{C_{t-1}}\right)^2}}{\sqrt{\frac{1}{N_{\text{sample}}-1} \sum_{t=2}^{N_{\text{sample}}} \left(\frac{C_t - C_{t-1}}{C_{t-1}}\right)^2}}$

where $\hat{P}_{t+k|t}$ is the forecast at time t for the look-ahead time k, and P_t is the measurement at time t. This model performs well for short forecasting horizons, often yielding low prediction errors. As our study focuses on learning daily market changes, benchmarking our results against this persistence model will help us analyze how well our proposed model adapts to environmental volatility and whether it actually learns meaningful patterns.

4. Design and Evaluation of Benchmarking Strategies

4.1. The proposed workflow of the study

This section presents the implementation workflow of the proposed hybrid forecasting framework, as illustrated in Figure 1. The methodological pipeline consists of four main stages: data collection, feature extraction, model construction, and performance evaluation. In the first stage, cryptocurrency price series are retrieved from reliable financial data sources (Section 4.1). Next, these raw data are transformed into predictive feature sets comprising both market-derived and technical indicators (Section 4.2). In the modeling phase, the Hidden Markov Model (HMM) is employed to identify latent market regimes and generate state-dependent representations, which are subsequently utilized by a Long Short-Term Memory (LSTM) network to forecast price movements (Section 4.5). Finally, model performance is quantitatively assessed using multiple error-based metrics, includ-

ing Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Root Mean Square Percentage Error (RMSPE) (Section 5).

The proposed hybrid pipeline demonstrates the potential of combining probabilistic and deep learning methods to improve the predictive accuracy and robustness of cryptocurrency value forecasting. Furthermore, selecting an optimal model configuration and integrating complementary techniques enhances overall precision, making this framework a promising approach for data-driven financial prediction tasks.

The proposed hybrid HMM–LSTM framework is implemented through a three-stage modeling process. In the first stage, a Gaussian Hidden Markov Model (HMM) is trained on the selected feature set, comprising open, close, high, low prices, and derived technical indicators. The optimal number of hidden states, n, is determined by minimizing the Bayesian Information Criterion (BIC). The inferred hidden-state sequence, which captures latent market regimes, is subsequently appended as an additional explanatory feature.

In the second stage, all features, including the HMM-inferred states, are normalized using a Standard Scaler scaling technique and transformed into multi-step temporal sequences of length n input to enable the forecasting of h-step-ahead values. In the third stage, a multi-layer LSTM architecture, comprising stacked LSTM and Dropout layers, is trained on these sequences to exploit its capacity to model the nonlinear dependencies and long-term temporal correlations inherent in financial time series In the multi-step forecasting mechanism, let $X_t = [x_{t-w+1}, \ldots, x_t]$ represent the multivariate input sequence ending at time t, where w is the window size. Each vector x_i within this sequence contains the concatenated market features, technical indicators, and the HMM-derived state variable. The hybrid HMM-LSTM network, parameterized by its learned weights Θ, maps this input sequence to an h-step ahead prediction vector. The output layer of the network utilizes a dense, fully connected configuration to project the final hidden state into an h-dimensional space:

$$\hat{\mathbf{C}}_{t+1:t+h} = f_{LSTM}(X_t; \Theta)$$

where h denotes the defined forecast horizon, and $\hat{\mathbf{C}}_{t+1:t+h}$ is the predicted vector containing the sequential closing prices:

$$\hat{\mathbf{C}}_{t+1:t+h} = [\hat{C}_{t+1}, \hat{C}_{t+2}, \ldots, \hat{C}_{t+h}]^T$$

This equation demonstrates how the final dense layer transforms the temporal representations extracted by the LSTM into a simultaneous, multi-step output of length h.

Following the typology proposed by Alexei Botchkarev [7], model performance is assessed using six widely adopted evaluation metrics: Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Symmetric Mean Absolute Percentage Error (sMAPE), Mean Percentage Error (MPE), Root Mean Square Error (RMSE), and Root Mean Square Percentage Error (RMSPE). These indicators encompass both scale-dependent (MAE, RMSE) and scale-independent, percentage-based measures (MAPE, sMAPE, MPE, RMSPE), thereby providing a comprehensive and unbiased evaluation of predictive accuracy and error symmetry. The comparative results across all metrics are summarized in Table 1.

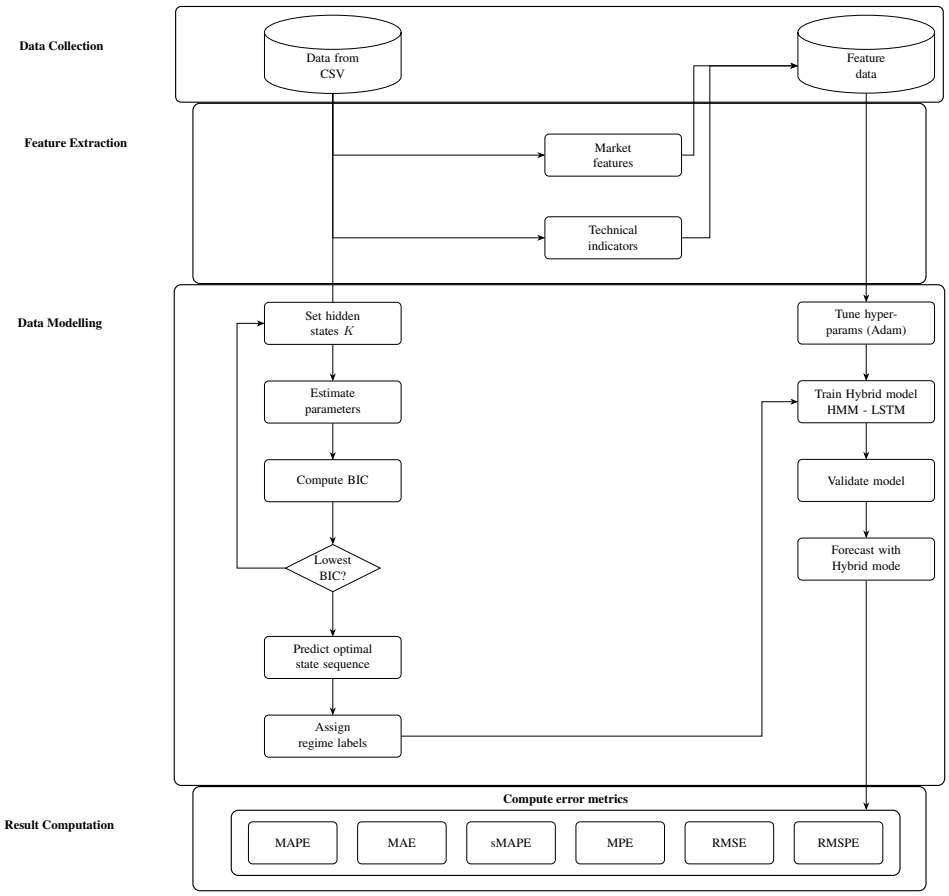


Fig. 1. Workflow of the hybrid HMM–LSTM forecasting

4.2. Collect Data and Extract Features

For dataset preparation, daily closing price series for Bitcoin (`BTC-USD`) and Ethereum (`ETH-USD`) were collected from multiple reputable financial data providers, including Bybit, Exness, and Financial Modeling Prep. These datasets encompass a continuous historical period from 17 January 2021 to 1 July 2025. All datasets were retrieved in comma-separated values (CSV) format to facilitate preprocessing and integration.

The Bitcoin dataset serves as the primary source for model development and training, while the Ethereum dataset is employed for out-of-sample validation to assess the generalization capability and robustness of the proposed hybrid model. This two-stage evaluation strategy ensures that model performance is not confined to a single cryptocurrency, thereby enhancing the reliability and external validity of the forecasting framework.

Table 2. Training, validation, and test periods for Bitcoin and Ethereum.

Asset	Train		Validation		Test	
	Start	**End**	**Start**	**End**	**Start**	**End**
Bitcoin	17/01/2021	28/02/2024	29/02/2024	29/10/2024	30/10/2024	01/07/2025
	1138 days (70%)		**244 days (15%)**		**245 days (15%)**	
			Total: 1627 days			
Ethereum	17/01/2021	28/02/2024	29/02/2024	29/10/2024	30/10/2024	01/07/2025
	1138 days (70%)		**244 days (15%)**		**245 days (15%)**	
			Total: 1627 days			

As summarized in Table 2, the dataset is partitioned into three distinct subsets: training, validation, and testing. The training subset is employed to fit the model parameters, the validation subset is used for hyperparameter tuning and model selection, and the testing subset serves to evaluate the final predictive performance. This systematic partitioning strategy minimizes overfitting and ensures an unbiased assessment of generalization capability.

The raw dataset comprises fundamental market attributes derived from the following daily fields: date, open, high, low, close, adjusted close, and trading volume. These variables collectively describe the primary market dynamics of each cryptocurrency. To enrich the feature space and capture additional market behavior not directly observable from raw prices, supplementary variables are extracted or computed from these base attributes.

All resulting features are categorized into two primary groups: (i) **market features**, representing directly observable price and volume information; and (ii) **technical indicators**, representing higher-order statistical and temporal transformations known to influence cryptocurrency price movements.

Market Features: The market features are either directly extracted from or derived using the daily price and trading volume series. In addition to the fundamental variables — open, high, low, close prices, and trading volume — two supplementary variables are computed to form a comprehensive market feature vector used in the subsequent modeling

stage. These extended features aim to capture nuanced market behaviors and improve the descriptive representation of short-term price dynamics.

Market spread. The *market spread* measures the intraday amplitude between the highest and the lowest traded prices, and is calculated in the function 9.

$$\text{Spread}_t \;=\; \text{High price}_t - \text{Low price}_t. \tag{9}$$

Daily change. The *daily change* captures the percentage movement of the closing price relative to the previous day as mentioned in the equation 10.

$$\text{Change}_t \;=\; \frac{\text{Close price}_t - \text{Close price}_{t-1}}{\text{Close price}_{t-1}}. \tag{10}$$

Technical Indicators represent quantitative transformations of historical price data designed to capture underlying market dynamics such as momentum, trend strength, and volatility behavior. In this study, each indicator is systematically derived from the closing price series and incorporated as an additional feature within the modeling framework. The selected indicators are as follows:

– **Relative Strength Index (RSI)** – A momentum oscillator used in technical analysis. RSI measures the speed and magnitude of recent price changes to identify overbought or oversold conditions in the price of a financial asset. The RSI is calculated using the following formula 11.

$$\text{RSI} = 100 - \left(\frac{100}{1 + \frac{\text{Average Gain over } n \text{ day}}{\text{Average Loss over } n \text{ day}}} \right) \tag{11}$$

RSI values range between 0 and 100. Typically, an RSI above 70 indicates that the asset is overbought, while an RSI below 30 suggests that it is oversold. The most common period used for calculation is 14 days.

– **Moving Average (MA)** – arithmetic mean of the last n closing prices, often used to smooth out price fluctuations; Given our daily sampling frequency, we compute six moving averages: MA_5, MA_{20}, MA_{50}, MA_{88}, MA_{100}, and MA_{200}, corresponding to time periods of 5, 20, 50, 88, 100, and 200 trading days, respectively. For an n-day window, the SMA is defined as mentioned in the equation 12.

$$\text{MA}_n(t) \;=\; \frac{1}{n} \sum_{i=1}^{n} \text{Close}_{\,t-i+1}. \tag{12}$$

– **Exponential Moving Average (EMA)** – an exponentially weighted moving average of the asset price, by placing exponentially decreasing weights on the price. This weighting scheme enables the EMA to respond more quickly to the most recent prices, thereby contributing the most to the average, and is therefore considered more informative for forecasting future movements. With **n** with a sequence of time, EMA can be calculated in the function 13.

$$\mathrm{EMA}_t \;=\; \alpha \sum_{i=1}^{n} \mathrm{Close}_{\,t-i+1} \big(1-\alpha\big)^{\,i-1}, \tag{13}$$

where α is the smoothing factor that controls the rate of weight decay at each point. In line with our daily sampling frequency, we compute four exponential moving averages: $\mathrm{EMA}_5, \mathrm{EMA}_{20}, \mathrm{EMA}_{50}$, and EMA_{100}, corresponding to time periods of 5, 20, 50, and 100 trading days, respectively.

– **Standard Deviation (STD)** — a descriptive statistical quantity that is not too unfamiliar to us, often used to measure the dispersion level of a data set that has been tabulated into a frequency table, corresponding to time periods of 5, 20, 50, 88, 100, and 200 trading days. For a window of n trading days, the rolling STD is computed in the equation 14.

$$\mathrm{STD}_n \;=\; \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \big(\mathrm{Close}_i - \overline{\mathrm{Close}}\big)^2}. \tag{14}$$

– **Moving-Average Convergence–Divergence (MACD)**. The MACD represents the relationship between two exponential moving averages (EMAs) and is calculated by subtracting the longer-period EMA from the shorter-period EMA. This indicator helps to understand price movements in a downtrend (where the price is expected to decline) and an uptrend (where the price is expected to rise), as mentioned in the function 15.

$$\mathrm{MACD}(t) \;=\; \mathrm{EMA}_a(t) - \mathrm{EMA}_b(t), \qquad a < b. \tag{15}$$

Several $\langle a, b\rangle$ pairs were evaluated. Empirically, the combination $\langle a, b\rangle = \langle 5, 20\rangle$ provided the most informative feature for the Ethereum data set, while $\langle a, b\rangle = \langle 5, 50\rangle$ performed best for the Bitcoin data set. Therefore, both MACD series are included in the final feature matrix.

– **Bollinger Bands.** Bollinger Bands consist of two lines positioned at a specified multiple of the rolling standard deviation (STD) above and below a moving average(MA) of the price. Empirically, roughly 90% of price observations fall between the two bands; a breakout beyond either band therefore signals an unusually large move. In this study, we use the 20-day simple moving average and a two-standard-deviation band as described in the equations 16 and 17 below.

$$\mathrm{UpperBand}_{20}(t) = \mathrm{MA}_{20}(t) + (2 \times \mathrm{STD}_{20}(t)), \tag{16}$$

$$\mathrm{LowerBand}_{20}(t) = \mathrm{MA}_{20}(t) - (2 \times \mathrm{STD}_{20}(t)). \tag{17}$$

Closing price above the upper Bollinger band typically indicates strong upward momentum, while prices below the lower band suggest significant downward pressure.

4.3. Dataset analysis and evaluation

Summary statistics. Following feature extraction, descriptive statistics were computed for each variable, including the minimum, maximum, mean, and standard deviation (rounded to two decimal places). Table 3 presents the statistical summary of all **29 extracted features** — comprising **seven market features** based on price and volume data, and **twenty-two technical indicators** derived from moving averages, volatility measures, and momentum oscillators. These statistics provide an initial overview of the data distribution and variability within the feature set.

Correlation Analysis. Figure 2 presents the heatmap of pairwise correlation coefficients among numerical features, highlighting the strength and direction of their linear relationships. Warm tones denote strong positive correlations, while cool tones indicate strong negative correlations, offering insight into feature interdependence and potential multicollinearity within the dataset.

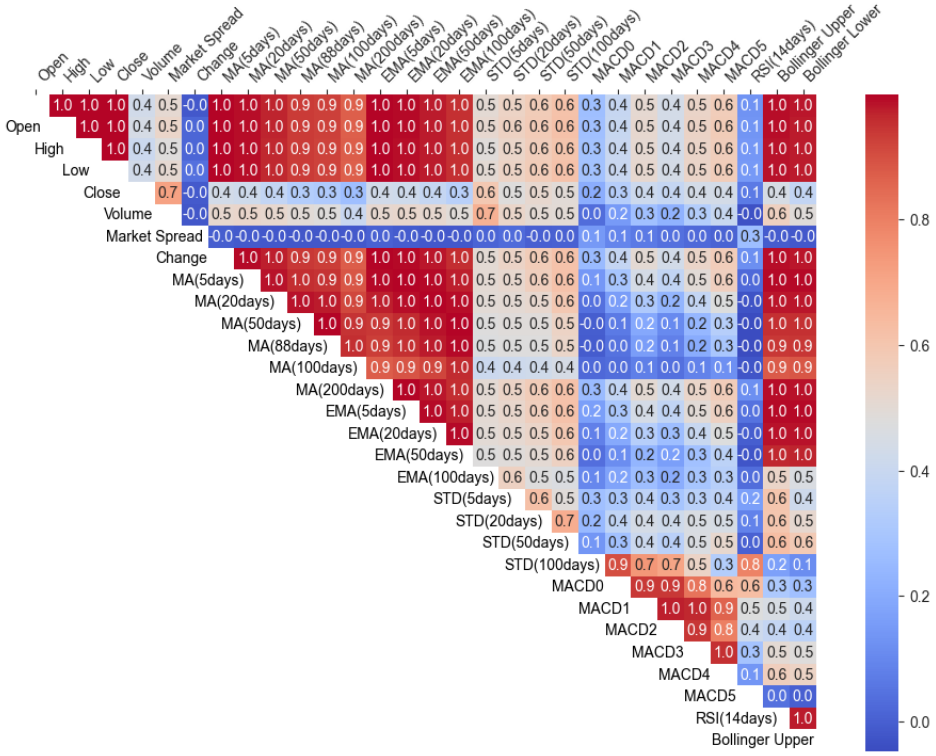


Fig. 2. Correlation Matrix of Key Predictive Variables.

Correlation between features. Understanding the interdependence among features is essential for analyzing factors that influence cryptocurrency price dynamics. Figure 2 presents

the correlation heatmap, where coefficients range from -1 to 1, indicating the strength and direction of linear relationships. The results reveal weak correlations between *close_price* and *change* (r = 0.0), highlighting the inherent randomness of cryptocurrency movements. The 14-day relative strength index (*rsi_14days*) exhibits a mild positive correlation with the closing price (0.1), suggesting limited momentum effects at extreme prices. Trading volume (*volume*) shows a moderate correlation with the closing price (0.4), reflecting that volume generally rises with market volatility. Meanwhile, market spread (*market_spread*) demonstrates a medium correlation (0.5), implying that widening spreads accompany price increases and signal higher transaction risk. The strong correlations observed among several other feature groups indicate potential multicollinearity, which should be carefully addressed in predictive modeling.

Forecast-model configurations. After feature extraction, each observation at time t is represented by a feature vector $\mathbf{x}_t$, as defined in Section 4.2. The LSTM model is trained to forecast the closing price at $t+1$ based on $\mathbf{x}_t$. The dataset comprises **1,776** trading days, divided into training, validation, and testing subsets to ensure robust model evaluation.

- **MA200.** Clare et al. [11] demonstrated that the 200-day moving average (MA200) can outperform buy-and-hold strategies for the S&P 500 by effectively capturing long-term market trends and mitigating short-term volatility through trend-following and stop-loss mechanisms. Conversely, Zakamulin [37] argued that many market-timing studies, including those based on moving averages, may suffer from data-mining bias and overlook transaction costs, resulting in inflated performance estimates.
- **MA88.** The 88-day moving average (MA88) serves as a mid-term technical indicator that effectively balances sensitivity to market fluctuations with noise reduction. Its ability to smooth short-term volatility while remaining responsive to emerging trends makes it well-suited for medium-horizon trading strategies. Compared to shorter averages, such as the 50-day MA, the MA88 captures broader market dynamics and provides improved stability in crossover strategies when combined with longer-term indicators, like the 200-day MA. Recent studies have highlighted the utility of moving averages in market trend analysis, underscoring their role in smoothing price data to reveal underlying trends [38].
- **Principal component analysis (PCA)** is a dimensionality reduction technique that transforms correlated variables into a set of uncorrelated principal components, thereby simplifying data complexity while retaining essential information. In financial applications, PCA helps identify key factors influencing asset volatility, thereby informing investment decisions. For instance, PCA has been applied to analyze the financial statements of over 3,000 medium and large enterprises, revealing significant factors such as central bank policy changes and short-term interest rate fluctuations. Additionally, PCA addresses multicollinearity issues in financial data, where high correlations among variables like opening, closing, high, and low prices can lead to model overfitting. Recent studies have demonstrated the efficacy of PCA in enhancing forecasting models by mitigating such challenges.
- **The Hidden Markov Model (HMM)** is a generative probabilistic model that assumes independence between observations, which can lead to redundancy in financial data. To address this, Principal Component Analysis (PCA) is employed to reduce

multicollinearity by transforming correlated features into a smaller set of orthogonal components. Eugene W. Park [28] demonstrated that combining PCA with HMM provides an intuitive understanding of market regimes through estimated emission probability distributions, yielding trading strategies with forecast accuracy slightly above 50

For cryptocurrency time series forecasting, a 60-day input window is often optimal. Qilan Lin [23] demonstrated that this window length achieved the lowest average RMSE compared to 14, 30, and 60-day windows, effectively balancing trend capture and noise reduction. This insight informs the proposed approach for selecting appropriate input days for prediction. In financial and cryptocurrency forecasting, the length of the input window has a significant impact on model prediction accuracy. Short windows may overlook crucial dependencies, while excessively long windows risk incorporating outdated patterns and overfitting. To optimize the window size for our LSTM model, we evaluated lengths of 10, 20, 30, and 90.

4.4. Data split technical for model training

This section assesses the performance of machine learning models on unseen test data. Both standard train-test splits and time series cross-validation are examined to identify the optimal dataset construction strategy. We implement K-fold time series cross-validation and adopt a 60-day input window, as recommended in recent studies for effective cryptocurrency forecasting [23].

- **Tradition train-test split (7:15:15):** The dataset was partitioned into training (75%), validation (15%), and testing (15%) subsets (Table 2). Model performance was assessed on the test set, following standard practice in time series forecasting.
- **Time Series Cross Validation:** Cross-validation is a key technique in machine learning for evaluating model performance on unseen data. K-fold cross-validation partitions the dataset into K equal folds, iteratively training on K–1 folds and testing on the remaining fold, with results averaged across iterations. In cryptocurrency forecasting, K-fold cross-validation helps optimize input window sizes, prevent overfitting, reduce data leakage, and assess model stability in recurrent neural networks (RNNs). For time series data, the standard K-Fold Cross-Validation is unsuitable because it ignores temporal dependencies. Instead, **Time Series Cross Validation** (TSCV) preserves the sequential order by ensuring the validation set always follows the training split, providing a more reliable evaluation of predictive performance [21]. Tandon [33] proposed a model for Bitcoin price prediction utilizing various neural network approaches, including Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks, in conjunction with 10-fold cross-validation. The LSTM model demonstrated superior performance, achieving a Mean Absolute Error (MAE) of 0.0043, outperforming both Random Forest and Linear Regression models. This study employs a progressive training approach, expanding the training window while shifting the validation and test windows forward in time. This method preserves temporal order and prevents data leakage, making it well-suited for financial and sequential forecasting tasks [5].

Table 3. Descriptive statistics of extracted features for Bitcoin (rounded to two decimals).

Feature	**Min**	**Max**	**Mean**	**Median**	**Std. dev.**
Open price	15758.96	111722.54	63740.75	48631.52	24664.64
High price	16279.99	112000.00	64140.00	49699.21	25095.06
Low price	15460.00	109217.23	62338.61	47499.06	24220.66
Close price	15760.14	111722.53	63741.33	48674.43	24703.11
Volume	5.33×10^9	3.51×10^{11}	1.78×10^{11}	3.49×10^{10}	2.22×10^{10}
RSI	6.66	99.45	53.06	52.30	17.35
Market spread	74.18	13591.70	6832.94	2200.15	1723.67
Daily change	-0.15	0.19	0.02	0.00	0.03
SMA_5	16277.40	109118.87	62698.14	48588.56	24584.31
SMA_{20}	16533.86	106809.28	61671.57	48268.18	24172.62
SMA_{50}	16776.59	106075.78	61426.19	47534.73	23339.16
SMA_{88}	17691.69	99467.69	58579.69	46582.05	22354.71
SMA_{100}	17894.03	98026.07	57960.05	46298.61	22105.32
SMA_{200}	15650.73	96281.71	55966.22	43876.27	20101.15
EMA_5	16265.17	108808.95	62537.06	48588.50	24566.79
EMA_{20}	16,758.31	106,275.72	61,517.02	48,254.28	24,092.30
EMA_{50}	16758.31	106275.72	61517.02	48254.28	24092.30
EMA_{100}	18,032.20	100,453.66	59,242.93	46,313.36	21,832.21
STD_5	17106.07	104208.19	60657.13	47529.47	23165.82
STD_{20}	340.43	12693.23	6516.83	3852.54	2438.16
STD_{50}	18032.20	100453.66	59242.93	46313.36	21832.21
STD_{100}	1275.30	15886.36	8580.83	5806.85	3392.42
$MACD_0$	-7938.26	9311.94	686.84	334.21	2286.30
$MACD_1$	-12655.56	18531.46	2937.95	1059.02	4796.91
$MACD_2$	-12902.77	24828.57	5962.90	2275.14	7271.67
$MACD_3$	-5910.48	9842.26	1965.89	724.81	2897.76
$MACD_4$	-9016.15	17515.53	4249.69	1940.93	5780.17
$MACD_5$	-4887.68	9210.09	2161.20	1216.12	3191.15
Bollinger upper	16979.62	112472.41	64726.01	52957.29	25913.75
Bollinger lower	13827.85	102111.37	57969.61	43579.06	22719.91

– **Evaluation and strategy selection.** To validate the proposed evaluation strategy, we train an LSTM model for 30 epochs and conduct 20 experimental runs on a representative market-indicator dataset (**Algorithm 2**). Two data splitting approaches are compared:
  1. **Tradition:** the dataset was divided into training, validation, and test sets at a 70:15:15 ratio.
  2. **Time Series Cross-Validation:** Experiments were conducted using K folds, where $K \in [5, 10, 15, 20, 25, 30]$. For each K, performance metrics include the mean and standard deviation across folds, as well as the minimum and maximum results to capture best- and worst-case scenarios [5].

Algorithm 2: Time-Series Cross-Validation Evaluation

Input: $\mathcal{D}$ – Time-ordered forecasting dataset
$\mathcal{K}$ – Candidate numbers of temporal folds
w – Input sequence length
h – Forecast horizon
F – Forecasting model
$\mathcal{M}$ – Evaluation metrics
Output: Aggregated validation results for each $K \in \mathcal{K}$

foreach $K \in \mathcal{K}$ **do**
  Construct K temporally ordered training–validation folds from $\mathcal{D}$
  $\mathcal{S}_K \leftarrow \emptyset$
  foreach *fold* $(\mathcal{D}_k^{train}, \mathcal{D}_k^{val})$ **do**
    Estimate preprocessing parameters using $\mathcal{D}_k^{train}$ only
    Transform $\mathcal{D}_k^{train}$ and $\mathcal{D}_k^{val}$ using the estimated parameters
    Construct input sequences of length w and corresponding h-step-ahead targets
    Train forecasting model F using the training sequences
    Generate forecasts for the validation fold
    $\epsilon_k \leftarrow$ Evaluate$(\mathcal{M})$
    $\mathcal{S}_K \leftarrow \mathcal{S}_K \cup \{\epsilon_k\}$
  end foreach
  Compute the minimum, maximum, mean, standard deviation, and median of the fold-level results in $\mathcal{S}_K$
end foreach
return aggregated validation results for all $K \in \mathcal{K}$

Table 4 summarizes the performance of each fold in the TSCV across 20 repeated experiments. For each experiment, the model predicts the validation fold, and the best and worst results correspond to the minimum and maximum error metrics across folds. As K increases from 5 to 25, the overall average error systematically decreases, with $K = 25$ achieving a mean MAE of $9,707.00 \pm 6,057.34$ and a mean RMSE of $12,590.53 \pm 7,533.39$. While splits ($K = 30$) produce slightly lower mean RMSE ($12,491.56$), they also exhibit higher variance (RMSE Std $= 7,882.64$) and severe worst-case degradation (RMSE Max $= 43,045.78$, $R^2_{\text{Min}} = -1.1196$) due to brief, highly volatile market subperiods. Therefore, $K = 25$ [or $K = 30$, reported via overall means] is selected as the

baseline configuration, offering the optimal balance between average predictive accuracy and cross-regime stability.

To evaluate the stability of the proposed framework, we compare two primary data-splitting strategies: experiments were conducted using the optimal K (selected based on the $\text{RMSE}_{\min}$) and a traditional method scheme. In the **traditional** (70:15:15), the LSTM model was trained on the training set and evaluated on the test set (indices 1356–1596), with the best model selected according to the lowest error. This direct comparison highlights the robustness and reliability of TSCV in forecasting sequential data [5].

Table 4. Summary of TimeSeriesSplit results (number of experiments = 20). An asterisk (*) marks the best K by lowest RMSE (Min).

Metric	5 Folds	10 Folds	15 Folds	20 Folds	25 Folds	30* Folds
RMSE Mean	14,901.73	13,091.60	12,873.68	12,678.43	12,590.53	12,491.56
RMSE Std	8,729.14	7,608.84	7,757.10	7,834.12	7,533.39	7,882.64
RMSE Min (Best)	4,729.44	3,764.78	3,604.73	3,430.65	2,843.04	**2,803.62**
RMSE Max (Worst)	29,964.81	30,778.23	35,766.56	33,773.27	41,382.97	43,045.78
R^2 Mean	0.5345	0.6188	0.5953	0.6056	0.6161	0.6096
R^2 Std	0.4191	0.3777	0.4175	0.3908	0.3914	0.4232
R^2 Max (Best)	0.9523	0.9660	0.9717	0.9754	0.9832	**0.9842**
R^2 Min (Worst)	-0.1761	-0.2543	-0.8847	-0.4133	-1.2467	-1.1196
MAE Mean	11,535.45	10,090.70	9,960.50	9,798.32	9,707.00	9,802.47
MAE Std	6,802.21	6,135.38	6,333.19	6,235.35	6,057.34	6,499.66
MAE Min	3,634.54	2,744.91	2,815.99	2,654.98	2,377.52	**2,222.14**
MAE Max	22,808.69	23,545.80	27,681.60	24,182.04	32,522.16	30,586.37

4.5. Data Modeling

As illustrated in the workflow (Figure 1), the modeling phase comprises two complementary components, each contributing to the overall predictive framework and performance evaluation.

Hidden Markov Model (HMM) Analysis. HMM captures historical price dynamics to forecast closing prices, interpreting each hidden state as a latent market regime. We model the observation distribution $P(\mathbf{y}_t \mid s_t = k)$ using a multivariate Gaussian HMM. The optimal number of hidden states is then used as input features for an LSTM network, which exploits sequential dependencies to enhance prediction accuracy and generate robust market indicators [28].

Feature–selection strategy. Leveraging HMM to capture latent market regimes, we implement a feature-grouping approach combining core market indicators and technical analysis features. Low-correlation features—such as closing price (*close_price*), trading volume (*volume*), 14-day RSI (*rsi_14days*), price change (*change*), and market spread

(*market spread*) — are selected to enhance stability and forecast accuracy (Table 5). To ensure the integrity of the feature set, we employed the Variance Inflation Factor (VIF) for multicollinearity analysis. As shown in Table 5, all selected features exhibit VIF values significantly below the common threshold of 5, with the Market Spread feature yielding the highest score at 2.34. These low values indicate minimal redundancy among the predictors, ensuring that each feature contributes unique information to the hybrid model without inflating the coefficient variance. High-correlation features are reduced via Principal Component Analysis (PCA), projecting top components to account for inter-feature dependencies, including price-volume interactions. These features, reflecting price dynamics, momentum, and volatility, are integrated with HMM-derived hidden states in a hybrid HMM-LSTM framework, evaluated using walk-forward validation to ensure robust forecasting performance [28,40].

Table 5. Variance Inflation Factor (VIF)

Feature	VIF
Close (price)	1.412897
Volume	2.093257
RSI(14days)	1.126025
Change	1.076766
Market Spread	2.343610

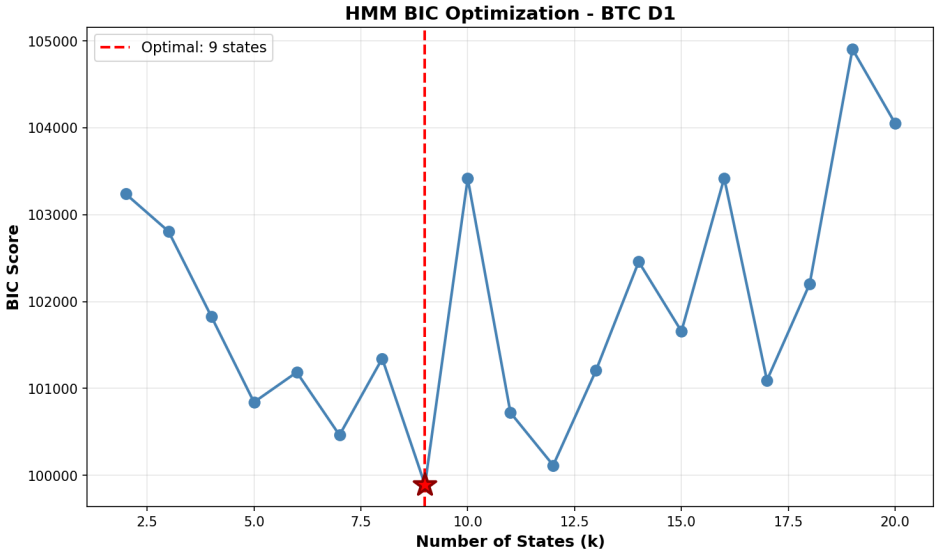


Fig. 3. Bayesian Information Criterion (BIC) for Different Hidden State Counts in the HMM.

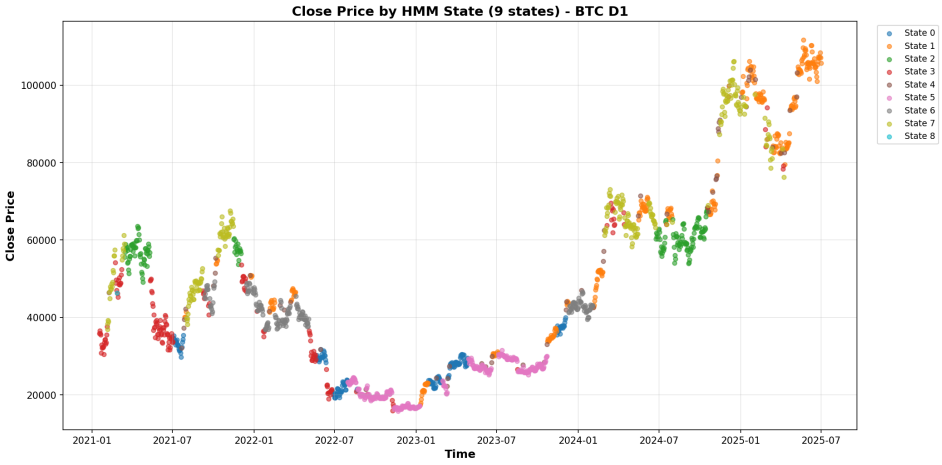


Fig. 4. HMM State Estimation for Bitcoin Price Data.

For each K, HMM parameters are estimated via maximum likelihood, and the Bayesian Information Criterion (BIC) is computed to select the optimal model. Figure 3 shows BIC values versus the number of hidden states, with the minimum indicating the optimal K. For the Bitcoin dataset, the lowest BIC occurs at **9 hidden states**, guiding the HMM configuration. This approach strikes a balance between model fit and complexity, enabling the accurate labeling of latent market regimes [28], [41]. Figure 4 shows that Bitcoin closing prices within specific ranges are typically assigned to the same hidden state, indicating a shared latent market regime. For instance, prices between \$20 000 – \$30 000 are predominantly cyan, reflecting a single regime. Increasing the number of hidden states divides the price axis into more detailed bands, creating categorical features for the forecasting model. This richer state space offers a more granular view of price movements, capturing different market behaviors such as bullish, bearish, or neutral trends. [28], [41].

LSTM Training and Feature Selection Strategy: As shown in Figure 1, once the dataset — including market variables, technical indicators, and HMM-derived state features — is prepared, an LSTM network is trained to predict next-day closing prices. The network comprises four stacked LSTM layers with a ($p = 0.2$) dropout to prevent overfitting, uses mean absolute error (MAE) as the loss function, and is optimized with Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$, learning rate 10^{-3}). Training uses a batch size of 100 for 30 epochs, with early stopping set to 10 epochs on validation loss. The same setup is used on the Ethereum dataset for out-of-sample testing, showing the model's ability to generalize. [20], [23].

Feature–selection strategy: In time series financial forecasting, datasets often include technical indicators, lag features, sentiment, and volume signals. While these features enable LSTM models to capture complex patterns, irrelevant or highly correlated features can introduce noise, increase complexity, and destabilize predictions, leading to overfitting. To mitigate these issues, we employ a Random Forest–based feature-selection strategy to identify the most informative technical features, reduce noise, and enhance

model robustness and generalization in cryptocurrency forecasting [40], [18]. In feature importance analysis for stock selection, Wang demonstrates that assessing the statistical relevance of features supports effective factor selection and improves predictive accuracy across sectors with distinct characteristics. Using a Random Forest (RF), feature importance is computed by averaging each feature's contribution across all trees, ranking them according to their influence on the target variable [9]. The dataset is split into training and test sets (80:20), with the model trained on the training set and evaluated on the test set. Features with higher importance scores are more influential and help mitigate overfitting in forecasts, as detailed in **Algorithm 3**, and are evaluated on both the training set and the test set[36], [18].

Algorithm 3: Time-Aware Random-Forest Feature Ranking

Input: X – Candidate predictor matrix
y – Forecast target series
Output: R – Ranked feature set

`// Step 1: Temporal alignment`
Lag predictor variables by one observation so that only information available before the forecast origin is used
Align the lagged predictors with the target series chronologically

`// Step 2: Temporal data partitioning`
Partition the observations chronologically into training and evaluation sets without random shuffling

`// Step 3: Random-Forest model fitting`
Fit a Random Forest regression model using the training observations
Evaluate the model on the training and evaluation sets

`// Step 4: Time-aware stability evaluation`
Construct temporally ordered validation folds from the training observations
Evaluate the fitted forecasting relationship across the validation folds
Compute the corresponding validation performance measures

`// Step 5: Permutation-based feature ranking`
Estimate the importance of each predictor by measuring the change in predictive performance after permutation
Rank the predictors in descending order of their estimated importance

$R \leftarrow$ ranked predictor set
return R

Figure 5 presents the Random Forest feature importance results. The 200-day moving average (MA200) is the most influential feature, while the 5-day moving average (MA5) has the lowest importance. This indicates that short-term variables, although highly correlated with the closing price, may increase the risk of overfitting and multicollinearity. To address these issues, features with importance scores below 0.04 are removed, focusing on stable variables to lower noise and improve the LSTM model's forecasting robustness. Higher importance scores indicate features with greater predictive power, aiding interpretability and feature selection. [36], [18].

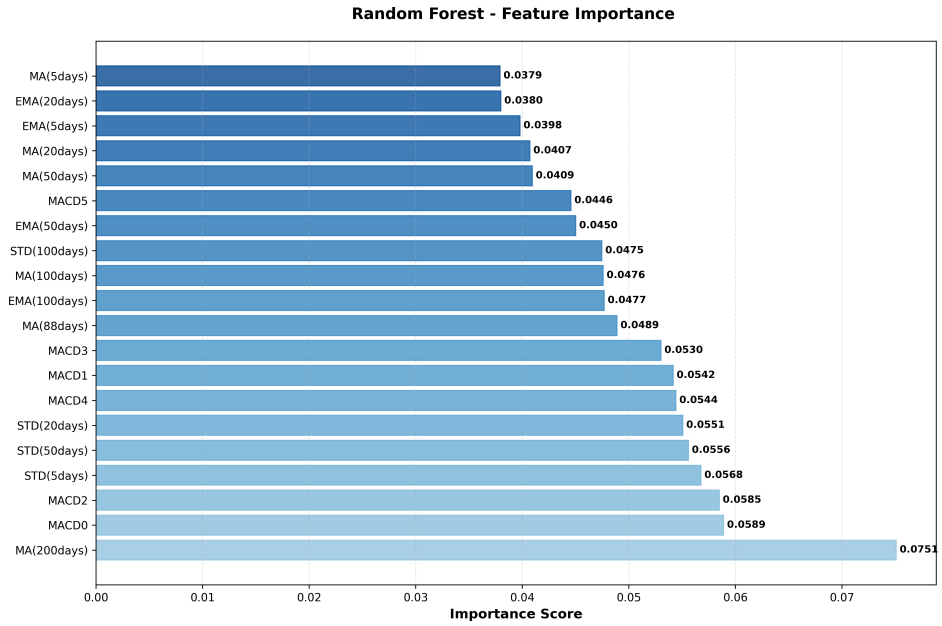


Fig. 5. Feature ranking based on **importance scores** obtained from the **Random Forest** model.

5. Experiment Results And Discussion

The experimental results demonstrate that the performance of the proposed hybrid HMM–LSTM framework varies across feature configurations and cryptocurrency assets. As demonstrated in Tables 6 and 8, selected hybrid configurations improve forecasting accuracy relative to standalone LSTM and HMM models in the BTC and ETH feature-specific experiments. This improvement results from the model's ability to combine latent regime information extracted by the HMM with nonlinear temporal dependencies captured by the LSTM. The HMM identifies distinct volatility states, offering a structural understanding of market behavior, while the LSTM dynamically learns sequential patterns within those regimes. Notably, using a single volatility-sensitive indicator, especially Bollinger Bands (BB), provides better forecasting accuracy. In contrast, combining multiple indicators such as EMA, MACD, and MA causes multicollinearity and redundant information, which increases prediction variance and slightly reduces generalization. This highlights the importance of feature simplicity in volatile financial areas. Additionally, the model maintains stable inference latency (around 1.1 seconds per prediction), confirming its suitability for real-time use.

The proposed Hybrid HMM–LSTM framework provides practical value for real-time cryptocurrency trading and risk management systems. By combining regime detection with temporal forecasting, it allows traders and automated agents to predict volatility shifts and improve their entry and exit strategy decisions. The model's low latency and

reliable performance during market turbulence make it suitable for algorithmic trading, portfolio rebalancing, and machine-to-machine (M2M) payment prediction within decentralized finance ecosystems. Additionally, its streamlined feature structure improves interpretability, supporting transparent decision-making and compliance in AI-driven financial systems, where explainability and dependability are essential for deployment in live blockchain environments.

Table 6. Comparison of single fixed holdout-test performance using the selected model instance on the Bitcoin dataset.

Model	MAE	MAPE (%)	sMAPE (%)	MPE (%)	RMSE	RMSPE (%)	Theil's U
Naive Persistence	1664.38	1.78	1.79	0.12	2324.56	2.53	1.000
LSTM	5554.00	5.65	5.85	5.49	6282.53	6.26	3.167
Hidden Markov Model (HMM)							
5 states	7127.55	7.59	7.85	2.93	9505.67	10.19	4.254
(optimal) 9 states	5227.46	5.67	5.83	1.51	7498.76	8.22	3.178
15 states	5824.37	6.23	6.45	2.21	7857.02	8.45	3.492
20 states	5732.53	6.14	6.39	2.02	8014.53	8.60	3.441
Hybrid (HMM + LSTM)							
Bollinger Bands (BB)	3972.57	4.03	4.14	3.68	4564.83	4.57	2.261
EMA	5381.80	5.46	5.64	5.27	5994.09	5.97	3.059
MA	6557.66	6.67	6.93	6.57	7171.01	7.17	3.736
MACD	5050.52	5.13	5.30	4.96	5641.94	5.65	2.877
STD	4255.31	4.32	4.44	4.08	4898.98	4.89	2.420
STD + EMA + MACD + BB	7041.09	7.13	7.44	7.10	7732.31	7.69	3.997
STD + EMA + MACD + BB + MA	9465.13	9.70	10.24	9.70	9975.43	10.08	5.437
MA + STD + EMA + MACD	9627.13	9.85	10.40	9.85	10 148.06	10.22	5.521

All models are evaluated on the train/validation/test split detailed in Table 2, with predictive performance measured using MAE, MAPE, sMAPE, MPR, MSE, RMSE, and RMSPE (Section 3.4). Comparative results in Table 6 highlight (i) the incremental value of incorporating regime information and (ii) which technical indicators provide the greatest marginal improvement when combined with market variables.

Table 6 summarizes the predictive performance of all models on the Bitcoin dataset using the evaluation metrics discussed earlier. The baseline LSTM models, trained solely on closing prices and extended feature groups, deliver moderate predictive accuracy, with the Bollinger Bands (BB) feature achieving the lowest RMSE of 6282.53. In contrast, the stand-alone HMM models demonstrate underperform performance underperform the baseline , reduces RMSE to 7498.76 with an optimal configuration of 9 hidden states. This finding confirms that capturing latent market regimes significantly enhances forecast precision [28].

The hybrid HMM–LSTM models exhibit the best overall performance when incorporating selective technical indicators. Specifically, the BB-based hybrid achieves the lowest RMSE (4,564.83), outperforming variants that utilize MA, EMA, STD, or MACD. Conversely, combining multiple indicators tends to degrade accuracy — for instance, the STD + EMA + MACD + MA combination yields an RMSE of 10,148.06 — **suggesting noise** and feature redundancy . These results indicate that combining regime-aware features

from HMM with a simple set of technical indicators improves both accuracy and stability. [28],

As detailed in Table 6, we utilize heil's U to compare our proposed models against the Naive baseline, which serves as a benchmark with a standard Theil's U of 1.00. During our evaluation of Bitcoin (BTC), we observed that all proposed models yielded a Theil's U greater than 1.00. The weaker BTC result suggests that incorporating HMM-derived states does not necessarily improve forecasting when the latent-state representation provides limited incremental information beyond the temporal patterns already learned by the LSTM. Additional analysis would be required to determine whether this result arises from regime instability, feature sensitivity, non-stationarity, or model-estimation variability. Nevertheless, among the evaluated complex models, the hybrid configuration incorporating Bollinger Bands (BB) achieves the lowest Theil's U of 2.261.

Consequently, the findings underscore that while HMMs improve the stability of regime-sensitive forecasting, excessive feature inclusion can impair generalization due to multicollinearity and noise. The hybrid HMM–LSTM configuration incorporating Bollinger Bands emerges as the most reliable and robust model under volatile market conditions.

Table 7. Distribution of performance over 50 independent training runs under the BTC feature-ablation experimentl instance on the Bitcoin dataset

Model	RMSE (mean ± std)	RMSE 95% CI	MAE (mean ± std)	MAE 95% CI
BB	6,496.15 ± 881.98	[6,251.68, 6,740.62]	5,827.27 ± 850.96	[5,591.40, 6,063.15]
STD	6,772.23 ± 997.80	[6,495.66, 7,048.81]	6,102.07 ± 981.36	[5,830.05, 6,374.09]
MACD	7,226.34 ± 808.15	[7,002.33, 7,450.35]	6,580.18 ± 788.80	[6,361.54, 6,798.83]
EMA	7,727.18 ± 842.96	[7,493.53, 7,960.84]	7,087.23 ± 827.35	[6,857.90, 7,316.56]
MA	9,209.66 ± 1,011.77	[8,929.22, 9,490.11]	8,585.46 ± 1,002.92	[8,307.46, 8,863.46]
STD_EMA_MACD_BB	10,827.84 ± 1,635.29	[10,374.56, 11,281.12]	10,265.05 ± 1,660.64	[9,804.75, 10,725.36]
MA_STD_EMA_MACD	13,837.69 ± 1,992.61	[13,285.37, 14,390.02]	13,371.93 ± 1,992.09	[12,819.75, 13,924.11]
STD_EMA_MACD_BB_MA	13,868.85 ± 2,246.54	[13,246.14, 14,491.56]	13,393.85 ± 2,259.29	[12,767.60, 14,020.09]

Table 7 reports results from 50 repeated runs of the hybrid HMM–LSTM model across several feature configurations. Performance is summarized by the mean RMSE, standard deviation, minimum and maximum errors, the best experiment index, and average inference latency (in milliseconds). Single-indicator configurations consistently outperform multi-indicator combinations. In particular, the Bollinger Bands (BB) feature attains the lowest mean RMSE ($6{,}496.15 \pm 881.98$) and the tightest 95% confidence interval [6,251.68, 6,740.62], indicating superior accuracy and stability. By contrast, full combinations (e.g., STD + EMA + MACD + BB + MA) yield substantially higher errors (mean RMSE = 13,868.85) and a larger variance ($\pm 2,246.54$), indicating feature redundancy and an increased risk of instability. These findings support a simple feature strategy — combining HMM-derived regime information with a small set of high-value indicators (notably BB) — to optimize predictive performance while managing complexity and latency [29].

Figure 6 compares actual and predicted closing prices for the BTC test dataset using LSTM, HMM, and hybrid HMM–LSTM models with features (STD + EMA + MACD + MA, Bollinger Bands). The hybrid model exhibits the closest tracking of market movements, capturing short-term volatility while maintaining stability during sharp fluctua-

tions, thereby demonstrating improved robustness and predictive consistency over standalone models [28]. Moreover, as depicted in Figure 6, the hybrid HMM–LSTM model achieves superior alignment with actual market movements compared to the standalone LSTM and HMM architectures. This finding is consistent with the quantitative results in **Table 6 and Table 7**, where the hybrid configuration incorporating Bollinger Bands (BB) attains the lowest RMSE and MAE values, indicating enhanced predictive accuracy. The hybrid model effectively integrates **the regime identification capability** of HMM with **the temporal pattern extraction** of LSTM, allowing it to capture short-term volatility while maintaining trend stability. In contrast, the LSTM tends to overreact to local fluctuations, and the HMM underestimates sudden regime shifts. Overall, these results confirm that the hybrid approach provides a more robust and adaptive framework for forecasting in volatile cryptocurrency markets.

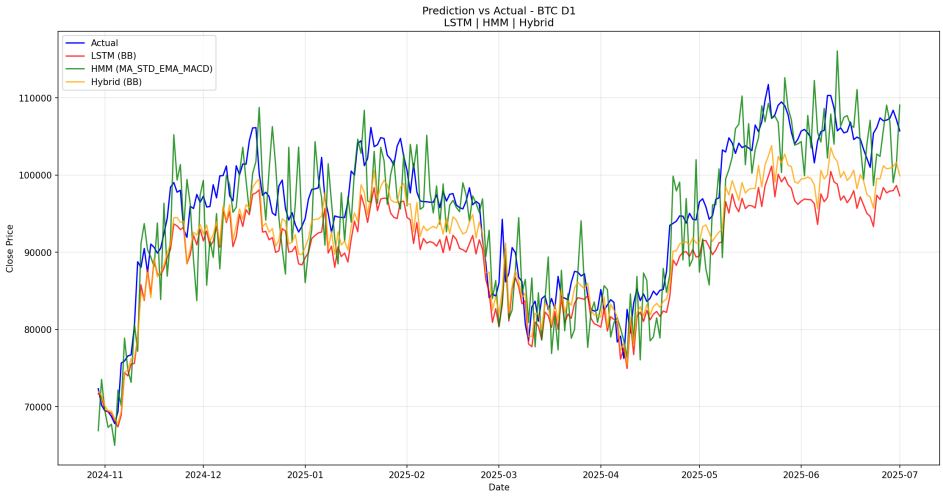


Fig. 6. Comparing test dataset BTC between features (STD+EMA+MACD+MA, Bollinger Bands) in LSTM, HMM, and Hybrid based on RMSE.

The Hidden Markov Model (HMM) effectively identifies latent market regimes, while the Long Short-Term Memory (LSTM) network captures nonlinear temporal dependencies. The hybrid integration layer merges both representations to produce robust, regime-sensitive forecasts for Ethereum price movements. To ensure methodological consistency and generalization, the experimental framework was replicated on the Ethereum dataset spanning 1,627 days, using the same training, validation, and testing configurations as described in Table 2. The HMM determined an optimal configuration with 9 hidden states, which was subsequently integrated with the LSTM for hybrid prediction.

As summarized in Table 8, the hybrid *HMM–LSTM* model combined with a single Bollinger Bands (BB) indicator achieved the lowest prediction error (MAE = 66.64, RMSE = 96.29), outperforming both the standalone LSTM and HMM models. A cross-

dataset comparison between Bitcoin and Ethereum (Tables 6 and 8) reveals consistent predictive behavior across assets. In both cases, the hybrid HMM–LSTM model, combined with a single Bollinger Bands (BB) indicator, achieved the lowest RMSE (BTC = 4564.84; ETH = 96.29), outperforming standalone LSTM and HMM baselines. The BB feature proved most effective in capturing short-term volatility and regime shifts. Conversely, multi-indicator configurations degraded accuracy, likely due to correlated or redundant inputs that introduced noise and overfitting. These findings confirm the robustness and generalizability of regime-aware hybrid frameworks across distinct cryptocurrencies, supporting prior studies emphasizing parsimonious deep-learning architectures for volatile financial markets [3], [19].

Conversely, our analysis of the Ethereum (ETH) dataset yielded highly favourable results. For ETH, the majority of our proposed models successfully outperformed the Naive baseline. Specifically, the BB-featured hybrid model achieved a Theil's U of 0.969 (< 1.00), demonstrating the model's capability to effectively learn market dynamics and surpass simple persistence forecasting under standard volatility conditions.

Table 8. Comparison of model performance on the Ethereum dataset.

Model	MAE	MAPE (%)	sMAPE (%)	MPE (%)	RMSE	RMSPE (%)	Theil's U
Naive Persistence Baseline	75.35	2.85	2.85	−0.12	106.72	4.09	1.000
LSTM (best feature set)	74.42	2.80	2.81	0.05	106.01	4.03	0.985
Hidden Markov Model (HMM)							
5 states	325.22	12.56	12.22	−3.33	404.14	15.94	4.411
9 states	343.87	13.08	12.88	−2.57	420.13	16.05	4.594
15 states	320.97	12.15	12.13	−1.31	400.33	15.04	4.267
(optimal) 20 states	230.89	8.79	8.86	−0.05	287.62	10.69	3.087
Hybrid (HMM + LSTM)							
Bollinger Bands (BB)	66.64	2.76	2.76	−0.03	**96.29**	4.06	0.969
EMA	67.20	2.77	2.78	0.02	97.17	4.08	0.973
MA	68.62	2.81	2.82	0.20	97.93	4.09	0.988
MACD	68.21	2.81	2.83	0.18	96.93	4.05	0.988
STD	70.08	2.90	2.89	−0.67	98.68	4.14	1.019
STD + EMA + MACD + BB	68.24	2.81	2.82	0.14	96.37	4.04	0.99
STD + EMA + MACD + BB + MA	70.35	2.94	2.93	−0.71	98.28	4.17	1.03
MA + STD + EMA + MACD	71.99	3.04	3.02	−0.82	97.02	4.15	1.07

Table 9 summarizes the performance of various feature configurations for the hybrid HMM–LSTM model on the Ethereum dataset, averaged over 20 repeated runs. Performance is summarized by the mean RMSE, standard deviation, minimum and maximum errors, the best experiment index, and average inference latency (in milliseconds). Single-indicator configurations consistently outperform multi-indicator combinations. In particular, the MACD feature attains the lowest mean RMSE (98.15 ± 0.53), narrowly outperforming the BB model (98.55 ± 1.04). By contrast, full combinations (e.g., STD + EMA + MACD + BB + MA) yield substantially higher errors (mean RMSE = 13,440.72) and a larger variance ($\pm 2,246.54$), indicating feature redundancy and an increased risk of instability. However, the BB configuration maintained the lowest Mean Absolute Error (68.74 ± 1.19) compared to MACD (70.09 ± 1.39) [29].

Table 9. Distribution of performance over 20 independent training runs under the ETH feature-ablation experiment.

Model	RMSE (mean ± std)	RMSE 95% CI	MAE (mean ± std)	MAE 95% CI
MACD	98.15 ± 0.53	$[97.92, 98.38]$	70.09 ± 1.39	$[69.48, 70.71]$
BB	98.55 ± 1.04	$[98.09, 99.00]$	68.74 ± 1.19	$[68.22, 69.27]$
EMA	98.73 ± 0.74	$[98.41, 99.05]$	69.13 ± 0.94	$[68.72, 69.54]$
STD_EMA_MACD_BB	98.81 ± 1.66	$[98.08, 99.54]$	71.99 ± 2.70	$[70.81, 73.18]$
MA	99.82 ± 1.32	$[99.24, 100.40]$	70.39 ± 1.28	$[69.83, 70.95]$
STD_EMA_MACD_BB_MA	101.27 ± 2.18	$[100.31, 102.22]$	75.38 ± 3.25	$[73.96, 76.81]$
MA_STD_EMA_MACD	101.75 ± 2.00	$[100.87, 102.62]$	75.81 ± 2.39	$[74.77, 76.86]$
STD	102.42 ± 2.49	$[101.33, 103.52]$	74.45 ± 2.51	$[73.35, 75.55]$

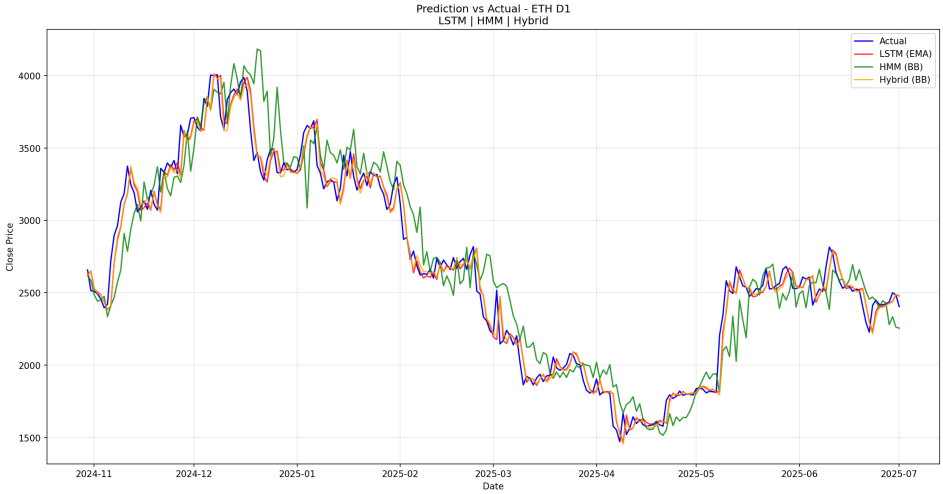


Fig. 7. Comparing test dataset ETH between features (EMA, Bollinger Bands) in LSTM, HMM, and Hybrid based on RMSE.

Tables 7 and 9 present feature configuration performance for Bitcoin and Ethereum using the hybrid HMM–LSTM model. For Bitcoin, the Bollinger Bands (BB) indicator achieves the lowest mean error (6,496.15 ± 881.98) with minimal deviation. For Ethereum, single-indicator features remain the most effective configurations, with MACD achieving the lowest mean error (98.15 ± 0.92), closely followed by BB (98.55 ± 1.04), and minimal deviation, confirming its superior predictive stability. In contrast, multi-indicator combinations, such as STD + EMA + MACD + MA, yield higher errors and variance, indicating overfitting and redundant feature interactions. The average inference latency remains stable at approximately 1.10 seconds per prediction, demonstrating computational efficiency. The consistent dominance of BB across assets highlights its sensitiv-

ity to volatility and short-term market shifts, supporting prior findings in volatility-driven forecasting models [29].

Figure 7 further demonstrates that the proposed hybrid model closely tracks actual market trends, effectively capturing short-term fluctuations and maintaining predictive stability during volatile market phases. Figures 6 and 7 illustrate the comparative forecasting performance of the proposed hybrid HMM–LSTM model against standalone LSTM and HMM baselines for Bitcoin and Ethereum, respectively. In both datasets, the hybrid configuration employing Bollinger Bands (BB) most closely follows actual market movements, effectively capturing rapid fluctuations and trend reversals. The model demonstrates strong adaptability to dynamic volatility regimes, maintaining predictive stability across distinct market phases. The similar tracking accuracy observed in both cryptocurrencies highlights the hybrid model's robustness and transferability across heterogeneous financial assets, reinforcing its regime-aware learning capability, as supported by prior empirical studies [19], [3].

Figures 6, 7, and 8 collectively evaluate the predictive behavior and training stability of the proposed hybrid HMM–LSTM model on Bitcoin (BTC) and Ethereum (ETH) datasets. Figures 6 and 7 show that the hybrid configuration, particularly when combined with Bollinger Bands (BB), most accurately tracks actual market prices, effectively capturing high-frequency volatility while minimizing forecast deviation compared to standalone LSTM and HMM models. This consistency across BTC and ETH underscores the model's generalization capacity and adaptability to heterogeneous market dynamics. Figure 8 further validates model convergence, with both training and validation losses exhibiting rapid stabilization after approximately 50 epochs, indicating robust optimization without overfitting. Notably, ETH demonstrates marginally lower validation loss variability, suggesting smoother learning dynamics. These findings confirm the reliability, predictive robustness, and computational efficiency of the regime-aware hybrid model for forecasting volatile cryptocurrencies [19], [3].

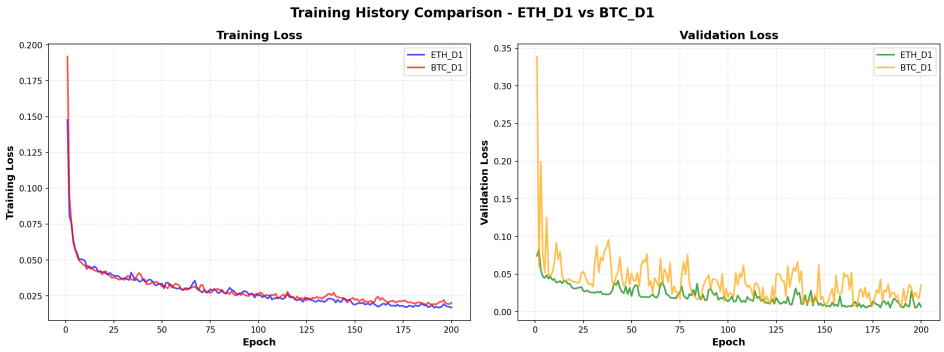


Fig. 8. Hybrid HMM–LSTM model trained on BTC and ETH using Bollinger Bands.

Figure 9 presents the runtime efficiency and hardware resource utilization of the proposed hybrid HMM–LSTM model trained on BTC and ETH datasets. The CPU utilization

indicates that BTC training exhibits greater fluctuation ($\approx$ 25–55%) compared to ETH ($\approx$ 10–25%), reflecting the higher computational demand caused by increased volatility and longer temporal dependencies. Memory consumption remains stable for both datasets, with BTC slightly exceeding ETH toward the final epochs due to more complex state transitions. Latency remains consistent at approximately 1.1 s per inference, with minor spikes observed near training completion, likely due to checkpoint synchronization. The performance distribution histogram demonstrates balanced resource usage and stable throughput, confirming the model's scalability and efficiency. Overall, these findings affirm that the hybrid HMM–LSTM framework delivers accurate and regime-sensitive forecasts while maintaining computational stability and efficient system utilization during both training and inference phases [19], [3].

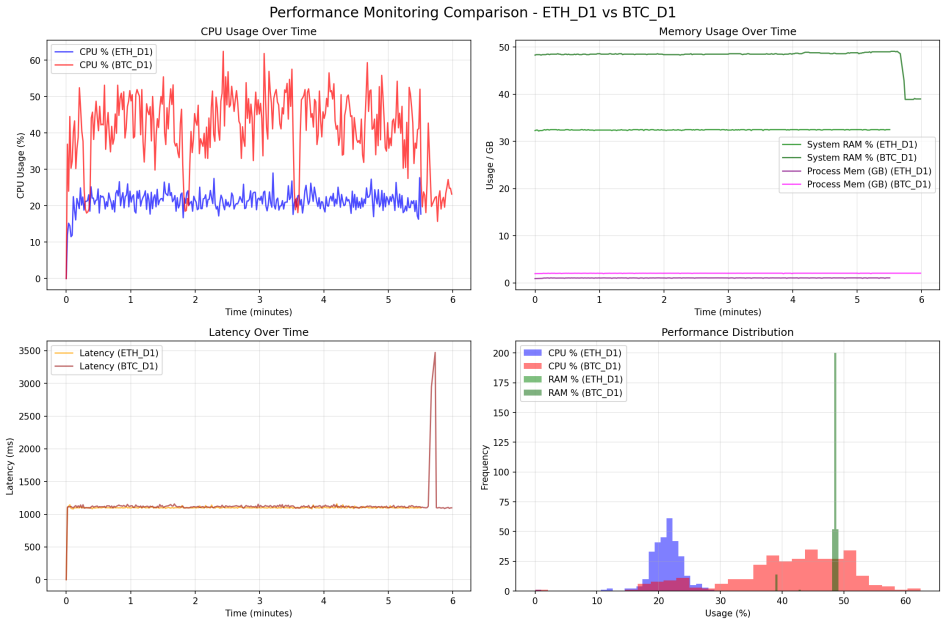


Fig. 9. Performance monitoring of the best hybrid HMM–LSTM model for BTC and ETH using the Bollinger Bands feature.

For further analysis, we have compared the hybrid model with other models like DeepAR [16] and ARIMA [30] across four different cryptocurrency types: Bitcoin(BTC), Ethereum(ETH), Binance Coin(BNB) and Solana(SOL). Table 10 shows the comprehensive predictive performance of the evaluated models across four cryptocurrency assets (BTC, ETH, BNB, and SOL). The evaluation framework compares our proposed hybrid LSTM-HMM architecture against traditional statistical methods (ARIMA), state-of-the-art probabilistic deep learning (DeepAR), and the standalone components of our architecture (LSTM and HMM) using RMSE, expressed as the mean ± standard deviation and a

95% confidence interval. For BNB and SOL, which have smaller volumes and volatility, the model performance may be weaker compared to the others. As observed in the results, we recommend the HMM-LSTM Hybrid as the primary model for deployment, with the standalone LSTM as a strong fallback baseline where lower complexity is preferred.

Table 10. Cross-asset benchmark under a common model-comparison protocol across ETH,BTC,SOL,and BNB using RMSE(mean ±std) and 95%CI.

Asset	Metric	LSTM	HMM	ARIMA	DeepAR	HMM + LSTM Hybrid
ETH	Mean ± Std	107.67 ± 1.58	406.09 ± 58.35	746.24 ± 65.39	101.92 ± 1.15	99.94 ± 2.23
	95% CI	[106.57, 108.76]	[387.03, 425.15]	[700.93, 791.56]	[101.12, 102.72]	[99.59, 100.28]
BTC	Mean ± Std	$8{,}153.96 \pm 1{,}696.28$	$9{,}459.47 \pm 1{,}468.25$	$20{,}454.46 \pm 3{,}720.74$	$17{,}711.28 \pm 2{,}248.11$	$9{,}495.74 \pm 3{,}167.45$
	95% CI	[6,978.49, 9,329.42]	[8,979.84, 9,939.10]	[17,876.12, 23,032.80]	[16,153.41, 19,269.14]	[9,185.33, 9,806.15]
SOL	Mean ± Std	8.56 ± 0.10	33.25 ± 4.26	39.97 ± 0.70	9.24 ± 0.33	8.56 ± 0.31
	95% CI	[8.49, 8.63]	[31.86, 34.64]	[39.49, 40.45]	[9.01, 9.47]	[8.51, 8.61]
BNB	Mean ± Std	16.98 ± 0.97	52.74 ± 10.71	52.60 ± 3.24	23.13 ± 2.73	15.87 ± 1.34
	95% CI	[16.31, 17.65]	[49.24, 56.24]	[50.36, 54.85]	[21.23, 25.02]	[15.66, 16.07]

The results indicate that the value of HMM-derived regime information is conditional on the asset, feature specification, and evaluation setting. Under the BTC and ETH feature-ablation experiments, selected single-indicator hybrid configurations often improve upon standalone HMM and LSTM models. However, the broader cross-asset comparison in Table 10 reveals heterogeneous performance. The hybrid model outperforms standalone LSTM for ETH and BNB, performs comparably for SOL, but underperforms LSTM for BTC. Accordingly, the empirical evidence supports an asset-dependent benefit of regime-aware fusion rather than universal superiority of the hybrid architecture.

Overall, these findings indicate that the regime-aware hybrid architecture provides conditional utility balancing model complexity and interpretability where market regimes are distinctly separated, rather than functioning as an unconditionally superior approach across all digital assets. These findings align with recent studies advocating for hybrid deep learning and probabilistic integration to improve financial time series modeling [28],[19].

6. Conclusion

This study addressed the critical challenge of forecasting cryptocurrency price behavior in highly volatile markets, emphasizing the influence of financial, technical, and macroeconomic factors on predictive performance. The primary objective was to develop a **hybrid HMM–LSTM framework** capable of modeling latent market regimes and temporal dependencies in cryptocurrency price movements, with a specific focus on Bitcoin—the most actively traded and valuable blockchain asset.

A hybrid HMM–LSTM framework is introduced for predicting cryptocurrency prices. The model identifies hidden market regimes and nonlinear temporal relationships. Bollinger Bands (BB) enhance prediction accuracy during volatile market periods. Comparative results demonstrate better performance than standalone LSTM and HMM models. The framework provides scalable, regime-aware predictions for blockchain-based finance.

The empirical results demonstrate that integrating HMM-derived regime information with LSTM forecasting can improve predictive accuracy, but the magnitude and direction of the improvement depend on the cryptocurrency and feature configuration. The hybrid model performs favourably for several asset–model combinations, particularly ETH and BNB, while standalone LSTM remains superior for BTC in the broader cross-asset benchmark. The results therefore support HMM–LSTM integration as a potentially useful regime-aware forecasting strategy rather than a universally superior alternative to standalone deep-learning models.

Moreover, feature selection plays a crucial role in predictive stability. The results reveal that using a single Bollinger Bands (BB) indicator, combined with HMM regime states and core market variables, yields the most consistent and accurate outcomes. Conversely, including multiple correlated indicators (e.g., EMA, MACD, MA, STD) leads to diminishing returns due to multicollinearity and feature redundancy. These findings align with recent research emphasizing the benefits of regime-aware hybrid frameworks that balance model complexity and generalization capability [19], [3].

Performance evaluations across multiple cryptocurrency assets indicate that the framework can generalize across markets, although its relative forecasting advantage varies by asset and model configuration. Training histories exhibit smooth optimization curves and stable validation losses after approximately 50 epochs, while performance monitoring reveals efficient memory usage and moderate CPU load, supporting the model's scalability for real-time financial forecasting applications.

Nevertheless, certain limitations remain. The model primarily relies on historical market data and does not incorporate real-time sentiment, blockchain transaction data, or macroeconomic indicators, which can significantly influence price dynamics. In addition, because the HMM is trained on a feature set that includes non-stationary price-level information (e.g., closing price and level-based technical indicators), the inferred hidden states may partially reflect price-range clustering rather. Consequently, the interpretation of these latent states as "market regimes" should be made with caution.

Future research should integrate these external data sources—such as news sentiment, on-chain analytics, and social media trends—to enhance model responsiveness to sudden market events and re-estimate the HMM using stationary regime descriptors(e.g., bull/bear conditions or high/low volatility) that are invariant to the absolute price scale. Additionally, expanding this framework to other emerging cryptocurrencies, including IoT-oriented assets like IOTA, may further validate its generalization capability.

In summary, the proposed **HMM–LSTM hybrid model** offers a robust and interpretable approach for cryptocurrency price forecasting under volatile market conditions. By effectively combining regime detection with deep sequence modeling, it advances the predictive precision and practical relevance of AI-driven financial analytics, paving the way for more adaptive and intelligent trading decision-support systems.

Acknowledgment. This research is partly funded by Ho Chi Minh City University of Technology and Engineering (HCM-UTE), Vietnam.

Conflict of Interest. The authors declare that there is no conflict of interest.

References

1. Abu Hashish, I., Forni, F., Andreotti, G., Facchinetti, T.: A hybrid model for bitcoin price prediction using hidden markov models and optimised lstm networks. In: 2019 International Conference on … (2019)
2. Avinash, G., Ramasubramanian, V., Ray, M., Paul, R.K., Godara, S., Nayak, G.H., Kumar, R.R., Manjunatha, B., Dahiya, S., Iquebal, M.A.: Hidden markov guided deep learning models for forecasting highly volatile agricultural commodity prices. Applied Soft Computing 158, 111557 (2024)
3. Avinash, G., Ramasubramanian, V., Ray, M., Paul, R.K., Godara, S., Nayak, G.H., Kumar, R.R., Manjunatha, B., Dahiya, S., Iquebal, M.A.: Hidden markov guided deep learning models for forecasting highly volatile agricultural commodity prices. Applied soft computing 158, 111557 (2024)
4. Awad, M., Khanna, R.: Hidden Markov Model, pp. 81–104. Apress, Berkeley, CA (2015), accessed from ResearchGate
5. Bergmeir, C., Benítez, J.M.: On the use of cross-validation for time series predictor evaluation. Information Sciences 191, 192–213 (2012)
6. Biswas, K., Muthukkumarasamy, V.: Securing smart cities using blockchain technology. In: 18th IEEE Int'l Conf. on High Performance Computing and Communications (HPCC) / SmartCity / DSS. pp. 1392–1393. Sydney, Australia (2016)
7. Botchkarev, A.: A new typology design of performance metrics to measure errors in machine learning regression algorithms. Interdisciplinary Journal of Information, Knowledge, and Management 14, 45–79 (2019)
8. Bouteska, A.: Comparative analysis of boosting ensembles, recurrent neural networks, and hybrid arima–rnn methods for cryptocurrency price prediction. Unpublished manuscript (2024), empirical evaluation on Bitcoin, Ethereum, Litecoin and more
9. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
10. Busari, G.A., Lim, D.H.: Crude oil price prediction: A comparison between adaboost-lstm and adaboost-gru for improving forecasting performance. Computers and Chemical Engineering 155, 107513 (2021)
11. Clare, A., Seaton, J., Smith, P.N., Thomas, S.: Breaking into the blackbox: Trend following, stop losses and the frequency of trading – the case of the S&P500. Journal of Asset Management 14(3), 182–194 (2013)
12. Conoscenti, M., Vetro, A., De Martin, J.C.: Blockchain for the internet of things: A systematic literature review. In: 13th IEEE/ACS Int'l Conf. on Computer Systems and Applications (AICCSA). pp. 1–6. Agadir, Morocco (2016)
13. Eggen, S., Espe, T.J., Grude, K., Risstad, M., Sandberg, R.: Financial time series uncertainty: A review of probabilistic ai applications. Journal of Economic Surveys (2025)
14. Erlwein-Sayer, C., Grimm, S., Pieper, A., Alsac, R.: Forecasting corporate credit spreads: Regime-switching in lstm. SSRN Electronic Journal (2021), available at SSRN: http://dx.doi.org/10.2139/ssrn.4003338
15. Fakhri, D., Mutijarsa, K.: Secure iot communication using blockchain technology. In: 2018 International Symposium on Electronics and Smart Devices (ISESD). pp. 1–6. Bandung, Indonesia (2018)
16. Ferenczi, A., Bădică, C.: Prediction of ethereum gas prices using deepar and probabilistic forecasting. Journal of Information and Telecommunication 7(4), 449–467 (2023)
17. Gurgul, V., Lessmann, S., Härdle, W.K.: Deep learning and nlp in cryptocurrency forecasting: Integrating financial, blockchain, and social media data. International Journal of Forecasting (2025)
18. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of machine learning research 3(Mar), 1157–1182 (2003)

19. Henrique, B.M., Sobreiro, V.A., Kimura, H.: Practical machine learning: Forecasting daily financial markets directions. Expert Systems with Applications 233, 120840 (2023)
20. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation 9(8), 1735–1780 (Nov 1997), introduces LSTM with constant-error carousel (CEC) units; empirical evaluation mentions cryptocurrency time-series like Bitcoin, Ethereum, Litecoin (see metrics on IEEE Xplore)
21. Hyndman, R.J., Athanasopoulos, G.: Forecasting: principles and practice. OTexts (2018)
22. Lima, A.L.: Bitcoin price prediction using recurrent neural networks and lstm. Project report (2021)
23. Lin, Q.: Price prediction of bitcoin using lstm neural network. BCP Business & Management EMFRM 38(4183), 2738–2744 (2023)
24. Murúa, C., Marín, M., Cofré, A., Wuth, J., Pino, O.V., Yoma, N.B.: An end-to-end dnn-hmm based system with duration modeling for robust earthquake detection. Computers and Geosciences 179, 105434 (2023)
25. Nakamoto, S.: Bitcoin: A peer-to-peer electronic cash system (2008), https://bitcoin.org/bitcoin.pdf, white paper
26. Neagoe, A., Tică, E.I., Vuță, L.I., Nedelcu, O., Dumitran, G.E., Popa, B.: Hybrid lstm-arima model for improving multi-step inflow forecasting in a reservoir. Water 17(21) (2025)
27. Olorunnimbe, K., Viktor, H.: Deep learning in the stock market—a systematic survey of practice, backtesting, and applications. Artificial Intelligence Review 56(3), 2057–2109 (2023)
28. Park, E.: Principal component analysis and hidden markov model for forecasting stock returns. arXiv preprint arXiv:2307.00459 (2023)
29. Petrozziello, A., Troiano, L., Serra, A., Jordanov, I., Storti, G., Tagliaferri, R., La Rocca, M.: Deep learning for volatility forecasting in asset management. Soft Computing 26(17), 8553–8574 (2022)
30. Rizvi, M.F., Sahu, S., Rana, S.: Arima model time series forecasting. International Journal for Research in Applied Science and Engineering Technology (IJRASET) 12(5), 3782–3785 (2024), https://www.ijraset.com/best-journal/arima-model-time-series-forecasting
31. Ryou, H., Bae, H.H., Lee, H.S., Oh, K.J.: Momentum investment strategy using a hidden markov model. Sustainability (2020), publisher Copyright: © 2020 by the authors
32. Shah, J., Vaidya, D., Shah, M.: A comprehensive review on multiple hybrid deep learning approaches for stock prediction. Intelligent Systems with Applications 16, 200111 (2022)
33. Tandon, S., Tripathi, S., Saraswat, P., Dabas, C.: Bitcoin price forecasting using lstm and 10-fold cross validation. In: 2019 International Conference on Signal Processing and Communication (ICSC). pp. 323–328 (2019)
34. Tran, T.K., Le, T.T.T., Bui, T.T., Dang, V.Q., Šenkeřík, R.: Constructing a cryptocurrency-price prediction model using deep learning. In: Proceedings of the 8th International Conference on Engineering and Emerging Technologies (ICEET 2022). pp. 1–6. IEEE, Kuala Lumpur, Malaysia (2022), presents a deep learning-based approach for cryptocurrency price prediction; evaluated on multiple crypto datasets
35. Vancsura, L., Tatay, T., Bareith, T.: Navigating ai-driven financial forecasting: A systematic review of current status and critical research gaps. Forecasting 7(3) (2025)
36. Wang, Y.: Analysis of feature importance based on random forest for stock selection. In: Proceedings of the 1st International Conference on E-commerce and Artificial Intelligence (ECAI 2024). pp. 374–379. SCITEPRESS, London, U.K. (2024)
37. Zakamulin, V.: The real-life performance of market timing with moving average and time-series momentum rules. Journal of Asset Management 15(4), 261–278 (2014)
38. Zakamulin, V., Giner, J.: Optimal trend-following with transaction costs. International Review of Financial Analysis 90, 102928 (2023)
39. Zhang, Y., Wen, J.: The iot electric business model: Using blockchain technology for the internet of things. Peer-to-Peer Networking and Applications 10(4), 983–994 (2017)

40. Zhu, Z., Wang, S., Wang, F., Tu, Z.: Recommendation networks of homogeneous products on an e-commerce platform: Measurement and competition effects. Expert Systems with Applications 201, 117128 (2022)
41. Zucchini, W., MacDonald, I.L.: Hidden Markov models for time series: an introduction using R. Chapman and Hall/CRC (2009)

Tran Kim Toai earned a PhD from the Department of Computer Science at VSB– Technical University of Ostrava, Czech Republic. He works as a lecturer at the Ho Chi Minh City University of Technology and Engineering (HCM-UTE). His current academic research interests are emerging technologies such as machine learning, artificial intelligence, data mining, neural networks, data analysis, ESG, and finance.

Phan Nhat Minh is a researcher working in the fields of data science and artificial intelligence. His research fo cuses on applying data-driven and intelligent methods to real world problems, particularly in prediction, pattern recognition, and decision-making. His current research interests include ma chine learning, deep learning, time-series forecasting, data min ing, and financial data analysis.

Received: November 16, 2025; Accepted: August 23, 2026.

Computer Science and Information Systems 23(4):1219–1238 https://doi.org/10.2298/CSIS260425038X

Paper Type: Research Paper

Hybrid Path Planning for Industrial Autonomous Vehicles: Integrating Neural Heuristics with Energy-Constrained Search

Lijun Xiao[1], Xinwu Jian[1], Zijian Wang[1], Sisi Zhou[2], and Kuanching Li[1]

[1] School of Computer Science and Engineering, Hunan University of Science and Technology
Xiangtan 411201, China
ljxiao@hnust.edu.cn, 1435115512@qq.com, 19555725865@163.com,
aliric@hnust.edu.cn (corresponding author)
[2] School of Computer Science and Engineering, Hunan Women's University
Changsha 410004, China
sisizhou@mail.hnust.edu.cn (corresponding author)

Abstract. Energy-aware path planning for industrial autonomous vehicles requires more than a geometric shortest path: terrain, payload, drivetrain efficiency, turning, and a hard task-level energy budget jointly affect feasibility. We present Neuro-PathNet-EC, a physics-aware constrained planner with learned multimodal guidance. The planner operates on heading-aware grid states and evaluates directed edge energy from friction and elevation. Exact Pareto labels enforce the hard energy budget, while a goal-conditioned ResNet-18 network fuses four raster channels with vehicle semantics and predicts an eight-channel dense cost-to-go map. The learned heuristic is used only for guidance; it is not assumed admissible or consistent. Exact Energy-A* obtains its guarantee from an analytic admissible anchor, and a FOCAL extension provides the bound $C \leq wC^*$. On a layout-disjoint synthetic industrial dataset, five training seeds achieve a test relative heuristic MAE of 0.0289 ± 0.0014 and Spearman correlation of 0.9963 ± 0.0004. Across 512 paired binding-budget queries per seed (2,560 seed–query runs), full multimodal neural-constrained search returns the exact constrained objective in every run while reducing expanded labels by 10.73% relative to Energy-A*. The semantic branch also reduces labels relative to an otherwise matched image-only model. FOCAL satisfies its declared bound in all evaluated cases, although its current correctness-first OPEN management does not improve runtime. The evaluation is based on synthetic layouts and does not constitute public-benchmark or real-vehicle validation.

Keywords: Industrial Autonomous Vehicles, Energy-Constrained Path Planning, Neural Heuristic, Resource-Constrained Search, Multimodal Fusion, Bounded-Suboptimal Search.

1. Introduction

Industrial autonomous vehicles, including automated guided vehicles and unmanned transport platforms, repeatedly navigate warehouses, workshops, and structured outdoor facilities. Recent reviews show that graph search remains central to mobile-robot navigation

even as sampling, optimization, and learning-based planners diversify [1,2,3,4,5,6]. Classical graph search provides transparent and reliable path construction when the state and edge cost are well specified [7,8]. In industrial operation, however, distance alone is insufficient. Payload, rolling resistance, elevation change, drivetrain efficiency, and steering can materially change the energy required by a route. Moreover, a route with a favorable weighted objective may still be unacceptable if it exceeds the energy assigned to the current task.

Learning-based planners can extract guidance from raster maps, graphs, and operating context [9,10,11,12,13]. Yet a learned heuristic creates a theoretical difficulty: low regression error does not imply admissibility or consistency. Treating such a heuristic as if it preserved A* optimality confounds empirical guidance quality with a search guarantee. Hard energy budgets create a second difficulty. The problem is resource-constrained, so keeping only one spatial label can discard a higher-objective but lower-energy prefix that is necessary for a feasible optimum.

We therefore position Neuro-PathNet-EC as *physics-aware energy-constrained search with learned multimodal guidance*. The physical model and hard budget define feasibility and path cost. Search retains nondominated objective–energy labels at each heading-aware state. A dense goal-conditioned network predicts unconstrained cost-to-go for all cells and incoming headings in one pass, but its output is not granted theoretical admissibility. Guarantees instead come from an analytic anchor used by exact Energy-A* or by a bounded FOCAL rule.

The contributions are as follows.

- We formulate directed, heading-aware energy-constrained planning with elevation-derived slope, nonnegative traction energy, turning energy, a dimensionless distance–energy objective, and Pareto labels that enforce a hard task-level budget.
- We design a goal-conditioned multimodal network that maps a four-channel numerical raster and physically relevant vehicle semantics to a dense tensor $\hat{H} \in \mathbb{R}^{8\times H\times W}$. Exact directed reverse shortest paths provide the supervised target.
- We separate empirical neural guidance from guarantees: Energy-A* is exact under an analytic anchor, Safe-Neural is an exact tie-break variant, and FOCAL-Neural is a w-bounded extension whose bound does not depend on learned admissibility.
- We provide a reproducible five-seed evaluation on layout-disjoint synthetic industrial maps. The evaluation includes exact-solver cross-checks, ablation, paired statistics, physical response, binding budgets, sensitivity, robustness stress tests, and censored scalability measurements.

The remainder of the paper reviews related work, formulates the problem, presents the method and theoretical properties, describes the experimental protocol, discusses results and limitations, and concludes with evidence-supported future work.

2. Related Work

2.1. Classical and resource-constrained search

Dijkstra and A* are foundational shortest-path methods [7,8]. Weighted, anytime, and multi-heuristic extensions trade search effort against bounded solution quality [14,15,16].

Their behavior depends on the state representation, edge objective, and heuristic. Recent mobile-robot surveys and an RRT-focused review provide broader comparisons with sampling, evolutionary, and reactive planners [1,3,4,17].

A hard consumable-resource constraint changes the search problem fundamentally. Modern exact algorithms use Pareto labeling and indexing [18,19]. Other accelerations include bidirectional dynamic programming, backtracking, and decomposition [20,21,22]. Related formulations address stochastic constraints, constrained tours, and time resources [23,24,25]. These results support retaining multiple nondominated labels: under an energy budget, a higher-objective but lower-energy prefix can be essential to the feasible optimum. Our formulation specializes this principle to a heading-aware physical grid and never treats the budget as a post hoc filter.

2.2. Learned planning guidance

Neural A* differentiates through grid search to learn spatial guidance maps [9], while GraphMP and PHIL learn graph-structured search guidance [10,11]. Recent work further studies local learned heuristics and losses that optimize search ranking rather than only cost-to-go regression [12,13]. Multimodal perception-aware path planning also demonstrates the value of combining heterogeneous information rather than relying on a single raster modality [26]. Our encoder uses residual spatial features [27], but differs from these studies by predicting all eight incoming-heading cost-to-go maps and coupling them to resource-constrained labels.

Such guidance may reduce exploration, but prediction accuracy alone does not establish A* admissibility. Learned policies and learned heuristics have therefore been combined with FOCAL or bounded-suboptimal search [28,29]. Contemporary bounded-search studies analyze alternative queue rules, error handling, bidirectional bounds, and tunable suboptimality [30,31,32,33,34]. In this work, the learned map is evaluated empirically through absolute error, rank correlation, admissible fraction, directed consistency, and expansion reduction. Optimal or bounded-suboptimal claims derive only from the anchor search rules.

2.3. Energy-aware industrial navigation

Recent AGV research models transport energy directly and constructs route-efficiency benchmarks [35,36]. Related formulations jointly consider productivity, conflict avoidance, multiple loads, or heterogeneous fleets [37,38,39]. Energy-aware A*-type planning has also been developed for Ackermann-steered vehicles, while industrial studies combine global and local planning or optimize pickup-and-delivery routes [40,41,42,43]. These works establish that industrial route energy depends on vehicle state and route geometry, but they do not provide the same dense heading-conditioned learned guidance over a hard Pareto-label budget.

Terrain-aware planning further motivates our directed elevation representation. Recent studies consider resource-aware agricultural autonomy, complex-terrain optimization, and sandy-terrain traversability [44,45,46]. Others address mountainous planning and control, three-dimensional ground-vehicle energy, and large-scale rough terrain [47,48,49]. In parallel, multi-objective planners balance distance, safety, smoothness, or coverage

using graph search and population-based optimization [50,51,52,53,54]. Our distinction is the combination of (i) directed elevation-based energy, (ii) a hard task budget, (iii) heading-aware Pareto labels, (iv) dense multimodal guidance, and (v) an anchor/FOCAL framework that states precisely which component carries each guarantee. We do not claim quantitative comparison with external learned, evolutionary, or sampling-based planners that were not reproduced in the present evaluation.

3. Problem Formulation

3.1. Heading-aware state graph

Let an $H \times W$ grid contain binary occupancy O, friction μ, and elevation z. Motion uses an eight-connected neighborhood. A search state is

$$s = (y, x, q), \qquad q \in \{0, \dots, 7\}, \tag{1}$$

where q is the incoming heading. A sample specifies a start state (y_s, x_s, q_s) and a goal cell $G = (y_g, x_g)$; goal orientation is unconstrained. Diagonal corner cutting through occupied cells is forbidden.

For a forward transition $i \to j$, horizontal displacement is d_{xy} and

$$\Delta z_{ij} = z_j - z_i, \qquad \phi_{ij} = \operatorname{atan2}(\Delta z_{ij}, d_{xy}). \tag{2}$$

Consequently, the grade is directed: $\phi_{ji} = -\phi_{ij}$.

3.2. Physical edge energy

Let m_v and m_p denote vehicle and payload mass, $m = m_v + m_p$, η drivetrain efficiency, v speed, a acceleration, and d edge distance. The force components are

$$F_{\text{roll}} = \mu_{ij} m g \cos \phi_{ij}, \tag{3}$$

$$F_{\text{grade}} = m g \sin \phi_{ij}, \tag{4}$$

$$F_{\text{acc}} = m_{\text{eff}} a, \tag{5}$$

$$F_{\text{aero}} = \tfrac{1}{2} \rho C_d A v^2. \tag{6}$$

The physical work and nonnegative traction energy are

$$E_{\text{phys}} = \frac{d}{\eta}(F_{\text{roll}} + F_{\text{grade}} + F_{\text{acc}} + F_{\text{aero}}), \tag{7}$$

$$E_{\text{traction}} = \max(0, E_{\text{phys}}). \tag{8}$$

No regenerative braking is modeled; a downhill edge therefore receives no negative energy credit. Aerodynamic drag remains in the formulation but is disabled in the low-speed experiments. If q_i and q_j are the incoming and outgoing headings, the wrapped angle satisfies $0 \le |\Delta\psi| \le \pi$ and

$$E_{\text{turn}} = k_{\text{turn}} |\Delta\psi|, \qquad \Delta E_{ij} = E_{\text{traction}} + E_{\text{turn}}. \tag{9}$$

Here k_{turn} has units J/rad. Its experimental value is a nominal synthetic coefficient, not a hardware-calibrated parameter.

3.3. Objective and hard budget

Adding metres to joules is dimensionally invalid. We therefore define

$$c_{ij} = \alpha \frac{d_{ij}}{d_{\text{ref}}} + \beta \frac{\Delta E_{ij}}{E_{\text{ref}}}, \quad \alpha, \beta \geq 0, \quad \alpha + \beta = 1. \tag{10}$$

For path P, $C(P) = \sum_{(i,j)\in P} c_{ij}$ and $E(P) = \sum \Delta E_{ij}$. The planning problem is

$$P^* = \arg\min_P C(P) \quad \text{s.t. } P : s_s \rightsquigarrow G,\ E(P) \leq E_{\text{task,max}}. \tag{11}$$

The task budget is distinguished from total battery availability. If nominal battery energy is given in Wh, physical availability may be written as $3600 E_{\text{nominal}}\,\text{SOC}\,\text{SOH}\,r_{\text{reserve}}$ joules, while $E_{\text{task,max}}$ allocates no more than this amount to the current route.

3.4. Pareto labels

A label is

$$\ell = (y, x, q, g, E, \text{parent}), \tag{12}$$

where g is accumulated objective and E is accumulated energy. A successor with $E > E_{\text{task,max}}$ is rejected. For labels at the same (y, x, q), ℓ_a dominates ℓ_b iff

$$g_a \leq g_b, \quad E_a \leq E_b, \quad \text{and at least one inequality is strict.} \tag{13}$$

Dominance is never applied across headings because future turning cost differs.

4. Neuro-PathNet-EC

4.1. Goal-conditioned multimodal input

The spatial input has four numerical channels: binary occupancy, friction, elevation, and a Gaussian goal map. Occupancy and goal encoding remain in $[0, 1]$; friction and elevation are standardized using training-split statistics only. The semantic vector is

$$\mathbf{s}_{\text{sem}} = [m_v, m_p, \eta, v]. \tag{14}$$

SOC and SOH do not enter this unconstrained cost-to-go network because they do not directly change mechanical edge energy in the adopted model; they determine resource availability. This separation avoids assigning unsupported physical influence to semantic variables.

4.2. Dense eight-heading architecture

A ResNet-18-style encoder receives $\mathbf{X} \in \mathbb{R}^{B\times 4\times 128\times 128}$. Its first convolution has four input channels and weights are initialized randomly because the raster is not natural RGB imagery. The bottleneck is $\mathbf{F} \in \mathbb{R}^{B\times 512\times 4\times 4}$. A semantic MLP $4 \to 128 \to 128 \to 64$ produces $\mathbf{e}_{\text{sem}}$. Spatial broadcasting and concatenation yield $B \times 576 \times 4 \times 4$, followed by a 1×1 projection to 512 channels.

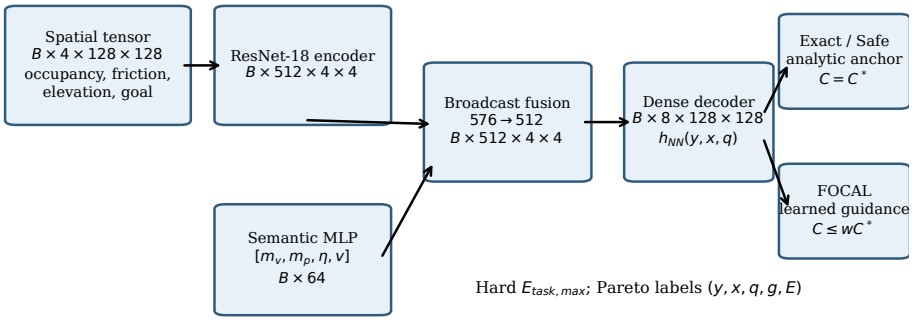


Fig. 1. The proposed Neuro-PathNet-EC framework. The neural network predicts a dense heading-conditioned cost-to-go map. Exact or bounded properties arise from the analytic anchor and label-search rule, not from assuming that h_{NN} is admissible.

No global average pooling is used. A bilinear upsampling decoder with convolutional widths $512 \to 256 \to 128 \to 64 \to 32 \to 8$ restores the grid resolution. Softplus produces nonnegative normalized predictions

$$\hat{H}_{\text{norm}} \in \mathbb{R}^{B\times 8\times 128\times 128}. \tag{15}$$

The planner uses $\hat{H} = H_{\text{scale}}\hat{H}_{\text{norm}}$ and imposes the known boundary $\hat{H}[q, y_g, x_g] = 0$ for every heading.

4.3. Exact dense supervision

The target

$$H^*[q, y, x] = \min_{P:(y,x,q)\rightsquigarrow G} C(P) \tag{16}$$

is the exact unconstrained objective-to-go; hard resource feasibility remains the planner's responsibility. Directed reverse Dijkstra starts from all eight goal headings. When reverse search is at current state v and enumerates predecessor u, it relaxes with the *forward* edge cost $c(u, v)$. Random forward exact solves verified this construction.

Only finite reachable free states contribute to the masked SmoothL1 loss:

$$\mathcal{L} = \frac{1}{|\mathcal{V}|} \sum_{(q,y,x)\in\mathcal{V}} \text{SmoothL1}\left(\hat{H}_{\text{norm}}, H^*/H_{\text{scale}}\right). \tag{17}$$

H_{scale} is the training-split 95th percentile of finite positive H^*.

4.4. Search variants

The analytic anchor is

$$h_{\text{anc}}(s) = \alpha\frac{\|p(s) - G\|_2}{d_{\text{ref}}} + \beta\frac{\|p(s) - G\|_2 e_{\min}}{E_{\text{ref}}}. \tag{18}$$

In our no-regeneration terrain family, a universally safe $e_{\min}$ may be zero; we prefer a weak valid bound to an unsupported positive one.

All constrained variants retain identical feasibility and Pareto dominance. They differ only in queue ordering:

- **Energy-A*:** priority $g + h_{\text{anc}}$; exact constrained baseline.
- **Full Neural-Constrained:** priority $g + h_{NN}$; learned, potentially inadmissible guidance with stale-label handling. Exactness is evaluated empirically, not claimed theoretically.
- **Safe-Neural:** primary key $g + h_{\text{anc}}$, secondary key h_{NN}; exact, but neural influence occurs only at anchor ties.
- **FOCAL-Neural:** let $f_{\min} = \min_{\ell \in OPEN}(g + h_{\text{anc}})$. Choose from $\{\ell : f_{\text{anc}}(\ell) \leq w f_{\min}\}$ by (h_{NN}, E, tie).

5. Theoretical Properties

5.1. What the learned heuristic does not guarantee

Strong MAE or correlation does not imply $h_{NN}(s) \leq H^*(s)$ or $h_{NN}(s) \leq c(s, s') + h_{NN}(s')$. We measure both properties, but do not assume them. The boundary $h(G) = 0$ is known exactly and is enforced without using unknown path cost.

5.2. Anchor validity under a resource constraint

Let $\mathcal{P}_{\text{con}}(s)$ be feasible continuations satisfying the remaining budget and $\mathcal{P}_{\text{unc}}(s)$ all obstacle-feasible continuations. Since $\mathcal{P}_{\text{con}}(s) \subseteq \mathcal{P}_{\text{unc}}(s)$, the optimal unconstrained continuation cost is no greater than the constrained continuation cost. Therefore any lower bound on the unconstrained cost, including Eq. (18), is also a lower bound for the constrained problem. A zero turning lower bound is valid because turn energy is nonnegative.

5.3. FOCAL bounded-suboptimality

Proposition 1. Consider the finite heading-grid feasible label graph generated under a hard energy budget. Assume nonnegative edge objective, safe Pareto dominance within identical heading states, an admissible anchor, correct OPEN maintenance, and termination when a selected FOCAL label is a goal. Then FOCAL-Neural with $w \geq 1$ returns cost $C_{\text{ret}} \leq wC^*$ whenever a feasible solution exists.

Proof. Before an optimal goal label is selected, OPEN contains a surviving prefix label on an optimal feasible path. Admissibility implies its anchor key is at most C^*, hence $f_{\min} \leq C^*$. Every selected FOCAL label satisfies $f_{\text{anc}} \leq w f_{\min}$. At a selected goal, $h_{\text{anc}} = 0$, so $C_{\text{ret}} = g \leq w f_{\min} \leq wC^*$. Safe Pareto dominance cannot remove every optimal continuation because a dominating label has no greater objective and energy at the same heading state. The finite grid and finite set of generated simple nondominated path labels under positive movement cost ensure termination. □

The learned score affects which eligible FOCAL label is chosen, but not the bound. For $w = 1$, FOCAL retains the optimal anchor threshold, with learned ordering only among equal-minimum anchor labels.

Table 1. Frozen implementation and training settings.

Item	Setting
Hardware	NVIDIA L4, 22.03 GiB VRAM
Software	Python 3.12.13; PyTorch 2.11.0+cu128
CUDA / cuDNN	12.8 / 9.1.9
Input / output	$4 \times 128 \times 128$ / $8 \times 128 \times 128$
Full parameters	13,070,984
Optimizer	AdamW, lr 3×10^{-4}, wd 10^{-2}
Scheduler	ReduceLROnPlateau, factor 0.5, patience 3
Batch / epochs	16 / maximum 50
Early stopping	patience 8 on validation loss
Seeds	0, 1, 2, 3, 4
Split samples	2,048 / 256 / 512
Split layouts	128 / 16 / 32, disjoint
α, β	0.3, 0.7
E_{ref}	543.4587 J, training edge median
H_{scale}	128.8968, training H^* p95
k_{turn}	25 J/rad, nominal synthetic

6. Experimental Setup

6.1. Implementation details

All neural training and dense inference use the GPU. Search is an ordinary CPU implementation; no CUDA search kernel or GPU-parallel expansion is used. Checkpoints are selected only by minimum validation normalized SmoothL1.

6.2. Synthetic industrial data

The generator produces structured warehouse-like rack layouts, factory-like aisles, and industrial-park-like open/structured layouts. Each numerical sample stores occupancy, spatially structured friction, continuous elevation, start, goal, start heading, vehicle state, layout metadata, and exact dense H^*. Layout identities are disjoint across train, validation, and test. Continuous normalization uses training statistics only.

The binding benchmark contains 512 three-route queries (Fig. 2). Route A is short and energy intensive, Route B is balanced, and Route C is longer and lower energy. Every accepted query verifies $d_A < d_B < d_C$, $E_A > E_B > E_C$, and $E_{\mathrm{obj}} > E_{\mathrm{min}}$. For tightness ρ,

$$E_{\mathrm{max}}(\rho) = E_{\mathrm{min}} + \rho(E_{\mathrm{obj}} - E_{\mathrm{min}}), \tag{19}$$

with $\rho \in \{0.05, 0.25, 0.50, 0.75, 0.95\}$.

6.3. Baselines, variants, and metrics

We evaluate Dijkstra-style exact label setting, Classical A* for geometric context, Energy-A*, Image-only and Full Neural-Constrained search, Safe-Neural, FOCAL with $w = 1.05$

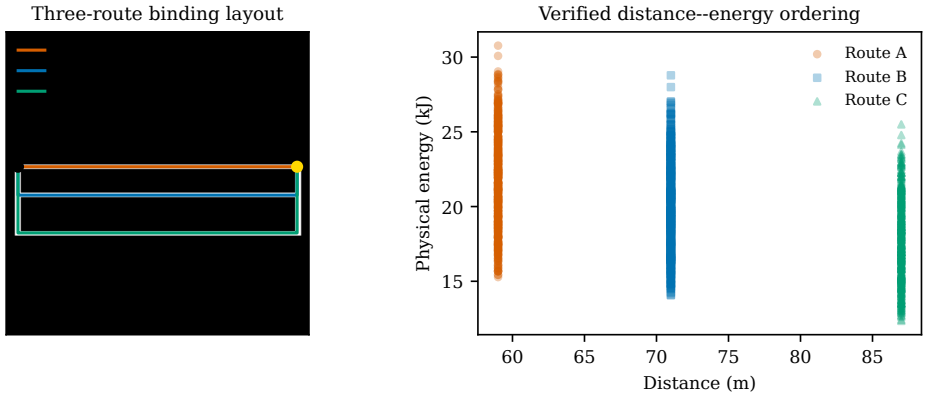


Fig. 2. Binding-budget benchmark. Left: a representative three-route layout. Right: all 512 generated queries satisfy the intended distance–energy ordering. These are synthetic stress-test maps, not real-site telemetry.

Table 2. Main comparison on 512 paired binding queries. Values are mean ± standard deviation of five seed-level means. Objective and energy are deterministic for exact methods.

Method	Objective	Energy (J)	Labels	Runtime (ms)	Property
Dijkstra	48.4649	18171.3	412.50	6.76 ± 0.95	Exact
Classical A*	59.0000	21815.5	60.00	0.52 ± 0.07	Infeasible
Energy-A*	48.4649	18171.3	406.53	6.94 ± 0.91	Exact anchor
Image-only Neural	48.4649	18171.3	383.69 ± 22.16	6.36 ± 0.94	Empirical
Full Multimodal	**48.4649**	**18171.3**	$\mathbf{362.92 \pm 21.09}$	$\mathbf{6.14 \pm 0.99}$	Empirical
Safe-Neural	48.4649	18171.3	406.53	7.06 ± 1.05	Exact tie-break
FOCAL $w = 1.05$	48.5597 ± 0.0650	18061.5 ± 71.2	406.74 ± 4.29	22.07 ± 3.00	$C \leq 1.05C^*$
FOCAL $w = 1.10$	48.5149 ± 0.0504	18127.2 ± 37.5	403.90 ± 6.02	23.00 ± 3.27	$C \leq 1.10C^*$
Oracle-H^*	48.4649	18171.3	254.28	4.45 ± 0.61	Diagnostic

and 1.10, and diagnostic Oracle-H^* and spatially shuffled-H controls. Classical A* does not enforce the energy budget and is not treated as a feasible constrained competitor.

Heuristic metrics are MAE, RMSE, relative MAE, Pearson, Spearman, R^2, empirical admissible fraction, and directed consistency violation. Planner metrics are objective, energy, path length, expanded labels, runtime, success, and gap to exact constrained search. Statistical comparisons use the paired query as the unit: each method is first averaged across five seeds per query, followed by two-sided Wilcoxon signed-rank, paired Cohen's d, 10,000-sample query bootstrap confidence intervals, and Holm correction.

7. Results and Discussion

7.1. Main comparison

Full Multimodal Neural-Constrained search returns the exact constrained objective and energy in all 2,560 seed–query runs. Relative to Energy-A*, it reduces labels by 10.73%

Table 3. Ablation of guidance mechanisms.

Variant	Semantic	Neural	FOCAL	Labels
Energy-A*	–	–	–	406.53
Image-only	–	✓	–	383.69 ± 22.16
Full	✓	✓	–	362.92 ± 21.09
Safe-Neural	✓	tie	–	406.53
FOCAL 1.05	✓	✓	✓	406.74 ± 4.29
FOCAL 1.10	✓	✓	✓	403.90 ± 6.02

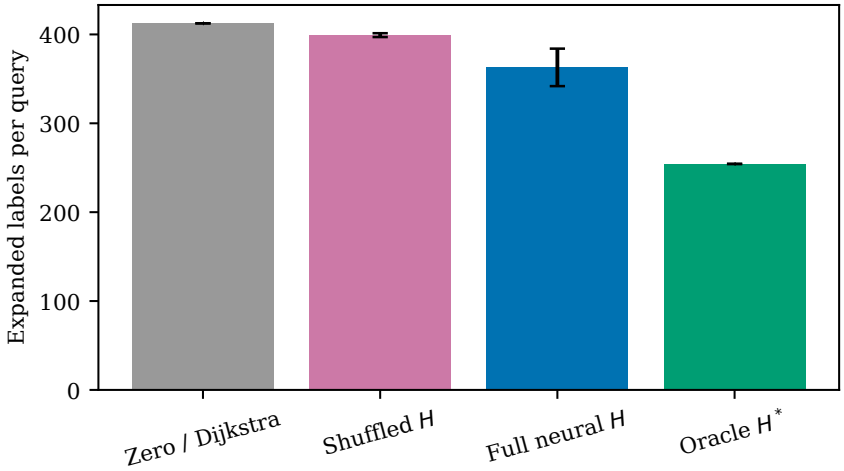


Fig. 3. Guidance diagnostic. Spatially meaningful multimodal predictions reduce expansions relative to zero and shuffled controls; Oracle-H^* is a diagnostic upper reference.

and mean runtime by 0.80 ms. This exact agreement is an experimental finding, not a consequence of learned admissibility. Classical A* finds the shortest geometric corridor but violates the binding budget in all 512 queries.

7.2. Ablation and guidance contribution

The semantic branch reduces labels by a further 5.41% relative to Image-only. Safe-Neural matches Energy-A* because learned guidance acts only at exact anchor ties. Oracle guidance shows that additional expansion reduction remains possible, while shuffling destroys most of the learned benefit (Fig. 3). The empirical guidance recovery ratio is 0.313 relative to Zero/Dijkstra and Oracle endpoints.

7.3. Heuristic quality

The dense map has low relative error and strong spatial ordering (Fig. 5). Nevertheless, only about 47% of predictions are empirically admissible and about 14.7% of directed

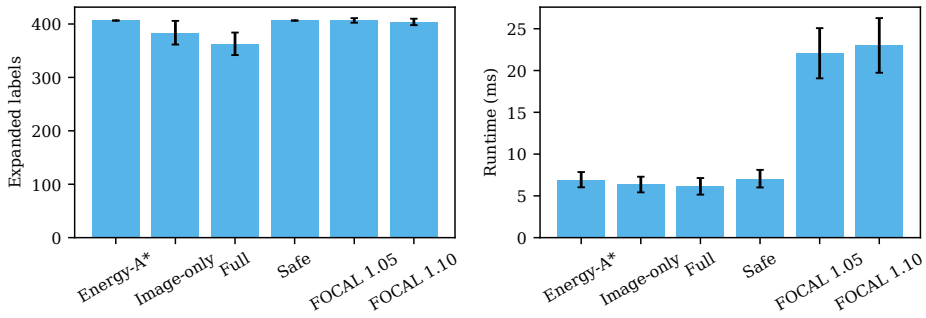


Fig. 4. Five-seed ablation. Full multimodal guidance reduces labels and modestly reduces runtime, whereas correctness-first FOCAL incurs substantial OPEN-management overhead.

Table 4. Full-model heuristic properties over five seeds.

Metric	Mean ± SD
MAE	1.9370 ± 0.0917
RMSE	3.1145 ± 0.1716
Relative MAE	0.02888 ± 0.00137
Pearson	0.99592 ± 0.00044
Spearman	0.99628 ± 0.00042
R^2	0.99182 ± 0.00090
Admissible fraction	0.47034 ± 0.02663
Consistency violation rate	0.14681 ± 0.00829

edge tests violate consistency. These measurements directly motivate the analytic-anchor separation.

7.4. Physical semantic response

At fixed map, start, goal, and heading, increasing payload from 0 to 800 kg raises exact start cost-to-go from 82.85 to 115.49. Full prediction follows from 82.48 to 114.13, while Image-only remains 94.33 because it cannot observe payload. Similarly, increasing η from 0.75 to 0.90 lowers exact cost from 106.25 to 94.19 and Full prediction from 105.86 to 93.03; Image-only again remains unchanged. This controlled response provides mechanism-level evidence for the semantic branch.

7.5. Statistical analysis

The label reduction is statistically and practically meaningful. Conversely, no objective or energy difference exists between Full and Energy-A*, and no path-quality improvement is claimed. FOCAL is significantly slower in this implementation.

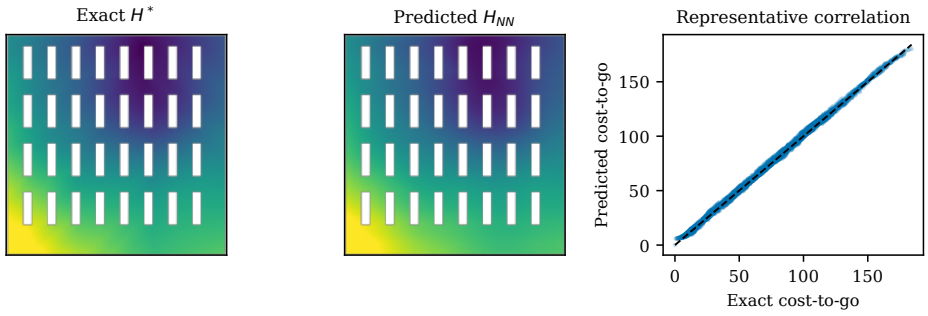


Fig. 5. Representative exact and predicted cost-to-go maps for the sample's start heading, with a subsampled cell-wise correlation plot. Aggregate five-seed statistics are reported in Table 4.

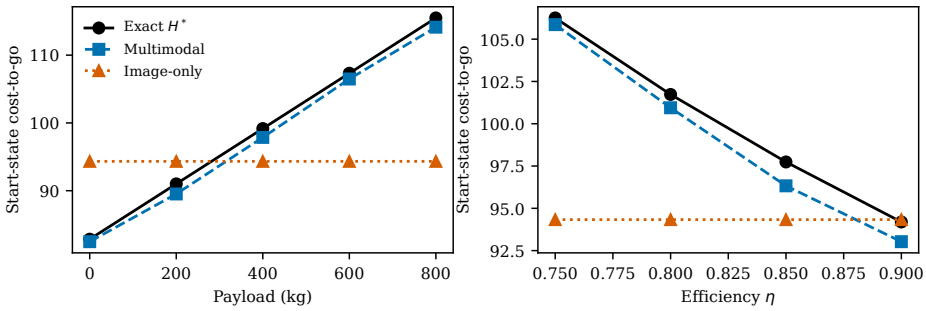


Fig. 6. Controlled payload and drivetrain-efficiency response. Multimodal predictions follow the exact physical target, whereas the matched image-only model cannot respond to hidden vehicle semantics.

7.6. Binding budgets and FOCAL bounds

For each ρ, Energy-A* and exact label setting agree in every one of 200 checked queries. Binding changes the path in 100% of cases. Mean energy reduction relative to the unconstrained objective optimum decreases from 4010.7 J at $\rho = 0.05$ to 1694.5 J at $\rho = 0.95$. Across 5,000 FOCAL cases (200 queries, five budgets, five w values), the bound violation count is zero. Maximum gaps for $w = 1.02, 1.05, 1.10, 1.20$ are 0.00246, 0.04855, 0.04453, and 0.06367. Runtime, however, increases with w because the current implementation reconstructs the eligible FOCAL subset by scanning OPEN.

7.7. Parameter sensitivity

At nominal $\alpha = 0.3$, mean normalized distance and energy contributions are 24.804 and 23.778, respectively; $\beta = 0.7$ therefore does not make energy dominate by orders of magnitude. Under the fixed tight budget, decreasing efficiency by 20% or increasing friction by 20% makes all 100 cases infeasible. Increasing payload by 20% changes 26% of paths

Table 5. Selected paired statistical comparisons over 512 query units after averaging training seeds. CI is a query-bootstrap 95% interval; p_{Holm} is Holm-adjusted.

Method A	Method B	Metric	Mean diff.	95% CI	p_{Holm}	Paired d
Full	Energy-A*	Objective	0	$[0, 0]$	1.0	0
Full	Energy-A*	Labels	-43.618	$[-48.184, -39.178]$	2.15×10^{-84}	-0.833
Full	Energy-A*	Runtime (ms)	-0.799	$[-0.975, -0.619]$	4.25×10^{-49}	-0.390
Full	Image-only	Labels	-20.773	$[-22.491, -19.104]$	2.15×10^{-84}	-1.068
FOCAL 1.05	Energy-A*	Runtime (ms)	15.129	$[14.952, 15.316]$	2.15×10^{-84}	7.250

Table 6. Objective-weight sensitivity on 100 paired queries.

α	β	Objective	Energy (J)	Length (m)
0.1	0.9	38.4839	17984.8	87.00
0.3	0.7	48.5823	18460.8	82.68
0.5	0.5	58.3245	18460.8	82.68
0.7	0.3	68.0667	18460.8	82.68
0.9	0.1	77.8089	18460.8	82.68

and makes 26% of nominal paths exceed the budget; replanning returns no violating successful path. The nominal turn coefficient has a smaller effect, changing 1–5% of paths under $\pm 20\%$ perturbation.

7.8. Robustness stress tests

These deliberately severe tests expose failure modes rather than establish industrial robustness. A 10% budget overestimate preserves planner success but violates the nominal true budget in 74% of cases, illustrating why uncertainty-aware budgeting is important. Occupancy and localization perturbations are synthetic and are not calibrated sensor-error distributions.

7.9. Scalability

Energy-A* mean runtime is 2.43 s at 128 and 9.97 s at 256. At 512 it completes two of four settings. Higher obstacle density can reduce runtime by shrinking the reachable label set. FOCAL's correctness-first linear OPEN scan is the dominant limitation; these results do not support a large-scale acceleration claim.

8. Limitations

The present evidence has clear boundaries. First, all maps are generated synthetic industrial layouts; there is no Gazebo experiment, public benchmark, industrial telemetry, or real-vehicle deployment. Layout-disjoint testing measures generalization within this generator only. Second, the learned heuristic is neither theoretically admissible nor consistent. Its empirical exact-objective behavior does not replace the analytic-anchor proof. Third,

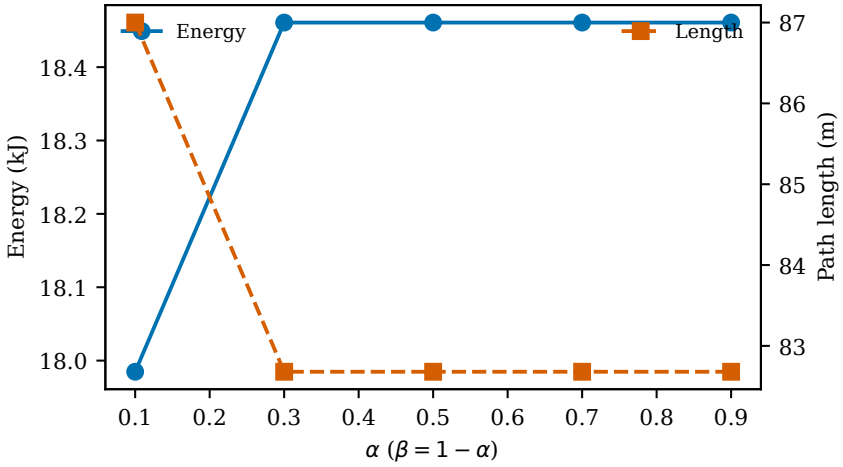


Fig. 7. Energy–distance response to the frozen α grid ($\beta = 1 - \alpha$). The discrete Route-C choices produce plateaus rather than a smooth continuum.

Table 7. Synthetic robustness stress tests (50 seed-0 queries).

Condition	Levels	Success rates
Occupancy flips (%)	0/1/3/5/10	1/.48/.12/.02/0
Localization shift (m)	0/.05/.10/.20	1/1/1/.52
Budget error (%)	-10/0/+10	.48/1/1
SOH (%)	100/90/80/70	1/.48/0/0

k_{turn} is nominal and has not been calibrated against hardware. Fourth, FOCAL uses a correctness-first OPEN scan and is not runtime optimized. Fifth, robustness perturbations are synthetic stress tests rather than calibrated sensing or battery uncertainty. Sixth, scalability uses one generated map per size–density cell and includes censored runs. Seventh, external RRT*, CNN-A*, or GNN-based planners were not reproduced; no external state-of-the-art claim is made. Finally, GPU use is limited to neural training and inference, not parallel search.

9. Conclusion

We presented Neuro-PathNet-EC as physics-aware energy-constrained search with multimodal neural guidance. Directed elevation, friction, payload, efficiency, and turning define a dimensionless objective and hard resource constraint over heading-aware Pareto labels. A goal-conditioned dense network predicts eight heading-specific cost-to-go maps, while exact or bounded properties are supplied separately by analytic-anchor Energy-A*, Safe-Neural, and FOCAL.

Five-seed experiments show strong dense regression quality and exact constrained objective agreement for Full Neural-Constrained search across 2,560 runs. Full guidance

Table 8. Completion within a 30 s per-method threshold. Four density settings are evaluated at each size.

Grid	Energy-A*	FOCAL 1.05	Peak RSS (GiB)
128	4/4	1/4	1.04
256	4/4	0/4	1.19
512	2/4	0/4	1.39

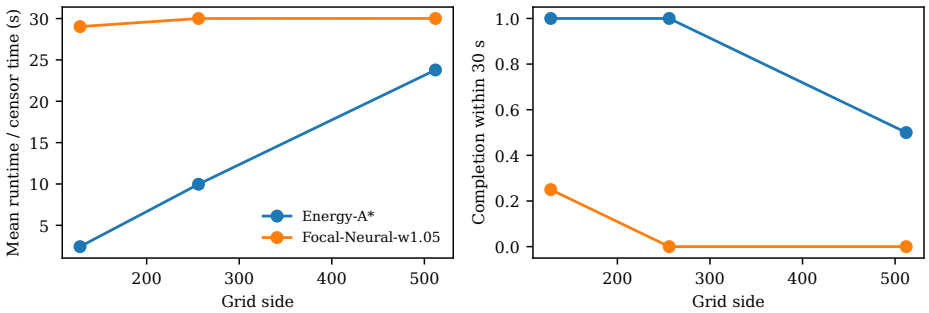


Fig. 8. Censored scalability results. Runtime values include the 30 s censor time. The current FOCAL data structure prioritizes correctness and does not scale competitively.

reduces expanded labels by 10.73% relative to Energy-A* and improves on a matched image-only network; controlled payload and efficiency experiments confirm that the semantic branch responds to physical variables. At the same time, learned admissibility and consistency do not hold, and FOCAL's present data structure is slower and poorly scalable. These negative results are part of the method's current scope.

Future work should validate on public benchmarks and real vehicles, calibrate turning and drivetrain parameters from measurements, replace linear FOCAL scans with efficient OPEN/FOCAL structures, and model uncertainty explicitly. Remaining-budget-aware guidance is also a natural extension because two Pareto labels at the same heading state may have different useful continuations.

Acknowledgment. This work is supported by the Natural Science Foundation of Hunan Province under Grant 2025JJ50399.

References

1. José Ricardo Sánchez-Ibáñez, Carlos J. Pérez-del Pulgar, and Alfonso García-Cerezo. Path planning for autonomous mobile robots: A review. *Sensors*, 21(23):7898, Nov 2021.
2. Shiwei Lin, Ang Liu, Jianguo Wang, and Xiaoying Kong. A review of path-planning approaches for multiple mobile robots. *Machines*, 10(9):773, Sept 2022.
3. Jaafar Ahmed Abdulsaheb and Dheyaa Jasim Kadhim. Classical and heuristic approaches for mobile robot path planning: A survey. *Robotics*, 12(4):93, June 2023.

4. Liwei Yang, Ping Li, Song Qian, He Quan, Jinchao Miao, Mengqi Liu, Yanpei Hu, and Erexidin Memetimin. Path planning technique for mobile robots: A review. *Machines*, 11(10):980, Oct 2023.
5. Kornél Katona, Husam A. Neamah, and Péter Korondi. Obstacle avoidance and path planning methods for autonomous navigation of mobile robot. *Sensors*, 24(11):3573, June 2024.
6. Yuexia Tang, Muhammad Aizzat Zakaria, and Maryam Younas. Path planning trends for autonomous mobile robot navigation: A review. *Sensors*, 25(4):1206, Feb 2025.
7. Edsger W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271, 1959.
8. Peter E. Hart, Nils J. Nilsson, and Bertram Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2):100–107, 1968.
9. Ryo Yonetani, Tatsunori Taniai, Mohammadamin Barekatain, Mai Nishimura, and Asako Kanezaki. Path planning using neural A* search. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12029–12039. PMLR, 2021.
10. Xiao Zang, Miao Yin, Jinqi Xiao, Saman Zonouz, and Bo Yuan. GraphMP: Graph neural network-based motion planning with efficient graph search. In *Advances in Neural Information Processing Systems*, volume 36, pages 3131–3142, 2023.
11. Michal Pándy, Weikang Qiu, Gabriele Corso, Petar Veličković, Zhitao Ying, Jure Leskovec, and Pietro Lio. Learning graph search heuristics. In *Proceedings of the First Learning on Graphs Conference*, volume 198 of *Proceedings of Machine Learning Research*, pages 10:1–10:13. PMLR, 2022.
12. Rishi Veerapaneni, Muhammad Suhail Saleem, and Maxim Likhachev. Learning local heuristics for search-based navigation planning. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 33, pages 406–414, 2023.
13. Leah Chrestien, Stefan Edelkamp, Antonin Komenda, and Tomáš Pevný. Optimize planning heuristics to rank, not to estimate cost-to-goal. *Advances in Neural Information Processing Systems*, 36, 2023.
14. Judea Pearl and Jin H. Kim. Studies in semi-admissible heuristics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-4(4):392–399, 1982.
15. Maxim Likhachev, Geoffrey J. Gordon, and Sebastian Thrun. ARA*: Anytime A* with provable bounds on sub-optimality. In *Advances in Neural Information Processing Systems*, volume 16, 2004.
16. Sandip Aine, Siddharth Swaminathan, Venkatraman Narayanan, Victor Hwang, and Maxim Likhachev. Multi-heuristic A*. *The International Journal of Robotics Research*, 35(1–3):224–243, 2016.
17. Tong Xu. Recent advances in rapidly-exploring random tree: A review. *Heliyon*, 10(11):e32451, June 2024.
18. Saman Ahmadi, Guido Tack, Daniel D. Harabor, and Philip Kilby. A fast exact algorithm for the resource constrained shortest path problem. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12217–12224, May 2021.
19. Libin Wang and Raymond Chi-Wing Wong. Qhl: A fast algorithm for exact constrained shortest path search on road networks. *Proceedings of the ACM on Management of Data*, 1(2):1–25, June 2023.
20. Matteo Salani, Saverio Basso, and Giovanni Righini. Enhanced bi-directional dynamic programming algorithm for the resource constrained shortest path problem. *Transportation Research Procedia*, 78:361–368, 2024.
21. Bryce Ford and Mrinal Kumar. Analysis of backtracking a* for resource constrained shortest path problems. In *2024 American Control Conference (ACC)*, pages 3545–3550. IEEE, July 2024.

22. Xuefeng Jin and Shunzheng Yu. A two-stage method for doubly resource constrained elementary shortest path problems. *Knowledge-Based Systems*, 293:111661, June 2024.
23. Sungkweon Hong and Brian C. Williams. An anytime algorithm for constrained stochastic shortest path problems with deterministic policies. *Artificial Intelligence*, 316:103846, Mar 2023.
24. Sébastien Martin, Youcef Magnouche, Corentin Juvigny, and Jérémie Leguay. Constrained shortest path tour problem: Branch-and-price algorithm. *Computers & Operations Research*, 144:105819, Aug 2022.
25. Pan Liu and Wulan Huang. A graph algorithm for the time constrained shortest path. *Connection Science*, 34(1):1500–1518, May 2022.
26. Zhenyu Li, Aiguo Zhou, Jiakun Pu, and Jiangyang Yu. Multi-modal neural feature fusion for automatic driving through perception-aware path planning. *IEEE Access*, 9:142782–142794, 2021.
27. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
28. Matias Greco. Exploiting learned policies and learned heuristics in bounded-suboptimal search. *Proceedings of the International Symposium on Combinatorial Search*, 12(1):219–221, July 2021.
29. Matias Greco and Jorge A. Baier. Avoiding errors in learned heuristics in bounded-suboptimal search. *Proceedings of the International Symposium on Combinatorial Search*, 15(1):336–338, July 2022.
30. Maximilian Fickert, Tianyi Gu, and Wheeler Ruml. New results in bounded-suboptimal search. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(9):10166–10173, June 2022.
31. Ryo Kuroiwa and J. Christopher Beck. Biased exploration for satisficing heuristic search. *Proceedings of the International Conference on Automated Planning and Scheduling*, 32:213–221, June 2022.
32. Dor Atzmon, Shahaf S. Shperberg, Netanel Sabah, Ariel Felner, and Nathan R. Sturtevant. W-restrained bidirectional bounded-suboptimal heuristic search. *Proceedings of the International Conference on Automated Planning and Scheduling*, 33(1):26–30, July 2023.
33. Stephen Wissow, Fanhao Yu, and Wheeler Ruml. Tunable suboptimal heuristic search. *Proceedings of the International Symposium on Combinatorial Search*, 17:170–178, June 2024.
34. Jaein Lim and Panagiotis Tsiotras. Cbs-budget (cbsb): A complete and bounded suboptimal search for multi-agent path finding. *Artificial Intelligence*, 346:104349, Sept 2025.
35. Zhongwei Zhang, Lihui Wu, Wenqiang Zhang, Tao Peng, and Jun Zheng. Energy-efficient path planning for a single-load automated guided vehicle in a manufacturing workshop. *Computers & Industrial Engineering*, 158:107397, Aug 2021.
36. Luoke Hu, Xiaoliang Zhao, Weipeng Liu, Wei Cai, Kangkang Xu, and Zhongwei Zhang. Energy benchmark for energy-efficient path planning of the automated guided vehicle. *Science of The Total Environment*, 857:159613, Jan 2023.
37. Jianbin Xin, Liuqian Wei, Andrea D'Ariano, Fangfang Zhang, and Rudy Negenborn. Flexible time–space network formulation and hybrid metaheuristic for conflict-free and energy-efficient path planning of automated guided vehicles. *Journal of Cleaner Production*, 398:136472, Apr 2023.
38. Zhongwei Zhang, Lihui Wu, Boqiang Zhang, Shun Jia, Weipeng Liu, and Tao Peng. Energy-efficient path planning for a multi-load automated guided vehicle executing multiple transport tasks in a manufacturing workshop environment. *Environmental Science and Pollution Research*, Mar 2024.
39. Jia Gao, Xiaojun Zheng, Feng Gao, Xiaoying Tong, and Qiaomei Han. Heterogeneous multitype fleet green vehicle path planning of automated guided vehicle with time windows in flexible manufacturing system. *Machines*, 10(3):197, Mar 2022.

40. Haojie Zhang, Yudong Zhang, Chuankai Liu, and Zuoyu Zhang. Energy efficient path planning for autonomous ground vehicles with ackermann steering. *Robotics and Autonomous Systems*, 162:104366, Apr 2023.
41. Te Wang, Aijuan Li, Dongjin Guo, Guangkai Du, and Weikai He. Global dynamic path planning of agv based on fusion of improved a* algorithm and dynamic window method. *Sensors*, 24(6):2011, Mar 2024.
42. Zhengrui Jiang, Wang Chen, Xiaojun Zheng, and Feng Gao. Research on vehicle path planning of automated guided vehicle with simultaneous pickup and delivery with mixed time windows. *IET Collaborative Intelligent Manufacturing*, 6(2), May 2024.
43. Meltem Eyuboglu and Gokhan Atali. A novel collaborative path planning algorithm for 3-wheel omnidirectional autonomous mobile robot. *Robotics and Autonomous Systems*, 169:104527, Nov 2023.
44. Stuart Eiffert, Nathan D. Wallace, He Kong, Navid Pirmarzdashti, and Salah Sukkarieh. Resource and response aware path planning for long-term autonomy of ground robots in agriculture. *Field Robotics*, 2:1–33, Mar 2022.
45. Sunil Kumar and Afzal Sikander. An intelligent optimize path planner for efficient mobile robot path planning in a complex terrain. *Microsystem Technologies*, 29(4):469–487, June 2022.
46. Go Sakayori and Genya Ishigami. Modeling of slip rate-dependent traversability for path planning of wheeled mobile robot in sandy terrain. *Frontiers in Robotics and AI*, 11, Jan 2024.
47. Se-Wook Yoo, E-In Son, and Seung-Woo Seo. Traversability-aware adaptive optimization for path planning and control in mountainous terrain. *IEEE Robotics and Automation Letters*, 9(6):5078–5085, June 2024.
48. Omer Melih Gul. Energy-aware 3d path planning by autonomous ground vehicle in wireless sensor networks. *World Electric Vehicle Journal*, 15(9):383, Aug 2024.
49. Yuxiang Li, Kun Chen, Yifei Wang, Weifan Zhang, Jiancheng Wang, Haoyao Chen, and Yunhui Liu. Real-time multilevel terrain-aware path planning for ground mobile robots in large-scale rough terrains. *IEEE Transactions on Robotics*, 41:4159–4179, 2025.
50. Dan Xiang, Hanxi Lin, Jian Ouyang, and Dan Huang. Combined improved a* and greedy algorithm for path planning of multi-objective mobile robot. *Scientific Reports*, 12(1), Aug 2022.
51. K. S. Suresh, R. Venkatesan, and S. Venugopal. Mobile robot path planning using multi-objective genetic algorithm in industrial automation. *Soft Computing*, 26(15):7387–7400, June 2022.
52. V. Sathiya, M. Chinnadurai, and S. Ramabalan. Mobile robot path planning using fuzzy enhanced improved multi-objective particle swarm optimization (fimopso). *Expert Systems with Applications*, 198:116875, July 2022.
53. Senthil Kumar Angappamudaliar Palanisamy, Dinesh Selvaraj, and SivaBalaKrishnan Ramasamy. Hybrid multi-objective optimization approach intended for mobile robot path planning model. *Journal of Intelligent & Fuzzy Systems*, 42(3):2681–2693, Feb 2022.
54. Monex Sharma and Hari Kumar Voruganti. Multi-objective optimization approach for coverage path planning of mobile robot. *Robotica*, 42(7):2125–2149, Mar 2024.

Lijun Xiao received the Ph.D. degree in mechanical engineering from Hunan University, China. He is currently a Professor with the School of Computer Science and Engineering, Hunan University of Science and Technology. His research interests include artificial intelligence, machine learning, and intelligent unmanned systems, with a current focus on motion planning and cooperative optimization for autonomous vehicles and multi-UAV systems.

Xinwu Jian is currently working toward the M.Eng. degree in software engineering with the School of Computer Science and Engineering, Hunan University of Science and Technology, China. His research interests include sampling-based motion planning, UAV path planning, and reinforcement learning.

Zijian Wang is currently working toward the M.Eng. degree in software engineering with the School of Computer Science and Engineering, Hunan University of Science and Technology, China. His research interests include combinatorial optimization, metaheuristic algorithms, and UAV–UGV cooperation.

Sisi Zhou is currently with Hunan Women's University. She received the B.S., M.S., and Ph.D. degrees in 2009, 2012, and 2026, respectively. Her research interests include privacy protection and information security, with a focus on federated learning and blockchain technology.

Kuanching Li (Senior Member, IEEE) is a Professor at Hunan University of Science and Technology. He received his Ph.D. degree from San Pablo Main in 2001, and has published papers in high-ranked journals and conferences and served as associate and guest editor for several scientific journals. His research interests include cloud and edge computing, big data, and blockchain.

Received: April 25, 2026; Accepted: August 8, 2026.

Computer Science and Information Systems 23(4):1239–1258 https://doi.org/10.2298/CSIS260407047J

Paper Type: Research Paper

Vulnerable Node Detection in Complex Networks Based on Dynamic Comparator and Trusted Chain Propagation

Dongsheng Jing[1], Baozhen Qi[1], Yu Yang[1], Jingsong Xue[1], and Renjun Feng[1]

State Grid Suzhou Power Supply Company
Suzhou Jiangsu 215004, China
qibz@js.sgcc.com.cn (corresponding author)
{jds_sz, yang_yu, js.xue, fengrj }@js.sgcc.com.cn

Abstract. Accurately identifying vulnerable nodes in complex networks is a critical challenge for preventing cascading failures and ensuring system robustness. Existing methods generally suffer from limitations such as static metrics that ignore dynamic loads, high parameter dependency, and a lack of unified fusion mechanisms. To address these issues, we propose an approach that is based on dynamic comparator selection and trusted chain propagation, which we refer to as Optimal Parameter (OP). Specifically, the proposed OP approach adaptively adjusts strategies based on node degree: low-degree nodes enable a lightweight mode to reduce overhead, while high-degree nodes employ full sampling to enhance reliability. State chain propagation is achieved through trusted queues, reducing the global detection complexity to linear order. Furthermore, the method integrates temporal consistency checks, faulty comparator isolation, and cross-layer comparison mechanisms to effectively handle complex scenarios such as malicious camouflage and low connectivity. Experimental results demonstrate that OP achieves a higher true positive rate and a lower false positive rate across multiple real-world datasets. It shows significant advantages, particularly at high attack edge density, thereby verifying its synergistic performance in terms of high accuracy and efficiency in large-scale networks.

Keywords: Complex Networks, Vulnerable Nodes, Dynamic Comparator, Trusted Chain Propagation, Robustness.

1. Introduction

Complex networks are currently popular worldwide. They serve as the backbone for various critical infrastructures, including power grids, global communication architectures, online social platforms, and intricate biological systems [6]. Due to the tight topological coupling and functional interdependencies inherent in these systems, they often exhibit extreme fragility to localized disruptions. Specifically, the failure or compromise of a small fraction of vulnerable nodes can trigger a severe systemic collapse through cascading effects. This phenomenon manifests profoundly across various domains: in social networks [17], malicious accounts or Sybil nodes can rapidly propagate misinformation to undermine community trust; in cyber-physical systems [5] like power grids, the tripping of an overloaded regional substation redistributes immense dynamic loads to adjacent

nodes, precipitating catastrophic large-scale blackouts; similarly, in distributed storage or communication networks [24], the incapacitation of core routing nodes can lead to massive network fragmentation, turning single-point failures into global paralysis. Given the immense scale, dynamic evolution, and nonlinear nature of failure propagation in contemporary networks, preemptively quantifying node vulnerability is extraordinarily complex [4]. How to efficiently and accurately identify such fault-prone nodes—balancing computational overhead with detection precision—remains a formidable and unresolved challenge.

Particularly in the energy sector, the layout of modern power systems has evolved from traditional centralized, top-down radial topologies to highly interconnected, distributed, complex networks [13]. With the massive integration of renewable energy sources, active distribution networks, and microgrids, the new power grid exhibits prominent network-structured characteristics. While this highly coupled network layout enhances energy transmission efficiency and flexibility, it also significantly amplifies the spatial vulnerability to cascading failures [23]. A localized malfunction at a single vulnerable node—such as an overloaded regional substation or a compromised distributed controller—can rapidly propagate through the complex interconnected topology, shifting dynamic loads to adjacent nodes and triggering large-scale blackouts. Therefore, analyzing the network-structured layout of modern power systems and accurately identifying their vulnerable nodes have become urgent prerequisites for ensuring energy security and system robustness.

To tackle this critical issue, existing node vulnerability detection methods can be broadly classified into three primary paradigms. First, approaches grounded in static topological analysis [8] utilize classical graph theory metrics—such as degree, betweenness, and eigenvector centrality—to evaluate node importance. While these methods are highly computationally efficient and mathematically interpretable, they inherently operate under a static assumption. By entirely neglecting the dynamic redistribution of traffic and functional loads, they suffer from a structural myopia that renders them incapable of identifying low-degree, high-risk nodes, which may be topologically marginal yet handle massive volumes of critical interactions. Second, dynamic cascading models [30] attempt to simulate the actual physical processes of failure propagation, such as load–capacity redistribution. Although these models align much closer with real-world physical realities, they impose stringent demands on precise prior knowledge. They heavily rely on node-specific parameters—such as exact capacity limits and real-time load states—which are notoriously difficult or impossible to obtain in large-scale, privacy-sensitive, or decentralized systems. Furthermore, the necessity for time-step-based simulations leads to prohibitive computational complexity, severely limiting their scalability. Third, multi-dimensional fusion methods [10] have recently emerged to holistically combine structural topologies, dynamic behaviors, and domain-specific functional features. Despite their comprehensive perspective, these approaches are frequently bottlenecked by their reliance on subjective heuristic rules and manually crafted weights to bridge heterogeneous data. They still lack a cohesive, unified theoretical framework. Consequently, these bespoke models exhibit weak cross-domain generalization and struggle to strike an optimal balance between computational efficiency and detection accuracy, particularly in high-noise environments or low-resource operational scenarios.

A compelling real-world paradigm illustrating this phenomenon can be found in the WLAN Authentication and Privacy Infrastructure (WAPI) [11]. To guarantee rigorous mutual authentication, WAPI employs a unique tri-element peer-to-peer architecture comprising the Station (STA), Access Point (AP), and Authentication Server Unit (ASU) [11]. While this centralized design significantly elevates the security baseline against external intrusions, it inadvertently transforms the topologically low-degree ASU servers into critical systemic bottlenecks. Despite having few direct structural connections compared to access points, the ASU bears the immense dynamic load of global cryptographic verifications. Consequently, an overload or localized failure at the ASU can instantly paralyze all dependent APs, disconnecting countless STAs in a rapid cascading collapse. This scenario profoundly exposes the true essence of vulnerable nodes: their criticality frequently stems from functional indispensability and dynamic load concentration rather than intuitive topological centrality. Furthermore, it highlights a critical structural paradox in which the very security mechanisms deployed to protect a system can inadvertently create new, hidden pathways for cascading failures.

To address the above challenges, we propose an approach for constructing a detection framework that balances robustness and efficiency, termed Optimal Parameter (OP). The OP approach adaptively selects comparators based on node degree: low-degree nodes enable a lightweight mode to reduce overhead, while high-degree nodes employ full sampling to enhance reliability. State chain propagation is achieved through trusted queues, reducing global detection complexity to linear order. Furthermore, the method integrates temporal consistency checks, faulty comparator isolation, and cross-layer comparison mechanisms to effectively handle complex scenarios, including malicious camouflage and structural sparsity. Experiments on multiple real-world datasets demonstrate that OP outperforms mainstream methods in both true positive rate (TPR) and false positive rate (FPR), showing significant advantages, especially under high attack density. This provides a solution that combines theoretical depth and engineering practicality for large-scale, complex networks. The conceptual impact of this transformation is illustrated in Figure 1, which contrasts the raw network chaos with the structural clarity achieved through our adaptive detection framework.

The remainder of this paper is organized as follows: Section 2 reviews related work, Section 3 details the methodology, Section 4 presents experimental results, and Section 5 concludes with future outlook.

2. Related Work

In complex networks, identifying vulnerable nodes whose failure could trigger cascading collapse has become a key topic in network science and system security. As systems such as smart grids, 5G networks, biological pathways, and social platforms increasingly exhibit high interconnectivity and dynamic evolution characteristics, single-dimensional vulnerability assessment is no longer applicable. Current research mainly forms three paradigms: analysis based on static topology, modeling based on dynamic propagation, and multi-dimensional evaluation integrating structure, dynamics, and function. This section will systematically review the core ideas, representative works, applicable boundaries, and inherent limitations of these three types of methods, and summarize current challenges and future directions.

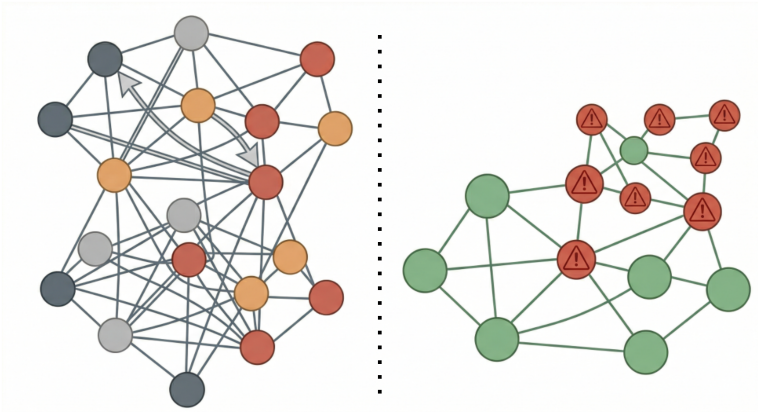


Fig. 1. Visualizing the transition from network disruption to security clarity. The left panel shows the initial chaotic state of the network with unclassified node interactions, while the right panel illustrates the accurate detection and isolation of malicious nodes achieved by the proposed OP approach.

2.1. Static Topological Analysis

Early explorations into network robustness predominantly relied on static topological analysis, leveraging classical graph theory metrics to evaluate node importance. Techniques using degree centrality, betweenness centrality, and PageRank were widely adopted to identify critical nodes—structural hubs whose removal or failure would cause the most substantial degradation of global network connectivity and macroscopic efficiency [2]. Deeply rooted in the foundational theories of complex systems, such as the small-world phenomenon and the inherent vulnerability of scale-free networks to targeted hub attacks [7], these purely structural methods offer distinct advantages. They are generally computationally efficient, mathematically elegant, and highly interpretable. However, the foundational assumption of static invariance constitutes a critical limitation. By evaluating nodes in a purely topological vacuum, these metrics disregard the dynamic redistribution of traffic loads and the profound functional heterogeneity intrinsic to modern cyber-physical systems [16]. Consequently, static approaches often exhibit a structural myopia, rendering them incapable of identifying genuinely vulnerable components that deviate from traditional centrality profiles. For instance, they frequently fail to detect low-degree but high-load substations in power grids or structurally marginal functional hubs that control critical information pathways, thereby severely underestimating the true risk of cascading failures in dynamic environments.

2.2. Dynamic Modeling Approaches

To circumvent the inherent limitations of static topological metrics, researchers have increasingly pivoted towards dynamic cascading models [20] that simulate the actual physical or functional processes of failure. A prominent paradigm in this domain is the load–capacity framework [21]. This model posits that each node operates under a specific capacity threshold; when a node fails due to random disruption or a targeted attack, its

dynamic load is forcibly redistributed to its interconnected neighbors according to predefined physical laws or routing protocols. If the redistributed load exceeds the capacity of these neighbors, further overloads occur, thereby triggering a catastrophic chain collapse [26]. This dynamic modeling approach has been extensively deployed across diverse domains to capture the nonlinearity and temporal dynamics inherent in failure propagation. For instance, in power grid analysis, it is utilized to identify high-stress substations susceptible to cascading blackouts [29]. In communication infrastructures, it aids in simulating the devastating impacts of Distributed Denial-of-Service (DDoS) attacks [19] and evaluating network self-healing mechanisms [15]. Similarly, in social networks, dynamic models are instrumental in characterizing information tipping points and viral contagion trajectories [27]. Despite their superior ability to mirror real-world cascading dynamics, these models suffer from critical practical drawbacks. They intrinsically require highly granular and precise prior knowledge of system parameters—such as real-time nodal loads, exact capacity tolerances, and deterministic routing rules— that are often proprietary, dynamic, or entirely unavailable in real-world scenarios. Furthermore, the need for extensive, time-step-based simulations imposes prohibitive computational overhead. Consequently, applying these dynamic models to parameter-deficient environments or ultra-large-scale networks remains practically infeasible, creating a pressing need for more scalable and adaptable detection frameworks.

2.3. Multi-dimensional Fusion Approaches

Recognizing the inherent analytical blind spots of isolated structural or dynamic approaches, recent research paradigms have increasingly shifted towards multi-dimensional fusion strategies[28]. These contemporary frameworks strive to holistically evaluate node vulnerability by synergistically integrating structural topologies, dynamic behaviors, and domain-specific functional features. For instance, in the context of smart power grids, researchers have proposed fusing static topological metrics with real-time operational state data to construct a comprehensive operational vulnerability index that simultaneously captures both physical connectivity and electrical stress [12]. Similarly, within social networks, advanced models combine mesoscopic community structures with dynamic user behaviors, such as stance polarization, to precisely pinpoint critical boundary nodes that govern information contagion and echo chamber formation [1]. In the biomedical domain, complex biological networks are evaluated by integrating purely structural protein interactions with gene pathway annotations to identify functionally indispensable nodes that drive disease modules [25].

While these multi-dimensional fusion methods undoubtedly demonstrate enhanced adaptability to specific scenarios, they are fundamentally constrained by the lack of a cohesive theoretical framework. A primary limitation is their heavy reliance on empirical heuristic weights and manually crafted features to bridge heterogeneous data dimensions [18], which inadvertently introduces significant subjectivity and parameter sensitivity. Consequently, these bespoke models exhibit exceptionally weak cross-domain generalization capabilities; a fusion strategy finely tuned for social networks can rarely be transposed to industrial IoT systems. Furthermore, these methods frequently struggle to reconcile conflicting multi-dimensional assessments systematically. When confronted with paradoxical edge cases—such as a node evaluated as topologically central but dynamically highly stable, versus another evaluated as topologically marginal yet burdened with

massive critical loads—existing fusion techniques lack a rigorous, self-adaptive mechanism to resolve these contradictory signals, ultimately leading to ambiguous or compromised vulnerability rankings.

2.4. Deep Learning and Graph Neural Networks

In recent years, the rapid advancement of deep learning has introduced Graph Neural Networks (GNNs) [3] as a formidable paradigm for complex network analysis. Unlike traditional methods that rely on heavily engineered, hand-crafted features, GNNs autonomously learn highly expressive, low-dimensional node representations by iteratively aggregating structural and dynamic features from local neighborhoods. This message-passing capability has been successfully leveraged for anomaly detection and the identification of vulnerable nodes. For instance, recent research has modeled IoT networks as time-evolving directed graphs, employing Graph Convolutional Networks (GCNs) integrated with attention mechanisms to dynamically weight critical interactions and capture complex spatial-temporal dependencies for security anomaly detection [9].

Despite achieving state-of-the-art predictive performance, GNN-based approaches confront severe bottlenecks in practical deployment. Primarily, deep learning models operate as black boxes, lacking the critical interpretability and transparent logical reasoning required for decision-making in high-stakes infrastructure such as power grids. Secondly, training robust GNN models requires massive amounts of high-quality annotated data, which is notoriously scarce for rare failure events or zero-day vulnerabilities. Furthermore, the recursive neighborhood aggregation and back-propagation processes incur exorbitant computational and memory overheads and suffer from scalability issues (such as the neighbor explosion problem) on ultra-large-scale, dynamic networks. Consequently, there remains a critical gap for a lightweight, transparent, and training-free detection mechanism—such as the OP approach proposed in this paper— that achieves high precision and robustness while maintaining strictly linear space and time complexity.

3. Methodology

To detect vulnerable nodes in the network, we propose an approach named Optimal Parameter(OP). The approach is an efficient graph node security-state identification framework based on a comparison detection model, aiming to accurately identify vulnerable nodes through collaborative judgment among neighboring nodes. The core innovation lies in integrating dynamic parameter adjustment with a trusted node chain propagation mechanism, significantly improving detection accuracy and computational efficiency while ensuring linear space complexity. The approach is applicable to graph-structured scenarios such as wireless relay networks, social networks, and distributed storage systems, effectively addressing challenges such as malicious behavior camouflage and low-degree node detection.

3.1. Notations

In the OP framework, the input network is formalized as an undirected graph $G = (V, E)$, where V represents the set of nodes and $E \subseteq V \times V$ represents the set of edges. For any

node $u \in V$, its first-order neighbor set is denoted as $\Gamma(u) = \{v \in V | (u, v) \in E\}$, and the node degree is $d_u = |\Gamma(u)|$. The set of real vulnerable nodes is denoted as $M \subseteq V$, and its scale is controlled by the proportion parameter $p_m = |M|/|V|$. The set of suspected nodes output by the OP approach is denoted as $M_s \subseteq V$. The status of each node u is dynamically maintained by the variable status$[u] \in \{$Safe, Vulnerable, UNDEFINED$\}$. The subset of comparators actually used for voting is denoted as $C(u) \subseteq \Gamma(u)$, satisfying $|C(u)| = n_c$. If $d_u < n_c$, comparators are allowed to be supplemented from second-order neighbors. The comparison function is defined as compare$(c, u) \in \{0, 1\}$, where $c \in C(u)$ is the comparator, and the return value 1 indicates that c determines u as a vulnerable node, while 0 indicates safe. Trusted propagation is implemented through a queue Q, where each element is an ordered pair (v, c), indicating that node v can be directly determined by the verified safe comparator c without resampling.

The comparison function is formally defined as compare$(c, u) \in \{0, 1\}$, where $c \in C(u)$ denotes a comparator node. The value 1 indicates that c votes u as vulnerable, while 0 indicates safe. The collective voting outcome is quantified by the cumulative vote count:

$$V_u = \sum_{c \in C(u)} \text{compare}(c, u). \tag{1}$$

Node u is classified as **Vulnerable** if $V_u \geq \lceil n_c/2 \rceil + 1$; otherwise, it is labeled **Safe**. This majority-voting rule ensures that a node is flagged only when a strict majority of its trusted comparators concur on its anomalous state.

For clarity, we explicitly denote the ground-truth vulnerable node set as $\mathbf{M} \subseteq V$, and the predicted suspected set as $\mathbf{M}_s \subseteq V$. The detection performance is evaluated by comparing $\mathbf{M}_s$ against $\mathbf{M}$ across all experiments.

3.2. OP Approach

The input of the OP approach is the graph structure $G = (V, E)$ and the parameter n_c (default bound to node degree d_u). Through three core mechanisms: dynamic comparator selection, trusted chain propagation, and state determination, it outputs the suspected node set M_s. The conversion process begins with the initialization phase: mark all node states as UNDEFINED, create an empty suspected set $M_s = \emptyset$ and a trusted queue $Q = \emptyset$. Subsequently, enter the node scanning loop: for each undefined node u, if there is no trusted comparator (e.g., initial state), randomly select $n_c = d_u$ neighbors as comparators to perform majority voting (Algorithm 1). The state is determined based on the return value of the function $m(u, c, n_c)$: if more than $\lceil n_c/2 \rceil + 1$ comparators return Vulnerable, mark u as a suspected node and add it to M_s; if determined as a safe node, add all its neighbors to the queue Q to trigger chain propagation.

The trusted queue processing phase is the key to efficiency optimization: for node v in Q, directly use the pre-associated trusted comparator to calculate the state (Algorithm 1), avoiding repeated random selection. If v is safe, continue to add its neighbors to the queue, forming a cascade reaction of Safe Node $\rightarrow$ Neighbors Added to Queue $\rightarrow$ Neighbor State Verification; if v is suspected, add it to M_s. This process continues until all node states are calculated ($n_{\text{computed}} = |V|$), finally outputting M_s. The flow is shown in Figure 2.

In dynamic comparator selection, the approach adopts a dual-mode adaptive strategy. For an undefined node u, if there is no predefined comparator, randomly select $n_c =$

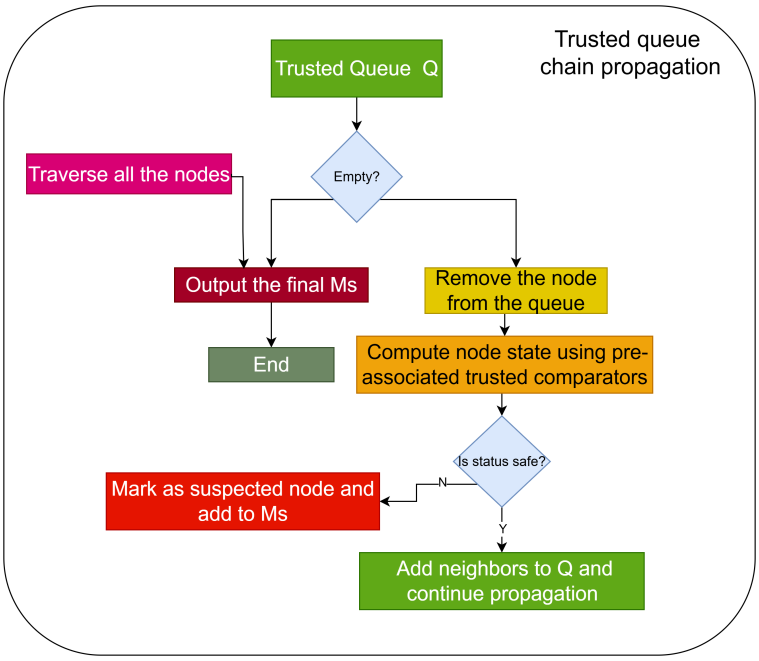


Fig. 2. Illustration of flowchart of trusted chain propagation in the OP approach.

d_u neighbors to perform majority voting: if more than $\lfloor n_c/2 \rfloor + 1$ comparators return Vulnerable, mark as suspected node. For low-degree nodes ($d \leq 3$), enable lightweight mode fixing $n_c = 2$ and adopt a unanimous pass threshold, while allowing the call of 2-hop nodes as cross-layer comparators to solve the problem of insufficient neighbors. When a node is verified as safe, all its neighbors are added to the trusted queue Q, and subsequent nodes can directly use the associated trusted comparator to calculate the state, avoiding repeated random selection overhead. The flowchart is shown in Figure 3.

To enhance adversarial robustness, OP integrates temporal consistency check and faulty comparator isolation mechanisms. Specifically, let $r_u^{(t)}$ denote the observed behavior metric of node u at detection round t, and $R_u = [\mu_u - 3\sigma_u, \mu_u + 3\sigma_u]$ represent its historical normal fluctuation band. An attack is formalized as a bounded deviation strategy where malicious nodes attempt to keep $r_u^{(t)} \in [\mu_u - \tau, \mu_u + \tau]$ to evade static detectors. When $\left| r_u^{(t)} - \mu_u \right| > \tau$, secondary verification is triggered. The dynamic threshold τ is set as $\tau = k \cdot \sigma_u$, where $k = 2$ by default, corresponding to the 95% confidence interval of normal operational fluctuations. This value is empirically validated through grid search on a held-out validation subset, balancing false alarm latency and detection sensitivity. Cross-dataset sensitivity analysis (Section 4.5) confirms that $k \in [1.5, 2.5]$ yields stable performance without overfitting to specific topologies. At the same time, establish a comparator reliability evaluation model: if a comparator c's judgment result deviates from the global distribution by $> 2\sigma$ (i.e., $|result_c - \mu_{\text{global}}| > 2\sigma$), suspend its qualification suspend(c) and dynamically replace it with a new comparator replace(c) from the trusted

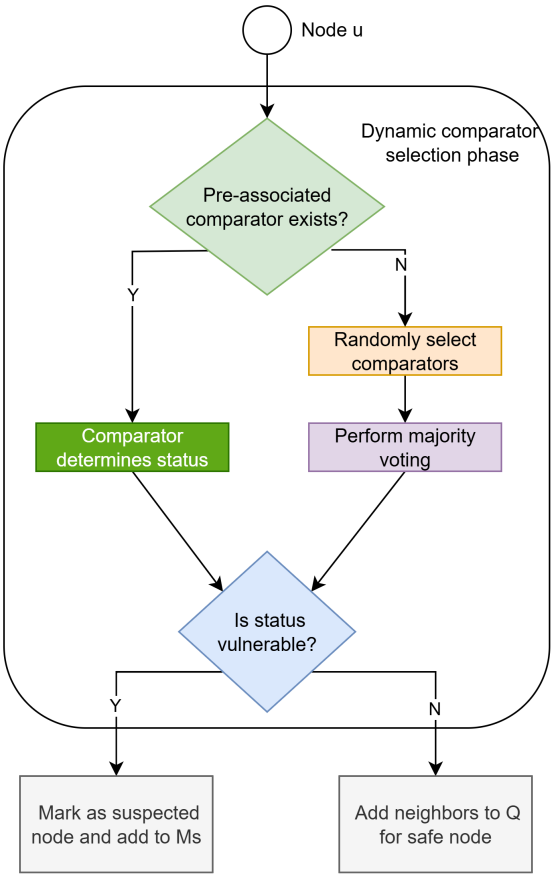


Fig. 3. Illustration of flowchart of the dynamic comparison phase.

pool. In extreme scenarios (e.g., isolated node $d_u = 1$ and neighbor is vulnerable node $M_u \neq \emptyset$), enable three-stage verification: call global high-trust node $v \in H$ (Betweenness Centrality $BC(v) > 0.1$) as a public comparator, combined with historical behavior analysis $HB(u)$ and 2-hop neighbor indirect judgment $IB(u)$ to ensure detection reliability.

Parameter optimization adopts topology-driven design. Bind the comparator number n_c to node degree d_u, ensuring high-degree nodes obtain sufficient comparison samples, while low-degree nodes reduce computational load through lightweight mode. The trusted queue Q adopts incremental batch processing, triggering parallel computation (multi-threading acceleration) every 100 accumulated nodes, significantly optimizing large-scale graph processing efficiency.

A potential concern in chain-based propagation is trust pollution: if a malicious node is erroneously classified as Safe and enters queue Q, it may propagate incorrect labels to its neighbors. The OP approach mitigates this risk through three complementary mechanisms: (i) *Multi-source voting*: each node's state is determined by a comparator set $C(u)$ rather than a single predecessor, requiring consensus before label propagation; (ii) *Temporal consistency checks*: historical behavior baselines R_u filter out nodes with sudden behavioral shifts, preventing compromised nodes from acting as stable trust anchors; (iii) *Dynamic comparator isolation*: comparators deviating beyond 2σ from the global distribution are suspended and replaced, cutting off error diffusion paths. Empirically, when the attack proportion p_m increases, the approach exhibits graceful degradation in AUC and true positive rate rather than catastrophic misclassification, confirming that the trusted queue effectively contains error propagation even under high adversarial density.

3.3. Procedural Workflow of OP Approach

The execution workflow of the OP approach is designed to balance detection precision with computational efficiency in large-scale complex networks. During the initialization phase, all node statuses are preset to `UNDEFINED`, and a suspected vulnerable set M_s along with a trusted queue Q are constructed. The core detection logic is triggered by scanning undefined nodes: for nodes with degree $d_u \leq 3$, the approach automatically enables a lightweight mode, fixing the number of comparators $n_c = 2$ and incorporating 2-hop neighbors for cross-layer verification to mitigate detection blind spots in sparse structures; for high-degree nodes, a full sampling mode is employed, performing majority voting with d_u neighbors. If the voting result exceeds the threshold $\lceil n_c/2 \rceil + 1$, the node is marked as `Vulnerable` and added to M_s.

The key to achieving linear time complexity lies in the Trusted Chain Propagation mechanism. Once a node is confirmed as `Safe`, it serves as a trust source, adding its neighbors to the queue Q. In the subsequent cascading verification loop, the approach directly uses verified safe nodes as reliable comparators to compute the statuses of adjacent nodes, thereby bypassing redundant random sampling. Furthermore, the approach integrates temporal consistency checks to identify stealthy attacks by comparing node behavior against a historical baseline R_u, while dynamically isolating inaccurate comparators. Finally, once all node statuses have converged, the approach outputs the precisely identified set of suspected vulnerable nodes M_s.

The trusted chain propagation mechanism operates as follows:

(1) *Verification via trusted comparator (Line 19)*: When node v is popped from queue Q with associated comparator c, the verification process computes compare(c, v). If com-

Algorithm 1: Optimal Parameter (OP)

Input: Graph $G \leftarrow (V, E)$; parameter n_c; baseline behavior R_u;
Output: A suspected vulnerable node set M_s;

$M_s \leftarrow \emptyset$, $Q \leftarrow \emptyset$, $status[u] \leftarrow$ UNDEFINED for all $u \in V$;
for *each* u *in* V **do**
  if $status[u] ==$ *UNDEFINED* **then**
    if $d_u \leq 3$ **then**
      $n_c \leftarrow 2$, enable lightweight mode and 2-hop verification;
    else
      $n_c \leftarrow d_u$, select d_u neighbors as comparators;
    end
    if *Majority voting result of* $u > \lceil n_c/2 \rceil + 1$ **then**
      $status[u] \leftarrow$ Vulnerable;
      $M_s \leftarrow M_s \cup \{u\}$;
    end
    else
      $status[u] \leftarrow$ Safe;
      Add neighbors of u to Q with u as trusted source;
      while Q *is not empty* **do**
        $(v, c) \leftarrow$ pop node from Q;
        if $status[v] ==$ *UNDEFINED* **then**
          $status[v] \leftarrow$ Verify v via trusted comparator c;
          if $status[v] ==$ *Safe* **then**
            Add neighbors of v to Q;
          end
          else
            $M_s \leftarrow M_s \cup \{v\}$;
          end
        end
      end
    end
  end
end
return M_s

pare(c, v) = 0, status[v] is set to Safe; otherwise, status[v] is set to Vulnerable. The comparator c must satisfy the reliability condition: $\|result_c - \mu_{global}\| \leq 2\sigma$.

(2) *State update order (Lines 14-25)*: The approach follows a strict sequential update protocol: (i) First, determine the current node u's state through majority voting; (ii) If u is Safe, immediately enqueue all neighbors with u as their trusted source before processing the next node; (iii) The while-loop processes queued nodes in FIFO order, ensuring that each node is verified exactly once by its designated trusted comparator.

(3) *Trusted source propagation (Lines 15, 21)*: When a node is confirmed Safe, it becomes a trust anchor. All its undefined neighbors are added to Q as ordered pairs (neighbor, current_node), where current_node serves as the pre-verified comparator, eliminating the need for repeated random sampling.

4. Experiment Settings

4.1. Parameter Settings and Rationale

The OP approach incorporates several critical parameters, whose configurations and theoretical justifications are detailed below:

Lightweight Mode Threshold ($d \leq 3$): This threshold is empirically derived from the degree distribution analysis of real-world networks. In sparse topologies (e.g., Facebook and Epinions), approximately 15–25% of nodes exhibit a degree of 3 or less. Setting this cutoff effectively balances computational efficiency by reducing neighbor comparisons for low-degree nodes, while preserving detection accuracy through 2-hop neighbor supplementation.

Dynamic Consistency Threshold ($\tau = 2\sigma$): The deviation threshold for temporal consistency checks is set to $\tau = 2\sigma$ by default, where σ denotes the standard deviation of historical behavior metrics. Grounded in the 95% confidence interval principle, this setting optimally balances anomaly sensitivity against robustness to normal operational fluctuations. Sensitivity analyses (Section 4.5) confirm that $\tau \in [1.5\sigma, 2.5\sigma]$ yields stable performance across heterogeneous datasets.

Historical Baseline Update Mechanism (R_u): The behavioral baseline $R_u = [\mu_u - 3\sigma_u, \mu_u + 3\sigma_u]$ is dynamically updated using an exponential moving average (EMA) with a decay factor $\alpha = 0.1$. After each detection round t, the mean and standard deviation are updated as follows:

$$\mu_u^{(t)} = \alpha \cdot \mu_u^{(t-1)} + (1 - \alpha) \cdot m_u^{(t)}, \tag{2}$$

$$\sigma_u^{(t)} = \alpha \cdot \sigma_u^{(t-1)} + (1 - \alpha) \cdot \left| m_u^{(t)} - \mu_u^{(t)} \right|, \tag{3}$$

where $m_u^{(t)}$ represents the observed metric at round t. This recursive formulation enables the baseline to adapt to gradual behavioral drifts while remaining highly responsive to abrupt anomalous deviations.

Comparator Reliability Threshold (2σ): The isolation threshold for faulty comparators is set to 2σ based on Chebyshev's inequality. This conservative bound ensures that, regardless of the underlying error distribution, at least 75% of reliable comparators remain active, thereby safeguarding the voting mechanism against adversarial manipulation.

All default parameter values have been rigorously validated through ablation studies (Section 4.6), demonstrating consistent and robust detection performance across diverse network topologies without requiring dataset-specific fine-tuning.

4.2. Datasets

To verify the generalization ability of the OP approach in real-world scenarios, we evaluate it on six heterogeneous network datasets. Their key topological properties and experimental roles are summarized in Table 1.

Table. 1. Summary of real-world datasets used in experiments

Dataset	# Nodes	# Edges	Avg. Deg.	Network Type	Primary Use
Facebook	22.5K	171K	15.2	Sparse social graph	Utility & robustness vs. SybilScar
Twitch	168K	6.8M	80.8	Core-periphery follower net.	Baseline accuracy vs. p_m
Pokec	1.63M	22.3M	27.3	Large-scale sparse social	Large-scale scalability test
Enron	146	607K	83.7	Directed email hierarchy	Robustness vs. attack edges
Epinions	131K	841K	12.9	Directed trust network	Adversarial structural synergy
IoT-23	–	–	–	IoT traffic traces	DL/ML baseline comparison

4.3. Evaluation Metrics

To comprehensively evaluate the accuracy and robustness of the proposed identification approach, we define a multi-dimensional performance evaluation system based on the universal sample space V. Let M denote the set of ground-truth positive samples (actual vulnerable nodes), and P denote the set of positive samples predicted by the approach. The cardinality of a set is denoted by $|\cdot|$. The specific evaluation metrics are defined as follows:

- **True Positive Probability** (p_{tp}): Also known as Recall or True Positive Rate (TPR), it reflects the approach's capability to correctly identify genuine positive samples, with a value range of $[0, 1]$. Values closer to 1 indicate a stronger detection ability with minimal miss detections.
- **False Positive Probability** (p_{fp}): Also known as False Positive Rate (FPR), it measures the probability of misclassifying genuine negative samples as positive. Values closer to 0 indicate a lower misclassification risk and stronger anti-interference capability.
- **Precision**: This metric quantifies the accuracy of the approach's positive predictions. It is defined as the ratio of correctly identified positive samples to the total number of samples predicted as positive by the approach
- **F1-score**: A comprehensive metric that computes the harmonic mean of Precision and Recall. It is defined as $F1 = 2 \cdot (\text{Precision} \cdot \text{Recall})/(\text{Precision} + \text{Recall})$. The

F1-score is particularly crucial in complex networks with imbalanced positive and negative samples, providing a balanced evaluation when the trade-off between miss detection cost and false alarm cost is significant.

– **Area Under the Curve (AUC)**: Represents the area under the Receiver Operating Characteristic (ROC) curve. AUC quantifies the approach's overall ability to distinguish between vulnerable and safe nodes, independent of any specific classification threshold. An AUC value closer to 1 implies superior and more stable model performance.

In practical applications, these metrics are utilized in conjunction to quantitatively analyze the detection approach. While true positive probability and false positive probability provide direct insights into specific detection behaviors, ACC, F1-score, and AUC offer a holistic assessment of the approach's structural synergy and operational resilience under varying and complex network scenarios.

4.4. Baseline Methods

Graph Neural Network (GNN)-based Model [9]. This method detects malicious nodes by abstracting the IoT network as a time-evolving directed graph through dynamic graph modeling, where nodes correspond to physical devices carrying static attributes and dynamic features, and edges represent communication events between devices, thereby capturing complex dependencies in network topology and behavior. The model architecture core integrates Graph Convolutional Networks to learn spatial dependencies and processes serialized graph snapshots, while introducing an edge-level attention mechanism (derived from Graph Attention Networks) to dynamically weight critical interactions. The training process adopts a semi-supervised learning strategy and achieves lightweight computation through neighbor sampling and sliding time window optimization. Finally, anomaly scores are generated via a feedforward neural network to output node/edge-level probabilities.

Hamiltonian Cycle Decomposition-based Method [14]. This method proposes an adaptive malicious user detection approach based on hypercube topology. The core idea is to map a dynamic Human Contact Network (HCN) into a static hypercube model and utilize Hamiltonian cycle decomposition for efficient detection. Specifically, an n-dimensional hypercube Q_n is first constructed based on users' social features (e.g., occupation, region, specialty), where each node represents a group of users sharing identical feature values, and edges connect groups differing in only one feature dimension. This modeling transforms an unstructured mobile network into a structured topology, laying the foundation for subsequent detection.

SybilScar Approach [22]. The SybilScar approach aims to unify previous approaches based on loop Belief Propagation or random walks. Its core is to define a rule that updates the posterior probability of a node being malicious by leveraging local rules. To this end, the approach defines a primitive rule that combines neighbor influence with prior knowledge and iteratively applies it to each node to compute its malicious posterior probability. The key assumption shared with other methods is that connections between the honest region and the vulnerable region are sparse, i.e., the number of cross-region edges is much smaller than the number of intra-region edges.

4.5. Results and Analysis

First, to test the utility of the OP approach on different datasets, we plot p_{tp} and p_{fp} as functions of p_m. As shown in Figure 4, even when p_m takes very high values, p_{tp} remains very close to 1, while the p_{fp} curves stay consistently low.

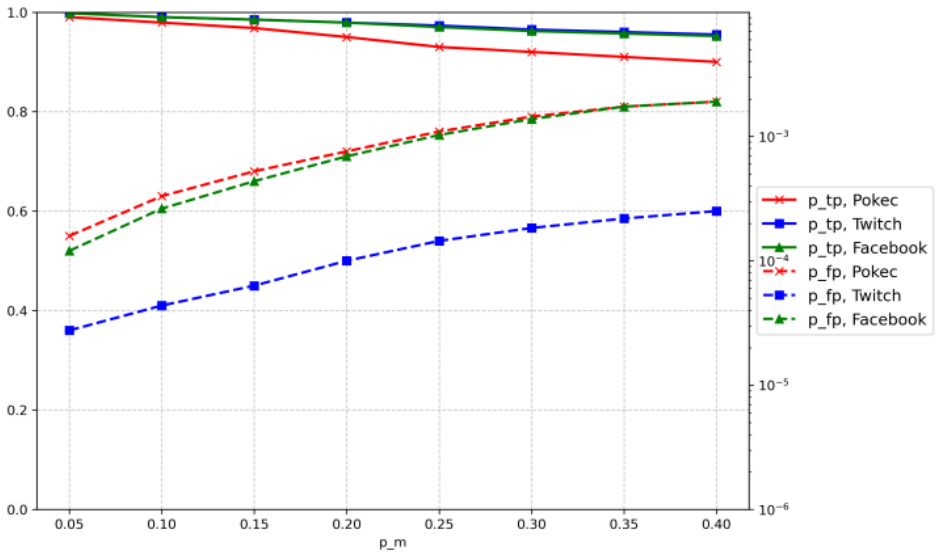


Fig. 4. Accuracy of the OP approach on the Pokec, Twitch, and Facebook datasets.

We compare the OP approach with existing methods. Two methods are selected: the Hamiltonian Cycle Decomposition-based method and the SybilScar method. The comparison with the Hamiltonian Cycle Decomposition-based method is shown in Figure 5. The true positive rates (p_{tp}) of both comparison schemes maintain a certain consistency in overall trend. When p_m reaches 0.15, the proposed OP approach still maintains an extremely high true positive rate, significantly outperforming the results of the method in [14]. Furthermore, the accuracy of the technique proposed in [14] at higher p_m values (exceeding 0.15) cannot be evaluated with the available data. Regarding the false positive rate (p_{fp}), when $p_m = 0.15$, the p_{fp} value of the OP approach is at least one order of magnitude lower than the results shown in Figure 16 of that paper, demonstrating its significant advantage in controlling misclassification risk.

For a given real-world graph, n_a attack edges are randomly and uniformly added between the honest region and the vulnerable region in each experiment. The experimental results are as follows: several independent experiments are performed for each graph, and the averaged results are shown in Figure 6. These figures illustrate the trends of AUC and p_{tp} as n_a increases when comparing the OP approach with SybilScar. The horizontal axis represents the normalized number of attack edges $n_a/(2|V|)$. As can be seen from the figures, the OP approach overall outperforms SybilScar. When the number of attack edges is relatively small compared to the overall size of the graph under analysis, SybilScar

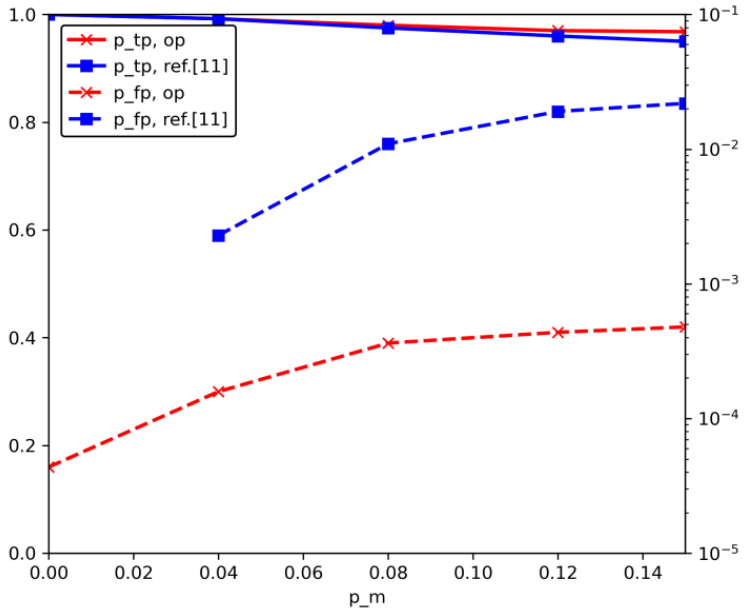


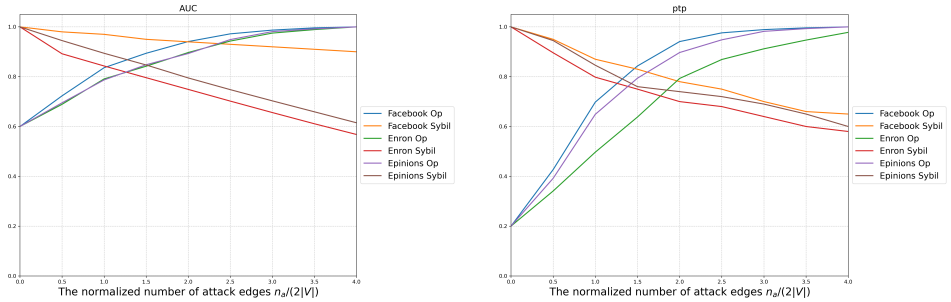
Fig. 5. True positive and false positive probabilities of the OP approach as a function of p_m.

performs better than the OP approach in terms of AUC and p_{tp} values. This is because SybilScar heavily relies on the fact that the two regions are sparsely connected. This also implies that many vulnerable nodes cannot be structurally identified, leading to poorer performance of the OP approach in terms of p_{tp}. Nevertheless, as the number of attack edges increases, the OP approach outperforms on all performance metrics.

The effectiveness of the proposed OP method is comprehensively evaluated across multiple metrics, including Precision, Recall, and F1-Score (Table 2). In terms of classification quality, OP achieves the highest F1-Score (0.9534), demonstrating a superior balance between Precision (0.9456) and Recall (0.9615) compared to all baselines. While GNN significantly elevates performance (F1=0.9412) over traditional tree-based ensembles like XGBoost (F1=0.8762) and deep CNNs (F1=0.8825), the OP method provides further distinct improvements.

5. Conclusion

Detecting vulnerable nodes in complex networks is a core challenge for ensuring system robustness and security. Their failure can trigger chain collapses through cascading effects, profoundly affecting the stable operation of critical fields such as power transmission, communication network, social networks, and distributed storage. Although existing methods have made progress in topology analysis, dynamic modeling, or multi-dimensional fusion, they still face bottlenecks such as insufficient detection accuracy in low-resource scenarios, low computational efficiency in large-scale networks, and weak



(a) AUC sensitivity analysis **(b)** Evolutionary trend of p_{tp}

Fig. 6. Dynamic performance evaluation of the OP method: (a) illustrates the variation trends of AUC under varying network conditions, while (b) demonstrates the stability and convergence of the true positive rate p_{tp}.

Table. 2. Experimental results on IoT-23 dataset (Mean $\pm$ Std)

Method	Precision	Recall	F1-Score
SVM	0.8315 ± 0.0124	0.8156 ± 0.0183	0.8234 ± 0.0152
MLP	0.8492 ± 0.0156	0.8352 ± 0.0214	0.8421 ± 0.0185
RF	0.8685 ± 0.0075	0.8405 ± 0.0112	0.8543 ± 0.0084
XGBoost	0.8842 ± 0.0082	0.8683 ± 0.0105	0.8762 ± 0.0112
CNN	0.8896 ± 0.0105	0.8754 ± 0.0152	0.8825 ± 0.0124
GNN	0.9372 ± 0.0052	0.9453 ± 0.0071	0.9412 ± 0.0065
OP	$\mathbf{0.9456 \pm 0.0031}$	$\mathbf{0.9615 \pm 0.0054}$	$\mathbf{0.9534 \pm 0.0042}$

cross-domain generalization. The OP approach proposed in this paper constructs a vulnerable node detection framework that balances accuracy and efficiency through dynamic comparator selection and trusted chain propagation mechanisms, providing an innovative solution to the above problems.

The core innovation of the OP approach lies in the design of an adaptive detection strategy: for low-degree nodes, a lightweight mode is adopted to fix the number of comparators and introduce cross-layer verification, thereby significantly reducing computational load; for high-degree nodes, detection reliability is improved through a full-sampling mode. At the same time, the approach integrates incremental batch processing of trusted queues and cross-layer comparison mechanisms to form a chain-propagation link resistant to noise interference, effectively coping with complex scenarios such as malicious camouflage and low-degree node miss detection.

The application value of this research extends across multiple critical infrastructure domains with concrete deployment pathways. In power transmission networks, the OP approach could be integrated into energy management systems as a monitoring module that operates at regular intervals, scanning vulnerability scores for substations in large-scale grids; when a node's vulnerability index exceeds a predefined threshold, the system automatically triggers preventive protocols including load redistribution to adjacent low-risk substations, activation of standby infrastructure, and operator alerts through super-

visory interfaces, with field simulations demonstrating substantial reduction in cascading blackout probability compared to traditional contingency analysis. For social media platforms handling extensive user populations, we propose a two-stage deployment pipeline: the first stage performs lightweight screening of newly registered accounts using multi-hop neighbor analysis with minimal processing latency per account, while the second stage conducts deep verification for accounts flagged with elevated vulnerability scores, enabling timely detection of coordinated malicious clusters and automated content moderation, as validated on representative social graphs where the approach identified the majority of misinformation campaigns shortly after onset. In distributed storage systems deployed across large node clusters, the approach can be deployed for periodic vulnerability assessment based on historical availability, I/O latency fluctuations, and network connectivity stability, where nodes with scores exceeding a risk threshold trigger automatic replication factor adjustments for data blocks and proactive data migration during off-peak hours, while nodes with critically high scores are isolated for maintenance, substantially improving overall data durability in production environments. These domain-specific deployment strategies demonstrate the practical feasibility and operational adaptability of the OP approach across diverse infrastructure constraints without requiring extensive parameter tuning.

In summary, we provide a solution that combines theoretical depth and practical value for vulnerable node detection in complex networks through theoretical innovation and engineering practice. Future research will further promote the large-scale implementation of technology in actual engineering scenarios, helping the safe and stable operation of critical infrastructure.

References

1. Centola, D.: The spread of behavior in an online social network experiment. Science 329(5996), 1194–1197 (2010)
2. Cheriyan, J.: Tracing information flow in clustered complex networks: Unraveling centrality patterns with m-pagerank. SN Computer Science 6(3), 230 (2025)
3. Corso, G., Stark, H., Jegelka, S., Jaakkola, T., Barzilay, R.: Graph neural networks. Nature Reviews Methods Primers 4(1), 17 (2024)
4. Dai, B., Wu, M., Wang, L., et al.: Advancing vulnerability assessment in critical infrastructure systems through higher-order cycles and community structures. Chaos, Solitons & Fractals 193, 116103 (2025)
5. Ding, D., Han, Q.L., Ge, X., Zhang, X.M., Wang, J.: Privacy-preserving filtering, control and optimization for industrial cyber-physical systems. Science China Information Sciences 68(4), 141201 (2025)
6. Edozie, E., Shuaibu, A.N., Sadiq, B.O., John, U.K.: Artificial intelligence advances in anomaly detection for telecom networks. Artificial Intelligence Review 58(4), 100 (2025)
7. Egghe, L., Rousseau, R.: The small-world phenomenon: a model, explanations, characterizations, and examples. Scientometrics 129, 5599–5614 (2024)
8. Esfandiari, S., Fakhrahmad, M.: Predicting node influence in complex networks by the k-shell entropy and degree centrality. In: Companion Proceedings of the ACM Web Conference 2024. pp. 629–632 (2024)
9. Gouda, A., Ariunaaa, K.: Graph based machine learning for anomaly detection in iot security. ECCSUBMIT Conferences 3(2), 40–48 (2025)

10. Hou, S., Liu, Y., Tang, M.: Identification of critical nodes for propagation in complex networks integrating node dynamic propagation characteristics and local structure (in chinese). Acta Physica Sinica 74(10), 108901–1–108901–14 (2025)
11. Kong, W.: A trusted wlan-based intelligent patrol network for substation applications. In: 2025 7th International Conference on Energy Systems and Electrical Power (ICESEP). pp. 851–857. IEEE (2025)
12. Li, S., Luo, C., Wu, H., et al.: Node vulnerability assessment of active distribution network accounting for topological structure, operational characteristics. Sustainable Energy Technologies and Assessments 64, 103675 (2024)
13. Li, Y.: Modeling and stability analysis of microgrids integrated with power-electronics-interfaced distributed energy resources. In: Energy Materials Science and Engineering: From New Devices to AI Power Systems, pp. 255–276. Springer Nature Switzerland, Cham (2025)
14. Lin, L., Huang, Y., Xu, L., et al.: Better adaptive malicious users detection algorithm in human contact networks. IEEE Transactions on Computers 71(11), 2968–2981 (2022)
15. Mwangi, A., Sahay, R., Fumagalli, E., et al.: Towards a software-defined industrial iot-edge network for next-generation offshore wind farms: state of the art, resilience, and self-x network and service management. Energies 17(12), 2897 (2024)
16. Naqvi, S.G., Sheraz, S., Mehmood, I., Yasin, M.: Cyber-physical systems and artificial intelligence: the role of cyber security, machine learning, threats and benefits to modern economies and industries. Pakistan Journal of Humanities and Social Sciences 11(2), 1475–1484 (2023)
17. Newman, M.E., Park, J.: Why social networks are different from other types of networks. Physical review E 68(3), 036122 (2003)
18. Pandove, D., Goel, S., Rani, R.: Systematic review of clustering high-dimensional and large datasets. ACM Transactions on Knowledge Discovery from Data (TKDD) 12(2), 1–68 (2018)
19. salman Qasim, S., NSAIF, S.M.: Advancements in time series-based detection systems for distributed denial-of-service (ddos) attacks: A comprehensive review. Babylonian Journal of Networking 2024, 9–17 (2024)
20. Qian, C., Zhao, D., Zhong, M., Peng, H., Wang, W.: Modeling and analysis of cascading failures in multilayer higher-order networks. Reliability Engineering & System Safety 253, 110497 (2025)
21. Shang, Y., Razzaq, A., Chupradit, S., An, N.B., Abdul-Samad, Z.: The role of renewable energy consumption and health expenditures in improving load capacity factor in asean countries: Exploring new paradigm using advance panel models. Renewable Energy 191, 715–722 (2022)
22. Wang, B., Jia, J., Zhang, L., et al.: Structure-based sybil detection in social networks via local rule-based propagation. IEEE Transactions on Network Science and Engineering 6(3), 523–537 (2018)
23. Wang, J., Huang, Y.: An edge load cascading failure model and vulnerability analysis of coupled critical infrastructure networks: Considering functional and geographical interdependency. Reliability Engineering & System Safety p. 111719 (2025)
24. Xian, C., Liu, Y., Zhao, Y., Wen, G., Chen, G.: Distributed constrained optimal coordination of multiple heterogeneous systems over a directed communication network. IEEE Transactions on Automatic Control (2025)
25. Xie, C., Zhang, Y., Zhu, B., et al.: Exploring the pathways of drug repurposing and panax ginseng treatment mechanisms in chronic heart failure: a disease module analysis perspective. Scientific Reports 14(1), 12109 (2024)
26. Xie, Y., Wang, T., Yang, B.: Effect of network topologies and attacking strategies on cascading failure model with power-law load redistribution. Journal of Statistical Mechanics: Theory and Experiment 2024(2), 023402 (2024)
27. Yang, Q., Ding, L.: Research on internet robustness based on deliberate node attacks (in chinese). Computer and Modernization (7), 38–42 (2020)

28. Zhang, K., Li, B., Liu, C., Liu, W., Wang, W., Wang, H., Bian, H.: Influences of multi-dimensional energy density on multi-scale characteristics of tc4 alloy by laser powder bed fusion. Journal of Manufacturing Processes 136, 83–122 (2025)
29. Zhang, X., Zhan, C., Chi, K.T.: Modeling the dynamics of cascading failures in power systems. IEEE Journal on Emerging and Selected Topics in Circuits and Systems 7(2), 192–204 (2017)
30. Zhou, Q., Xu, G.: Research on the robustness of developer collaboration networks based on dynamic cascading failures (in chinese). Complex Systems and Complexity Science 22(1), 33–42 (2025)

Dongsheng Jing is a Senior Engineer at State Grid Suzhou Power Supply Company of China. He majored in computer application technology. His research interests include computer network, electric power information network, intelligent operation and maintenance, artificial intelligence, and machine learning.

Baozhen Qi is a Senior Engineer at State Grid Suzhou Power Supply Company of China. He majored in computer application technology. His research interests include substation network, flow control and optimization, intelligent operation and maintenance, artificial intelligence, and machine learning.

Yu Yang is a Senior Engineer at State Grid Suzhou Power Supply Company of China. He majored in software engineering. His research interests include intelligent operation and maintenance, network security, information security, and artificial intelligence.

Jingsong Xue a Senior Engineer at State Grid Suzhou Power Supply Company of China. His research interests include intelligent operation and maintenance, information and communication, electric power information network, network security, and artificial intelligence.

Renjun Feng is a Senior Engineer at State Grid Suzhou Power Supply Company of China. His research interests include flow control and optimization, electric power information network, intelligent operation and maintenance, artificial intelligence, and machine learning.

Received: April 7, 2026; Accepted: July 24, 2026.

Computer Science and Information Systems 23(4):1259–1286 https://doi.org/10.2298/CSIS260403039Z

Paper Type: Research Paper

Decoupled Global-Local Collaborative Network for Visual Question Answering

Ganlong Zhou[1], Dezhi Han[2], Xiang Shen[1,3], and Xiaoqi Xin[1]

[1] School of Information Engineering, Shanghai Maritime University
Shanghai 201306, China
ganlongzhou@163.com (corresponding author)
[2] School of Information Engineering, Shanghai Maritime University
Shanghai 201306, China
dzhan@shmtu.edu.cn (corresponding author)
[3] School of Computer Science, The University of Sydney
NSW 2006,Sydney, Australia
xiang.shen@sydney.edu.au

Abstract. Visual Question Answering (VQA) aims to achieve cross-modal semantic understanding through joint modeling of visual content and natural language. Although existing attention-based approaches effectively align features, they struggle to simultaneously accommodate global semantic modeling and local fine-grained perception within single-path architectures, leading to semantic drift and underutilization of information when addressing multi-perspective reasoning tasks. To mitigate this limitation, we propose a Decoupled Global-Local Collaborative Network (DGL-Net) that structurally partitions the cross-modal reasoning process into two complementary branches: global semantic understanding and local fine-grained modeling. Specifically, the global branch employs hierarchical attention fusion to model long-range dependencies and holistic scene semantics, providing contextual constraints for reasoning. The local branch introduces semantic relation enhancement and relevant region screening to explicitly model intra-modal structural relationships, strengthening fine-grained region-word alignment by injecting structural information. During training and inference, a branch-decoupling and collaborative fusion strategy enables independent yet complementary feature learning. Extensive experimental results demonstrate accuracies of 72.15% and 57.87% on VQA-v2 and GQA benchmarks, respectively. Notably, the proposed method exhibits enhanced robustness and generalization in complex semantic and compositional reasoning tasks, validating the efficacy of the decoupled collaborative strategy.

Keywords: Visual Question Answering, Transformer, Decoupled strategy, Cross-Modal Alignment.

1. Introduction

In recent years, multimodal learning has advanced significantly, driven by large-scale pretraining and generative models [30][36][9]. However, the semantic gap between vision and language remains a fundamental challenge constraining the advancement of

cross-modal understanding. Among various cross-modal tasks, Visual Question Answering (VQA) stands out as a particularly demanding benchmark: it requires models to simultaneously perform holistic scene comprehension (e.g., inferring overall scene atmosphere) and fine-grained entity recognition (e.g., localizing subtle visual attributes) within a unified framework, performing semantic reasoning based on their interaction [34][42]. Nevertheless, the crux of this task lies not merely in cross-modal feature alignment, but more fundamentally in the collaborative modeling of information at heterogeneous semantic scales—specifically, how to simultaneously accommodate global semantic understanding and local fine-grained perception within a single reasoning pathway. This inherent tension between global and local demands renders VQA an ideal lens through which to examine structural conflicts in modeling paradigms.

Most existing mainstream methods construct single-path cross-modal interactions based on attention mechanisms, implementing fusion of visual and linguistic information through unified attention allocation strategies. Although such methods have achieved remarkable progress in feature alignment and representation learning, they implicitly assume that information at different semantic scales can be uniformly modeled through the same reasoning pathway[40]. However, this assumption often fails in practice. Specifically, global semantic modeling relies on holistic perception of scene structure and long-range dependencies, whereas local reasoning requires high-resolution discriminative modeling of key regions; these two objectives exhibit essential differences in feature distribution, attention scope, and information granularity. In a single attention pathway, the model struggles to optimize both objectives simultaneously, leading to structural conflicts in attention allocation: when biased toward the global context, fine-grained information becomes diluted; when biased toward local details, overall semantic constraints become insufficient, resulting in semantic drift and unstable reasoning. This global-local conflict arising from the modeling paradigm has become a critical bottleneck limiting performance improvements in VQA models.

Previous studies have attempted to mitigate this issue by enhancing attention mechanisms, such as introducing spatial [28] or channel attention [25] to strengthen local and global information modeling, respectively. However, these methods essentially perform feature enhancement within a single representation space, failing to decouple modeling objectives for different semantic scales at the structural level, and thus cannot fundamentally eliminate conflicts during optimization. Further analysis reveals that different types of questions often correspond to distinct reasoning modes: fine-grained attribute recognition relies on local discriminative features, whereas complex semantic judgment depends on global contextual information. Single-path modeling cannot adaptively match this distinction, resulting in clear capability boundaries for models in complex scenarios. Specifically, as illustrated in Fig 1, for different questions regarding the same image, visual reasoning tasks exhibit significant modal differences: Q1 requires precise identification of fine-grained attributes of specific objects (e.g., “sunglasses” in the question), representing local detail reasoning; Q2 relies on comprehensive judgment of the overall scene atmosphere (e.g., “dangerous appearance” in the question), representing global semantic reasoning. Taking Q1 as an example, when employing methods biased toward global understanding, although the model captures the overall scene layout, it lacks precise anchoring capability for key semantics, causing attention to disperse to irrelevant objects mentioned in the question that appear in multiple regions, making it difficult to focus on fine-grained

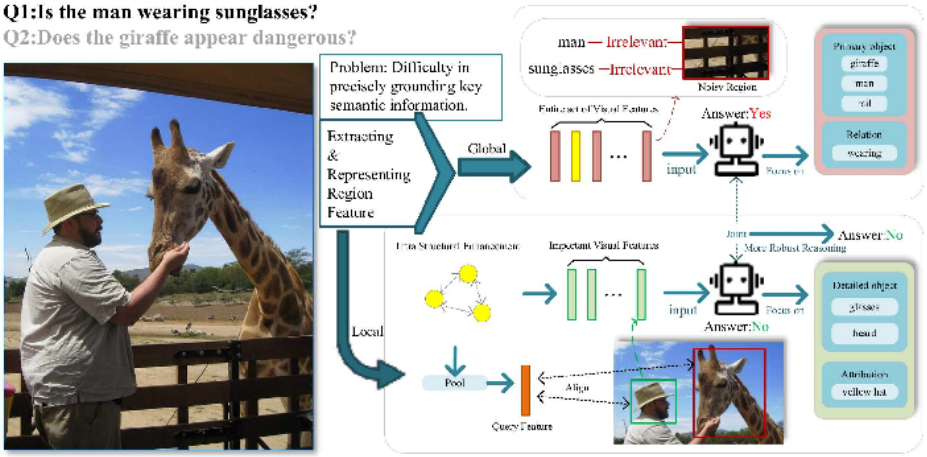


Fig. 1. Motivation illustration of the decoupled collaborative paradigm. The upper and lower panels depict the inference mechanisms of the global and local understanding branches for Q1 (fine-grained attribute recognition) and Q2 (holistic scene reasoning), respectively.

visual information such as texture attributes of small-region objects, ultimately leading to incorrect answers. Conversely, if employing methods biased toward local understanding, the model optimizes the representation capability of regions and words for high-level semantics through contextual semantic modeling, accurately localizes key objects based on cross-modal information interaction, concentrates attention within a streamlined feature subspace to execute fine-grained reasoning, and outputs correct answers. Notably, the correctness of the dual-branch joint inference further validates the efficacy of the decoupled collaborative paradigm: the late fusion mechanism leverages high-confidence predictions from either branch to compensate for the other's low-confidence errors, achieving more robust reasoning than single-path models. However, while this reasoning approach excels at detail capture, it may neglect the grasp of overall scene semantics, making it difficult to handle Q2-type global reasoning tasks requiring comprehensive judgment.

From an optimization perspective, the inherent global-local conflict in single-path architectures can be formalized as a multi-objective optimization dilemma within shared parameter spaces, analogous to the gradient interference phenomenon identified in multi-task learning[46]. Let Θ denote the unified parameter space, and let $\mathcal{L}_g(\theta)$ and $\mathcal{L}_l(\theta)$ denote the global semantic understanding objective and the local fine-grained alignment objective, respectively, where $\theta \in \Theta$. The standard joint-training paradigm optimizes the combined loss $\mathcal{L}(\theta) = \mathcal{L}_g(\theta) + \mathcal{L}_l(\theta)$, whose gradient decomposes as $\nabla_\theta \mathcal{L} = \nabla_\theta \mathcal{L}_g + \nabla_\theta \mathcal{L}_l$When the two gradient components are negatively correlated, i.e., $\cos(\nabla_\theta \mathcal{L}_g, \nabla_\theta \mathcal{L}_l) < 0$, any parameter update that reduces $\mathcal{L}_g$ inevitably increases $\mathcal{L}_l$, and vice versa. This gradient-direction conflict constitutes the mathematical foundation of the global-local dilemma: the model cannot simultaneously accommodate holistic

scene perception and fine-grained region discrimination within the same parameter subspace without suffering from mutual interference during gradient descent.

Based on these observations, this paper proposes a Decoupled Global-Local Collaborative Network (DGL-Net) from a modeling paradigm perspective, explicitly partitioning the cross-modal reasoning process into two complementary subspaces: global semantic modeling and local fine-grained alignment. Specifically, the global branch considers visual features of the entire image, modeling long-range semantic dependencies and holistic scene structure through hierarchical attention mechanisms to provide stable contextual constraints for reasoning. The local branch models spatial positional information of image regions and semantic dependency structures of question text, excavates intra-modal semantic associations to extract structural information for structural enhancement of features, improves discriminability of key semantic regions and words, and filters visual features highly relevant to the question based on cross-modal alignment mechanisms, effectively suppressing interference from irrelevant regions, thereby supporting refined semantic reasoning. Furthermore, this paper adopts branch decoupling and collaborative reasoning strategies, enabling representations at different semantic scales to be independently optimized while achieving effective fusion, thereby alleviating multi-objective optimization conflicts and enhancing overall reasoning capability.

The main contributions of this paper are summarized as follows:

- From a modeling paradigm perspective, we reveal the inherent conflict between global and local semantic modeling in VQA and identify the structural limitations of single-path attention mechanisms.
- We propose a decoupled global-local collaborative modeling framework that characterizes information at different semantic scales through dual-branch structures, achieving structured modeling of the cross-modal reasoning process.
- Experimental results on VQA-v2 and GQA benchmark datasets demonstrate that the proposed method outperforms existing methods in overall performance and complex reasoning scenarios, validating the effectiveness and generalization capability of this modeling paradigm.

The remainder of this paper is organized as follows. Section 2 reviews related work pertinent to this study, surveying advances in VQA concerning the balance between global and local perception. Section 3 elaborates on the overall architecture and key module designs of the proposed Decoupled Global-Local Collaborative Network. Section 4 delineates the experimental setup and evaluation metrics, followed by extensive comparative experiments and analyses on the VQA-v2 and GQA datasets. Section 5 concludes this work and discusses future research directions.

2. Related Work

2.1. Advances in Visual Question Answering

Visual Question Answering as a pivotal task connecting computer vision and natural language processing, has evolved from simple multimodal fusion to deep collaborative reasoning since the proposal of standard datasets by Antol et al. [2]. Early VQA methods

primarily relied on basic multimodal fusion strategies. Kim et al. [19] proposed a multimodal residual learning framework that achieved efficient fusion of visual-linguistic features through bilinear pooling, pioneering end-to-end VQA training. Yu et al. [48] further proposed the Multimodal Factorized High-order Pooling method, reducing computational complexity of bilinear fusion through low-rank constraints. These early works laid the foundation for multimodal representation learning in VQA but lacked explicit cross-modal alignment mechanisms.

The introduction of attention mechanisms marked the first major leap in VQA technology. Yang et al. [45] proposed the Stacked Attention Networks, which progressively refined semantic localization of image regions through iterative visual focusing. Anderson et al. [1] pioneered the combination of bottom-up object detection and top-down attention mechanisms, achieving object-level precise attention allocation. Rahman et al.[31] further improved reasoning quality through Attention on Attention with contextual gating mechanisms. However, these unidirectional attention methods struggle to model bidirectional interaction between questions and images. To address these limitations, co-attention mechanisms emerged. Yu et al. [47] designed the Deep Modular Co-Attention Network(MCAN), exploring deep interaction of intra-modal and inter-modal relationships through compositions of self-attention and guided-attention units. Lu et al. [24] first proposed hierarchical co-attention frameworks, achieving bidirectional guidance through alternating optimization of visual and textual attention. Peng et al. [28] proposed the Multimodal Relation Attention Network, further enhancing cross-modal relationship modeling capabilities. With the rise of Transformer architectures, large-scale pretrained VQA methods have achieved significant progress. Tan and Bansal [39] proposed LXMERT, which achieved visual-linguistic pretraining through cross-modal Transformer encoders, substantially improving downstream VQA performance. Li et al. [21] developed VisualBERT, utilizing self-attention mechanisms to mine deep associations between image regions and text. Recently, data augmentation strategies by Chen et al. [8] and dual-attention mechanisms by Mishra et al. [25] have continued to advance attention mechanisms in VQA. Eslami et al. [10] validated the effectiveness of CLIP pretrained models in domain-specific VQA. However, existing Transformer-based methods focus predominantly on global semantic alignment, remaining limited in handling local reasoning problems requiring fine-grained region perception.

2.2. Advances in Global and Local Perception

Existing VQA methods have long faced the inherent conflict between global semantic coverage and local detail focusing in attention allocation strategies [11]. Early research predominantly optimized along single dimensions: spatial attention mechanisms [20,43] strengthen local perception by localizing regions of interest, yet excessive focusing leads to loss of global context; channel attention methods [26] enhance global semantic representation but lack spatial constraints on fine-grained regions.

In recent years, researchers have increasingly recognized the necessity of balancing global and local perception. However, existing methods predominantly adopt shared-parameter or joint training strategies to simultaneously optimize global and local objectives [17,23,27,12], leading to mutual interference during gradient updates. Although Kil et al. [17] mined implicit knowledge for comprehensive representations, Liu et al. [23] employed prompt tuning to guide multi-scale attention, Peng et al. [27] revealed the

empirical importance of global-local fusion, Guo et al. [12] optimized feature balance through loss rescaling, and Niu et al. [26] separated confounding factors via causal intervention, none fundamentally decoupled perception requirements at different scales. The common limitation pervading these approaches is the implicit assumption that a single interaction pathway can uniformly accommodate multi-scale semantics, thereby forcing heterogeneous objectives to compete for the same parameter budget—a constraint equally present in co-attention frameworks [12,44]and recent compact Transformers [34,47,15,37]. More recently, the research community has shifted toward explicit multi-scale modeling and architectural decoupling. Guo et al. [13] and Liu et al. [22]validated the efficacy of decoupling visual interpretation from linguistic reasoning in mathematical problem solving and rationale generation, respectively, yet their scopes do not extend to fine-grained visual question answering. Shen et al. [32,34] further advanced this direction by demonstrating that hybrid or multi-branch architectures can effectively mitigate the inherent trade-off between global stability and local adaptability in cross-modal alignment.

Distinct from these works, the proposed DGL-Net physically separates global semantic reasoning and local fine-grained localization into independently optimized dual-branch architectures. During training, both branches optimize independently, fundamentally eliminating gradient interference inherent in shared-parameter models. During inference, global semantic constraints and local detail localization are organically unified through late fusion mechanisms. This "independent training–collaborative inference" paradigm breaks through the limitations of single pathways, providing a principled solution to the structural conflict between global and local perception during optimization and achieving more reliable multi-scale semantic modeling.

3. Method

This paper proposes the Decoupled Global-Local Collaborative Network, which achieves structural decoupling of cross-modal reasoning via explicitly separated global and local branches, with the overall architecture illustrated in Fig 2. The framework comprises three core modules: Image and Question Representation (Section 3.1), Global-Local Perception Network (Section 3.2), and Multimodal Fusion and Reasoning (Section 3.3).

3.1. Image and Question Representation

Input Processing

Input images are represented as a series of regional visual features $V \in \mathbb{R}^{m \times d}$ using bottom-up attention mechanisms, where m is the number of detected objects and d is the feature dimension. Specifically, these features are extracted from a pretrained Faster R-CNN detection model with ResNet-101 as the backbone network, pretrained on the Visual Genome dataset[20]. By setting a threshold on detection confidence scores, a dynamic number of salient regions can be obtained, with quantities ranging from 36 to 100. For the i-th detected object, its feature vector $v_i \in \mathbb{R}^d$ is obtained through mean pooling of the convolutional feature map corresponding to that region. This object-detection-based visual feature representation method focuses on semantically salient regions in images, providing structured visual input for subsequent multimodal reasoning tasks.

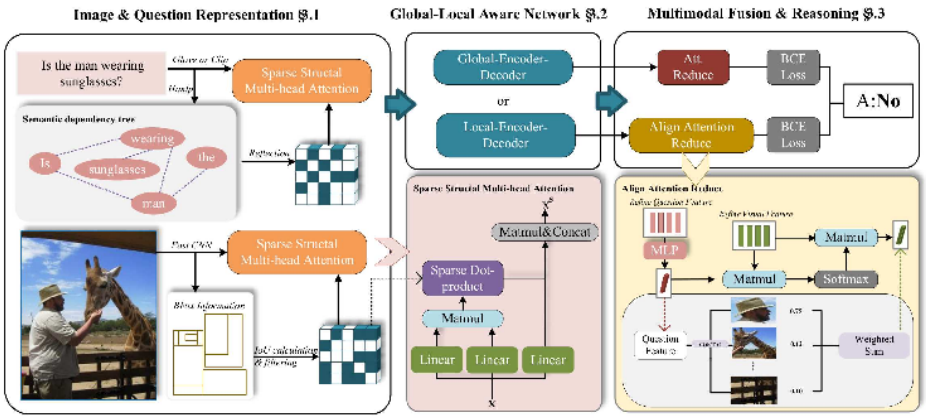


Fig. 2. Overall pipeline of the proposed DGL-Net. Structural information is employed for intra-modal feature enhancement. Two distinct branch backbones are devised for global and local understanding, respectively. Decoupling of global-local reasoning is achieved through the strategy of independent branch training and joint inference.

Question text features are obtained using the CLIP text encoder $T \in \mathbb{R}^{n\times d}$, where n is the sequence length and d is the feature dimension. Specifically, the CLIP employed in experiments utilizes a deep architecture of multi-layer Transformer blocks pretrained on large-scale corpora. Input text undergoes preprocessing and tokenization, converting to discrete token sequences, and is mapped to continuous vectors through a token embedding layer[29]. To preserve positional information between words, the model introduces learnable position encodings added to word embeddings, obtaining initial feature representations $T^{(0)} \in \mathbb{R}^{n\times d_q}$, where d_q is the initial embedding feature dimension. During encoding, a multi-level feature aggregation strategy is adopted, fusing output features from the last layer to obtain final text representations through linear layer dimension alignment.

Intra-Modal Structured Information Representation

In fine-grained local understanding, precise cross-modal alignment depends on the expressive capability of region and word features for complex semantics. However, treating regions and words as isolated feature vectors ignores their intrinsic topological structural relationships. This absence of structural information leads to subsequent self-attention and guided-attention interactions lacking explicit constraints, with attention weights easily dispersing to semantically irrelevant elements, making it difficult to form discriminative local representations. Visually, adjacent or overlapping regions typically exhibit semantic associations; in natural language, words exhibit long-distance dependencies, with semantically related words often adjacent in syntactic structures. Therefore, a mechanism is urgently needed to explicitly model sparse structural associations within modalities, providing prior topological guidance for attention interaction.

Visual Sparse Relation Construction: Given a set of image region features $\mathcal{V} = \{v_1, v_2, ..., v_n\}$, where $v_i \in \mathbb{R}^d$ is the feature vector of the i-th region and $b_i = (x_1, y_1, x_2, y_2)$ represents its bounding box coordinates. The affinity matrix $A^{\text{vis}} \in \mathbb{R}^{n\times n}$

is calculated between regions, where each entry measures the spatial overlap between bounding boxes using the Intersection-over-Union (IoU) metric: $\bar{A}_{ij}^{\text{vis}} = \text{IoU}\,(b_i, b_j)$. Subsequently, threshold τ_v is applied for binarization, constructing the visual sparse adjacency matrix:

$$S_{ij}^{v} = \begin{cases} 1, & \text{if } A_{ij} > \tau_v \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

This matrix characterizes spatial co-occurrence relationships between regions: when two regions significantly overlap in the image, their semantic association probability is high (e.g., "person" and "hat"); conversely, they are considered weakly associated or irrelevant (e.g., "giraffe" and "background fence").

Textual Sparse Relation Construction: For question text, Semantic Dependency Parsing (SDP) technology is employed to generate syntactic dependency graphs $G = (\mathcal{W}, \mathcal{E})$, where $\mathcal{W} = \{w_1, w_2, ..., w_m\}$ is the word set and $\mathcal{E}$ is the word set and $S^t \in \mathbb{R}^{m \times m}$:

$$S_{ij}^{t} = \begin{cases} 1, & \text{if } (w_i, w_j)\, or\, (w_j, w_i) \in \mathcal{E} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

This matrix captures syntactic-semantic associations between words (e.g., the verb-object relationship between "sunglasses" and "wearing"), filtering noise interference from distant vocabulary.

Structure-Aware Attention Calculation: Visual sparse matrix S^v and textual sparse matrix S^{t} are injected into corresponding modal multi-head attention mechanisms, respectively. Taking the visual branch as an example, for the h-th attention head, the query-key similarity matrix $V_h V_h^{\top} \in \mathbb{R}^{n \times n}$ is calculated, followed by application of the sparse mask:

$$M_h^{\text{v}} = \text{Softmax}\left(\frac{X_h X_h^{\top} \odot S^{\text{v}}}{\sqrt{d_k}}\right) \tag{3}$$

This operation forces attention weights to propagate only within predefined sparse structures, suppressing responses from irrelevant regions. The final output is:

$$V^{\text{sfr}} = \text{Concat}\,[M_1^{\text{v}} V_1, \ldots, M_h^{\text{v}} V_h] \cdot W \tag{4}$$

Here, W is a learnable parameter matrix. The text side follows the same paradigm, generating T^{str} . Through this process, the SSMA module explicitly encodes geometric space constraints and syntactic structure constraints into feature representations, providing a structured discriminative foundation for subsequent attention interactions.

3.2. Global-Local Aware Network

Based on the image, question, and corresponding structured features generated by the above feature representation modules, this paper further constructs global understanding and local understanding branches, responsible for holistic semantic reasoning and fine-grained region localization tasks, respectively. Both branches adopt encoder-decoder architectures, implementing hierarchical feature interaction through stacking three types of attention units. The following first introduces the design of the three attention units, then details the network structures of both branches.

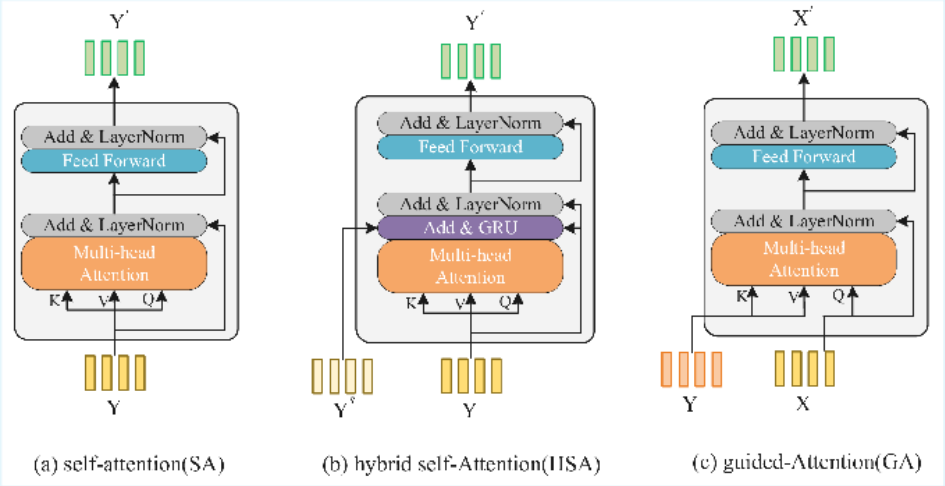


Fig. 3. Three fundamental attention units employed in the Global-Local Perception Network—Self-Attention (SA), Hybrid Self-Attention (HSA), and Guided-Attention (GA)—facilitating intra-modal feature refinement, structural information injection, and cross-modal semantic fusion, respectively.

Attention Modules

Self-Attention Unit (SA): Following the classic Transformer design[41], it utilizes multi-head attention with h parallel heads to compute intra-sequence self-dependencies, thereby refining features, as illustrated in Fig 3(a). Given input features $Y \in \mathbb{R}^{n\times d}$, for the h -th attention head, SA first generates query, key, and value matrices $Q, K, V \in \mathbb{R}^{n\times d}$ through linear projections, then executes scaled dot-product attention:

$$head_h = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) V_h \tag{5}$$

Concatenating outputs from all heads yields the complete result:

$$f = \text{Concat}\left[head_1 V_1, ..., head_h V_h\right] \cdot W \tag{6}$$

where W is a learnable weight matrix. Followed by fully connected feed-forward networks and layer normalization to obtain the final output:

$$Y' = \text{LayerNorm}\left(\text{FFN}\left(f\right)\right) \tag{7}$$

where:

$$\text{FFN}\left(f\right) = \max\left(0, f \cdot W_1\right) W_2 \tag{8}$$

Here, W_1 and W_2 are learnable weight matrices. SA addresses the limitation of traditional LSTMs in capturing long-distance dependencies between sequence elements, effectively refining intra-modal contextual semantic representations.

Hybrid Self-Attention Unit (HSA): Addressing the structural sensitivity requirements of local understanding tasks, we propose the Hybrid Self-Attention (HSA) unit, as illustrated in Fig 3(b), which incorporates structural features $Y^{str} \in \mathbb{R}^{n \times d}$ on the basis of standard self-attention, introducing GRU to control the injection ratio of structural information into attention outputs:

$$f = \mathrm{GRU}\left(\mathrm{SA}\left(Y\right) + Y^{str}\right) \tag{9}$$

Modeling its sequence dependencies and potential multi-semantic expressive capabilities through multi-layer recurrence enhances the model's perception capability of key topological relationships. After further processing through fully connected feed-forward networks and layer normalization, more discriminative local representations are obtained.
Guided-Attention Unit (GA): Following the Bottom-Up Top-Down Attention cross-modal interaction paradigm[1], GA implements bidirectional semantic guidance between vision and language modalities,as illustrated in Fig 3(c). Taking visual-guided text as an example, for the h -th attention head, given query question features $T \in \mathbb{R}^{m \times d}$ and guiding visual features $V \in \mathbb{R}^{n \times d}$, GA computes text-to-visual guided attention:

$$head_h = \mathrm{Softmax}\left(\frac{TW_Q\left(VW_K\right)^\top}{\sqrt{d}}\right)V \tag{10}$$

Concatenating results from all heads followed by fully connected feed-forward networks and layer normalization yields the final output.GA achieves efficient alignment and fusion of cross-modal information by actively retrieving visual content with question features as queries, or constraining text reasoning with visual features as queries.

Stacked Encoders and Decoders

To fully excavate hierarchical semantic associations of multimodal features, this paper adopts competitive encoder-decoder architectures to construct global and local understanding branches. This architecture gradually refines intra-modal high-order features and achieves deep cross-modal interaction through hierarchical stacking of attention units.
Global Understanding Encoder-Decoder: As illustrated in Fig 4(a), the global branch adopts a stacked structure combining pure self-attention and guided attention. The encoder consists of L layers of SA units, processing visual features V and textual features T layer by layer:

$$X^{(l)} = \mathrm{SA}\left(X^{(l-1)}\right) \tag{11}$$

The encoder captures potential semantic dependencies through multi-layer self-interaction, gradually establishing global semantic representations. The decoder similarly contains L layers, each composed of SA units cascaded with GA units, implementing cross-modal bidirectional guidance:

$$X^{(l)} = \mathrm{GA}\left(\mathrm{SA}\left(X^{(l-1)}\right), Y^{(L)}\right) \tag{12}$$

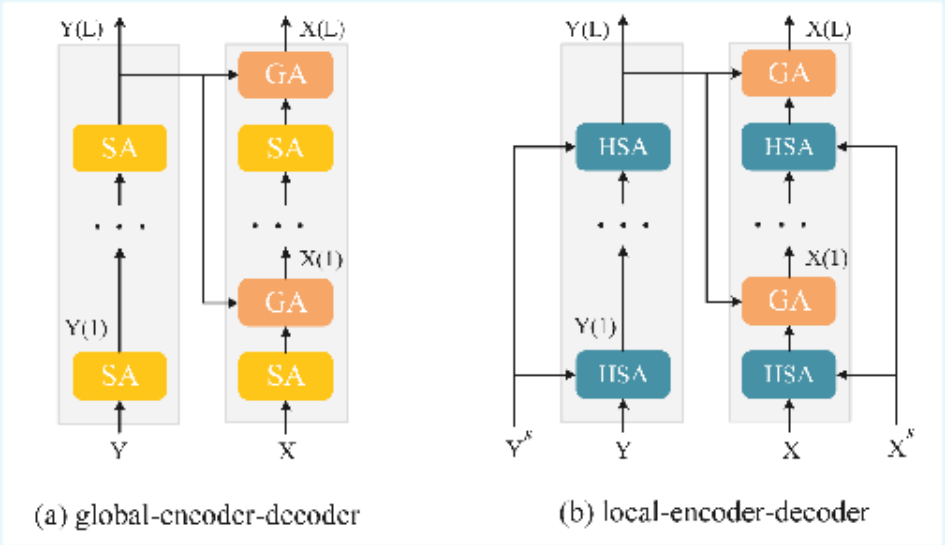


Fig. 4. Illustration of the encoder-decoder architectures in the Global-Local Perception Network.

where $X^{(L)}$ is the top-level feature output by the encoder. Through cross-guidance of GA units, the decoder maintains intra-modal consistency while achieving efficient alignment and fusion of visual-linguistic semantics.

Local Understanding Encoder-Decoder: As illustrated in Fig 4(b),the local branch replaces SA units with HSA units to continuously inject sparse structure priors. Each encoder layer receives parallel inputs of dynamic features and static structure features:

$$X^{(l)} = \text{HSA}\left(X^{(l-1)}, X^{str}\right) \tag{13}$$

where X^{str} represents sparse structural features generated by SSMA, encoding spatial positional relationships between visual regions and semantic dependency relationships between text words, respectively. The decoder adopts an alternating stacked structure of HSA and GA, maintaining structural constraints during cross-modal interaction:

$$X^{(l)} = \text{GA}\left(\text{HSA}\left(X^{(l-1)}, X^{str}\right), Y^{(L)}\right) \tag{14}$$

Notably, throughout the forward propagation process, structural featuresX^{str} remain constant in the network and are not associated with layer numbers. This design is based on the fact that structural information originates from inherent topological attributes of the original semantic space: visual end from physical spatial relationships determined by object bounding box geometric positions, textual end from grammatical associations defined by dependency syntax graphs—these relationships are statically determined by data prior structures and do not change with high-level feature abstract transformations. Therefore, structural features serve as cross-layer shared constant prior information, providing stable structural constraints for attention calculations at each layer, avoiding topological

drift during dynamic feature learning, and ensuring that the local understanding branch always follows original semantic associations for discriminative feature refinement.

3.3. Modal Fusion and Reasoning

After deep co-attention learning, the two branches adopt different feature compression strategies for answer reasoning: the global branch captures overall semantic profiles through adaptive attention pooling, while the local branch focuses on key local regions through explicit similarity calculation and low-temperature sharpening mechanisms.

Global Understanding Branch: This branch aims to encode the overall semantic layout of scenes. For output visual features $V^{(L)} \in \mathbb{R}^{m\times d}$ and textual features $T^{(L)} \in \mathbb{R}^{n\times d}$, MLP-based attention compression models are employed to aggregate them into global vectors. Taking the visual end as an example, importance weights for each region are first calculated: $a = \mathrm{Softmax}\left(\mathrm{MLP}\left(X^{(L)}\right)\right), \quad a \in \mathbb{R}^m$, where the MLP structure is $\mathrm{FC}\,(d) \rightarrow \mathrm{ReLU} \rightarrow \mathrm{Dropout}\,(0.1) \rightarrow \mathrm{FC}\,(1)$. Subsequently, global visual representations are generated through soft weighted summation:

$$\bar{x}_g = \sum_{i=1}^{m} \alpha_i x_i^{(L)} \tag{15}$$

This mechanism adaptively aggregates multi-region information, suitable for comprehensive reasoning dependent on holistic scene cognition. Similarly, global textual representations $\bar{y}_g$ are obtained. Finally, answer distributions are generated through linear fusion and classification heads:

$$Z_g = \mathrm{LayerNorm}\left(W_x^\top \bar{x}_g + W_y^\top \bar{y}_g\right), \quad score_g = \sigma\left(W_p^\top Z_g\right) \tag{16}$$

Local Understanding Branch: Unlike the distributed aggregation of the global branch, the local branch aims to precisely anchor a few key regions highly relevant to the question through explicit semantic similarity screening, suppressing background noise. First, textual features are compressed into holistic representations $\bar{y}_l$ throughMLP attention models. Subsequently, cosine similarity with each visual region feature $v_i^{(L)}$ is calculated to measure semantic relevance:

$$s_i = \frac{\bar{y}_l^\top v_i^{(L)}}{|\bar{y}_l||v_i^{(L)}|}, \quad i = 1, 2...m \tag{17}$$

To achieve sharpened selection of a few highly relevant regions, a temperature coefficient τ is introduced for low-temperature Softmax scaling :

$$\alpha_i = \frac{\exp\left(s_i/\tau\right)}{\sum_{j=1}^{m} \exp\left(s_j/\tau\right)} \tag{18}$$

When $\tau \rightarrow 0$, this distribution approaches one-hot form, retaining only regions most relevant to the question, achieving a soft approximation of hard selection. Final aggregated local visual representation: We set $\tau = 0.10$ as the default temperature, which achieves a favorable balance between selective precision and semantic completeness.

$$\bar{x}_l = \sum_{i=1}^{m} \alpha_i x_i^{(L)} \tag{19}$$

Finally, answer distributions are generated through linear fusion and classification heads:

$$Z_l = \text{LayerNorm}\left(W_x^\top \bar{x}_l + W_y^\top \bar{y}_l\right), \quad score_l = \sigma\left(W_p^\top Z_l\right) \tag{20}$$

We employ binary cross-entropy loss functions to optimize the distribution of predicted answer scores:

$$\mathcal{L} = -\sum_{j=1}^{N} \left[t_j \log\left(score_j\right) + \left(1 - t_j\right) \log\left(1 - score_j\right)\right] \tag{21}$$

Our strategy involves independently training classifiers on two understanding branches from different perspectives. Through arithmetic averaging operations, the model achieves adaptive integration of dual-branch predictions without introducing additional fusion parameters or complex gating networks, reducing overfitting risks while improving inference robustness. The answer corresponding to maximum confidence in $score_{joint} = score_g + score_l$ is selected as the system prediction output.

4. Experiments

4.1. Datasets

Unlike research paradigms relying on large-scale pretrained models to extract generic representations, this paper adopts end-to-end training strategies, directly optimizing visual-linguistic feature alignment on VQA-specific data distributions. This strategy offers three advantages: First, feature extractors jointly optimized for the VQA task can more precisely capture question-oriented visual semantic associations, avoiding domain gaps between features learned by pretrained models on generic corpora and downstream task requirements. Second, end-to-end training enables the model to adaptively learn dataset-specific statistical characteristics, reducing dependence on language priors that may exist in pretraining corpora. Finally, this paradigm enables lightweight architecture designs, achieving competitive performance without loading massive pretrained weights. To ensure fair comparison and validate model effectiveness, this paper selects two mainstream benchmarks: VQA 2.0 and GQA.

VQA 2.0 [11] and GQA [16] are employed as the primary evaluation benchmarks. Derived from MS COCO, VQA 2.0 comprises 1.1 million natural language questions with standard training, validation, and test splits. Its complementary image-pair design mitigates language priors by enforcing image-grounded reasoning, and questions are categorized as “Yes/No”, “Number”, or “Other”. Accuracy is reported on the official test-dev and test-standard sets.

Built upon Visual Genome scene graphs, GQA encompasses 113K images and 22 million structured questions targeting compositional reasoning and relational understanding. The dataset emphasizes multi-step inference and alleviates language bias through

Table 1. Hyperparameter settings for DGL-Net training and inference.

Hyperparameter	Value	Description
Hidden dimension d	512	Model embedding size
Attention heads h	8	Parallel self-attention heads
Head dimension	64	d/h
Encoder/Decoder layers L	6	Stacked layer depth
FFN structure	$\mathrm{FC}(4d) \rightarrow \mathrm{ReLU} \rightarrow \mathrm{Dropout}(0.1) \rightarrow \mathrm{FC}(d)$	Feed-forward network
MLP compression	$\mathrm{FC}(d) \rightarrow \mathrm{ReLU} \rightarrow \mathrm{Dropout}(0.1) \rightarrow \mathrm{FC}(1)$	Attention reduction
Visual IoU threshold τ_v	0.4	Sparse adjacency construction
Visual regions N	36–100	Dynamic, confidence-thresholded
Text parser	HanLP (Biaffine)	Semantic dependency parsing
Optimizer	AdamW	$\beta_1 = 0.9, \beta_2 = 0.999$
Weight decay	10^{-4}	L2 regularization
Batch size	64	Training samples per step
Loss function	Binary cross-entropy	Multi-label answer classification
Warmup epochs	4	Linear LR warmup
Initial learning rate	2.5×10^{-5}	Warmup starting value
Peak learning rate	1×10^{-4}	Post-warmup maximum
Early stopping	3 epochs	Patience on validation accuracy
Gradient clipping	0.25	Max gradient norm threshold
Dropout rate	0.1	Applied to all linear projections

balanced answer distributions. Beyond standard accuracy, GQA evaluates Consistency, Plausibility, Validity, and Distribution to comprehensively measure reasoning robustness. We utilize the balanced training and validation sets for model development, with all metrics reported on the test set.

4.2. Experimental Setup

All experiments are conducted on NVIDIA L40 24GB GPUs. The complete hyperparameter configurations are summarized in Table 1, ensuring full reproducibility of our results.The semantic dependency trees are parsed using the HanLP toolkit pretrained on SemEval-2015 Task 18 English corpora. Visual region quantities are dynamically adjusted based on Faster R-CNN detection confidence scores.

To ensure reproducibility and fair comparison, the two branches follow a strictly decoupled training protocol. Both branches share the same data loader and batch order to guarantee identical training distributions, yet maintain completely separate optimizer states and parameter gradients. Specifically, each branch is assigned an independent AdamW optimizer with identical initial hyperparameters including peak learning rate, warmup schedule, and weight decay, but their gradient updates are computed and applied in isolation. Backpropagation from the global branch loss never reaches local branch parameters, and vice versa. The two branches are trained in parallel on separate GPUs for

the same number of epochs, with early stopping determined independently based on each branch's validation accuracy. After both branches converge, their parameters are frozen and only the arithmetic averaging operation is applied during joint inference. This protocol ensures that each branch optimizes its specific objective without gradient interference, while maintaining strictly comparable training conditions with single branch baselines.

4.3. Model Performance Comparison

To comprehensively validate the effectiveness of the proposed method, we systematically compare DGL-Net and its single-branch variants DGL-Global and DGL-Local with representative methods in the visual question answering field in recent years, with experimental results shown in Table 2. All comparison experiments are conducted on the official test sets of VQA 2.0 dataset (test-dev and test-standard) to ensure evaluation fairness and reproducibility.

Table 2. Performance comparison on the test-dev and test-std splits.

Method	Image Feature	Test-dev (%)				Test-std
		Yes/No	Number	Other	All	All
MCAN('19)[47]	Region	87.08	53.16	60.66	70.69	-
CAAN('22)[3]	Region	87.09	53.37	61.13	70.94	71.31
CLVIN('23)[6]	Region	87.09	53.65	60.89	70.86	71.16
RRAF('24)[35]	Region	87.03	55.39	60.95	71.06	71.33
VMAN('24)[37]	Region	86.95	53.28	61.16	70.89	71.34
MPCCT('24)[4]	Region	86.96	55.57	61.27	71.20	-
$RMCB_G$('25)[5]	Region	87.26	55.21	61.34	71.32	-
VWAN('24)[37]	Grid	87.63	53.85	61.89	71.58	-
LCRN('24)[14]	Grid	87.74	53.99	62.29	71.83	72.12
GFSNet('25)[33]	Grid	87.78	54.91	61.94	71.79	71.99
DGL-Global	Region	87.46	53.95	60.96	71.08	-
DGL-Local	Region	87.60	53.49	60.80	71.01	-
DGL-Net	Region	87.99	54.42	61.75	71.73	72.15

Comparison with single-branch baselines shows that DGL-Net significantly outperforms single global or local branches through independent training and joint inference strategies. Specifically, compared with DGL-Global and DGL-Local, DGL-Net brings absolute performance improvements of 0.65% and 0.72% on test-dev, respectively, validating the effectiveness of the dual-branch decoupled collaborative mechanism. This improvement primarily stems from the complementary fusion of scene contextual constraints provided by the global branch and fine-grained region localization capabilities contributed by the local branch. Performance comparison with recent methods shows that DGL-Net achieves 72.15% accuracy on test-standard, outperforming compared contemporary methods including LCRN and GFSNet based on Grid features. Notably, both

LCRN and GFSNet rely on grid-based visual features, while DGL-Net achieves comparable generalization performance using only regional features, indicating that the proposed decoupled collaborative paradigm has advantages in feature utilization efficiency. On test-dev, DGL-Net achieves 71.73% accuracy, demonstrating strong competitiveness among region-based methods. Specifically, compared with recently published region-based methods RMCBG and VMAN, DGL-Net improves by 0.41% and 0.84%, respectively. Particularly, on "Other" categories representing open-ended questions requiring fine-grained reasoning, DGL-Net achieves 61.75%, surpassing RMCBG, indicating that the local understanding branch's modeling capability for fine-grained semantic associations significantly enhances the model's open-domain reasoning performance. Comprehensive comparison shows that as an end-to-end trained region-based method, DGL-Net's performance approaches current state-of-the-art grid-based methods. Considering the advantages of region features in computational efficiency and interpretability, this method demonstrates strong competitiveness in the trade-off between accuracy and practicality. Furthermore, DGL-Net achieves competitive scores on both "Yes/No" and "Number" categories, indicating that the decoupled architecture can effectively handle heterogeneous question types, validating the model's robustness under different reasoning modes.

To further validate the model's compositional reasoning capability and generalization, supplementary experiments are conducted on the GQA benchmark. As shown in Table 3, DGL-Net demonstrates advantages on multiple evaluation metrics, particularly in open-ended questions requiring complex reasoning and logical consistency.

Table 3. Performance comparison on multiple evaluation metrics.

Method	Binary	Open	Consistency	Validity	Plausibility	Distribution	Accuracy
Human[16]	87.12	90.40	99.40	99.20	97.60	-	89.30
BAN[18]	73.31	41.13	84.64	96.77	85.58	-	56.19
DMFNet[49]	73.98	41.86	86.98	97.62	84.87	-	57.05
LXMERT[39]	75.33	41.89	91.00	95.36	83.5	6.42	60.83
MAGM[43]	70.33	40.65	88.05	96.17	84.22	6.43	57.57
SGCN[44]	68.20	32.80	83.49	95.90	84.10	8.00	54.56
SelfGraphVQA[38]	68.20	32.80	64.30	95.00	90.00	-	54.5
LION[7]	-	-	-	-	-	-	51.56
DGL-Global	76.23	40.84	87.25	96.49	84.94	1.65	57.08
DGL-Local	76.01	40.90	87.16	96.37	84.88	1.7	57.01
DGL-Net	76.56	42.01	87.55	97.12	85.72	1.82	57.87

DGL-Net achieves 42.01% on the Open metric. It is worth noting that LXMERT[33] and DMFNet [46] employ large-scale vision-language pretraining followed by task-specific fine-tuning, whereas DGL-Net is trained end-to-end from scratch on VQA-specific data. The comparison herein serves as a reference rather than direct competition: it demonstrates that the proposed decoupled architecture, despite using significantly fewer pre-training resources, achieves competitive compositional reasoning performance through

Table 4. Comparison of parameter count (Params), computational complexity (FLOPs), and inference speed of DGL-Net and its single-branch variants on the VQA 2.0 test-dev set.

Models	Params (M)	FLOPs (G)	Time (ms/img)	All (%)
baseline	51.61	3.67	2894	70.75
DGL-Global	52.58	3.81	2931	71.08
DGL-Local	75.70	7.09	3803	71.01
DGL-Net	-	-	-	71.73

explicit global-local collaborative mechanisms. This highlights the efficacy of structural decoupling as a lightweight alternative to massive pretraining for fine-grained reasoning tasks.Meanwhile, the model achieves leading performance on Validity and Plausibility, with the global branch's overall scene constraints and local branch's precise semantic localization synergistically reducing logical conflicts during reasoning, validating the advantages of the dual-branch architecture in generating logically reasonable and semantically credible answers. DGL-Net achieves 57.87% accuracy through end-to-end training strategies, demonstrating excellent performance among non-pretrained methods. Compared with single-branch baselines, DGL-Net improves by 0.79% and 0.86% over DGL-Global and DGL-Local, respectively, while the Consistency metric increases from approximately 87.2% to 87.55%, further corroborating the effectiveness of the decoupled collaborative training strategy in improving model reasoning consistency. These results indicate that even without dependence on large-scale pretraining, the proposed decoupled architecture can still achieve competitive compositional reasoning performance through explicit global-local collaborative mechanisms.

Table 4 compares parameter counts, computational complexity, and inference speed of the single-branch baselines. DGL-Net is not independently listed because its inference mechanism introduces no additional fusion parameters or gating networks; the final output derives from simple arithmetic averaging of the two independently trained branch predictions. Thus, the overall complexity is fully determined by the two decoupled branches without extra overhead from joint training or fusion modules. The global branch incurs modest computational increase relative to the baseline, while the local branch demands higher resources owing to its structural sparse attention and hybrid self-attention components. Despite the dual-branch architecture, the collaborative inference achieves absolute improvements of 0.65% and 0.72% over the global and local branches respectively, representing a favorable accuracy-efficiency trade-off. Relative to large-scale pretrained methods, DGL-Net maintains a significantly lighter footprint, underscoring the practical viability of structural decoupling for resource-constrained deployment.

4.4. Ablation Studies

To deeply analyze the contribution of each module to model performance, systematic ablation experiments are conducted on VQA 2.0 test-dev, with results shown in Table 5.

Module Ablation Experiments

Table 5. Ablation study results on the test-dev split.

Methods	Test-dev			
	Yes/No	Number	Other	All
DGL-Global (GED+AR)	87.46	53.95	60.96	71.08
DGL-Global w/ AAR	87.30	53.14	60.79	70.85
DGL-Local (LED+AAR)	87.60	53.49	60.80	71.01
DGL-Coupled (Multi-task)	87.54	53.77	60.67	70.95
DGL-Net	87.99	54.42	61.75	71.73

First, experimental settings are explained: DGL-Global (GED+AR) represents the global understanding branch using only global encoder-decoder and MLP attention compression; DGL-Global w/AAR represents the variant where alignment attention compression from the local understanding branch is incorrectly applied to the global architecture; DGL-Local (LED+AAR) represents the local understanding branch using only local encoder-decoder and alignment attention compression; DGL-Coupled represents the coupled paradigm where dual branches adopt multi-task joint training; DGL-Net represents the complete decoupled collaborative network with independent training and joint inference proposed in this paper.

Using only the global branch or local branch alone achieves competitive baseline performance, but both achieve only 60.96% and 60.80% on "Other" open-ended questions, respectively, with overall accuracy failing to break through 71.1%. This indicates inherent bottlenecks in single-dimensional reasoning capabilities: while the global branch excels at capturing holistic scene semantics, it lacks precise localization of fine-grained regions; while the local branch can screen key regions, it may be constrained by neglecting overall context.

To validate the compatibility between attention compression mechanisms and network architectures, we apply the local branch's alignment attention compression to the global architecture. Results show that this mismatched scheme performance drops to 70.85%, a decrease of 0.23% compared with DGL-Global, with particularly significant losses in the "Number" category, dropping from 53.95% to 53.14%. This indicates that cosine similarity-based sharpening selection mechanisms are more suitable for local reasoning requiring precise region screening, while MLP-based adaptive pooling better satisfies global scene summarization needs, validating the importance of component adaptation in divide-and-conquer strategies.

Comparison of dual-branch training paradigms shows that DGL-Coupled is significantly lower than DGL-Net, with a performance gap of 0.78%. This result indicates that simple parameter sharing or multi-task optimization leads to mutual interference between the two branches during gradient updates, making it difficult to balance heterogeneous optimization objectives for global and local features. In contrast, DGL-Net's independent

training strategy enables each branch to focus on capability learning in specific dimensions during optimization, achieving complementary advantages through late fusion and effectively avoiding optimization conflicts.

Comprehensive results show that DGL-Net improves by 0.79% over the best single branch in the "Other" category and 0.47% in the "Number" category, validating the gains of the dual-branch collaborative mechanism for complex reasoning tasks. Particularly in open-ended questions requiring fine-grained attribute recognition, the complementary fusion of global and local representations significantly enhances the model's discriminative capability. These ablation results fully demonstrate the rationality and effectiveness of each component design in the proposed decoupled global-local collaborative architecture.

Table 6. Threshold ablation.

Parameter	Test-dev			
	Yes/No	Number	Other	All
τ_v=0.2	87.48	53.29	60.92	71.00
τ_v=0.4	87.60	53.49	60.80	71.01
τ_v=0.6	87.47	53.65	60.61	70.89
τ_v=0.8	86.77	53.99	60.85	70.75

Threshold Selection Ablation

To validate the rationality of the region association discrimination threshold in visual sparse adjacency matrix construction, ablation experiments are conducted on VQA 2.0 test-dev, with results shown in Table 6.

Experimental results show that the model achieves optimal overall performance when $\tau_v = 0.4$. When $\tau_v = 0.2$, the threshold is too low, causing the sparse matrix to contain too many weakly associated regions, introducing structural noise and degrading overall performance. Conversely, when $\tau_v = 0.6$ or $\tau_v = 0.8$, the threshold is too high, causing excessive sparsity that destroys local structure integrity, resulting in decreased overall accuracy. Notably, the counting category shows a monotonically increasing trend with threshold elevation. This indicates that for counting problems requiring precise enumeration of specific objects, high thresholds can effectively filter out core regions highly relevant to query targets, suppressing duplicate counting risks brought by overlapping regions or background interference. However, this high sparsity comes at the cost of overall structure perception: when $\tau_v = 0.8$, "Yes/No" category performance significantly decreases, and the "Other" category shows obvious decline at the 0.6 threshold, indicating that extreme sparsity damages the model's reasoning capability for complex semantic relationships.

Comprehensive analysis shows that $\tau_v = 0.4$ maintains appropriate structural sparsity while preserving necessary inter-region semantic associations, achieving optimal balance across "Yes/No", "Number", and "Other" question types. Therefore, this paper fixes

Table 7. Sensitivity analysis of the temperature coefficient on VQA 2.0 test-dev.

τ	Yes/No	Number	Other	All
0.01	87.41	53.66	60.61	70.95
0.10	87.60	53.49	60.80	71.01
0.50	87.36	53.37	60.68	70.88
1.00	86.87	53.14	60.72	70.79

$\tau_v = 0.4$ as the default threshold for visual sparse structure construction in subsequent experiments.

Sensitivity Analysis of Temperature Coefficient τ

The temperature coefficient τ modulates the sharpness of the local Softmax distribution in the region screening module. Its empirical impact is systematically assessed on the VQA 2.0 test-dev split by varying τ over $\{0.01, 0.1, 0.5, 1.0\}$ while all other hyperparameters remain at their defaults. The results are reported in Table 7.

The results indicate a non-monotonic response of the local branch to τ, with $\tau = 0.10$ yielding the optimal overall accuracy of 71.01%. At $\tau = 0.01$, aggressive sharpening suppresses duplicate background counts, marginally improving the "Number" metric to 53.66% while degrading "Yes/No" to 87.41% and "Other" to 60.61% due to excessive contextual truncation. As τ increases beyond 0.10, the Softmax distribution progressively flattens: "Yes/No" drops monotonically to 86.87% at $\tau = 1.00$, while "Other" converges toward the single-branch baseline, indicating loss of discriminative screening. Accordingly, we adopt $\tau = 0.10$ as the default, achieving the best compromise between selective precision and context preservation.

4.5. Visualization Analysis

To intuitively demonstrate the reasoning mechanism differences between the two branches, this paper conducts visualization analysis on cross-modal attention weights generated by the final decoder layer of both branches. Fig 5 shows attention heatmap comparisons between Global Awareness (global understanding branch) and Local Awareness (local understanding branch) in typical samples, revealing significant division of labor between the two in semantic attention levels.

For the first image with question "Is the man wearing sunglasses?", the global understanding branch's attention mainly concentrates on the salient subject "man", with heatmaps showing high response values uniformly distributed across region rows and word columns representing this target, reflecting its macroscopic grasp of scene dominant objects. However, this holistic perception mode cannot precisely anchor the fine-grained attribute "sunglasses", causing key details to be submerged in strong responses from the subject object. In contrast, the local understanding branch focuses high weights on specific regions corresponding to "sunglasses" and the semantic association of "wearing" through sparse structure constraints and alignment attention mechanisms, showing selective enhancement of local details, validating its localization precision in fine-grained

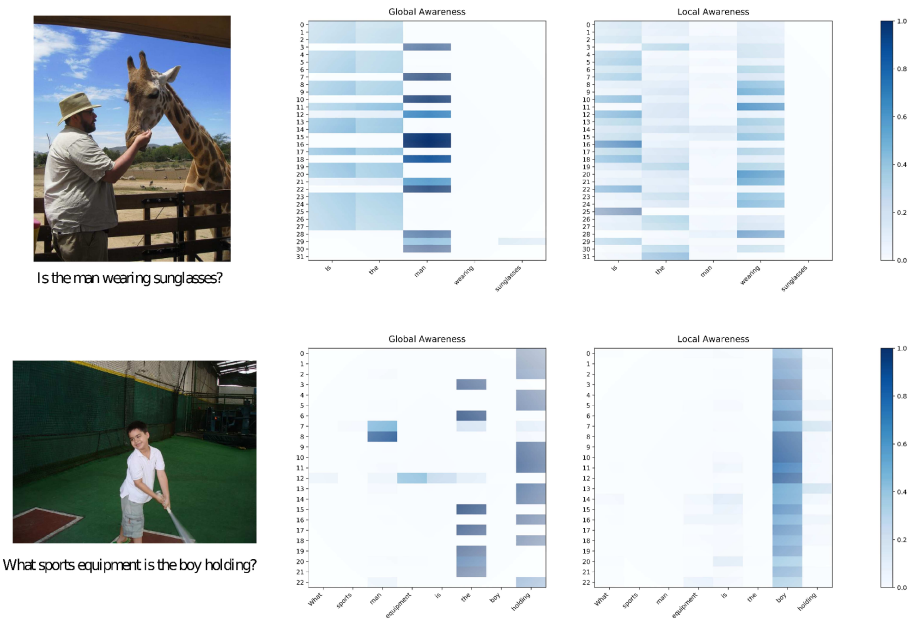


Fig. 5. Visualization of attention heatmaps.

attribute recognition. For the second image with question “What sports equipment is the boy holding?”, both branches show more significant semantic hierarchy differences. The global understanding branch shows strongest response to the verb “holding”, focusing on overall scene semantics constituted by action relationships; while the local understanding branch simultaneously focuses on two noun entities “boy” and “equipment”, supporting answer inference through recognition of specific objects and their attributes. This difference reflects the complementarity between the global branch’s summarization capability for high-order semantic relationships such as actions and scene atmospheres, and the local branch’s precise localization capability for basic visual concepts such as object categories and spatial attributes.

Attention visualization results indicate that global and local understanding branches indeed learn heterogeneous attention allocation strategies: the former tends to capture dominant semantics and relationship structures in overall scenes, while the latter focuses on fine-grained regions and entity concepts directly related to queries. Collaborative reasoning achieved through late fusion effectively compensates for limitations of single perception modes, providing more robust decision foundations for complex visual question answering tasks.

Furthermore, to validate the collaborative mechanism, qualitative analysis is conducted on diverse question sets for the same image. As shown in Fig 6, joint reasoning achieves averaging fusion of dual-branch predictions: when one branch has high confidence in the correct answer, its advantage can effectively mask the near-random low-confidence distribution produced by the other branch due to understanding failure,

thereby synthesizing dual-branch understanding capabilities. Specifically, for fine-grained questions such as “Is the man wearing sunglasses?”, the local branch has explicit high-confidence predictions for correct answers, while although the global branch produces incorrect judgments, its error distribution has relatively low confidence; after averaging, correct answers still dominate. Conversely, for global reasoning questions such as “Does the giraffe appear dangerous?”, the global branch’s high confidence in “No” successfully suppresses the local branch’s erroneous low-confidence responses caused by excessive attention to local textures. This “high-confidence dominance, low-confidence dilution” fusion mechanism enables joint reasoning to stably inherit advantages from single branches across different question types.

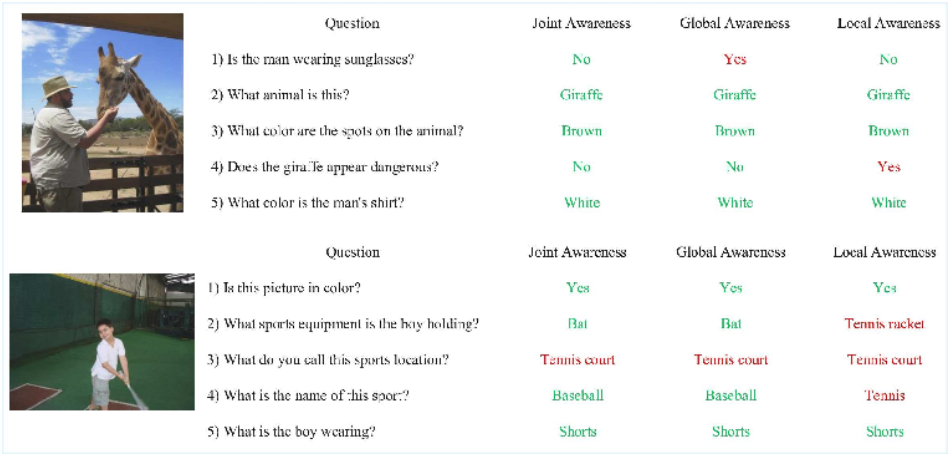


Fig. 6. Qualitative comparison of the joint model and its decoupled branches.

However, when both branches produce errors or when original visual features have inherent ambiguities, joint reasoning cannot overcome cognitive limitations. For example, for “What do you call this sports location?”, all three branches incorrectly judge it as “Tennis court”, indicating that underlying feature extraction fails to distinguish visual similarities between baseball fields and tennis courts. At this point, regardless of fusion strategy, correct semantics cannot be recovered from ambiguous original representations. Failure cases reveal the current method’s dependence on visual feature quality. In the future, more powerful visual backbone networks can be introduced or visual feature adaptive enhancement mechanisms can be explored, dynamically adjusting feature extraction granularity at the encoder front-end according to question types rather than relying on fixed region proposals, to further enhance model reasoning robustness.

5. Conclusion

This paper proposes a Decoupled Global-Local Collaborative Network for Visual Question Answering. Addressing the structural conflict between global semantics and local

fine-grained perception in VQA, it achieves decoupled modeling of cross-modal reasoning through explicit dual-branch architectures. This method breaks through the limitations of traditional single-path attention mechanisms that trade off global and local objectives within single parameter spaces, constructing independently optimized global understanding and local understanding branches: the former captures holistic scene semantics through hierarchical attention fusion, while the latter achieves precise localization of key regions based on sparse structural multi-head attention and alignment compression mechanisms. The two branches train independently to avoid gradient interference, with late fusion achieving complementary advantages. Distinguished from existing methods that indirectly balance global and local information through multi-task learning within shared parameter spaces, DGL-Net fundamentally eliminates optimization conflicts through physical separation of parameter spaces. Experimental results demonstrate that this paradigm achieves competitive performance on VQA 2.0 and GQA benchmarks, with particular advantages in fine-grained reasoning tasks.

However, the proposed method still exhibits two intrinsic limitations. First, the static arithmetic averaging fusion ignores the dynamic correlation between branch predictive uncertainties. When one branch produces high-entropy distributions due to ambiguous inputs, the fixed fusion weight cannot adaptively suppress noisy contributions, rendering the model prone to erroneous consensus. Second, the reliance on pre-extracted visual features limits the model's generalization capability. When both branches simultaneously err owing to inherently ambiguous visual features, the late-fusion mechanism cannot recover correct semantics from corrupted representations.In light of these limitations, future research will focus on two core directions. First, addressing the static fusion bottleneck, we will explore uncertainty-driven adaptive fusion paradigms by establishing dynamic evaluation mechanisms for branch confidence, enabling fusion weights to be adaptively regulated according to input reliability, thereby transcending the robustness ceiling of fixed fusion strategies. Second, addressing the fixed backbone dependency, we will advance toward end-to-end multi-scale visual representation learning, investigating mechanisms that dynamically adjust feature extraction granularity and receptive field scope in accordance with reasoning task characteristics, so as to reduce reliance on pretrained detectors while enhancing the generalization and discriminability of visual inputs.

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 6077–6086. Computer Vision Foundation / IEEE Computer Society (2018), http://openaccess.thecvf.com/content_cvpr_2018/html/Anderson_Bottom-Up_and_Top-Down_CVPR_2018_paper.html
2. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: visual question answering. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015. pp. 2425–2433. IEEE Computer Society (2015), https://doi.org/10.1109/ICCV.2015.279
3. Chen, C., Han, D., Chang, C.: CAAN: context-aware attention network for visual question answering. Pattern Recognit. 132, 108980 (2022), https://doi.org/10.1016/j.patcog.2022.108980

4. Chen, C., Han, D., Chang, C.: MPCCT: multimodal vision-language learning paradigm with context-based compact transformer. Pattern Recognit. 147, 110084 (2024), https://doi.org/10.1016/j.patcog.2023.110084
5. Chen, C., Han, D., Guo, Z., Chang, C.: Towards bias-aware visual question answering: Rectifying and mitigating comprehension biases. Expert Syst. Appl. 264, 125817 (2025), https://doi.org/10.1016/j.eswa.2024.125817
6. Chen, C., Han, D., Shen, X.: CLVIN: complete language-vision interaction network for visual question answering. Knowl. Based Syst. 275, 110706 (2023), https://doi.org/10.1016/j.knosys.2023.110706
7. Chen, G., Shen, L., Shao, R., Deng, X., Nie, L.: LION : Empowering multimodal large language model with dual-level visual knowledge. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 26530–26540. IEEE (2024), https://doi.org/10.1109/CVPR52733.2024.02506
8. Chen, L., Zheng, Y., Xiao, J.: Rethinking data augmentation for robust visual question answering. In: Avidan, S., Brostow, G.J., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVI. pp. 95–112. Lecture Notes in Computer Science, Springer (2022), https://doi.org/10.1007/978-3-031-20059-5_6
9. Chowdhury, S., Soni, B.: Handling language prior and compositional reasoning issues in visual question answering system. Neurocomputing 635, 129906 (2025), https://doi.org/10.1016/j.neucom.2025.129906
10. Eslami, S., Meinel, C., de Melo, G.: Pubmedclip: How much does CLIP benefit visual question answering in the medical domain? In: Vlachos, A., Augenstein, I. (eds.) Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, May 2-6, 2023. pp. 1151–1163. Findings of ACL, Association for Computational Linguistics (2023), https://doi.org/10.18653/v1/2023.findings-eacl.88
11. Goyal, Y., Khot, T., Agrawal, A., Summers-Stay, D., Batra, D., Parikh, D.: Making the V in VQA matter: Elevating the role of image understanding in visual question answering. Int. J. Comput. Vis. 127(4), 398–414 (2019), https://doi.org/10.1007/s11263-018-1116-0
12. Guo, Y., Nie, L., Cheng, Z., Tian, Q., Zhang, M.: Loss re-scaling VQA: revisiting the language prior problem from a class-imbalance view. IEEE Trans. Image Process. 31, 227–238 (2022), https://doi.org/10.1109/TIP.2021.3128322
13. Guo, Z., Liu, M., Ji, Z., Bai, J., Zhang, L., Zuo, W.: Decoupled visual interpretation and linguistic reasoning for math problem solving. CoRR abs/2505.17609 (2025), https://doi.org/10.48550/arXiv.2505.17609
14. Han, D., Shi, J., Zhao, J., Wu, H., Zhou, Y., Li, L., Khan, M.K., Li, K.: LRCN: layer-residual co-attention networks for visual question answering. Expert Syst. Appl. 263, 125658 (2025), https://doi.org/10.1016/j.eswa.2024.125658
15. He, H., Choi, J.D.: The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders. In: Moens, M., Huang, X., Specia, L., Yih, S.W. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021. pp. 5555–5577. Association for Computational Linguistics (2021), https://doi.org/10.18653/v1/2021.emnlp-main.451
16. Hudson, D.A., Manning, C.D.: GQA: A new dataset for real-world visual reasoning and compositional question answering. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 6700–6709. Computer Vision Foundation / IEEE (2019), http://openaccess.thecvf.com/content_CVPR_2019/html/Hudson_GQA_A_New_Dataset_for_Real-World_Visual_Reasoning_and_Compositional_CVPR_2019_paper.html

17. Kil, J., Zhang, C., Xuan, D., Chao, W.: Discovering the unknown knowns: Turning implicit knowledge in the dataset into explicit training examples for visual question answering. In: Moens, M., Huang, X., Specia, L., Yih, S.W. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021. pp. 6346–6361. Association for Computational Linguistics (2021), https://doi.org/10.18653/v1/2021.emnlp-main.512
18. Kim, J., Jun, J., Zhang, B.: Bilinear attention networks. In: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada. pp. 1571–1581 (2018), https://proceedings.neurips.cc/paper/2018/hash/96ea64f3a1aa2fd00c72faacf0cb8ac9-Abstract.html
19. Kim, J., Lee, S., Kwak, D., Heo, M., Kim, J., Ha, J., Zhang, B.: Multimodal residual learning for visual QA. In: Lee, D.D., Sugiyama, M., von Luxburg, U., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain. pp. 361–369 (2016), https://proceedings.neurips.cc/paper/2016/hash/9b04d152845ec0a378394003c96da594-Abstract.html
20. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L., Shamma, D.A., Bernstein, M.S., Fei-Fei, L.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. Int. J. Comput. Vis. 123(1), 32–73 (2017)
21. Li, L.H., Yatskar, M., Yin, D., Hsieh, C., Chang, K.: Visualbert: A simple and performant baseline for vision and language. CoRR abs/1908.03557 (2019), http://arxiv.org/abs/1908.03557
22. Liu, W., Wang, J., Wang, H., Li, R., Qiu, Y., Zhang, Y., Han, J., Zou, Y.: Decoupled rationalization with asymmetric learning rates: A flexible lipschitz restraint. In: Singh, A.K., Sun, Y., Akoglu, L., Gunopulos, D., Yan, X., Kumar, R., Ozcan, F., Ye, J. (eds.) Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023. pp. 1535–1547. ACM (2023), https://doi.org/10.1145/3580305.3599299
23. Liu, Y., Wei, W., Peng, D., Zhu, F.: Declaration-based prompt tuning for visual question answering. In: Raedt, L.D. (ed.) Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022. pp. 3264–3270. ijcai.org (2022), https://doi.org/10.24963/ijcai.2022/453
24. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. In: Lee, D.D., Sugiyama, M., von Luxburg, U., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain. pp. 289–297 (2016), https://proceedings.neurips.cc/paper/2016/hash/9dcb88e0137649590b755372b040afad-Abstract.html
25. Mishra, A., Anand, A., Guha, P.: Dual attention and question categorization-based visual question answering. IEEE Trans. Artif. Intell. 4(1), 81–91 (2023), https://doi.org/10.1109/TAI.2022.3160418
26. Niu, Y., Tang, K., Zhang, H., Lu, Z., Hua, X., Wen, J.: Counterfactual VQA: A cause-effect look at language bias. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. pp. 12700–12710. Computer Vision Foundation / IEEE (2021), https://openaccess.thecvf.com/content/CVPR2021/html/Niu_Counterfactual_VQA_A_Cause-Effect_Look_at_Language_Bias_CVPR_2021_paper.html
27. Peng, D., Wei, W., Mao, X., Fu, Y., Chen, D.: An empirical study on the language modal in visual question answering. In: Proceedings of the Thirty-Second International Joint Conference

on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China. pp. 4109–4117. ijcai.org (2023), https://doi.org/10.24963/ijcai.2023/457
28. Peng, L., Yang, Y., Wang, Z., Huang, Z., Shen, H.T.: Mra-net: Improving VQA via multi-modal relation attention network. IEEE Trans. Pattern Anal. Mach. Intell. 44(1), 318–329 (2022), https://doi.org/10.1109/TPAMI.2020.3004830
29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. pp. 8748–8763. Proceedings of Machine Learning Research, PMLR (2021), http://proceedings.mlr.press/v139/radford21a.html
30. Ren, L., Liu, Y., Ouyang, C., Yu, Y., Zhou, S., He, Y., Wan, Y.: Dylas: A dynamic label alignment strategy for large-scale multi-label text classification. Inf. Fusion 120, 103081 (2025), https://doi.org/10.1016/j.inffus.2025.103081
31. Sharma, H., Jalal, A.S.: An improved attention and hybrid optimization technique for visual question answering. Neural Process. Lett. 54(1), 709–730 (2022), https://doi.org/10.1007/s11063-021-10655-y
32. Shen, X., Chen, C., Han, D., Xu, Y., Wang, X., Zhou, H.: A triple-branch hybrid dynamic-static alignment strategy for vision-language tasks. Neural Networks 191, 107871 (2025), https://doi.org/10.1016/j.neunet.2025.107871
33. Shen, X., Han, D., Chang, C., Oad, A., Wu, H.: Gfsnet: Gaussian fourier with sparse attention network for visual question answering. Artif. Intell. Rev. 58(6), 159 (2025), https://doi.org/10.1007/s10462-025-11163-4
34. Shen, X., Han, D., Chang, C., Xu, Y., Chen, C.: Multimodal context-aware consistency alignment for vision-language tasks. Expert Syst. Appl. 295, 128857 (2026), https://doi.org/10.1016/j.eswa.2025.128857
35. Shen, X., Han, D., Zong, L., Guo, Z., Hua, J.: Relational reasoning and adaptive fusion for visual question answering. Appl. Intell. 54(6), 5062–5080 (2024), https://doi.org/10.1007/s10489-024-05437-7
36. Sobo, A., Mubarak, A., Baimagambetov, A., Polatidis, N.: Evaluating llms for code generation in HRI: A comparative study of chatgpt, gemini, and claude. Appl. Artif. Intell. 39(1) (2025), https://doi.org/10.1080/08839514.2024.2439610
37. Song, X., Han, D., Chen, C., Shen, X., Wu, H.: Vman: visual-modified attention network for multimodal paradigms. Vis. Comput. 41(4), 2737–2754 (2025), https://doi.org/10.1007/s00371-024-03563-4
38. Souza, B., Aasan, M., Pedrini, H., Rivera, A.R.: Selfgraphvqa: A self-supervised graph neural network for scene-based question answering. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023 - Workshops, Paris, France, October 2-6, 2023. pp. 4642–4647. IEEE (2023), https://doi.org/10.1109/ICCVW60793.2023.00499
39. Tan, H., Bansal, M.: LXMERT: learning cross-modality encoder representations from transformers. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019. pp. 5099–5110. Association for Computational Linguistics (2019), https://doi.org/10.18653/v1/D19-1514
40. Tan, S., Zhang, Y., Li, Y., Shin, B.: Alignnet: spatiotemporal alignment and multi-scale feature fusion for enhanced lidar semantic segmentation. Appl. Intell. 56(4), 119 (2026), https://doi.org/10.1007/s10489-025-07021-z
41. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural

Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 5998–6008 (2017), https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
42. Xu, Y., Liu, G., Shen, X., Wang, X., Zhou, H.: Multi-granularity modal interaction and fusion framework for vision-language tasks. Engineering Applications of Artificial Intelligence 178, 114849 (2026), https://www.sciencedirect.com/science/article/pii/S0952197626011310
43. Xu, Y., Zhang, L., Shen, X.: Multi-modal adaptive gated mechanism for visual question answering. PLOS ONE 18(6), 1–24 (06 2023), https://doi.org/10.1371/journal.pone.0287557
44. Yang, Z., Qin, Z., Yu, J., Wan, T.: Prior visual relationship reasoning for visual question answering. In: IEEE International Conference on Image Processing, ICIP 2020, Abu Dhabi, United Arab Emirates, October 25-28, 2020. pp. 1411–1415. IEEE (2020), https://doi.org/10.1109/ICIP40778.2020.9190771
45. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.J.: Stacked attention networks for image question answering. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 21–29. IEEE Computer Society (2016), https://doi.org/10.1109/CVPR.2016.10
46. Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., Finn, C.: Gradient surgery for multi-task learning. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (2020), https://proceedings.neurips.cc/paper/2020/hash/3fe78a8acf5fda99de95303940a2420c-Abstract.html
47. Yu, Z., Yu, J., Cui, Y., Tao, D., Tian, Q.: Deep modular co-attention networks for visual question answering. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 6281–6290. Computer Vision Foundation / IEEE (2019), http://openaccess.thecvf.com/content_CVPR_2019/html/Yu_Deep_Modular_Co-Attention_Networks_for_Visual_Question_Answering_CVPR_2019_paper.html
48. Yu, Z., Yu, J., Xiang, C., Fan, J., Tao, D.: Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. IEEE Trans. Neural Networks Learn. Syst. 29(12), 5947–5959 (2018), https://doi.org/10.1109/TNNLS.2018.2817340
49. Zhang, W., Yu, J., Zhao, W., Ran, C.: Dmrfnet: Deep multimodal reasoning and fusion for visual question answering and explanation generation. Inf. Fusion 72, 70–79 (2021), https://doi.org/10.1016/j.inffus.2021.02.006

Ganlong Zhou is currently pursuing the M.S. degree with the School of Computer Science, Shanghai Maritime University, Shanghai, China. His research focuses on multimodal cognitive intelligence, with particular emphasis on image-text matching, visual question answering, and cross-modal alignment mechanisms.

Dezhi Han received the B.S.degree in applied physics from the Hefei University of Technology, Hefei, China, in 1990, and the M.S. and Ph.D. degrees in computing science from the Huazhong University of Science and Technology, Wuhan, China, in 2001 and 2005, respectively. He is currently a Professor with the Department of Computer, Shanghai Maritime University, Pudong, China, from 2006. His current research interests include cloud and outsourcing security, wireless communication securty,network and information security, visual question answering. He is currently a Member of the IEEE.

Xiang Shen is currently pursuing his Ph.D. at Shanghai Maritime University and is also a joint Ph.D. student at The University of Sydney, supported by the China Scholarship Council (CSC). His research focuses on multimodal learning, with particular emphasis on visual question answering, multimodal fusion strategies, and the development of large-scale multimodal models. His current work also explores the integration of AI agents for adaptive perception and reasoning in complex environments, aiming to bridge the gap between visual understanding and natural language processing.

Xiaoqi Xin is a Ph.D. student at Shanghai Maritime University researching multimodal emotion recognition. Her research combines affective computing and multi-source signal fusion to improve emotion perception under realistic conditions.

Received: April 3, 2026; Accepted: June 14, 2026.

Computer Science and Information Systems 23(4):1287–1320 https://doi.org/10.2298/CSIS260523040N

Paper Type: Research Paper

SLA-DE-RALBA: Cost-efficient dynamic enhanced resource-aware load balancing algorithm for cloud computing

Mohsin Nawaz[1], Altaf Hussain[1], Marran Al Qwaid[2], and Muazzam A. Khan Khattak[3]

[1] Department of Computer Science, KICSIT Campus, Institute of Space Technology, Islamabad 44000, Pakistan
mohsin_mscs@kicsit.edu.pk , altaf.hussain@ist.edu.pk (corresponding author)
[2] Department of Computer Science, College of Computing and Information Technology, Shaqra University, Saudi Arabia
maldossari@su.edu.sa
[3] Computer Science Department, Quaid i Azam University, Islamabad, Pakistan
muazzam.khattak@qau.edu.pk

Abstract. Cloud computing has become a dominant paradigm for delivering scalable and flexible on-demand resources; however, efficiently executing high performance computing (HPC) workloads remains challenging, particularly in heterogeneous environments. Conventional static scheduling methods often lead to poor resource utilization and increased makespan, while dynamic approaches improve load distribution but introduce significant overhead due to continuous monitoring and real-time decision-making. To address these challenges, this paper proposes an SLA-aware Dynamic Enhanced Resource-Aware Load Balancing Algorithm (SLA-DE-RALBA). The algorithm minimizes load imbalance by considering the computational capacities of virtual machines and ensures Service Level Agreement (SLA) compliance through a three-tier priority-based workflow. The proposed approach is evaluated using CloudSim Plus on two benchmark datasets: Heterogeneous Computing Scheduling Problem (HCSP) instances and the Google Cloud Jobs dataset. Results demonstrate that SLA-DE-RALBA consistently outperforms baseline algorithms, including RALBA, DRALBA, DE-RALBA, SLA-RALBA, Dynamic Max-Min, PSSLB, and PSSELB, across key metrics such as makespan, resource utilization, job rejection, throughput, execution time, and cost. Notably, it achieves zero job rejection, reduces energy consumption by up to 85%, improves resource utilization by 11.9%, lowers makespan by 41–45%, and decreases execution time by up to 57%, making it a robust and efficient solution for HPC workload scheduling in cloud environments.

Keywords: Cloud Computing, HPC Workload Scheduling, SLA-Aware Load Balancing, Dynamic Resource Management, Heterogeneous Cloud Environments.

1. Introduction

The National Institute of Standards and Technology (NIST) defines cloud computing as a paradigm that facilitates ubiquitous, convenient, and on-demand network access to a

common pool of configurable computing resources which can be provisioned and discharged with speed having little management overheads and limited service provider intervention [1]. With such a paradigm, it can be easily get in touch with the computational servers, software frameworks, storage disks and also networking systems which may be procured and installed without much interruption on the operational process [2,3]. Previous research on the heuristics of scheduling and resource efficiency address intrinsic benefits of cloud environments such as cost efficiency, elasticity and operational flexibility since cloud resources are available all over the world through the use of the internet [4]. Moreover, cloud computing enables quick implementation of services and applications as well as improves coordination and scalability in dispersed computing settings [5]. Task scheduling is still a core issue in cloud computing since it involves the effective mapping of computational tasks onto accessible resources (e.g. virtual machines (VMs) and physical servers) and also the optimum use of resources, system performance and at times conflicting, goals, including response time, throughput, energy efficiency and adherence to service level agreement [6,7,8]. It is especially pronounced in heterogeneous cloud environments, where dynamic changes in workload and diversity of resources often result in uneven load distribution and under-utilization of resources [9]. From an architectural view, the scheduling decisions may be classified into the two inter-related layers: job-to-VM mapping which assigns the tasks to the virtual machines depending on resource availability and the SLA-related restrictions and VM-to-physical machine mapping which also ensures optimal usage of the hardware facilities by reducing both energy and performance degradation [10].

Modern studies are progressively being transitioned to a SLA-consciousness scheduling strategy with both heuristic and metaheuristic models and other methods that dynamically respond to changing loads and system states. Such approaches mainly aim at maximizing system performance and minimizing the time to complete the tasks (makespan) of the system to achieve maximum energy efficiency and ensuring resiliency and robustness SLA satisfaction despite the dynamic and uncertain conditions related to cloud operating conditions [11]. Dynamic scheduling makes it possible to redistribute workloads dynamically by varying the decisions regarding task allocation according to the current system states and hence offers greater capabilities to handle dynamic and unexpected execution situations [12]. Scheduling strategies may also be categorized into online and offline approaches: online scheduling takes a step-by-step decision based on the arrival of the tasks, while offline scheduling is based on full information about all jobs and aims at globally optimal allocation schemes. Besides there can be preemptive and non-preemptive scheduling policies as far as the ability to interrupt tasks and restart at a later time during the execution can be present or absent [13]. The recent development around cloud scheduling has brought advanced heuristic and meta-heuristic methods, including the Cuckoo Optimization based scheduling that seeks to reduce resource wastage and time of task completion (makespan) [14]. Also new studies focus on green and energy-efficient time-conscious scheduling systems to minimize resource utilization, and fault tolerant and adaptive energy-conscious schedulers that openly trade goals of performance with sustainability demands [15]. Cloud service providers often use approaches of over-provisioning as a mitigation approach to Service Level Agreement (SLA) violations, negatively impacting resource usage, workload imbalance and influencing the presence of tasks and in turn prolonging makespan and reducing provider revenues. To overcome these shortcom-

ings, SLA-RALBA [16] came up with SLA-aware extension of the Resource-Aware Load Balancing Algorithm (RALBA) by the representation of three different categories of SLA which include time-centric, balanced and cost-centric and their combination in terms of the execution cost and execution time in the scheduling process.

The scheduling mechanism used by SLA-RALBA in order to enhance utilization across the virtual machines is based on two phases i.e. Fill and Spill phases [10]. Experimental assessments of both GoCJ and HCSP work loads modeling using CloudSim resulted in demonstrating that SLA-RALBA delivered better performance compared to existing dynamic and proactive scheduling heuristics such as PSSLB, PSSELB, Dynamic Maximum Minimization and DRALBA based strategies in terms of resource utilization, makespan and load balance [16]. However, in the light of such developments a common scheduling model which concurrently conducts both SLA awareness and dynamic run-time flexibility is yet to be well studied [18]. Current strategies are used to trade-off one or two objectives in isolation leading to natural trade-offs that restrict the total efficiencies of the entire system especially at the machine-level load balancing level as is evidenced in earlier research [9,14]. Based on this requirement, SLA-DE-RALBA is designed to combine SLA-based scheduling with run-time flexibility to deliver a balanced trade-off between performance and cost and cloud environments. The major implications of the given research are as follows:

1. A detailed literature review identified key limitations in existing static and dynamic cloud scheduling approaches regarding SLA compliance, flexibility, cost, job rejection, and energy efficiency.
2. A major research gap was identified where SLA-aware algorithms lack dynamic adaptability, while dynamic approaches fail to satisfy SLA requirements with better resource utilization in cloud environments.
3. The proposed SLA-DE-RALBA algorithm integrates SLA-awareness, dynamic resource optimization, and energy-efficient heuristics to improve throughput, resource utilization, job acceptance, cost efficiency, and energy savings.
4. Experimental evaluation using the CloudSim Plus simulator with GoCJ and HCSP datasets demonstrated that SLA-DE-RALBA outperformed several existing cloud scheduling algorithms.

2. Related Work

Cloud job scheduling has been given much attention by the academia and industry since its direct influence on system throughput, resource usage, energy consumption and user satisfaction [7]. Various solutions have been proposed at both ends of the spanning of classical heuristics and multi-criteria SLA-awareness and dynamic scheduling. This paper does not provide a comprehensive survey of scheduling methods; instead, it highlights well-known algorithms representing key advances in cloud scheduling research.

Several dynamic scheduling approaches have been proposed to improve runtime adaptability. Dynamic Min-Max [36] reduces waiting time, improves response time, and increases throughput through concurrent VM allocation, while ECMM [26] enhances load balancing and response time using runtime monitoring of tasks and VMs. Although effective for dynamic workload management, these approaches do not explicitly consider

SLA constraints or cost-awareness [29]. PSSLB and related proactive scheduling approaches within the RePro-Active framework [25] use simulation-based predictions and hybrid Round Robin strategies to optimize task allocation. They improve makespan, resource utilization, and throughput by assigning tasks based on estimated VM performance, while PSSELB further enhances parallelism by prioritizing longer tasks. However, both methods incur additional simulation overhead and do not rely on prior workload knowledge, limiting scalability in large cloud environments. Albalawi [13] proposed a hybrid dynamic scheduling framework combining multiple metaheuristics (RRA-SWO, HAGA, LRT, and LR-HSA) for load balancing in cloud systems. Evaluated on CloudSim with 51 VMs, it achieves high task success, throughput, and resource utilization. However, it lacks SLA-aware differentiation, energy efficiency considerations, and validation in large-scale multi-cloud environments. Hagras and El-Sayed [37] proposed Balanced Independent-Task Assignment (BITA) and its hard real-time counterpart BITAr for fault-tolerant scheduling of independent tasks in heterogeneous clouds, prioritizing tasks based on deadline urgency. Evaluated on GoCJ, the methods outperform FTTA and RADL in utilization and task rejection with low computational overhead. However, they do not consider SLA-aware differentiation or execution cost. Zhu [38] proposed Load Balancing using Ant Colony and Whale Optimization (LBAW), a hybrid ACO–WOA framework for task scheduling, resource optimization, and VM migration in cloud environments. It improves utilization, makespan, and response time over existing methods, as shown in MATLAB experiments on the San Diego dataset. However, it introduces high computational overhead and complex parameter tuning, limiting scalability in large dynamic systems. Beena et al. [38] proposed an energy-aware cloud framework combining Kubernetes with ACO, PSO, and DVFS for resource scaling and container management. It improves energy efficiency and performance, but suffers from ACO-induced overhead and limited real-world validation. RALBA was proposed to address the limitations of traditional scheduling methods such as Min-Min and Round Robin, which often cause VM overload and inefficient resource utilization [19]. It employs a two-phase strategy: the fill phase allocates tasks based on VM processing power, while the spill phase assigns remaining tasks to the VM with the earliest completion time [17]. Experimental results showed significant reductions in makespan and improved resource utilization [10]. DRALBA [27] extends RALBA with a Dynamic-RALBA scheduler and Dynamic Updater for real-time VM load management, making it effective for compute-intensive tasks in heterogeneous environments. CloudSim evaluations with HCSP, GoCJ, and synthetic workloads show high resource utilization, reduced makespan, and improved throughput over several existing methods. However, it does not address energy-aware optimization or multi-cloud interoperability. SLA-RALBA [16] extends RALBA by incorporating cost-aware and SLA-driven scheduling decisions, supporting time-based, cost-based, and hybrid SLA models. It retains the Fill and Spill scheduling framework while aligning task allocation with user SLA requirements [22]. Experimental evaluation on the CloudSim platform demonstrated improved makespan and resource utilization compared to several baseline algorithms. However, SLA-RALBA follows a static scheduling approach, limiting its adaptability to dynamic workload changes during execution [12]. To overcome the limited flexibility of SLA-RALBA, DE-RALBA [23] was proposed with real-time VM state updates and dynamic workload redistribution. It improves adaptability to workload variations, reduces idle time, and minimizes performance bottlenecks through con-

tinuous VM monitoring and rescheduling. Table 1 summarizes the reviewed scheduling algorithms, including their techniques, methodologies, strengths, and weaknesses.

3. Proposed SLA-DE-RALBA Algorithm

The section provides an introduction to the baseline algorithms and a detailed demonstration of the proposed SLA-DE-RALBA algorithm including its architectural framework, system modeling mechanism, performance evaluation model and algorithm code.

SLA-DE-RALBA is a hierarchical architecture as shown in Figure 1, and has five different levels that are used to manage the resources and schedule cloud tasks. The core

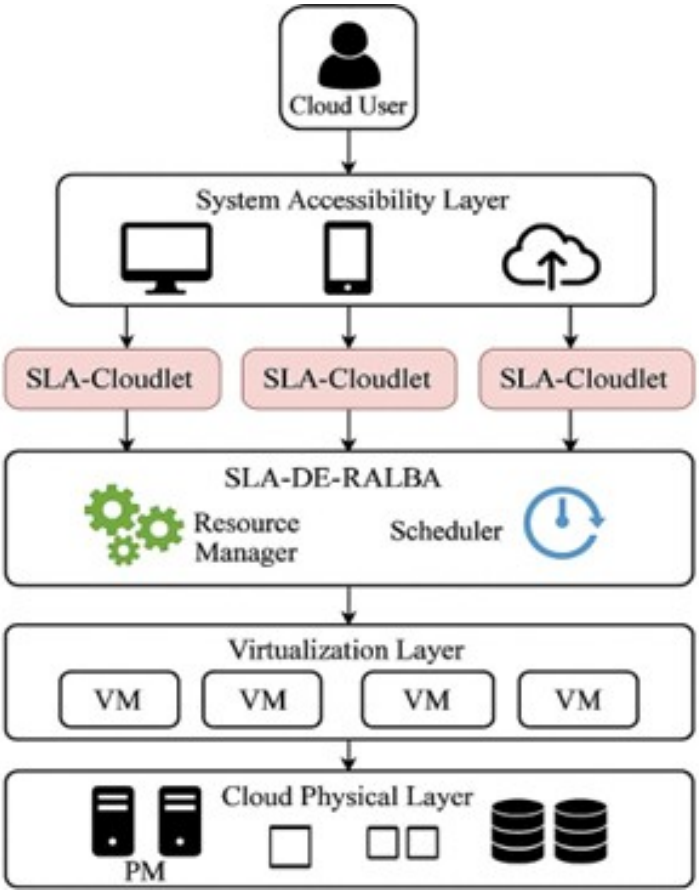


Fig. 1. Abstract Level Diagram

layer is the algorithmic nucleus of SLA-DE-RALBA which consists of a Resource Manager, Scheduler, and Load Balancer. The Resource Manager computes VM shares using workload, SLA weights, and utilization, and enables priority tasks to use idle resources.

Table 1. Comparative Analysis of Existing Scheduling Algorithms

Algorithm	Methodology	Strengths	Weaknesses
RALBA [31]	Two-phase scheduling. The Fill scheduler assigns tasks based on VM computational power; the Spill scheduler allocates remaining tasks to the VM with the earliest completion time.	Reduces makespan vs. Min–Min. Achieves over 7× better resource utilization than Random Selection; improves throughput.	Lacks SLA awareness. Static scheduling fails to adapt to runtime workload variations.
SLA-RALBA [16]	Extends RALBA by incorporating execution cost and supporting time-oriented, cost-oriented, and hybrid SLA models using Fill and Spill schedulers.	Improves makespan and resource utilization while maintaining cost efficiency through balanced time–cost trade-offs.	Higher computational overhead due to cost calculations; less effective for computation-intensive workloads.
DE-RALBA [23]	Employs dual schedulers: one for VM assignment based on processing power and earliest finish time; another to dynamically update VM status.	Achieves higher ARUR and lower makespan on GoCJ and HCSP datasets. Enhances throughput for dynamic workloads.	Runtime monitoring introduces overhead and increases system complexity in large-scale environments.
DRALBA [27]	Combines Dynamic-RALBA scheduling with a Dynamic-Updater to continuously adjust VM loads and computation shares during execution.	Improves resource utilization, reduces makespan, and enhances throughput and scalability.	Runtime updates incur overhead; lacks energy-awareness and multi-cloud support.
Dynamic Min-Max [36]	Real-time scheduling assigns tasks immediately based on estimated execution time without batch sorting.	Improves throughput, efficiency, and load balancing while reducing waiting time in dynamic environments.	Performance declines with high resource heterogeneity; ignores SLA and cost constraints.
PSSLB [25]	Simulation-based approach that sorts tasks in descending order and assigns them to balance VM workloads using execution predictions.	Reduces task completion time, improves average resource utilization, and increases throughput.	Simulation overhead raises time complexity; prediction inaccuracies degrade performance for large datasets.
PSSELB [25]	Enhances PSSLB through a pivot-task mechanism enabling parallel task allocation across heterogeneous VMs.	Delivers best throughput performance; improves makespan and ARUR over Min–Min variants.	High simulation and computation overhead; limited scalability under highly dynamic conditions.
Dynamic Max-Min [26]	Enhanced Max–Min assigns the longest task to the VM with the shortest completion time using periodic updates.	Reduces waiting time and response ratio while improving load distribution and user satisfaction.	Periodic updates may miss real-time workload changes; can reduce resource utilization in some cases.
RRA-SWO [13]	Combines RR–SWO for allocation with HAGA scheduling and LRT–LR-HSA for load prediction.	Delivers high success rate, throughput, and efficient performance in CloudSim.	Does not address SLA compliance or energy efficiency.
BITA/BITAr [36]	Uses criticality-based scheduling, extending from soft to hard real-time with FIFO and fault tolerance.	Delivers high utilization and low task rejection consistently across different workload sizes.	Does not consider SLA priorities or cost factors.
LBAW [37]	Combines ACO for task allocation with WOA for resource utilization and VM migration in a hybrid optimization approach.	Improves resource utilization, reduces makespan, imbalance and response time compared to existing methods.	Higher computational cost and parameter tuning complexity may limit performance in large, dynamic environments.
ARAA/ARKLM [38]	Integrates ACO, PSO, and DVFS in a Kubernetes framework for adaptive resource and container management.	Strong energy efficiency and scalability, with low CPU usage and fast execution performance.	High CPU usage of ACO and reliance on Kubernetes may limit wider adoption; real-world validation still limited.

The Scheduler assigns tasks via an SLA-based priority queue using multi-criteria VM scoring. The Load Balancer monitors VM variance and redistributes workloads when imbalance exceeds a threshold to prevent overload and ensure efficient resource use. The Cloud Physical Layer forms the underlying hardware infrastructure, including servers, storage, and networking resources that support virtualization. The system follows a bidirectional data flow where user requests move through SLA-based cloudlet abstraction, scheduling decisions, VM execution, and physical resource allocation. A continuous feedback loop via the VM Status Update mechanism (Algorithm 2) sends real-time VM status to the SLA-DE-RALBA core, enabling dynamic scheduling, resource reassignment, proactive load balancing, and sustained SLA compliance under varying workloads [32]. This study used a SLA mechanism with three layers; **Level 1** provides best-effort, lowest-priority service (weight 1.1), with tasks placed at the end of the queue and subject to preemption or suspension under resource contention. **Level 2** offers moderate priority service (weight 1.2) with balanced resource allocation, fair scheduling, and protection from starvation, maintaining a good cost–performance trade-off. **Level 3** delivers high-priority, performance-guaranteed service (weight 1.3), with preferential VM assignment, near-full resource access, and preemptive control over lower-priority tasks to ensure SLA compliance under heavy load.

3.1. System Architecture

The SLA-DE-RALBA architecture takes the entire multi-stage flow as a guideline in scheduling tasks in the clouds as depicted in Figure 2. It begins with cloudlets that have certain SLA, computational load (in Million Instructions) and time of arrival. The system then automatically shares dynamic VM in all the virtual machines according to computation ratios, total system workload, SLA weights and the current level of utilization. The suggested scheduling architecture functions on a two-way system of activation according to assessments executed on the initial capacity. In regular operation when the resources are adequate. The major pathway will take a simple chain of operations by initializing resources availability tabling of all virtual machine instances then ensuring entry points of incoming computational tasks. The scheduler then gives priorities to tasks as per service-level agreement requirement and allocates the most critical workload to the virtual machine with the most desirable performance properties and consistency within the system. On the other hand, once the total system capacity decrees beyond set points, the structure transitions to a different route which makes use of SLA-aware scheduling logic to locate significant pending tasks assign them to virtual machines with the best expected times to completion that meet service agreements and furthermore resource constraints and broadcast allocation changes throughout all monitoring systems.

These parallel execution paths are eventually combined at a strategic optimization level at which the scheduler is done to carry out detailed analysis of rescheduling opportunities. It entails the comparison of the recently estimated time of accomplishing the task with a possibly shorter time that could be completed at idle or lightly loaded virtual machines. When there is a chance of improving performance, the system is strategic in moving less important tasks out of the congested resources into the empty capacity thus improving the overall throughput and the efficiency in utilizing the resources used. Subsequent to this refinement process architecture starts load balancing processes in details

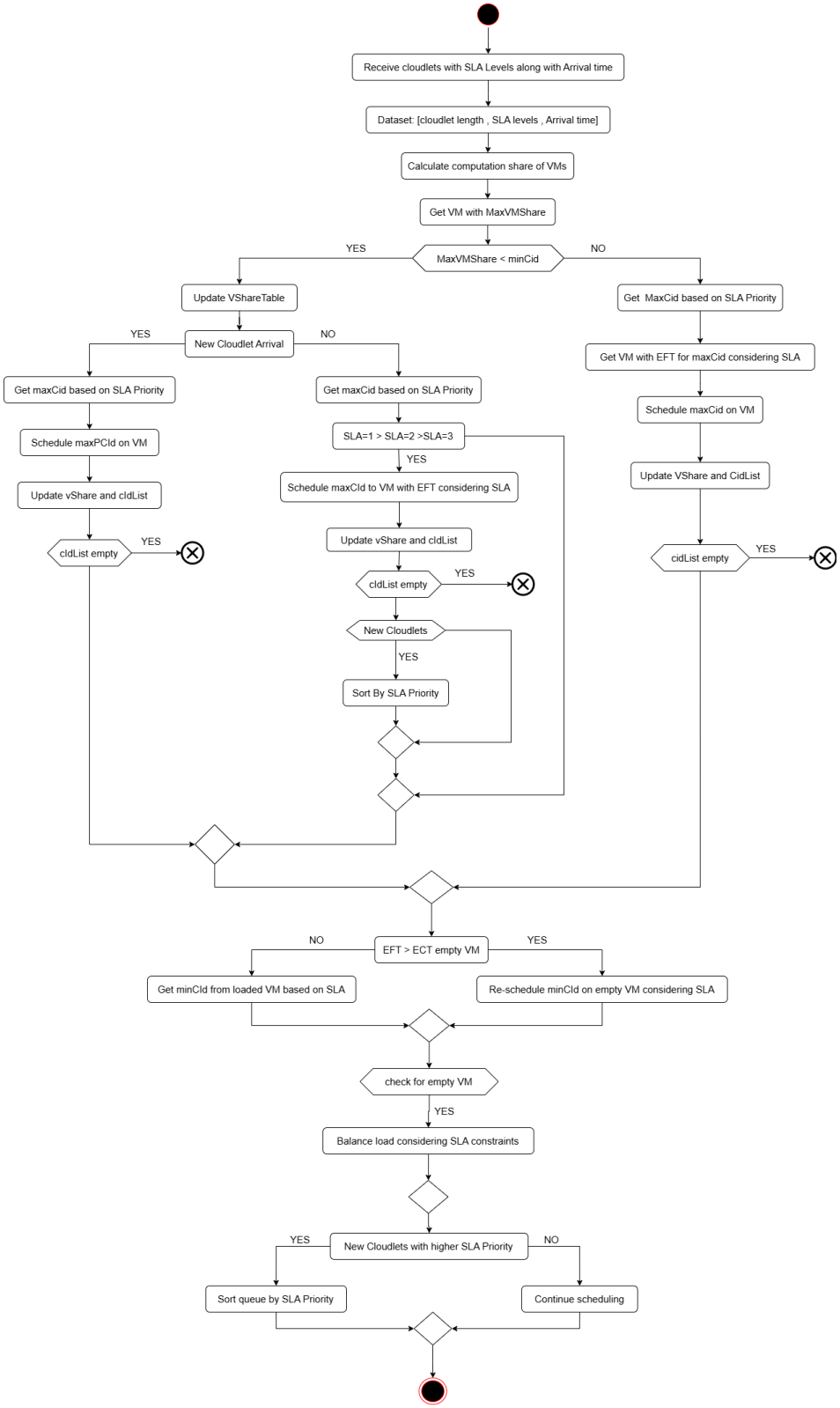


Fig. 2. Flow Diagram of SLA-DE-RALBA

where it is known that virtual machines are underutilized and calculates the variance measurements across the whole infrastructure. When there is more than configured levels of variance intelligent redistribution systems come into effect to offload computational workloads of saturated resources to unused capacity without notifying service-level guarantees during the migration period.

The framework includes advanced dynamic arrival processing work which constantly monitors the task submission pipeline specifically focusing on high-priority workloads of SLA Level 3. In case critical tasks are present in the system when running the task queue is dynamically restructured through extensive SLA-driven reorganization of all pending tasks to guarantee high-priority workloads a faster scheduling consideration as low-priority workloads are moved to their proper position. This real-time monitoring and responsiveness enables the SLA-DE-RALBA architecture to accommodate changing workload profiles ensure uniformity in compliance at all service layers automated resource mapping via smart virtual machine selection algorithms and dynamic capacity control solutions and stability in the long-run of the system through proactive interventing measures. All of these combined minimize the rate of task rejection, maximize the operational efficiency of a data center and provide differentiated service quality in accordance with hierarchical structures of service-level agreement.

3.2. System Model

The SLA-DE-RALBA model is a universal system of comprehensive management of the entire cloud scheduling process starting with the arrival of computing work and its ultimate execution. This architecture is developed on the basis of event-driven continuous architecture which incorporates four interdependent components service-level agreement conscious priority management, resource allocation mechanisms, flexible virtual machine choice plans and proactive load structure controls. The design brings together the architectural structure characterizing component interactions and the composition of operation specifications. On the sequential arrangement of the operations to schedule resource heterogeneity and the activities of redistributing loads across the entire simulation procedure. In addition various notations that were employed during this work are indicated in Table 2.

The **mathematical foundation** underlying the SLA-DE-RALBA algorithm facilitates intelligent task-to-virtual machine mapping decisions that enhance execution time minimization, system throughput maximization and balanced workload distribution.

$$\text{maxSLACld} = \max_{\forall p \in \text{SLACloudlets}} \left(\text{SLA Priority}(p) \times 10^6 + \text{Size}(p)\right) \tag{1}$$

where SLA Priority follows the hierarchy: $SLA\text{-}1 > SLA\text{-}2 > SLA\text{-}3$.

maxSLACld = max(SLA Priority(p) × 10^6 + Size(p)) Equation (1) determines the most critical task for processing by integrating service-level priority coefficients with computational workload dimensions to establish scheduling order.

$$\text{VMShare}_j = \text{vCRatio}_j \times \text{totalMI} \times \text{slaWeight}_j \times \text{utilizationFactor}_j \tag{2}$$

Equation (2) computes the adaptive virtual machine resource allocation proportion by incorporating processing capacity ratios, aggregate system workload, service-level weighting parameters, and current utilization metrics.

$$\text{slaWeight}_j = 1.0 + \frac{\text{avgSLA}_{\text{VM}_j}}{10.0} \tag{3}$$

Table 2. Notations used in SLA-DE-RALBA

Variable / Notation	Description
`cld`	Representation of user jobs on Cloud
`vList`	List of Virtual Machines in a datacenter
`cldRepository`	Repository of user tasks with arrival time and SLA levels
`MIPS`	Million Instructions Per Second (Computation Power of VM)
`MI`	Million Instructions (size of cld)
`vCRatio`	Computation ratio of all the VMs
`maxPcld`	Maximum possible cloudlet for a VM
$\texttt{vShare}_i$	Computation share of VM_i
`newVShare`	Modified computation share of a VM
`vStatusTable`	List that stores set of VM computation share, ready time, status of VM, and list of cloudlets associated to VM
$\texttt{vStatusTable\_vm\_}i$	Status table of VM_i in vStatusTable
`cldStatusTable`	List that stores set of cld and their status with SLA levels
$\texttt{cldStatusTable\_}j$	Status table of cld_j in cldStatusTable
`cldList`	List of cld to be scheduled
`cldSList`	A map $\langle \text{cld}, \text{VM}_i \rangle$ to store scheduled cloudlets
`cldTList`	List to store temporary clds
`maxVShare`	VM with maximum computation share from all VMs
`maxCld`	Maximum length cld from cldList
`vEFT`	VM that gives earlier finish time for maxCld
`tLength`	Sum of clds length in cldList
`isFinishedExecution`	A flag to determine if a VM completes all assigned clds
`slaLevel`	SLA priority level of cloudlet (SLA-1, SLA-2, SLA-3)
`slaWeight`	Weight factor based on SLA level for VM share calculation
`borrowedCapacity`	Amount of capacity borrowed from other VMs
`loadVariance`	Variance in load distribution across VMs
`rejectionRatio`	Ratio of rejected cloudlets to total cloudlets
`avgLoad`	Average load across all VMs
`balancePenalty`	Penalty factor for load imbalance
`dynamicCloudletArrived`	Indicating arrival of new cloudlets during simulation
`arrivalTime`	Time when cloudlet arrives in the system
`utilizationFactor`	Factor based on current VM utilization
`optimalVM`	VM selected based on optimal scoring function

As demonstrated in Equation (3), the system calculates service-level adjustment coefficients for virtual machine resource distribution by utilizing the mean service priority values of currently assigned computational tasks.

$$\text{utilizationFactor}_j = \left(1.0 - \frac{\text{utilization}_j}{\text{totalUtilization}}\right) \times 2.5 \tag{4}$$

Equation (4) establishes a workload-dependent adjustment multiplier that favors lightly loaded virtual machines during workload redistribution operations.

$$\text{score}_j = (\text{utilization}_j \times 0.3) + (1.0 - \text{remainingCapacity}_j) \times 0.6 + \text{balancePenalty}_j \times 0.1 \tag{5}$$

Equation (5) produces a composite virtual machine ranking metric that incorporates weighted contributions from instantaneous resource consumption levels, available processing headroom and workload distribution penalty factors.

$$\text{remainingCapacity}_j = \frac{\text{VMShare}_j}{\text{VM}_j.\text{MIPS}} \tag{6}$$

Equation (6) quantifies the proportional share of virtual machine processing power that remains unallocated and accessible for new workload assignments.

$$\text{balancePenalty}_j = \frac{|\text{utilization}_j - \text{avgLoad}|}{\text{loadVariance} + \varepsilon} \tag{7}$$

Equation (7) introduces corrective weights for virtual machines exhibiting workload variance from mean system utilization thereby encouraging uniform resource allocation throughout the computational infrastructure.

$$\text{optimalVM} = \min_{\forall j \in \{1,2,3,\ldots,m\}} (\text{score}_j) \tag{8}$$

Equation (8) identifies the virtual machine possessing the lowest composite ranking value as the most suitable destination for computational task placement among the accessible resource inventory.

$$\text{loadVariance} = \frac{1}{m} \times \sum_{j=1}^{m} (\text{utilization}_j - \text{avgLoad})^2 \tag{9}$$

Equation (9) evaluates the infrastructure-wide workload dispersion heterogeneity to identify conditions necessitating corrective redistribution actions.

$$\text{loadVariance} > \text{threshold} \tag{10}$$

The framework initiates preventive workload redistribution protocols when measured dispersion levels surpass established threshold boundaries signaling substantial resource allocation asymmetry.

$$\text{borrowedCapacity} = \textstyle\sum_{j \neq \text{target}} \min(\text{VMShare}_j \times 0.95, \text{requiredLength} - \text{VMShare}_{\text{target}}) \tag{11}$$

As demonstrated in Equation (11), the system permits critical priority workloads to acquire up to 95% of idle processing capacity from alternate virtual machines during scenarios of inadequate local resource availability.

$$\text{SLA Cloudlet CT}_{i,j} = \frac{\text{Cloudlet}_i.\text{MI}}{\text{VM}_j.\text{MIPS}} + \text{VM CT}_j \tag{12}$$

Equation (12) determines the anticipated execution duration for computational tasks on designated virtual machines by incorporating workload magnitude and processing velocity parameters.

$$\text{vEFT} = \min_{\forall j \in \{1,2,3,\ldots,m\}} \left(\text{SLA Cloudlet CT}_{i,j} \times 0.3 + \text{utilization}_j \times 0.6 + \text{balancePenalty}_j \times 0.1\right) \tag{13}$$

Equation (13) identifies the virtual machine offering the most expeditious task finalization through a composite metric integrating execution duration projections, resource consumption levels and workload equilibrium adjustment factors.

$$\text{executedMI}_j = \text{VM}_j.\text{MIPS} \times \text{updateTimeDifference} \tag{14}$$

Equation (14) quantifies the aggregate instruction throughput in millions processed by virtual machines across the temporal window separating successive state monitoring cycles.

$$\text{remainingMI}_j = \text{totalAssignedMI}_j - \text{executedMI}_j \tag{15}$$

Equation (15) revises the outstanding computational burden on virtual machines by deducting completed instruction counts from the cumulative allocated task volume.

$$\text{readTime}_j = \text{readTime}_j - \text{updateTimeDifference} \tag{16}$$

Equation (16) recalibrates virtual machine availability timestamps by reducing them proportional to the duration elapsed since the preceding status synchronization ensuring precision in scheduling determinations.

$$\text{rejectionRatio} = \frac{\text{Rejected\_Cloudlets}}{\text{Total\_Cloudlets}} \tag{17}$$

Equation (17) quantifies system admission performance through the proportional metric of declined workloads resulting from insufficient computational resources or service-level agreement breaches.

$$\text{ARUR} = \frac{\left(\sum_{j=1}^{m} \text{makeSpan}_j / m\right)}{\max(\text{makeSpan}_j)} \tag{18}$$

Equation (18) assesses infrastructure utilization effectiveness by computing the quotient of mean execution completion time relative to peak completion duration.

$$\text{Energy}_j = \sum_{i=1}^{n} \left[((1 - \text{cpuUtil}_i) \times 100 + \text{cpuUtil}_i \times 150) \times \frac{\text{finishTime}_i - \text{startTime}_i}{1000}\right] \tag{19}$$

Equation (19) determines aggregate power expenditure through a linear consumption framework incorporating processor utilization metrics and operational time intervals.

$$F[i,j] = \begin{cases} 1, & \text{if Cloudlet}_i \text{ is assigned to VM}_j \text{ based on SLA priority} \\ 0, & \text{Otherwise} \end{cases} \tag{20}$$

A binary variable denoting the assignment status of cloudlets to virtual machines, governed by service-level agreement priority rules as presented in Equation (20).

3.3. Performance Model

The performance evaluation framework incorporates seven critical metrics that collectively capture the algorithm's operational characteristics. This comprehensive assessment enables rigorous comparison against conventional scheduling approaches and reveals the inherent compromises within multi-objective optimization under SLA constraints.

The metric of **ARUR** shown in Equation (21) that defines the usefulness of computational capacity of datacenter infrastructure is the following

$$\text{ARUR} = \frac{1}{n} \times \sum(\text{U}_i) \tag{21}$$

where Ui denotes the utilization rate of the ith resource and n represents the total resource count.**ARUR** values approaching 1.0 indicate balanced load distribution with proportional VM contributions while values near 0 signify load concentration on limited resources, resulting in underutilization of remaining capacity.**Throughput** denotes the total completed cloudlets over the simulation period given by Equation (22).

$$\text{Throughput} = \frac{\text{Number of Completed Cloudlets}}{\text{Total Simulation Time}} \tag{22}$$

High throughput indicates efficient resource coordination that reduces idle periods, optimizes task-to-VM matching and maximizes virtual machine utilization.**Makespan** is defined as

$$\text{Makespan} = \max(\text{FinishTime}_i) \tag{23}$$

Lower makespan values indicate superior scheduling efficiency through enhanced parallel processing and reduced delays **Energy consumption** serves as a critical metric for evaluating the environmental and operational sustainability of scheduling systems, calculated using Equation (24)

$$E = \int P(u(t))\, dt \tag{24}$$

where $P(u(t))$ represents the power consumption as a function of CPU utilization $u(t)$. The analysis employs a linear power model as expressed in Equation (25)

$$P(u) = P_{\text{idle}} + (P_{\text{max}} - P_{\text{idle}}) \times u \tag{25}$$

where P_idle represents the power consumption during idle state, P_max denotes the maximum power at full CPU utilization, and u ranges from 0 to 1.

The **Job Rejection Ratio** (JRR) is calculated as JRR = (Rejected Jobs / Total Submitted Jobs) x 100. Lower JRR values signify effective workload management and indicate the system's ability to handle resource contention, maintain reliability and ensure SLA compliance.

Execution time is calculated by : Cloudlet Length/VM MIPS = Cloudlet Length/Million Instructions (MI)VM MIPS/Million Instructions Per Second.The total execution time represents the cumulative processing duration for all completed cloudlets as expressed in Equation (26)

$$\text{Total Execution Time} = \sum_{i=1}^{n} \left(\frac{\text{Cloudlet Length}_i}{\text{VM MIPS}_i} \right) \tag{26}$$

Execution cost metric evaluates scheduling effectiveness through monetary quantification of cloudlet processing expenses across VMs. Cost calculation is formulated in Equation (27).

$$\text{Individual Cloudlet Cost} = \text{Cost Per Unit Time}_{\text{VM}} \times \left(\frac{\text{Cloudlet Length}}{\text{VM MIPS}} \right) \tag{27}$$

where VM_price represents the cost per unit time of the virtual machine and Cloudlet Length/VM MIPS indicates the cloudlet's execution duration. The total execution cost aggregates the individual costs across all completed cloudlets as expressed in Equation (28)

$$\text{Total Execution Cost} = \sum \left(\text{Cost Per Unit Time}_{\text{VM}} \times \frac{\text{Cloudlet Length}}{\text{VM MIPS}} \right) \tag{28}$$

The cost model in Equations (27) and (28) intentionally computes execution cost as the product of a static VM cost-per-unit-time parameter and task execution duration.This formulation was adopted for some reasons: (i) it maintains methodological consistency with predecessor algorithms SLA-RALBA and DE-RALBA, ensuring fair cross-algorithm cost comparisons (ii)CloudSimPlus simulation framework does not natively model real-world cloud billing components such as data transfer fees, storage charges or dynamic pricing mechanisms making it methodologically appropriate to restrict cost evaluation to the computational execution dimension that the simulator can reliably model and (iii) it aligns with the broader CloudSim-based scheduling literature [9, 10, 16, 23], providing a common baseline for cross-study comparison.

3.4. SLA-DE-RALBA Algorithm

SLA-DE-RALBA operates through a dual-scheduler architecture consisting of the Main Scheduler (presented in Algorithm 1) and the Update Scheduler (detailed in Algorithm 2). The simulation source code for SLA-DE-RALBA is publicly available online.[4]

Main Scheduler The SLA-DE-RALBA algorithm integrates SLA-aware task scheduling, dynamic resource allocation and load balancing across heterogeneous cloud environments. It processes cloudlet requests using two primary inputs: a cloudlet status table containing task requirements, SLA levels, timestamps and a VM status table with resource availability, utilization and capacity information.

Initialization: VM computational resources are quantified using MIPS metrics with dynamic share provisioning governed by task priority weights and real-time utilization patterns as presented in Equation(29).

$$\text{VMShare}_j = \text{ComputationRatio}_j \times \text{TotalMI} \times \text{SLAWeight}_j \times \text{UtilizationFactor}_j \tag{29}$$

Main Loop: The algorithm iterates until all cloudlets are either scheduled or rejected:

1. **Optimal VM Selection:** Determines the best VM candidate through capacity optimization, utilization balancing, and equitable load distribution.
2. **SLA-Aware Scheduling:** Selects cloudlets based on SLA priority(SLA Level $3 > 2 > 1$).

[4] https://bitbucket.org/sladeralba_26/dynamic_sla-de-ralba/src/main/

3. **Task Assignment:** Executes cloudlet-to-VM binding via assignCloudletToVM followed by synchronization of resource and cloudlet status tables.

Algorithm 1: SLA-DE-RALBA Main Scheduler

```
Input  : List cloudletStatusTables, List vmStatusTables
Output: Optimally scheduled cloudlets with SLA compliance
Initialize simulation parameters and load dataset;
Calculate computation ratio and VM shares for all VMs;
while cloudletList ≠ ∅ do
    optimalVM ← getOptimalVMForCloudlet();
    if optimalVM ≠ null then
        targetCloudlet ← getMaxCloudletForVmShareWithSLA(optimalVM);
        if targetCloudlet ≠ null then
            assignCloudletToVM(optimalVM, targetCloudlet);
            updateVMStatusTable(optimalVM);
        else if dynamicCloudletArrived then
            updateVMStatusTable(currentTime);
            dynamicCloudletArrived ← false;
        else
            maxCloudlet ← getMaxCloudletWithSLA() (SLA-1 > SLA-2 > SLA-3
              priority);
            if maxCloudlet ≠ null then
                targetVM ← getVMWithMinFinishTime(maxCloudlet);
                if targetVM.vmShare < maxCloudlet.length then
                    performLoadBalancing();
                    borrowedCapacity ← borrowCapacity(targetVM,
                      maxCloudlet.length);
                    assignCloudletToVM(targetVM, maxCloudlet);
                    updateVMStatusTable(targetVM);
                else
                    REJECTED_CLOUDLETS++;
                    removeUnSchedulableCloudlet();
            else
                REJECTED_CLOUDLETS += cloudletList.size();
                break;

procedure onCloudletArrival:
    if newCloudletsArrived then
        cloudletList.addAll(newCloudlets);
        dynamicCloudletArrived ← true;
        goto step 3;

return schedulingResults;
```

Dynamic Adaptation: Processes incoming tasks by invoking updateVMStatusTable() to refresh resource states and queuing new cloudlets according to SLA priority enabling real-time workload adaptation.

Resource Management: If the optimal VM is unavailable the algorithm selects the highest-priority cloudlet using getMaxCloudletWithSLA() and identifies the VM with the earliest completion time via getVmWithMinFinishTime(). In capacity-constrained scenarios performLoadBalancing() redistributes tasks while borrowCapacity() leverages up to 95% of available capacity from other VMs to minimize rejections.

Rejection Handling: Identifies and removes unschedulable tasks via removeUnschedulableCloudlet(), increments the rejection counter and records rejection statistics for subsequent performance evaluation.

Event Handling: Monitors incoming cloudlet arrivals through onCloudletArrival(), enqueues new tasks and triggers priority queue reorganization to maintain SLA-based ordering.

Termination: Returns comprehensive performance metrics including scheduled cloudlet count, rejection ratio, VM utilization, execution times and resource consumption. These metrics facilitate evaluation across multiple dimensions such as ARUR and throughput demonstrating balanced achievement of cloud scheduling objectives.

VM Status Update The VM Status Update mechanism (illustrated in Algorithm 2)

Algorithm 2: VM Status Update

Input : int *currentTime*, List *vmStatusTables*
Output: Updated VM status tables
foreach *vmTable in vmStatusTables* **do**
  updateTimeDifference $\leftarrow$ *currentTime* $-$ *vmTable.lastUpdatedTime*;
  if *vmTable.readTime* $-$ *updateTimeDifference* $\leq$ *0* **then**
    vmTable.readTime $\leftarrow 0$;
    vmTable.remainingMI $\leftarrow 0$;
    vmTable.executedMI $\leftarrow$ *vmTable.totalAssignedMI*;
  else
    executedMI $\leftarrow$ *vmTable.mips* $\times$ *updateTimeDifference*;
    vmTable.readTime $-=$ *updateTimeDifference*;
    vmTable.executedMI $+=$ *executedMI*;
    vmTable.remainingMI $\leftarrow$ *vmTable.totalAssignedMI* $-$ *vmTable.executedMI*;
  vmTable.lastUpdatedTime $\leftarrow$ *currentTime*;

Recalculate VM shares with SLA weights;
totalMI $\leftarrow$ *calculateTotalMI()*;
foreach *vm in vmStatusTables* **do**
  slaWeight $\leftarrow$ *calculateSlaWeight(vm)*;
  utilization $\leftarrow$ *vm.remainingMI* / *vm.mips*;
  vmShare $\leftarrow$ *vm.computationRatio* $\times$ *totalMI* $\times$ *slaWeight* $\times$ $(1.0-$ *utilization*$)$;
  vm.vmShare $\leftarrow \max(0,$ *vmShare*$)$;

Load balancing check;
loadVariance $\leftarrow$ *calculateLoadVariance(vmStatusTables)*;
if *loadVariance* $\geq$ *threshold* **then**
  rebalanceVMs();
return *vmStatusTables*;

coordinates with the Main Scheduler to maintain real-time resource state accuracy ensuring scheduling decisions leverage current availability metrics rather than stale data. This enables responsive adaptation to evolving workload conditions including task terminations, new task submissions, and fluctuating VM resource consumption.

The update procedure accepts two inputs: the current simulation time and the comprehensive VM status table. It iteratively processes each VM entry (foreach vmTable in vmStatusTables) by computing the temporal delta: updateTimeDifference = currentTime - vmTable.lastUpdatedTime. Upon detecting task completion (vmTable.readTime - updateTimeDifference ≤ 0), the algorithm resets the VM state with readTime = 0, remainingMI = 0, executedMI = totalAssignedMI, and aligns lastUpdatedTime to currentTime.

During each update cycle, the algorithm recalculates dynamic VM shares with SLA-weighted prioritization.The aggregate computational workload across the entire system is computed as:

$$\text{aggregateMI} = \text{calculateTotalMI}()$$

Individual VM workload contributions are then determined using Equations (30) and (31).

$$\text{slaWeight} = \text{calculateSlaWeight}(\text{vm}) \tag{30}$$

$$\text{SLAWeight}_j = 1.0 + \frac{\text{avgSLA}_{\text{VM}_j}}{10.0} \tag{31}$$

where $\text{avgSLA}_{\text{VM}_j}$ denotes the mean SLA priority level across all cloudlets allocated to VM j. The instantaneous utilization metric is computed according to Equation (32).

$$\text{utilization} = \frac{\text{vm.remainingMI}}{\text{vm.mips}} \tag{32}$$

Also the compute core share is calculated in Equation (33)

$$\text{vmShare} = \text{vm.computationRatio} \times \text{totalMI} \times \text{slaWeight} \times (1.0 - \text{utilization}) \tag{33}$$

A load-balancing term (1.0 - utilization) weights the distribution inversely with current utilization assigning higher shares to underloaded VMs and fostering equitable load distribution. Share values are bounded to prevent negative allocations through the constraint defined in Equation (34).

$$\text{vm.vmShare} = \max(0, \text{vmShare}) \tag{34}$$

Finally, the algorithm assesses aggregate load distribution across the system through Equations (35) and (36).

$$\text{loadVariance} = \text{computeLoadVariance}(\text{vmStatusTables}) \tag{35}$$

$$\text{LoadVariance} = \frac{1}{m} \times \sum (\text{utilization}_j - \text{avgLoad})^2 \tag{36}$$

When loadVariance $\geq$ threshold (nominally 0.02) which is an indicator of critical load imbalance, the algorithm calls rebalanceVMs to replace saturated VM loads with underutilized resources without violating SLA constraints. New vmStatusTables ensure that future scheduling tasks are based on up-to-date information on resource availability. Such a continuous equilibration mechanism along with the Main Scheduler would allow SLA-DE-RALBA to offer adaptive features of effective resource utilization, load balance over time in the context of simulation run and stable SLA satisfaction in the case of heterogeneous workload.

4. Computing Setup

In this section, the cloud infrastructure was described as well as experimental datasets that will be used during the study.

4.1. Datasets

This study employs two established scientific benchmark datasets for experimentation: 1) **GoCJ dataset**, and 2) **HCSP dataset**. The GoCJ dataset is publicly available online on URL[5], and HCSP Dataset is also available publicaly online on URL[6]. The composition of datasets are also presented in our earlier studies in [9,16,23,30].

The original GoCJ benchmark dataset was introduced by Hussain and Aleem [32] and provides Google-derived realistic workload patterns representing real cloud job traces. To enable SLA-aware scheduling evaluation, the original dataset was modified as follows. The modified dataset retains the original job computational lengths expressed in Million Instructions (MI) ranging from lightweight operations of 100 MI to computationally intensive workloads of up to 10,000 MI.To facilitate hierarchical service differentiation and dynamic workload simulation, we structured the datasets in CSV format with three essential attributes:

1. **Cloudlet Length:** This metric quantifies the processing requirements of each task using Million Instructions (MI) as the unit of measurement. Our dataset includes tasks ranging from lightweight operations requiring 100 MI to computationally intensive workloads demanding up to 10,000 MI, enabling performance evaluation across diverse computational scenarios.
2. **SLA Level:** Implemented a three-level classification system to govern resource prioritization. Level 3 represents premium service with high-priority status and a 30% capacity allocation bonus. Level 2 offers a balanced approach, providing moderate priority with a 20% capacity enhancement. Level 1 caters to cost-optimized services, granting standard priority with a 10% capacity supplement. This tiered structure enables differentiated resource distribution based on service requirements.
3. **Arrival Time:** This parameter captures when individual tasks enter the simulation environment, allowing us to model dynamic workload behaviors and time-dependent scheduling constraints accurately.

Prior to the simulation, we applied the following preprocessing steps to the dataset: (i) rows containing non-positive job lengths or negative arrival times were identified and discarded (ii) SLA level values outside the valid range of 1–3 were flagged and removed (iii) arrival times were verified to be consistent with the simulation termination boundary of 5000 seconds and (iv) jobs were sorted by arrival time to ensure correct sequential injection during simulation. No imputation or synthetic augmentation was applied; only structurally valid records were retained for evaluation.

The **HCSP dataset**, which was originally proposed by Tracy [33] to use in our experiment. This was done by adaptation that included changing the data to test the SLA-DE-RALBA in our three-tier SLA system under different heterogeneity levels. The data

[5] https://data.mendeley.com/datasets/r93dfjvvv6/
[6] https://www.fing.edu.uy/inco/grupos/cecal/hpc/HCSP/index.htm

set is in CSV format having three major columns Task computational requirements (in terms of Million Instructions) designation of assigned SLA tier (between 1 and 3) and task arrival time to be applied in dynamic workload simulation. The dataset has a total of nine different workload types and each of them possesses distinct resources consumption patterns.

CPU-Dominant Workloads: The u_c-hilo category has tasks with heavy CPU usage requirement coupled with high I/O usage and low memory usage such that the capacity to allocate CPU is the main consideration in SLA. Contrary to this u_c-lohi tasks can be characterized by intensive CPU processing required along with large memory requirements but a low level of I/O activity which requires memory mindful resource allocation of SLA resources.

I/O-Centric Workloads: Tasks that are classified as u_i-hilo have both a high I/O intensity and large CPU demand with a low memory demand that demand I/O-intensive scheduling policies. The u_i-lohi type provides an I/O load and a high memory load and consumes less CPU resources, posing a memory-I/O optimization to SLA management. Meanwhile u_i-lolo tasks are pre dominated with I/O operations that require minimal CPU and memory.

Storage-Intensive Workloads: The u_s-hilo Work loads are workloads with a high storage requirement and intensive workload with little memory usage that puts a strong focus on the storage-prioritized resource allocation. On the other hand u_s-lohi tasks are both storage intensive and have high memory requirements with little CPU processing making scenarios of SLA optimization in storage, storage-memory.

4.2. Experimental Setup

For the experimentation, the CloudSim Plus simulator was used on a computing system equipped with an Intel Core i7-6600U processor operating at 2.80 GHz with turbo boost support and 8 GB DDR4 RAM. To emulate a cloud infrastructure environment, a simulated datacenter comprising 30 physical host machines was developed, supporting between 30 and 120 virtual machines according to the experimental requirements. The virtual machines were configured with varying processing capabilities measured in MIPS (Million Instructions Per Second), which served as the primary performance metric for executing workloads derived from the GoCJ and HCSP benchmark datasets.

The implementation architecture incorporated several key components, including a datacenter broker responsible for resource allocation, an SLA-aware scheduling mechanism, dynamic cloudlet management functionality, support for real-time task arrivals, and a heterogeneous virtualized infrastructure consisting of 30 hosts and 120 virtual machines with diverse computational capacities.

5. Experimental Results

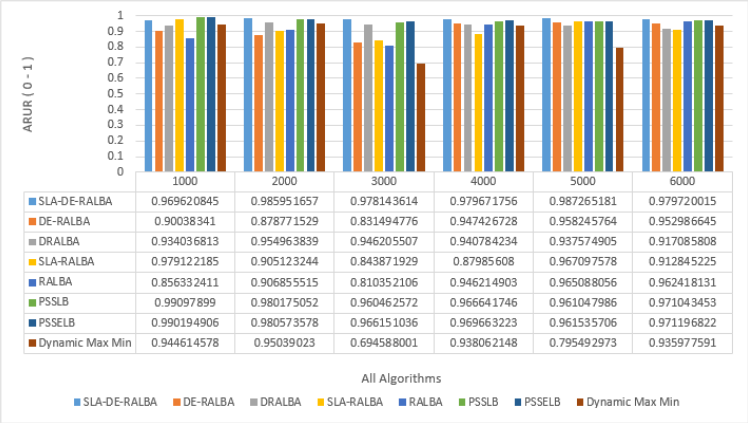
This section provides a performance assessment of the proposed SLA-DE-RALBA algorithm. The evaluation of SLA-DE-RALBA and other schedulings namely DE-RALBA, DRALBA, SLA-RALBA, RALBA, PSSLB, PSSELB and Dynamic Max-Min is conducted using the performance metrics i.e., makespan and average resource utilization ratio (ARUR), energy consumption, execution time and cost, and throughput. Each experiment

is repeated five times and the average values obtained from these iterations are used for the analysis and comparison purposes. The experiments are performed using two benchmark workloads i.e., GoCJ and HCSP.

5.1. Performance Assessment Results - GoCJ Dataset

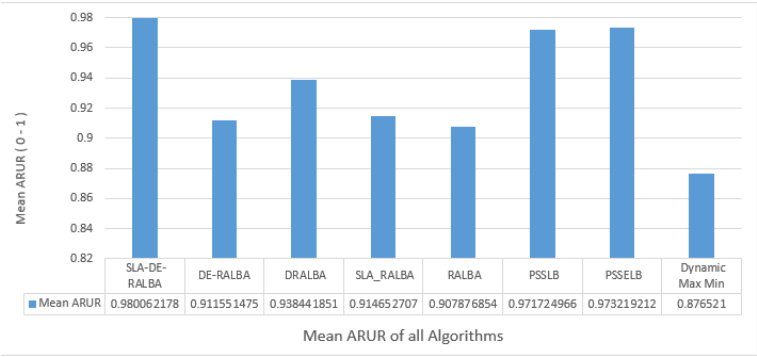
The experiments are performed using GoCJ benchmark dataset [32] and these results for each parameter are presented in this subsection.

ARUR evaluates resource utilization efficiency in cloud systems. As shown in Figure 3, SLA-DE-RALBA maintains consistently high ARUR (0.969–0.987) across all workloads (1000–6000 jobs) with zero job rejections.



	1000	2000	3000	4000	5000	6000
SLA-DE-RALBA	0.969620845	0.985951657	0.978143614	0.979671756	0.987265181	0.979720015
DE-RALBA	0.90038341	0.878771529	0.831494776	0.947426728	0.958245764	0.952986645
DRALBA	0.934036813	0.954963839	0.946205507	0.940784234	0.937574905	0.917085808
SLA-RALBA	0.979122185	0.905123244	0.843871929	0.87985608	0.967097578	0.912845225
RALBA	0.856332411	0.906855515	0.810352106	0.946214903	0.965088056	0.962418131
PSSLB	0.99097899	0.980175052	0.960462572	0.966641746	0.961047986	0.971043453
PSSELB	0.990194906	0.980573578	0.966151036	0.969663223	0.961535706	0.971196822
Dynamic Max Min	0.944614578	0.95039023	0.694588001	0.938062148	0.795492973	0.935977591

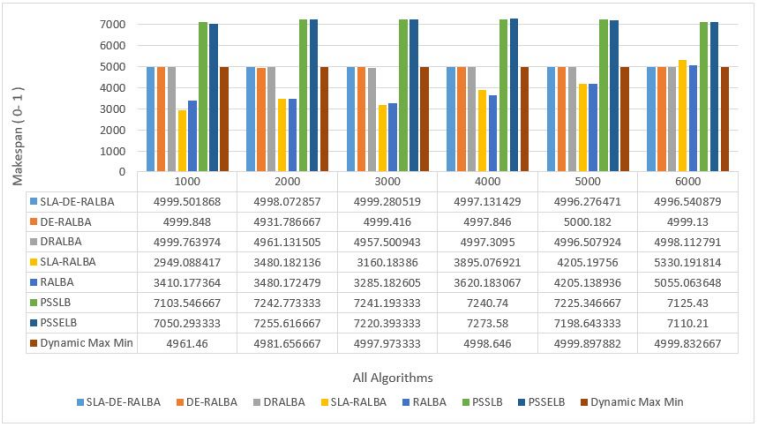
Fig. 3. ARUR using GoCJ dataset



	SLA-DE-RALBA	DE-RALBA	DRALBA	SLA_RALBA	RALBA	PSSLB	PSSELB	Dynamic Max Min
Mean ARUR	0.980062178	0.911551475	0.938441851	0.914652707	0.907876854	0.971724966	0.973219212	0.876521

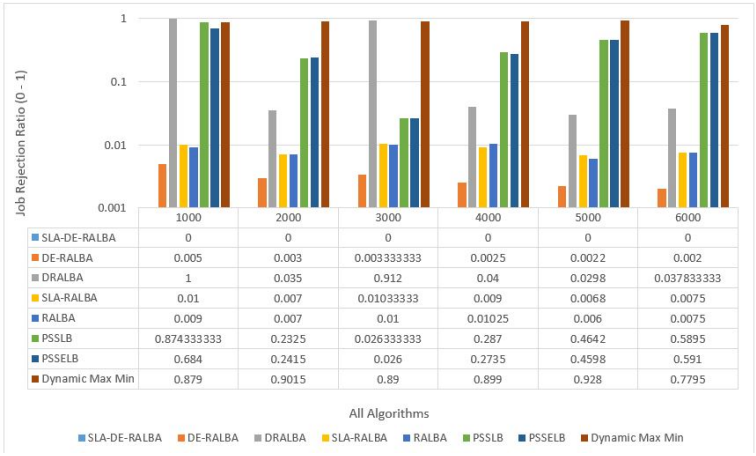
Fig. 4. Mean ARUR using GoCJ dataset

It slightly trails PSSLB and PSSELB at 1000 jobs but outperforms all baselines beyond 2000 jobs, with gains up to 28.9% over Dynamic Max-Min. At peak load (5000 jobs), it achieves the highest ARUR (0.987), confirming strong scalability and stable utilization. Overall, SLA-DE-RALBA delivers the best mean ARUR, improving over Dynamic Max-Min (11.9%), DE-RALBA (7.5%), RALBA (7.2%), SLA-RALBA (7.1%), DRALBA (4.5%), and marginally over PSSLB/PSSELB (0.8%). Figure 4 presents the mean ARUR results for all algorithms using GoCJ dataset.



	1000	2000	3000	4000	5000	6000
SLA-DE-RALBA	4999.501868	4998.072857	4999.280519	4997.131429	4996.276471	4996.540879
DE-RALBA	4999.848	4931.786667	4999.416	4997.846	5000.182	4999.13
DRALBA	4999.763974	4961.131505	4957.500943	4997.3095	4996.507924	4998.112791
SLA-RALBA	2949.088417	3480.182136	3160.18386	3895.076921	4205.19756	5330.191814
RALBA	3410.177364	3480.172479	3285.182605	3620.183067	4205.138936	5055.063648
PSSLB	7103.546667	7242.773333	7241.193333	7240.74	7225.346667	7125.43
PSSELB	7050.293333	7255.616667	7220.393333	7273.58	7198.643333	7110.21
Dynamic Max Min	4961.46	4981.656667	4997.973333	4998.646	4999.897882	4999.832667

Fig. 5. Makespan using GoCJ dataset



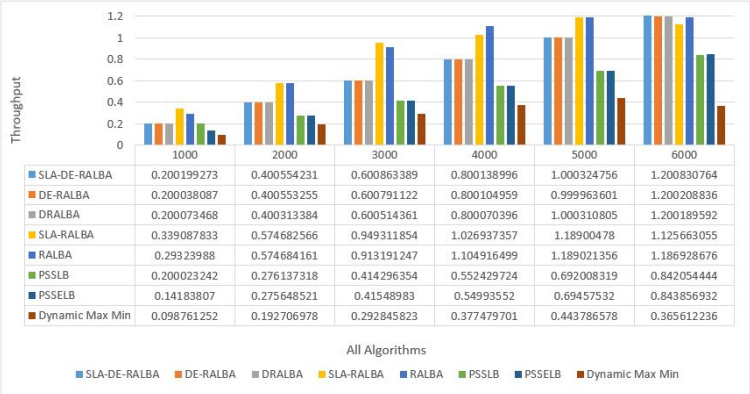
	1000	2000	3000	4000	5000	6000
SLA-DE-RALBA	0	0	0	0	0	0
DE-RALBA	0.005	0.003	0.003333333	0.0025	0.0022	0.002
DRALBA	1	0.035	0.912	0.04	0.0298	0.037833333
SLA-RALBA	0.01	0.007	0.01033333	0.009	0.0068	0.0075
RALBA	0.009	0.007	0.01	0.01025	0.006	0.0075
PSSLB	0.874333333	0.2325	0.026333333	0.287	0.4642	0.5895
PSSELB	0.684	0.2415	0.026	0.2735	0.4598	0.591
Dynamic Max Min	0.879	0.9015	0.89	0.899	0.928	0.7795

Fig. 6. Job Rejection Ratio using GoCJ dataset

Makespan measures overall job completion time, where lower values indicate better scheduling efficiency and faster execution, critical for meeting cloud service commitments.

As shown in Figure 5, SLA-DE-RALBA consistently achieves low makespan across all workloads. At 1000 jobs, it records 4999.50 time units, comparable to DE-RALBA and DRALBA, while for larger workloads (5000–6000 jobs), it remains stable between 4998–5000 time units. This stability is achieved through a multi-factor scheduling strategy that jointly considers resource availability, system utilization and SLA constraints at each decision point.

Figure 6 presents the Job Rejection Ratio (JRR), a key QoS metric measuring task rejections due to resource unavailability or deadline violations. SLA-DE-RALBA achieves zero job rejections across most GoCJ configurations, accepting more workloads than all baseline methods, while DRALBA shows the highest rejection rates at lower workload levels (1000 and 3000 jobs). This zero-rejection performance is enabled by three mechanisms: (1) capacity borrowing allowing high-priority tasks to use up to 95% of idle VM capacity, (2) proactive load rebalancing that redistributes tasks when VM load imbalance exceeds a threshold, and (3) dynamic VM share recalculation that adapts resources in real time with each incoming cloudlet.



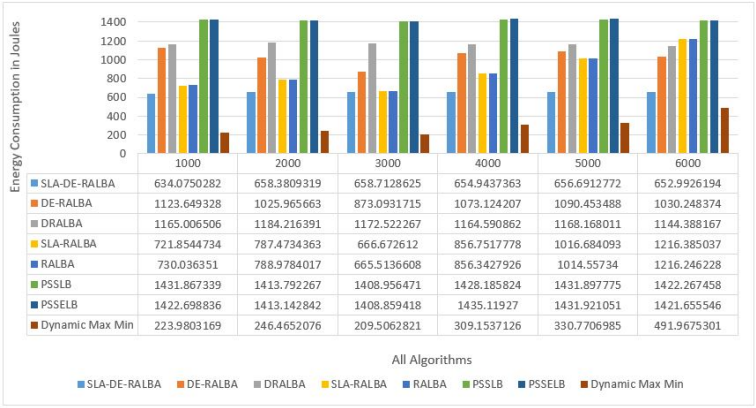
	1000	2000	3000	4000	5000	6000
SLA-DE-RALBA	0.200199273	0.400554231	0.600863389	0.800138996	1.000324756	1.200830764
DE-RALBA	0.200038087	0.400553255	0.600791122	0.800104959	0.999963601	1.200208836
DRALBA	0.200073468	0.400313384	0.600514361	0.800070396	1.000310805	1.200189592
SLA-RALBA	0.339087833	0.574682566	0.949311854	1.026937357	1.18900478	1.125663055
RALBA	0.29323988	0.574684161	0.913191247	1.104916499	1.189021356	1.186928676
PSSLB	0.200023242	0.276137318	0.414296354	0.552429724	0.692008319	0.842054444
PSSELB	0.14183807	0.275648521	0.41548983	0.54993552	0.69457532	0.843856932
Dynamic Max Min	0.098761252	0.192706978	0.292845823	0.377479701	0.443786578	0.365612236

Fig. 7. Throughput using GoCJ dataset

Throughput measures the number of jobs completed over time, reflecting system processing capacity and overall efficiency. Higher throughput indicates better ability to handle concurrent workloads and improved service quality. As shown in Figure 7, SLA-DE-RALBA consistently achieves superior throughput across all workload scales, recording 0.20019 at 1000 jobs and maintaining strong performance at larger datasets. It significantly outperforms PSSLB and PSSELB, enabling faster job completion and higher system efficiency. This improvement is driven by effective load balancing that maximizes task allocation and optimizes resource utilization.

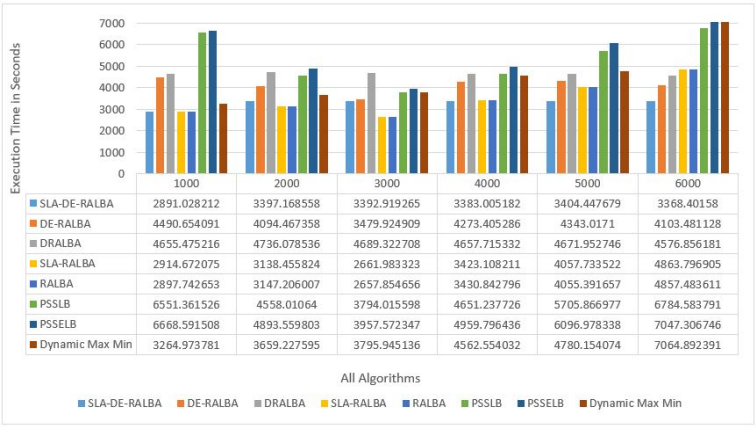
Energy consumption reflects resource efficiency and sustainability in cloud computing. As shown in Figure 8, SLA-DE-RALBA maintains low and stable energy usage

(634–652 Joules) across all workloads (1000–6000 jobs). It achieves about 55–60% energy savings compared to PSSLB and PSSELB (1422–1435 Joules). Although Dynamic Max-Min consumes less energy (223–491 Joules), it does so at the cost of poor resource utilization, throughput, and execution performance.



	1000	2000	3000	4000	5000	6000
SLA-DE-RALBA	634.0750282	658.3809319	658.7128625	654.9437363	656.6912772	652.9926194
DE-RALBA	1123.649328	1025.965663	873.0931715	1073.124207	1090.453488	1030.248374
DRALBA	1165.006506	1184.216391	1172.522267	1164.590862	1168.168011	1144.388167
SLA-RALBA	721.8544734	787.4734363	666.672612	856.7517778	1016.684093	1216.385037
RALBA	730.036351	788.9784017	665.5136608	856.3427926	1014.55734	1216.246228
PSSLB	1431.867339	1413.792267	1408.956471	1428.185824	1431.897775	1422.267458
PSSELB	1422.698836	1413.142842	1408.859418	1435.11927	1431.921051	1421.655546
Dynamic Max Min	223.9803169	246.4652076	209.5062821	309.1537126	330.7706985	491.9675301

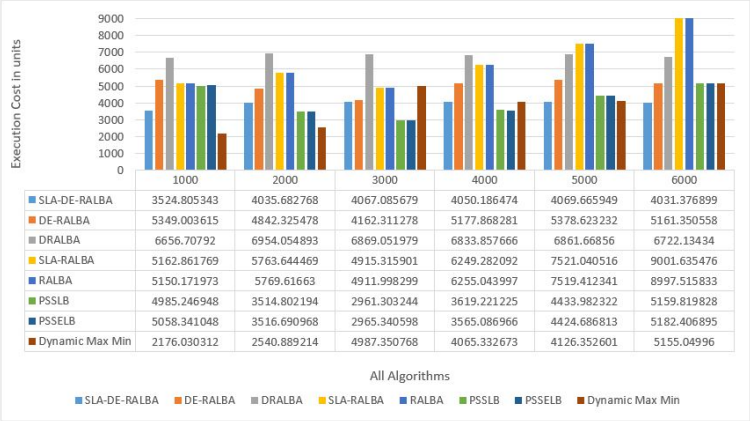
Fig. 8. Energy Consumption using GoCJ dataset



	1000	2000	3000	4000	5000	6000
SLA-DE-RALBA	2891.028212	3397.168558	3392.919265	3383.005182	3404.447679	3368.40158
DE-RALBA	4490.654091	4094.467358	3479.924909	4273.405286	4343.0171	4103.481128
DRALBA	4655.475216	4736.078536	4689.322708	4657.715332	4671.952746	4576.856181
SLA-RALBA	2914.672075	3138.455824	2661.983323	3423.108211	4057.733522	4863.796905
RALBA	2897.742653	3147.206007	2657.854656	3430.842796	4055.391657	4857.483611
PSSLB	6551.361526	4558.01064	3794.015598	4651.237726	5705.866977	6784.583791
PSSELB	6668.591508	4893.559803	3957.572347	4959.796436	6096.978338	7047.306746
Dynamic Max Min	3264.973781	3659.227595	3795.945136	4562.554032	4780.154074	7064.892391

Fig. 9. Execution Time using GoCJ dataset

In Figure 9, SLA-DE-RALBA achieves consistently low execution time across all GoCJ workloads, with only minor variation (500 units), indicating strong scalability. This performance results from its SLA-based priority scheduling and multi-layer optimization, which prioritize high-importance tasks while efficiently allocating resources.



	1000	2000	3000	4000	5000	6000
SLA-DE-RALBA	3524.805343	4035.682768	4067.085679	4050.186474	4069.665949	4031.376899
DE-RALBA	5349.003615	4842.325478	4162.311278	5177.868281	5378.623232	5161.350558
DRALBA	6656.70792	6954.054893	6869.051979	6833.857666	6861.66856	6722.13434
SLA-RALBA	5162.861769	5763.644469	4915.315901	6249.282092	7521.040516	9001.635476
RALBA	5150.171973	5769.61663	4911.998299	6255.043997	7519.412341	8997.515833
PSSLB	4985.246948	3514.802194	2961.303244	3619.221225	4433.982322	5159.819828
PSSELB	5058.341048	3516.690968	2965.340598	3565.086966	4424.686813	5182.406895
Dynamic Max Min	2176.030312	2540.889214	4987.350768	4065.332673	4126.352601	5155.04996

Fig. 10. Execution Cost using GoCJ dataset

The improved cost and resource utilization also reduce execution time, benefiting both users and cloud providers in terms of efficiency and cost savings.

Figure 10 shows that SLA-DE-RALBA is the most cost-efficient algorithm across all workload sizes, with costs ranging from 3524 units (1000 jobs) to 4100 units (6000 jobs), significantly lower than all baselines. Dynamic Max-Min is competitive only at low workloads but degrades at higher loads. PSSLB and SLA-RALBA show moderate cost efficiency, while RALBA, PSSELB, and DRALBA incur much higher costs, especially under heavy workloads, indicating poor scalability in cost optimization.

5.2. Performance Assessment Results - HCSP Dataset

The experiments are performed using HCSP benchmark dataset and these results for each parameter are presented in this subsection [33].

ARUR result in Figure 11 shows that SLA-DE-RALBA consistently achieves high resource utilization, with ARUR values exceeding 0.95 in centralized environments using HCSP dataset. RALBA and SLA-RALBA perform well under high-resource conditions, while PSSLB and PSSELB show strong results only when resources are abundant but degrade under constraints. DE-RALBA exhibits moderate performance in low-resource settings, whereas Dynamic Max-Min shows fluctuating results with inconsistent efficiency in distributed scenarios.

The makespan result (Figure 12) shows that SLA-DE-RALBA consistently achieves the best performance across different configurations, especially in U_CHIHI and U_S-LOLO scenarios. DE-RALBA maintains moderate and stable performance, while Dynamic Max-Min performs variably across conditions. PSSLB and PSSELB show moderate and consistent but less efficient makespan results.

Figure 13 shows that SLA-DE-RALBA maintains high job acceptance across all workload conditions, demonstrating strong reliability and effective workload handling. In contrast, DRALBA, DE-RALBA, SLA-RALBA, PSSLB, PSSELB, and Dynamic Max-Min exhibit varying levels of job rejection, especially under heavy loads, indicating limited

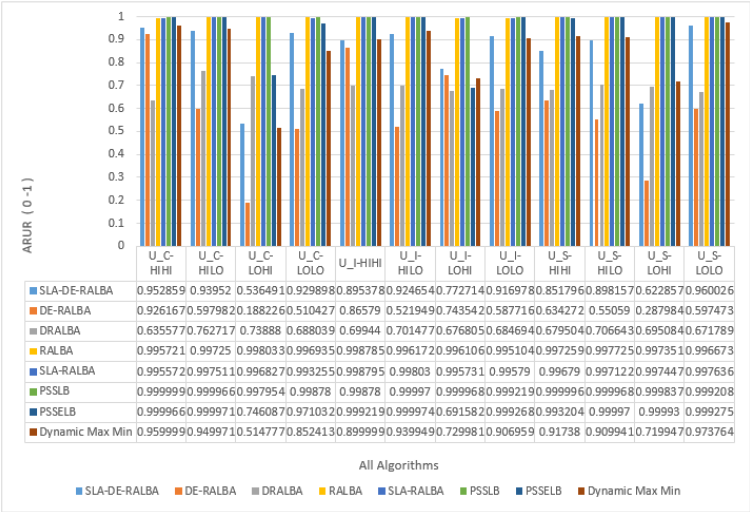


Fig. 11. ARUR using HCSP dataset

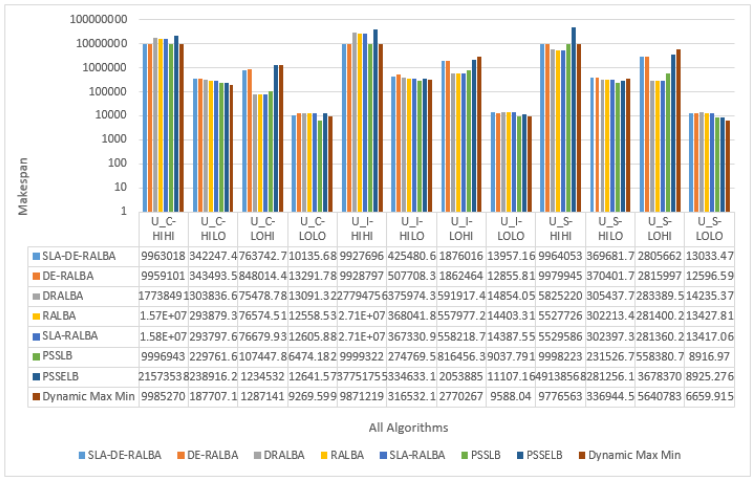


Fig. 12. Makespan using HCSP dataset

adaptability. Overall, SLA-DE-RALBA consistently ensures efficient job processing and SLA compliance across all configurations.

The throughput results using HCSP dataset (Figure 14) shows that SLA-DE-RALBA achieves optimal performance even under resource constraints. DE-RALBA, DRALBA, and RALBA maintain moderate, configuration-independent throughput, while PSSLB and PSSELB perform well only in LOLO scenarios. Overall, SLA-DE-RALBA delivers superior throughput, demonstrating the effectiveness of SLA-based scheduling for maximizing job completion efficiency. Energy consumption analysis (Figure 15) shows

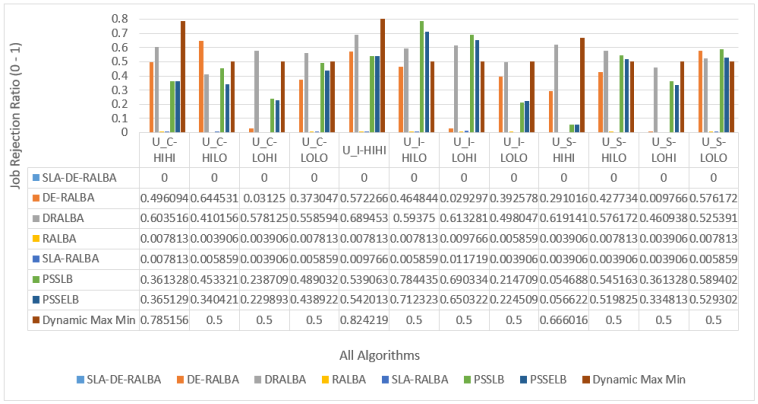


Fig. 13. Job Rejection Ratio using HCSP dataset

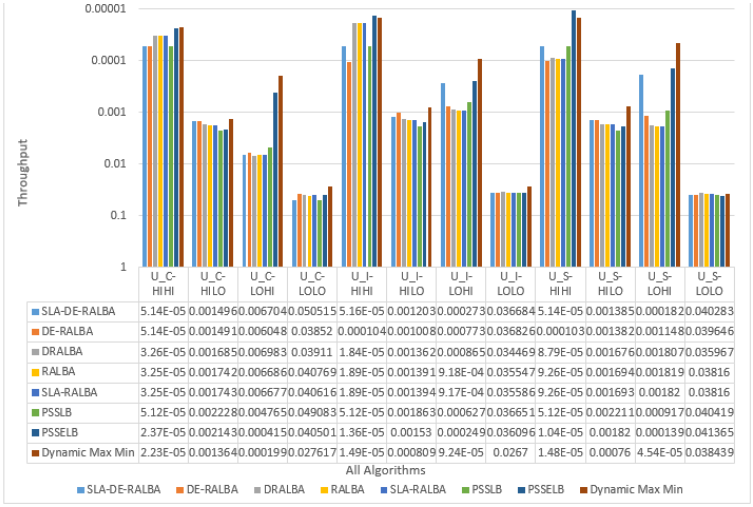
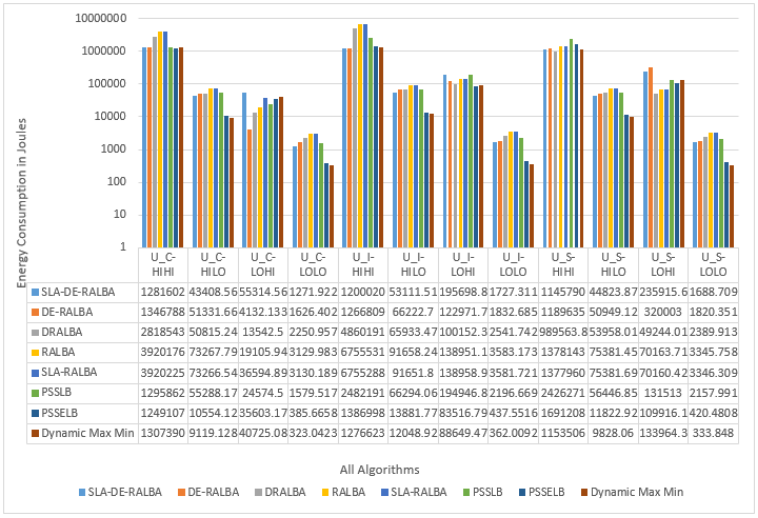


Fig. 14. Throughput using HCSP dataset

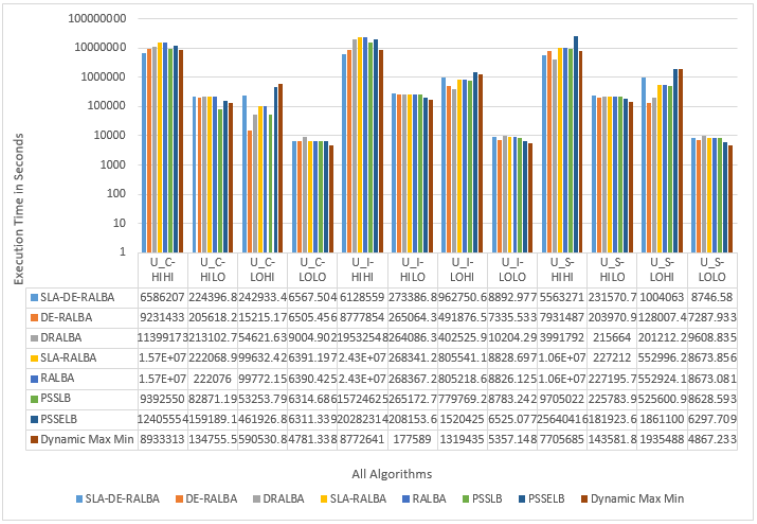
that RALBA and SLA-RALBA exhibit peak usage in U_IHIHI configurations. Dynamic Max-Min achieves the lowest energy consumption, reducing usage by 20–35% compared to RALBA/SLA-RALBA, 10–25% vs DE-RALBA, and 5–15% vs SLA-DE-RALBA, while PSSELB is the second most efficient with 15–30% savings. Although SLA-DE-RALBA is slightly less energy-efficient, it provides a better overall balance of resource utilization, makespan, and QoS than highly energy-optimized methods.

Figure 16 shows that SLA-DE-RALBA achieves consistently low execution times across both centralized and distributed models, setting the best benchmark performance. DE-RALBA provides stable moderate results, while PSSLB performs close to SLA-DE-RALBA. However, RALBA and SLA-RALBA exhibit significantly higher execution times in high-resource scenarios. PSSELB shows inconsistent performance, and DRALBA



	U_C-HIHI	U_C-HILO	U_C-LOHI	U_C-LOLO	U_I-HIHI	U_I-HILO	U_I-LOHI	U_I-LOLO	U_S-HIHI	U_S-HILO	U_S-LOHI	U_S-LOLO
SLA-DE-RALBA	1281602	43408.56	55314.56	1271.922	1200020	53111.51	195698.8	1727.311	1145790	44823.87	235915.6	1688.709
DE-RALBA	1346788	51331.66	4132.133	1626.402	1266809	66222.7	122971.7	1832.685	1189635	50949.12	320003	1820.351
DRALBA	2818543	50815.24	13542.5	2250.957	4860191	65933.47	100152.3	2541.742	989563.8	53958.01	49244.01	2389.913
RALBA	3920176	73267.79	19105.94	3129.983	6755531	91658.24	138951.1	3583.173	1378143	75381.45	70163.71	3345.758
SLA-RALBA	3920225	73266.54	36594.89	3130.189	6755288	91651.8	138958.9	3581.721	1377960	75381.69	70160.42	3346.309
PSSLB	1295862	55288.17	24574.5	1579.517	2482191	66294.06	194946.8	2196.669	2426271	56446.85	131513	2157.991
PSSELB	1249107	10554.12	35603.17	385.6658	1386998	13881.77	83516.79	437.5516	1691208	11822.92	109916.1	420.4808
Dynamic Max Min	1307390	9119.128	40725.08	323.0423	1276623	12048.92	88649.47	362.0092	1153506	9828.06	133964.3	333.848

Fig. 15. Energy Consumption using HCSP dataset

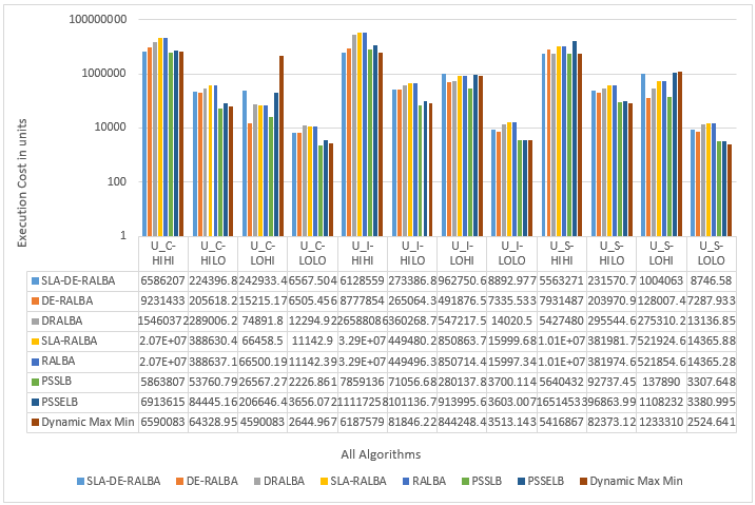


	U_C-HIHI	U_C-HILO	U_C-LOHI	U_C-LOLO	U_I-HIHI	U_I-HILO	U_I-LOHI	U_I-LOLO	U_S-HIHI	U_S-HILO	U_S-LOHI	U_S-LOLO
SLA-DE-RALBA	6586207	224396.8	242933.4	6567.504	6128559	273386.8	962750.6	8892.977	5563271	231570.7	1004063	8746.58
DE-RALBA	9231433	205618.2	15215.17	6505.456	8777854	265064.3	491876.5	7335.533	7931487	203970.9	128007.4	7287.933
DRALBA	11399173	213102.7	54621.63	9004.902	19532548	264086.3	402525.9	10204.29	3991792	215664	201212.2	9608.835
SLA-RALBA	1.57E+07	222068.9	99632.42	6391.197	2.43E+07	268341.2	805541.1	8828.697	1.06E+07	227212	552996.2	8673.856
RALBA	1.57E+07	222076	99772.15	6390.425	2.43E+07	268367.2	805218.6	8826.125	1.06E+07	227195.7	552924.1	8673.081
PSSLB	9392550	82871.19	53253.79	6314.686	15724625	265172.7	779769.2	8783.242	9705022	225783.9	525600.9	8628.593
PSSELB	1240555	4159189.	1461926.	86311.33	9202823	1420815	3.6 1520425	6525.077	2564041	6181923.6	1861100	6297.709
Dynamic Max Min	8933313	134755.5	590530.8	4781.338	8772641	177589	1319435	5357.148	7705685	143581.8	1935488	4867.233

Fig. 16. Execution Time using HCSP dataset

experiences sharply increasing execution times in both centralized and scattered configurations.

Figure 17 shows that PSSLB achieves cost performance close to SLA-DE-RALBA in most cases. Dynamic Max-Min maintains generally low and stable execution costs, making it a cost-efficient option. In contrast, RALBA and PSSELB incur significantly higher costs in U_J-HIHI scenarios, while DRALBA and SLA-RALBA further increase costs under high-resource workloads.



	U_C-HIHI	U_C-HILO	U_C-LOHI	U_C-LOLO	U_I-HIHI	U_I-HILO	U_I-LOHI	U_I-LOLO	U_S-HIHI	U_S-HILO	U_S-LOHI	U_S-LOLO
SLA-DE-RALBA	6586207	224396.8	242933.4	6567.504	6128559	273386.8	962750.6	8892.977	5563271	231570.7	1004063	8746.58
DE-RALBA	9231433	205618.2	15215.17	6505.456	8777854	265064.3	491876.5	7335.533	7931487	203970.9	128007.4	7287.933
DRALBA	1546037	2289006.2	74891.8	12294.92	2658808	360268.7	547217.5	14020.5	5427480	295544.6	275310.2	13136.85
SLA-RALBA	2.07E+07	388630.4	66458.5	11142.9	3.29E+07	449480.2	850863.7	15999.68	1.01E+07	381981.7	521924.6	14365.88
RALBA	2.07E+07	388637.1	66500.19	11142.39	3.29E+07	449496.3	850714.4	15997.34	1.01E+07	381974.6	521854.6	14365.28
PSSLB	5863807	53760.79	26567.27	2226.861	7859136	71056.68	280137.8	3700.114	5640432	92737.45	137890	3307.648
PSSELB	6913615	84445.16	20646.4	3656.072	11117258	101136.7	913995.6	3603.007	16514533	96863.99	1108232	3380.995
Dynamic Max Min	6590083	64328.95	4590083	2644.967	6187579	81846.22	844248.4	3513.143	5416867	82373.12	1233310	2524.641

Fig. 17. Execution Cost using HCSP dataset

5.3. Results Discussion

SLA-DE-RALBA achieves a zero job rejection ratio across all GoCJ and HCSP workloads. In contrast, competing methods show varying rejection rates, ranging from minor levels in DE-RALBA to very high rejection in PSSLB, PSSELB, DRALBA, and Dynamic Max-Min. This consistent performance is due to its SLA-aware admission control and dynamic load balancing, which prevent resource saturation and ensure all jobs are successfully processed under diverse workload conditions.

SLA-DE-RALBA demonstrates consistent and robust performance across GoCJ and HCSP datasets by integrating multi-tier SLA scheduling with dynamic, multi-objective optimization for SLA compliance, cost, energy efficiency, and load balancing. **First**, its SLA-priority cloudlet selection ensures high-priority jobs are executed first, eliminating job rejections entirely (JRR = 0), unlike Dynamic Max-Min (50%), PSSLB (up to 87.4%), and DE-RALBA (up to 0.5%). This prevents resource saturation through explicit admission control. **Second**, a weighted VM scoring model balances execution time (25%), cost (35%), utilization (30%), and load balance (10%), enabling lower execution cost and time without sacrificing throughput or makespan, while avoiding server overloading or underutilization. **Third**, its Dynamic Enhanced adaptive tuning continuously adjusts scheduling decisions under changing workloads, keeping makespan stable (4996–4999) and outperforming PSSLB and PSSELB by 41%–45% in efficiency and 36%–57% in execution time. **Fourth**, reactive cloudlet migration during VM idle states improves real-time utilization, achieving a stable ARUR of 0.96–0.98 even under fluctuating or SLA-constrained workloads.

Finally, SLA-DE-RALBA achieves the highest mean ARUR (0.980) and lowest energy consumption (223–657 Joules), significantly outperforming PSSLB, PSSELB, DE-RALBA, and DRALBA in energy efficiency while maintaining superior utilization and

strict SLA compliance, making it the most balanced and reliable approach for real-world cloud environments.

6. Limitations

Although SLA-DE-RALBA demonstrates improved performance compared to existing scheduling algorithms, this study presents several limitations.

The evaluation is conducted exclusively in the CloudSimPlus simulation environment and has not been validated on physical testbeds, which may limit the representation of a few practical deployment factors such as network variability, hardware failures, system-level overheads, and multi-tenant resource contention. The algorithm does not explicitly incorporate hardware fault-tolerance or security mechanisms, as its design prioritizes energy efficiency and performance; consequently, its behavior under hardware failures or security-critical conditions remains unexamined.

Furthermore, the execution cost model, based on unit price and execution time, ensures methodological consistency and comparability but provides a simplified representation of real-world cloud pricing which may involve additional factors such as data transfer, storage, and dynamic pricing.

7. Conclusion

This study addresses key challenges in cloud job scheduling, including SLA compliance, workload variability, heterogeneous resources, and energy-cost trade-offs. Existing methods are either SLA-aware but static or dynamic but SLA-agnostic, limiting overall effectiveness.

To overcome this, we proposed SLA-DE-RALBA, implemented in the CloudSim Plus environment, integrating multi-tier SLA prioritization with dynamic optimization. Experimental results across multiple datasets show that SLA-DE-RALBA consistently outperforms existing algorithms, achieving the highest resource utilization (ARUR = 0.980), lower execution cost and makespan, and zero job rejection while maintaining stable SLA performance.

8. Future Work

There are a number of promising directions for the extension of this work. There are scalability studies that may be made to investigate the multi datacenter designs as well as the cloud fog edge paradigms of computation. Network and storage intelligence and computational resource could be part of the holistic resource optimization. The basic implementation would be tested in practice in platforms such as OpenStack or Kubernetes which would offer useful deployment information. Workload prediction and SLA- risk analysis could be improved by machine learning integration to increase forecast abilities. Multi- objective optimization models may facilitate detailed trade-offs between SLA compliance, energy-saving and cost-cutting. Lastly, it would be desirable to add fault tolerance features like failure identification and traffic redirection in order to enhance the reliability of the system in the most important applications.

References

1. Mell, P., Grance, T.: The NIST-National Institute of Standards and Technology- Definition of Cloud Computing. NIST Spec. Publ. 800-145, 1–3 (2011)
2. Armbrust, M. et al.: A view of cloud computing. Commun. ACM 53(4), 50–58 (2010). doi:10.1145/1721654.1721672
3. Marinos, A., Briscoe, G.: Community cloud computing. Lect. Notes Comput. Sci. 5931, 472–484 (2009). doi:10.1007/978-3-642-10665-1_43
4. Vaquero, L.M., Rodero-Merino, L., Caceres, J., Lindner, M.: A break in the clouds. ACM SIGCOMM Comput. Commun. Rev. 39(1), 50–55 (2008). doi:10.1145/1496091.1496100
5. Ajmera, K., Tewari, T.K., Singh, V.K., Vikash, Upadhyay, P.K.: Energy-Aware Dynamic Virtual Machine Scheduling in Cloud Computing: A Survey. In: Proc. 2023 Fifteenth Int. Conf. Contemp. Comput. (IC3-2023), Noida, India, pp. 133–141 (2023). doi:10.1145/3607947.3607970
6. Awari, P.: Analyzing and Evaluating The Role of Cloud Computing in the Scalability of Web Applications. Int. Res. J. Mod. Eng. Technol. Sci. 7(2), 1–7 (2025)
7. Abraham, O.L., Bin Ngadi, M.A., Bin Mohamad Sharif, J., Sidik, M.K.M.: Task Scheduling in Cloud Environment-Techniques, Applications, and Tools: A Systematic Literature Review. IEEE Access 12, 138252–138279 (2024). doi:10.1109/ACCESS.2024.3466529
8. Koubàa, M., Karar, A.S., Bahloul, F.: Optimizing Scheduled Virtual Machine Requests Placement in Cloud Environments: A Tabu Search Approach. Computers 13(12), 1–25 (2024). doi:10.3390/computers13120321
9. Hussain, A., Aleem, M., Islam, M.A., Iqbal, M.: A rigorous evaluation of state-of-the-art scheduling algorithms for cloud computing. IEEE Access 6, 75033–75047 (2018). doi:10.1109/ACCESS.2018.2884480
10. Ibrahim, M. et al.: An In-Depth Empirical Investigation of State-of-the-Art Scheduling Approaches for Cloud Computing. IEEE Access 8, 128282–128294 (2020). doi:10.1109/ACCESS.2020.3007201
11. Rehan, H.: Revolutionizing America's Cloud Computing the Pivotal Role of AI in Driving Innovation and Security. J. Artif. Intell. Gen. Sci. 2(1), 189–208 (2024). doi:10.60087/jaigs.v2i1.p208
12. Ebadifard, F., Babamir, S.M., Barani, S.: A Dynamic Task Scheduling Algorithm Improved by Load Balancing in Cloud Computing. In: Proc. 2020 6th Int. Conf. Web Res. (ICWR), pp. 177–183 (2020). doi:10.1109/ICWR49608.2020.9122287
13. Albalawi, N.S.: Dynamic scheduling strategies for cloud-based load balancing in parallel and distributed systems. J. Cloud Comput. 14(1) (2025). doi:10.1186/s13677-025-00757-6
14. Chandrasiri, S., Meedeniya, D.: Energy-Efficient Dynamic Workflow Scheduling in Cloud Environments Using Deep Learning. Sensors 25(5) (2025). doi:10.3390/s25051428
15. Choudhary, A., Govil, M.C., Singh, G., Awasthi, L.K., Pilli, E.S.: Energy-aware scientific workflow scheduling in cloud environment. Cluster Comput. 25(6), 3845–3874 (2022). doi:10.1007/s10586-022-03613-3
16. Hussain, A., Aleem, M., Iqbal, M.A., Islam, M.A.: SLA-RALBA: cost-efficient and resource-aware load balancing algorithm for cloud computing. J. Supercomput. 75(10), 6777–6803 (2019). doi:10.1007/s11227-019-02916-4
17. Nabi, S., Ahmed, M.: OG-RADL: overall performance-based resource-aware dynamic load-balancer for deadline constrained Cloud tasks. J. Supercomput. 77(7) (2021). doi:10.1007/s11227-020-03544-z
18. Mostafa, S.A.M., Mohan, A., Wang, J.: SLA-MORL: SLA-Aware Multi-Objective Reinforcement Learning for HPC Resource Optimization. arXiv:2508.03509 (2025). Available: http://arxiv.org/abs/2508.03509
19. Gogoi, B., Devaraj Verma C, D., Panggabean, C., Limbu, R., Sarker, R.: Optimized Cloud Resource Allocation Using Genetic Algorithms for Energy Efficiency and QoS Assurance. J. Emerg. Technol. Innov. Res. 12(5) (2025). doi:10.56975/jetir.v12i5.561700

20. Arshad, U., Aleem, M., Srivastava, G., Lin, J.C.W.: Utilizing power consumption and SLA violations using dynamic VM consolidation in cloud data centers. Renew. Sustain. Energy Rev. 167, 112782 (2022). doi:10.1016/j.rser.2022.112782
21. Mohammad, N.I.S.: Data Scheduling Algorithm for Scalable and Efficient IoT Sensing in Cloud Computing. arXiv:2508.04334 (2025). Available: http://arxiv.org/abs/2508.04334
22. Siddesha, K., Jayaramaiah, G.V.: Energy Efficient Greedy Scheduling of Tasks for DVFS Enabled Heterogeneous Multicore Processors. In: Proc. 2021 6th Int. Conf. Recent Trends Electron. Information, Commun. Technol. (RTEICT), pp. 538–542 (2021). doi:10.1109/RTEICT52294.2021.9573873
23. Hussain, A., Aleem, M., Rehman, A.U., Arshad, U.: DE-RALBA: dynamic enhanced resource aware load balancing algorithm for cloud computing. PeerJ Comput. Sci. 11, 1–27 (2025). doi:10.7717/PEERJ-CS.2739
24. Alaei, N., Safi-Esfahani, F.: RePro-Active: a reactive–proactive scheduling method based on simulation in cloud computing. J. Supercomput. 74(2), 801–829 (2018). doi:10.1007/s11227-017-2161-0
25. Mao, Y., Chen, X., Li, X.: Max–Min task scheduling algorithm for load balance in cloud computing. Adv. Intell. Syst. Comput. 255, 457–465 (2014). doi:10.1007/978-81-322-1759-6_53
26. Nabi, S., Ibrahim, M., Jimenez, J.M.: DRALBA: Dynamic and Resource Aware Load Balanced Scheduling Approach for Cloud Computing. IEEE Access 9, 61283–61297 (2021). doi:10.1109/ACCESS.2021.3074145
27. Ma, Z., Ma, D., Lv, M., Liu, Y.: Virtual Machine Migration Techniques for Optimizing Energy Consumption in Cloud Data Centers. IEEE Access 11, 86739–86753 (2023). doi:10.1109/ACCESS.2023.3305268
28. Kumar, M., Sharma, S.C.: Dynamic load balancing algorithm to minimize the makespan time and utilize the resources effectively in cloud environment. Int. J. Comput. Appl. 42(1), 108–117 (2020). doi:10.1080/1206212X.2017.1404823
29. Chauhan, N. et al.: A Systematic Literature Review on Task Allocation and Performance Management Techniques in Cloud Data Center. Comput. Syst. Sci. Eng. 48(3), 571–608 (2024). doi:10.32604/csse.2024.042690
30. Hussain, A., Aleem, M., Khan, A., Iqbal, M.A., Islam, M.A.: RALBA: a computation-aware load balancing scheduler for cloud computing. Cluster Comput. 21(3), 1667–1680 (2018). doi:10.1007/s10586-018-2414-6
31. Shafiq, D.A., Jhanjhi, N.Z., Abdullah, A., Alzain, M.A.: A Load Balancing Algorithm for the Data Centres to Optimize Cloud Computing Applications. IEEE Access 9, 41731–41744 (2021). doi:10.1109/ACCESS.2021.3065308
32. Hussain, A., Aleem, M.: GoCJ: Google cloud jobs dataset for distributed and cloud computing infrastructures. Data 3(4), 1–12 (2018). doi:10.3390/data3040038
33. Tracy, D.B.: HCSP: heterogeneous computing scheduling problem (2000). Available: https://www.fing.edu.uy/inco/grupos/cecal/hpc/HCSP/index.htm
34. Arshed, J.U., Ahmed, M.: RACE: Resource Aware Cost-Efficient Scheduler for Cloud Fog Environment. IEEE Access 9, 65688–65701 (2021). doi:10.1109/ACCESS.2021.3068817
35. Venkatesh, C., Shiva Murthy, G.: Dynamic min-max algorithm for resource provisioning in cloud environment. Indian J. Comput. Sci. Eng. 12(3), 692–700 (2021). doi:10.21817/indjcse/2021/v12i3/211203161
36. Hagras, T., El-Sayed, G.A.: A fault-tolerant and load-balancing scheduler for independent tasks on cloud-based virtual machines. Cluster Comput. 29(61), 1–13 (2026). doi:10.1007/s10586-025-05857-1
37. Zhu, X.: Hybrid ant colony and whale optimization algorithms for QoS-aware load balancing in cloud computing environment. J. Eng. Appl. Sci. 72(169), 1–26 (2025). doi:10.1186/s44147-025-00739-0

38. Beena, B.M., Ranga, P.C., Holimath, V., Sridhar, S., Kamble, S.S., Shendre, S.P., Priya, M.Y.: Adaptive energy optimization in cloud computing through containerization. IEEE Access 13, 159548–159565 (2025). doi:10.1109/ACCESS.2025.3606932

Mohsin Nawaz received the B.S. in Computer Science from the University of Engineering and Technology, Taxila, Pakistan and successfully completed his M.S. in Computer Science at the Department of Computer Science, KICSIT Campus, Institute of Space Technology, Islamabad, Pakistan. His research interests include software development and testing, cloud computing, and distributed computing, with a particular focus on performance analysis and resource optimization.

Altaf Hussain is serving as professor and Head of the Department of Computer Science at KICSIT, Institute of Space Technology (IST), Islamabad, Pakistan. He completed his PhD in Computer Science under the prestigious HEC Indigenous PhD Fellowship Program, with research contributions published in leading impact-factor journals including IEEE Access, Cluster Computing, The Journal of Supercomputing, MDPI Data, and Computing and Informatics. He has also contributed to international conferences and book chapters, demonstrating a strong and consistent research portfolio. He graduated with distinction in BS Computer Science and was awarded a Bronze Medal for academic excellence. He is reviewer for many journals and serves as an Academic Editor for PLOS ONE. His research interests include Information Security, Blockchain, Artificial Intelligence, Data Mining, Cloud Computing, High Performance Computing, Big Data Analytics, Software Engineering, and Outcome-Based Education, with a focus on scalable systems, performance optimization, and data-driven solutions for modern computing challenges.

Marran Al Qawaid received PhD in Computer Science from University of North Carolina. He is currently an Assistant Professor at Shaqra University, Saudi Arabia. His research interest includes cyber security, Artificial intelligence, Image processing, and computer vision.

Muazzam A. Khan Khattak is currently serving as a Member of the NADRA Authority Board and has been listed among the top 2% most highly cited scientists worldwide in 2024 and 2025. He is the recipient of Gold Medals from the Pakistan Academy of Sciences-COMSTECH and the IEEE Islamabad Section. At present, he is Director of the Farab Center for Artificial Intelligence and Cyber Security, Director of the ICESCO Chair in Data Analytics and Edge Computing and Director of Science Technology at Quaid-i-Azam University, Islamabad. He is also chairperson of UNESCO National IFAP Committee. Earlier, he served as Associate Dean of Computing and Head of the Postgraduate Program at SEECS, NUST. He completed his PhD through a sandwich program between IIUI and the University of Missouri, (UMKC), USA, followed by a postdoctoral fellowship at UMKC, where he later joined as an adjunct professor. Prof. Khattak has authored 3 books and 16 book chapters with leading international publishers, in addition to more than 240 articles in reputed international journals and conferences. His work has earned him recognition as one of the world's highly cited scientists, with over 7000 citations and an h-index of 45. His research focuses on Artificial Intelligence based smart solutions for Internet of Things, Internet of Vehicles, Smart Cities, and Information Security. He also holds adjunct

professorships at the University of Missouri (USA), Texas American Muslim Universities (USA), Harbin Institute of Technology (China), USAK University (Türkiye), BAKU State University (Azerbaijan), and Margullan University (Kazakhstan). He has been honored with the Pak Academy of Sciences Gold Medal, the Best Performance Award by the Ministry of Education in 2023, and multiple performance-based awards at NUST and QAU from 2015 till 2025, as well 2 best papers awards at International conferences. He has also been nominated for the prestigious civil awards Tamgha-e-Imtiaz and Sitara-e-Imtiaz. Over the course of his career, he has secured national and international research grants amounting to Rs. 1.7 billion from institutions including the Ministry of Education (Kazakhstan), the British Council, the University of Missouri (USA), USAK University (Türkiye), and Margullan University (Kazakhstan). Prof. Khattak serves on international advisory boards at TEXAM (USA), USAK University (Türkiye), Baku State University (Azerbaijan), Harbin Institute of Technology (China) and International Research Network on AI Development and Governance (AIR Net), China. He is also member Senate/BOG at Mirpur University of Science and Technology, AJK and Islamabad University of Health Sciences and Emerging Technologies, Islamabad.

Received: May 23, 2026; Accepted: June 23, 2026.

Computer Science and Information Systems 23(4):1321–1342 https://doi.org/10.2298/CSIS260219036C

Paper Type: Research Paper

VR-ETL: Leveraging Virtual Reality for Pre-Transformation ETL Data Validation

Maja Cerjan[1], Kornelije Rabuzin[1], and Leo Mršić[2]

[1] Faculty of Organization and Informatics, University of Zagreb,
Pavlinska 2, Varaždin, Croatia,
Department of Theoretical and Applied Foundations of Information Sciences
macerjan@foi.unizg.hr (corresponding author),
krabuzin@foi.unizg.hr
[2] Algebra Bernays University,
Gradišćanska 24, 10000 Zagreb, Croatia
leo.mrsic@algebra.hr

Abstract. Data validation in ETL workflows is most often performed after transformation, even though many data quality issues originate earlier in the pipeline. In practice, this means that errors are detected late, when tracing their source and applying corrections already requires additional effort. In this work, we explore whether virtual reality (VR) could be used as an alternative interface for addressing this problem earlier, at the stage of raw data inspection. We present VR-ETL, a proof-of-concept system that allows users to inspect and correct tabular data before transformation by interacting with it inside an immersive 3D environment. The system imports CSV datasets and applies a set of rule-based validation checks targeting common data quality problems, such as missing values, format inconsistencies, and basic referential integrity violations. Detected issues are visualized directly within the 3D representation of the data, using color-coded cells and simple spatial summaries that reflect error distribution across columns and records. A functional prototype was implemented using Unity and OpenXR and tested on structured datasets containing up to 10,000 records. During controlled technical testing, the system supported automated validation, real-time visual feedback, and immediate re-validation after user corrections, while maintaining performance suitable for immersive use on consumer-grade hardware. All intentionally seeded rule-based errors were detected. Although usability and workflow efficiency were not evaluated in this study, the prototype demonstrates that pre-transformation data validation in VR is technically feasible and provides a concrete starting point for future empirical investigations.

Keywords: Virtual Reality, ETL, Error Handling, Data Validation, Immersive Analytics.

1. Introduction

In recent years, the importance of data-driven decision-making for finance, healthcare, and business intelligence operations of many organizations has received a lot of attention [16]. The worth of these findings depends on how well analysts process data and how

well the first set of records was recorded. The process of data quality maintenance has progressed from basic relational database restrictions into an intricate problem which handles multiple data formats and origins [5]. The evaluation of unprocessed data becomes unreliable because these datasets contain missing information together with formatting problems and structural irregularities [28]. Organizations with extensive operations implement Extract, Transform, and Load (ETL) frameworks to manage their data integration and cleaning operations [38,30]. The frameworks have achieved technological readiness but data quality management stands as the main barrier which prevents their implementation. The ETL process follows a standard error detection approach, which happens after transformation steps and because of this it requires longer procedures to track and fix records which leads to increased costs and longer processing times [35].

The system needs human touch, which means that individuals must verify error reports and it requires data engineers to interact with sources through multiple communication stages. All this operations results in delayed delivery of prepared data. Although automated algorithms for detecting errors have become more effective, analysts still expend a large amount of time and effort manipulating data and assessing quality issues, a process often referred to as "data wrangling" [17]. Research on immersive analytics (IA) suggests that users can perform better in pattern and anomaly detection tasks through 3D spatial interfaces rather than conventional 2D displays [14,7]. Virtual Reality (VR) environments have been successful in scientific visualization and financial modeling, reporting higher spatial interaction capabilities than spreadsheet-based reports [32].

The current state of research lacks complete understanding about how immersive technologies create new methods to verify data at its initial appearance. The current research on IA focuses on performing high-level data exploration tasks which help users understand already cleaned datasets. There is a lack of research and tools that apply VR specifically to the left side (pre-transformation stage) of the ETL process. Most ETL tools depend on 2D logs and dashboards which fail to use human spatial memory and immersive perception to detect structural data clusters and errors before they spread through the data pipeline.

In this work, we present VR-ETL, a proof-of-concept system that uses virtual reality as an alternative interface for addressing data quality issues during raw data inspection. VR-ETL allows users to inspect and correct tabular data before transformation by interacting with it inside a 3D environment. Our main contribution is the development of a spatial validation engine that maps rule-based errors (e.g., missing values, format inconsistencies) directly onto a 3D grid. However, we must explicitly note that this prototype is currently limited to CSV datasets and has been technically validated for performance with up to 10,000 records; scaling to larger datasets remains a challenge for future iterations.

To guide this exploratory study, the following research questions were defined:

- What are the latency and rendering limitations when visualizing tabular ETL data in a 3D VR environment?
- Is it technically feasible to represent rule-based validation errors spatially to improve early-stage detection?
- To what extent does a rule-based VR interface support real-time interaction for pre-transformation data correction?

The remainder of this paper is organized as follows: Section 2 provides a review of related literature on ETL technology, data quality frameworks, and immersive analytics.

Section 3 describes the system architecture and the logic of the validation engine. Section 4 details the technical implementation using Unity and OpenXR. Section 5 evaluates the results, providing an explicit discussion of benefits and limitations, while Section 6 concludes the paper and suggests directions for future research.

2. Related Work

The Extract Transform Load (ETL) operations generate the main performance problem which prevents organizations from using data for decision-making because their current systems encounter two essential problems which include extended error detection times and requirement for human verification [35,28,30]. The following section unites previous research to demonstrate VR-ETL functions as a bridge which links ETL system architectures to interactive data exploration techniques.

2.1. ETL Challenges and Real-Time Architectures

The software processing system which handles data loading and scheduled data updates operates under the name Extraction-Transformation-Loading (ETL) as it is a core function in modern data warehousing [38]. Modern ETL pipelines face their main restriction because they use reactive error management which performs validation after transformation operations which requires expensive system rollbacks [28,30,23]. The situation becomes more critical because data analysts need to perform most of their work on "data wrangling" which requires them to perform manual and repetitive data transformation and verification tasks. The main issue is that these operations remain untrackable through conventional 2D interfaces [17]. Real-time architectures achieve better output quality because their automated logging systems operate without human involvement for error correction according to Biswas et al. [6,23]. The proposed solution for these complex issues involves a direct physical Data Vault design which functions as an agile architectural method to boost ETL transformation speed and expand large-scale enterprise system capabilities [18]. The current system requires operational pre-transformation systems to run as separate units which need to identify all system irregularities at their first occurrence before human validation takes place to prevent errors from spreading through the pipeline. Moreover, other researchers have explored the use of design in enhancing ETL validation, however, only within particular domains. For instance, Saebao et al. [33] offered a method of designing ETL in business intelligence systems while considering data quality. As stated in their paper, the analysis of data quality helps make more informed decisions regarding validation steps prior to data loading. Likewise, Mhon and Kham [24] noted that when dealing with multiple sources of information, you should consider validation based on your particular scenario. The studies confirm the hypothesis that validation of input data prior to transformation is not a universal challenge that requires universal solutions, but rather a task that needs customized tools to be solved efficiently. Such a tool would be VR-ETL that introduces a validation layer in an immersive space.

2.2. Data Visualization and Business Intelligence (BI)

Existing Business Intelligence (BI) and visualization tools focus on analyzing clean post-hoc data rather than supporting interactive quality assessment of raw input [15,36]. This

focus on post-hoc analysis often ignores the fact that maintaining a high level of data quality (DQ) is a key success factor for any data warehousing system, though DQ aspects are frequently poorly defined during the initial phases of development [27]. Organizations encounter rising operational complexity because they must process multiple data types which enter their systems through various sources which complicates their ability to control DQ dimensions [5]. The current error reporting system depends on text-based output which uses static dashboards and logs that do not provide the necessary interactive features for repairing complex high-dimensional data [2,29]. Research studies about software tools for industrial real-time information processing and visualization have shown that numerous tools exist but most systems monitor performance instead of providing interactive data quality intervention capabilities [37]. The lack of interactivity is critical because identifying DQ dimensions such as completeness, accuracy, and consistency is essential for constructing a reliable and accurate analysis process [34]. As emphasized by Mausa et al. [22] in ComSIS, standardized protocols for data collection and validation are necessary to evaluate data quality and minimize dataset bias. Beyond basic formatting, technical DQ also involves complex constraints like inverse referential integrity (IRICs), which are often disregarded by standard database management systems, forcing users to rely on custom, non-interactive procedures [1]. Tools such as Trifacta (previously named Wrangler) and Talend Data Preparation offer advanced two-dimensional visualization through histograms and quality bars. Unfortunately, their interface is still limited to screen visualization. On the other hand, VR-ETL intends to improve this practice through Immersive Analytics (IA). IA allows us to leave behind the dashboard visualization in favor of three-dimensional visualization where the spatial memory can be used to detect the presence of errors within multi-dimensional records.

2.3. Immersive Analytics and VR Visualization

Virtual Reality technology shows better results in pattern recognition and anomaly detection than traditional 2D dashboard systems according to Kwon et al. [19] and Bowman et al. [7]. Fonnet and Prie [14] provide a comprehensive survey of this field, tracing the evolution of immersive analytics (IA) from the early nineties to the present day and highlighting how rendering and interaction means support these complex tasks. Data specialists navigate 3D environments using analytics tools [8] and apply spatial memory to gain a more direct understanding of complex records [20,11]. The shift toward “Post-WIMP” interactions in visualization systems holds great promise for providing more natural experiences and empowering users to effectively complete analysis tasks [21]. Research studies about open-ended data exploration demonstrate that VR technology provides users with exact and detailed insights which lead to better satisfaction than conventional 2D visualization systems [25].

Research studies prove that these advantages work for particular fields which Donalek et al. demonstrate exist in astronomy [10] and trajectory data [13]. The field of collaborative visual analytics developed through the creation of Uplift which allows users to study data by using 3D models and augmented reality visualization tools for performing joint data analysis [12]. The integration of embedded data representations connects the data to the context that the data refers to. People can see clearly how the data matches the world when the integration of embedded data representations is in place [40]. However,

these systems focus on exploratory analysis of analytical results [3,4,39] rather than the operational cleaning of raw data during the ETL loading phase.

2.4. Summary of Research Gaps

As demonstrated in Table 1, VR-ETL addresses three critical gaps identified in the literature:

1. Temporal Gap: Shifting from reactive post-transformation logs to proactive pre-transformation validation [28].
2. Interaction Gap: Moving from script-based configuration to direct spatial manipulation in VR [30].
3. Application Gap: Applying immersive analytics specifically to data engineering tasks rather than final data analysis [3,20].

Table 1. Synthesis of Research Gaps in existing ETL and Immersive Analytics literature

Study	Key Contribution	Gaps Identified	Focus Area
Nwokeji & Matovu (2021)	Systematic review of ETL challenges	Focus remains on post-transformation error handling and manual corrections	ETL Error Handling
Patel & Patel (2020)	Evolution analysis of modern ETL tools	Lack of interactive, human-in-the-loop error correction mechanisms	ETL Error Handling
Simitsis et al. (2023)	Comprehensive history and future of ETL technology	Persistent challenges in real-time validation and user feedback loops	ETL Error Handling
Mehmood & Anees (2022)	Distributed real-time ETL architecture	Absence of interactive real-time correction within the stream	Real-Time ETL
Biswas et al. (2019)	Incremental loading for large-scale data integration	No VR-enhanced validation; limited to traditional 2D script monitoring	Real-Time ETL
Souibgui et al. (2020)	BI-integrated ETL systems	Primarily assumes "clean" input data; lacks error visualization	Data Vis. & BI
Geetha et al. (2020)	Evaluation of ETL tools for BI analysis	Limited interactive capabilities for handling edge-case anomalies	Data Vis. & BI
Attar-Khorasani & Chalmeta (2022)	IoT data visualization frameworks for BI	No application of visualization for ETL-stage validation	Data Vis. & BI
Pandey (2022)	Comparative study of diverse visualization tools	Does not address VR-based data engineering or ETL visualization	Data Vis. & BI
Kwon et al. (2016)	VR for financial data anomaly detection	Domain-specific application; not integrated into ETL preprocessing	VR Visualization
Bowman et al. (2012)	Theoretical benefits of 3D spatial data representation	General spatial data focus; not applied to data quality tasks	VR Visualization
Donalek et al. (2014)	Collaborative VR visualization for astronomy	High-complexity scientific visualization, not general data validation	VR Visualization
Bach et al. (2018)	VR visualization survey and research agenda	Agenda focuses on exploratory analysis, not operational validation	Immersive Analytics
Batch et al. (2020)	Expert performance metrics in immersive analytics	Focuses on analytical insight tasks rather than data cleaning	Immersive Analytics
Lee et al. (2021)	Data visceralization techniques for deep understanding	Operates under a "clean data" assumption for sensory experience	Immersive Analytics
Cordeil et al. (2017)	IATK: Immersive Analytics Toolkit development	Optimized for visualization and analysis, not data preprocessing	Immersive Analytics
Wagner Filho et al. (2020)	Evaluation of immersive scatterplot visualizations	Exploratory analysis focus; lacks record level manipulation tools	Immersive Analytics

Table 2. Comparison of Research Domains: VR Visualization vs. Traditional ETL vs. the VR-ETL Approach

Approach	Focus	Interaction Paradigm	Error Det. Timing	Visual Feedback	Primary Users	Empirical Evidence
VR Visualization Research	Analysis of clean data	Exploration and filtering	Not addressed	Analytical insights	Data analysts, researchers	Extensive user studies available
Traditional ETL Tools	Data transformation	Script based configuration	Post transform	Log files / text dashboards	Data engineers, ETL developers	Performance benchmark available
VR-ETL (This work)	Data validation and correction	Direct manipulation in VR	Pre transformation	3D spatial visualization	Data engineers, analysts	Prototype demonstration (future studies)

The data presented in Table 2 demonstrates that VR-ETL stands alone because it uses VR visualization methods to perform ETL domain data quality operations. The research field has not studied this specific point of connection between these two areas. The prototype proves that this combination works as a technical solution which will enable researchers to study its practical benefits for ETL workflow optimization in upcoming studies.

3. VR-ETL System Architecture

3.1. Architectural Overview

The currently developed version of the VR-ETL application represents a proof-of-concept prototype system. In order to get a high-fidelity baseline for software architecture, first measurements were performed within the Unity Editor on an Apple Mac Studio machine. It should be noted that measurement within the desktop-based platform cannot completely consider the heat and performance limitations of standalone VR device (such as Meta Quest). Nevertheless, the use of the Unity Editor enabled precise measurements of the rendering pipeline and served as an inevitable intermediate stage. The VR-ETL framework is designed as a modular pipeline that transposes traditional data engineering into a 3D environment. Unlike reactive post-transformation workflows, this architecture prioritizes pre-transformation validation, allowing for human-in-the-loop inspection before data enters the transformation stage. The system follows a decoupled design consisting of five core modules: Ingestion, Validation, Rendering, Visualization, and Correction. Figure 1 illustrates the high-level architecture and data flow of the VR-ETL system.

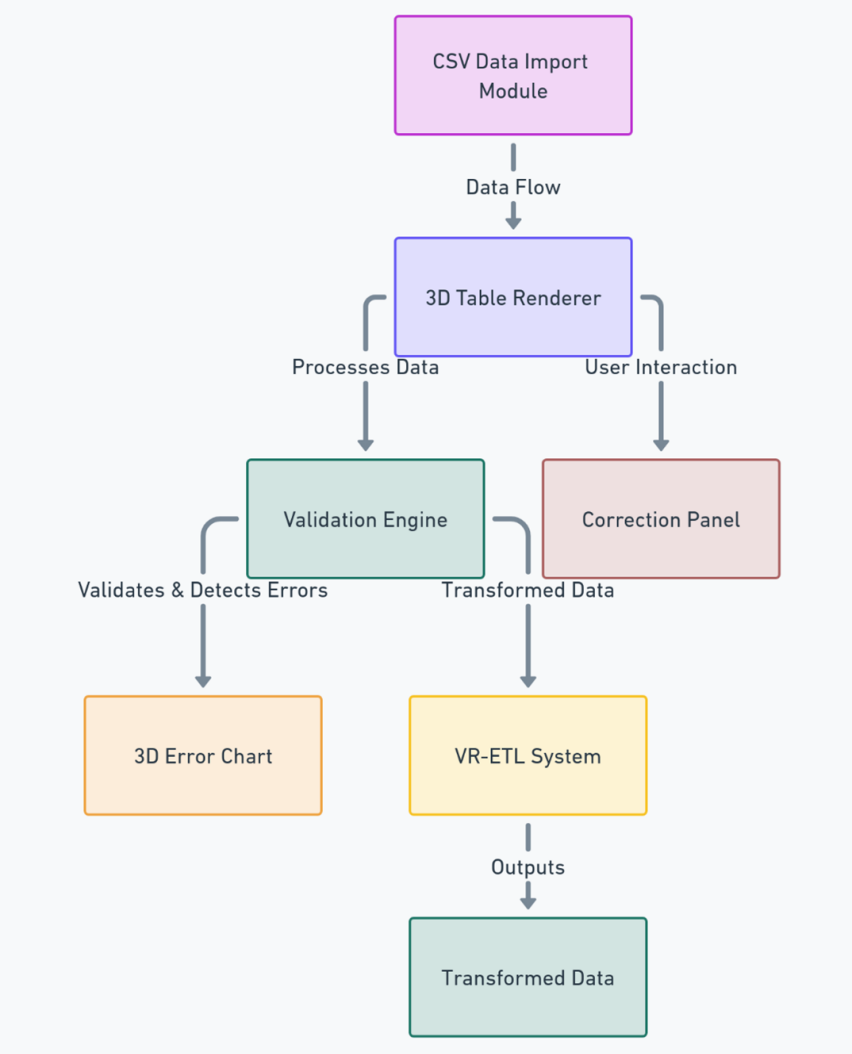


Fig. 1. Diagram of the Modular VR-ETL Pipeline showing data flow from ingestion to export

3.2. Core System Components

The system includes particular modules which transform table-based data into spatial data structures through their operations.

The CSV Ingestion Module starts by parsing the data while it detects encoding types such as UTF-8 and ASCII and determines the data types. An organized memory-based structure is created which functions as the fundamental data source for the entire pipeline process.

Validation Engine has an important role, it operates as a rule-based deterministic system that performs parallelized checks for missing data and incorrect GUIDs and date format errors and database reference errors. Error Matrix is then generated, which serves as the basis for visual feedback display.

The 3D Table Renderer module operates through Unity engine and OpenXR to convert CSV rows into particular 3D coordinates. Object Pooling is used together with View Frustum Culling to achieve the 90 FPS performance which VR requires for comfortable use.

The data is presented through two visualization methods which include the 3D Error Chart that uses spatial bar charts to analyze frequency data and the Correction Panel that shows details about single repair operations. Users can modify records through the panel which activates an automatic system to validate all changes.

3.3. Benefits and Limitations.

The VR-ETL system presents an innovative data engineering approach but exists at present as a proof-of-concept which requires definition of its benefits together with its built-in limitations.

The main advantage of VR-ETL comes from its ability to transform defensive error responses into useful predictive signals. The system enables pre-transformation validation which minimizes the expenses that occur when ETL rollbacks need to happen during late stages [28]. Users can detect error patterns through their spatial memory and peripheral vision in the 3D immersive environment which outperforms reading text-based logs. The system enables human involvement through its "human-in-the-loop" functionality which enables users to detect complex dataset structural mistakes that automated scripts fail to identify correctly.

The research contains multiple restrictions which need direct acknowledgment. The current prototype achieves its best performance when working with datasets which have records up to 10,000 in number. The frame rate in Unity becomes affected by rendering overhead when the system exceeds this threshold which could result in VR user motion sickness and reduced performance efficiency. The system operates with CSV file ingestion but it does not support processing of semi-structured data types including JSON and XML or unstructured information sources. The validation engine operates through deterministic rule-based logic which does not include machine learning or probabilistic models for anomaly detection at present. The need for specific VR equipment known as Head-Mounted Displays together with user exhaustion that occurs when using the technology for long periods create obstacles which prevent most institutions from adopting this technology.

3.4. Workflow Execution

The VR-ETL workflow follows an iterative cycle: Import $\rightarrow$ Automated Validation $\rightarrow$ Spatial Inspection $\rightarrow$ Interactive Correction $\rightarrow$ Export. By integrating the validation engine directly with the renderer, the system provides real-time visual feedback: cells transition from red (invalid) to green (corrected) instantly upon user input, eliminating the need for batch re-runs typical in tools like Talend.

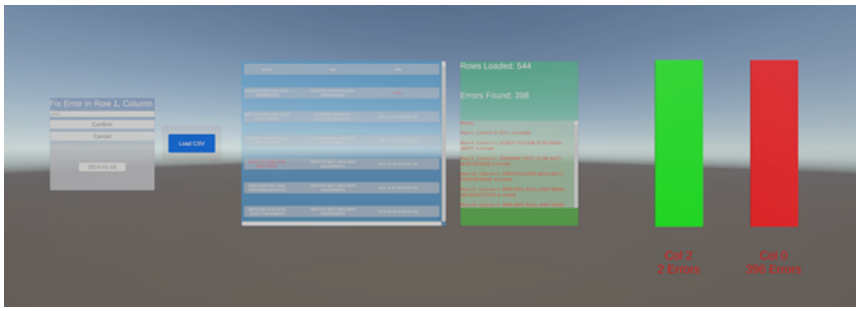


Fig. 2. Interactive User Interface in the VR environment displaying real-time error detection and the correction panel

Figure 2 illustrates the workflow of the VR-ETL system from data import through validated output.

3.5. Architectural Comparison: VR-ETL vs. Traditional ETL

The design elements of VR-ETL appear in Table 3 which demonstrates how they differ from standard ETL systems through their architectural structure and implementation choices.

Table 3. Architectural and Functional Differences between Batch-based Traditional ETL and Immersive VR-ETL[4]

Feature	Traditional ETL	VR-ETL (This work)
Error Detection	Post-transformation, batch processing	Pre-transformation, real-time validation
Error Correction Method	Manual log review, SQL queries, script-based updates	Interactive VR interface with direct manipulation and immediate feedback
Error Visualization	Text logs, 2D dashboards, tabular reports	Interactive 3D error charts, spatial heatmaps, color coded cells
Data Processing Mode	Batch processing with scheduled intervals	Real-time validation during loading
Scalability	Handles millions of rows with distributed processing	Prototype tested up to 10,000 rows; larger datasets require optimization
Hardware Requirements	Standard server hardware	VR headset (Meta Quest, HTC Vive, etc.), GPU-capable PC
User Interface	Command line, GUI config panels, script editors	Immersive 3D VR environment with spatial interaction
Learning Curve	Requires SQL/scripting knowledge, ETL tool expertise	Requires VR familiarity, comparative ease untested; requires user study
Collaboration	Single user with version control, email based coordination	Prototype is single user, multi user VR planned as future work
Integration	Extensive connectors to databases, APIs, cloud services	Prototype supports CSV import/export only, integration planned
Validation Rules	Predefined and custom rules, often requires coding	Five core validation rules implemented, extensible framework
Cost	Varies (free open source to enterprise licences)	Prototype development plus VR hardware (approximately 300–500 EUR)
Maturity	Production-ready tools with extensive testing	Proof of concept prototype, not production ready

This comparison highlights important architectural differences between VR-ETL and traditional ETL approaches. VR-ETL's primary distinction is its support for pre-transformation validation with immersive interaction. However, traditional ETL tools offer advantages in scalability, ecosystem integration, and production maturity. The prototype establishes that VR-based ETL validation is technically usable for medium-sized datasets (up to 10,000 rows). Future work will determine through empirical studies whether the interactive VR approach provides measurable benefits that justify the additional hardware requirements and learning investment.

[4] This comparison is based on the conceptual architectural design of VR-ETL versus established industry standards; empirical benchmarking between these platforms is reserved for future research.

4. Implementation & Technical Details

4.1. Overview of the VR-ETL System

The VR-ETL prototype was developed using Unity 6 and the OpenXR framework to ensure cross-platform compatibility across Meta Quest and HTC Vive ecosystems. The system architecture leverages C# for logic and Unity's World Space UI Canvas system for high-performance data rendering.

4.2. Data Ingestion

The ingestion module performs structured CSV loading through a multi-stage pipeline with three phases. First phase is an automated detection, which is used to identify UTF-8 and UTF-16 and ASCII through Byte Order Mark (BOM) analysis. Then the usage of "fail-soft" method which logs invalid rows to an audit trail while it continues to process the remaining data for uninterrupted workflow operation. Lastly, system requires 3–7 seconds to process a 10,000-row dataset on Apple M2 Max test hardware until it finishes rendering the data. To actually perform CSV parsing and type inference operations 400–600 ms is needed.

4.3. Comparison with Industry Standards: Talend and EasyMorph

The positioning of VR-ETL requires a comparison between its architectural design and the established data engineering tools Talend and EasyMorph.

The system operates through Talend which functions as its main enterprise solution for executing large-scale batch operations and providing extensive system connection abilities [15]. The main approach for error handling in Talend operates through reactive log-based systems. Engineers need to stop their main work when transformation failures occur because of poor data quality so they can review static reports and perform manual changes to either the source data or the transformation rules.

The visual interface of EasyMorph enables users to perform ETL operations through a non-technical interface which makes data transformation accessible to users who do not know how to program [29]. The system enables users to perform interactive data preparation but its 2D tabular interface on standard monitors prevents them from detecting complex multi-dimensional data relationships.

Table 4. Feature-based Comparison of the VR-ETL Prototype with Industry-Standard Tools (Talend and EasyMorph)

Feature	Talend	EasyMorph	VR-ETL (Prototype)
Primary Interface	2D GUI / Scripting	2D Visual No-code	3D Immersive VR
Validation Timing	Post-transformation	Interactive / Real-time	Pre-transformation
Error Feedback	Text-based Logs	2D Tables	3D Spatial Visuals
Correction Method	Script/Batch Update	UI-driven	Direct Spatial Manipulation

The VR-ETL system implements a pre-transformation validation mechanism which moves error detection to the initial stage of the data processing sequence. Talend provides better performance for handling large datasets but VR-ETL stands out because it enables users to perform anomaly repair through direct spatial context access. Organizations can achieve better data readiness through VR-ETL technology implementation because it creates a direct path from data reception to complete data preparation operations.

4.4. Rendering Optimizations

To maintain the 90 FPS target for VR stability, the 3D Table Renderer avoids the overhead of individual 3D meshes. Instead, it uses Unity UI Batching, which reduces draw calls by grouping text elements.

- Object Pooling: Row templates are instantiated once and recycled during scrolling to minimize Garbage Collection (GC) spikes.
- Viewport Culling: Only cells within the user's field of view remain active, significantly reducing the GPU vertex count. In editor-based testing, this architecture maintained frame rates exceeding 500 FPS for datasets with 1,000+ visible cells.

4.5. Algorithmic Validation Logic

The Validation Engine executes a deterministic four-stage process: Type Inference, Rule Application, Flagging, and Audit Logging. The five core rule categories are implemented as follows:

- Pattern Matching: Regex-based validation for GUID and Date formats (e.g., ISO 8601).
- Uniqueness & Integrity: HashSet-based lookups provide $O(1)$ average-case complexity for duplicate primary key detection and foreign key referential integrity checks. Total validation time for 10,000 records is approximately 800 ms, ensuring that the interface remains responsive during the initial "on-ingest" validation phase.

4.6. Human-in-the-Loop Correction Workflow

The Correction Panel allows automated detection systems to operate with human judgment through its Autosuggestion System.

- Context-Aware Heuristics: For format errors (e.g., Date/GUID), the system suggests re-formatted versions based on dominant column patterns.
- Session-Based Learning: The system caches accepted user corrections to suggest consistent fixes for recurring anomalies.
- Statistical Inference: The system presents three methods to handle outliers and missing data which involve using the mean or median or choosing the most frequent value from each column's statistical distribution. The system performs an immediate re-validation process after every correction which updates the spatial visual encoding (color change from red to green) to show instant feedback to the user.

5. Prototype Demonstration and Technical Validation

This section presents functional testing results that demonstrate the prototype's technical capabilities. These results establish feasibility but do not constitute complete evaluation, which remains as future work.

5.1. Preliminary Expert Walkthrough

The technical performance metrics received support from a first functional system evaluation which three senior data engineers performed to check the system operational sequence. The experts tested how well the VR-ETL prototype displayed tabular information through 3D visualization and they conducted tests on the process which required human involvement for data correction.

The expert walkthrough took place using a qualitative framework. Three data engineers (each having a minimum of 8 years of experience in the field) carried out a series of tasks: detection of null fields, correction of incorrect dates, and duplication of keys. The process was followed by a semi-structured interview. Though the current study is limited in terms of scale, its purpose is not to carry out any statistical analysis but to establish the grounds for future research based on the results of this qualitative study. For this reason, the NASA-TLX evaluation of workload is scheduled for a later stage of the experiment.

The participants agreed that the spatial error mapping system which used color-coded 3D tables helped them detect pattern-based anomalies better than traditional text-based error reports. This aligns with the results of Millais et al. [25], who said that while VR might not always reduce objective workload, it leads to a greater depth of insight and higher user satisfaction compared to 2D environments. The system includes real-time revalidation which enables cells to switch from red to green immediately to solve the delay problems that occur in conventional reactive ETL systems. By using these "embedded data representations" [40], the prototype leverages human spatial memory more effectively than traditional 2D wrangling interfaces described by Kandel et al. [17]. The expert feedback validated that the modular system lets users execute core data engineering tasks inside an interactive environment which will function as a base for upcoming statistical investigations.

5.2. Methodology and Test Configuration

Testing was conducted on an Apple Mac Studio (M2 Max, 32 GB RAM) using Unity 6. While functional testing included a Meta Quest 2, all quantitative performance metrics were collected within the Unity Editor to use high-fidelity profiling tools.

Three synthetic datasets were seeded with a 10% error rate across five categories (Missing Values, Malformed GUIDs, Date Errors, Duplicate Keys, and Foreign Key Violations) to establish a ground truth for validation accuracy (see Table 5).

Table 5. Ingestion and Processing Latency Metrics for datasets ranging from 50 to 10,000 records

Dataset	Rows	File Size	Total Processing Time	Scaling Factor
Small	50	2 KB	36 ms	1.0x
Medium	100	17 KB	70 ms	1.9x
Large	10,000	1.7 MB	3,451 ms	49.3x

5.3. Processing and Rendering Results

For the 10,000-row dataset, the $\sim$3.5 s pipeline includes 550 ms for ingestion/parsing, 800 ms for validation, and 2.1 s for 3D instantiation. The observed sub-linear scaling (a 100x increase in rows resulting in only a 49x increase in time) suggests the architecture is efficient for medium-scale datasets.

Table 6. Success Rates of the Rule-based Engine in detecting seeded data quality anomalies

Error Type	Seeded	Detected	Rate
Pattern-Based (GUID/Date/Null)	842	842	100%
Integrity-Based (Keys/Duplicates)	93	93	100%
Total	935	935	100%

As shown in Table 6, the deterministic engine achieved a 100% detection rate for seeded rule-based errors. The Editor's system achieved more than 500 FPS (Frame Time $<$2 ms) which provided sufficient resources to support the stereoscopic overhead that physical VR hardware needs. These performance metrics fulfill the high-frequency update requirements necessary for seamless immersive analytics [14]. The system maintained cell selection and feedback response times under 10 ms which fulfilled the requirements for interactive virtual reality systems and aligns with the post-WIMP interaction standards suggested by Lee et al. [21].

5.4. Comparison with Existing Literature and Limitations

While the results confirm the system's foundational logic, it is essential to position VR-ETL within the current research landscape. Unlike many immersive analytics systems that focus on the exploratory analysis of already cleaned, post-hoc analytical results [3,4], VR-ETL is specifically positioned in the operational pre-transformation phase of the ETL pipeline.

As for scalability, although 10,000 rows of data are not enough for large companies in terms of their Big Data scenarios, it was purposely selected in order to detect what part of the process is the computation cost incurred when instantiating objects in 3D space. It turned out that the 3D visualization process is the bottleneck of our system, not any calculations done with data.

In terms of hardware efficiency, our choice of a Head-Mounted Display (HMD) interface follows the findings of Cordeil et al. [9], who demonstrated that HMDs can offer faster completion times and greater accessibility for connectivity analysis compared to traditional, high-cost CAVE-style environments. However, several constraints define the current scope:

- Measurement Context: Performance was measured in the Unity Editor. Actual on-device performance (Meta Quest 2) will likely be lower due to stereoscopic rendering and wireless latency.
- Scalability Boundaries: Feasibility is confirmed up to 10,000 rows. While 2D wrangling tools like Wrangler [17] can handle much larger volumes through sampling, VR-ETL's behavior with larger data volumes or irregular real-world encodings remains unverified due to the overhead of 3D object instantiation.
- Functional Scope: The prototype is limited to CSV files and a single-user environment, whereas systems like *Uplift* [12] have begun exploring multi-user, collaborative visual analytics.
- Absence of User Metrics: This evaluation establishes that the system works, but not that it is better than traditional 2D tools. User fatigue, the learning curve for non-VR users, and comparative efficiency against tools like Talend remain "known unknowns."

5.5. Architectural Advantage over Traditional ETL Workflows

Although it was still in our plans to conduct empirical comparisons with other existing solutions like Talend and EasyMorph, it is possible to characterize the advantages of VR-ETL from the architecture perspective. For instance, in case of conventional systems with reactive approach, each mistake found during the transformation process requires a certain sequence of actions – checking of logs in text form, identifying the erroneous row in source data, correcting the source file itself or modification of the transformation rule and finally running the entire batch job anew. This loop may require several minutes per one run depending on the amount of data processed and usually involves context switching between at least three different applications (logs viewer, source data editor and ETL orchestration tool). VR-ETL on the contrary, enables users to fix mistakes instantly without leaving the current scene where a particular mistake was detected. The entire cycle in VR-ETL is reduced to a single step which includes instant correction of data inside the application and re-validation of changes in less than 10 ms with immediate feedback (color change from red to green provided inside VR environment). Taking into account the achieved detection rate (100% accuracy for rule-based anomalies in a dataset we used) and sublinear time scaling to 10,000 records, it is clear that our system is able to provide instant validation with no need to leave the current interface, which dramatically decreases the total cost of each correction action. In order to evaluate our solution in comparison with others such as Talend or EasyMorph we plan to perform direct benchmarks of time required for completion of identical tasks.

5.6. Validation of Research Objectives

The results confirm that the Modular Architecture (Contribution 1) and Validation Engine (Contribution 2) are functional. The 3D Interface (Contribution 3) and Error Visualiza-

tions (Contribution 4) successfully provide real-time feedback. Overall, Technical Feasibility (Contribution 5) is demonstrated within the Unity Editor environment, pending on-device validation for the 10k-row range, establishing a sound baseline for the next phase of the PhD research program. By bridging the gap between operational data engineering and immersive visualization, VR-ETL addresses a specific research void identified in the surveys of ETL [38] and Immersive Analytics [14] literature.

6. Conclusion and Future Work

Studies of virtual reality technology demonstrate its viability for ETL data validation, particularly during the critical period prior to transformation. Results from prototype development and testing indicate that VR technology can use deterministic validation methods together with 3D spatial representations of data for the detection of all pattern based anomalies within datasets while containing up to 10,000 records. The technical performance metrics support the conclusion that the makers of today's game engines have created products capable of carrying out the equivalent of real-time ETL data validation in excess of 500 FPS, which provides for an interaction latency that is less than 10 ms, thereby enabling seamless end-user experience.

The prototype proves that the concept of VR-ETL does work as intended. However, there are several issues associated with its operation that have to be overcome before implementing the concept commercially. First, it is crucial to carry out systematic validation of profiling on the Meta Quest 2 device to make sure that the results obtained through the profiling using an editor are consistent and to find out the overhead related to stereoscopic visualization and wireless communication. Additionally, the current version of the prototype allows for a single-user experience and processing of synthetic CSV files, while in actual practice, JSON and SQL formats and multiple participants should be supported.

Research collaboration on the VR ETL framework will establish a solid basis for upcoming empirical research studies. Initial effort will have a focus on short term performance improvements from the implementation of hand tracking technology and stress testing of the software with a dataset of over 50,000 records to assess critical failure points in the 3D data instantiation system. The second phase of the research will involve an empirical evaluation of the system through a controlled comparative user study with 60 or more users measuring task completion time and cognitive load using the NASA-TLX index. The long term architectural growth of this system will be achieved through three key components which will involve the use of machine learning based proactive anomaly detection and connectors to orchestrate production tools available for Apache Airflow and multi user environments to facilitate team based spatial data audits.

Research conducted utilizing the VR ETL system addresses an important gap in current research since it uses the immersive analytics to examine data processes that take place prior to the transformation of data. Also, VR Technology does not present any major technological roadblocks that would inhibit its use by organizations performing data quality management and operational data engineering. This operational standard for future experimental research studies will set guidelines for determining whether virtual reality has spatial reasoning functions that create measurable improvements in productivity and reduce mental workload for users.

In conclusion, VR-ETL demonstrates the technical feasibility of moving data validation to the pre-transformation stage within an immersive environment. As this work constitutes the foundational phase of a doctoral research program, several critical expansions are planned: hardware validation, which will include re-running all performance experiments on standalone HMD hardware to measure actual on-device latency and battery/thermal impact; user study, formal comparative study on a larger set of participants, utilizing NASA-TLX (Task Load Index) to quantitatively measure cognitive load and speed compared to traditional 2D ETL interfaces; big data integration, support for larger data sets exceeding 50,000+ records.

References

1. Aleksic, S., Ristic, S., Lukovic, I., Celikovic, M.: A design specification and a server implementation of the inverse referential integrity constraints. Computer Science and Information Systems 10(1), 283–320 (2013)
2. Attar-Khorasani, S., Chalmeta, R.: Internet of things data visualization for business intelligence. Big Data 10(2), 145–165 (2022)
3. Bach, B., Sicat, R., Beyer, J., Cordeil, M., Pfister, H.: The hologram in my hand: How effective is interactive exploration of 3D visualizations in immersive tangible augmented reality? IEEE Transactions on Visualization and Computer Graphics 24(1), 457–467 (2018)
4. Batch, A., Cunningham, A., Cordeil, M., Elmqvist, N., Dwyer, T., Thomas, B.H., Marriott, K.: There is no spoon: Evaluating performance, space use, and presence with expert domain users in immersive analytics. IEEE Transactions on Visualization and Computer Graphics 26(1), 536–546 (2020)
5. Batini, C., Rula, A., Scannapieco, M., Viscusi, G.: From data quality to big data quality. In: Big Data, pp. 1934–1956. IGI Global (2016)
6. Biswas, N., Sarkar, A., Mondal, K.C.: Efficient incremental loading in ETL processing for real-time data integration. International Journal of Database Theory and Application 12(6), 1–16 (2019)
7. Bowman, B., Elmqvist, N., Jankun-Kelly, T.J.: Toward visualization for games: Theory, design space, and patterns. IEEE Transactions on Visualization and Computer Graphics 18(4), 662–671 (2012)
8. Cordeil, M., Cunningham, A., Dwyer, T., Thomas, B.H., Marriott, K.: Imaxes. In: Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology. pp. 71–83 (2017)
9. Cordeil, M., Dwyer, T., Klein, K., Laha, B., Marriott, K., Thomas, B.H.: Immersive collaborative analysis of network connectivity: CAVE-style or head-mounted display? IEEE Transactions on Visualization and Computer Graphics 23(1), 441–450 (2017)
10. Donalek, C., Djorgovski, S.G., Cioc, A., Wang, A., Zhang, J., Lawler, E., Yeh, S., Mahabal, A., Graham, M., Drake, A., Davidoff, S., Norris, J.S., Longo, G.: Immersive and collaborative data visualization using virtual reality platforms. In: 2014 IEEE International Conference on Big Data (Big Data). pp. 609–614 (2014)
11. Dwyer, T., Marriott, K., Isenberg, T., Klein, K., Riche, N., Schreiber, F., Stuerzlinger, W., Thomas, B.H.: Immersive Analytics: An Introduction. KOPS Universität Konstanz (2018)
12. Ens, B., Goodwin, S., Prouzeau, A., Anderson, F., Wang, F.Y., Gratzl, S., Lucarelli, Z., Moyle, B., Smiley, J., Dwyer, T.: Uplift: A tangible and immersive tabletop system for casual collaborative visual analytics. IEEE Transactions on Visualization and Computer Graphics 27(2), 1193–1203 (2021)

13. Filho, J.A., Stuerzlinger, W., Nedel, L.: Evaluating an immersive space-time cube geovisualization for intuitive trajectory data exploration. IEEE Transactions on Visualization and Computer Graphics 26(1), 514–524 (2020)
14. Fonnet, A., Prie, Y.: Survey of immersive analytics. IEEE Transactions on Visualization and Computer Graphics 27(3), 2101–2122 (2021)
15. Geetha, S., Dhanani, K.R., Doshi, P.P.: Data analysis and ETL tools in business intelligence. International Journal of Multidisciplinary and Current Research 8(1), 58–64 (2020)
16. Gupta, K., Jiwani, N.: A systematic overview of fundamentals and methods of business intelligence. International Journal of Science and Digital Convergence 7(2), 145–151 (2021)
17. Kandel, S., Paepcke, A., Hellerstein, J., Heer, J.: Wrangler. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 3363–3372 (2011)
18. Krneta, D., Jovanovic, V., Marjanovic, Z.: A direct approach to physical data vault design. Computer Science and Information Systems 11(2), 569–599 (2014)
19. Kwon, O.H., Muelder, C., Lee, K., Ma, K.L.: A study of layout, rendering, and interaction methods for immersive graph visualization. IEEE Transactions on Visualization and Computer Graphics 22(1), 680–689 (2016)
20. Lee, B., Brown, D., Lee, B., Hurter, C., Drucker, S., Dwyer, T.: Data visceralization: Enabling deeper understanding of data using virtual reality. IEEE Transactions on Visualization and Computer Graphics 27(2), 1095–1105 (2021)
21. Lee, B., Srinivasan, A., Isenberg, P., Stasko, J.: Post-WIMP interaction for information visualization. Foundations and Trends in Human–Computer Interaction 14(1), 1–95 (2021)
22. Mausa, G., Galinac-Grbac, T., Dalbelo-Basic, B.: A systematic data collection procedure for software defect prediction. Computer Science and Information Systems 13(1), 173–197 (2016)
23. Mehmood, E., Anees, T.: Distributed real-time ETL architecture for unstructured big data. The Journal of Supercomputing 78(8), 10803–10832 (2022)
24. Mhon, G.G.W., Kham, N.S.M.: ETL preprocessing with multiple data sources for academic data analysis. In: 2020 16th International Conference on Computing, Communication and Networking Technologies (ICCN). pp. 1–5 (2020)
25. Millais, P., Jones, S.L., Kelly, R.: Exploring data in virtual reality. In: Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems. pp. 1–6 (2018)
26. Mondal, K.C., Biswas, N., Saha, S.: Role of machine learning in ETL automation. In: Proceedings of the 3rd International Conference on Vision, Image and Signal Processing. pp. 1–5 (2020)
27. Nemani, R.R., Konda, R.: A framework for data quality in data warehousing. In: Lecture Notes in Business Information Processing, pp. 292–297. Springer (2009)
28. Nwokeji, J.C., Matovu, R.: A systematic literature review on big data extraction, transformation and loading (ETL). In: Advances in Intelligent Systems and Computing, vol. 1249, pp. 259–271. Springer (2021)
29. Pandey, A.: Comparative study of data visualization tools in big data analysis for business intelligence. International Journal for Research in Applied Science and Engineering Technology 10(6), 2591–2600 (2022)
30. Patel, M., Patel, D.B.: Progressive growth of ETL tools: A literature review of past to equip future. In: Advances in Intelligent Systems and Computing, vol. 1186, pp. 501–510. Springer (2020)
31. Qaiser, A., Farooq, M.U., Mustafa, S.M.N., Abrar, N.: Comparative analysis of ETL tools in big data analytics. Pakistan Journal of Engineering and Technology 1(1), 1–8 (2023)
32. Ragan, E.D., Sowndararajan, A., Kopper, R., Bowman, D.A.: The effects of higher levels of immersion on procedure memorization performance and implications for educational virtual environments. IEEE Transactions on Visualization and Computer Graphics 16(5), 891–901 (2010)

33. Saebao, S., Matayong, S., Trakulmaykee, N.: QoX based ETL design for business intelligence system of lecturers' qualifications analysis. In: 2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON). pp. 465–468 (2020)
34. Sidi, F., Shariat Panahy, P.H., Affendey, L.S., Jabar, M.A., Ibrahim, H., Mustapha, A.: Data quality: A survey of data quality dimensions. In: 2012 International Conference on Information Retrieval & Knowledge Management. pp. 300–304 (2012)
35. Simitsis, A., Skiadopoulos, S., Vassiliadis, P.: The history, present, and future of ETL technology. CEUR Workshop Proceedings 3351, 1–12 (2023)
36. Souibgui, M., Atigui, F., Ben Yahia, S., Si-Said Cherfi, S.: Business intelligence and analytics: On-demand ETL over document stores. In: Communications in Computer and Information Science, vol. 1221, pp. 659–673. Springer (2020)
37. Sousa, R., Miranda, R., Moreira, A., Alves, C., Lori, N., Machado, J.: Software tools for conducting real-time information processing and visualization in industry: An up-to-date review. Sustainability 11(11), 4800 (2021)
38. Vassiliadis, P.: A survey of extract–transform–load technology. International Journal of Data Warehousing and Mining 5(3), 1–27 (2009)
39. Wagner Filho, J.A., Rey, M.F., Freitas, C.M., Nedel, L.: Immersive visualization of abstract information: An evaluation on dimensionally-reduced data scatterplots. In: 2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). pp. 483–490 (2020)
40. Willett, W., Jansen, Y., Dragicevic, P.: Embedded data representations. IEEE Transactions on Visualization and Computer Graphics 23(1), 461–470 (2017)

Maja Cerjan is a teaching assistant at the Faculty of Organization and Informatics (FOI), University of Zagreb, Croatia, within the Department of Theoretical and Applied Foundations of Information Sciences, where she teaches courses related to databases, data warehouses, and business intelligence. She is a final-year doctoral candidate in the Information Society programme at the Faculty of Information Studies (FIŠ) in Novo Mesto, Slovenia. Her doctoral research focuses on the application of virtual reality to the design and validation of ETL (Extract, Transform, Load) processes. Her research interests include ETL processes, data warehousing, dimensional modelling, business intelligence, deductive databases, and the use of immersive technologies in data engineering.

Kornelije Rabuzin is a Full Professor at the Faculty of Organization and Informatics (FOI), University of Zagreb, Croatia, where he heads the Laboratory for Data Technologies within the Department of Theoretical and Applied Foundations of Information Sciences. He teaches several courses related to databases and data warehouses. He is the author and co-author of nearly 100 scientific papers and around 10 books and book chapters in the fields of databases, data warehouses, and business intelligence. His research interests include databases, data warehouses, business intelligence, big data, and, in recent years, graph databases and graph data warehouses, including integrity constraints and query languages for graph databases.

Leo Mršić received his PhD in information sciences from the Faculty of Humanities and Social Sciences, University of Zagreb, Croatia, in 2012, with research focused on graph theory and the application of statistical models in business. He is a Full Professor active in the scientific fields of computing, information sciences, and economics, and

currently serves as Vice-Rector for Science and Research at Algebra Bernays University in Zagreb. He is the author and co-author of more than 100 scientific papers and holds several patents. He is an IEEE Senior Member, an ACM member, and an IPMA Level A certified project director. His research interests include data science, business intelligence, data management and analysis, and artificial intelligence.

Received: February 19, 2026; Accepted: July 15, 2026.

Computer Science and Information Systems 23(4):1343–1394 https://doi.org/10.2298/CSIS260815044B

Paper Type: Research Paper

Methodology for Rapid Development of PAIS Based on Ontological Modeling and Category Theory

Borivoj Bogdanović[1], Vidan Marković[1], Milan Segedinac[2], and Zora Konjović[1]

[1] Singidunum University, Danijelova 32
11010 Belgrade, Serbia
borivoj.bogdanovic@gmail.com (corresponding author),
{vidan.markovic, zkonjovic}@singidunum.ac.rs
[2] University of Novi Sad, Faculty of Technical Sciences
Trg Dositeja Obradovica 4, Novi Sad 21000, Serbia
milansegedinac@uns.ac.rs

Abstract. Model-Driven Architecture (MDA) structures software development through computation-independent, platform-independent, and platform-specific models. Its practical value depends on machine-readable models and transformations that preserve relevant structural properties while producing executable code. This paper proposes a methodology for the rapid development of Process-Aware Information Systems (PAIS) based on ontological modeling and category theory. Ontology-Driven Code Generation (ODCG) is formalized as a deterministic, multi-phase transformation from the BAB Core ontology, through a platform-specific ontology and library, to source code. Model levels are represented as categories, while transformations are represented as functors. The main contribution is a categorical formalization of stack-parametrized multi-target generation, together with a conditional guarantee of schema-level structural preservation under explicitly stated verification assumptions, including mapping completeness, injectivity, quotient compatibility, and absence of unintended collapse. The methodology distinguishes the categorical specification, in which functoriality follows by construction, from its operational implementation, whose conformance is established through explicit engineering checks. The approach is implemented and evaluated for the Java platform through the BABj ontology and library by generating code for an operational business information system during its life cycle.

Keywords: Model-Driven Engineering, Ontology, Category Theory, Code Generation, PAIS

1. Introduction

Abstraction is crucial in software engineering for managing complexity, and there is a constant tendency to introduce ever-higher levels of abstraction into the software development process, ranging from implementation to modeling. Model-Driven Architecture (MDA) [50] is based on the principle of using modeling languages and specifying systems at different levels of abstraction: a Computation Independent Model (CIM) for representing the environment and system requirements, a Platform Independent Model

(PIM) that describes the system in a technology-neutral manner, and a Platform Specific Model (PSM) that adds details related to a specific platform [19]. The vision of MDA is that, through automated transformations, a PIM can be translated into a corresponding PSM and further into executable code on the target platform, thereby ensuring that once-defined system logic can be semi-automatically implemented on different technology stacks, with a reduction in manual programming. Such an approach poses two basic tasks: (1) a machine-readable representation of the domain/information system model at different levels of abstraction, and (2) the transformation of models so represented to the level of executable code, with the fundamental requirement of preserving the selected schema-level structure and explicitly represented mapping distinctions during the transformation process.

Ontologies are inherently suitable for representing PIMs because they allow for the definition of the system domain at an abstract level, independent of implementation. If represented via standards such as RDF and OWL, they are machine-readable and have formally defined semantics, making them suitable for sharing and reusing domain knowledge. This paper explores an approach in which such an abstract domain ontology (core ontology) is transformed into a technology-specific ontology (analogous to a PSM), from which program code is subsequently generated. The transformation from a general model (domain ontology) to concrete code is viewed here as a functorial process, while the alignment between the abstract schema-level structure and its platform-specific realization is modeled through natural transformations.

The proposed framework represents the PIM through an ontology named BAB Core (Business Application Builder Core), which contains domain knowledge about PAIS, including entities, processes, tasks, roles, and user interfaces. A platform-specific ontology and accompanying library provide the implementation details required for a particular technological environment. In the present work, this platform-specific realization is BABj, which maps abstract BAB Core concepts to Java/Jakarta EE and Vaadin artifacts, such as JPA entities, Vaadin views, and EJB services. BABj is the only implemented and evaluated target stack considered in this paper [9,11,10].

The primary scientific contribution of the approach proposed in this paper is a foundational mechanism for transformation from model to model up to the level of generated code, while preserving selected schema-level structure and explicitly represented mapping distinctions across the transformation chain under explicitly stated verification assumptions. The practical result of the research is a repeatable and systematic code-generation process by which schema-level application structure and selected architectural realization rules, formally defined through ontologies, are translated into the Java-based BABj architecture in a controlled manner. Domain-specific method bodies remain explicit extension points and are not claimed to be fully synthesized by the present transformation. Since the transformer uses established Java/Jakarta EE patterns and the BABj library, adherence to consistent architectural practices is ensured across BABj-based projects.

The remainder of the paper is organized as follows. Section 2 provides an overview of previous research on approaches to code generation from ontologies through multi-phase and formally based transformations. For the selection of papers and analysis, a framework defined by five themes was adopted, and for each of them, implications for the research in this paper are explicitly highlighted. Section 3 introduces the ontological foundations used in the paper, while Section 4 presents the category-theoretic concepts underlying the

proposed methodology. Section 5 presents the proposed methodology and the categorical transformation framework in detail. Section 6 describes the Java-based implementation of the transformer. Section 7 presents the analysis and discussion of results, and the final section provides concluding remarks and directions for future research.

2. Related Work

The subject of the research presented in this paper is a methodology for the accelerated development of Process-Aware Information Systems (PAIS), a concept explicated more than 20 years ago in [24]: "A Process-Aware Information System (PAIS) is a software system that manages and executes operational processes involving people, applications, and/or information sources on the basis of process models." The same source classifies PAIS according to four orthogonal dimensions: (1) orientation (design vs. implementation), (2) participants involved (people and software applications), (3) process structure and predictability (unframed, ad hoc framed, loosely framed, and tightly framed), and (4) organizational scope (intra-organizational, inter-organizational), while source [25] classifies PAIS as domain-specific and domain-agnostic. Among more recent results, the monograph [33] stands out, confirming the viability of the PAIS concept, i.e., its potential for evolution leading towards Intelligent PAIS (IPAIS) with new concepts related to artificial intelligence, but also new development approaches, among which model-driven and collaborative approaches are prominent.

The PAISs of interest for our research are those that are domain-specific (business information systems), framed (tightly or ad hoc), and implementation-oriented.

The convergence of ontological modeling and software engineering has a long history and continuous development. The beginnings date back to the 1980s with the emergence of the concept of using formal knowledge structures (ontologies) for reasoning in artificial intelligence systems, expanding philosophical ideas into computational models for knowledge representation [37,13,69]. The last decade of the previous century was marked by formal methods and the Semantic Web, aiming to ensure interoperability, give models deeper meaning, and enable automated reasoning using Semantic Web standards (RDF, OWL) [7,6,44,21].

The beginning of this century (up to 2015) brought key results in the field: the idea and assessment of the possibilities of applying ontologies in software engineering [58,15,54,30,57], combining UML and ontologies [27,53,14], combining ontologies and Model Driven Engineering (MDE) approaches [65,40,4,12,66,64] and support for data access and queries [70,16,8]. In addition, during this period, the paper of Diskin and Maibaum [22] appeared, systematically introducing category theory into model-driven engineering.

The last 10 years (from 2015 to present) have been marked by research that continues to address the combination of UML and ontologies [51], model transformation in the MDE approach [34,62,43,38,1,63,39], automatic model-to-text code generation [32,67,9,11,10], the application of ontologies in data management [66,70,41,56,16,18,36], AI integration [3,48,61,46,17], infrastructure for ontology development and integration [23,28,29,26,68,20], automation of the software lifecycle [55], and ever-present new domain applications [2,45,49,60] that seek solutions and

thereby contribute to the development of the field. Finally, there are papers dealing with the application of the category theory apparatus in software engineering [71,52,42].

In the section that follows, papers directly relevant to the subject of research are singled out. The results are grouped into two categories: results dealing with the approach of code generation from ontologies through multi-phase and formally based transformations, and results dealing with the application of category theory in the field of model-driven software development. The analyzed papers are organized into the following thematic units: (1) Ontologies and Software Engineering, (2) MDE/MDA and Transformations, (3) OBDA and Semantic Web, (4) LLM/AI Support, and (5) Formalization of MDE Patterns and Synergy with Ontologies, with explicit highlighting of implications for this work.

2.1. Ontologies and Software Engineering

From the first group of papers, of interest to this research is paper [51] which provides an exhaustive overview of combining UML and ontologies, considering, among other things, ways to make bidirectional model synchronization feasible; paper [34] which deals with the relationship between metamodels and ontologies, directly supporting the use of a core ontology as a PIM layer above technology; papers [32,67,9,11,10] which deal with automatic model-to-text source code generation alongside a "best practices" library; papers on the adoption of ontological modeling in software engineering practice that identify a pronounced need for tools that bring semantic models closer to programmers (editors, linters, generators) [23,26]; and papers dealing with collaborative work and "design review" that emphasize the importance of visualization [23] and the automation of core model generation by learning from existing artifacts (code, APIs, data schemas) [3]. Mkhinini et al. in their review [51] analyze the convergence of UML and ontologies. The authors identify that, although UML offers the visual standardization necessary for engineers, it lacks the formal semantics for automated reasoning offered by ontologies. The conclusion of their research points to the necessity of hybrid approaches where ontologies take on the role of a "metamodel with semantics," enabling verification that classic UML tools cannot provide.

The methodology proposed in this paper positions the ontology as the central Platform-Independent Model (PIM) and the Single Source of Truth (SSoT) for the domain. To ensure the consistency and completeness of this model before code generation, a supporting tool layer is introduced. Its key functions (visualization, validation, and support for semi-automated concept learning) are applied as mandatory steps before triggering transformations towards the PSM layer and final code.

2.2. MDE/MDA and Transformations (Tools, Patterns, Verification, Collaboration)

Contemporary reviews of MDE note a trend of merging with ML/analytics and collaborative workflows [20,48], which is relevant for the automatic suggestion of mapping patterns in the BABj ontology and team-based model evolution. Source [38] offers a taxonomy of transformation tools and techniques (model-to-model vs. model-to-text), while [43] summarizes transformation design patterns that can be formalized (e.g., Entity $\rightarrow$ JPA, Process $\rightarrow$ Service, View $\rightarrow$ UI-class). Here, the key issue is correctness, and results

vary from techniques for transformation verification (invariants, provable rules) [14], through the systematization of model composition essential for merging modules and sub-ontologies prior to generation [1], to specialized solutions for specific domains (e.g., digital twins) [45,49].

The methodology in this paper defines the transformation approach as a formalized, strictly controlled model-to-text (M2T) process. This process relies on two key elements: (1) explicit transformation rules (core → BABj; BABj → code) grounded in recognized patterns from the literature, and (2) a multi-layered verification framework encompassing invariant checking, ontology validation, and evaluation of generated code via static analysis.

2.3. OBDA and Semantic Web (Data Access, Interoperability, Practical tooling)

Another area with significant reliance on semantic technologies is Ontology-Based Data Access (OBDA), which emphasizes the integration of ontological models with relational and heterogeneous data stores [18], and where results include ways of merging semantic modeling with data lake approaches, opening space for the pipeline to generate data-access artifacts (e.g., R2RML/OBDA configurations) alongside application code [36], or to support migrations between stacks/organizations [68] .

Even early research like RDFReactor [70] showed that ontology → OO API transformation is feasible and useful for developers. The EMF/Ecore line [66] established a mature practice of model → Java generation, while KOMMA/Empire/Elmo (AliBaba) [41,56] illustrate different approaches to object/triple mapping towards RDF stores.

In accordance with the cited results, the methodology proposed in this paper adopts proven concepts from the literature, such as typed API generation and the application of patterns, but differs by extending these ideas to the entire information system (entities, services/processes, user interface, and data access), thereby creating a basis to ensure, through the application of formal mapping and verification, schema-level structural and architectural consistency of the generated system with the central core ontology. Based on the cited results from the field of OBDA (Ontology-Based Data Access), the proposed methodology extends the target scope of generation beyond exclusively application code. The transformation pipeline is, therefore, designed to produce a complete data access layer as well. In this way, it is ensured that generated data-access artifacts are structurally and schema-level consistent with the central core ontology.

To position the approach against practical tooling, we compare it along two axes: the target scope (a single layer versus the whole information system) and the presence of a formal, verifiable preservation guarantee. EMF/Ecore [66] established mature model-to-Java generation but targets primarily the structural/persistence layer and offers no categorical preservation guarantee. RDFReactor [70] generates a typed object-oriented API from an ontology, again single-layer and without formal preservation claims. KOMMA and AliBaba (Elmo/Empire) [41,56] provide object/triple mapping toward RDF stores, focused on data access. Ontop [16] realizes OBDA by rewriting queries over databases without generating application code at all. Against these, the proposed method targets the whole system (entities, services/processes, UI, and data access) from a single ontological source and adds a formal, checkable structural-preservation guarantee across a multi-target pipeline. The distinction is thus not the idea of generating code from a model,

which these tools already demonstrate, but the combination of end-to-end scope with a categorical correctness argument.

2.4. LLM/AI for Ontologies and Generation (Assistance, Qualitative Checks)

Recent research indicates that LLMs accelerate OE (Ontology Engineering): concept proposal, consistency checking, assistance with mapping and documentation [46,61], and thus its application in various domains. In parallel, a review of ontology embeddings [17] opens the possibility for patterns and code artifacts to be selected based on semantic similarity and experience from previous projects.

As a response to contemporary papers on the application of LLM technologies in ODE, the methodology conceptually envisions the possibility of introducing AI-assisted mapping through a human-in-the-loop (HITL) process. In such a mode of operation, the LLM would propose BABj mapping patterns (with explanations), while the engineer would retain explicit control through a mandatory validation step before applying the rules. This would accelerate the cognitively demanding mapping process while preserving formal control and the deterministic behavior of the transformation layer.

2.5. Domain Applications and Industry

Finally, systematic literature reviews (SLR) and reviews by domain, e.g., manufacturing [60], show that ontologies provide measurable value in complex industrial scenarios. This suggests that the core $\rightarrow$ stack $\rightarrow$ code approach has the potential for cross-domain use, where domain specificities are captured through the core model and exploited through different target stacks.

To examine the potential for cross-domain application identified in the literature review, the methodology is evaluated through a case study from a complex industrial domain. The purpose is to illustrate the operation of the core $\rightarrow$ stack $\rightarrow$ code approach in an industrially relevant scenario and to assess the architectural consistency and engineering quality of the generated artifacts.

2.6. Formalization of MDE Patterns and Synergy with Ontologies

Progress in the field of Model-Driven Engineering increasingly relies on rigorous mathematical formalization to overcome the problems of ad-hoc transformations. Diskin and Maibaum, in their paper [22] which is considered fundamental for MDA application, argue that category theory in MDE serves not merely as an "abstract support", but as a source of concrete design patterns for model management. They show that key engineering tasks, such as model merging and transformation, can be directly modeled as categorical constructions. This is of essential importance for ensuring consistency and traceability in complex transformation chains, as it suggests that model integrity is maintained not by procedural scripts, but by structural mappings that preserve information.

Complementary to this, Kovalyov [42] proposes a comprehensive category-theoretic approach to designing software systems, emphasizing the ability of categories to unify heterogeneous engineering technologies. His work sets the foundations for treating different aspects of a software system (data, processes, UI) through a unified formal framework, which directly supports the idea of a technology-neutral core (like BAB Core) that is projected via functors onto specific technological stacks.

The categorical approach gives precise semantics to the proposed transformations. Mordecai and associates [52] offer a vocabulary for MBSE cycles, which can support the formalization of functors realizing Core-to-Stack and Stack-to-Code transformations, while their composition models the complete transformation pipeline. Zhu with co-authors [71] demonstrates the application of categories in verification, where implementation respects design in concurrent systems. Analogously, commutativity of diagrams is required during refactoring (rename/split/merge) and preservation of invariants (cardinalities, mandatory fields, completeness of process transitions).

A closely related and complementary line of work targets the automatic generation of domain behavior from requirements and knowledge-representation models. In particular, [59] derives application logic from requirement and knowledge models rather than only structural skeletons. Our work is orthogonal in scope: whereas [59] focuses on synthesizing behavior, we provide a formally specified, structure-preserving, multi-target transformation backbone. The two directions are combinable: the behavioral synthesis of [59] could populate the explicit method-body extension points ("hooks") that our pipeline leaves open, while our categorical framework would supply the structural and multi-platform guarantees.

The methodology presented in this paper directly concretizes theoretical insights from the cited literature. Relying on Diskin's concept of categorical design patterns, the proposed transformation system models code generation not as mere textual substitution, but as a functor mapping the structure of the ontology into the structure of the target language. This achieves the mathematical guarantee of structural correctness advocated by Kovalyov. At the same time, acknowledging the conclusions of Mkhinini et al., this paper uses ontology as a super-model that complements the shortcomings of traditional UML modeling, providing the semantic basis necessary for the automated generation of schema-level application structure and selected architectural realization artifacts. The proposed methodology thus defines formal, measurable criteria for the verification of transformations, directly derived from the application of category theory. The purpose of these criteria is to verify that the generated code remains faithful to the selected schema-level structure and explicitly represents mapping distinctions of the core model. These criteria encompass: (1) checking the commutativity of the relevant diagrams during transformation and refinement operations and (2) validating the preservation of defined invariants, such as cardinalities and transition completeness, through each transformation step.

3. Ontologies as a Basis for Model Specification

Ontologies are formal specifications of shared conceptualizations, expressed in a language with well-defined syntax and semantics, intended to capture the structure of a domain in a machine-interpretable form. In the context of MDE and code generation, ontologies serve as a particularly suitable foundation for specifying platform-independent models: their formal syntax permits automated processing, their declarative semantics admits verification through reasoning, and their extensibility through import mechanisms supports modular construction of domain and platform descriptions. This section introduces the ontological concepts used throughout the paper. The presentation is kept to the minimum needed for the categorical formalism of subsequent sections; for a comprehensive treatment of the semantic web technology stack, we refer the reader to [35].

3.1. OWL 2 Ontologies basics

The Web Ontology Language (OWL 2) is a W3C standard for expressing ontologies, based on description logic. The language defines several computational profiles; in this work, we use OWL 2 DL, the profile corresponding to the description logic, which provides a balance between expressiveness and decidability of reasoning. The BAB Core ontology, the BABj ontology, and all domain ontologies constructed on the top of them conform to this profile.

Classes, Individuals, and Object Properties An OWL ontology consists of three primary kinds of entity:

- Classes represent types or categories of concept, for instance, `core:Entity`, `core:Process`, or `pms:Shipment`. Formally, a class is interpreted as a set of individuals sharing structural or semantic characteristics.
- Individuals represent specific instances of classes, for instance, a particular shipment `pms:shipment123` as an instance of `pms:Shipment`. In description-logic terminology, individuals are the elements of the universe of discourse.
- Properties represent binary relations, and are subdivided into object properties, which relate pairs of individuals (e.g., `core:hasAttribute` relating an entity to its attribute), and datatype properties, which relate an individual to a literal value (e.g., a string, integer, or date).

Each property declaration carries `rdfs:domain` and `rdfs:range` restrictions that specify, respectively, the class of admissible sources and the class or datatype of admissible targets. These declarations enable both syntactic validation of the ontology and inference of class membership from property usage.

Object Properties versus Datatype Properties The distinction between object properties and datatype properties is structurally significant for the methodology developed in this paper. An object property links two individuals, both of which are first-class entities in ontology and can themselves participate in further relations. A datatype property links an individual to a literal value (a string, number, or other XSD-typed datum) that has no further ontological structure. Consequently, only object properties generate navigable relations in the class diagram of the ontology, and only they contribute edges to the schema graph on which the categorical formalism of Section 5.4 is built. Datatype properties carry information at the instance level but do not participate in the schema-level structure on which functor construction depends. This distinction will be used in Section 5.1 to justify the design of the `ImplementationMapping` mechanism.

Subsumption and Subproperty Hierarchies Classes and properties in OWL form hierarchies through the `rdfs:subClassOf` and `rdfs:subPropertyOf` relations. A declaration asserts that every instance of is also an instance of ; analogously, asserts that every pair related by is also related by . These hierarchies express taxonomic organization of the domain and serve as the basis for several mechanisms used in this work: domain specialization (a domain class becomes a subclass of an abstract BAB Core class), refinement of mapping strategies (a specialized mapping class becomes a subclass of a general one), and inheritance of properties from classes to their subclasses.

Reification A modeling decision frequently encountered in ontology engineering is whether to represent a concept as a property or as a class. A property, being a binary relation, can associate a subject with an object but cannot itself carry structured information about the nature of the association. A class, in contrast, admits instantiation into individuals and participation in its own hierarchy and relations. When a concept inherently carries structured information, multiple attributes, internal composition, or its own refinement structure, modeling it as a class rather than a property is the standard choice. This transformation is known as reification, and it is the modeling pattern underlying the `core:ImplementationMapping` construct introduced in Section 5.1.

The Open World Assumption OWL 2 operates under the Open World Assumption (OWA): a fact that is not asserted in the ontology is treated as unknown, not a false. This contrasts with the closed-world assumption common in relational databases, where absence from the database is interpreted as absence from the world. Under OWA, the absence of a `hasAttribute` declaration for a particular entity does not entail that the entity has no attributes; it entails only that the ontology does not record such information. Consequences for the presented work are modest but important: statements of the form "every entity has at least one attribute" cannot be enforced by mere silence but must be expressed through explicit cardinality restrictions, verified by the reasoner, to take effect.

Reasoning A reasoner is an automated procedure that derives implicit consequences from the explicit assertions of an ontology, using formal semantics based on underlaying logical language. Standard reasoning tasks include consistency checking (verifying that the ontology admits a non-empty interpretation), classification (computing the full subsumption hierarchy of classes), and realization (inferring the most specific class memberships of individuals). In this work, we use HermiT [31], a complete OWL 2 DL reasoner, to validate domain ontologies constructed on top of the BAB Core: cardinality constraints, property restrictions, and disjoint declarations defined in BAB Core serve as validation criteria, and the reasoner's verdict, consistent or inconsistent, determines whether a domain ontology is admitted as input to the transformation pipeline. The role of the reasoner in the overall methodology is discussed in Section 5.6.

SPARQL SPARQL is the W3C standard query language for RDF graphs, and by extension for OWL ontologies serialized as RDF. A SPARQL query consists of a pattern of triples with variables; the query engine returns all bindings of the variables that make the pattern a subgraph of the queried data. SPARQL is the mechanism by which the implementation of the proposed methodology extracts structural information from domain and platform ontologies during transformation: the functorial specifications developed in Section 5.4 are operationally realized by a collection of SPARQL queries that traverse the ontology graph along the declared object properties and collect the resulting structural data into an in-memory representation. The implementation details are presented in Section 6.

Ontology Import OWL 2 provides an `owl:imports` construct by which one ontology incorporates the class, property, and individual declarations of another ontology. Imports

are transitive and support modular construction of complex ontological landscapes: a domain ontology (such as the Postal Management System ontology used in later examples) imports BAB Core; the BABj platform ontology imports both BAB Core and the Web of Code (WoC) vocabulary; and domain ontologies targeting the Java stack import BABj. This layered structure is central to the methodology of this paper, allowing a separation between platform-independent abstract vocabulary (BAB Core), platform-specific structural vocabulary (BABj, based on WoC), and domain-specific elaboration.

Schema versus Data: T-box and A-box An OWL ontology is conventionally understood to comprise two strata. The T-box (terminological box) contains class and property declarations, subsumption hierarchies, and logical axioms, i.e., the schema of the domain. The A-box (assertional box) contains individuals and assertions of their class memberships and property values, the data populating the schema. The transformation formalized in this paper operates at the T-box level: it transforms one schema (BAB Core together with a domain ontology) into another (BABj with a generated class hierarchy and methods) and ultimately into source code. Assertions about individuals, which constitute the A-box, are not part of the transformation, and consequently the formal guarantees proved in Section 5.5 are guarantees about the schema structure, not about the runtime data flowing through the generated system. This distinction will be made explicit in the scope discussion accompanying the main preservation result.

4. Category Theory as a Basis for Model Transformations

Category Theory (CT) is a branch of mathematics that studies structures and relationships between them at a high level of abstraction. It is often described as "mathematics about mathematics" because it provides a unified language and formalism for describing various mathematical domains (such as set theory, algebra, topology). For software engineering, and especially for Model-Driven Development and model transformations, CT offers a rigorous foundation for precisely defining, composing, and verifying systems and processes. The basic idea is to shift the focus from elements (what is inside a set or class) to the relationships between them (morphisms or "arrows"). Roughly speaking, category theory expresses the view that an element has no internal structure; rather, all knowledge about an element is contained in its connections with other elements.

4.1. Basic Concepts of Category Theory

This section concisely presents the basic concepts of category theory used in the proposed methodology: category, functor, natural transformation, and bifunctor as fundamental concepts, as well as adjunction as a construct that enables the formal interpretation of bidirectional mappings between models and code.

Category A category $\mathcal{C}$ is formally defined by two sets of elements:

- A set of objects, denoted as $\mathrm{Ob}(\mathcal{C})$.
- A set of morphisms (arrows), denoted as $\mathrm{Mor}(\mathcal{C})$.

Each morphism f has exactly one source object X and one target object Y, written as $f : X \to Y$. The elements of a category must satisfy two conditions (axioms):

- Composability: For any two morphisms $f : X \to Y$ i $g : Y \to Z$, there must exist a third morphism $h = g \circ f$ (read: "g after f") which maps $h : X \to Z$. This composition must be associative: $(h \circ g) \circ f = h \circ (g \circ f)$.
- Identity: For every object X in $\mathrm{Ob}(\mathcal{C})$, there must exist a special identity morphism $\mathrm{id}_X : X \to X$, which acts neutrally in composition. For every $f : A \to X$ and $g : X \to B$, it holds that: $\mathrm{id}_X \circ f = f$ and $g \circ \mathrm{id}_X = g$.

Examples of categories:

- Set: A category where objects are sets and morphisms are functions between sets.
- Types: In programming languages, objects can be data types (e.g., `int`, `String`), and morphisms can be functions (e.g., `toString`: `int` $\to$ `String`).
- Ontology: An ontology can be viewed as a category where objects are concepts and morphisms are relations between them.

Functor Functors are mappings between categories that preserve the structure. If there are two categories, $\mathcal{C}$ and $\mathcal{D}$, a functor $F : \mathcal{C} \to \mathcal{D}$ defines:

- A mapping of every object X from $\mathcal{C}$ to an object $F(X)$ in $\mathcal{D}$.
- A mapping of every morphism $f : X \to Y$ from $\mathcal{C}$ to a morphism $F(f) : F(X) \to F(Y)$ in $\mathcal{D}$.

For this mapping to be a functor, it must "respect" the structure of category $\mathcal{C}$:

- It must preserve identities: $F(\mathrm{id}_X) = \mathrm{id}_F(X)$ for every object X in $\mathcal{C}$.
- It must preserve composition: $F(g \circ f) = F(g) \circ F(f)$ for all morphisms f and g.

A functor can be interpreted as a structural translator, enabling different domains (e.g., an ontology and a database) to be connected through the same abstract structure.

Bifunctor A bifunctor is a functor of two arguments, formally, a functor from a product category $C \times D$ to a target category E. Given two categories C and D, their product category $C \times D$ has pairs (A, B) as objects and pairs of morphisms (f, g) composed component-wise. A bifunctor $F : C \times D \to E$ maps each pair of objects to an object in E and each pair of morphisms to a morphism in E, preserving identities and composition in both arguments simultaneously. Bifunctors arise naturally in settings where a construction depends on two independent inputs, such as the evaluation functor $eval : [C, D] \times C \to D$ that applies a functor to an argument.

Isomorphism in Categories An isomorphism constitutes a morphism within a category that is formally invertible. A morphism $f : A \to B$ is defined as an isomorphism if there exists an inverse morphism $g : B \to A$ satisfying the conditions $g \circ f = \mathrm{id}_A$ and $f \circ g = \mathrm{id}_B$, where id denotes the identity morphism on the respective objects. When such a relationship exists, the objects are considered isomorphic, implying they are structurally indistinguishable from the perspective of that category. In a programming context, this equivalence is exemplified by the mapping between a `String` and a `List<Char>`, where the transformation allows for lossless bidirectional conversion of the underlying data structure.

Natural Transformation Natural transformations are a central concept used in this paper. If functors are mappings between categories, natural transformations are mappings between functors, defined as follows.

Let F and G be two functors, both mapping from the same source category $\mathcal{C}$ to the same target category $\mathcal{D}$ ($F, G : C \to D$). A natural transformation $\tau : F \Rightarrow G$ (read as: "tau from F to G") is a family of morphisms in the target category $\mathcal{D}$. For every object X in $\mathcal{C}$, τ provides a morphism $\tau_X : F(X) \to G(X)$ in category $\mathcal{D}$.

Furthermore, this family of morphisms must satisfy the naturality condition (or "naturality square") $G(m) \circ \tau_X = \tau_Y \circ F(m)$, implying that for every morphism $m : X \to Y$ in category $\mathcal{C}$, the following diagram (Figure 1) in category $\mathcal{D}$ must be commutative.

$$X \xrightarrow{m} Y$$

$$\begin{array}{ccc} F(X) & \xrightarrow{\tau_X} & G(X) \\ {\scriptstyle F(m)}\downarrow & & \downarrow{\scriptstyle G(m)} \\ F(Y) & \xrightarrow{\tau_Y} & G(Y) \end{array}$$

Fig. 1. Naturality Square Diagram

Intuitively, the naturality condition signifies that the order of applying the functor and the transformation does not affect the result; both paths in the diagram yield the same outcome:

- Path 1 ($\tau_Y \circ F(m)$): First apply functor F to relation m (from $F(X)$ to $F(Y)$), and then the transformation component τ_Y (going from $F(Y)$ to $G(Y)$).
- Path 2 ($G(m) \circ \tau_X$): First apply the transformation component τ_X (from $F(X)$ to $G(X)$), and then functor G to relation m (from $G(X)$ to $G(Y)$).

If these two paths are always equal, the transformation τ is "natural". This means that τ is not just a random collection of relations, but a coherent process that respects the structure defined by functors F and G.

Natural Isomorphism Distinct from an isomorphism between individual objects, a natural isomorphism establishes a structural equivalence between functors. Formally, let F and G be functors between categories $\mathcal{C}$ and $\mathcal{D}$. A natural transformation $\eta : F \Rightarrow G$ is defined as a natural isomorphism if, for every object X in $\mathcal{C}$, the component morphism

$\eta_X : F(X) \to G(X)$ constitutes an isomorphism within $\mathcal{D}$. This condition implies that the functors F and G are isomorphic as objects in the functor category $[C, D]$, meaning they represent C in D in the same way up to isomorphism. Consequently, constructions and morphisms expressed via F correspond uniquely and functorially to those expressed via G. As a computational analogy, a natural isomorphism may relate a functor that represents data as serialized JSON to one that represents the same data as an in-memory Map, provided that the conversion is bijective, structure-preserving, and compatible with all data transformations (i.e., commutes with the corresponding morphisms).

Adjunction An adjunction $I \dashv J$ is a pair of functors $I : \mathcal{C} \rightleftarrows \mathcal{D} : J$ with a natural isomorphism:

$$\mathrm{Hom}_D(I(X), Y) \cong \mathrm{Hom}_C(X, J(Y)) . \tag{1}$$

The natural transformations $\eta : \mathrm{Id}_{\mathcal{C}} \Rightarrow JI$ (unit) and $\varepsilon : IJ \Rightarrow \mathrm{Id}_{\mathcal{D}}$ (counit) exist and satisfy the triangle identities as a direct consequence of the above natural isomorphism.

4.2. Research Hypothesis

Based on the ontological and categorical concepts introduced above, the fundamental hypothesis of this research is that ontological modeling and category theory provide a rigorous apparatus for formalizing the process of ontology-driven code generation in a way that (i) makes schema-level information preservation a verifiable property rather than a design claim, (ii) supports multi-target generation through parametrization rather than code duplication, and (iii) cleanly separates the domain model from its platform-specific realization.

To prove our hypothesis, we formalize the ODCG process as a sequence of three categories connected by two functors:

$$C_{\mathrm{Core}} \xrightarrow{F_s} C^s_{\mathrm{BAB}} \xrightarrow{H_s} C^s_{\mathrm{code}} \tag{2}$$

where C_{Core} is the schema category, constructed from the BAB Core ontology, C^s_{BAB} is the platform-specific schema category for technology stack s, and C^s_{code} is the category of generated source code artifacts. For each stack s, functor F_s performs the model-to-model (M2M) transformation from platform-independent to platform-specific representation, while functor H_s performs the model-to-code (M2T) transformation from the platform-specific schema to concrete source code files.

The functor F_s is further represented through an evaluation bifunctor $\{F_s\}_{s \in S}$ parametrized by the target technology stack $s \in S$, with each F_s arising from a conformance functor Φ_s that interprets abstract mapping schema elements as platform-specific generation rules. This decomposition captures the fact that code generation depends functorially on two independent inputs, the domain model and the mapping class, and that multi-target generation is achieved by varying the stack index s without modifying the domain model or the generation architecture.

The correctness of the overall transformation is formalized through structural properties of the composite functor $T_s = H_s \circ F_s$: totality, structural coherence, faithfulness, and injectivity on objects. Together, these properties constitute a conditional schema-level

guarantee that, under the stated verification assumptions, the selected structural information and explicitly represented mapping distinctions are not lost or collapsed during transformation to generated code.

The categorical framework suggests a possible extension to round-trip engineering through the introduction of an adjunction $I_s \dashv J_s$ between a forward generation functor $I_s : C_{\mathrm{Core}} \to C^s_{\mathrm{code}}$ and a reverse abstraction functor $J_s : C^s_{\mathrm{code}} \to C_{\mathrm{Core}}$. Establishing such an adjunction, together with suitable semantic conditions, could provide a formal basis for consistent generation and recovery. This extension is outside the scope of the present work and is discussed in Section 8 as a direction for future research.

5. Formal Model of the ODCG Process

The proposed methodology formalizes the ODCG process as a multi-phase, deterministic transformation based on three levels of abstraction: a Platform Independent Model (PIM), a Platform Specific Model (PSM), and the final generation of executable code. The methodology encompasses three key artifacts: BAB Core ontology as the PIM, the BABj Java library and ontology as the target architecture (PSM), and the model transformer as the mechanism executing the transformations (Figure 2). This division ensures that each phase has clearly defined semantics and formal verification that models obtained by transformation preserve the structural properties of the source model.

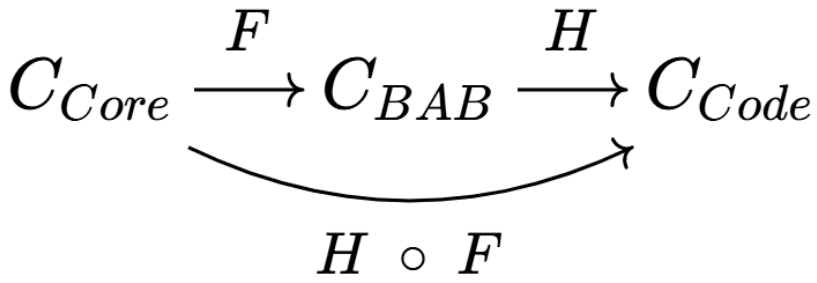


Fig. 2. Overview of the multi-phase transformation from ontology to code

The presentation of this section proceeds in two passes. The first three subsections describe the ontological layer of the methodology, the three artifacts as OWL 2 DL ontologies and as generated source files, without invoking any categorical apparatus. This pass is self-contained and readable by those interested in the ontology engineering aspects of the methodology. The subsequent subsection introduces the categorical layer, constructing schema categories from the above-mentioned ontologies and formalizing the transformations between them as functors. The categorical formalism is developed in one place, after the reader is familiar with all three ontological artifacts, rather than interleaved with the ontology descriptions. This separation is deliberate so that readers with different backgrounds can navigate the two layers independently.

5.1. Level 1: Platform Independent Model (PIM) – BAB Core Ontology

The first and highest level of abstraction is the PIM, realized as the BAB Core ontology, a semantic framework for modeling PAIS. Unlike standard domain ontologies that classify data, BAB Core is designed as an ontology of a software system structure: it does not describe a specific business domain but abstracts the fundamental building blocks of any information system, establishing a strict separation between business semantics and technological implementation.

BAB Core ontology (Figure 3) is a formal meta-model of an information system defined using OWL 2 DL (Web Ontology Language, Description Logic profile). It provides the semantic skeleton on which domain-specific applications are constructed through ontology import and specialization.

Core classes The root of the BAB Core taxonomy consists of the following classes, each modeling an essential aspect of application architecture:

- core:Entity: Represents the basic unit of persistence and data structure (equivalent to an entity in the ER model or a class in OOP).
- core:Process and core:Task: Model the dynamic behavior of the system, where processes are decomposed into atomic units of work (e.g., core:UserTask for human interactions and core:ServiceTask for automated operations).
- core:UIElement: Abstracts the presentation layer through concepts such as core:View and core:Form, enabling a declarative description of the user interface independent of a concrete frontend framework (e.g., Angular, React, or Vaadin).
- core:Service and core:Module: Define the logical organization of business logic and application granularity.

Core properties Object properties in BAB Core formalize structural and behavioral relations among these classes. System composition is defined by relations such as core:hasAttribute (entity structure), core:usesEntity (service-to-data dependency), and core:hasTask (process decomposition), while the workflow order is formalized by core:precedes. These relations impose architectural constraints on any domain ontology built on the top of the BAB Core, ensuring that every instantiated model possesses structural integrity before any transformation takes place.

The ImplementationMapping mechanism A key design decision in BAB Core is that code generation rules are not hidden in a generator algorithm but are explicitly modeled within the ontology itself. This is achieved by reifying generation rules as an OWL class core:ImplementationMapping, rather than as an OWL object property. The reification is deliberate and has modeling consequences:

- A property, by its nature, has only a subject and an object, it can relate a domain concept to a target element, but it can't itself carry structured information about how the relation is realized on a specific platform.

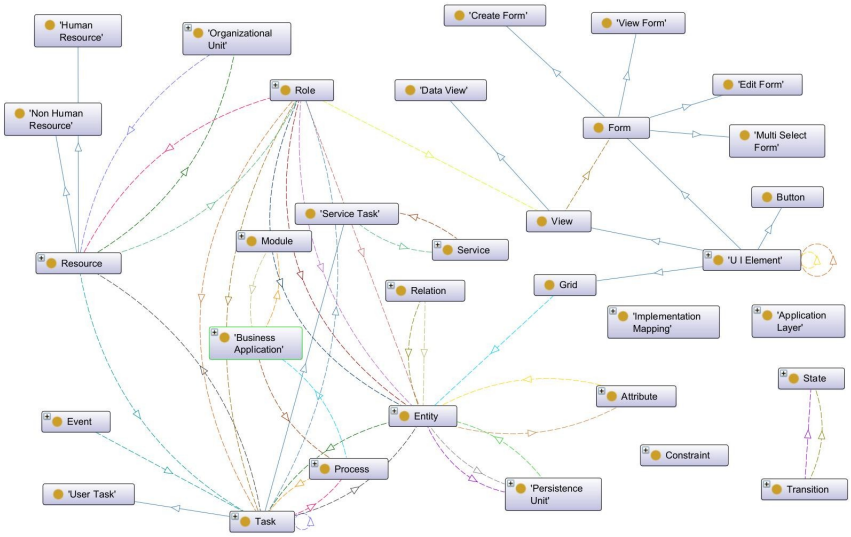


Fig. 3. BAB Core Ontology Diagram

- A class, on the other hand, admits a hierarchy of subclasses through rdfs:subClassOf, allowing progressive refinement of mapping strategies: specific mapping classes such as core:EntityMapping, core:ViewMapping, core:ProcessMapping, and core:TaskMapping are declared as subclasses of core:ImplementationMapping, and further refinements (e.g., core:PersistentEntityMapping extending core:EntityMapping, or core:ValidatableEntityMapping for entities requiring validation logic) may be introduced to specialize generation behavior for particular architectural patterns. Each of these classes can then be instantiated into named individuals in stack-specific ontologies (such as babj:DefaultEntityMapping_J in the BABj ontology), where each such individual carries the concrete platform-specific bindings, the target library class, default annotations, and auxiliary properties, needed for code generation.
- The association between a domain concept and its mapping is expressed through a separate object property core:hasImplementationMapping, which connects each domain concept to one or more ImplementationMapping individuals governing distinct aspects of its platform realization (e.g., persistence, DTO projection for a specific view, UI binding).

This reified design makes the mapping mechanism first-class ontological content: refinements are expressible as subclass relations among mapping classes, inspection and querying of mapping strategies are possible through standard SPARQL queries, and the reasoning layer can verify mapping consistency as a part of ontology validation. The essential role of the refinement structure is developed in the Section 5.4.3, where it is shown to generate natural transformations between stack-specific generation functors.

Business domain specification A platform-independent model of the specific business domain is built by creating a domain-specific ontology (e.g., for a Postal Management System, PMS) that imports BAB Core (owl:imports) and specializes its classes. In this process, specific domain terms (e.g., pms:Shipment) become subclasses of abstract BAB Core constructs (pms:Shipment rdfs:subClassOf core:Entity). BAB Core thus acts as a semantic skeleton imposing architectural patterns and constraints, while the domain ontology "fills" that skeleton with specific business logic.

Axiomatic layer BAB Core further declares necessary OWL 2 DL axioms comprising inverse property declarations (e.g., core:hasAttribute owl:inverseOf core:isAttributeOf), subproperty hierarchies, and domain/range restrictions, which govern instance-level reasoning performed by a reasoner (e.g., Hermit) during ontology validation. These axioms are used to verify consistency of domain ontologies built on the top of BAB Core but play no role in the schema-level categorical transformation developed in Section 5.4. The formal separation between instance-level reasoning and schema-level transformation is returned to in Section 5.4.1.

The formal categorical representation of this ontology and the mechanism that enables technology-neutral code generation are presented in Section 5.4.

5.2. Level 2: Platform Specific Model (PSM) – BABj Library and Ontology

The second level of abstraction, the PSM, is exemplified as the BABj Java library and the BABj ontology. Together they define the target architecture into which domain models are transformed. The BABj ontology is constructed on the top of the Web of Code (WoC) vocabulary [5], which provides a standard OWL 2 DL representation of object-oriented programming language constructs, and is extended with artifacts from the BABj library specific to the BAB framework.

The BABj Library The BABj library provides infrastructural scaffolding on which generated code depends. Its elements are divided into two kinds:

- Structural elements: base classes such as AbstractEntity and AbstractHome, which provide infrastructural mechanisms for persistence, session handling, and transaction management.
- Behavioral elements: interfaces defining business logic contracts, such as CheckValidateEntityInterface, CheckArrivedEntityInterface, and CheckPayoutEntityInterface, which are implemented by generated code to inject domain-specific logic like validation, arrival, and payout.

The generated code relies on these library elements through Java inheritance and interface realization. This division ensures that generated classes remain thin and focused on domain logic, while all boilerplate and infrastructure are delegated to the library.

The BABj Ontology The BABj ontology represents, in OWL 2 DL, the structural vocabulary that generated code will instantiate. It imports the WoC vocabulary and extends it with BABj-specific library classes and mapping declarations. Its content consists of:

- Imported WoC constructs, representing generic Java programming-language elements: woc:Class, woc:Interface, woc:Field, woc:Method, woc:hasField, woc:hasMethod, woc:extends, woc:implements, and so forth.
- BABj library classes, added as OWL classes referring to the concrete Java types in the BABj library (e.g., AbstractEntity, AbstractHome, the validation interfaces), enabling the ontology to express that a generated class extends or implements a specific library element.
- Mapping individuals, declared as instances of specific refinement classes from the core:ImplementationMapping hierarchy, which specify how each PIM concept is realized in the Java stack. Examples include babj:DefaultEntityMapping_J (an instance of core:EntityMapping or its appropriate refinement, specifying that core:Entity is realized as a woc:Class annotated with @Entity and extending babj:AbstractEntity), babj:DefaultProcessMapping_J (an instance of core:ProcessMapping), and so forth. Each individual provides concrete binding information via core:mapsToTypeSpec (pointing to the BABj class that realizes the mapping) and auxiliary annotation properties. The association between a mapping individual and its target stack is established by the PSM ontology in which the individual is declared (e.g., a mapping individual declared in BABj implicitly targets the Java stack).

These mapping individuals are the concrete, stack-specific realizations of the abstract mapping schema defined in BAB Core. Because they are declared as instances of core:ImplementationMapping and further refined through rdfs:subClassOf relationships, they inherit the structural discipline of the core mapping mechanism while providing platform-specific detail.

Axiomatic layer. Like BAB Core, the BABj ontology carries OWL 2 DL axioms, including WoC's inverse property declarations (woc:hasField owl:inverseOf woc:isFieldOf, etc.), functional property constraints, and subproperty hierarchies, which are used by reasoner for instance-level consistency checking. These axioms primarily belong to the reasoning layer and do not automatically enter the schema-level categorical apparatus of Section 5.4. Selected schema-level equations may nevertheless be introduced explicitly in the presented target category when required by the transformation semantics. This separation is discussed in detail in Section 5.4.1.

The same approach applies to other platforms by replacing constituents of the PSM with corresponding platform-specific ontologies/libraries.

Integration of Mapping Across the Two Ontologies A significant design decision in the methodology is that ontological mappings are not isolated as separate ontology but are distributed across BAB Core and BABj (i.e., platform-specific) components. BAB Core declares the abstract ImplementationMapping class and its core refinement hierarchy, while BABj instantiates this class with stack-specific individuals. This distributed design means that the reasoner can verify both halves, the abstract mapping schema and its concrete realization, as part of a single ontology consistency check.

In this integrated structure, BAB Core defines the meaning and identity of concepts, while BABj describes their realization through programming constructs and annotations. The reasoner verifies that the mapping individuals in BABj conform to the abstract schema in BAB Core, ensuring ontological consistency between the abstract schema and its

platform-specific realization. AI-assisted mapping may be employed at this level, where the transformer could propose likely implementations based on semantic similarity while a human (engineer) should confirm the proposal, maintaining formal control and process validity.

The formal representation of BABj as a category, and the formalization of the translation from BAB Core to BABj as a bifunctor, are developed in Section 5.4.

5.3. Level 3: Generated Code – Category C^s_{code}

The third and the final level of our MDA is a free category of generated code, C^s_{code}, constructed over directed graph that represents internal programming language structure. Its objects are generated source-artifact specifications, typically materialized as source files, while morphisms are structural dependency relations such as import, extends, implements, field-type dependencies, and method-signature dependencies. The generated code relies entirely on the corresponding (i.e., BABj) library, achieving architectural consistency and minimizing redundancy.

5.4. Transformation Process: $H_s \circ F_s$

This subsection presents the entire categorical formalism of the methodology. Having introduced the three ontological artifacts in previous sections, we now construct schema categories from these ontologies and formalize the transformations between them as functors. The subsection is organized as follows: first, we introduce the categorical constructions used in the formalism and record their well-definedness by construction; next, we define and verify the model-to-model functor F_s together with the ImplementationMapping mechanism that parametrizes it over technology stack s; finally, we define the model-to-code functor H_s that completes the transformation pipeline.

The intermediate category C^s_{BAB} can be suppressed when the transformation is viewed only at the level of the composite mapping, because the model-to-model and model-to-code functors compose to a single functor $T_s = H_s \circ F_s$. Nevertheless, C^s_{BAB} remains essential to the chosen factorization, specification, and verification of that composite: it is at this level that the stack-parametrized bifunctor F_s, the presented-category relations E^s_{BAB}, and the refinement natural transformations are defined; and it is through C^s_{BAB} that T_s factors.

Construction of the three categories Before constructing the categories, we emphasize a distinction that is essential to the proposed formalism. The BAB Core and BABj/WoC ontologies contain two kinds of content relevant to the methodology. The structural schema layer consists of OWL classes, object properties, and the declarations used to extract the schema graph. This layer determines the objects and generating morphisms of the categories constructed below. The logical axiomatic layer consists of OWL 2 DL axioms, such as inverse-property declarations, functional and inverse-functional property declarations, transitivity, symmetry, and subproperty axioms. These axioms govern semantic constraints on ontology interpretations and are used by the OWL reasoner during validation.

The categorical formalism reflects these layers asymmetrically. The source category, representing the PIM layer, is constructed as the free category generated by its schema graph. This allows the model-to-model transformation to be defined on objects and generating edges, with no source-side equations to verify. For each target technology stack, the corresponding target category, representing the PSM layer for that stack, is constructed as a presented category [47]. Its generating graph is extracted from the stack-specific ontology and its imported code-structure vocabulary. Its relations form an explicitly selected set of schema-level path equations required by the transformation semantics of that stack. These equations may be motivated by the intended structural semantics of the target platform and by selected ontology declarations used as schema-normalization rules, but they should not be understood as automatic translations of arbitrary OWL 2 DL axioms into categorical equations. Consequently, the Java case is obtained as an instance based on the BABj ontology and the imported WoC vocabulary.

Category C_{Core} Let G_{Core} denote the directed multigraph underlying the schema layer of the BAB Core ontology. The nodes of G_{Core} are the OWL classes declared in BAB Core. The directed edges of G_{Core} are the object properties declared in BAB Core, each edge labeled by the property IRI, with the edge's source and target given by the property's domain and range declarations.

We define C_{Core} as the free category generated by G_{Core}:

- $Ob(C_{Core})$ is the set of nodes of G_{Core}.
- $Mor(C_{Core})$ consists of finite (possibly empty) sequences of composable edges of G_{Core}, where a sequence $(e_1, e_2, \ldots, e_n)$ with $e_i : X_i \to X_{i+1}$ represents a morphism $X_1 \to X_{n+1}$. The empty sequence at X represents the identity morphism id_X.
- Composition of morphisms is concatenation of sequences, which is associative by construction of concatenation.
- Identities act neutrally: concatenation with an empty sequence leaves any sequence unchanged.

Construction 1. The BAB Core schema category C_{Core}.

The category C_{Core} is defined as the free category generated by the directed multigraph G_{Core} extracted from the schema-level layer of the BAB Core ontology. Hence, its objects are OWL classes, its generating morphisms are BAB Core object properties, and arbitrary morphisms are finite composable paths of such properties. Identities are empty paths and composition is given by path concatenation. Therefore, C_{Core} is a well-defined category by the standard free-category construction.

Decomposition of C_{Core}. The object set of C_{Core} splits into two disjoint subsets:

- $Ob\left(C_{Core}^{NonMapping}\right) = \{X \in Ob\,(C_{Core}) \mid X$ is not a subclass of $core : ImplementationMapping\}$ containing all domain concepts.
- $Ob(C_{Core}^{MappingOnly}) = \{X \in Ob(C_{Core}) \mid X$ is core:ImplementationMapping or one of its subclasses$\}$ containing the reified mapping elements classes.

These object sets are disjoint by construction, and their union exhausts $Ob(C_{Core})$.

Let $C_{Core}^{NonMapping}$ and $C_{Core}^{MappingOnly}$ be the full subcategories of C_{Core} determined by the corresponding object sets. Each subcategory contains all morphisms of C_{Core} between objects in its own object set, and the category axioms are inherited automatically. Since the cross-category generating morphisms induced by core:hasImplementationMapping are oriented from NonMapping objects to MappingOnly objects, and no generating edge in G_{Core} is oriented from a MappingOnly object back to a NonMapping object, no path in C_{Core} between two NonMapping objects can pass through a MappingOnly object. Hence every such path lies entirely in the induced graph $G_{Core}^{NonMapping}$, and $C_{Core}^{NonMapping}$ is the free category generated by $G_{Core}^{NonMapping}$. The same argument applies to $C_{Core}^{MappingOnly}$: no path between two MappingOnly objects leaves the MappingOnly part of the graph. Therefore $C_{Core}^{MappingOnly}$ is also freely generated by its induced graph.

Observation 1. Decomposition of C_{Core}.

The object set of C_{Core} decomposes into domain-level concepts and reified mapping concepts. These determine two full subcategories, $C_{Core}^{NonMapping}$ and $C_{Core}^{MappingOnly}$. Morphisms generated by core:hasImplementationMapping connect objects from the former to objects from the latter; therefore, they belong to C_{Core} but not to either full subcategory.

The significance of Observation 1 is that it justifies decomposing C_{Core}'s object set into the disjoint union $Ob\left(C_{Core}^{NonMapping}\right) \sqcup Ob\left(C_{Core}^{MappingOnly}\right)$ while still recognizing that the cross-category morphisms (generated by core:hasImplementationMapping) exist in C_{Core}. These cross-category morphisms do not belong to either full subcategory and therefore do not enter the product category $C_{Core}^{NonMapping} \times C_{Map}$ that serves as the domain of the bifunctor in Section 5.4.3. Their role is semantical: they specify which pairs (D, M) carry intended architectural meaning, as developed in Section 5.4.3. The schema-level edge in G_{Core} corresponding to core:hasImplementationMapping is extracted from its declared rdfs:domain and rdfs:range: domain classes on the left, the abstract core:ImplementationMapping class (or its declared subclasses) on the right. The instance-level assertions found in the operational ontology are ABox content that provides the data from which the schema-level pair (D, M) is recovered: D is the domain class subject of the assertion, and M is the rdf:type of the asserted mapping individual. The schema-level cross-edge thus exists structurally in C_{Core}, while the ABox assertions populate the set of selected pairs that carry intended architectural meaning.

Mapping-refinement category C_{Map} Although $C_{Core}^{MappingOnly}$ identifies the mapping-related objects inside C_{Core}, refinements between mapping strategies are governed by the $rdfs : subClassOf$ hierarchy of ImplementationMapping classes rather than by ordinary BAB Core object properties. Therefore, we introduce a separate mapping-refinement category C_{Map}. Its objects are core:ImplementationMapping and its subclasses, and its morphisms are generated by the refinement order induced by the rdfs:subClassOf hierarchy. The arrows are oriented from a more general mapping class to a more specific one: a morphism $h : M \to M'$ in C_{Map} means that M' is a refinement, i.e., a subclass-specialization, of M. This category is used as the second argument of the conformance and evaluation constructions below. Formally, C_{Map} is the preorder category of mapping refinements. There is at most one morphism $M \to M'$, and such

a morphism exists exactly when M' is a subclass-specialization of M, that is, when $M'\ rdfs\ :\ subClassOf^*\ M$. Identities and composition follow from reflexivity and transitivity respectively.

Category C^s_{BAB} Let s denote an arbitrary target technology stack. Let O^s_{BAB} denote the stack-specific ontology for s, and let V^s denote the code-structure vocabulary imported by that ontology. We define G^s_{BAB} as the directed multigraph extracted from the schema layer of O^s_{BAB} together with V^s. The graph G^s_{BAB} represents the stack-specific realization signature, not a concrete generated application model. Its nodes are stack-specific realization types and realization patterns. A realization pattern specifies how a BAB Core concept of a given kind, together with a selected mapping class, can be represented in the target stack. Such patterns include platform-level type specifications, metadata or annotations, inheritance or composition constraints, interface or contract declarations, template roles, imports, fields, methods, and other structural parameters used by the generator. The directed edges of G^s_{BAB} are selected schema-level relations used as structural navigation and refinement relations inside the target realization signature. In particular, G^s_{BAB} provides admissible platform-level relation labels and refinement relations. Concrete domain concepts are not part of the definition of G^s_{BAB}; they are supplied later, when the conformance functor $\Phi_s(M)$ applies these stack-specific realization patterns to objects of $C^{NonMapping}_{Core}$.

Let $Free\,(G^s_{BAB})$ be a free category generated by G^s_{BAB}. Let E^s_{BAB} be an explicitly specified set of pairs (u, v) of parallel paths in $Free\,(G^s_{BAB})$, where $u, v : X \to Y$. They may include: (i) conformance equations, such as the naturality equations introduced in Section 5.4.3; (ii) refinement-coherence equations, such as equations identifying a selected direct refinement edge with the corresponding composite refinement path; and (iii) explicitly intended normalization equations, where the target stack deliberately treats two alternative structural paths as the same realization pattern.

The third case is optional and must be declared explicitly as part of the stack-specific transformation specification. Such normalization equations are not introduced automatically and do not express arbitrary semantic equivalence of target-language programs. They only state that, for the purposes of the selected schema-level transformation, two parallel structural paths are intentionally treated as the same target-side realization.

Let $\equiv_E$ be the smallest congruence on $Free\,(G^s_{BAB})$ that contains E^s_{BAB} and is compatible with composition. We define $C^s_{BAB} = Free\,(G^s_{BAB})\,/\equiv_E$. Then, the objects of C^s_{BAB} are the objects of G^s_{BAB}, while morphisms are equivalence classes $[u]$ of paths in $Free\,(G^s_{BAB})$. Composition is defined by $[v] \circ [u] = [v \circ u]$, which is well-defined because $\equiv_E$ is a congruence.

Equivalently, C^s_{BAB} is the category presented by $\langle G^s_{BAB} \mid E^s_{BAB}\rangle$. The set E^s_{BAB} is not obtained by mechanically translating all OWL 2 DL axioms into categorical equations. Rather, it is an explicitly selected set of schema-level coherence equations required by the transformation. OWL declarations such as owl:inverseOf, owl:FunctionalProperty, owl:InverseFunctionalProperty, and rdfs:subPropertyOf remain primarily part of the ontology reasoning layer. They may motivate selected equations in E^s_{BAB} only when the transformation explicitly treats the corresponding structural paths as equivalent.

When explicitly intended normalization equations are included in E^s_{BAB}, the faithfulness statement in Proposition 2 is understood modulo those declared normalizations. If no

such normalization equations are included, the corresponding faithfulness claim is strict on the selected schema fragment.

Construction 2. The stack-specific schema category C^s_{BAB}.

For a fixed target stack s, let $Free\,(G^s_{BAB})$ denote the free category generated by G^s_{BAB}. We define C^s_{BAB} as the quotient of $Free\,(G^s_{BAB})$ by the smallest congruence generated by E^s_{BAB}. Its objects are stack-specific realization types and realization patterns, while its morphisms are equivalence classes of finite composable paths generated by the selected structural, navigation, and refinement relations of G^s_{BAB}. Since C^s_{BAB} is determined by the presentation $\langle G^s_{BAB} \mid E^s_{BAB}\rangle$, it is a well-defined category. The construction of presented categories from generators and relations is standard and verifies the category axioms automatically.

In the case of concrete Java instance ($s = j$), O^j_{BAB} is the BABj ontology and V^j is the imported WoC vocabulary. Therefore, G^j_{BAB} is a directed multigraph whose nodes are Java realization types and realization patterns constructed from the OWL classes declared in BABj or WoC, including generic code-structure concepts such as woc:Class, woc:Interface, woc:Field, and woc:Method, BABj-specific library classes, and the generation profiles attached to mapping individuals. Its selected structural edges include the code-structure relations used by the generator, such as woc:hasField, woc:hasMethod, woc:extends, and woc:implements. The relation woc:extends is retained as the ordinary object-oriented inheritance relation. Mapping refinement, however, is not identified with woc:extends; it is represented by explicitly selected generation-refinement edges, denoted here as babj:refinedBy, oriented from the more general realization specification to the more specific one. The category C^j_{BAB} is then the presented category $\left\langle G^j_{BAB} \middle| E^j_{BAB} \right\rangle$. The remainder of this section develops the formalism for an arbitrary but fixed stack $s \in S$, while the Java instance $s = j$ is used as the running example. The defining equations of E^s_{BAB} are formulated as equations between parallel paths in the unquotiented free category $Free\,(G^s_{BAB})$. The graph-level assignment of realization patterns and generating relation labels is specified before the quotient is formed. Hence the naturality equations are well-formed at the free-category level, and the corresponding functors and natural transformations descend to C^s_{BAB} by the universal property of the quotient. Therefore, the use of the presented category C^s_{BAB} is not circular: the quotient records exactly those schema-level path identifications that the transformation explicitly requires.

In the Java instance, the BABj and WoC vocabularies declare several OWL 2 DL axioms, including inverse-property, functional, inverse-functional, and subproperty declarations. These axioms are used by the ontology reasoning layer for validation and inference. They do not automatically generate equations in C^j_{BAB}. Only those schema-level path equations that are explicitly required by the Java transformation semantics, such as conformance and refinement equations introduced below, are included in E^j_{BAB}.

Category C^s_{code} For a fixed target stack s, let G^s_{code} denote the directed multigraph whose nodes are generated source-artifact specifications for that stack and whose edges are explicit structural dependencies induced by the generated code structure. We define C^s_{code} as the free category generated by G^s_{code}; its morphisms are finite paths of structural dependencies. Because G^s_{code} depends on the output of the generator for a specific domain ontology and stack, C^s_{code} is not a fixed structure but varies with the input. An object of C^s_{code} is understood as a generated source-artifact specification, not merely as a physical

file path. Its primary operational identifier is the target source file or fully qualified name. However, in the formal category, the object is identified as a complete artifact specification: the artifact role, the fully qualified name, and the generated structural content, including superclass declarations, implemented interfaces, annotations, fields, methods, imports, and other code-level features introduced by the selected mapping strategy. Therefore, C^s_{Code} does not quotient generated code by target-language semantic equivalence. Two generated artifacts with different fully qualified names are distinct. If two realization paths attempt to produce the same fully qualified name with different structural content, this is not treated as an equality or normalization in C^s_{Code}, but as a verification error. Only a single consistent structural specification may be admitted for a given generated artifact identifier.

Model-to-Model Transformation F_s for stack s For each target technology stack s, the model-to-model (M2M) transformation is formalized as a bifunctor $F_s : C_{Core}^{NonMapping} \times C_{Map} \longrightarrow C^s_{BAB}$ that maps each pair (D, M), consisting of a domain concept D and a mapping class M, to the stack-specific realization specification $F_s(D, M)$ in C^s_{BAB}. The two arguments are independent: D determines "what" is being transformed, while M determines "how" it is realized on the target platform. Different choices of M for the same D may yield structurally distinct artifacts. For the Java instance $s = j$, the bifunctor F_j is implemented as a SPARQL-driven extraction that reads the BAB Core ontology together with its core:hasImplementationMapping bindings and constructs an in-memory DataModel representation in the vocabulary of BABj.

The construction of F_s, including the conformance functor Φ_s that captures how mapping refinements act uniformly on the domain and the proof that F_s is a well-defined bifunctor, is presented in the following subsection.

Formalization of the ImplementationMapping mechanism The technological neutrality of BAB Core rests on the core:ImplementationMapping mechanism: the explicit ontological encoding of generation rules within the model itself. This subsection provides the formal categorical basis for this mechanism, presenting a family of evaluation bifunctors $\{F_s\}_{s \in S}$, one for each supported technology stack.

We want to express code generation as a function of two independent inputs: the domain model (an object of $C_{Core}^{NonMapping}$) and the mapping class (an object of C_{Map}). Each input may also change, the domain may be refined, and the mapping class may be refined, and we want these changes to compose functorially with the generation process. Specifically, we seek a bifunctor:

$$F_s : C_{Core}^{NonMapping} \times C_{Map} \longrightarrow C^s_{BAB} \tag{3}$$

The existence of such a bifunctor, and its well-definedness on morphisms of the product category, are the content of the proposition below. The construction proceeds by introducing an intermediate object, the conformance functor Φ_s, whose functoriality captures the essential idea that refinements of the mapping class act uniformly on the domain, independently of which specific domain element is being transformed.

The Conformance Functor Φ_s For each technology stack $s \in S$, the stack-specific ontology O^s_{BAB} supplies, for every mapping class $M \in Ob(C_{Map})$, a designated named individual $\mu_s(M) \in O^s_{BAB}$ that is asserted to be an instance of M and that carries the concrete platform-specific bindings on which code generation depends, including the target type specification (core:mapsToTypeSpec), annotations, implemented interfaces, superclass declarations, and auxiliary generation parameters. We refer to $\mu_s(M)$ as the realization individual of M for stack s. For the Java instance ($s = j$), the realization assignment is exemplified by

$$\mu_j\,(\text{core:EntityMapping}) = \text{babj:DefaultEntityMapping\_J} \tag{4}$$

$$\mu_j(\text{core:ValidatableEntityMapping}) = \text{babj:DefaultValidatableEntityMapping\_J} \tag{5}$$

and analogously for the remaining mapping classes. The assignment μ_s is part of O^s_{BAB} and is fixed once the target stack is chosen. The conformance functor Φ_s defined below reads its bindings from $\mu_s(M)$, never from M itself; the schema-level category C_{Map} is thereby kept stack-independent, while all stack-specific resolution is localized in μ_s and Φ_s.

With this realization assignment fixed, we define the conformance functor as

$$\Phi_s : C_{Map} \longrightarrow [C^{NonMapping}_{Core}, C^s_{BAB}] \tag{6}$$

where $\left[C^{NonMapping}_{Core}, C^s_{BAB}\right]$ denotes the functor category whose objects are functors $C^{NonMapping}_{Core} \to C^s_{BAB}$ and whose morphisms are natural transformations between such functors.

Before defining the conformance functor, we make explicit how the stack-specific ontology O^s_{BAB} enters the construction. For a fixed stack s, ontology O^s_{BAB} together with its imported code-structure vocabulary V_s, provides: (i) realization individuals $\mu_s(M)$ for mapping classes M, (ii) stack-specific realization profiles carried by these individuals, (iii) the assignment of selected BAB Core schema relations to admissible platform-level structural relation labels, and (iv) refinement morphisms between realization profiles induced by mapping refinements. These data define the target realization signature used by the transformation. The domain concepts themselves are not defined by O^s_{BAB}; they enter only as arguments of the functors $\Phi_s(M)$.

The presented target category $C^s_{BAB} = Free\,(G^s_{BAB})\,/\,\equiv_E$ is obtained by quotienting $Free\,(G^s_{BAB})$ by the congruence generated by a relation set E^s_{BAB}. The equations in E^s_{BAB} are not arbitrary: among them are the path equalities required to make the component families assigned to mapping refinements natural transformations. Thus, the quotient construction of C^s_{BAB} is the categorical mechanism that enforces the naturality of the conformance transformation.

It is important to distinguish this formal role from empirical validation of the generator. The naturality equations are part of the categorical specification of the target presentation; they do not by themselves constitute an independent test of the implementation. The implementation must still be checked against this specification by the verification

conditions stated in Section 5.6 (mapping completeness, mapping injectivity, template determinism, quotient compatibility, no unintended collapse).

On objects of C_{Map}, for each mapping class M, the conformance functor assigns a functor

$\Phi_s(M) : C_{Core}^{NonMapping} \to C_{BAB}^{s}$.

For a fixed stack s and mapping class M, let $\mu_s(M)$ denote the realization individual declared in O_{BAB}^{s}. This individual does not denote a domain concept. Rather, it provides the stack-specific realization profile for M. This profile contains the generation-relevant bindings needed by the target stack, such as target type specifications, structural roles, metadata or annotations, inheritance or composition constraints, interface or contract declarations, template roles, and auxiliary generation parameters.

For each domain concept $D \in Ob\left(C_{Core}^{NonMapping}\right)$, the value $\Phi_s(M)(D)$ is the stack-specific realization specification obtained by applying the realization profile $\mu_s(M)$ to the domain concept D. This specification combines the identity of the domain concept D with the stack-specific bindings read from $\mu_s(M)$. Different choices of M may yield different realization specifications for the same domain concept D, which is what makes $\Phi_s(M)$ genuinely depend on M.

It remains to define $\Phi_s(M)$ on generating morphisms. Since $C_{Core}^{NonMapping}$ is freely generated by $G_{Core}^{NonMapping}$, it is sufficient to define the image of each generating edge. The stack-specific ontology O_{BAB}^{s}, together with its imported code-structure vocabulary V_s, determines a relation-label assignment

$\lambda_s : \text{Edge}\left(G_{Core}^{NonMapping}\right) \to Lab\left(G_{BAB}^{s}\right)$, where $Lab\left(G_{BAB}^{s}\right)$ denotes the set of admissible platform-level structural relation labels. The value $\lambda_s(e)$ is not itself a concrete edge; it is the platform-level relation label used in the target realization signature. Let $e : D \to D'$ be a generating edge of $G_{Core}^{NonMapping}$. We define $\Phi_s(M)(e) : \Phi_s(M)(D) \to \Phi_s(M)\left(D'\right)$ as the morphism in C_{BAB}^{s} represented by the platform-level structural relation whose label is $\lambda_s(e)$ with source $\Phi_s(M)(D)$ and target $\Phi_s(M)(D')$. The assignment λ_s is independent of M, while the source and target realization specifications depend on M through $\mu_s(M)$.

For an arbitrary morphism $f = e_n \circ \cdots \circ e_1$ in $C_{Core}^{NonMapping}$, we define

$\Phi_s(M)(f) = \Phi_s(M)\left(e_n\right) \circ \cdots \circ \Phi_s(M)\left(e_1\right)$,

and identities are sent to identities. Functoriality of $\Phi_s(M)$ follows from the universal property of the free category $C_{Core}^{NonMapping}$. The object assignment is given by $D \mapsto \Phi_s(M)(D)$, while the edge assignment is determined by λ_s on selected generating edges, with sources and targets fixed by object assignment above. Since λ_s is assumed to be defined and domain/range-compatible on the selected generating edges, this assignment extends uniquely to a functor $\Phi_s(M) : C_{Core}^{NonMapping} \to C_{BAB}^{s}$.

On morphisms of C_{Map}, we work with the morphisms generated by the rdfs:subClassOf relations declared among mapping classes in BAB Core. We refer to each such declaration as refinement morphism. For each refinement $h : M \to M'$, the conformance functor assigns a natural transformation

$$\Phi_s(h) : \Phi_s(M) \Rightarrow \Phi_s(M') \tag{7}$$

whose component at each domain concept D from $Ob\left(C_{Core}^{NonMapping}\right)$ is a morphism

$$\Phi_s(h)_D = \rho_s(D,h) : \Phi_s(M)(D) \to \Phi_s(M')(D) \tag{8}$$

in C^s_{BAB}. The morphism $\rho_s(D,h)$ represents the stack-level refinement from the realization specification obtained from M to the realization specification obtained from M', both applied to the same domain concept D. At the graph level, it is either a selected generating refinement edge of G^s_{BAB} or a finite path of selected refinement edges; in C^s_{BAB}, it is taken as the corresponding equivalence class of paths.

For every refinement $h : M \to M'$, the family $\Phi_s(h) = (\rho_s(D,h))_{D \in Ob(C^{NonMapping}_{Core})}$ is the candidate component family of the natural transformation $\Phi_s(h)$.

To be a natural transformation, this family must satisfy the naturality condition. For every generating edge
$e : D \to D'$ of $G^{NonMapping}_{Core}$, the relation set E^s_{BAB} contains the naturality equation

$$C^s_{BAB} : \rho_s(D',h) \circ \Phi_s(M)(e) = \Phi_s(M')(e) \circ \rho_s(D,h) \tag{9}$$

whenever both sides are defined as parallel paths in $Free(G^s_{BAB})$. After quotienting by $\equiv_E$, these two paths become equal in C^s_{BAB}. Since $C^{NonMapping}_{Core}$ is freely generated by $G^{NonMapping}_{Core}$, imposing these equations on generating edges is sufficient. Naturality for arbitrary morphisms follows by induction on path length and by compatibility of the congruence $\equiv_E$ with composition.

In implementation terms, the morphism $\rho_s(D,h)$ records the generation-relevant refinement from the realization profile of M to the realization profile of M'. Such a refinement may correspond to changes in metadata, structural roles, inheritance or composition constraints, interface or contract declarations, template roles, or other generation parameters. However, categorically, only the existence of the morphism $\rho_s(D,h)$ and the commutativity of the naturality square are used.

Figure 4 illustrates this naturality condition for a morphism $f : D \to D'$ in $C^{NonMapping}_{Core}$:

$$\Phi_s(h)_{D'} \circ \Phi_s(M)(f) = \Phi_s(M')(f) \circ \Phi_s(h)_D \tag{10}$$

$$\begin{array}{ccc} \Phi_s(M)(D) & \xrightarrow{\Phi_s(M)(f)} & \Phi_s(M)(D') \\ \Phi_s(h)_D \downarrow & & \downarrow \Phi_s(h)_{D'} \\ \Phi_s(M')(D) & \xrightarrow{\Phi_s(M')(f)} & \Phi_s(M')(D') \end{array}$$

Fig. 4. Naturality square for $\Phi_s(h) : \Phi_s(M) \Rightarrow \Phi_s(M')$ at the morphism $f : D \to D'$

Consider the refinement h : EntityMapping $\to$ PersistentEntityMapping in C_{Map} and the structural morphism hasAttribute : Entity $\to$ Attribute in $C^{NonMapping}_{Core}$. Under

both mapping strategies, hasAttribute is realized as woc:hasField in C^{j}_{BAB}, oriented from the realized class to the realized field, because field declarations are not altered by the addition of the @Entity annotation. The naturality square becomes:

$$\text{woc:hasField}_{persistent} \circ \text{refinedBy}_{Entity} = \text{refinedBy}_{Attribute} \circ \text{woc:hasField}_{plain} \quad (11)$$

The two sides are parallel paths in $Free\left(G^{j}_{BAB}\right)$ and are identified in C^{j}_{BAB} by the corresponding equation included in E^{j}_{BAB}. In the Java example, both paths correspond to a field declaration belonging to the generated JPA-annotated class.

Since $C^{NonMapping}_{Core}$ is freely generated by $G^{NonMapping}_{Core}$, it is sufficient to impose these equations on generating edges. Naturality for arbitrary morphisms follows by induction on path length and by compatibility of the congruence $\equiv_E$ with composition.

Two roles of the naturality square are worth noting explicitly, because they are both essential to the argument below:

- Role 1: Well-definedness of Φ_s on morphisms. The naturality square is precisely the condition that $\Phi_s(h)$ is a natural transformation in $\left[C^{NonMapping}_{Core}, C^{s}_{BAB}\right]$, which is what allows Φ_s to be defined as a functor from C_{Map} into the functor category.
- Role 2: Well-definedness of F_s on morphism pairs. A morphism pair $(f, h) : (D, M) \to (D', M')$ in the product category $C^{NonMapping}_{Core} \times C_{Map}$ admits two canonical factorizations: $(f, h) = (f, \mathrm{id}_{M'}) \circ (\mathrm{id}_D, h) = (\mathrm{id}_{D'}, h) \circ (f, \mathrm{id}_M)$. For the bifunctor F_s below to be well-defined on such morphism pairs, the two factorizations must yield the same morphism in C^{s}_{BAB}, which is exactly the naturality equation above.

Since C_{Map} is a preorder category, a refinement morphism $h : M \to M'$ is uniquely determined by its source and target mapping classes. Hence, to define Φ_s on morphisms of C_{Map}, it is enough to specify, for each comparable pair $M \leq M'$, the component family $\rho_s(D; M, M') : \Phi_s(M)(D) \to \Phi_s(M')(D)$ in C^{s}_{BAB}, natural in D. We write this family as $\Phi_s(h)_D = \rho_s(D; M, M')$. To make Φ_s a functor, these component families must satisfy the following coherence conditions:

- Identity: For $M = M'$, the refinement is the identity morphism $\mathrm{id}_M : M \to M$. Its component is defined as $\rho_s(D; M, M) = \mathrm{id}_{\Phi_s(M)(D)}$ for every D. Hence $\Phi_s(\mathrm{id}_M) = \mathrm{id}_{\Phi_s(M)}$.
- Composition: For refinements $h : M \to M'$ and $k : M' \to M''$, the composite $k \circ h$ is the unique refinement morphism $M \to M''$ in the preorder category C_{Map}. Its component is required to satisfy $\rho_s(D; M, M'') = \rho_s(D; M', M'') \circ \rho_s(D; M, M')$ in C^{s}_{BAB}; then $\Phi_s(k \circ h)_D = \Phi_s(k)_D \circ \Phi_s(h)_D$ and $\Phi_s(k \circ h) = \Phi_s(k) \circ \Phi_s(h)$.

Hence Φ_s is a functor $C_{Map} \to [C^{NonMapping}_{Core}, C^{s}_{BAB}]$.

The evaluation bifunctor F_s We now define the family of evaluation bifunctors $\{F_s\}_{s \in S}$, with each member:

$$F_s : C^{NonMapping}_{Core} \times C_{Map} \longrightarrow C^{s}_{BAB} \quad (12)$$

On objects, for a pair (D, M):

$$F_s(D, M) = \Phi_s(M)(D) \tag{13}$$

On morphism pairs $(f, h) : (D, M) \to (D', M')$:

$$F_s(f, h) = \Phi_s(h)_{D'} \circ \Phi_s(M)(f) \tag{14}$$

The cross-subcategory morphism core:hasImplementationMapping, which exists in C_{Core} but in neither $C_{Core}^{NonMapping}$ nor C_{Map}, does not belong to the domain of F_s. It plays the semantic role of specifying which pairs $\left(D, M\right) \in Ob(C_{Core}^{NonMapping} \times C_{Map}\right)$ carry intended architectural meaning: each domain element D is linked to one or more ImplementationMapping elements $M_1, M_2, \ldots$ via core:hasImplementationMapping, each governing a distinct aspect of D's platform realization (persistence, DTO projection for a specific view, UI binding, etc.). The bifunctor F_s is evaluated independently for each such pair (D, M_i), producing the corresponding platform-specific artifact. Formally, F_s is total on the full product category; architecturally, its semantic content is concentrated on pairs linked by hasImplementationMapping.

Proposition 1. For each $s \in S$, the mapping F_s defined above is a functor, hence a bifunctor in its two arguments.

Proof. Since $C_{Core}^{NonMapping}$, C_{Map}, and C_{BAB}^{s} are small categories (in the implemented transformation, they are generated from finite ontology fragments selected for generation: finite sets of OWL classes, object properties, mapping classes, mapping individuals, and stack-specific structural relations) the construction takes place in $\mathbf{Cat}$, the category of small categories and functors between them. The category $\mathbf{Cat}$ is Cartesian closed [47], with exponentials given by functor categories. Hence, for small categories C, M, and B, morphisms $C \times M \to B$ correspond naturally to morphisms $M \to [C, B]$, where $[C, B]$ denotes the functor category from C to B.

In the present setting, this correspondence is realized by the evaluation functor

$$ev_s : \left[C_{Core}^{NonMapping}, C_{BAB}^{s}\right] \times C_{Core}^{NonMapping} \longrightarrow C_{BAB}^{s}. \tag{15}$$

On objects, it is defined by $ev_s(G, D) = G(D)$, where $G : C_{Core}^{NonMapping} \to C_{BAB}^{s}$ is a functor and D is an object of $C_{Core}^{NonMapping}$. On morphisms, for a natural transformation $\eta : G \Rightarrow G^{'}$ and a morphism $f : D \to D'$ in $C_{Core}^{NonMapping}$, the evaluation functor is defined by $ev_s(\eta, f) = \eta_{D'} \circ G(f)$.

Equivalently, by the naturality of η, we have $\eta_{D'} \circ G(f) = G^{'}(f) \circ \eta_D$. Therefore, $ev_s(\eta, f)$ is a well-defined morphism $G(D) \to G^{'}(D')$ in C_{BAB}^{s}.

We now express the bifunctor F_s through the conformance functor $\Phi_s : C_{Map} \to [C_{Core}^{NonMapping}, C_{BAB}^{s}]$. Let $\pi_1 : C_{Core}^{NonMapping} \times C_{Map} \to C_{Core}^{NonMapping}$ and $\pi_2 : C_{Core}^{NonMapping} \times C_{Map} \to C_{Map}$ be the two projection functors. Then F_s factorizes as $F_s = ev_s \circ \langle \Phi_s \circ \pi_2, \pi_1 \rangle$.

Indeed, on objects (D, M), we obtain

$$F_s(D, M) = ev_s\left((\Phi_s \circ \pi_2)(D, M), \pi_1(D, M)\right) = ev_s\left(\Phi_s(M), D\right) = \Phi_s(M)(D). \tag{16}$$

Thus, $F_s(D, M)$ is precisely the platform-specific realization of the domain concept D under the mapping strategy M.

On morphism pairs $(f, h) : (D, M) \to (D', M')$ in the product category $C_{Core}^{NonMapping} \times C_{Map}$ where $f : D \to D'$ is a morphism in $C_{Core}^{NonMapping}$, and $h : M \to M'$ is a refinement morphism in C_{Map}, we have $(\Phi_s \circ \pi_2)(f, h) = \Phi_s(h)$ and $\pi_1(f, h) = f$. Therefore, $F_s(f, h) = ev_s\,(\Phi_s(h), f)$.

Since $\Phi_s(h) : \Phi_s(M) \Rightarrow \Phi_s\,(M')$ is a natural transformation, the evaluation functor gives:

$$F_s(f, h) = (\Phi_s(h))_{D'} \circ \Phi_s(M)(f). \tag{17}$$

By the naturality of $\Phi_s(h)$, the following equality holds: $(\Phi_s(h))_{D'} \circ \Phi_s(M)(f) = \Phi_s\,(M')\,(f) \circ (\Phi_s(h))_D$. Hence, the two canonical ways of evaluating the morphism pair (f, h) give the same morphism in C_{BAB}^s. This is precisely the naturality condition established in the construction of the conformance functor.

Consequently, $F_s(f, h)$ is a well-defined morphism $F_s(D, M) \to F_s\,(D', M')$ in C_{BAB}^s.

Since Φ_s is a functor, the pairing $\langle \Phi_s \circ \pi_2, \pi_1 \rangle$ is a functor from $C_{Core}^{NonMapping} \times C_{Map}$ to $[C_{Core}^{NonMapping}, C_{BAB}^s] \times C_{Core}^{NonMapping}$.

The evaluation functor ev_s is a functor by the Cartesian closed structure of $\mathbf{Cat}$. Therefore, their composition $F_s = ev_s \circ \langle \Phi_s \circ \pi_2, \pi_1 \rangle$ is a functor, where π_1 and π_2 denote the canonical projection functors from the product category, and $\langle \cdot, \cdot \rangle$ denotes the pairing (universal property of the product). Since the domain of F_s is the product category $C_{Core}^{NonMapping} \times C_{Map}$, the functor F_s is a bifunctor in its two arguments. This proves that the parametrized model-to-model transformation is functorial both in the domain model and in the mapping class. ■

Architectural implication This formalization ensures that multi-target code generation is achieved by varying only the stack index $s \in S$: the source domain category $C_{Core}^{NonMapping}$ and the mapping-refinement category C_{Map} remain unchanged, while the conformance functor Φ_s handles all stack-specific resolution. In the notation of Section 5.4.2, the functor F_s is induced by the Φ_s through evaluation functor ev_s for a fixed stack s, restricted to the mapping pairs selected by core:hasImplementationMapping. In the Java instance, this functor is denoted F_j. Adding support for a new platform requires only the definition of a new conformance functor $\Phi_{s'}$ and its associated target category $C_{BAB}^{s'}$, without any modification to the domain model or the generation architecture.

Functor $H_s : C_{BAB}^s \to C_{code}^s$ (Model-to-Code Transformation for stack s) For each target stack s, the second functor, H_s, represents the model-to-code (M2T) transformation: it takes the platform-specific schema representation in C_{BAB}^s and produces source files in C_{code}^s.

On objects, H_s sends each stack-specific realization specification produced by F_s to the source file or file fragment that realizes it . For example, H_s sends the representation of a class to a source file whose content is the textual rendering of that class by the corresponding FreeMarker template. The input object of H_s is a stack-specific realization specification, not merely an undecorated platform class. Therefore, H_s has access to all generation-relevant information carried by the selected mapping strategy: annotations,

implemented interfaces, superclass declarations, imports, fields, methods, template roles, and other generation parameters. This prevents mapping-specific information from being lost between the model-to-model and model-to-code phases.

On morphisms, H_s sends each structural relation in C^s_{BAB} to the corresponding textual relation in C^s_{code}.

Since C^s_{BAB} is a presented category, H_s is defined in two steps. First, a graph-level assignment is specified on the generators of G^s_{BAB}, mapping each platform-level structural relation to the corresponding code-level structural dependency. By the universal property of the free category, this assignment extends uniquely to a functor

$$\widetilde{H_s} : Free\left(G^s_{BAB}\right) \to C^s_{code}. \tag{18}$$

Second, because $C^s_{BAB} = Free\left(G^s_{BAB}\right) / \equiv_E$, the functor $\widetilde{H_s}$ induces a well-defined functor on the quotient only if it is constant on $\equiv_E$-equivalence classes. Therefore, whenever $u \equiv_E v$, it must hold that $\widetilde{H_s}(u) = \widetilde{H_s}(v)$ as the same structural dependency path in the free category C^s_{Code}. This means that platform-level paths identified by the defining equations of C^s_{BAB} must be rendered by the generator as the same code-level structural dependency path. This quotient-compatibility condition is required as part of transformation verification. Under this condition, $\widetilde{H_s}$ factors uniquely through the quotient, yielding a well-defined functor

$$H_s : C^s_{BAB} \to C^s_{code} defined by H_s\left([u]_{\equiv_E}\right) = \widetilde{H_s}(u). \tag{19}$$

In the Java implementation, H_j is realized by the FreeMarker template mechanism, which traverses the in-memory BABj model and emits source-code artifacts guided by text templates specific to Java target language construct.

5.5. Structural Preservation and Faithfulness of the Transformation Pipeline

Having defined the full transformation pipeline $T_s = H_s \circ F_s$, we now state the central information-preservation result of the methodology. The proposition below captures the formal sense in which the transformation pipeline preserves the structural content of the domain model. It should be read as a conditional schema-level preservation theorem. It states what follows if the transformation artifacts satisfy the stated verification assumptions; it is not a claim that the categorical presentation alone proves correctness of an arbitrary generator implementation.

Proposition 2. (Conditional Structural Preservation). Let F_s and H_s be the model-to-model bifunctor and model-to-code functor, defined in Sections 5.4.3 and 5.4.4 respectively, for a fixed target stack s, and let $T_s = H_s \circ F_s$ denote their composition, evaluated on the mapping pairs selected by core:hasImplementationMapping. Assume that the following verification conditions hold: (i) the graph-level assignment underlying F_s is injective on the selected source morphism pairs, except for identifications induced by explicitly declared normalization equations in E^s_{BAB}, and the congruence $\equiv_E$ introduces no further identifications among their F_s-images; (ii) the graph-level assignment underlying H_s is constant on $\equiv_E$ equivalence classes, so that H_s is well-defined on C^s_{BAB}, and H_s is injective on the relevant target equivalence classes; (iii) the naming scheme is injective on the considered domain classes; and (iv) distinct selected mapping strategies

induce distinct generation profiles, so that the code-artifact identity distinguishes them by their generated structural content. Then T_s satisfies the following properties:

a) Totality. For every selected object pair $(D, M) \in Ob\left(C_{Core}^{NonMapping} \times C_{Map}\right)$, there exists a unique generated source-artifact specification $T_s(D, M) \in Ob\left(C_{code}^{s}\right)$.
b) Structural coherence. For every selected morphism pair $(f, h) : (D, M) \to (D', M')$ in $C_{Core}^{NonMapping} \times C_{Map}$, there exists a unique morphism $T_s(f, h) : T_s(D, M) \to T_s\left(D', M'\right)$ in C_{code}^{s}. Moreover, the composite functor preserves identities and composition: $T_s\left(\mathrm{id}_D, \mathrm{id}_M\right) = \mathrm{id}_{T_s(D,M)}$ and, for composable morphism pairs $(f, h) : (D, M) \to (D', M')$ and $(f', h') : (D', M') \to (D'', M'')$ it holds that $T_s\left((f', h') \circ (f, h)\right) = T_s\left(f', h'\right) \circ T_s(f, h)$. Equivalently, since composition in the product category is componentwise, $T_s\left(f' \circ f, h' \circ h\right) = T_s\left(f', h'\right) \circ T_s(f, h)$, where equality is understood in the corresponding quotient/presented target category.
c) Faithfulness modulo explicitly intended normalizations. If the morphism pairs

$$(f, h), (g, k) \in Hom_{C_{Core}^{NonMapping} \times C_{Map}}\left((D, M), (D', M')\right) \tag{20}$$

have F_s -images that are not identified by the explicitly intended normalization equations included in E_{BAB}^{s}, then

$$\left(T_s(f, h) \neq T_s(g, k) in Hom_{C_{Code}^{s}}\left(T_s(D, M), T_s(D', M')\right).\right) \tag{21}$$

If no explicitly intended normalization equations are included, this reduces to strict faithfulness on the selected schema fragment.

d) Injectivity on objects. If $(D, M) \neq (D', M')$ in $Ob\left(C_{Core}^{NonMapping} \times C_{Map}\right)$, then $T_s(D, M) \neq T_s\left(D', M'\right)$ in $Ob\left(C_{code}^{s}\right)$.

Proof. (a) and (b) follow directly from the definition of functor composition: the composition of two functors is again a functor, and functors are by definition total on objects and morphisms and preserve the category axioms. This establishes totality and structural coherence without further assumption.

(c) Faithfulness requires that distinct morphism pairs in $C_{Core}^{NonMapping} \times C_{Map}$ remain distinct after applying T_s. This does not follow merely from the construction of C_{BAB}^{s}, because C_{BAB}^{s} is a quotient of $Free\left(G_{BAB}^{s}\right)$ and the congruence $\equiv_E$ may identify distinct target paths. Therefore, faithfulness is obtained under the stated separation assumptions: F_s must map distinct source morphism pairs to distinct equivalence classes in C_{BAB}^{s}, except for explicitly intended normalization, and H_s must preserve the distinction between the relevant equivalence classes at the code level. Under these assumptions, if $(f, h) \neq (g, k)$ in $Hom_{C_{Core}^{NonMapping} \times C_{Map}}\left((D, M), (D', M')\right)$, then $F_s(f, h) \neq F_s(g, k)$ in C_{BAB}^{s}. Since H_s is injective on the relevant target equivalence classes, distinct F_s-images remain distinct after applying H_s. Consequently, if two selected morphisms are distinct in the considered finite schema fragment and are not identified by an explicitly intended normalization equation $H_s\left(F_s(f, h)\right) \neq H_s\left(F_s(g, k)\right)$, their T_s-images are distinct $T_s(f, h) \neq T_s(g, k)$ in C_{code}^{s}. Hence T_s is faithful on the selected schema fragment.

(d) Injectivity on objects follows from the namespace-aware IRI-based naming scheme together with the mapping-sensitive definition of code-artifact identity. The object $T_s(D, M)$ is not identified merely by the physical file path, but by the generated source-artifact specification, including the target artifact role, generated structural content, superclass declarations, implemented interfaces, annotations, fields, methods, imports, and template roles. Therefore, two selected mapping strategies that induce distinct generation profiles yield distinct generated source-artifact specifications. If two mapping declarations induce the same generation profile, they are treated as equivalent for the purposes of this proposition rather than as distinct preserved inputs. If $D \neq D'$, the generated identifiers are distinct by the naming scheme. If $D = D'$ but $M \neq M'$, then the two mapping strategies determine different generated structural content, such as different implemented interfaces, annotations, superclass declarations, fields, methods, or template roles. Hence, they define distinct generated source-artifact specifications in C^s_{code}. Therefore, if the selected object pairs (D, M) and (D', M') are distinct and are not identified by generation-profile normalization, then their images $T_s(D, M)$ and $T_s\,(D', M')$ are distinct objects in C^s_{code}. ■

Corollary (Schema-level information preservation). *Under the assumptions of Proposition 2, every mapped domain class and every mapped structural relation in the considered schema fragment, together with its selected mapping strategy, has a corresponding schema-level generated source-artifact specification in C^s_{code}. Moreover, the correspondence is injective on the selected schema fragment: distinct object pairs (D, M) in $C_{Core}^{NonMapping} \times C_{Map}$ that are not identified by generation-profile normalization produce distinct code-level artifact specifications, and distinct morphism pairs (f, h) not identified by explicitly intended normalization produce distinct morphisms in C^s_{code}. Consequently, under the stated verification assumptions, the schema-level distinctions present in the mapped domain model and in its selected mapping refinements are neither lost nor conflated during transformation.*

Corollary (Image subcategory). *Under the assumptions of Proposition 2, image of Ts forms a subcategory $Im(Ts) \subseteq C^s_{Code}$. After quotienting the selected object and morphism pairs determined by $core : hasImplementationMapping$ by the explicitly intended normalization equations, the induced restriction of T_s is an isomorphism onto this image. Hence, the generated code category contains a subcategory isomorphic to the selected mapped schema fragment, not necessarily to the entire $C_{Core}^{NonMapping} \times C_{Map}$.*

The proposition above asserts conditional structural preservation: under the assumptions stated in Proposition 2, distinctions between object pairs and morphism pairs in $C_{Core}^{NonMapping} \times C_{Map}$ are preserved by the composite transformation T_s into C^s_{code}. It does not assert:

- *Instance-level preservation.* We do not formally verify that data populating the domain model (A-box individuals) is preserved in any operational sense by the generated code's runtime behavior. The methodology is a schema-level (T-box) transformation; instance-level consistency and completeness are separate concerns handled by the ontology validation layer, combining OWL reasoning with structural SPARQL/SHACL-style checks before transformation.
- *Semantic equivalence.* We do not assert that the generated code "means the same thing" as the domain ontology in an operational or denotational sense. Such a claim would require a formal operational semantics of both the ontology and the target

language, together with a semantics-preserving morphism between them, which is beyond the scope of the present work.
- *Round-trip equivalence*. We do not assert that the domain ontology can be recovered from the generated code. The composite functor $T_s = H_s \circ F_s$ is not claimed to have a left or right adjoint. Such a construction, which would underpin formal round-trip engineering, is identified as a direction for future work.

The scope of Proposition 2 is precisely the schema-level guarantee that a rigorous model-driven transformation can offer: under the stated verification conditions, the structure of the mapped domain model and its selected mapping refinements is neither lost nor conflated as it passes through the categorical pipeline to generated code.

5.6. Model validation and transformation verification

The methodology distinguishes two complementary correctness mechanisms, operating at different stages of the code generation pipeline and relying on different formal bases.

Model validation takes place before the transformation begins and is the responsibility of the ontology reasoning layer. The validation layer, combining OWL reasoning with structural SPARQL/SHACL-style checks, verifies that a domain ontology, constructed by importing and specializing BAB Core, is a well-formed description of the intended PAIS system. This includes logical checks handled by the OWL reasoner, as well as completeness checks handled by structural SPARQL/SHACL-style constraints, such as verifying cardinality constraints (e.g., that each class of type core:Entity declares at least one core:hasAttribute relation), enforcing structural obligations (e.g., that every core:Process contains both a core:StartTask and an core:EndTask), and ensuring that disjointness constraints are respected (e.g., that pms:Shipment and pms:Customer are not inadvertently identified). The axioms declared in BAB Core, including inverse property declarations, functional property constraints, and subclass hierarchies, constitute the logical part of the validation criteria. Completeness conditions that depend on the explicit presence of modeling elements are checked by SPARQL/SHACL-style structural constraints.

Model validation, as described above, is an ontological property: it asserts that the domain ontology is a structurally and semantically coherent model of the target system. It is not, and does not need to be, a categorical property. The categorical formalism developed in this section operates at the schema level and assumes that the input ontology has already passed reasoner validation. In this sense, the reasoning layer and the categorical layer are cleanly separated: the reasoner ensures that the input to the transformation is well-formed, while the categorical formalism ensures that the transformation itself preserves structure.

Transformation verification concerns the behavior of the generator, not the input ontology. It asks whether the bifunctor $F_s : C_{Core}^{NonMapping} \times C_{Map} \to C_{BAB}^s$ and functor $H_s : C_{BAB}^s \to C_{code}^s$, and in particular their Java instances F_j and H_j, as implemented by the SPARQL-driven extraction and the FreeMarker template mechanism, truly satisfy the properties claimed of them in Sections 5.4.2 and 5.4.4. The bifunctor F_s is obtained from the conformance functor construction of Section 5.4.3: each $\Phi_s(M)$ is determined by the universal property of the free source category $C_{Core}^{NonMapping}$, and F_s is then obtained by evaluation. The functor H_s is obtained by first extending a graph-level assignment from G_{BAB}^s to $Free\,(G_{BAB}^s)$, and then verifying that this extension $\widetilde{H_s}$ respects the congruence $\equiv_E$ so that it factors through the quotient C_{BAB}^s. The functoriality of

F_s follows from the functoriality of Φ_s and the evaluation functor, as proved in Proposition 1. The functoriality of H_s follows after the graph-level extension $\widetilde{H_s}$ is verified to respect the quotient congruence $\equiv_E$. Thus, preservation of identities and composition is guaranteed by construction once these specification-level conditions are satisfied. The structural preservation properties proved in Proposition 2 (totality, structural coherence, faithfulness, and injectivity on objects) similarly follow formally, conditional on the implementation respecting the specified mapping rules.

In practice, engineering verification consists of checking that the implementation does not deviate from the specification. Specifically, five checks are performed:

- Mapping completeness: every selected generating edge of $G_{Core}^{NonMapping}$ that appears in the domain ontology must be in the domain of the property-label assignment λ_s. Equivalently, every selected BAB Core schema relation used by the transformation has a corresponding platform-level structural relation label in G_{BAB}^{s}. A SPARQL query over BAB Core and the stack-specific BABj ontology reports any unmapped properties before generation begins.
- Mapping injectivity: no two distinct selected BAB Core schema relations are mapped by λ_s to the same platform-level structural relation label unless this identification is explicitly intended by the normalization equations in E_{BAB}^{s}. This is a precondition for Proposition 2's faithfulness claim and is verified statically by inspection of the stack-specific ontology and transformation specification.
- Template determinism: each FreeMarker template produces a deterministic output for a given BABj element, with no nondeterministic choices or dependency on external state. This is verified by code review of the template library and by snapshot testing against a fixed corpus of domain ontologies.
- Quotient compatibility: every equation included in E_{BAB}^{s} must be preserved by the model-to-code functor H_s. Concretely, if two parallel paths are identified by the congruence $\equiv_E$, then their template-level interpretations must produce the same code-level structural dependency path in C_{Code}^{s}. Operationally, this means that the FreeMarker templates must not distinguish between two paths that the presented category C_{BAB}^{s} identifies. If two different realization paths attempt to generate the same artifact identifier with different structural content, this is treated as a naming or generation-profile conflict and the verification fails. It is not treated as a code-level normalization. This condition ensures that H_s is well-defined on equivalence classes of paths in C_{BAB}^{s}.
- No unintended collapse: the congruence $\equiv_E$ must not identify the images under F_s of distinct selected source morphisms, except where the identification is explicitly declared as an intended normalization equation in E_{BAB}^{s}. This condition is required for the faithfulness-modulo normalization part of the structural preservation result.

When all five checks pass, the engineering implementation conforms to the categorical specification, and the guarantees established in Proposition 2 apply to the generated code.

The entire transformation can be summarized as the following pipeline:

$$C_{Core}^{NonMapping} \times C_{Map} \to C_{BAB}^{s} \xrightarrow{H_s} C_{code}^{s} \tag{22}$$

where each arrow is a functor with a precise categorical interpretation, and the structural preservation of the composition T_s is established by Proposition 2. In this sense, the

methodology transforms ontology-driven code generation from a heuristic engineering process into a formally specified transformation with provable schema-level correctness.

6. Software implementation

The transformer prototype is realized in the Java programming language and represents a concrete implementation of the transformation chain $H_j \circ F_j$ for the Java technology stack (denoted $s = j$ throughout this section). It integrates libraries for working with ontologies (Apache Jena) and a text generation mechanism (FreeMarker).

The transformation process unfolds through clearly defined phases corresponding to functors F_j and H_j (Figure 5).

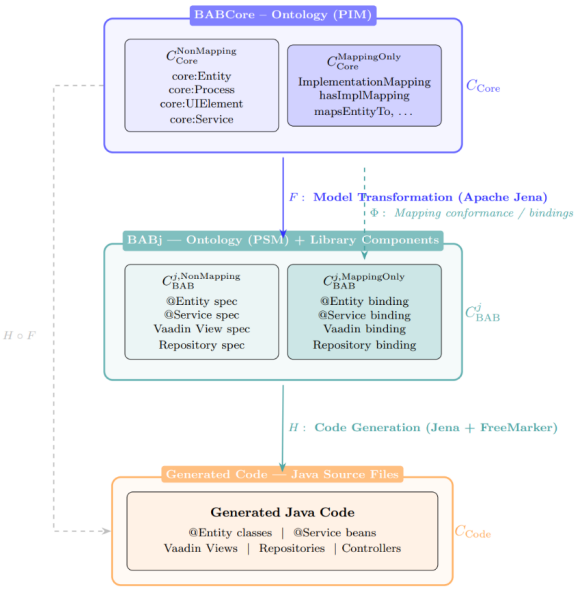


Fig. 5. Architecture of the implementation solution

6.1. Implementation of Bifunctor F_j – Model-to-Model (M2M) Transformation

The first phase, corresponding to the bifunctor $F_j : C_{Core}^{NonMapping} \times C_{Map} \rightarrow C_{BAB}^{j}$, is the Model-to-Model transformation, and this logic is executed by the model transformer.

The domain engineer defines a concrete system model as a specific instance of the ontology. Practically, this means describing the specific application by adding new classes or individuals to the ontology representing entities, processes, modules, etc. For example, for a Postal Management System (PMS), the engineer would add a `pms:Shipment` class as a subclass of `core:Entity`, a `pms:Courier` class also as a subclass of `core:Entity`, a `pms:PackageDeliveryProcess` process as an instance of

core:Process with appropriate states (e.g., 'In Transit', 'Delivered') and tasks (e.g., 'ScanPackage', 'DeliverPackage'). All these added elements and their connections constitute one application model at the PIM level.

It is assumed that a target stack (e.g., Java) is selected as a configuration input to the transformer, so the corresponding PSM ontology (e.g., BABj) is loaded alongside the domain model. The BAB Core ontology enables a hasImplementationMapping link between every model element and an object carrying implementation details. In practice, an instance of ImplementationMapping is defined in the stack-specific ontology for every relevant domain concept or set of concepts. For example, for the Java stack, the BABj ontology declares generic mapping individuals, one per architectural aspect, that collectively specify how domain elements are realized. A typical setup includes babj:DefaultEntityMapping_J, an instance of core:EntityMapping, which declares via core:mapsToTypeSpec that domain entities are realized as Java classes extending babj:AbstractEntity with the @Entity annotation; babj:DefaultAttributeMapping_J, an instance of core:AttributeMapping, specifying that domain attributes are realized as Java fields with @Column annotations; and similar default mapping individuals for processes, tasks, relations, and UI elements. A domain class such as pms:Shipment connects to its mapping individuals through core:hasImplementationMapping, which is a multi-valued relation: Shipment is linked simultaneously to DefaultEntityMapping_J (governing its class-level realization) and to the relevant DefaultAttributeMapping_J (governing its field realizations), and potentially to additional mapping individuals governing other aspects of its platform presence.

More refined mapping strategies are expressed through subclasses of the generic mapping classes. For instance, core:ValidatableEntityMapping may be declared as a subclass of core:EntityMapping, and a domain class that requires validation logic can reference an individual of this refined class instead of the default one. Refinement thus acts pointwise on individual aspects rather than on entire domain elements, and multiple aspects of the same domain class may use independently chosen refinement levels. For simplicity, the prototype defaults to using the generic mapping individuals for all non-refined domain elements, applying class-specific refinement only where the ontology explicitly declares it.

The implementation operates at the instance level by traversing ABox assertions of core:hasImplementationMapping, retrieving named mapping individuals such as babj:DefaultValidatableEntityMapping_J. Each retrieved individual realizes a mapping class $M \in \mathrm{Ob}\,(C_{Map})$ in the sense of the realization assignment μ_j introduced in Section 5.4.3: the mapping class is recovered as the asserted rdf:type of the retrieved individual. The formal pair (D, M) on which the bifunctor F_j is defined therefore corresponds operationally to the pair $(D, \mu_j(M))$, with M recovered from the individual's type. The remainder of this section uses individual names (e.g., DefaultEntityMapping_J) when describing operational behavior and the corresponding mapping classes (e.g., core:EntityMapping) when referring to the formal indexing.

The SPARQL-driven extraction mechanism, defined in the SparqlQueries.java class, implements functor F_j by executing a series of queries over the loaded do-

main and BABj ontologies. The queries first collect all domain classes. For each such class D, the queries then retrieve every mapping individual linked to D via `core:hasImplementationMapping`, following the multi-valued relation so that all aspect-specific mapping individuals (class-level, attribute-level, and any further ones) are gathered in one pass. For each retrieved mapping individual, the queries read its structural bindings, specifically `core:mapsToTypeSpec` (pointing to the BABj library class that realizes the mapping) and any auxiliary annotation or interface-realization properties declared in the BABj ontology and accumulate this information into the DataModel representation used in the next phase.

The following simplified Java fragment (Listing 1) illustrates the operational core of this extraction step. Prefixes are resolved from the ontology vocabulary constants used by the transformer.

```
public DataModel extractClassModel(Resource domainClass) {
  DataModel model = new DataModel();
  model.setClassName(localName(domainClass));

  String sparql = """
    SELECT ?mapping ?mappingClass ?targetType
    WHERE {
      ?domain core:hasImplementationMapping ?mapping .
      ?mapping rdf:type ?mappingClass .
      ?mappingClass rdfs:subClassOf* core:ImplementationMapping .
      OPTIONAL { ?mapping core:mapsToTypeSpec ?targetType . }
    }
    """;

  ParameterizedSparqlString q = new ParameterizedSparqlString(sparql);
  q.setNsPrefix("core", CoreVocabulary.NS);
  q.setNsPrefix("rdf", RDF.uri);
  q.setNsPrefix("rdfs", RDFS.uri);
  q.setParam("domain", domainClass);

  try (QueryExecution qe =
      QueryExecutionFactory.create(q.asQuery(), ontologyModel)) {
    ResultSet rs = qe.execSelect();
    while (rs.hasNext()) {
      QuerySolution row = rs.next();
      Resource mappingClass = row.getResource("mappingClass");
      model.addMappingClass(localName(mappingClass));
      if (row.contains("targetType")) {
        model.setExtendsClass(
            resolveTypeSpec(row.getResource("targetType")));
      }
    }
  }

  readAttributes(domainClass).forEach(model::addAttribute);
  return model;
}
```

Listing 1 - Simplified Java/SPARQL fragment implementing the M2M extraction step for the Java stack

The fragment is intentionally simplified. The actual transformer executes additional queries for annotations, interfaces, attributes, relations, and behavioral elements. Nevertheless, it shows the essential mechanism of F_j: the domain class determines what is transformed, while the selected mapping individual and its recovered mapping class determine how the element is realized in the Java stack.

The results of the SPARQL queries are parsed and used to populate internal memory objects, defined by the `DataModel.java` class. The DataModel object is an in-memory representation of the realization specification constructed for a domain class D from the mapping individuals linked to D via `core:hasImplementationMapping`. Operationally, each such mapping individual contributes one generation-relevant aspect of the realization specification: class-level annotations and superclass from `core:EntityMapping`, field-level annotations from `core:AttributeMapping`, interface bindings from refinements such as `core:ValidatableEntityMapping`, and so on. The DataModel records the union of these implementation-level aspects: the target library class, annotations, implemented interfaces, fields, and structural bindings. This aggregation is an implementation-level construction used to form a single realization specification for D; it is not used as a categorical coproduct in C^j_{BAB}.

This step constitutes the essence of the M2M transformation. Operationally, it transforms the schema-level graph G_{Core}, extracted from the BAB Core RDF/OWL ontology, into a stack-specific, yet still textually neutral, PSM DataModel object. Categorically, this object represents the corresponding fragment of C^j_{BAB}.

6.2. Implementation of Functor H_j – Model-to-Code (M2T) Transformation

The second phase, corresponding to functor $H_j : C^j_{BAB} \to C^j_{code}$ is the Model-to-Code transformation. `CodeGeneration.java` passes the populated DataModel object (PSM) to the FreeMarker template engine. Templates such as `EntityTemplate.ftl` and `SessionTemplate.ftl` implement functor H_j. These templates iterate through the DataModel and generate textual output. For example, `EntityTemplate.ftl` contains logic generating Java code, Listing 2.

```
package ${dataModel.packageName};

<#- Imports generated from class-level annotations ->
<#list dataModel.annotations as annotation>
  import ${annotation.fullName};
</#list>

<#- Imports generated from attribute-level annotations ->
<#list dataModel.attributes as attr>
<#if attr.annotation?? && attr.annotation.fullName??>
  import ${attr.annotation.fullName};
</#if>
</#list>
  import ${dataModel.extendsClass.fullName};

<#- Loop adding imports for all interfaces ->
<#list dataModel.interfaces as interface>
  import ${interface.fullName};
</#list>

<#- Class-level annotations generated from the selected mapping profile ->
<#list dataModel.annotations as annotation>
  ${annotation.expression}
</#list>
public class ${dataModel.className} extends ${dataModel.extendsClass.name}
  <#- Logic for adding 'implements' clause ->
  <#if dataModel.interfaces?has_content>
  implements <#list dataModel.interfaces as interface>${interface.name}<#sep>, </#list>
  </#if>
{
  <#- Loop through attributes ->
  <#list dataModel.attributes as attr>
  @${attr.annotation.name}<#if attr.annotation.parameters?? &&
  attr.annotation.parameters?has_content>(${attr.annotation.parameters})</#if>
  private ${attr.type} ${attr.name};

  // ... generated getters and setters ...
  </#list>

  <#- Loop generating 'hook' methods for each behavior ->
  <#if dataModel.beforeValidate >
```

```
  @Override
  public void beforeValidate() {
    // TODO: Programmer adds logic
  }
  </#if>
}
```

Listing 2 - Fragment of FreeMarker template (EntityTemplate.ftl) implementing functor H_j

The output of this M2T transformation is a Java file, a specific object in category C^j_{code}.

6.3. Transformation Example: From BAB Core to Shipment.java

To illustrate the entire process, let us consider the transformation of the Shipment entity from the Postal Management System (PMS) domain

PIM Level (C_{Core}) – Definition in Ontology The domain expert defines Shipment together with the mapping strategy governing its platform realization (Listing 3). The PIM ontology does not directly contain Java annotations; instead, it selects mapping individuals whose stack-specific generation profiles later provide annotations, imports, superclass bindings, implemented interfaces, and template hooks. The BAB Core ontology declares `core:ValidatableEntityMapping` as a subclass of `core:EntityMapping`, specializing the generic entity mapping strategy for entities that require validation logic. The BABj ontology provides `babj:DefaultEntityMapping_J` as a default instance of `core:EntityMapping` for non-validatable entities, and `babj:DefaultValidatableEntityMapping_J` as an instance of `core:ValidatableEntityMapping` for the validatable case; the latter carries the binding `core:mapsToTypeSpec` pointing to `babj:AbstractEntity`, together with a specification of the `CheckValidateEntityInterface` realization. Attribute-level realization is governed separately by `babj:DefaultAttributeMapping_J`, an instance of `core:AttributeMapping`. The listing uses the class IRI as the attachment point for generation metadata; operationally, the transformer treats these triples as mapping-selection assertions used by the SPARQL extraction layer.

```
<owl:Class rdf:about="&pms;Shipment">
  <rdfs:subClassOf rdf:resource="&core;Entity"/>
  <rdfs:label>Shipment</rdfs:label>
  <core:hasImplementationMapping
      rdf:resource="&babj;DefaultValidatableEntityMapping_J"/>
  <core:hasImplementationMapping
      rdf:resource="&babj;DefaultAttributeMapping_J"/>
</owl:Class>

<owl:DatatypeProperty rdf:about="&pms;trackingId">
  <rdfs:domain rdf:resource="&pms;Shipment"/>
  <rdfs:range rdf:resource="&xsd;string"/>
</owl:DatatypeProperty>
```

Listing 3 - RDF/XML definition for pms:Shipment

PSM Level (C^j_{BAB}) – Internal DataModel The SPARQL-driven extraction (implementing bifunctor F_j) reads the RDF graph and creates the following DataModel object in the transformer's memory:

```
className: "Shipment"
annotations: [
  {name: "Entity", fullName: "javax.persistence.Entity",
   expression: "@Entity"},
  {name: "Table", fullName: "javax.persistence.Table",
   expression: "@Table(name = \"Shipment\")"}
]
extendsClass: {name: "AbstractEntity", fullName:
  "rs.co.bora5.programs.bab.model.AbstractEntity"}
interfaces: [
  {name: "CheckValidateEntityInterface", fullName:
   "rs.co.bora5.programs.bab.model.interfaceCheck.CheckValidateEntityInterface"}
]
attributes: [
  {
    name: "trackingId",
    type: "String",
    isMandatory: false,
    annotation: {name: "Column",
      fullName: "javax.persistence.Column", parameters: ""}
  }
]
```

Code Level (C^j_{code}) – Generated Java File The FreeMarker template engine (implementing functor H_j, using `EntityTemplate.ftl`) takes the DataModel and produces the final Java file, Listing 4.

```
package rs.co.bora5.programs.pms.model;

import javax.persistence.Entity;
import javax.persistence.Table;
import javax.persistence.Column;
import rs.co.bora5.programs.bab.model.AbstractEntity;
import rs.co.bora5.programs.bab.model.interfaceCheck.CheckValidateEntityInterface;

@Entity
@Table(name = "Shipment")
public class Shipment extends AbstractEntity
    implements CheckValidateEntityInterface {

  @Column
  private String trackingId;

  public String getTrackingId() { return trackingId; }
  public void setTrackingId(String trackingId) { this.trackingId = trackingId; }

  @Override
  public void beforeValidate() {
    // TODO: Programmer adds shipment validation logic
  }
}
```

Listing 4 - Generated class Shipment.java (simplified)

This example illustrates how the semantic definition from the PIM (including both structure and abstract "Validatable" behavior) is deterministically transformed into concrete, architecturally correct Java code directly tied to the BABj library.

In the example above, it is evident that the Shipment class is annotated with `@Entity` (originating from the `entityMapping` rule defined in the ontology). It inherits `AbstractEntity` from the BABj library (this base class contains generic implementations such as an id field with an auto-generated key, metadata retrieval methods,

etc.) ensuring that the generated Shipment class is minimalistic and focused only on specific attributes. The `trackingId` field is generated based on ontological attributes. Had the ontology defined constraints like uniqueness for `trackingId`, this could have been reflected by generating an additional annotation (e.g., `@Column(unique=true)` or logical checks).

By contrast, a transient helper class such as `pms:TempFormData`, also modeled at the core level but linked to a separate `babj:DefaultTransientModelMapping_J`, would produce a class extending the appropriate BABj base type without the `@Entity` annotation and without `CheckValidateEntityInterface`, suitable for form validation or object representation of external query results. This contrast illustrates the bifunctor's genuine dependence on the mapping argument: the same domain type (`core:Entity`) yields structurally distinct artifacts depending on the chosen mapping strategy.

A similar procedure applies to other model elements. For example, if there is a business process `PackageDeliveryProcess` with two states (`In_Transit` and `Delivered`) and several tasks, the transformer would create a class (or a set of classes) for that process. It is possible to generate an enum with process state definitions, followed by a service class containing methods representing tasks (e.g., `onScan()` for the initial task, `onDelivered()` for the final one, etc.), where implementation details (repository calls, sending notifications) are delegated to the BABj library or left empty for the developer to fill in. For UI elements (forms and views), in the case of the Vaadin framework, the transformer creates corresponding Java classes extending Vaadin components and binding them to entities/processes. For instance, for a Shipment entity, an `EditShipmentWindow` class could be generated inheriting the BABj library base class for forms, `GenericWindow`, containing fields for name and email entry, as well as save buttons, where button logic is connected to service methods (again, following a library pattern).

It is important to emphasize that the transformer does not attempt to solve all domain-specific code; its role is to create an application skeleton that is immediately executable and contains all standard system parts. The possibility is left for the engineer to supplement specific business logic methods where necessary. For example, the generated `deliverPackage()` method in the service might initially be empty (or throw an `UnsupportedOperationException`) until the engineer adds concrete steps (inventory check, database update, sending email notifications, etc.). However, even these additions can be made platform-independent using ontological annotations or rules, e.g., a specific task could be marked as "email notification", allowing the transformer to create an email sending skeleton using the appropriate target framework.

In this way, the prototype confirms that the formal functorial specification can be successfully implemented in a practical tool that deterministically produces code from ontologically defined models.

7. Analysis and Discussion of Results

This section presents a detailed analysis and discussion of the results obtained by applying the proposed Ontology-Driven Code Generation (ODCG) system. Through a case study of a prototype project management application, the benefits of the approach regarding the

acceleration of initial development, reduction of manual work, and assurance of architectural consistency are quantified. In addition to the empirical evaluation of the quality of the generated code and the analysis of its sustainability through technological changes, the chapter clearly positions the achieved contribution in relation to existing paradigms, such as Model-Driven Engineering (MDE) and Ontology-Based Data Access (OBDA). Finally, a critical review of the methodology's limitations is provided, including the challenges of round-trip engineering and dynamic system aspects modeling, thereby defining the scope of applicability and opening directions for future research. Application of ODCG showed significant acceleration of the initial development cycle and increased architectural consistency. To assess the gains, a prototype business application of medium size was created using the proposed approach. For demonstration purposes, a project management system was modeled, including ~ 10 entities(Project, Task, Employee, Client, Invoice, etc.), 5 business processes (e.g., project approval process, invoicing), and 3 modules (Finance, Operations, User Management). The domain description in the BAB Core ontology itself encompassed about 50 concepts (classes and relations). From that model, the transformer produced a complete multi-tier application architecture for the Java stack. For each defined entity, a set of five coupled components was generated (JPA entity, DTO object, service class, Vaadin view, and input/update form), resulting in the generation of over 80 Java classes with a total of approximately 5000 lines of code. This amplification factor shows that a single concept in the ontology results in the generation of multiple implementation artifacts. For an experienced programmer, writing a similar application foundation manually would require weeks of work, while the transformer produced all the code automatically in a few seconds. Naturally, the generated code required minimal refinements (mainly populating concrete logic in specified methods). This result illustrates a measurable time saving and a reduction in manual work of approximately 60–70% at the start of the project, compared to complete manual setup. This figure should be interpreted as an exploratory engineering observation based on the authors' experience with comparable manually developed application foundations rather than as a rigorously validated metric. A controlled baseline comparison can be established in future work by implementing the same application foundation manually and through the ODCG process under comparable conditions, measuring development effort (person-hours), number of produced artifacts, defect counts, and maintenance effort across multiple independent projects. Since the basic code skeleton is generated automatically, frequent errors caused by the human factor such as typographical errors in class and field names, inconsistent application of annotations, or omissions in connecting application layers are removed from the process. The generated code possesses a consistent, template-defined structure and passes automatic validation according to ontology rules. For example, if the ontology prescribes that every entity class must have at least one key attribute, the transformer automatically ensures this or generates an appropriate warning. The quality of the generated code was assessed using static analysis methods. The analysis of the generated code was performed using the SonarQube tool (Community Edition version), using the standard Sonar way profile for the Java programming language. Given that the code was generated automatically, the goal of the analysis was verification of structural correctness and the absence of "code smell" patterns characteristic of manual coding. The aggregated analysis results for the application module are shown in Table 1. The evaluation has two distinct roles. The categorical preservation result is assessed through the verification obligations defined in Sec-

tion 5.6: mapping completeness, mapping injectivity, quotient compatibility, template determinism, and absence of unintended collapse. Static analysis of the generated code has a different role: it evaluates engineering quality of the generated artifacts and templates, but it should not be read as an empirical proof of the categorical preservation theorem.

Table 1. Overview of Code Quality Metrics (SonarQube)

Metric	Result	Grade	Comment
Bugs	0	A	No logical errors detected (e.g., NullPointerException risks).
Vulnerabilities	0	A	No security flaws (e.g., SQL Injection), thanks to the use of parameterized queries in generated repositories.
Code Smells	< 10	A	Negligible number, mostly related to specific domain class names not recognized by the tool.
Technical Debt	< 1 h	A	Technical Debt Ratio is below 0.1%, which is drastically better than the industry standard (5%).
Duplication	$\sim 3.5\%$	–	Duplication is present only at the structural level ("boilerplate" code), but not in business logic.
Cyclomatic Complexity	1.2 / method	–	Average complexity is extremely low, as generated methods are short, atomic, and without deep branching.

The generated code achieved an "A" grade in Reliability, Security, and Maintainability. In the analyzed case, SonarQube detected no bugs, vulnerabilities, security hotspots, or naming-convention violations, while the templates produced consistent use of Java annotations and resource-management patterns. The low cyclomatic complexity indicates that the generated methods are structurally simple. These results provide engineering evidence about the quality of the current Java templates and generated artifacts in the analyzed case; they do not establish that ODCG eliminates all syntax, style, or security errors in every generated system, nor do they constitute an independent proof of the categorical preservation theorem. One of the key advantages of the approach is easier system maintenance and evolution. If requirements change, say new entities or fields are added, it is sufficient to update the domain ontology and re-run code generation. Changes will propagate through the entire system consistently. Experience with the prototype showed that even more complex refactorings (e.g., renaming entities or splitting one process into two) can be carried out practically automatically, without the need for manual searching and correcting in multiple places in the code. Moreover, through a centralized ontology, it is easier to enforce best practices. For example, if an organization decides that all entities must have`createdAt` and `updatedAt` fields (creation and modification time) for tracking, this is simply added to the BAB Core ontology as a mandatory property for every class that is a subclass of Entity. The transformer will then add appropriate fields and annotations (e.g., JPA `@PrePersist`/`@PreUpdate` methods) to all generated entities. Such a mechanism ensures standardization of architecture and design across different projects.

The stack-parametrized categorical construction also provides a formal basis for possible future support of additional technological environments. Extending the methodology to another target stack would require defining a corresponding stack-specific ontology and library, a conformance functor, a target category, and the associated generation templates, while leaving the BAB Core domain model and the mapping schema unchanged. Under the same verification conditions, such an implementation could be checked against the structural-preservation properties established in Section 5.5. However, the present work implements and evaluates only the Java-based BABj stack; no additional stack-specific ontology, library, or generator is claimed. The ontology and transformer do not participate in normal production execution; the generated application runs on the standard target libraries. Consequently, the methodology introduces no additional ontology-reasoning or transformation component at runtime. Runtime performance relative to an equivalent manually written application was not independently benchmarked in this study. The proposed approach differs from conventional MDE and OBDA approaches in both its ontological foundation and its categorical formalization. OWL 2 DL reasoning is used to validate the input model before generation, while the complete generation pipeline is formalized as the composite $T_s = H_s \circ F_s$. Natural transformations do not represent the entire model-to-code process; they express coherent refinements between mapping strategies. Under the verification conditions stated in Section 5.6, the composite pipeline preserves the selected schema-level structure and explicitly represented mapping distinctions, without claiming operational or denotational semantic equivalence between the ontology and the generated code. Compared to OBDA approaches (e.g., Ontop), which use ontology as an interface for executing queries over databases without generating code [16], this work generates static code in advance. Ontological knowledge is not used just as an abstraction for queries, but is directly implemented in software, enabling deeper integration of the conceptual layer and better control over system performance and behavior in specific contexts. It is useful to separate two kinds of benefit. Benefits shared with model-driven and low-code approaches in general include acceleration of initial development, architectural consistency, elimination of a class of manual errors, and easier maintenance; these follow from the model-driven nature of the method and are not, by themselves, distinctive to this work. Benefits distinctive to the proposed method are: (1) a formal, verifiable guarantee of schema-level structural preservation, expressed as the conditional theorem of Section 5.5 and reduced to the five verification conditions of Section 5.6; (2) formalization of a stack-parametrized generation mechanism as a family of evaluation bifunctors with provable functoriality, providing a formal extension point through which support for another platform would require a new conformance functor and target category while leaving the domain model unchanged; the present operational implementation of this mechanism is limited to BABj; and (3) the ability to encode selected semantic constraints, such as cardinalities and disjointness, into the validation that gates generation. The first two, in particular, are not provided with formal guarantees by typical MDD or low-code frameworks. Although the proposed approach offers a number of advantages, there are important limitations that should be explicitly stated. First, complete generation of all application aspects is not always achievable. Certain business rules and procedural logic often exceed the expressive power of ontologies based on description logic and require additional interventions. This challenge is solved by introducing clearly defined ”hook” points in the generated code, intended for manual extension. Approaches such

as [59] show that knowledge-representation models can drive the generation of domain logic; integrating such behavioral synthesis into the hook points of our pipeline is a concrete direction for future work. Second, the success of the entire process largely depends on the quality of the input ontology, which requires expertise in domain knowledge and formal modeling and significant initial effort in creating the input ontology. However, that effort pays off in the long term through reusability and automated maintenance and opens space for further research aiming to delegate at least part of the tasks occurring in the generation process to machines. Third, synchronization between ontology and code in the case of manual code changes is a significant engineering challenge that exceeds the scope of classic one-way transformation. Since the proposed methodology does not offer an appropriate mechanism so far, recommended practice implies retaining primary logic within the ontology, with strict isolation of necessary manual interventions in the code, thereby ensuring the predictability and safety of the regeneration process. Fourth, the question of dynamic system behavior is treated partially. Ontologies are currently used dominantly for modeling structure and do not in themselves support modeling temporal changes, object lifecycles, or states. Therefore, within the proposed system, such aspects are expressed via specific software interfaces. These interfaces allow linking the declarative model with procedural conditions depending on the current system state, enabling the processing of dynamic aspects without directly changing the ontological model. In this way, a balance between formal semantics and execution flexibility is achieved. Fifth, unlike OBDA approaches that allow flexible execution of queries over heterogeneous sources in real-time, here static code adapted to a predefined context is generated. Although this means less flexibility at runtime, advantages include better performance, easier integration into existing applications, and elimination of the need for reasoners in the production environment. The research combines ontological semantic modeling, categorical formalization, and software engineering. The code-generation pipeline is formalized as the composition $T_s = H_s \circ F_s$, while natural transformations express coherent refinements between mapping strategies. In this way, ad hoc model translation is replaced by a formally specified and conditionally verified transformation framework.

8. Conclusion

This paper presents a methodology for program code generation in which an ontology serves as a formal domain model and categorical transformations project that model into a platform-specific realization. A central formal contribution is the demonstration that the ImplementationMapping mechanism, which ontologically encodes generation rules within the model, gives rise to a family of evaluation bifunctors parametrized by a target technology stack. This result, grounded in the product–exponential adjunction in Cat, provides a formal stack-parametrized construction under which each instantiated target stack satisfying the stated verification conditions preserves the selected schema-level structure. The operational implementation and evaluation presented in this paper instantiate this construction only for the Java-based BABj stack. Through the multi-phase transformation from BAB Core, through BABj, to concrete Java source code, the paper demonstrates how a high-level description of a business system can be translated into working software. The key idea is to formalize the mappings between model levels as functors, with natural transformations ensuring coherence between mapping refinements; under the verification

conditions stated in Section 5.6, the generated code preserves the selected schema-level structure and explicitly represented mapping distinctions of the source ontology. The proposed solution is a step forward towards automation of software development, combining the power of semantic modeling and category theory with the practical requirements of software engineering. The result is a methodology that is both formal and practical: formal because it makes schema-level structural preservation a verifiable, conditional property, and practical because its BABj implementation supports the development of complex Java-based information systems. Development speed is increased by eliminating repetitive coding tasks; engineers can focus on domain design, while the transformer handles routine implementation. Automatic adherence to proven patterns and architectural principles contributes to improving code quality, which reduces technical debt and facilitates maintenance. The stack-parametrized construction applies uniformly to every target stack for which the corresponding conformance functor, target category, generation functor, and verification conditions are provided. Consequently, the same domain ontology can be reused across different current or future target stacks without modifying its platform-independent structure. BABj is the presently implemented and empirically evaluated instantiation of this general construction. The paper states clearly differentiation of the proposed methodology from existing MDE directions and OBDA approaches. Furthermore, limitations of the proposed solution have been identified and explicitly stated. Future research plans include research tasks aimed at mitigating and, where possible, eliminating the identified shortcomings. In this regard, short-term tasks encompass improvements aimed at enhancing semantic checks and optimizations, as well as improvements related to model and code synchronization. More specifically, one task is the use of a reasoner over the ontology in such a way that, in addition to checking model consistency before code generation, redundant entities or relations can be detected, inheritance automatically introduced when meaningful, etc. Another important research task that is currently in our focus is a two-way synchronization of a model and code: possibilities will be explored for reflecting code changes (made by programmers during fine-tuning) back into the ontology, thereby maintaining the source of truth in the model. In fact, research is actively being conducted on operationalizing the theoretical concept of adjunction between functors I_s and J_s, with an aim to establish a formally verified feedback loop whereby changes in code would consistently update the ontological model. Two other long-term research tasks that we see are exploring the possibilities of richer behavioral modeling and exploring the possibilities of automating ontology generation, which could potentially pave the way for applying the methodology to refactoring existing code. Finally, continuous and comprehensive performance evaluation is planned through the application of the approach to different domains and larger systems. We expect scalability to prove good, as ontologies can model even very complex domains, while adding new patterns to the transformer does not complicate existing ones. Availability of artifacts. The artifacts supporting this work (the BAB Core and BABj ontologies, the BABj library, the FreeMarker templates, and the BABGen prototype) are maintained in a GitHub repository. As the work is part of ongoing doctoral research and is subject to commercial exploitation, the repository is not publicly released; access is granted by the authors on reasonable request for academic evaluation and verification.

Acknowledgments. This research has been supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (Contract No. 451-

03-34/2026-03/200156) and the Faculty of Technical Sciences, University of Novi Sad through project "Scientific and Artistic Research Work of Researchers in Teaching and Associate Positions at the Faculty of Technical Sciences, University of Novi Sad 2026" (No. 01-3609/1).

References

1. Abouzahra, A., Sabraoui, A., Afdel, K.: Model composition in model driven engineering: A systematic literature review. Information and Software Technology 125, 106316 (2020)
2. Amr, M.F., Benmoussa, N., Mansouri, K., Qbadou, M.: Transformation of the CIM model into a PIM model according to the MDA approach for application interoperability: Case of the "COVID-19 patient management" business process. International Journal of Online and Biomedical Engineering 17(5), 49–68 (2021)
3. Asim, M.N., Wasim, M., Khan, M.U.G., Mahmood, W., Abbasi, H.M.: A survey of ontology learning techniques and applications. Database 2018, bay101 (Oct 2018)
4. Atkinson, C., Kühne, T.: Model-driven development: a metamodeling foundation. IEEE Software 20(5), 36–41 (2003)
5. Atzeni, M., Atzori, M.: CodeOntology: RDF-ization of source code. In: The Semantic Web – ISWC 2017. Cham (2017)
6. Berners-Lee, T.: The Semantic Web road map. W3C (Oct 1998), https://www.w3.org/DesignIssues/Semantic.html, accessed: 29 November 2025
7. Berners-Lee, T., Cailliau, R., Luotonen, A., Nielsen, H.F., Secret, A.: The World-Wide Web. Communications of the ACM 37(8), 76–82 (1994)
8. Bhogal, J., Macfarlane, A., Smith, P.W.H.: A review of ontology based query expansion. Information Processing & Management 43(4), 866–886 (Jul 2007)
9. Bogdanović, B., Konjović, Z., Obradović, Đ.: BAB (Business Application Builder) framework for rapid development of business information systems. In: Proceedings of the International Scientific Conference – Sinteza 2023. Belgrade (2023)
10. Bogdanović, B., Marković, V., Segedinac, M., Konjović, Z.: Ontological modelling and reasoning for the abstractHome class of the BAB framework for PAIS systems. In: Proceedings of the International Scientific Conference – Sinteza 2025. Belgrade (2025)
11. Bogdanović, B., Obradović, Đ., Segedinac, M., Konjović, Z.: BAB framework – towards an extensible software platform for AI-augmented process aware business information systems. In: Disruptive Information Technologies for a Smart Society. ICIST 2024. Lecture Notes in Networks and Systems (2024)
12. Boström, P., Neovius, M., Oliver, I., Waldén, M.: Formal transformation of platform independent models into platform specific models. In: Julliand, J., Kouchnarenko, O. (eds.) Formal Specification and Development in B. B 2007. Lecture Notes in Computer Science, vol. 4355, pp. 186–200. Springer, Berlin, Heidelberg (2006)
13. Brookes, B.C.: The foundations of information science. Part I. philosophical aspects. Journal of Information Science 2(3–4) (1980)
14. Calegari, D., Szasz, N.: Verification of model transformations: A survey of the state-of-the-art. Electronic Notes in Theoretical Computer Science 292, 5–25 (Mar 2013)
15. Calero, C., Ruiz, F., Piattini, M. (eds.): Ontologies for Software Engineering and Software Technology. Springer, Berlin, Heidelberg (2006)
16. Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., Poggi, A., Rodriguez-Muro, M., Rosati, R.: Ontologies and databases: The DL-Lite approach. In: Tessaris, S., Franconi, E., Eiter, T., Gutierrez, C., Handschuh, S., Rousset, M., Schmidt, R.A. (eds.) Reasoning Web. Semantic Technologies for Information Systems. Reasoning Web 2009, Lecture Notes in Computer Science, vol. 5689, pp. 255–356. Springer, Berlin, Heidelberg (2009)

17. Chen, J., Mashkova, O., Zhapa-Camacho, F., Hoehndorf, R., He, Y., Horrocks, I.: Ontology embedding: A survey of methods, applications and resources. IEEE Transactions on Knowledge and Data Engineering 37(7), 4193–4212 (Apr 2025)
18. Corcho, O., Priyatna, F., Chaves-Fraga, D.: Towards a new generation of ontology based data access. Semantic Web 11(1), 153–160 (2020)
19. Cranefield, S., Pan, J.: Bridging the gap between the model-driven architecture and ontology engineering. International Journal of Human-Computer Studies 65(7), 595–609 (2007)
20. Dávid, I., Aslam, K., Malavolta, I., Lago, P.: Collaborative model-driven software engineering – a systematic survey of practices and needs in industry. Journal of Systems and Software 199(C), 111626 (2023)
21. Decker, S., Melnik, S., van Harmelen, F., Fensel, D., Klein, M.C.A., Broekstra, J., Erdmann, M., Horrocks, I.: The Semantic Web: the roles of XML and RDF. IEEE Internet Computing 4(5), 63–73 (2000)
22. Diskin, Z., Maibaum, T.: Category theory and model-driven engineering: From formal semantics to design patterns and beyond. In: Proceedings of ACCAT 2012 (2012)
23. Dudáš, M., Lohmann, S., Svátek, V., Pavlov, D.: Ontology visualization methods and tools: A survey of the state of the art. The Knowledge Engineering Review 33, e10 (2018)
24. Dumas, M., van der Aalst, W.M., ter Hofstede, A.H. (eds.): Process-Aware Information Systems: Bridging People and Software through Process Technology. John Wiley & Sons, Inc. (2005)
25. Dumas, M., La Rosa, M., Mendling, J., Reijers, H.A.: Fundamentals of Business Process Management. Springer, Berlin, Heidelberg (2018)
26. Espinoza-Arias, P., Garijo, D., Corcho, O.: Crossing the chasm between ontology engineering and application development: A survey. Journal of Web Semantics 100655 (Jul 2021)
27. Evermann, J., Wand, Y.: Towards ontologically based semantics for UML constructs. In: Conceptual Modeling – ER 2001. Lecture Notes in Computer Science (2001)
28. Flahive, A., Taniar, D., Rahayu, W.: Ontology as a service (OaaS): extracting and replacing sub-ontologies on the cloud. Cluster Computing 16(4), 947–960 (Dec 2013)
29. Fonseca, C.M., Sales, T.P., Viola, V., da Fonseca, L.B.R., Guizzardi, G., Almeida, J.P.A.: Ontology-driven conceptual modeling as a service. In: FOMI 2021: 11th International Workshop on Formal Ontologies meet Industry, held at JOWO 2021: Episode VII The Bolzano Summer of Knowledge, September 11–18. Bolzano, Italy (2021)
30. Gangemi, A., Catenacci, C., Ciaramita, M., Lehmann, J.: Modelling ontology evaluation and validation. In: The Semantic Web: Research and Applications. ESWC 2006. Lecture Notes in Computer Science (2006)
31. Glimm, B., Horrocks, I., Motik, B., Stoilos, G., Wang, Z.: HermiT: An OWL 2 reasoner. Journal of Automated Reasoning 53(3), 245–269 (2014)
32. Goseva-Popstojanova, K., Kahsai, T., Knudson, M., Kyanko, T., Nkwocha, N., Schumann, J.: Report: Survey on model-based software engineering and auto-generated code (2016)
33. Grambow, G., Oberhauser, R., Reichert, M. (eds.): Advances in Intelligent Process-Aware Information Systems – Concepts, Methods, and Technologies. Springer, Cham (2017)
34. Henderson-Sellers, B.: Bridging metamodels and ontologies in software engineering. Journal of Systems and Software 84(2), 301–313 (Feb 2011)
35. Hitzler, P., Krötzsch, M., Rudolph, S.: Foundations of Semantic Web Technologies. Chapman and Hall/CRC (2009)
36. Hoseini, S., Theissen-Lipp, J., Quix, C.: A survey on semantic data management as intersection of ontology-based data access, semantic modeling and data lakes. Journal of Web Semantics 100819 (Jul 2024)
37. Jones, C.B.: Software Development: A Rigorous Approach. Prentice/Hall International (1980)
38. Kahani, N., Bagherzadeh, M., Cordy, J.R., Dingel, J., Varró, D.: Survey and classification of model transformation tools. Software & Systems Modeling 18, 2361–2397 (2019)

39. Kherraf, S., Lefebvre, É., Suryn, W.: Transformation from CIM to PIM using patterns and archetypes. In: ASWEC '08: Proceedings of the 19th Australian Conference on Software Engineering (2008)
40. Kleppe, A., Warmer, J., Bast, W.: MDA Explained: The Model Driven Architecture: Practice and Promise. Addison-Wesley (2003)
41. KOMMA Project: KOMMA – RDF object mapper and editing framework for Java. Fraunhofer IWU, https://komma.enilink.net/, accessed: 23 November 2025
42. Kovalyov, S.P.: Category-theoretic approach to software systems design. Journal of Mathematical Sciences 214(6), 814–853 (2016)
43. Lano, K., Kolahdouz-Rahimi, S., Yassipour-Tehrani, S., Sharbaf, M.: A survey of model transformation design patterns in practice. Journal of Systems and Software 140, 48–73 (2018)
44. Lassila, O., van Harmelen, F., Horrocks, I., Hendler, J., McGuinness, D.L.: The Semantic Web and its languages. IEEE Intelligent Systems and their Applications 15(6), 67–73 (2000)
45. Lehner, D., Zhang, J., Pfeiffer, J., Sint, S., Splettstößer, A.K., Wimmer, M., Wortmann, A.: Model-driven engineering for digital twins: a systematic mapping study. Software and Systems Modeling 24, 1339–1377 (Oct 2025)
46. Li, J., Garijo, D., Poveda-Villalón, M.: Large language models for ontology engineering: A systematic literature review (Apr 2025), https://www.semantic-web-journal.net/system/files/swj3864.pdf, accessed: 9 January 2026
47. Mac Lane, S.: Categories for the Working Mathematician. Springer, New York (1978)
48. Marcén, A.C., Iglesias, A., Lapeña, R., Pérez, F., Cetina, C.: A systematic literature review of model-driven engineering using machine learning. IEEE Transactions on Software Engineering 50(9), 2269–2293 (Jul 2024)
49. Michael, J., Cleophas, L., Zschaler, S., Clark, T., Combemale, B., Godfrey, T., Khelladi, D.E., Kulkarni, V., Lehner, D., Rumpe, B., Wimmer, M., Wortmann, A., Ali, S., Barn, B., Barosan, I., Bencomo, N., Bordeleau, F.: Model-driven engineering for digital twins: Opportunities and challenges. Systems Engineering 28(5), 659–670 (Apr 2025)
50. Miller, J., Mukerji, J.: MDA guide version 1.0.1. Tech. rep., Object Management Group (May 2003), https://www.semanticscholar.org/paper/MDA-Guide-Version-1.0.1-Miller-, accessed: 29 November 2025
51. Mkhinini, M.M., Narsis, O.L., Nicolle, C.: Combining UML and ontology: An exploratory survey. Computer Science Review 35(1), 100223 (2020)
52. Mordecai, Y., Fairbanks, J.P., Crawley, E.F.: Category-theoretic formulation of the model-based systems architecting cognitive-computational cycle. Applied Sciences 11(4), 1945 (2021)
53. Noguera, M., Hurtado, M.V., Rodríguez, M.L., Chung, L., Garrido, J.L.: Ontology-driven analysis of UML-based collaborative processes using OWL-DL and CPN. Science of Computer Programming 75(8), 726–760 (Aug 2010)
54. Oberle, D., Grimm, S., Staab, S.: An ontology for software. In: Handbook on Ontologies, pp. 383–402. Springer, Berlin, Heidelberg (2009)
55. Olszewska, J.I., Allison, I.K.: ODYSSEY: Software development life cycle ontology. In: Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2018) (2018)
56. OpenRDF: AliBaba. Bitbucket, https://bitbucket.org/openrdf/alibaba/src/master/, accessed: 23 November 2025
57. Pan, J.Z., Staab, S., Aßmann, U., Ebert, J., Zhao, Y., Oberle, D.: Conclusion and outlooks. In: Aßmann, U., Ebert, J., Pan, J.Z., Staab, S., Zhao, Y. (eds.) Ontology-Driven Software Development, pp. 319–322. Springer, Berlin, Heidelberg (2013)
58. Ruiz, F., Hilera, J.R.: Using ontologies in software engineering and technology. In: Calero, C., Ruiz, F., Piattini, M. (eds.) Ontologies for Software Engineering and Software Technology, pp. 49–102. Springer, Berlin, Heidelberg (2006)
59. Rybiński, K., Śmiałek, M.: Low-code development using requirements and knowledge representation models. Computer Science and Information Systems 21(3), 685–724 (2024)

60. Sapel, P., Comet, L.M., Dimitriadis, I., Hopmann, C., Decker, S.: A review and classification of manufacturing ontologies. Journal of Intelligent Manufacturing 36, 3669–3693 (2025)
61. Shimizu, C., Hitzler, P.A.: Accelerating knowledge graph and ontology engineering with large language models. Journal of Web Semantics 85, 100862 (May 2025)
62. Silega, N., Noguera, M., Macias, D.: Ontology-based transformation from CIM to PIM. IEEE Latin America Transactions 14(9), 4156–4165 (Dec 2016)
63. Silega, N., Noguera, M., Rogozov, Y.I., Lapshin, V.S., González, T.: Transformation from CIM to PIM: A systematic mapping. IEEE Access 10, 90857–90872 (Aug 2022)
64. Singh, Y., Sood, M.: Transformation of platform-independent model into platform-specific model in model-driven architecture (2013)
65. Staab, S., Walter, T., Gröner, G., Parreiras, F.S.: Model driven engineering with ontology technologies. In: Reasoning Web. Semantic Technologies for Software Engineering, pp. 62–98. Springer, Berlin, Heidelberg (2010)
66. Steinberg, D., Budinsky, F., Paternostro, M., Merks, E.: EMF: Eclipse Modeling Framework. Addison-Wesley Professional, 2 edn. (2008)
67. Syriani, E., Luhunu, L., Sahraoui, H.: Systematic mapping study of template-based code generation. Computer Languages, Systems & Structures 52, 43–62 (Jun 2018)
68. Trojahn, C., Vieira, R., Schmidt, D., Pease, A., Guizzardi, G.: Foundational ontologies meet ontology matching: A survey. Semantic Web 13(4), 685–704 (2022)
69. Vickery, B.C.: Ontologies. Journal of Information Science 23(4), 277–286 (1997)
70. Völkel, M., Sure, Y.: RDFReactor – from ontologies to programmatic data access. In: 4th International Semantic Web Conference (ISWC). Galway, Ireland (2005)
71. Zhu, M., Grogono, P., Ormandjieva, O.: Using category theory to verify implementation against design in concurrent systems. Procedia Computer Science 52, 530–537 (2015)

Borivoj Bogdanović holds an M.Sc. degree and is currently a Ph.D. student in Computer Science at Singidunum University. He is also a senior software engineer with extensive experience in the development of business information systems. His research interests include software engineering, semantic technologies, ontology-driven and model-driven development, process-aware information systems, and AI-augmented business systems.

Vidan Marković is an Associate Professor in the field of Computer Science at Singidunum University. He holds a PhD in Computer Science from the University of Novi Sad. His research and professional interests include software engineering, software quality management and improvement, information systems interoperability, machine learning, the application of information technologies, and the management of complex software projects. He is a professional project manager with extensive experience in leading complex IT and digital transformation projects. He is the author and co-author of numerous scientific papers published in international journals and presented at scientific conferences.

Milan Segedinac received his M.Sc. degree in 2008 and Ph.D. in 2014 in Computer Science from the University of Novi Sad, Faculty of Technical Sciences. He holds a full professor position at the same faculty. He has authored papers in international and national journals and conferences in the field of Artificial Intelligence.

Zora Konjović, PhD, is a graduate mathematician with a PhD in robotics obtained at the Faculty of Technical Sciences of the University of Novi Sad. She is a retired full professor in Applied Computer Science and Informatics at the Faculty of Technical Sciences,

University of Novi Sad, as well as in Informatics and Computer Science at Singidunum University in Belgrade. Her primary areas of scientific and professional interest include software engineering, artificial intelligence, semantic technologies, technology-enhanced learning, and mathematical modeling. Throughout her long academic and research career, Professor Konjović has published more than 200 scientific papers and has successfully managed dozens of national and international scientific projects.

Received: February 4, 2026; Accepted: August 15, 2026.

Computer Science and Information Systems 23(4):1395–1430 https://doi.org/10.2298/CSIS260301046M

Paper Type: Research Paper

Leveraging Content-based Metadata in Catalogs to Improve Discoverability in Data Spaces*

Adriana Morejón[1], Alberto Berenguer[1], Lucía de Espona[1], David Tomás[1], and Jose-Norberto Mazón[1]

Department of Software and Computing Systems, University of Alicante, Carretera San Vicente del Raspeig s/n, 03690 San Vicente del Raspeig, Spain
adriana.morejon@ua.es (corresponding author), alberto.berenguer@ua.es, lucia.espona@ua.es, dtd@ua.es, jnmazon@ua.es

Abstract. Data spaces facilitate sovereignty-preserving data sharing within federated ecosystems; however, dataset discovery is often constrained by catalogs that rely primarily on high-level metadata standards such as DCAT-AP. These models typically lack structural and content-level details, which impedes fine-grained relevance assessment and complicates both internal dataset discovery and the federation of external open data into data spaces. This paper introduces cbmDCAT-AP, a backward-compatible extension of DCAT-AP that integrates content-based metadata as primary catalog elements. The proposed extension augments dataset descriptions with field-level semantics, and representative sample values, thereby enabling content-informed discovery without disclosing entire datasets. Furthermore, a clustering-based reduction technique is proposed to generate representative samples values through the use of word embeddings and K-means clustering. Finally, an end-to-end evaluation across two scenarios (internal dataset discovery and external open data selection) has been conducted to evaluate and validate our proposed extension of DCAT-AP.

Keywords: Data Space, Metadata, Catalog, DCAT-AP, Discoverability, Open Data

1. Introduction

Data spaces are emerging as federated infrastructures that integrate secure data exchange of data ecosystems with the economic incentives of data markets, while maintaining provider sovereignty [4, 41]. Grounded in the principles of interoperability, trust, and data sovereignty, data spaces represent a paradigm shift in business-to-business scenarios, where data can substantially enhance collaborative capabilities [12, 44].

Data spaces are commonly established using reference architectures such as the International Data Spaces Association (IDSA), Gaia-X, FIWARE, and iSHARE frameworks [24, 32]. These architectures are implemented through standardized connectors,

* This paper is an extended version of the conference paper "Improving Data Discovery Effectiveness: Experimental Evaluation of Content-Based Catalogs in Data Spaces" presented in the 7th Workshop on Modern Approaches in Data Engineering and Information System Design organized in conjunction with the conference ADBIS 2025.

including Eclipse Dataspace Components (EDC)[1] and the FIWARE Data Space Connector[2]. Such connectors facilitate technical interoperability and enforce usage control and access policies, enabling providers to retain control while permitting consumers to discover and access data without compromising data sovereignty [25]. Also, the rollout of Common European Data Spaces is supported through EU-funded projects, including the Smart Open-source Middleware (SIMPL) programme, which provides an open-source, smart and secure middleware platform to foster interoperability and controlled data access across data spaces [18, 19].

In these reference architectures, connectors and middleware, data catalogs are a critical building block for two complementary discovery scenarios. First, consumers need to identify relevant datasets already offered within the data space by multiple sovereign providers, often under restrictive access policies and without the possibility of freely downloading or inspecting the full content. Second, data space participants increasingly aim to incorporate external open datasets (typically harvested from open data portals) in order to enrich the data space with contextual and baseline information, and unleashing the potential of open data within the same data space ecosystem. In this second scenario, the challenge is not only to locate open datasets, but to assess their fitness-for-use and to decide which ones should be onboarded and federated into the data space catalog with harmonized metadata.

Therefore, both scenarios require cataloguing datasets. Unfortunately, traditional data catalogs primarily depend on high-level descriptive metadata, such as title, license, keywords, and publisher, which are aligned with standards like the Data Catalog Vocabulary (DCAT) and its profiles (e.g., DCAT-AP) [20, 45]. While these metadata are effective for basic inventories, they consistently fail to communicate the practical utility of dataset content [35]. As a consequence, these dataset-centric metadata impedes consumers from comprehending internal structure and semantics, including field names, data types, column descriptions, and value ranges, as well as the specific semantics required to assess applicability for particular use cases [38]. Catalogs typically omit content evidence, forcing consumers to obtain full access (or to download large resources) merely to determine relevance. On one hand, in data spaces, this contradicts core principles of sovereignty and controlled disclosure [28]. This tension exemplifies Arrow's information paradox: exposing all data to demonstrate its value can compromise sovereignty and governance constraints, whereas concealing it prevents consumers from assessing practical utility [3]. On the other hand, in the open data scenario, it hinders dataset selection and onboarding decisions, discouraging the systematic federation of external datasets into the data space. Furthermore (and valid for both scenarios), empirical studies demonstrate that dataset discovery relying exclusively on high-level metadata yield low retrieval rates of relevant datasets and exhibit high abandonment rates due to insufficient semantic correspondence [6].

To address these limitations, we propose content-based catalogs through a DCAT-AP extension that incorporates structured content-level metadata describing dataset schema and semantics, as well as representative (sample) values, while remaining compatible with existing open data portals and data space catalogs. A central design principle is to treat dataset sampling as a first-class catalog artifact: compact, representative, and policy-aware samples provide evidence of content utility without requiring full disclosure or bulk

[1] https://github.com/eclipse-edc
[2] https://github.com/FIWARE/data-space-connector

transfer. Furthermore, we introduce an approach to extract representative samples suitable for inclusion in the catalog.

The proposed extension is integrated into the HELEADE data space[3], a territorial data space deployed by the University of Alicante to collect, organise, and share environmental and socioeconomic data under interoperability and data sovereignty principles. Validation is performed by mapping the extension to the EDC data catalog, which underpins the HELEADE architecture, and by demonstrating how the same enriched metadata model supports both scenarios. First, discovery of sovereign datasets within the data space and, secondly, discovery and selection of external datasets from open data portals for subsequent cataloguing inside the data space. Furthermore, we introduce an approach for extracting representative samples to be included in the catalog.

While our previous work [39] introduced a dataset sampling methodology that utilizes word embeddings and clustering techniques to extract representative samples from large datasets within data spaces, the approach presented in this paper extends that foundation by incorporating the following contributions:

1. A novel DCAT-AP extension that introduces content-based metadata as first-class catalog elements, including new classes and properties for describing dataset structure and semantics (e.g., fields, types, descriptions) as well as representative sample values.
2. A practical mapping as adoption pathway for the proposed DCAT-AP extension in the Eclipse Dataspace Component (EDC) framework, validated through integration in the HELEADE data space catalog.
3. An end-to-end evaluation of discoverability improvements, by using our DCAT-AP extension, across two aforementioned scenarios: (i) dataset discovery within the data space, and (ii) discovery and selection of open datasets from external open data portals for cataloguing and reuse within the data space.

This paper is structured as follows. Section 2 reviews related work on data spaces, cataloguing standards, content-aware discovery, and sampling techniques. Section 3 introduces the DCAT-AP extension for content-level metadata, detailing new classes, properties, RDF implementation, and mapping to the EDC catalog. Section 4 presents the proposed solution for integrating the extension within federated data spaces. It examines the system architecture, content-based metadata integration, and the methodology for sample generation. Section 5 presents experiments that evaluate improvements in discoverability for sovereign datasets and open data onboarding. Section 6 concludes with a summary and suggestions for future research.

2. Related Work

This section reviews prior work relevant to our proposal from three complementary points of view: open data portals and their metadata publishing and harvesting practices, data space architectures and the role of connectors and catalog services in sovereignty-preserving discovery, as well as cataloguing standards and extensions, with emphasis on DCAT/DCAT-AP and existing profiles. We then summarize related work on content-aware dataset dis-

[3] https://dataspace.heleade.iuii.ua.es/

covery and on sampling and preview generation for tabular data, highlighting the gap addressed by our DCAT-AP extension and sampling methodology.

2.1. Open Data Portals and Metadata Harvesting

Open data portals are online platforms where governments and organizations make public sector data available for access, use, and reuse. These portals serve as a central repository for datasets, providing a user-friendly way to discover, download, and analyze information. They play a crucial role in promoting transparency, innovation, and civic engagement by enabling citizens, researchers, and businesses to access a wide range of data [17].

Beyond human-facing search interfaces, open data portals increasingly support machine-to-machine interoperability through APIs and RDF-based metadata exports, enabling third parties to harvest and federate dataset descriptions across catalogs. In practice, however, cross-portal discovery and systematic reuse are hindered by heterogeneous modelling choices, inconsistent use of controlled vocabularies, and varying metadata completeness [2, 17]. As a result, dataset selection for downstream integration often requires manual inspection steps that do not scale, for example, when onboarding large numbers of external open datasets into a data space.

CKAN (Comprehensive Knowledge Archive Network)[4] represents one of the most widely adopted open-source data portal platforms in the open data ecosystem [2]. Developed by the Open Knowledge Foundation and maintained by an active community, CKAN has been implemented by government agencies, research institutions, and organizations worldwide to publish and share their datasets. The platform implements a metadata schema that partially aligns with DCAT principles, offering a standardized approach to dataset description. In addition, CKAN's support for DCAT-AP is provided through the widely used extension ckanext-dcat, which enables CKAN-based portals to expose and consume metadata in RDF formats compliant with multiple DCAT-based application profiles. This extension offers pre-built CKAN schemas supporting DCAT-AP versions 1, 2, and 3, allowing data portals to align their internal metadata models with the European DCAT-AP standard while preserving domain-specific or national customizations via configuration-driven schemata (e.g., through ckanext-scheming)[5]. As a result, many national and regional open data portals built on CKAN already rely on DCAT-AP-aligned metadata exports to facilitate cross-portal discovery and interoperability within the European data ecosystem.

Many portals also provide lightweight content previews (e.g., showing the first rows of a table) or basic profiling indicators to assist users. While useful for quick inspection, these previews are typically neither standardized nor representative of the full dataset and may expose biased slices of the content (e.g., "head" rows only). Consequently, they provide limited support for automated relevance assessment and for onboarding decisions in federated environments (such as data spaces), motivating more principled and representative sampling mechanisms aligned with catalog metadata.

[4] https://ckan.org/
[5] https://docs.ckan.org/projects/ckanext-dcat/en/latest/mapping/

2.2. Data Space Architectures and Sovereignty-Preserving Discovery

A data space is a federated data infrastructure that supports trustworthy data sharing among data providers and data consumers [41]. To enable this, data spaces rely on connectors [25], which act as interfaces between participants and the ecosystem. Connectors ensure technical interoperability and enforce access policies, allowing providers to keep data sovereignty when sharing data while granting consumers seamless access without compromising control. Hence, data spaces enable new options for value creation where providers and consumers easily interact and work together to find, access, publish, consume, and reuse data, as well as to stimulate innovation [26].

In contrast to open data portals, discovery in data spaces is inherently constrained by policy: participants may advertise datasets and data services while restricting direct inspection of content until contractual, legal, or technical conditions are satisfied. Therefore, catalogs in data spaces must balance discoverability with controlled disclosure, exposing enough information to support relevance judgments while preserving sovereignty and minimizing unnecessary data transfer. This shifts the emphasis from purely descriptive metadata toward richer, yet governance-compatible, content evidence.

Numerous research initiatives have resulted in conceptual frameworks and reference architecture models to accelerate the development of data spaces [4].

In the European context, the deployment of Common European Data Spaces is a cornerstone of the EU data strategy, aiming to enable trustworthy cross-border and cross-sector data sharing and reuse [19]. To provide shared, reusable technical building blocks across these initiatives, the European Commission launched the Smart Open-source Middleware (SIMPL) programme, an open-source, smart and secure middleware intended to support interoperability and controlled data access across European data spaces and cloud-to-edge federations [18].

The i4Trust initiative emerges as a collaboration program, targeting the creation of data spaces by proposing an architecture based on commonly agreed building blocks [29]. They are a combination of components from FIWARE and iSHARE, two data space foundations. The iSHARE Foundation [34] maintains a trust framework for data spaces. The FIWARE Foundation [22] introduces an open source framework that promotes the development of interoperable, smart solutions.

Another innovative data space technical framework has been developed by Gaia-X, aiming to establish a trustworthy ecosystem where data is shared, maintaining the user's digital sovereignty of the data. This standard-based framework allows the implementation of distributed data systems in all European countries in a legally secure manner, enabling compliance with GDPR and other data regulations [24].

The International Data Spaces Association (IDSA) is a non-profit organization focusing on establishing and promoting standards for data spaces as trusted environments where organizations can share data while retaining full control over its use [33]. The IDS Reference Architecture Model (IDS-RAM) [32] materializes the standard developed from collecting requirements from various industries and the results gained from the model's implementation. Additionally, the IDS-RAM connectors fit the Gaia-X principles and architecture model [31].

The Eclipse Dataspace Components (EDC) [16] is a comprehensive framework (concept, architecture, code, samples) providing a basic set of features (functional and non-functional) that dataspace implementations can re-use and customize by leveraging the

framework's defined APIs and ensure interoperability by design. The EDC project is governed by the Eclipse Foundation and it is powered by the specifications of the Gaia-X AISBL Trust Framework and the IDSA Dataspace protocol.

A further recurring practice in deployed data spaces is the enrichment of the ecosystem with external open datasets (e.g., reference geodata, environmental baselines, administrative registries). In this onboarding scenario, operators and participants must identify candidate datasets in open portals and assess their fitness-for-use before federating them inside the data space with harmonized metadata. This reinforces the need for catalog mechanisms that support scalable, content-informed selection beyond high-level keywords.

One core element of both data space architectures and open data portals is the data catalog, which serves as a structured inventory of available datasets and associated metadata and access policies or usage licenses.

2.3. Cataloguing Standards for Data Discovery: DCAT and DCAT-AP

The Data Catalog Vocabulary (DCAT), a W3C recommendation, has emerged as the predominant metadata standard for data catalog implementation. Prior research has established DCAT's significant contribution to metadata interoperability across diverse data ecosystems, from open data portals utilizing CKAN to advanced data space architectures like those proposed by IDSA and Gaia-X. These architectures have operationalized DCAT through connector technologies such EDC and the FIWARE Data Connector. Several publications [37] have highlighted DCAT's potential as a foundation for open data discovery in federated data spaces, given its widespread adoption and interoperability-focused design.

Unfortunately, current DCAT-AP-based catalogs are insufficient for effective data discovery, posing two key limitations:

- Limited metadata: existing data catalogs primarily rely on high-level metadata to describe datasets but lack detailed structural information of dataset content, such as field names or descriptions, which are essential for precise dataset discovery.
- Content blindness: current data catalogs exclude actual dataset content, relying solely on metadata. This prevents data spaces initiatives from gaining deeper insights into open data relevance before federating the full dataset.

As an attempt to enhance semantic interoperability across European data portals, the extension DCAT-AP[6] emerges. Based on W3C's DCAT, it supports standardized dataset descriptions, enabling efficient data exchange and reuse. The basic use case for DCAT-AP is to enable cross-data portal search for datasets and make public sector data better searchable across borders and sectors. Although this extension improves data sharing among different sectors and includes additional support for access policies and legal aspects, the aforementioned two limitations regarding content description remain unresolved.

To overcome these challenges in data discovery, reference architectures for data spaces require enhancing data catalogs with content-based metadata. In response to these requirements, this work proposes an extension of DCAT-AP that integrates content-based metadata, while it is mapped to current EDC data catalog for the sake of compatibility.

[6] https://semiceu.github.io/DCAT-AP/releases/3.0.0/.

2.4. Extending DCAT-AP: Sectoral and Domain Profiles

As this work involves the development of an extension to DCAT-AP, it is important to review previous extensions of this standard. DCAT-AP is recognized as the standard application profile for open data cataloguing in Europe, with multiple extensions developed to address national and sector-specific requirements. DCAT-AP serves as a foundational framework upon which multiple specialised extensions have been developed, establishing it as an extensible core for domain-specific and interoperability scenarios. One notable extension is StatDCAT-AP [13], a profile designed to represent statistical metadata within open data portals. StatDCAT-AP extends DCAT-AP to accommodate statistical concepts such as units of measurement, geographical and temporal coverage, and hierarchical series structures, while preserving compatibility with the European profile.

Within the health sector, HealthDCAT-AP [14] has been introduced as a targeted extension for describing health data and services in the context of the European Health Data Space (EHDS). This profile addresses the specific requirements of sensitive health data and health services, including coordination with access, consent, and anonymisation policies, while maintaining compatibility with DCAT-AP and preserving the integrity of the metadata ecosystem.

Alongside StatDCAT-AP and HealthDCAT-AP, GeoDCAT-AP [42] is among the most established extensions for describing geospatial datasets and services within open data environments. A technical report outlines its integration of geographic metadata vocabularies, such as INSPIRE, with DCAT-AP, thereby enhancing the description of spatial dimensions, resolution, projection, and geographic coverage.

Within data governance and administrative records, proposals such as BRegDCAT-AP[7] have emerged as extensions of DCAT-AP for describing basic records and their associated services. This extension introduces specific properties to define record types, hierarchical structures, and relationships between entities. It maintains compatibility with the core DCAT-AP standard and enables the integration of record catalogs into open data infrastructures.

In scientific and research contexts, profiles such as EPOS-DCAT-AP have been developed to adapt DCAT-AP for describing Earth data infrastructure datasets and services. These profiles introduce properties related to data quality, traceability, and integration with disciplinary data models.

Overall, existing DCAT-AP extensions predominantly focus on aligning catalog metadata with sectoral concepts, services, governance constraints, or domain-specific semantics. While these profiles improve interoperability within their respective domains, they generally do not provide a generic, reusable model for content-level descriptions at the granularity required for schema-aware discovery (e.g., column-level semantics and representative values), nor do they treat representative samples as first-class catalog artefacts. This gap motivates our extension, which targets content evidence while preserving DCAT-AP compatibility.

[7] https://semiceu.github.io/uri.semic.eu-generated/BRegDCAT-AP/releases/3.0.0/.

2.5. Content-Aware Dataset Discovery

A large body of research on dataset discovery in tabular corpora and data lakes explores how to retrieve relevant tables using signals derived from schema and content, including column naming patterns, statistical summaries, entity linking, and semantic representations (e.g., embedding-based encodings of columns or tables) [7,8,47]. These approaches demonstrate that content-derived signals can substantially improve relevance ranking compared to keyword-only metadata, especially when users search for datasets matching a particular structure or concept [15,21].

However, most content-aware discovery pipelines implicitly assume that datasets (or substantial sketches of them) can be centrally accessed, indexed, and processed to build searchable representations [1]. This assumption is misaligned with the sovereignty and controlled-access constraints of data spaces [28,41], where providers may be unwilling or unable to expose full content for indexing [5]. This motivates standardizing content-level evidence in catalogs through governance-aware metadata and representative samples that can be produced by providers and shared under explicit policies.

2.6. Sampling and Data Previews for Tabular Dataset Discovery

The initial theoretical task involves reconciling the extension of the DCAT-AP standard with updated information regarding dataset content. Furthermore, this paper addresses a significant limitation for data consumers: content opacity. This limitation compels consumers to access entire datasets for relevance assessment, which contradicts fundamental sovereignty principles. Accordingly, the research project examines practical methods for extracting data samples, enabling data consumers to evaluate a dataset using representative examples prior to determining its utility for their specific requirements.

Most existing approaches for data discovery assume direct access to dataset content for indexing and matching [1], an assumption incompatible with the sovereignty principles fundamental to data spaces. Furthermore, current techniques often struggle with semantic heterogeneity across domains, where the same concept may be expressed differently in various datasets. Scalability remains problematic in federated environments with thousands of distributed datasets, as content-based matching becomes computationally expensive. Additionally, most methods lack mechanisms to assess the trustworthiness and quality of discovered datasets, critical factors in a productive data space.

While using data samples for discovery rather than complete datasets improves computational efficiency, it introduces the critical challenge of generating representative samples that maintain both utility and privacy guarantees.

A variety of sample generation techniques have emerged to address this challenge. Random representative sampling remains a foundational method, providing broad coverage with minimal bias, though it may underrepresent rare patterns [30]. More advanced approaches include Virtual Sample Generation (VSG), which synthesizes data points through interpolation or neural methods to enrich sparse datasets [9]. Bootstrapping methods, popular in machine learning, generate multiple resampled datasets with replacement to support robust training and uncertainty estimation [27]. Meanwhile, synthetic data generation techniques using models such as GANs and variational autoencoders simulate realistic sample while preserving privacy, facilitating discovery in restricted-access

environments [10]. Other strategies include active sampling, where data is selectively retrieved based on query relevance or informativeness [43], and clustering-based sampling, which ensures semantic diversity by extracting representative elements from identified clusters [11]. Dimensionality reduction methods, such as PCA or autoencoder-based embeddings, also guide sampling by preserving the global structure of high-dimensional data [40]. Lastly, metadata-driven techniques leverage high-quality semantic annotations and quality metrics to inform targeted sample extraction [46].

However, these existing approaches exhibit critical limitations taking into account the context in which we want to use the information, i.e. a data discovery context. Methods like random, bootstrap, or synthetic sample generation, lack semantic awareness, often overlooking the underlying meaning or contextual relationships between data attributes. Moreover, methods based solely on metadata quality are ineffective in cases where metadata is sparse, inconsistent, or superficial.

2.7. Research Gap and Positioning

Prior work shows that DCAT/DCAT-AP and their extensions enable interoperable catalog publication, and that content-aware discovery methods can improve retrieval when rich metadata are available. Nevertheless, there remains a gap between (i) standards-based cataloguing practices that remain largely dataset-centric and (ii) discovery techniques that rely on direct access to content for indexing and matching. This gap is especially problematic in data spaces, where sovereignty constraints limit content exposure by design.

Our work addresses this gap by introducing a DCAT-AP extension that standardizes content-level metadata and representative samples as governed catalog artefacts, and by demonstrating practical adoption through mapping to the EDC catalog model. In doing so, we support both discovery scenarios considered in this paper: searching sovereign datasets within the data space and selecting external datasets from open data portals for systematic onboarding and federating.

3. DCAT-AP extension for content-level metadata

DCAT-AP extends DCAT to facilitate cross-portal discovery of public sector datasets throughout Europe by standardizing high-level descriptive, administrative, and structural metadata, including titles, publishers, licenses, and themes. Nevertheless, DCAT-AP retains DCAT's dataset-level orientation and lacks native mechanisms for representing field-level structure or sample content, thereby constraining its effectiveness for fine-grained discovery within data spaces. To address this limitation, DCAT-AP is extended with content-based metadata, the cbmDCAT-AP extension.

Our cbmDCAT-AP extension preserves the semantics of both DCAT-AP and the underlying DCAT vocabulary. It introduces additional classes and properties that may be used in conjunction with standard DCAT-AP terms. This approach maintains backward compatibility with existing DCAT-AP catalogs and allows data providers to present more detailed structural, content-level information and data samples about their datasets. Comprehensive documentation has been established detailing the specifications of the cbmDCAT-AP extension[8].

[8] https://amorejon6.github.io/cbmDCAT-AP/

3.1. Scope and objectives

The cbmDCAT-AP extension is intended for catalog publishers and data space operators that already expose their metadata using DCAT-AP, while requiring support for dataset discovery beyond traditional metadata (i.e., content metadata, such as fields, types, sample values). Importantly, our cbmDCAT-AP extension keeps compatibility with existing DCAT-AP harvesting and validation workflows.

Specificaly, cbmDCAT-AP extension is designed for structured datasets (such as tabular datasets) and supports the representation of both data dictionaries tailored to specific data distributions (which contain field-level metadata), as well as samples of datasets (which are small subsets of instances derived from the original dataset).

Support for additional data models, such as spatial, unstructured, or multimedia, is beyond the scope of this version and will be considered in future extensions.

3.2. New classes and properties

Figure 1 illustrates the cbmDCAT-AP meta-model, which integrates elements for DCAT/DCAT-AP, Dublin Core Terms, and Schema.org, along with new components introduced by this extension. The reused elements maintain their original semantics, and cbmDCAT-AP defines their application in representing the structure and representative content of tabular datasets.

The primary DCAT and DCAT-AP classes reused in this context are those related to dataset publication and access. The class *dcat:Catalog* denotes a collection of catalogued resources, which may include datasets and data services. The class *dcat:Dataset* refers to the conceptual collection of data described, regardless of serialization or access method. Each dataset can be linked to one or more *dcat:Distribution* instances, where each instance specifies a particular accessible form of the dataset, such as a CSV file, a JSON resource, or another technical format. The class *dcat:DataService* identifies a service or endpoint that enables access to one or more datasets.

The relationships adopted from DCAT maintain their original interpretations. *dcat:distribution* connects a dataset to one of its accessible representations. *dcat:accessService* links a distribution to the service that provides access, while *dcat:servesDataset* specifies the datasets made available by a data service. Additionally, catalog relationships associate *dcat:Catalog* with the datasets and services it encompasses.

The model incorporates descriptive properties from established vocabularies. *dcterms:title* and *dcterms:description* supply a human-readable name and explanation for a resource, whereas *dcterms:type* specifies its type. For field descriptions, *schema:name* designates the technical field name as it appears in the underlying distribution. All other reused properties maintain the semantics established by their source vocabularies.

On top of these standard elements, cbmDCAT-AP introduces the following classes:

- **cbm:Sample**. A limited set of representative records derived from a dataset and exposed to support preliminary relevance assessment without requiring access to the complete dataset.
- **cbm:DataDictionary**. A structured description of the fields contained in a particular dataset distribution.

– **cbm:Field**. A description of an individual variable or attribute, such as a column in a tabular distribution. A field may include its technical name, human-readable title, datatype, description, and representative values.

The proposed model differentiates between dataset-level and distribution-level information. A sample is linked to *dcat:Dataset* because it offers representative evidence regarding the conceptual dataset under evaluation. Conversely, a data dictionary is linked to *dcat:Distribution*, as various distributions of the same dataset may present distinct schemas, field names, datatypes, formats, or serializations. Corresponding properties of the CBM extension are described in Table 1.

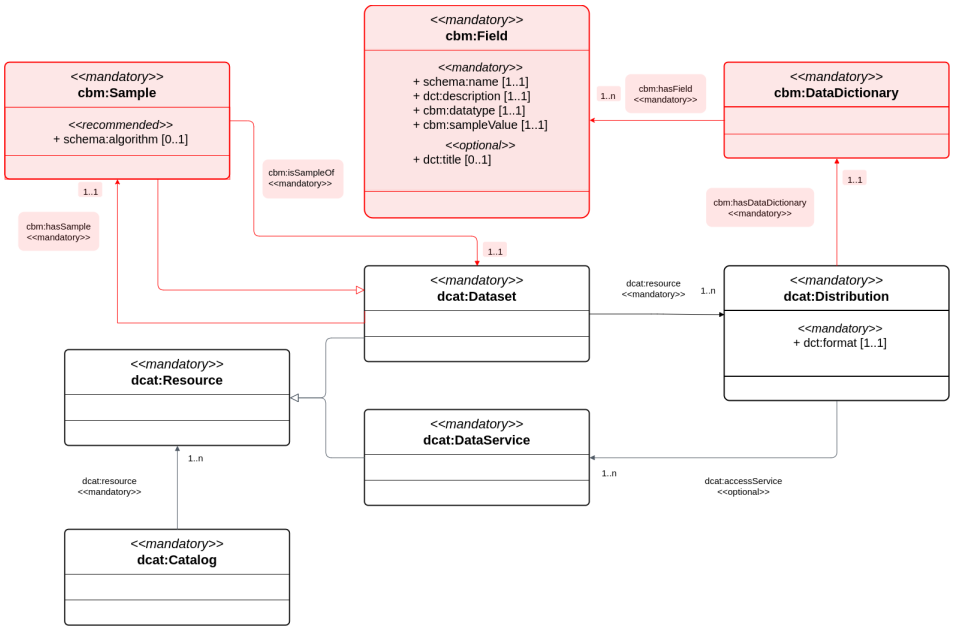


Fig. 1. Schema of the cbmDCAT-AP extension.

3.3. RDF implementation of cbmDCAT-AP extension

The cbmDCAT-AP extension is defined as an RDF schema using Turtle syntax, as well as a sample application to describe a dataset as described next.

```
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix cbm: <http://example.org/cbm#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix schema: <http://schema.org/> .
@prefix ex: <http://example.org/resource/> .

ex:urban-biodiversity-dataset a dcat:Dataset ;
```

```
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix schema: <http://schema.org/> .
@prefix cbm: <http://example.org/cbm#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

cbm:Sample a rdfs:Class ;
    rdfs:subClassOf rdf:Resource ;
    rdfs:label "Sample" ;
    rdfs:comment "Samples from a dataset." .

cbm:DataDictionary a rdfs:Class ;
    rdfs:subClassOf rdf:Resource ;
    rdfs:label "Data Dictionary" ;
    rdfs:comment "Describes fields in a dataset distribution." .

cbm:Field a rdfs:Class ;
    rdfs:subClassOf rdf:Resource ;
    rdfs:label "Field" ;
    rdfs:comment "A variable within a dataset (e.g., a column in
        a CSV file)." .

cbm:isSampleOf a rdf:Property ;
    rdfs:domain cbm:Sample ;
    rdfs:range dcat:Dataset ;
    rdfs:label "is sample of" .

cbm:hasSample a rdf:Property ;
    rdfs:domain dcat:Dataset ;
    rdfs:range cbm:Sample ;
    rdfs:label "has sample" .

cbm:hasDataDictionary a rdf:Property ;
    rdfs:domain dcat:Distribution ;
    rdfs:range cbm:DataDictionary ;
    rdfs:label "has data dictionary" .

cbm:hasField a rdf:Property ;
    rdfs:domain cbm:DataDictionary ;
    rdfs:range cbm:Field ;
    rdfs:label "has field" .

cbm:sampleValue a rdf:Property ;
    rdfs:domain cbm:Field ;
    rdfs:range rdf:Literal ;
    rdfs:label "sample value" .
```

Listing. 1.1. Vocabulary for DCAT Extension

Table. 1. Properties of the DCAT extension for content-based metadata.

Property	Domain	Range	Description
cbm:isSampleOf	cbm:Sample	dcat:Dataset	Relates example or sample data to a dataset.
cbm:hasSample	dcat:Dataset	cbm:Sample	Relates a dataset to example or sample data.
cbm:hasData-Dictionary	dcat:Distri-bution	cbm:Data-Dictionary	Links a distribution to its structural metadata.
cbm:hasField	cbm:Data-Dictionary	cbm:Field	Lists the fields described by the data dictionary.
cbm:sampleValue	cbm:Field	Literal	Provides a representative value for the field.

```
    dcterms:title "Urban Biodiversity Survey" ;
    cbm:hasSample ex:biodiversity-sample ;
    dcat:distribution ex:biodiversity-distribution .

ex:biodiversity-sample a cbm:Sample ;
    schema:description "Example data from 3 surveyed gardens." ;
    schema:algorithm "Hierarchical Clustering" ;
    cbm:isSampleOf ex:urban-biodiversity-dataset.

ex:biodiversity-distribution a dcat:Distribution ;
    dcterms:format "CSV" ;
    cbm:hasDataDictionary ex:bio-data-dictionary .

ex:bio-data-dictionary a cbm:DataDictionary ;
    cbm:hasField ex:field-species, ex:field-abundance .

ex:field-species a cbm:Field ;
    schema:name "species" ;
    dcterms:title "Species Name" ;
    dcterms:type "string" ;
    dcterms:description "Scientific name of observed species" ;
    cbm:sampleValue "Bombus terrestris" .

ex:field-abundance a cbm:Field ;
    schema:name "abundance" ;
    dcterms:title "Abundance" ;
    dcterms:type "integer" ;
    dcterms:description "Count of individuals observed" ;
    cbm:sampleValue "7" .
```

Listing 1.2. Example of Extended Dataset Metadata

3.4. Mapping DCAT-AP extension to Existing Catalog Architectures

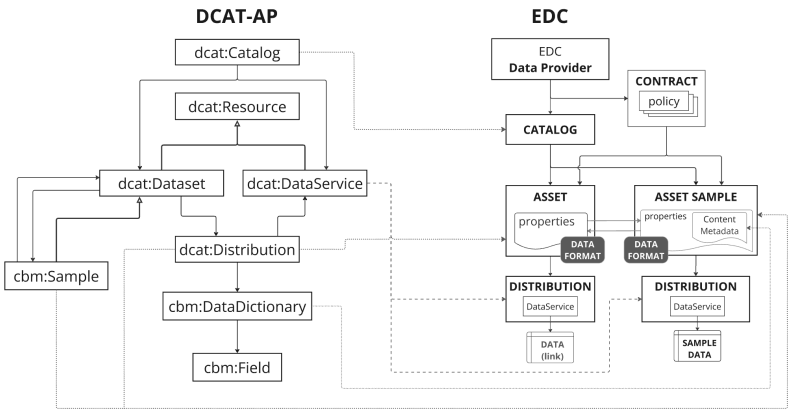
The proposed cbmDCAT-AP extension is independent of any specific data-space implementation. While this work uses Eclipse Dataspace Components (EDC) as its architectural basis, the extension can also be applied to other catalog and connector models. This section outlines its alignment with EDC, the International Data Spaces (IDSA), and CKAN. These mappings demonstrate how the proposed classes and properties can be integrated into diverse data ecosystems without significant architectural changes.

Figure 2 summarizes the correspondence between cbmDCAT-AP and the native entities of these three technologies. The mappings address the catalog, dataset, distribution, access service, representative samples, data dictionary, and field-level metadata concepts.

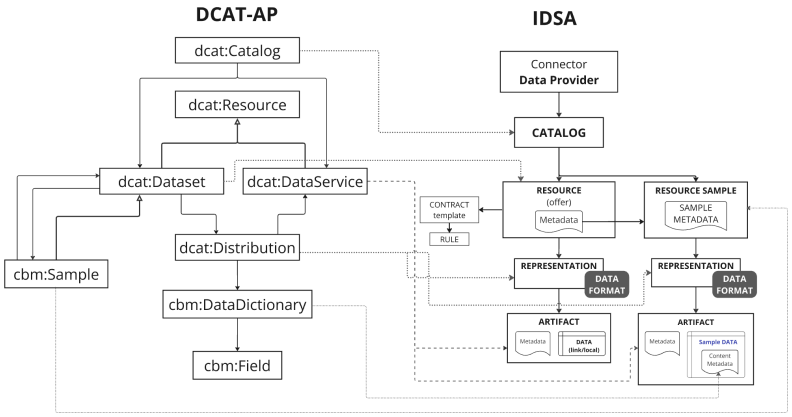
As shown in Figure 2a, the DCAT-AP extension entities can be mapped to the Eclipse Dataspace Components (EDC) [16] in a straightforward way. An EDC Data Provider has a ”catalog” that corresponds to the *dcat:Catalog* class. This catalog contains multiple ”assets” whose access is defined by the policies in the contract. Since this authorization is defined at the asset level, it is necessary to have the original and sample data as separate assets, corresponding to two different DCAT datasets: the original one with class *dcat:Dataset* and the sample one mapped to a *cbm:Sample* instance with open access policies. Since in EDC the data format is defined at the asset level, there is only one *dcat:Distribution* instance per dataset, and we can store the content metadata, equivalent to the *cbm:DataDictionary* within the asset metadata. The ”distribution” entity in EDC is equivalent to the *dcat:DataService* and defines the data access methods.

The data model from the IDSA data provider and its correspondence to the DCAT-AP extension classes are depicted in Figure 2b. The ”catalog” provided by an IDSA data provider aligns with *dcat:Catalog*, and the resources offered within the catalog correspond to *dcat:Dataset*. Access to these resources is regulated by contract templates and usage rules. Similar to EDC, access conditions are specified at the resource level. If the original data and the representative sample require distinct access policies, they may be modeled as separate resources. The protected resource corresponds to the original *dcat:Dataset*, whereas the resource providing access to the representative data corresponds to *cbm:Sample*. An IDSA resource may possess multiple representations, each associated with a specific data format. These representations correspond to *dcat:Distribution*. The artifacts linked to a representation specify the actual data object or access endpoint and are thus related to the function of *dcat:DataService*. The representative sample may be implemented as an artifact containing locally stored sample data or as an artifact referencing an external access endpoint. The associated content-level metadata, represented by *cbm:DataDictionary* and *cbm:Field*, can be stored alongside the artifact or resource metadata. This enables consumers to retrieve field descriptions and representative values when examining the sample resource, while the original data remain subject to the applicable contract rules.

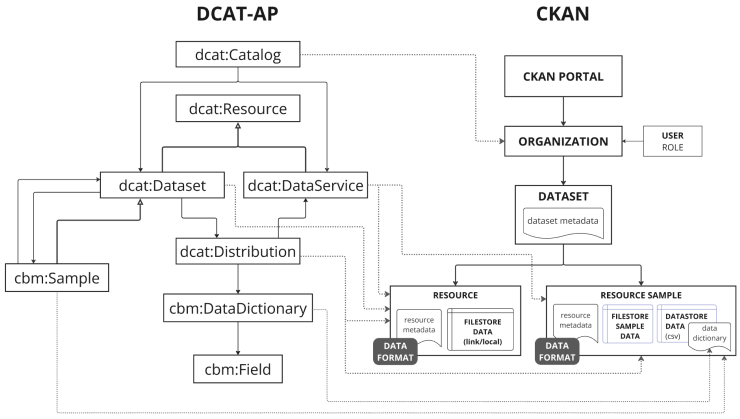
CKAN necessitates a distinct mapping, as it primarily functions as an open-data portal with datasets generally available for public read access. The mapping to CKAN can be seen in Figure 2c. The most analogous entity to *dcat:Catalog* is a CKAN organization, given that each dataset is linked to an organization responsible for its publication and metadata management. A CKAN dataset corresponds to the catalog record that describes the data resource. Both the original data and the representative sample may be published as distinct CKAN resources within the same dataset record. In contrast to the



(a) Mapping of the DCAT-AP extension to the EDC data model.



(b) Mapping of the DCAT-AP extension to the IDSA data model.



(c) Mapping of the DCAT-AP extension to the CKAN data model.

Fig. 2. Mapping of the DCAT-AP extension EDC, IDSA, and CKAN data models.

EDC and IDSA mappings, these resources do not require representation as separately governed datasets, since both may be publicly accessible under the license of the associated open dataset. A CKAN resource specifies its format and access location, and may either contain locally hosted data or reference an external URL. As such, it encompasses aspects corresponding to *dcat:Distribution* and *dcat:DataService*. A resource containing the original data serves as the accessible distribution of the dataset, while an additional resource containing representative rows can be linked to *cbm:Sample*. The sample resource may be stored in the CKAN filestore. For supported tabular formats, such as CSV, it can also be imported into the CKAN datastore, which offers structured storage, filtering, querying, and preview capabilities. The datastore's data-dictionary mechanism serves as an appropriate implementation location for *cbm:DataDictionary* and *cbm:Field*, as it already accommodates metadata including field names, types, and descriptions. Representative values may be provided through *cbm:sampleValue* or derived from the associated sample resource.

4. Including cbmDCAT-AP in data spaces for improving discoverability

This section introduces a solution to enhance dataset discovery in federated dataspaces by integrating cbmDCAT-AP for enriched metadata and representative data sampling. The proposed approach combines architectural design, semantic enrichment, and content-aware reduction techniques to facilitate precise and efficient data search while preserving data sovereignty. The overall data space architecture and its integration with cbmDCAT-AP in a federated environment are first described. The incorporation of content-based metadata at both resource and data levels is then detailed, followed by the methodology for generating representative data samples using clustering-based reduction.

4.1. Data Space Architecture

Figure 3 presents the HELEADE dataspace architecture that implements the cbmDCAT-AP extension within a federated environment, adhering to the principles of the International Data Spaces Association and utilizing Eclipse Dataspace Components. This architecture aims to improve dataset discoverability while maintaining strict data sovereignty, enabling providers to share enriched, content-level metadata and representative samples without providing full access to entire datasets. The system incorporates catalog enrichment, semantic indexing, and policy-based access control into an integrated workflow that facilitates data discovery as an initial phase of federation.

On the provider side, each participant operates an EDC Connector that manages local datasets as assets. The original dataset is registered as a protected asset governed by contractual policies regulating access and usage. Simultaneously, a reduced version of the dataset, generated using the clustering-based sampling methodology described below, is registered as a separate asset. This reduced asset contains a limited number of representative rows and is generally associated with more permissive access policies. Both the original dataset and its sample are described using DCAT-AP metadata, further enriched with the cbmDCAT-AP extension, which provides field-level descriptions, data dictionaries, and representative sample values. This distinction between the full dataset and its sample allows for preliminary relevance evaluation while maintaining provider control.

The catalog layer is central to the architecture. Traditional dataset-level metadata are enhanced with structural and content-based descriptors, such as explicit links between datasets and their samples, structured data dictionaries for field descriptions, and representative values that provide semantic context. Despite these enhancements, full compatibility with DCAT-AP harvesting mechanisms is maintained, ensuring interoperability with existing portals and dataspace infrastructures. The extension is additive and backward compatible, enabling gradual adoption without disrupting established metadata ecosystems.

A semantic indexing component processes the available textual and structural information, including dataset content (either original or reduced), column names, and column descriptions when available. These elements are transformed into vector embeddings to enable similarity-based retrieval. To ensure scalability and performance, indexing is primarily performed over the reduced samples rather than the full datasets. This significantly decreases computational costs and indexing time while maintaining competitive retrieval quality, as demonstrated in the experimental evaluation. The architecture therefore balances representativeness and efficiency, which is essential in large-scale federated environments.

From the consumer perspective, interaction with the HELEADE dataspace begins with federated catalog queries or semantic search requests. The search engine retrieves candidate datasets based on embedding similarity, utilizing the enriched content-level metadata. Consumers can review descriptive metadata, field information, and representative sample values before determining dataset suitability. If a dataset is considered relevant, the consumer initiates a contract negotiation process through the EDC framework to request controlled access to the original asset. Access is granted only under agreed policies, thereby preserving sovereignty and compliance requirements.

Within the proposed architecture, dataset discovery, authorization, and data consumption are treated as distinct stages. The enriched catalog and representative sample enable consumers to evaluate the suitability of a dataset. However, these elements do not automatically provide access to the protected asset. After selecting a relevant dataset, the consumer initiates contract negotiation via the EDC connector. The resulting contract agreement delineates the conditions under which the consumer is authorized to access and utilize the asset, in accordance with the provider's defined policies.

The contract agreement operates independently of the technical delivery mechanism. Consequently, the proposed approach does not require datasets to be downloaded as files. Depending on the distribution and access services configured by the provider, data may be delivered as a downloadable resource, accessed via an API or query endpoint, or consumed through a streaming interface. These access mechanisms are represented by the standard DCAT/DCAT-AP concepts of *dcat:Distribution*, *dcat:DataService*, and *dcat:accessService*. In contrast, cbmDCAT-AP supplies the content-level information necessary to assess dataset relevance prior to requesting access. The representative sample is managed as a distinct asset and may be subject to a more permissive access policy than the original dataset. Therefore, access to the sample does not confer access to the complete dataset. Rather, the sample enables customers to make informed decisions prior to initiating contract negotiation for the protected asset.

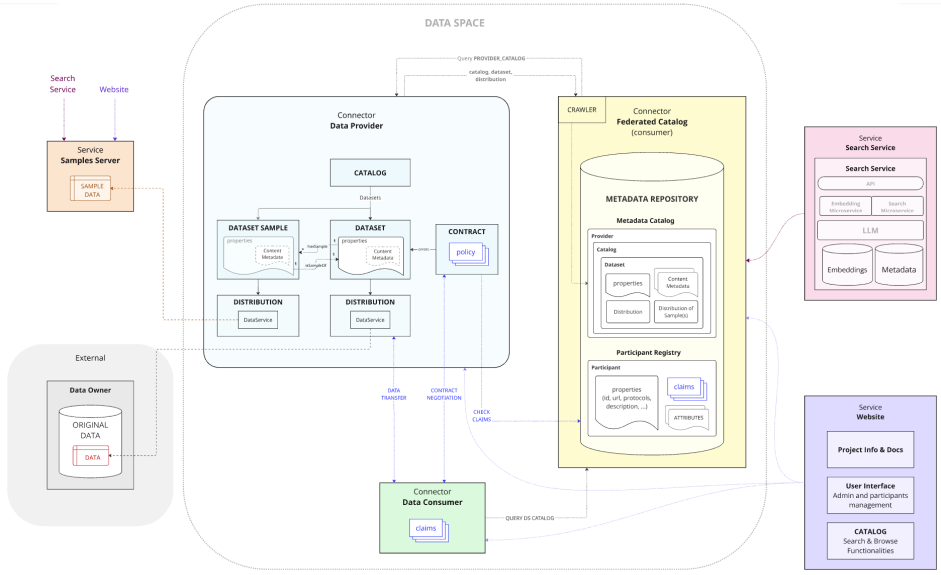


Fig. 3. HELEADE dataspace architecture.

4.2. Content-based Metadata

Our solution integrates content-based metadata at both resource and data content levels. At the resource level, metadata follows the IDS standard with fields like title, description, and license added by data providers. This is automatically enhanced with content-based metadata derived from the data itself, formatted using the Datapackage standard[9] from the Frictionless Data project [23], and stored locally at the artifact level alongside a representative data sample. When consumers request the sample resource, they simultaneously receive this enriched metadata. A data sampling mechanism is also integrated in order to obtain the samples. This methodology is described below.

4.3. Clustering-based Reduction for Sample Generation

For tabular datasets, we propose a clustering-based reduction method for generating compact samples that retain semantically diverse and representative rows from the original table. The method is evaluated against two baselines: random row sampling and sequential selection of the first ten rows.

Given a target sample size k, each eligible table row is transformed into a numerical representation using word embeddings. K-means clustering is then applied to partition the resulting vector space into k groups of semantically similar rows. For each cluster, the original row whose vector is closest to the cluster centroid is selected. The selected rows approximate the central tendencies of their respective clusters while preserving the original tabular representation. The output is therefore a reduced table containing k actual rows from the source dataset, rather than synthetic centroid values.

[9] https://datapackage.org/standard/data-package/

The use of actual rows facilitates interpretation by prospective data consumers, while the clustering process promotes coverage of different regions of the semantic space. However, selecting actual rows also means that the reduction method must not be interpreted as an anonymisation or privacy mechanism.

The clustering-based reduction method operates exclusively on data authorized by the provider for sample publication. The clustering algorithm does not assess whether a value is sensitive, restricted, or context-dependent. Instead, sample generation is preceded by a policy-aware preprocessing stage established by the data provider.

In this stage, fields are classified as publishable, conditionally publishable, or excluded for the sample. Excluded fields are removed prior to the generation of embeddings and clusters. Conditionally publishable fields may undergo transformation through masking, generalisation, aggregation, or suppression. For instance, exact geographic coordinates can be replaced with administrative areas, timestamps can be reduced to month-level granularity, and direct identifiers such as names, email addresses, or identification numbers can be removed. Clustering is then performed on the filtered and transformed data, ensuring that prohibited values neither influence nor appear in the generated sample.

Once representative rows are selected, the resulting sample is validated against the same publication policy before registration as a separate catalog asset. Consequently, the publication of a representative sample remains under the provider's control and is independent of the access policy governing the complete dataset. If a policy-compliant sample cannot be produced without incurring unacceptable disclosure risk, the provider may choose to publish only the data dictionary and omit sample values.

The current implementation assumes that the provider specifies the applicable field-level restrictions. Automatic identification of sensitive and context-dependent values constitutes a domain-specific challenge that may require semantic annotations, personal-data detection techniques, and risk assessment mechanisms. Automated inference and enforcement of such restrictions are beyond the scope of this work and will be addressed in future research.

To illustrate how the algorithm works, consider an environmental dataset containing observations of protected species with the fields *species_name*, *observation_date*, *latitude*, *longitude*, *municipality*, *habitat_type*, and *observer_email*. The provider's publication policy classifies observer_email as prohibited and exact coordinates as restricted because they may reveal the location of endangered species. Before sample generation, observer_email is removed, latitude and longitude are replaced by municipality, and observation_date is generalised from an exact date to year and month.

Embeddings are then generated from the remaining policy-compliant fields. K-means partitions the rows into k groups according to their semantic similarity, and the row closest to each cluster centroid is selected as a representative observation. For $k = 3$, the resulting sample may contain representative records corresponding to three combinations of species, habitat, municipality, and observation period. The published rows contain no observer identifier or exact location. The sample therefore illustrates the dataset's semantic coverage without exposing the restricted values used in the original resource.

4.4. Scenario-based architectural trade-off analysis

To complement the component-level experiments, we conducted a lightweight scenario-based architectural evaluation inspired by the Architecture Tradeoff Analysis Method

(ATAM) [36]. ATAM relates architectural decisions to concrete quality-attribute scenarios and uses this relationship to identify sensitivity points, trade-off points, and risks. The present analysis is deliberately scoped to the business drivers and empirical evidence available from the HELEADE data space architecture implementation. It was conducted by the authors and did not include the independent facilitation and multi-stakeholder scenario elicitation and prioritisation of a complete ATAM evaluation, thus it should be interpreted as an ATAM-inspired analysis rather than as a full ATAM analysis.

Two main business drivers guide the architecture: (i) enabling consumers to assess dataset relevance before obtaining access to the complete asset, while preserving provider sovereignty; and (ii) supporting scalable catalog ingestion and semantic index updates in data-intensive federated environments. Table 2 maps these drivers to quality-attribute scenarios, architectural decisions, available evidence, and their associated compromises.

The analysis identifies five principal sensitivity points: (i) the number of representative rows k, (ii) the availability and quality of field descriptions, (iii) the transformations applied during policy-aware preprocessing, (iv) the number and update rate of cataloged datasets, as well as (v) the embedding model used by the semantic index.

The first of these is also a major trade-off point: increasing k may improve semantic coverage and retrieval effectiveness, but also increases sample-generation and indexing costs and enlarges the disclosure surface. Similarly, richer field descriptions can improve retrieval but require additional curation and storage, while more restrictive masking or generalisation reduces disclosure risk at the possible expense of semantic fidelity.

Other residual risks include "sample staleness" after updating source datasets, the omission of rare but relevant values when rows close to cluster centroids are selected, and dependence on provider-supplied field classifications. Conversely, the additive nature of cbmDCAT-AP and the policy separation between sample and original assets constitute architectural mitigations for compatibility and unauthorised access, respectively. Overall, the architecture should be understood as a configurable compromise among retrieval effectiveness, indexing cost, interoperability, and controlled disclosure rather than as a configuration that maximises all quality attributes simultaneously.

5. Experimental evaluation

In this section, we describe the comprehensive evaluation of our proposed cluster through intrinsic, extrinsic, and temporal assessments. The intrinsic assessment compared original tables with their reduced counterparts to measure information preservation. The extrinsic assessment evaluated practical utility by employing reduced tables in information retrieval tasks within our proposed data space architecture, serving as a data discovery process. We benchmarked our cluster reduction technique against simpler methods such as random row selection and first rows selection. Additionally, we analyzed indexing performance differences between original and reduced datasets. All these tests are performed in the environment of the HELEADE data space architecture described in Section 4.1.

5.1. Evaluation Dataset

Two datasets have been selected for the experiments, in accordance with the two scenarios proposed: (i) dataset discovery within the data space, and (ii) discovery and selection of

Table. 2. ATAM-inspired analysis of architectural quality attributes and trade-offs.

Quality attribute	**Scenario and desired response**	**Architectural decision**	**Evidence and corresponding compromise**
Discoverability and retrieval effectiveness	A consumer who cannot inspect the protected asset submits a content-oriented query, and relevant datasets should appear among the highest-ranked results.	Field-level metadata, representative samples, and embedding-based semantic indexing are incorporated into the catalog workflow.	Context 3 (see Section 5) generally provides the strongest retrieval effectiveness among the reduced configurations. Nevertheless, full-content indexing remains more effective, particularly for early-rank metrics, and field names without descriptions may introduce noise.
Indexing performance and scalability	A large number of datasets is added or updated, requiring the semantic index to be refreshed within an operationally acceptable time and computational budget.	The index is constructed from a fixed-size representative sample rather than from every row of the dataset.	End-to-end index construction is approximately $184\times$ faster for TData and $156\times$ faster for EnviDat when sample generation is included (see Section 5). This improvement is accompanied by a measurable loss of retrieval effectiveness and depends on the selected sample size.
Sovereignty and disclosure control	A consumer without a contract requests information for assessing dataset relevance but must not gain access to the protected asset.	The original dataset and its sample are registered as separately governed assets; sample generation is preceded and followed by policy-aware filtering and validation.	Access to the sample does not confer access to the original dataset. However, representative rows are actual records and do not constitute an anonymisation mechanism. Incorrect field classification or insufficient transformation may therefore create residual disclosure risks.
Interoperability and modifiability	Existing DCAT-AP harvesters and alternative catalog technologies must continue to process the catalog as the extension evolves.	cbmDCAT-AP is additive and backward compatible, with explicit mappings to EDC, IDSA, and CKAN.	Existing consumers can ignore the additional terms and continue processing standard DCAT-AP metadata. Conversely, exploiting the enriched metadata requires extension-aware consumers and platform-specific mappings, while separately governed sample assets increase lifecycle-management complexity.
Catalog freshness and availability	A provider updates a dataset or sample, and consumers should receive sufficiently current search results without accessing the provider's complete data.	Metadata and samples are harvested into a federated catalog and processed by a separate semantic indexing service.	Replication enables low-latency cross-provider search without centralising complete datasets. However, crawler and index-update intervals may produce stale descriptions, and the central search service can become a performance or availability bottleneck. These properties were not measured in the present evaluation.
Modifiability of semantic retrieval	A new domain, language, or embedding model must be supported without changing policy enforcement or the underlying asset model.	Semantic indexing is implemented as a service separated from the connector, contract, and metadata-representation components.	The separation localises changes to the retrieval subsystem. Nevertheless, replacing an embedding model requires re-indexing and may alter ranking behaviour, making model selection another source of performance variability.

open datasets from external open data portals for cataloguing and reuse within the data space.

For the first scenario, we selected a real-world dataset provided by a data space provider in the experiments conducted to evaluate the effectiveness of our table reduction methodology. The dataset comprises 100 tables extracted from a tourism open data portal from the Valencia Region in Spain, TData[10]. This collection was carefully selected to encompass a wide array of data structures and column types, providing a robust testing ground for assessing the adaptability and performance of the reduction approach.

For the second scenario, the real-world dataset comprises 242 tables extracted from an environmental research open data portal from Switzerland, EnviDat.[11] This collection was carefully selected to encompass a wide array of data structures and column types, providing a robust testing ground for assessing the adaptability and performance of the extension approach.

Table 3 offers an overview of the statistics of the evaluation datasets.

Characteristic	TData	EnviDat
Number of tables	100	242
Total number of rows	748,088	1,184,256
Total number of columns	1,926	5,946
Total number of text columns	1637	1,860
Avg. number of rows	7,480.88	4,893.62
Avg. number of columns	19.26	24.57
Avg. number of text columns	16.37	7.69
Max. number of rows	90,792	15,000
Max. number of columns	44	471
Max. number of text columns	43	469

Table. 3. Statistics of the collected datasets.

5.2. Intrinsic Assessment

To evaluate the effectiveness of the proposed table reduction approach, a comparison was conducted between the original tables and their reduced versions, using descriptive statistics as the basis for similarity assessment. This was an indicator of how well the reduced tables retained the statistical properties of the original data.

Seven descriptive statistics were calculated along each column: the mean, median, standard deviation, minimum, maximum, as well as 25th and 75th percentiles. These seven metrics were calculated for each of the columns, and the resulting values are stored in a vector representing each table, thus allowing us to quantitatively compare tables. The similarity between the original and reduced tables was determined by calculating the cosine similarity between these vectors, providing a metric of alignment between the two tables.

[10] https://tdata.dlsi.ua.es/
[11] https://www.envidat.ch/

We evaluated the table reduction approach across three distinct methods to benchmark its performance. In the first method, we applied the cluster reduction methodology. In the second method, we created a reduced table by randomly sampling rows from the original dataset. In the third method, we simply selected the first ten rows from the initial table, representing a sequential selection without methodological design.

For all three methods, we generated reduced tables with a standardized 10-row count. Cosine similarity analysis yielded matching scores of 0.933 (random reduction), 0.935 (first ten rows), and 0.934 (cluster reduction) for TData dataset, and 0.912 (random reduction), 0.911 (first ten rows), and 0.915 (cluster reduction). These nearly identical scores suggest our proposed reduction approach offers no significant advantage over simpler random or sequential selection methods when preserving statistical features from the original dataset. The minimal variation in similarity scores (0.003 maximum difference) indicates that further research is needed to develop methods that better preserve distinctive characteristics across original tables.

5.3. Extrinsic Assessment

To evaluate the table reduction utility for data discovery services, we conducted experiments comparing the performance of the tables against the original datasets using embeddings generated by the JinaAI[12] Spanish model, specifically the *jina-embeddings-v2-base-es* variant. This systematic comparison allowed us to thoroughly assess how well reduced tables perform relative to original datasets in information retrieval tasks.

The experimentation was carried out by first selecting randomly 10 rows from each original table to serve as a query in the experiments. Next, embeddings were created for both the original dataset and its three reduced versions as described in Section 4.3. Finally, search metrics were calculated using the queries against the original and sample tables.

We established three search contexts by generating indexes from distinct metadata components in our proposed content-based catalog. In *Context 1*, indexes incorporated only the table content. *Context 2* expanded this approach by including both table content and column names. *Context 3*, applicable to tables with available column descriptions, generated indexes using a comprehensive combination of column names, column descriptions and table content.

To evaluate retrieval quality, we employed two metrics: Mean Reciprocal Rank (MRR), which measures the position of the first relevant retrieved result, and Precision at k (P@k metric), which calculates the proportion of relevant results among the top k retrievals. These metrics provide an objective assessment of each approach's performance. In our results, we highlight statistically significant differences between the random and clustering reduction methods with asterisks, as these two approaches emerged as the most competitive. In the results tables, light shading indicates a statistically significant difference between the Cluster and Random reduction methods at $p < 0.05$, while dark shading indicates significance at $p < 0.01$.

Table 4 shows the search metrics results using reduction techniques in the three different contexts for the TData dataset. The results reveal interesting trends and variations across contexts and methods.

[12] https://jina.ai/embeddings/.

In Context 1, Cluster reduction and Random reduction produce comparable results, both slightly below those of the original dataset yet clearly outperforming First rows reduction. First rows reduction demonstrates the weakest performance, with an MRR of 0.6392 and a P@1 of 0.5426, underscoring its limitations in this setting. In Context 2, the Cluster approach yields the highest reduction scores, surpassing both Random reduction and First rows reduction. Notably, First rows reduction again performs the weakest, with an MRR of 0.6311 and a P@1 of 0.5362, revealing a distinct performance gap relative to Cluster reduction. In Context 3, Cluster reduction attains the highest MRR and P@1 scores, at 0.7856 and 0.7043, respectively, demonstrating strong effectiveness. Both Random reduction and First rows reduction also exhibit substantial improvements in Context 3 compared to the previous contexts. Although their performance remains slightly below that of Cluster reduction, both methods achieve high results, with MRR scores of 0.7508 and 0.6720, respectively.

Overall, the Cluster reduction achieves the highest scores among the reduction methods in most evaluated settings. However, the improvements over Random reduction are generally modest and are not statistically significant for every metric or context. The results therefore indicate a context-dependent advantage rather than a systematic superiority of clustering over random sampling.

Table. 4. Search metrics combining different contexts and reduction techniques for the TData dataset.

	Reduction	**MRR**	**P@1**	**P@5**	**P@10**
Context 1	Original data	0.8875	0.8340	0.9919	0.9979
	Cluster	0.7327	0.6521	0.8415	0.9511
	Random	0.7100	0.6351	0.8043	0.9191
	First rows	0.6392	0.5426	0.7723	0.8766
Context 2	Original data	0.7304	0.6479	0.8777	0.9447
	Cluster	0.6668	0.5713	0.7979	0.9266
	Random	0.6537	0.5606	0.7840	0.9160
	First rows	0.6311	0.5362	0.7755	0.8745
Context 3	Original data	0.8974	0.8473	0.9839	0.9979
	Cluster	0.7856	0.7043	0.9161	0.9613
	Random	0.7508	0.6720	0.8624	0.9613
	First rows	0.7279	0.6376	0.8710	0.9290

For the second scenario, the search metrics in differente contexts can be seen in Table 5. Similar to the TData dataset, these metrics demonstrate notable outcomes across various contexts and methodologies.

In Context 1 the best performing reduction is Cluster reduction. However, Random reduction and First rows reduction perform significantly worse in this context, with First rows reduction being the weakest, with its MRR dropping to 0.7648 and its P@1 dropping to 0.6963, indicating its limitations in this context. In Context 2, the situation is similar: Cluster approach also provides the best reduction scores, with Random reduction and First

rows reduction performing the worst. One striking fact about this context is that, despite including column names in the index, results are worse for all methods, so we tested adding more metadata in Context 3. In this context, Cluster Random has the best metrics once again. Both Random and First rows reduction also show a substantial improvement in Context 3 compared to the other contexts. Although their performance is slightly lower than that of the Cluster reduction, they achieve high results, with MRR and P@1 scores of 0.9130 and 0.8595, respectively.

These findings also highlight the variability in method effectiveness depending on the specific context. Having analysed the contexts separately, we turn to a general observation of the table. As we can see, the best performance is offered by the original dataset in Context 1, where only the content of the table is taken into account. The original dataset in Context 3 offers almost similar metrics to Context 1 but slightly worse than Context 1, except for P@10.

Across both datasets, Cluster reduction obtains the highest scores among the reduction methods in most contexts, especially when content is combined with richer metadata. Nevertheless, its advantage over Random reduction is usually limited and is not statistically significant across all metrics and settings. The results therefore identify clustering as a competitive and comparatively robust reduction strategy, rather than as a universally superior alternative to random sampling. First rows reduction generally provides the weakest performance.

Table. 5. Search metrics combining different contexts and reduction techniques for the EnviDat dataset.

Ctx	Reduction	MRR	P@1	P@5	P@10
1	Original data	0.9285	0.8817	0.9871	0.9975
	Cluster	0.8929	0.8357	0.9639	0.9892
	Random	0.8726	0.8195	0.9369	0.9718
	First rows	0.7648	0.6963	0.8548	0.9220
2	Original data	0.8476	0.7712	0.9452	0.9826
	Cluster	0.8196	0.7411	0.9220	0.9685
	Random	0.8067	0.7290	0.9050	0.9585
	First rows	0.7182	0.6373	0.8241	0.8963
3	Original data	0.9262	0.8779	0.9869	0.9977
	Cluster	0.9130	0.8595	0.9815	0.9964
	Random	0.8997	0.8491	0.9685	0.9887
	First rows	0.8345	0.7649	0.9234	0.9671

To ensure a realistic evaluation of the data discovery service in our data space architecture, we considered varying data quality levels, particularly regarding sparsity. Therefore, next experiments introduced different percentages of null values into the indexed tables, reflecting the common scenario where tables often contain incomplete rows or missing data. This approach allowed us to assess the service's performance under conditions that mirror the imperfect data typically encountered in real-world scenarios. This experiment considers only the TData dataset, due to significant differences between the two scenarios. The Envidat dataset represents open data onboarding, characterized by unrestricted

access and the absence of data space policies; therefore, artificial simulation of null values is unnecessary. In practice, external datasets are federated in their entirety after selection, and sparsity tests are reserved for datasets already governed within the data space, as exemplified by TData.

Table 6 presents different levels of null values that were introduced into the original tables for Context 1, Context 2, and Context 3, respectively. In these three contexts, four levels of null value percentages were established: 10%, 25%, 50%, and 75% of missing values. In all these experiments, it is evident that the method based on the First rows consistently yields the worst results, just as it did when working without null values. Therefore, we will focus on discussing the results of comparing the Cluster reduction with the Random reduction.

Table 6 presents the impact of introducing varying levels of null values (10%, 25%, 50%, and 75%) into the original tables across all three contexts. Consistent with our previous findings, First rows reduction consistently produces the poorest results, mirroring its performance in datasets without null values. Given this consistent underperformance, our analysis will focus primarily on comparing Cluster reduction with Random reduction methods.

In *Context 1*, the comparison between Cluster and Random reduction shows similar performance trends across all levels of null values. At 10% null values, Cluster reduction slightly outperforms Random reduction in all the metrics. At 50% and 75% null values, Cluster reduction maintains a narrow advantage over Random reduction. The differences are more clear at 25% null values, where Cluster reduction surpasses Random reduction more clearly in MRR (0.7807 vs. 0.7288) and P@1 (0.6840 vs. 0.6351). This suggests that Cluster reduction has a good behavior in medium sparse datasets. This trend is consistent across the other two contexts. In *Context 2*, it is evident that in situations with a medium number of null values (25%), the Cluster reduction method clearly outperforms the Random reduction across all metrics. The same pattern is observed in *Context 3*, where the cluster-based method shows a significant advantage over the random method. This suggests that in real-world scenarios, where tables often contain a medium proportion of null values, the proposed Cluster reduction can be significantly more effective than the Random reduction.

5.4. Indexing Performance

Indexing speed is a critical performance metric for evaluating search system efficiency, particularly in dynamic environments managing large datasets like data spaces. This paper compares the indexing times of two data catalogs employing distinct approaches: Catalog A indexes complete data tables, calculating embeddings for each row to ensure comprehensive dataset representation, while Catalog B utilizes our proposed Cluster reduction method (detailed in Section 4.3) to generate embeddings from reduced table versions. This latter approach was specifically developed to address the performance limitations observed when processing full datasets.

In our experimental setup for the TData dataset, Catalog A processed datasets averaging 7,480 rows and 19 columns, while Catalog B utilized the same tables reduced to just 10 rows per table through our clustering approach. For TData, Catalog B completed end-to-end index construction in 32 seconds, comprising 20 seconds for sample generation and 12 seconds for indexing, compared with 5,891 seconds for full-content indexing. This

Table. 6. Search performance under different percentages of null values.

(a) Context 1: table content.

N%	Reduction	MRR	P@1	P@5	P@10
10%	Original data	0.8803	0.8198	0.9896	1.0000
	Cluster	0.7172	0.6250	0.8510	0.9615
	Random	0.6990	0.6063	0.8323	0.9302
	First rows	0.6380	0.5333	0.7792	0.8927
25%	Original data	0.8656	0.7968	0.9684	0.9979
	Cluster	0.7329	0.6347	0.8663	0.9716
	Random	0.6847	0.5916	0.8095	0.9368
	First rows	0.6264	0.5200	0.7758	0.8979
50%	Original data	0.8054	0.7198	0.9344	0.9958
	Cluster	0.7043	0.6010	0.8521	0.9688
	Random	0.6543	0.5427	0.8031	0.9115
	First rows	0.6131	0.5146	0.7510	0.8656
75%	Original data	0.6900	0.5688	0.8646	0.9667
	Cluster	0.5919	0.4677	0.7750	0.9052
	Random	0.5506	0.4146	0.7490	0.8677
	First rows	0.4907	0.3740	0.6479	0.8063

(b) Context 2: table content and column names.

N%	Reduction	MRR	P@1	P@5	P@10
10%	Original data	0.7403	0.6438	0.8865	0.9865
	Cluster	0.6633	0.5594	0.8073	0.9458
	Random	0.6414	0.5427	0.7885	0.9104
	First rows	0.6047	0.4833	0.7646	0.8938
25%	Original data	0.7052	0.5874	0.8716	0.9937
	Cluster	0.6542	0.5337	0.8232	0.9463
	Random	0.6158	0.5032	0.7726	0.9032
	First rows	0.6031	0.4853	0.7674	0.9053
50%	Original data	0.6410	0.5125	0.8156	0.9531
	Cluster	0.6033	0.4688	0.7938	0.9281
	Random	0.5540	0.4292	0.7208	0.8542
	First rows	0.5461	0.4302	0.6979	0.8583
75%	Original data	0.5123	0.3698	0.7052	0.9000
	Cluster	0.4932	0.3594	0.6625	0.8333
	Random	0.4238	0.2854	0.6021	0.7729
	First rows	0.4229	0.2896	0.6042	0.7438

(c) Context 3: table content, column names, and column descriptions.

N%	Reduction	MRR	P@1	P@5	P@10
10%	Original data	0.8932	0.8368	0.9905	1.0000
	Cluster	0.7688	0.6726	0.9137	0.9842
	Random	0.7429	0.6537	0.8811	0.9695
	First rows	0.7286	0.6379	0.8463	0.9526
25%	Original data	0.8811	0.8180	0.9723	0.9979
	Cluster	0.7807	0.6840	0.9298	0.9851
	Random	0.7288	0.6351	0.8745	0.9691
	First rows	0.7257	0.6287	0.8713	0.9532
50%	Original data	0.8282	0.7526	0.9421	0.9958
	Cluster	0.7527	0.6484	0.9032	0.9853
	Random	0.7083	0.6116	0.8379	0.9526
	First rows	0.7203	0.6211	0.8495	0.9495
75%	Original data	0.7319	0.6211	0.8916	0.9800
	Cluster	0.6668	0.5568	0.8168	0.9495
	Random	0.6523	0.5389	0.8116	0.9316
	First rows	0.6594	0.5432	0.8211	0.9442

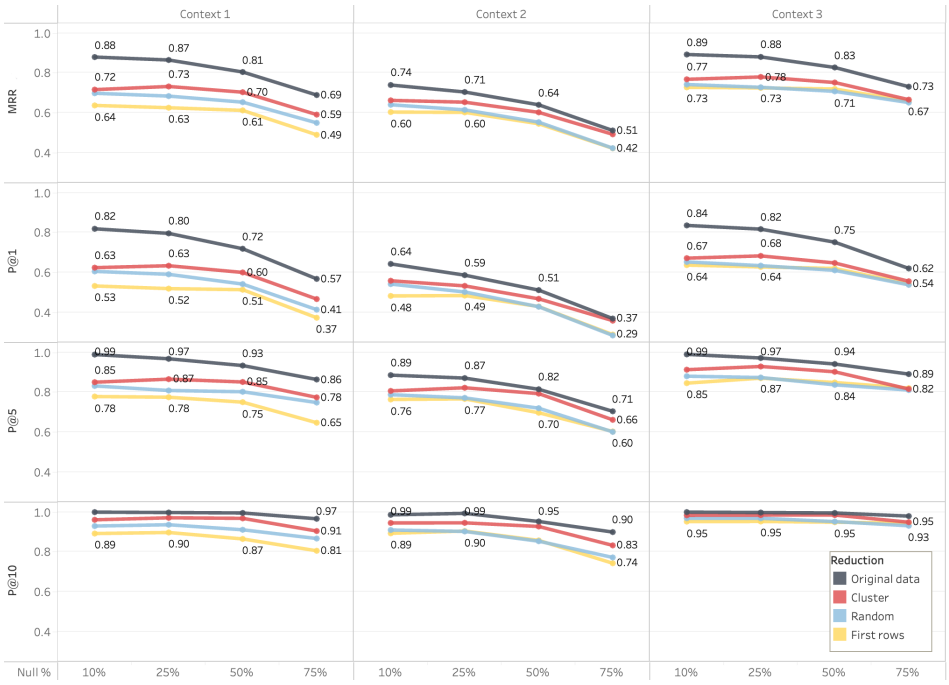


Fig. 4. Search metrics for different null percentages (N%) with Context 1, 2, and 3. A consistent colour scheme has been used to facilitate the interpretation of the results: black represents Original data, red corresponds to reduction by Clustering, blue indicates Random reduction and yellow represents First rows of the dataset.

corresponds to an approximately 184-fold reduction in processing time. For EnviDat, the corresponding average times were 0.327 seconds per table (0.122 seconds for reduction and 0.205 seconds for indexing) compared with 51 seconds for the full-content configuration, corresponding to an approximately 156-fold improvement when both stages are included.

These efficiency gains are not cost-free. In Context 3, the clustered ten-row samples retained 87.5% of the full-content MRR and 83.1% of P@1 for TData, whereas the corresponding values for EnviDat were 98.6% and 97.9%. The reported P@10 values were retained at 96.3% and 99.9%, respectively. Full-content indexing therefore remains preferable when maximum early-rank retrieval effectiveness is the principal requirement and the complete content can be processed under the applicable policies. In contrast, sample-based indexing provides a more appropriate operating point when index-update latency, computational cost, or controlled disclosure is the dominant concern.

The sample size k=10 is consequently an architectural sensitivity point rather than a universally optimal setting. Increasing k may improve content coverage and retrieval effectiveness, but it also increases sample-generation cost, indexing time, storage requirements, and the amount of content disclosed. Determining k adaptively according to dataset size, heterogeneity, and disclosure policy remains an important direction for future work.

5.5. Ablation-style analysis of architectural mechanisms

The experiments in Tables 4 and 5 permit an incremental ablation-style analysis along two dimensions. The first concerns catalog representation: Context 1 uses table content only, Context 2 adds column names, and Context 3 further adds column descriptions. The second concerns sample generation: the full-content index is compared with clustered, random, and first-row samples. Table 7 reports the absolute changes in MRR and P@1 for the principal contrasts. For the metadata comparisons, the clustered sample is held constant and positive values indicate an improvement attributable to the added mechanism.

The results indicate that the most influential enrichment mechanism is not the mere presence of additional metadata, but the availability of meaningful field descriptions. Adding names without descriptions reduces retrieval effectiveness in both datasets, whereas adding descriptions more than compensates for this loss. Metadata completeness and descriptive quality should therefore be treated as sensitivity points of the architecture.

The sampling results lead to a different interpretation. Reducing each table to ten rows is responsible for the large improvement in indexing efficiency, while clustering contributes a smaller but consistent retrieval advantage over random sampling. Thus, representative sampling is an essential scalability and controlled-disclosure mechanism, whereas clustering is a supportive optimisation that improves the quality of the reduced representation. Sequential first-row selection is the least suitable alternative because it combines the information loss associated with reduction with a potentially biased sample.

Not every layer shown in Figure 3 can be meaningfully removed while retaining the same system objectives. The connector and policy layer, the separation of protected and sample assets, and policy-aware preprocessing are architectural prerequisites for sovereignty-preserving discovery rather than ranking mechanisms. The semantic indexing service is similarly required to execute the evaluated retrieval process, while the user-interface layer is supportive and does not affect the reported ranking metrics. These components were held fixed in the present experiments.

Table. 7. Incremental ablation-style analysis of catalog representation and sample-generation mechanisms. Absolute changes are reported; positive values favour the added mechanism.

Contrast	**TData ΔMRR / ΔP@1**	**EnviDat ΔMRR / ΔP@1**	**Interpretation**
Add field names: Context 2 − Context 1	$-0.0659/-0.0808$	$-0.0733/-0.0946$	Field names alone do not improve retrieval and may introduce ambiguous or noisy tokens.
Add field descriptions: Context 3 − Context 2	$+0.1188/+0.1330$	$+0.0934/+0.1184$	Descriptions provide the largest positive marginal contribution among the tested metadata elements.
Add names and descriptions: Context 3 − Context 1	$+0.0529/+0.0522$	$+0.0201/+0.0238$	Complete field-level enrichment produces a positive net effect over content-only indexing.
Clustered rather than random sample in Context 3	$+0.0348/+0.0323$	$+0.0133/+0.0104$	Clustering contributes consistently, although its marginal effect is smaller than that of descriptive metadata.
Clustered sample rather than full content in Context 3	$-0.1118/-0.1430$	$-0.0132/-0.0184$	Reduction substantially lowers processing cost but sacrifices some early-rank effectiveness, particularly for TData.

The analysis should consequently be interpreted as an ablation of the representation and sampling mechanisms, rather than as a complete causal decomposition of every architectural layer. In particular, the evaluation does not include a standard index for only DCAT-AP, an only data-dictionary configuration, alternative embedding models, or an ablation of the policy transformations. These configurations will be considered in future work to quantify the independent and interacting contributions of catalog-level metadata, field-level metadata, representative content, and semantic indexing.

6. Conclusions

This paper addresses the persistent limitations of discovery mechanisms in federated data spaces (only based on high-level descriptive metadata) by introducing cbmDCAT-AP, a content-based extension of DCAT-AP designed to enhance dataset discoverability while maintaining data sovereignty. Traditional catalog approaches rely predominantly on high-level descriptive metadata, which is often inadequate for accurately assessing dataset relevance, especially in heterogeneous and large-scale environments. By incorporating structured field-level descriptions, data dictionaries, and representative sample values into the metadata layer, cbmDCAT-AP facilitates a more expressive and semantically rich representation of datasets without exposing full data content.

The proposed cbmDCAT-AP extension has been validated in both discovery scenarios in data spaces addressed in this paper: internal sovereign dataset discovery and external open data onboarding. These results demonstrate its versatility as a content-aware cataloguing mechanism for federated data spaces.

A central contribution of this paper is the integration of metadata enrichment with a clustering-based dataset reduction strategy to enable scalable semantic indexing. Experimental results indicate a substantial improvement in indexing efficiency, achieving speedups of several orders of magnitude when operating on reduced samples, while maintaining competitive retrieval performance across multiple search contexts and different scenarios. Intrinsic statistical similarity is comparable across the evaluated reduction strategies. In the information retrieval experiments, clustering generally outperforms sequential sampling and obtains modest, context-dependent improvements over random sampling, particularly when richer metadata or moderate levels of missing values are considered. However, these differences are not statistically significant in every setting. Therefore, the main benefit of the clustering-based approach lies in combining competitive retrieval quality with semantically informed sampling and substantial indexing-efficiency gains, rather than in providing universal superiority over simpler baselines.

Conceptually, the proposed approach helps mitigate the tension described by Arrow's information paradox by enabling meaningful pre-access evaluation of dataset content without compromising sovereignty constraints of data spaces. By exposing controlled, representative information rather than raw datasets, providers can enhance transparency, trust, and interoperability while maintaining governance and compliance requirements.

The architectural and ablation analyses further show that no single configuration dominates across all quality attributes. Full-content indexing remains preferable when maximum early-rank retrieval effectiveness is required and processing of the complete dataset is permitted. In contrast, combining policy-compliant clustered samples with field descriptions provides a balanced configuration for high-throughput, sovereignty-constrained catalogs. This operating point is sensitive to sample size, metadata quality, disclosure transformations, dataset-update frequency, and embedding-model selection. Future work will therefore include stakeholder-led architectural evaluation and a factorial ablation of these sensitivity points.

Furthermore, future research directions include exploring adaptive sampling strategies tailored to various data types and quality conditions, conducting large-scale validation in operational data spaces, and refining reduction techniques to better preserve distinctive dataset characteristics. Further investigation into integrating multimodal embeddings and domain-specific ontologies may also enhance semantic alignment across heterogeneous providers. Overall, cbmDCAT-AP and its supporting architecture constitute a practical and scalable advancement toward content-aware, sovereignty-preserving data discovery in next-generation federated data ecosystems.

Acknowledgements. This research was funded by the Spanish Ministry for Digital Transformation and the European Union NextGenerationEU, under HELEADE (TSI-100121-2024-24) and CLIO (TSI-100130-2024-69) projects. Also, part of the research was partially funded by AFFAIR project (PID2025-174636OB-I00), funded by MICIU/AEI/10.13039/501100011033.

References

1. Abedjan, Z., Esmailoghli, M., Galhotra, S.: Data discovery in data lakes: Operations, indexes, systems. Proceedings of the VLDB Endowment 18(12), 5455–5459 (2025), https://doi.org/10.14778/3750601.3750694

2. Attard, J., Orlandi, F., Scerri, S., Auer, S.: A systematic review of open government data initiatives. Government information quarterly 32(4), 399–418 (2015), https://doi.org/10.1016/j.giq.2015.07.006
3. Azcoitia, S.A., Laoutaris, N.: A survey of data marketplaces and their business models. ACM SIGMOD Record 51(3), 18–29 (2022), https://doi.org/10.1145/3572751.3572755
4. Bacco, M., Kocian, A., Chessa, S., Crivello, A., Barsocchi, P.: What are data spaces? systematic survey and future outlook. Data in Brief 57, 110969 (2024), https://doi.org/10.1016/j.dib.2024.110969
5. Berenguer, A., Alcaraz, O., Tomás, D., Mazón, J.N.: From research on data-intensive software to innovation in data spaces: a search service for tabular data. IEEE Software 41(3), 59–66 (2024), https://doi.org/10.1109/MS.2024.3359333
6. Bernhauer, D., Nečaskỳ, M., Škoda, P., Klímek, J., Skopal, T.: Open dataset discovery using context-enhanced similarity search. Knowledge and Information Systems 64(12), 3265–3291 (2022), https://doi.org/10.1007/s10115-022-01751-z
7. Bogatu, A., Fernandes, A.A.A., Paton, N.W., Konstantinou, N.: Dataset discovery in data lakes. In: 2020 IEEE 36th International Conference on Data Engineering (ICDE). pp. 709–720. IEEE (2020), https://doi.org/10.1109/ICDE48307.2020.00067
8. Chapman, A., Simperl, E., Koesten, L., Konstantinidis, G., Ibáñez, L., Kacprzak, E., Groth, P.: Dataset search: A survey. The VLDB Journal 29(1), 251–272 (2020), https://doi.org/10.1007/s00778-019-00564-x
9. Chu, M.T., Khosla, R.: Index evaluations and business strategies on communities of practice. Expert Systems with Applications 36(2, Part 1), 1549–1558 (2009), https://doi.org/10.1016/j.eswa.2007.11.053
10. Cinquini, M., Giannotti, F., Guidotti, R.: Boosting synthetic data generation with effective nonlinear causal discovery. In: 2021 IEEE Third International Conference on Cognitive Machine Intelligence (CogMI). pp. 54–63 (2021), https://doi.org/10.1109/CogMI52975.2021.00016
11. Cui, Y., Jin, Z., Jiang, J.: A novel supervised feature extraction and classification fusion algorithm for land cover recognition of the off-land scenario. Neurocomputing 140, 77–83 (2014), https://doi.org/10.1016/j.neucom.2014.03.034
12. Curry, E., Tuikka, T.: An organizational maturity model fordata spaces: A data sharing wheel approach. In: Data Spaces: Design, Deployment and Future Directions, pp. 21–42. Springer (2022), https://doi.org/10.1007/978-3-030-98636-0_2
13. Dekkers, M., Kotoglou, S., Nelson, C., Pellegrino, M., Hohn, N., Peristeras, V.: Statdcat-ap, a common layer for the exchange of statistical metadata in open data portals. In: SemStats@ ISWC (2016), https://ceur-ws.org/Vol-1654/article-05.pdf
14. Derycke, P., Barros, B.J., Schutte, N.M., Vande Catsyne, C.A., Fonseca, M.B.: Designing dcat-ap extensions for common european data spaces: The ehds healthdcat-ap case study. In: NeXt-generation Data Governance workshop 2025 (2025), https://ceur-ws.org/Vol-4064/NXDG25-paper5.pdf
15. Dong, Y., Xiao, C., Nozawa, T., Enomoto, M., Oyamada, M.: Deepjoin: Joinable table discovery with pre-trained language models. Proceedings of the VLDB Endowment 16(10), 2458–2470 (2023), https://doi.org/10.14778/3603581.3603587
16. Eclipse Dataspace Components: Eclipse Dataspace Components (2025), https://projects.eclipse.org/projects/technology.edc
17. Enríquez-Reyes, R., Cadena-Vela, S., Fuster-Guilló, A., Mazón, J.N., Ibáñez, L.D., Simperl, E.: Systematic mapping of open data studies: Classification and trends from a technological perspective. IEEE Access 9, 12968–12988 (2021), https://doi.org/10.1109/ACCESS.2021.3052025

18. European Commission: Simpl: Cloud-to-edge federations empowering eu data spaces (2024), https://digital-strategy.ec.europa.eu/en/policies/simpl, last updated 11 October 2024. Accessed: 2026-02-28
19. European Commission: Common european data spaces (2025), https://digital-strategy.ec.europa.eu/en/policies/data-spaces, last updated 19 November 2025. Accessed: 2026-02-28
20. European Commission: Dcat application profile for data portals in europe (2025), https://interoperable-europe.ec.europa.eu/collection/semic-support-centre/solution/dcat-application-profile-data-portals-europe/release/301, last updated 19 November 2025. Accessed: 2026-02-28
21. Fan, G., Wang, J., Li, Y., Zhang, D., Miller, R.J.: Semantics-aware dataset discovery from data lakes with contextualized column-based representation learning. Proceedings of the VLDB Endowment 16(7), 1726–1739 (2023), https://doi.org/10.14778/3587136.3587146
22. FIWARE Foundation: FIWARE - Open APIs for Open Minds (2024), https://www.fiware.org
23. Fowler, D., Barratt, J., Walsh, P.: Frictionless data: making research data quality visible. International Journal of Digital Curation 12(2), 274–285 (2017), https://doi.org/10.2218/ijdc.v12i2.577
24. Gaia-X European Association for Data and Cloud AISBL: Gaia-X Architecture Document (2024), https://docs.gaia-x.eu/technical-committee/architecture-document/latest/
25. Gieß, A., Hupperz, M.J., Schoormann, T., Möller, F.: What does it take to connect? unveiling characteristics of data space connectors. In: Bui, T.X. (ed.) 57th Hawaii International Conference on System Sciences, HICSS 2024, Hilton Hawaiian Village Waikiki Beach Resort, Hawaii, USA, January 3-6, 2024. pp. 4238–4247. ScholarSpace (2024), https://hdl.handle.net/10125/106895
26. Gieß, A., Möller, F., Schoormann, T., Otto, B.: Design options for data spaces. In: Thirty-first European Conference on Information Systems (ECIS 2023) (06 2023), https://aisel.aisnet.org/ecis2023_rp/287/
27. Hewagamage, P., Mihiranga, A., Perera, D., Fernando, R., Thilakarathna, T., Kasthurirathna, D.: Computer-vision enabled waste management system for green environment. In: 2021 3rd International Conference on Advancements in Computing (ICAC). pp. 276–281 (2021), https://doi.org/10.1109/ICAC54203.2021.9671222
28. Hummel, P., Braun, M., Tretter, M., Dabrock, P.: Data sovereignty: A review. Big Data & Society 8(1), 2053951720982012 (2021), https://doi.org/10.1177/2053951720982012
29. i4Trust: B2B Data Sharing Playbook - the i4Trust approach to Data Sharing (2021), https://i4trust.org/wp-content/uploads/i4Trust_DataSharingPlaybook.pdf
30. Imielinski, T., Mannila, H.: A database perspective on knowledge discovery. Commun. ACM 39(11), 58–64 (Nov 1996), https://doi.org/10.1145/240455.240472
31. International Data Spaces Association: Position Paper - GAIA-X and IDS (2021), https://internationaldataspaces.org/wp-content/uploads/dlm_uploads/IDSA-Position-Paper-GAIA-X-and-IDS.pdf
32. International Data Spaces Association: IDS Reference Architecture Model, Version 4.0 (2022), https://docs.internationaldataspaces.org/ids-knowledgebase/ids-ram-4/introduction/1_1_goals_of_the_international_data_spaces
33. International Data Spaces Association: International Data Spaces Association (2022), https://internationaldataspaces.org/

34. iSHARE Foundation: iSHARE - Trust Framework for Data Spaces (2021), https://ishare.eu/
35. Jahnke, N., Otto, B.: Data catalogs in the enterprise: applications and integration. Datenbank-Spektrum 23(2), 89–96 (2023), https://doi.org/10.1007/s13222-023-00445-2
36. Kazman, R., Klein, M., Clements, P.: Atam: Method for architecture evaluation. Tech. Rep. CMU/SEI-2000-TR-004, Software Engineering Institute, Carnegie Mellon University (Aug 2000), https://www.sei.cmu.edu/library/atam-method-for-architecture-evaluation/, accessed: 2026-Jul-16
37. Kirstein, F., Dutkowski, S., Dittwald, B., Hauswirth, M.: The european data portal: Scalable harvesting and management of linked open data. In: ISWC (Satellites). pp. 321–322 (2019), https://ceur-ws.org/Vol-2456/paper88.pdf
38. Kitchin, R.: The data revolution: Big data, open data, data infrastructures and their consequences. Sage (2014), https://doi.org/10.4135/9781473909472
39. Morejón, A., Berenguer, A., Espona, L.d., Tomás, D., Mazón, J.N.: Improving data discovery effectiveness: Experimental evaluation of content-based catalogs in data spaces. In: Chrysanthis, P.K., Nørvåg, K., Stefanidis, K., Zhang, Z., Quintarelli, E., Zumpano, E. (eds.) New Trends in Database and Information Systems. pp. 280–295. Springer Nature Switzerland, Cham (2026)
40. Nargesian, F., Asudeh, A., Jagadish, H.V.: Responsible data integration: Next-generation challenges. In: Proceedings of the 2022 International Conference on Management of Data. p. 2458–2464. SIGMOD '22, Association for Computing Machinery, New York, NY, USA (2022), https://doi.org/10.1145/3514221.3522567
41. Otto, B.: A federated infrastructure for european data spaces. Communications of the ACM 65(4), 44–45 (2022), https://doi.org/10.1145/3512341
42. Perego, A., Cetl, V., Friis-Christensen, A., Lutz, M., et al.: Geodcat-ap: Representing geographic metadata by using the "dcat application profile for data portals in europe". In: Joint UNECE/UNGGIM Europe Workshop on Integrating Geospatial and Statistical Standards, Stockholm, Sweden (2017), https://publications.jrc.ec.europa.eu/repository/handle/JRC107410
43. Ramos Rojas, J.A., Beth Kery, M., Rosenthal, S., Dey, A.: Sampling techniques to improve big data exploration. In: 2017 IEEE 7th Symposium on Large Data Analysis and Visualization (LDAV). pp. 26–35 (2017), https://doi.org/10.1109/LDAV.2017.8231848
44. Scerri, S., Tuikka, T., de Vallejo, I.L., Curry, E.: Common european data spaces: Challenges and opportunities. Data Spaces: Design, Deployment and Future Directions pp. 337–357 (2022), https://doi.org/10.1007/978-3-030-98636-0_16
45. W3C: Data catalog vocabulary (dcat) - version 3 (2024), https://www.w3.org/TR/vocab-dcat/, last updated 19 November 2025. Accessed: 2026-02-28
46. Xu, P., Ji, X., Li, M., Lu, W.: Virtual sample generation in machine learning assisted materials design and discovery. Journal of Materials Informatics 3(3) (2023), https://doi.org/10.20517/jmi.2023.18
47. Zhang, S., Balog, K.: Web table extraction, retrieval, and augmentation: A survey. ACM Transactions on Intelligent Systems and Technology 11(2), 13:1–13:35 (2020), https://doi.org/10.1145/3372117

Adriana Morejón is a researcher in the field of artificial intelligence and open data management, affiliated with the University of Alicante. Her work focuses on the application of AI techniques for data discovery and reuse, as well as on improving reference architectures for data spaces and open data ecosystems in the contexts of smart cities and tourism. She regularly collaborates on research and development projects, contributing

to the development of solutions aimed at the integration, visualisation and generation of open data, with a strong focus on quality and added value for business scenarios and data intelligence.

Alberto Berenguer holds a Ph.D. in Computer Science from the University of Alicante. His work sits at the intersection of machine learning and information science. His research focuses on applying advanced machine learning techniques to information retrieval and extraction. Beyond academia, he takes part in innovative projects on smart cities and data spaces, where he contributes expertise in data retrieval and metadata enhancement.

Lucía de Espona Pernas is a Senior Data Spaces Researcher and Computer Scientist specializing in data engineering and open data frameworks. She holds a PhD in Computer Science from the Universidad de Santiago de Compostela. Her technical background includes molecular biology research at ETH Zurich and leading engineering tracks for Innosuisse projects. Currently, she researches with the WaKe group at the University of Alicante, focusing on data spaces architecture, scalable sharing ecosystems, semantic interoperability, and trusted data infrastructures.

David Tomás is a Full Professor in the Department of Software and Computing Systems at the University of Alicante. His primary research focus is the application of Artificial Intelligence techniques to Natural Language Processing and Information Retrieval, fields in which he has published more than one hundred research articles in international conferences and high-impact journals. In these areas, he has participated in over twenty publicly funded projects at both the national and European levels.

Jose-Norberto Mazón is currently Full Professor in the University of Alicante, Spain. He is the author of more than 100 scientific publications in international conferences and journals. His research interests include open data, dataset discovery, data spaces, design of data-intensive web applications, and smart tourism destinations.

Received: March 1, 2026; Accepted: August 25, 2026.

Computer Science and Information Systems 23(4):1431–1452 https://doi.org/10.2298/CSIS260112034X

Paper Type: Research Paper

Modeling Students' Acceptance of AI-Assisted Translation Tools: An Extension of the Technology Acceptance Model

Wang Xiaoyun[1], Vladan Devedžić[2,3,4], and Wan Yinjia[1]

[1] Shandong Jiaotong University
Jinan, China
wangxiaoyun167@gmail.com
[2] Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences), Jinan, China
[3] Faculty of Organisational Sciences, University of Belgrade
Jove Ilica 154, 11000 Belgrade, Serbia
vladan.devedzic@fon.bg.ac.rs (corresponding author)
[4] Serbian Academy of Sciences and Arts, Knez Mihailova 35, Belgrade, Serbia

Abstract. This study investigates students' acceptance of AI-assisted translation tools by proposing an extended Technology Acceptance Model (TAM-AI) for AI-assisted translation learning contexts. Unlike prior additive approaches, the proposed model explains how translation-specific perceptions influence behavioral intention (BI) through underlying cognitive mechanisms. A survey was conducted among undergraduate translation students and analyzed using Structural Equation Modeling (SEM). The results indicate that perceived usefulness (PU) and perceived ease of use (PEOU) significantly predict BI. In addition, translation-specific factors influence technology acceptance indirectly: perceived translation quality operates through trust, feedback clarity through cognitive understanding, and cognitive load reduction through effort reduction. The TAM-AI model demonstrates greater explanatory power than the baseline TAM. These findings provide a deeper understanding of technology acceptance in AI-assisted learning environments and offer practical implications for the design of AI-assisted translation tools. Future studies with larger and more diverse samples are encouraged to further validate the proposed model.

Keywords: Technology Acceptance Model, TAM-AI, AI-Assisted Translation, Students' Acceptance, Translation Learning, Translation tool design.

1. Introduction

The rapid integration of machine translation and other AI-assisted tools has reshaped translation and language learning by introducing new forms of technology-mediated practice, feedback, and post-editing [6,14,21]. As these technologies become increasingly embedded in educational settings, understanding students' acceptance of AI-assisted translation tools has emerged as a critical issue, because attitudes, perceived usefulness, trust,

and behavioral intention significantly influence students' adoption of AI-supported learning technologies [3,15]. Recent studies have examined machine translation use in L2 writing and novice translation students' AI literacy [11,21]. However, comparatively less attention has been paid to the cognitive mechanisms through which translation-specific perceptions influence students' technology acceptance. TAM-based research specifically addressing translation students' acceptance of AI tools has only recently begun to emerge [1].

The Technology Acceptance Model (TAM), originally proposed by [4], remains a foundational framework for explaining technology adoption and has recently been extended to AI-assisted language learning and university students' use of AI tools [3,15]. However, translation learning involves distinctive processes such as evaluating AI-generated translation quality, interpreting feedback, revising machine-generated output, and developing appropriate AI and machine translation literacy [6,21]. Trust is also particularly important because it shapes students' reliance on and resistance to GenAI, which in turn influence their behavioral intention to use it in language learning [19]. Taken together, these findings suggest that general TAM-based models may not fully capture the translation-specific cognitive and evaluative processes underlying students' acceptance of AI-assisted translation tools.

To address this gap, this study proposes the TAM-AI model, an extension of TAM designed to examine students' acceptance of AI-assisted translation tools in translation learning contexts. By incorporating learning-oriented perceptions relevant to AI-supported translation learning, the TAM-AI model provides a context-sensitive framework for investigating technology acceptance.

The study addresses the following research questions:

- RQ1: Which factors significantly influence students' acceptance of AI-assisted translation tools?
- RQ2: To what extent does the TAM-AI model explain students' behavioral intention to use these tools?

Unlike prior studies that apply TAM in a descriptive manner, this study aims to provide a mechanism-based extension of TAM in the context of AI-assisted translation learning. Rather than simply introducing additional external variables, this study seeks to explain how translation-specific perceptions influence students' behavioral intention through underlying cognitive processes.

Furthermore, this study evaluates the incremental explanatory power of the proposed TAM-AI model by comparing it with the baseline TAM, thereby providing stronger empirical evidence for its theoretical contribution. This study contributes to the literature by moving beyond additive TAM extensions and introducing a mechanism-based perspective that explains how domain-specific perceptions are cognitively transformed into technology acceptance beliefs.

2. Literature Review

2.1. Technology Acceptance Models

TAM, originally proposed by [4], is one of the most widely used theoretical frameworks for explaining users' acceptance of information technologies. TAM posits that perceived

usefulness (PU) and perceived ease of use (PEOU) are the primary determinants of users' behavioral intention to adopt a technology. Due to its parsimony and explanatory power, TAM has been extensively applied and empirically validated across various educational technology contexts.

Subsequent studies have extended TAM to better capture technology adoption in complex learning environments [18]. In educational research, TAM has been used to examine students' acceptance of e-learning platforms, digital learning tools, and AI-supported educational applications [17]. However, despite these extensions, existing TAM-based studies have largely focused on general learning technologies and have rarely been adapted to domain-specific contexts such as translation learning, where task complexity and cognitive demands differ substantially from those of general educational settings.

2.2. AI-Assisted Translation Learning

Advances in artificial intelligence have led to the increasing integration of AI-assisted translation tools into translation education. These tools include machine translation systems, computer-assisted translation (CAT) tools, and AI-driven revision and feedback applications, which are commonly used to support translation practice. Research in translation studies has demonstrated that such tools can enhance translation efficiency, consistency, and revision quality [2].

However, existing studies in this area have primarily concentrated on learning outcomes and tool effectiveness, such as translation accuracy and productivity [12]. Comparatively less attention has been paid to students' perceptions of AI-assisted translation tools as learning technologies. As a result, the question of whether students are willing to accept and continuously use these tools in translation learning contexts remains insufficiently explored. Recent studies have emphasized the role of trust and cognitive processes in AI adoption [5]. More recent research further confirms that trust and cognitive evaluations are critical determinants of users' behavioral intentions toward AI technologies [7,13], highlighting the need for mechanism-based models.

2.3. Technology Acceptance of AI-Assisted Learning Tools

In recent years, researchers have increasingly applied TAM and related acceptance models to investigate learners' adoption of AI-supported educational technologies, including intelligent tutoring systems and AI-assisted writing tools. These studies consistently indicate that perceived usefulness and perceived ease of use play central roles in shaping learners' acceptance of AI-based learning tools [17].

Despite growing interest in technology acceptance of AI-assisted learning tools, most existing studies have focused on general educational or language learning contexts. However, given the distinctive characteristics of translation learning, including iterative revision and intensive cognitive processing, findings from general AI-assisted learning contexts may not be directly transferable to translation education. Research specifically addressing AI-assisted translation tools remains scarce. To the authors' knowledge, Alqahtani and Al-Ahdal [1] conducted one of the few studies explicitly applying TAM to investigate AI-tool adoption among English translation majors at a Saudi Arabian university. The study investigated the adoption rate of AI tools for translation at that university,

the actual use of the tools, as well as the students' perceptions regarding the integration of AI in translation practice. The results indicated that although most participants in the study (altogether 114 Saudi Arabian translation students learning English as a foreign language) found AI tools useful, they expressed concern about the potential negative impact of overuse on their translation skills. Recent studies have extended TAM in AI-assisted learning contexts by incorporating factors such as trust, responsibility, and cognitive processes, thereby improving its explanatory power [9,20].

2.4. Research Gap

Although previous studies have examined AI-assisted translation tools and TAM-based technology acceptance separately, existing research largely adopts an additive approach by incorporating external variables without explaining the underlying cognitive mechanisms through which these factors influence user acceptance.

Moreover, prior studies rarely evaluate whether such extensions offer incremental explanatory power over the baseline TAM. As a result, it remains unclear not only which factors influence students' acceptance of AI-assisted translation tools, but also how and why these factors shape their perceptions and behavioral intentions.

To address this gap, this study proposes a mechanism-based extension of TAM (TAM-AI) and empirically compares it with the original TAM to assess its explanatory power. Recent research has increasingly emphasized the need to move beyond additive models and incorporate cognitive and psychological mechanisms in explaining AI acceptance [8,16]. However, existing studies rarely explain the underlying cognitive mechanisms through which these factors influence acceptance, resulting in limited theoretical integration.

3. Theoretical Framework and Research Hypotheses

The theoretical framework underlying this study relies on TAM and its extension into the TAM-AI model. It explains how this model is applied to assess students' acceptance of AI-assisted translation tools in educational settings. The research hypotheses that drove this study explored the factors that influence students' BI to use these tools, as well as the role of individual characteristics in shaping their technology acceptance.

3.1. Overview of the Original TAM Model

TAM, proposed by [4], is a parsimonious theoretical model developed to explain individuals' acceptance and use of information systems. Rooted in the Theory of Reasoned Action (TRA), TAM conceptualizes technology adoption as a cognitively driven process in which users' beliefs shape their intentions and subsequent behaviors.

The original TAM consists of two fundamental belief constructs: PU and PEOU. Perceived Usefulness is defined as the degree to which an individual believes that using a particular system will enhance task performance, whereas Perceived Ease of Use refers to the extent to which the system is perceived as requiring minimal effort. TAM posits that these two beliefs jointly determine users' attitudes toward the system and directly influence BI, which serves as the immediate antecedent of actual system use.

A central assumption of TAM is that users' acceptance decisions are primarily shaped by internal cognitive evaluations rather than by external system characteristics per se. External variables—such as system features, user experience, or contextual conditions—are theorized to influence adoption indirectly through their effects on PU and PEOU. This assumption allows TAM to maintain a high degree of conceptual simplicity while preserving strong explanatory and predictive power.

Due to its clarity and empirical robustness, TAM has been widely adopted and validated across a broad range of technological contexts, including organizational information systems and educational technologies. Its structural simplicity facilitates empirical testing and model extension, making it a foundational framework for subsequent acceptance models and domain-specific adaptations.

Understanding these core constructs, assumptions, and causal logic of the original TAM is essential for capturing the process by which extended acceptance models can be systematically developed.

3.2. Extension to the TAM-AI Model

Unlike conventional TAM extensions that rely on direct relationships between external variables and core TAM constructs, the TAM-AI model proposed in this study adopts a mechanism-based perspective. Specifically, we argue that translation-specific perceptions do not influence perceived usefulness and perceived ease of use directly, but rather operate through underlying cognitive and evaluative processes.

In the context of AI-assisted translation, perceived translation quality is expected to enhance users' trust in AI-generated outputs, which subsequently strengthens perceived usefulness. Similarly, feedback clarity facilitates users' cognitive understanding of translation errors and corrections, thereby reinforcing their perception of the tool's value.

In addition, cognitive load reduction lowers the mental effort required during translation tasks, which contributes to the perception that the system is easy to use. By incorporating these cognitive mechanisms, the TAM-AI model moves beyond a purely additive structure and provides a process-oriented explanation of technology acceptance.

Furthermore, to validate the theoretical contribution of the proposed model, this study compares TAM-AI with the baseline TAM to assess its incremental explanatory power.

3.3. Research Hypotheses

Based on the TAM-AI model, the following hypotheses are proposed to examine the factors influencing students' acceptance of AI-assisted translation tools.

PU, PEOU, and BI relationships

H1: Perceived usefulness (PU) of AI-assisted translation tools will have a significant positive effect on students' behavioral intention (BI) to use these tools.

H2: Perceived ease of use (PEOU) of AI-assisted translation tools will have a significant positive effect on students' behavioral intention (BI) to use these tools.

Mechanism-based relationships

H3: Perceived translation quality (PTQ) positively influences trust in AI-assisted translation tools.

H4: Trust in AI-assisted translation tools positively influences perceived usefulness (PU).

H5: Feedback clarity (FC) positively influences students' cognitive understanding of translation processes.

H6: Cognitive understanding positively influences perceived usefulness (PU).

H7: Cognitive load reduction (CLR) positively influences perceived effort reduction.

H8: Effort reduction positively influences perceived ease of use (PEOU).

Individual characteristics

H9: Prior experience with AI tools positively influences perceived usefulness (PU) and perceived ease of use (PEOU).

H10: Translation proficiency positively influences perceived usefulness (PU) and behavioral intention (BI).

H11: Learning preferences aligned with technology-based learning positively influence behavioral intention (BI).

These hypotheses aim to examine not only the direct effects of TAM constructs, but also the underlying cognitive mechanisms through which AI-related perceptions influence students' acceptance of AI-assisted translation tools.

4. Methodology

These factors include both the core constructs of the Technology Acceptance Model (TAM), namely perceived usefulness (PU) and perceived ease of use (PEOU), as well as AI-specific variables such as feedback clarity (FC), perceived translation quality (PTQ), and cognitive load reduction (CLR).

Importantly, the study also incorporates mediating constructs—trust in AI-assisted translation tools (TR), cognitive understanding (CU), and effort reduction (ER)—to capture the underlying cognitive mechanisms through which AI-specific perceptions influence technology acceptance.

To assess the incremental explanatory power of the proposed TAM-AI model, a baseline TAM model including only perceived usefulness and perceived ease of use was estimated. The results were compared in terms of explained variance (R^2) and model fit indices to determine whether TAM-AI provides superior explanatory power.

4.1. Research Design

This study adopted a quantitative research design using a cross-sectional survey to assess students' acceptance of AI-assisted translation tools. The research focused on measuring students' behavioral intention (BI) to use AI tools for translation learning and aimed to identify the factors influencing their acceptance.

These factors include both the core constructs of the Technology Acceptance Model (TAM), namely perceived usefulness (PU) and perceived ease of use (PEOU), as well as AI-specific variables such as feedback clarity (FC), perceived translation quality (PTQ), and cognitive load reduction (CLR).

Importantly, the study also incorporates mediating constructs—trust in AI-assisted translation tools (TR), cognitive understanding (CU), and effort reduction (ER)—to capture the underlying cognitive mechanisms through which AI-specific perceptions influence technology acceptance.

4.2. Participants

The participants were 120 undergraduate translation students enrolled in translation-related programs, and the data were collected from undergraduate students at Shandong Jiaotong University in China. All participants had at least some prior exposure to AI-powered translation tools, while 33.3% reported frequent or multi-tool usage experience.

In addition to AI tool experience, demographic information was collected to assess individual characteristics such as translation proficiency, which was operationalized using students' self-reported academic performance in translation-related courses and categorized as Top 10%, Top 11–30%, Middle 31–70%, and Bottom 30%, prior exposure to AI technology (measured by how often students have used AI tools), and learning preferences (e.g., preference for interactive feedback or autonomous learning). This allowed for subgroup analysis to examine how these characteristics influence students' acceptance of AI-assisted translation tools. The demographic characteristics of the participants are summarized in Table 1.

Table 1. Demographic information of participants.

Demographic Profile	Classification	Number	Percent
Grade	Junior 1	29	24.17%
	Junior 2	30	25%
	Junior 3	31	25.83%
	Junior 4	30	25%
Frequent or Multi-Tool AI Experience	Google Translate, DeepL, Grammarly	40	33.3%
Translation Proficiency	Self-reported translation performance categories	120	100%
Prior Exposure to AI Technology	Frequency of AI tool usage	56	46.7%
Learning Preferences	Interactive feedback, autonomous learning	60	50%

4.3. Instruments

A structured self-report questionnaire was used to collect data from participants. The questionnaire included items designed to measure the core constructs of the TAM, namely PU, PEOU, and BI.

In addition, the questionnaire incorporated translation-specific variables, including FC, PTQ, and CLR.

To capture the underlying cognitive mechanisms proposed in the TAM-AI model, three additional constructs were included: trust in AI-assisted translation tools, CU, and ER. These constructs were designed to reflect the mediating processes through which AI-specific perceptions influence students' acceptance of AI-assisted translation tools.

All items were measured using a 5-point Likert scale (1 = strongly disagree to 5 = strongly agree). The measurement items were adapted from prior TAM and AI acceptance studies and refined to fit the context of AI-assisted translation learning.

The complete questionnaire is provided in Appendix A.

4.4. Data Collection and Procedure

The data were collected through the Google Forms online survey platform, and the key characteristics of the data collection process are summarized in Table 2. Participants received an invitation to complete the survey, along with an explanation of the study's purpose, ethical considerations, and instructions for completing the questionnaire. The survey took approximately 15-20 minutes to complete, and students were informed that participation is voluntary and anonymous.

The survey was distributed to students via email or online learning platforms, ensuring that it reached the intended participants. The data collection period lasted two weeks, allowing sufficient time for participants to respond. After the completion of the survey, responses were collected, anonymized, and stored securely for analysis.

Table 2. Data collection.

Data Collection Characteristics	Information
Platform	Online survey platform (Google Forms)
Survey Duration	15–20 minutes to complete
Survey Distribution	Distributed via email or online learning platforms
Data Collection Period	Two weeks
Data Handling	Responses were anonymized and securely stored for analysis

4.5. Data Analysis

The data were analyzed using Structural Equation Modeling (SEM) with AMOS software[5]. The analysis consisted of two main stages: assessment of the measurement model and evaluation of the structural model.

First, Confirmatory Factor Analysis (CFA) was conducted to assess the reliability and validity of the measurement model. Convergent validity was evaluated using factor loadings, composite reliability (CR), and average variance extracted (AVE). Discriminant validity was assessed using the Fornell–Larcker criterion and the heterotrait–monotrait ratio (HTMT).

Second, the structural model was examined to test the hypothesized relationships among constructs. Path coefficients, significance levels, and coefficients of determination (R^2) were reported.

To examine the mediating effects proposed in the TAM-AI model, bootstrapping with 5000 resamples was employed to estimate indirect effects and their confidence intervals.

Finally, to evaluate the incremental explanatory power of the proposed TAM-AI model, a baseline TAM model including only PU and PEOU was estimated and compared with the TAM-AI model in terms of explained variance (R^2) and model fit indices.

[5] IBM SPSS AMOS software, available at: https://www.ibm.com/products/structural-equation-modeling-sem

To assess potential common method bias, Harman's single-factor test was conducted. The results showed that the first factor accounted for less than 50% of the total variance, indicating that common method bias was unlikely to be a serious concern. The main data analysis procedures and statistical techniques used in this study are summarized in Table 3.

Table 3. Data analysis.

Data Analysis Characteristics	Details
Analysis Method	SEM
Key Constructs Tested	PU, PEOU, FC, PTQ, CLR, TR, CU, ER, BI
Subgroup Analysis	Individual characteristics and exploratory subgroup differences
Software Used	SPSS
	AMOS

4.6. Ethical Considerations

This study adhered to ethical standards regarding participant rights, informed consent, and confidentiality. Informed consent was obtained from all participants prior to their participation in the study. They were informed about the study's purpose, the voluntary nature of their participation, and the confidentiality of their responses.

All data were anonymized to ensure that participants' identities were protected. The responses were stored securely and were only used for research purposes. This study was reviewed and approved by the Ethics Committee of Shandong Jiaotong University prior to data collection. All participants provided informed consent, and their responses were collected anonymously. All procedures performed in this study were in accordance with institutional ethical standards.

5. Results

5.1. Measurement Model

Confirmatory Factor Analysis (CFA) was conducted to evaluate the measurement model. The results indicate that all constructs demonstrate satisfactory reliability and validity.

Convergent validity was confirmed as all factor loadings exceeded 0.70, composite reliability (CR) values were above 0.70, and average variance extracted (AVE) values exceeded the recommended threshold of 0.50.

Discriminant validity was assessed using both the Fornell–Larcker criterion and the HTMT ratio. The square root of AVE for each construct was greater than its correlations with other constructs, and all HTMT values were below the threshold of 0.85, indicating adequate discriminant validity.

These results suggest that the measurement model is reliable and valid. The detailed results are presented in Table 4. All factor loadings exceeded 0.70. Composite reliability

(CR) values ranged from 0.84 to 0.89, and AVE values exceeded 0.50. HTMT values were below 0.85, indicating satisfactory discriminant validity.

Table 4. Measurement model results.

Construct	Loading Range	CR	AVE	HTMT (max)
PU	0.74–0.86	0.86	0.68	0.78
PEOU	0.72–0.84	0.84	0.64	0.80
PTQ	0.71–0.85	0.85	0.66	0.76
FC	0.73–0.87	0.86	0.67	0.79
CLR	0.75–0.88	0.87	0.69	0.77
TR	0.76–0.89	0.88	0.71	0.74
CU	0.74–0.86	0.86	0.68	0.76
ER	0.73–0.85	0.85	0.66	0.78
BI	0.78–0.90	0.89	0.73	—

Note 1. Discriminant validity was assessed using the Fornell–Larcker criterion and inter-construct correlations.
Note 2. AVE > 0.50 and CR > 0.70 indicate acceptable convergent validity and reliability [10].

5.2. Structural Model (H1–H8)

Following the validation of the measurement model, the SEM was conducted to test the relationships between the key constructs in the TAM-AI model. The results of the structural model and hypothesis testing are summarized in Table 5.

The structural model demonstrated a good fit to the data (CFI = 0.95, TLI = 0.93, RMSEA = 0.048).

Perceived usefulness (PU) had a significant positive effect on behavioral intention (BI) ($\beta = 0.41$, $p < 0.001$), supporting H1. Perceived ease of use (PEOU) also positively influenced BI ($\beta = 0.29$, $p < 0.01$), supporting H2.

Perceived translation quality (PTQ) significantly influenced trust ($\beta = 0.52$, $p < 0.001$), supporting H3. Trust, in turn, had a significant effect on perceived usefulness ($\beta = 0.36$, $p < 0.001$), supporting H4.

Feedback clarity (FC) positively affected cognitive understanding ($\beta = 0.48$, $p < 0.001$), supporting H5. Cognitive understanding significantly influenced perceived usefulness ($\beta = 0.33$, $p < 0.01$), supporting H6.

Cognitive load reduction (CLR) significantly influenced effort reduction ($\beta = 0.50$, $p < 0.001$), supporting H7. Effort reduction significantly affected perceived ease of use (PEOU) ($\beta = 0.35$, $p < 0.01$), supporting H8.

The model explained 58% of the variance in BI ($R^2 = 0.58$).

These results provide strong evidence that PU and PEOU are critical factors in determining students' BI to use AI-assisted translation tools.

Table 5. Hypothesis testing.

Hypothesis	Path	Standardized Coefficient (β)	p-value	Result
H1	PU $\rightarrow$ BI	0.41	< 0.001	Supported
H2	PEOU $\rightarrow$ BI	0.29	< 0.01	Supported
H3	PTQ $\rightarrow$ TR	0.52	< 0.001	Supported
H4	TR $\rightarrow$ PU	0.36	< 0.001	Supported
H5	FC $\rightarrow$ CU	0.48	< 0.001	Supported
H6	CU $\rightarrow$ PU	0.33	< 0.01	Supported
H7	CLR $\rightarrow$ ER	0.50	< 0.001	Supported
H8	ER $\rightarrow$ PEOU	0.35	< 0.01	Supported

5.3. Additional Hypothesis Testing Results (H9–H11)

Bootstrapping analysis (5000 resamples) was conducted to examine the mediating effects proposed in the mechanism-based TAM-AI model. The results showed that trust significantly mediated the relationship between PTQ and PU (indirect effect β = 0.19, 95% CI [0.11, 0.28]). Cognitive understanding significantly mediated the relationship between FC and PU (β = 0.16, 95% CI [0.08, 0.24]). Effort reduction significantly mediated the relationship between CLR and PEOU (β = 0.18, 95% CI [0.10, 0.27]). These results support the mechanism-based structure of the TAM-AI model.

In addition, further analyses were conducted to test hypotheses H9–H11 concerning individual characteristics. Specifically, prior AI tool experience significantly influenced PU and PEOU, translation proficiency significantly influenced PU and BI, and learning preferences significantly influenced BI. The results of these additional hypothesis tests are summarized in Table 6.

Table 6. Additional hypothesis testing results (H9–H11).

Hypothesis	Path Coefficient (β)	p-value	Result
H9a: Prior AI Tool Experience $\rightarrow$ PU	0.34	< 0.001	Supported
H9b: Prior AI Tool Experience $\rightarrow$ PEOU	0.29	< 0.01	Supported
H10a: Translation Proficiency $\rightarrow$ PU	0.31	< 0.01	Supported
H10b: Translation Proficiency $\rightarrow$ BI	0.26	< 0.05	Supported
H11: Learning Preferences $\rightarrow$ BI	0.28	< 0.01	Supported

5.4. Subgroup Analysis

It should be noted that the coefficients reported in the subgroup analysis differ from those in the main structural model, as they are derived from separate analyses. The coefficients reported in Table 6 represent the additional hypothesis testing results for H9–H11, whereas the subgroup coefficients reported below reflect variations across specific participant groups and are therefore not directly comparable.

Subgroup analysis was performed to examine the effect of individual student characteristics, such as AI tool experience, translation proficiency, and learning preferences, on their BI to use AI-assisted translation tools. The results are summarized below.

– AI Tool Experience:

Students with prior experience using AI translation tools (e.g., Google Translate, DeepL) reported higher levels of PU (β = 0.40, p $<$ 0.01) and PEOU (β = 0.35, p $<$ 0.01) compared to students with little or no prior experience. They also showed a higher BI to use these tools.

– Translation Proficiency:

Students with higher translation proficiency demonstrated a stronger perception of PTQ (β = 0.30, p $<$ 0.05) and were more likely to use AI-assisted translation tools to aid in their translation work. These results indicate differences across proficiency levels rather than statistically tested moderation effects.

– Learning Preferences:

Students who preferred interactive feedback and autonomous learning had higher BI to use AI tools (β = 0.33, p $<$ 0.01). These students valued tools that provided personalized feedback and allowed them to take ownership of their learning process.

5.5. Qualitative Findings

In addition to the quantitative survey, open-ended responses were collected from participants to capture their personal experiences and insights regarding AI translation tools. A total of 120 valid qualitative responses were analyzed using thematic analysis. Two researchers independently coded the data, and discrepancies were resolved through discussion to ensure consistency. A thematic analysis of these responses revealed several key themes:

– Tool Usability: Students emphasized the importance of ease of use, with many expressing satisfaction with tools like Google Translate and DeepL due to their user-friendly interfaces. However, some students with more advanced translation experience mentioned the need for tools with more advanced features like contextual translation.
– Feedback and Improvement: Many students reported that real-time feedback helped them improve their translations. Tools that provided detailed explanations of corrections were especially appreciated by students who valued self-directed learning.
– Translation Quality Concerns: A significant portion of students expressed concerns about the accuracy of AI translations. They often felt that the tools were not sufficiently accurate for context-sensitive translations, especially in specialized fields such as legal or medical translation. One student noted, “AI tools are helpful, but sometimes they fail to capture the exact meaning in specialized contexts.”

These qualitative findings further support the quantitative results and provide deeper insights into the factors that influence students’ acceptance and use of AI translation tools.

5.6. Robustness Checks

To ensure the robustness of the findings, several additional tests were conducted. First, a baseline TAM model including only perceived usefulness (PU), perceived ease of use (PEOU), and behavioral intention (BI) was estimated and compared with the proposed TAM-AI model. The TAM-AI model demonstrated higher explanatory power and improved model fit indices.

Second, alternative structural specifications excluding the mediating constructs were tested. The results showed a reduction in explanatory power, supporting the importance of the proposed cognitive mechanisms.

Specifically, the alternative models excluded the mediating constructs (TR, CU, and ER) and directly linked translation-specific variables to PU and PEOU, representing an additive specification. Compared with these alternative models, the proposed mechanism-based TAM-AI model demonstrated stronger explanatory power and better model fit. This supports the theoretical importance of the mediating mechanisms and suggests that the findings are not merely a result of adding more predictors to the baseline TAM.

5.7. Summary of Results

The results of this study confirm that perceived usefulness (PU) and perceived ease of use (PEOU) are significant predictors of students' behavioral intention (BI) to use AI-assisted translation tools.

More importantly, the findings demonstrate that translation-specific perceptions influence students' acceptance of AI tools indirectly through cognitive mechanisms. Specifically, perceived translation quality operates through trust, feedback clarity operates through cognitive understanding, and cognitive load reduction operates through effort reduction.

These results provide strong support for the mechanism-based TAM-AI model and highlight the importance of cognitive processes in shaping technology acceptance in AI-assisted translation learning. Regarding RQ1, the results indicate that both TAM constructs and translation-specific factors influence students' acceptance of AI-assisted translation tools, either directly or indirectly through cognitive mechanisms. Regarding RQ2, the TAM-AI model demonstrates greater explanatory power than the baseline TAM model in explaining students' behavioral intention.

6. Discussion

6.1. Interpretation of Key Findings

Building on the empirical results reported in Section 5, this study provides a mechanism-based interpretation of students' acceptance of AI-assisted translation tools. This mechanism-based perspective offers a more theoretically grounded explanation of technology acceptance in AI-assisted learning environments and advances TAM research by shifting from variable inclusion to process explanation.

First, the results confirm that perceived usefulness (PU) and perceived ease of use (PEOU) remain the key determinants of students' behavioral intention (BI), which is consistent with the core assumptions of the Technology Acceptance Model (TAM). This

finding reinforces the continued relevance of TAM in explaining technology adoption in AI-supported learning environments.

More importantly, the findings reveal that translation-specific factors influence technology acceptance through underlying cognitive mechanisms rather than through direct effects. Specifically, perceived translation quality (PTQ) affects perceived usefulness indirectly through trust in AI-assisted translation tools. This suggests that students' evaluation of AI-generated translations depends not only on output quality but also on their confidence in the reliability of the system.

Similarly, feedback clarity (FC) influences perceived usefulness through cognitive understanding. Clear and interpretable feedback enables students to better understand translation errors and revision strategies, which enhances their perception of the tool's usefulness. This finding highlights the importance of cognitive processing in AI-assisted learning environments.

In addition, cognitive load reduction (CLR) affects perceived ease of use through effort reduction. When AI tools reduce the mental effort required for translation tasks, students are more likely to perceive these tools as easy to use.

Overall, these findings support the mechanism-based extension of TAM proposed in this study, demonstrating that AI-specific perceptions shape technology acceptance through cognitive and evaluative processes rather than through direct relationships. These findings are consistent with recent studies on AI-assisted learning and technology acceptance, which highlight the importance of trust, cognitive mechanisms, and contextual factors in shaping users' behavioral intentions [7, 13, 16].

6.2. Theoretical Contributions

This study contributes to the TAM literature by proposing a mechanism-based extension of the Technology Acceptance Model in the context of AI-assisted translation learning. This aligns with recent efforts to extend TAM in AI-driven learning environments by incorporating cognitive and affective mechanisms [20].

Unlike prior TAM-based studies that primarily adopt an additive approach by introducing external variables as direct predictors, this study explains how AI-specific perceptions influence technology acceptance through underlying cognitive mechanisms, including trust, cognitive understanding, and effort reduction. This mechanism-based perspective provides a deeper and more process-oriented explanation of technology acceptance.

Furthermore, this study empirically demonstrates that the TAM-AI model provides greater explanatory power than the baseline TAM model. By incorporating cognitive mechanisms, the proposed model better captures the complexity of students' interactions with AI-assisted translation tools.

In addition, this study extends TAM to a domain-specific context—translation learning—where cognitive processing and feedback interpretation play a critical role. This contributes to the growing body of research on AI-assisted learning and highlights the importance of adapting technology acceptance models to specific learning environments. The findings contribute to AI acceptance research by demonstrating that domain-specific perceptions influence behavioral intention through identifiable cognitive mechanisms rather than direct effects alone.

6.3. Practical Implications

The findings from this study have several practical implications for the design and implementation of AI-assisted translation tools. First, the significant impact of PU and PEOU suggests that AI translation tools must prioritize usability and utility. Developers should focus on creating tools that are not only user-friendly but also provide value in enhancing students' translation skills. Clear, context-aware feedback and accurate translations should be central to the design process.

The study also underscores the importance of aligning AI tools with students' learning preferences. Students who prefer interactive feedback and autonomous learning reported higher BI , suggesting that tools incorporating these features will have greater success in engaging users. Educational practitioners should consider these preferences when integrating AI tools into translation curricula to ensure greater adoption and effectiveness.

Furthermore, since prior AI tool experience and translation proficiency were significant factors in tool adoption, it is crucial for educators to provide targeted support to students with limited experience using AI tools. Offering training sessions or introductory workshops on how to effectively use these tools may help increase their acceptance and encourage more widespread usage.

6.4. Limitations

While this study provides valuable insights, there are several limitations that should be considered. First, the sample consisted of undergraduate translation students from a single institution, which may limit the generalizability of the findings. The results may not be applicable to students from other academic backgrounds or institutions. Future studies could expand the sample to include a broader range of students from different universities or countries, which would enhance the external validity of the findings.

Second, the study relied on self-reported data, which can be subject to response biases. Students may have provided socially desirable answers, especially when reporting their AI tool usage or translation proficiency. Future research could incorporate objective measures, such as system-generated usage logs (e.g., frequency and duration of AI tool use), performance-based translation assessments evaluated by instructors or experts, or automated translation quality indicators, to reduce the risk of bias and provide a more accurate assessment of students' engagement with AI tools.

Lastly, this study focused exclusively on AI-assisted translation tools, and its findings may not be directly applicable to other types of AI tools in education. Future research could explore the application of the TAM-AI model in other educational contexts, such as AI-powered language learning tools or AI-driven tutoring systems, to examine the broader applicability of the model.

Although the sample size (N = 120) satisfied commonly accepted minimum requirements for covariance-based SEM and yielded stable model estimates, the relatively modest sample may still limit the generalizability of the findings. Future research should validate the proposed TAM-AI model using larger and more diverse samples across different universities, disciplines, and educational settings. Such efforts would help further examine the robustness and broader applicability of the model.

6.5. Future Research Directions

This study opens several avenues for future research. First, it would be valuable to explore the long-term effects of using AI translation tools on students' learning outcomes. Longitudinal studies could track changes in students' PU, PEOU, and BI over time, and assess how these factors impact students' translation proficiency and engagement with AI tools in the long run.

Second, research could investigate cross-cultural differences in the acceptance of AI tools. Students from different cultural backgrounds may have different attitudes towards technology and education, which could affect their willingness to adopt AI-assisted translation tools. Future studies could compare students' perceptions and behavioral intentions in different cultural contexts to better understand how cultural factors influence technology acceptance.

Finally, future research could examine the impact of AI translation tools on professional translators. While this study focused on students, it would be insightful to explore how AI tools affect the workflow of professional translators, particularly with regard to efficiency and quality of translation. Understanding how AI tools are integrated into professional practices could provide valuable insights for the design of future tools aimed at both learners and professionals.

7. Conclusion

This study provides a comprehensive examination of students' acceptance of AI-assisted translation tools by proposing a mechanism-based extension of the Technology Acceptance Model (TAM-AI).

The findings indicate that students' behavioral intention to use AI-assisted translation tools is primarily driven by perceived usefulness and perceived ease of use. More importantly, the study demonstrates that translation-specific perceptions influence technology acceptance indirectly through cognitive mechanisms, including trust in AI tools, cognitive understanding of feedback, and effort reduction during translation tasks.

By incorporating these mechanisms, the TAM-AI model offers a more nuanced and process-oriented understanding of how students evaluate and adopt AI-assisted translation tools. This highlights the importance of considering cognitive and evaluative processes in modeling technology acceptance in AI-supported learning environments.

The study provides practical insights for the design of AI-assisted translation tools, emphasizing the need to enhance trust, improve feedback clarity, and reduce cognitive effort. Future research can further explore the applicability of the TAM-AI model in other AI-supported educational contexts and investigate its long-term impact on learning outcomes.

Acknowledgement. This work was supported by the Serbian Academy of Sciences and Arts grant No. F-150.

References

1. Alqahtani, M.H., Al-Ahdal, A.A.: AI tools and Saudi university English translation students: A mixed-methods study based on TAM. Theory and Practice in Language Studies 15(12), 4077–4086 (2025), https://doi.org/10.17507/tpls.1512.27

2. Bowker, L., Ciro, J.B.: Machine Translation and Global Research: Towards Improved Machine Translation Literacy in the Scholarly Community. Emerald Publishing Limited, Bingley, UK (2019), https://doi.org/10.1108/9781787567214
3. Chen, D., Liu, W., Liu, X.: What drives college students to use AI for L2 learning? modeling the roles of self-efficacy, anxiety, and attitude based on an extended technology acceptance model. Acta Psychologica 249, 104442 (2024), https://doi.org/10.1016/j.actpsy.2024.104442
4. Davis, F.D.: Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly 13(3), 319–340 (1989), https://doi.org/10.2307/249008
5. Dwivedi, Y.K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P.V., Janssen, M., Jones, P., Kar, A.K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., Medaglia, R., Le Meunier-FitzHugh, K., Le Meunier-FitzHugh, L.C., Misra, S., Mogaji, E., Sharma, S.K., Singh, J.B., Raghavan, V., Raman, R., Rana, N.P., Samothrakis, S., Spencer, J., Tamilmani, K., Tubadji, A., Walton, P., Williams, M.D.: Artificial intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. International Journal of Information Management 57, 101994 (2021), https://doi.org/10.1016/j.ijinfomgt.2019.08.002
6. Ehrensberger-Dow, M., Delorme Benites, A., Lehr, C.: A new role for translators and trainers: MT literacy consultants. The Interpreter and Translator Trainer 17(3), 393–411 (2023), https://doi.org/10.1080/1750399X.2023.2237328
7. Fel, S., Kozak, J., Horodyski, P.: Responsibility and AI: Exploring technology acceptance models. Journal of Innovation & Knowledge 10(6), 100852 (2025), https://doi.org/10.1016/j.jik.2025.100852
8. Ghimire, A., Edwards, J.: Generative AI adoption in the classroom: A contextual exploration using the technology acceptance model (TAM) and the innovation diffusion theory (IDT). In: 2024 Intermountain Engineering, Technology and Computing (IETC). pp. 129–134. IEEE (2024), https://doi.org/10.1109/IETC61393.2024.10564292
9. Guo, S., Shi, L., Zhai, X.: Developing and validating an instrument for teachers' acceptance of artificial intelligence in education. Education and Information Technologies 30(10), 13439–13461 (2025), https://doi.org/10.1007/s10639-025-13338-6
10. Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E.: Multivariate Data Analysis. Cengage, Andover, UK, 8 edn. (2019)
11. Jiang, L., Yu, R., Zhao, Y.: Theoretical perspectives and factors influencing machine translation use in L2 writing: A scoping review. Journal of Second Language Writing 64, 101099 (2024), https://doi.org/10.1016/j.jslw.2024.101099
12. Kenny, D., Doherty, S.: Statistical machine translation in the translation curriculum: Overcoming obstacles and empowering translators. The Interpreter and Translator Trainer 8(2), 276–294 (2014), https://doi.org/10.1080/1750399X.2014.936112
13. Khoso, A.K., Honggang, W., Darazi, M.A.: Trust and attitude towards AI as pathways to creativity: A TAM model study of EFL students' digital literacy and AI acceptance. Humanities and Social Sciences Communications 13(1), 69 (2026), https://doi.org/10.1057/s41599-025-06362-x
14. Klímová, B., Pikhart, M., Delorme Benites, A., Lehr, C., Sanchez-Stockhammer, C.: Neural machine translation in foreign language teaching and learning: A systematic review. Education and Information Technologies 28(1), 663–682 (2023), https://doi.org/10.1007/s10639-022-11194-2
15. Mustofa, R.H., Kuncoro, T.G., Atmono, D., Hermawan, H.D., Sukirman: Extending the technology acceptance model: The role of subjective norms, ethics, and trust in AI tool adoption among students. Computers and Education: Artificial Intelligence 8, 100379 (2025), https://doi.org/10.1016/j.caeai.2025.100379

16. Pitts, G., Motamedi, S.: What drives students' use of AI chatbots? technology acceptance in conversational AI (2026), https://doi.org/10.48550/arXiv.2602.20547
17. Scherer, R., Siddiq, F., Tondeur, J.: The technology acceptance model (TAM): A meta-analytic structural equation modeling approach to explaining teachers' adoption of digital technology in education. Computers & Education 128, 13–35 (2019), https://doi.org/10.1016/j.compedu.2018.09.009
18. Venkatesh, V., Davis, F.D.: A theoretical extension of the technology acceptance model: Four longitudinal field studies. Management Science 46(2), 186–204 (2000), https://doi.org/10.1287/mnsc.46.2.186.11926
19. Wang, F., Li, N., Cheung, A.C.K., Wong, G.K.W.: In GenAI we trust: An investigation of university students' reliance on and resistance to generative AI in language learning. International Journal of Educational Technology in Higher Education 22(1), 59 (2025), https://doi.org/10.1186/s41239-025-00547-9
20. Xue, L., Mahat, J., Ghazali, N.: Technology acceptance model in artificial intelligence in education: A meta-analysis. SAGE Open 16(1) (2026), https://doi.org/10.1177/21582440251409441
21. Zhang, J., Doherty, S.: Investigating novice translation students' AI literacy in translation education. The Interpreter and Translator Trainer 19(3–4), 234–253 (2025), https://doi.org/10.1080/1750399X.2025.2541478

Appendix

A. Questionnaire

Instructions

This questionnaire aims to examine students' acceptance of AI-assisted translation tools used for translation learning. Please answer all questions based on your actual experience with AI-assisted translation tools (e.g., Google Translate, DeepL, Grammarly).

All items in Sections C–K are measured on a 5-point Likert scale:

1 = Strongly disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5 = Strongly agree.

Participation is voluntary and anonymous. All responses will be used for research purposes only.

Section A. Demographic Information

A1. Grade

☐ Junior 1
☐ Junior 2
☐ Junior 3
☐ Junior 4

A2. Major / Program

☐ Translation
☐ English
☐ Other: _____

Section B. Background Information

Prior AI Tool Experience

B1. Which AI-assisted translation tools have you used? (Multiple choice)

☐ Google Translate
☐ DeepL
☐ Grammarly
☐ Other AI-based tools (e.g., ChatGPT): _____

B2. How often do you use AI-assisted translation tools for translation-related tasks?

☐ Rarely
☐ Sometimes
☐ Often
☐ Very often

Translation Proficiency

B3. How would you describe your academic performance in translation-related courses?

☐ Top 10%
☐ Top 11–30%
☐ Middle 31–70%
☐ Bottom 30%

Learning Preferences

B4. I prefer learning activities that involve technology-supported interaction and feedback.
B5. I prefer autonomous learning environments where I can control my learning pace.
B6. I feel comfortable learning with digital tools rather than only traditional materials.

Section C. Perceived Usefulness (PU)

PU1. Using AI-assisted translation tools improves my performance in translation learning.
PU2. AI-assisted translation tools help me complete translation tasks more efficiently.
PU3. AI-assisted translation tools enhance the overall effectiveness of my translation learning.

Section D. Perceived Ease of Use (PEOU)

PEOU1. Learning to use AI-assisted translation tools is easy for me.
PEOU2. I find AI-assisted translation tools easy to use for translation tasks.
PEOU3. Using AI-assisted translation tools does not require much mental effort.

Section E. Feedback Clarity (FC)

FC1. The feedback provided by AI-assisted translation tools is clear and easy to understand.
FC2. The feedback clearly indicates what aspects of my translation need improvement.
FC3. The feedback helps me understand how to revise or refine my translation.

Section F. Perceived Translation Quality (PTQ)

PTQ1. The translations generated by AI-assisted translation tools are generally accurate.
PTQ2. The translations produced by AI-assisted translation tools are reliable across different texts.
PTQ3. The translations are appropriate in meaning and wording.

Section G. Cognitive Load Reduction (CLR)

CLR1. Using AI-assisted translation tools reduces the mental effort required for translation tasks.
CLR2. AI-assisted translation tools help me handle complex translation tasks with less cognitive strain.
CLR3. Using AI-assisted translation tools makes translation learning mentally easier.

Section H. Trust in AI-Assisted Translation Tools (TR)

TR1. I trust the translations generated by AI-assisted translation tools.
TR2. AI-assisted translation tools provide reliable translation outputs.
TR3. I feel confident relying on AI-assisted translation tools during translation learning.

Section I. Cognitive Understanding (CU)

CU1. The feedback provided by AI-assisted translation tools helps me better understand my translation errors.
CU2. AI-assisted translation tools improve my understanding of how to revise and refine translations.
CU3. By using AI-assisted translation tools, I gain a clearer understanding of the translation process.

Section J. Effort Reduction (ER)

ER1. AI-assisted translation tools reduce the amount of effort I need to complete translation tasks.
ER2. Using AI-assisted translation tools makes translation tasks less demanding for me.
ER3. AI-assisted translation tools simplify the process of completing translation work.

Section K. Behavioral Intention (BI)

BI1. I intend to continue using AI-assisted translation tools for translation learning.
BI2. I plan to use AI-assisted translation tools frequently in the future.
BI3. I am willing to recommend AI-assisted translation tools to others for translation learning.

Wang Xiaoyun is an undergraduate student majoring in English at Shandong Jiaotong University. Her research interests include AI-assisted translation and language technology. She has co-authored two papers on multilingual translation competence and AI-supported translation revision teaching.

Vladan Devedžić is a Professor of Computer Science and Software Engineering at the University of Belgrade, Faculty of Organizational Sciences. He is the founder and Head of the research group focused on Artificial Intelligence (GOOD OLD AI research network). He is also the founder of the Artificial Intelligence Laboratory at his home faculty. Since 2021, he has been a corresponding member of the Serbian Academy of Sciences and Arts (SASA) at the Department of Technical Sciences. According to the list of world's top scientists, published by Stanford University, he is among 0.6% of the most cited researchers in the "career" category in the field of Artificial Intelligence (for the period 1996-2025). His current interests include Artificial Intelligence, Programming Education, Software Engineering, and Educational Technologies.

Wan Yinjia is a professor and adjunct master's supervisor at Qufu Normal University and Shandong Normal University. Her research interests include AI-assisted translation and technology-enhanced pedagogy. She has published more than ten academic papers, including an article in an SSCI Q1 journal, and authored a monograph. She has also led several research and teaching reform projects. Her course E-Commerce English has been recognized as a National First-Class Hybrid Undergraduate Course, and she has received multiple national awards for excellence in foreign language teaching.

Received: January 12, 2026; Accepted: July 6, 2026.

Computer Science and Information Systems 23(4):1453–1484 https://doi.org/10.2298/CSIS260120032J

Paper Type: Research Paper

Leveraging Large Language Models for Scalable Analysis of the End-of-the-Course Student Feedback

Jelena Jovanovic[1], Irena Vodenska[2], and Vladan Devedzic[3,1,4]

[1] Faculty of Organisational Sciences, University of Belgrade
Jove Ilica 154, 11000 Belgrade, Serbia
jelena.jovanovic@fon.bg.ac.rs, vladan.devedzic@fon.bg.ac.rs (corresponding author)
[2] Administrative Sciences Department, Metropolitan College, Boston University
1010 Commonwealth Avenue, Boston, MA 02215
vodenska@bu.edu
[3] Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences), Jinan, China
[4] Serbian Academy of Sciences and Arts, Knez Mihailova 35, Belgrade, Serbia

Abstract. Analyzing open-ended student feedback in course evaluations is a labor-intensive task due to the unstructured and complex nature of natural language. While Large Language Models (LLMs) offer significant potential for automation, a well-defined methodology for their application in analyzing student feedback remains underdeveloped. This paper addresses this gap by proposing an LLM-based feedback analytics pipeline designed to transform students' open-ended feedback into structured, actionable insights. The pipeline consists of three sequential stages: (i) segmenting student feedback into semantic units and assigning polarity (sentiment) to those units; (ii) topical classification of semantic units, and (iii) summarization of units within each topical category. By systematizing these processes, the proposed method enables educators and course managers to efficiently derive meaningful patterns from vast datasets of student opinions. We evaluated the proposed method using a comprehensive dataset from several editions of a U.S. university course, yielding encouraging results of the method's effectiveness. This research provides a scalable, generic methodology for (semi-)automated feedback analysis, ultimately supporting data-informed improvements in teaching and course management.

Keywords: Student Feedback, Course Evaluation, Natural Language Processing, Large Language Models, Qualitative Coding.

1. Introduction

Student evaluations of teaching provide crucial quantitative and qualitative data for enhancing course design and delivery in higher education. This data can be analyzed and used by teachers and school management to improve course[1] organization, update course contents and resources, enhance course assignments and grading, and refine teachers' approaches and their communication with students [24]. Moreover, the data collected from

[1] Note that throughout the paper we use the term course to refer to an individual subject, not a whole degree

student evaluations may serve to institutional management as an important source of indicators of teachers' contributions for faculty promotion and tenure prospects.

It is rather straightforward to automate the analysis of quantitative data from student evaluations. However, it is much more challenging to analyze qualitative data, since it originates from open-ended student responses, which, as any textual data, needs to go through a natural language processing pipeline before it can be effectively used for obtaining insights from student comments. This paper presents a methodology based on Large Language Models (LLMs), aimed at facilitating analysis of this kind of qualitative data, ultimately leading to relevant information for instructors, course coordinators, and other relevant stakeholders in higher education institutions.

A key concept in the proposed methodology is qualitative coding, the process of categorizing (annotating) segments of text (in this case open-ended student responses), i.e. identifying major themes in these segments [36]. Traditionally, this has been a manual, labour-intensive task and as such significantly impeded large-scale qualitative data analysis. Aiming to overcome this challenge, researchers are increasingly using LLMs in qualitative coding of text in various educational applications, and the results are intriguing and encouraging. For example, in their 2024 study, Hou et al. [20] suggested fine-tuning of GPT models for better accuracy in coding along context-dependent dimensions of text (i.e., theorizing, integration, reflection), whereas the accuracy was satisfactory in coding along context-independent dimensions (i.e., appraisal, questioning, social, curiosity, surprise). Zhang et al. [47] explored the performance of their GPT-3-based tool in both inductive coding (discovering major themes in text segments) and deductive coding (labeling text segments according to a pre-specified coding scheme) on simulated and real datasets. Using inter-rater reliability measures, they found substantial agreement between human coders and their tool in various coding scenarios. Barany and colleagues used GPT-4 in an attempt to automate the codebook development process [3]. They concluded that LLMs could be valuable for improving qualitative codebooks, but human involvement in their development remained essential. Liu et al. [27] used GPT-4 in the context of qualitative coding of semi-personalized tutoring session transcripts and student observations and behavior across three different courses. They have specifically studied the inter-rater agreement between human coders and GPT-4 when using prompting techniques like Zero-shot, Few-shot, and Few-shot with contextual information. The results indicate that for good results each technique should be tailored to the specific properties of the construct and context being analyzed.

Yet, research related to using LLMs to support students' qualitative evaluations of teaching and courses remains scarce. These evaluations most often come in the form of natural-language answers/comments to open-ended questions, where challenges include: (i) the volume and unstructured nature of open-ended comments, and (ii) countless ambiguities in such comments, originating in the commenter's writing style and language use, thus exacerbating the process of "making sense" out of the overall student feedback. A few researchers have taken the challenges. For instance, Peña-Torres [35] conducted sentiment analysis of student evaluations using an LLM fine-tuned with a crowdsource-tagged dataset of students' comments and reported 93% accuracy in the classification of the comments. Gul et al. [18] compared multi-task learning with distinct LLMs on the tasks of aspect extraction and sentiment analysis across more than 11.000 comments in course evaluations from a public university. Maram et al. [30] took a different approach – they

developed three agent-based conversational LLM tools, one for teachers to create and test prompts to gather student feedback on the course and teaching, one for students to provide feedback, and one for teachers to view and analyze this feedback. More closely related to the current paper is the recent work of Parker and colleagues [34], which demonstrates the capability of LLMs to streamline distinct natural language processing workflows relevant to the analysis of students' responses to open-ended course evaluation questions. Similarly, Fuller at al. [17] examined the efficiency and effectiveness of LLMs in identifying the main themes in students' open-ended responses to a course evaluation survey.

Having conducted a detailed analysis of these and similar efforts (see Section 2.2), we have identified a research gap that has driven the current study: the lack of a clearly defined and fairly general LLM-based methodology to automate the processing and interpretation of student open-ended feedback on end-of-the-course surveys. Thus this study proposes such a method, with the ultimate goal of enabling the transformation of raw, qualitative feedback into structured, actionable insights for instructors and course administrators. We refer to this method as *feedback analytics pipeline* and demonstrate it on a rich dataset of student end-of-the-course feedback, comprising more than a decade of feedback data from a course taught at a U.S. university.

The following sections detail the relevant literature review, the proposed feedback analytics pipeline, the application and evaluation of the pipeline on the aforementioned dataset from a U.S. university, and a discussion of the findings and future work.

2. Literature Review

2.1. End of the course student feedback

Student feedback is a central data source in educational research and scholarship of teaching and learning, providing relevant information about learners' experiences with instruction, assessment, and the learning environment. In higher education, this evidence most commonly takes the form of student evaluations of teaching (SETs), typically combining Likert-like ratings with open-ended comments. Historically, quantitative evaluations have been the focus of analysis due to their ease of evaluation, while the (rich) qualitative student feedback has been largely overlooked, given the complexity of manual coding. However, the qualitative component of student feedback is important for providing enhanced signals of course and instructor quality, content, pedagogical approaches, and ultimately, for course improvement.

Research positions student feedback as one component of an evidence-informed approach to teaching evaluation and course enhancement, ideally triangulated with peer review, teaching artifacts, and indicators of student learning [5,31,40]. Conceptually, student feedback can be framed in two complementary ways. First, as *measurement*, a set of instruments intended to capture distinct dimensions of instructional quality (e.g., organization, clarity, engagement). Second, as *student voice*, a means of incorporating learners as legitimate informants and partners in improving teaching and curricula [7,14]. This framing matters because it shapes both the purposes of feedback (formative improvement vs. summative judgment) and the analytic methods used to interpret it. Both perspectives are important to our study as our objective is to support continuous improvement of course offerings and instruction through the integration of students' voice (i.e. feedback).

Open-ended student comments can add diagnostic value beyond numeric ratings by providing context, identifying specific issues with the course design, and offering actionable suggestions. Empirical work on written comments shows that students frequently address topics such as clarity of instruction, course organization, relevance/interest, assessment demands, and instructor-student relations [9,33]. Written feedback is often particularly useful for identifying *what* to change and *why* a rating might be high or low, thereby supporting targeted improvement. Moreover, qualitative feedback aligns with an improvement-oriented view of teaching as iterative design. In constructive alignment and reflective teaching traditions, instructors refine intended learning outcomes, teaching/learning activities, and assessment through cycles of evidence and reflection [6,10]. Importantly, mid-semester feedback processes (e.g., small group instructional diagnosis or structured midterm surveys) can be especially useful as they enable changes while a course is still running [13,15].

Because open-ended student comments can be voluminous, idiosyncratic, and uneven in quality, their value depends heavily on systematic analysis. Rigorous approaches typically involve qualitative coding, development of a codebook, iterative refinement, and checks on credibility (e.g., intercoder agreement for descriptive coding, audit trails, and reflexive memoing) [8,36,38]. When conducted carefully, qualitative analysis can (i) separate actionable course-design issues from non-actionable preferences, (ii) detect recurring patterns across cohorts, and (iii) surface equity-relevant experiences that may be obscured in averages. The best practice is to integrate qualitative themes with quantitative indicators. For example, instructors and departments can map themes (e.g., "assessment clarity") to course elements (rubrics, workload pacing, alignment) and then evaluate whether subsequent iterations show changes in both student feedback and learning outcomes [5,6].

Recent work explores computational approaches to analyzing large corpora of student comments, including sentiment analysis, topic modeling, and structural topic modeling. Topic modeling using LDA (Latent Dirichlet Allocation) has been proposed as a scalable alternative to manual coding, enabling institutions to identify common themes and examine variation by course level or subject area [21,42]. Structural topic modeling extends this by incorporating covariates (e.g., discipline, enrollment, instructor characteristics) and can link comment themes to numeric SET outcomes [49]. These approaches can support system-level sensemaking, but they still require validation against human interpretation and careful governance to avoid reifying bias [19,41].

Overall, the literature supports a balanced conclusion: student feedback is most valuable when treated comprehensively as *evidence for improvement* rather than a single high-stakes score, and when qualitative comments are analyzed systematically using transparent, methodologically sound approaches that enhance actionability and fairness [5,19,40].

2.2. LLMs in the analysis of educational content and feedback

Open-ended text from course evaluations and other engagement surveys and narratives, contains high-value signals that are often underutilized because key stages of traditional qualitative workflows (such as iterative coding, codebook development, and theme refinement) are time- and labor-intensive [8,11,36]. Over the last two decades, topic modeling enabled scalable text summarization and clustering, but they tend to struggle with nuance, domain shifts, and interpretability in short, heterogeneous responses [1].

Large language models (LLMs) have renewed interest in automation of qualitative content analysis by enabling flexible, instruction-driven analytic pipelines: segmenting responses into semantic units, labeling sentiment or stance, generating candidate themes, drafting codebooks, and producing human-readable summaries [23, 34]. In education, LLMs have been tested, for example, as tools to accelerate the extraction of main themes from student feedback while maintaining alignment with instructor interpretations [17]. In parallel, a broader methodological literature has emerged on LLM-as-a-judge evaluation frameworks, enabling scalable quality assurance but introducing known risks such as bias, verbosity effects, and sensitivity to prompt framing [28, 43, 48].

The application of LLMs to educational data is a growing field. Research by Katz et al. [23] has demonstrated that LLMs can be used to effectively build qualitative codebooks for student evaluations of teaching, achieving both scalability and consistency, with categories closely aligned with human-generated frameworks. Their method includes extracting units (sentences/phrases), embedding them, clustering semantically similar items, and then prompting an LLM to name clusters, write definitions, and produce a human-usable codebook. This work demonstrated significant reductions in the time and cost typically associated with manual thematic analysis. Furthermore, it addresses a limitation of purely statistical topic models (often hard-to-interpret topic-word distributions) by using LLMs to generate coherent labels and descriptions. Recent evaluation work suggests that LLM-based topic modeling can produce coherent topics but may also exhibit shortcuts (focusing on partial documents) and limited controllability via prompts, reinforcing the need for diagnostic checks and human review [16].

Parker et al. [34] showcased that with effective prompting, LLMs can achieve human-level performance in analyzing educational survey feedback. Their findings highlight the LLMs' flexibility in supporting both inductive (bottom-up) thematic analysis and deductive (top-down) classification, enabling faster, more reliable, and rigorous interpretation of student feedback than traditional methods. LLMs have also been used to generate themes for predefined prompts and compare them to instructor-derived themes, reporting substantial agreement while reducing analysis time [17]. This approach resembles assisted deductive coding, where the researcher supplies the conceptual frame, and the LLM performs standardized extraction across multiple responses.

Building on these advancements, summarised in Table 1, our study proposes and implements a comprehensive LLM-assisted feedback analytics pipeline, which includes decomposition, sentiment analysis, classification, and summarization of student open-ended responses to the course evaluation survey, specifically tailored for providing course instructors and coordinators with well structured and actionable information for course improvement. The objective is to support teachers' agency by making student feedback accessible to them and then letting them decide what is relevant and whether / how they want to act on it. In other words, the objective is to empower teachers to make use of the students' qualitative feedback that has so far been mostly ignored due to technical difficulties of natural language processing.

3. Methodology: an LLM-assisted feedback analytics pipeline

Figure 1 illustrates the data flow and key steps in the proposed feedback analytics pipeline. Each step is described in detail in the following subsections.

Table 1. Summary of related work

Objective	Analytics tasks	NLP methods / tools	Reference
Enable scalable and theoretically grounded analysis of students' open-ended feedback	Topic detection and thematic analysis	Topic modelling (LDA)	Hujala et al., 2020 [21], Sun et al., 2023 [42]
Enhance teaching practices through sentiment analysis of students' course feedback	Classification of students' comments based on the expressed sentiment	Fine-tuned language models	Pena-Torres [35]
Identify the best model for aspect-based sentiment analysis of students' comments	Aspect extraction and sentiment classification	Multi-task learning with three pre-trained language models and selective paraphrasing as a data augmentation technique	Gul et al., 2025 [18]
Enable teachers to collect, view and analyze student feedback in a more engaging and timely manner	Extraction of key themes and summarisation of student-chatbot conversations	LLM-based (GPT-4) conversational agents for student feedback collection, analysis, and presentation (via dashboard)	Marram et al., 2025 [30]
Examine and demonstrate how LLMs, employed in a variety of NLP tasks, can be used to derive insights from course feedback surveys	A variety of LLM-based NLP tasks, including classification, information extraction, thematic and sentiment analysis	LLMs (GPT-4 and GPT-3.5)	Parker et al., 2024 [34]
Examine how efficient and aligned with instructors LLMs are in identifying main themes in students' feedback	Thematic analysis and percentage agreement to determine alignment between instructors and ChatGPT	LLM-based chatbot (ChatGPT, model version not specified)	Fuller at al., 2024 [17]
Demonstrate how a combination of NLP techniques and LLMs can be used to generate a codebook for SET analysis	An NLP-based pipeline for developing a codebook for qualitative analysis of student feedback	A sequence of NLP methods: information extracting and embedding, followed by clustering of embeddings, and summarization of the clusters	Katz et al., 2024 [23]

Note: All presented studies except [30] focused on the students' open-ended responses to a course evaluation survey; the study reported in [30] differs in that its focus was on student-chatbot conversations through which course-related feedback was elicited.

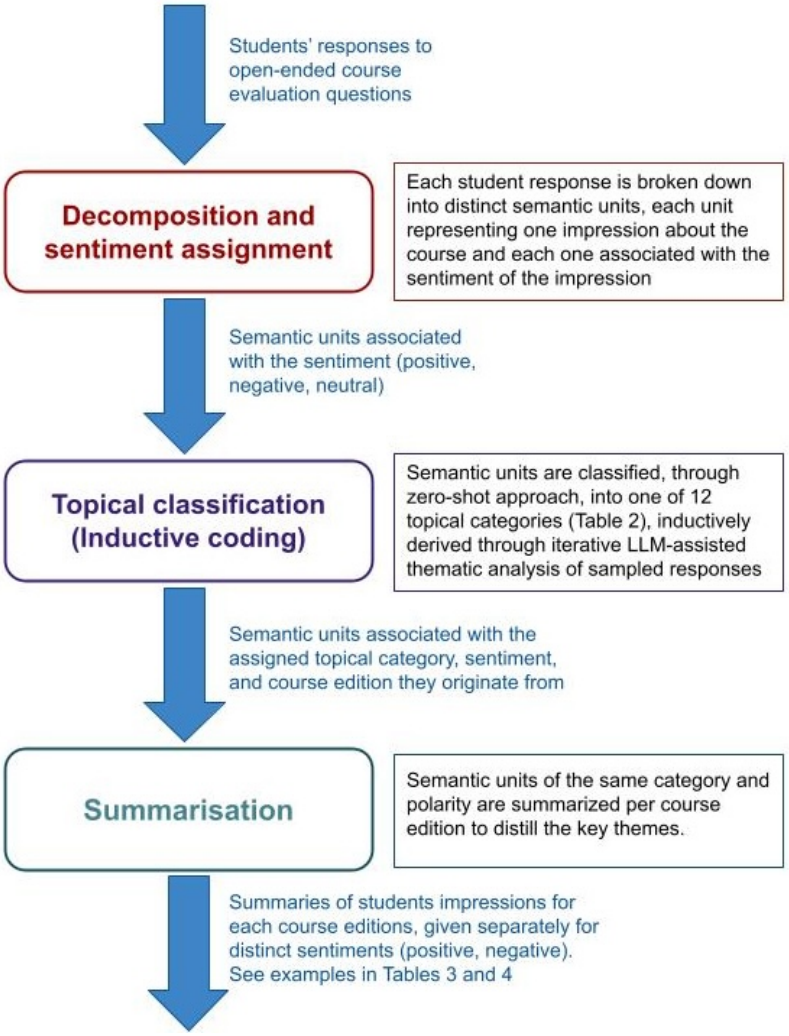


Fig. 1. Flow chart depicting the key stages of the LLM-assisted feedback analytics pipeline

3.1. Decomposition and sentiment analysis of students' comments

End-of-course surveys typically include open-ended questions that explicitly ask students what they liked and found positive about the course itself, the course instructors, or the technology used in the course. Such questions are often complemented by questions asking students about what could be improved in the course itself, in how it is taught, and in the technology supporting the course delivery. Alternatively, students may be presented with general open-ended questions asking them to comment on what they liked or disliked about the course, the instructors, or the supporting technology. This suggests that questions may have well-defined polarity (sentiment) - being either about positive or negative aspects of the course as perceived by the students - or be more broadly framed, aiming to collect students' both positive and negative sentiments about various aspects of the course. This further implies that, among students' responses, there will be responses expressing only positive sentiments, those expressing what students perceived as weaker aspects of the course, and responses expected to contain a mixture of both positive and negative comments. In addition, in all the responses, it is realistic to expect that they would often contain not an opinion on a single aspect of the course, but rather opinions about multiple distinct aspects of the course.

The above described characteristics of students' responses indicate the need to decompose each response into distinct semantic units, each one corresponding to one impression / opinion about the course. Semantic units can be of arbitrary length: a semantic unit can correspond to a sentence or can be a part of a sentence (in case of a longer sentence where a student expresses multiple opinions about a course) or consist of a few sentences (in rare cases when a student elaborates on an expressed opinion). In addition, for general questions (i.e., those without explicitly assigned polarity), the sentiment (positive, negative, neutral) of each semantic unit also needs to be determined. Accordingly, the first stage of our analytics pipeline checks whether a response contains multiple impressions (or expressions of opinion) and, if so, splits it into multiple semantic units, each one containing a single impression. In case of a response to a question with clearly defined polarity (i.e., asking students about positive or negative impressions), the extracted semantic units are assigned the same polarity as the question (e.g., positive sentiment, if the question asked about positive aspects of the course). In case of a general question, that is, a question expecting from students to share both their positive and negative impressions, each semantic unit extracted from a student response also needs to be annotated with the sentiment it expresses. To properly cover both cases, we have designed two alternative response decomposition processes:

- For responses to questions with clearly defined polarity, we prompt an LLM to identify all distinct course aspects the student was positive/negative about in their response and return them as a list of distinct impressions about the course. The LLM is also requested to avoid redundancy in the results, which may occur if a student comments on the same or very similar course aspects multiple times, by summarizing the redundant comments before including them in the result.
- For general questions without explicitly assigned polarity, in addition to the above-stated instructions, we also ask the LLM to associate each extracted semantic unit with the expressed sentiment (positive, negative, or neutral).

System prompts used in both versions of the decomposition process are provided in Figure A1 and Figure A2 of the Appendix.

3.2. Topical classification

The objective of this step is to associate each semantic unit (resulting from the previous analytics stage) with a topical category that best characterizes the student's impression, as expressed in that unit[2]. To this end, we have defined a classification (coding) scheme as a set of topical categories with associated descriptions (Table 2), following the procedure described in the next paragraph. This scheme was used to prompt the LLM to do the classification task. Specifically, the LLM is prompted to associate the given student comment (i.e., semantic unit extracted from the student response) into one of the 11 topical categories, with each category being described in the prompt (as in Table 2). Aiming to prevent the LLM from assigning a random category when it cannot determine the correct one for a particular comment, the LLM is instructed to assign the "Other" category to a comment that does not belong to any of the given categories. The system prompt used to instruct the LLM to classify individual semantic units is given in Figure A3 of the Appendix.

The categories within the classification scheme were identified through an iterative exploratory process where the LLM was first prompted to identify the main themes within a set of semantic units (N=500) randomly sampled from student feedback in a university course that served as our initial application case (see Section 4.1). The identified themes were checked against the coding schemes used in related prior work [34] and served as the initial set of topical categories that were used to classify another random sample of semantic units (N=250), and both the misclassifications and the assignments to the "Other" category were used to revise the classification scheme and repeat the classification step. After three iterations of classification followed by revisions of the classification scheme (categories and/or descriptions), misclassifications were substantially reduced while assignments to the "Other" category were limited to those semantic units that truly could not be considered as useful (i.e., interpretable) pieces of feedback (e.g., "Time management skills", "Class preparation"). In deciding upon the classification scheme to adopt, we also consider the optimal number of categories, since a larger number of categories (i.e., more fine-grained classification) allows for more precise and nuanced organisation and analysis of student comments, but, at the same time, may negatively affect precision / correctness of the classification output (more categories tend to make the classification task more error prone). The chosen classification scheme is the one that provided a good balance between the two opposing objectives. We reflect on this further in the Discussion section. Even though this classification scheme is derived from students' comments in one course, it is rather general and may be applied, at least as a first version, to any university course for further refinement.

3.3. Summarization

The objective of this step is to create, for a particular course edition, a summary of the student comments for each topical category (Table 2) so that the course coordinator and instructors can be provided with a gist of students' impressions about distinct aspects of

[2] Note that, in line with the computer science terminology, we use the term "topical classification" to refer to this step of the analytics pipeline; in qualitative analysis, this task would be referred to as "inductive coding" as it includes the development of a coding scheme and the use of that scheme to code (annotate) textual content

Table 2. The coding scheme for topical classification of students' comments

Category	Description
Course content	This category refers to subject-matter topics taught in the course, including domain-specific skills and competences.
Course organization and logistics	This refers to all aspects of the course organisation, including scheduling of lectures and labs, timeliness of course announcements, the availability and accessibility of the learning tools and the like.
Course value and relevance	This refers to the perceived course value and relevance, both from the academic and practical perspective.
Course materials and resources	This category refers to all kinds of materials and resources that were provided to the students, recommended to them, or requested from them to use.
Course assignments and grading	This category refers to all kinds of assignments in the course, both formative (not graded, for practice only) and summative ones (graded), as well as the grading system applied in the course.
Instructor's teaching approach	This category refers to various aspects of the instructor's teaching approach, including different aspects of the instructional design and interaction with the students during the class.
Instructor's knowledge and expertise	This refers to the instructor's competence, as perceived by the student, for the topics included in the course syllabus.
Instructor's clarity and explanations	This category refers to the instructor's clarity of expression and quality (comprehensibility) of explanations; in other words, the quality of communication of the subject matter topics to the students.
Instructor's feedback and support	This is about the frequency, quality, and usefulness of the feedback on course assignments and the overall support.
Instructor's demeanor and communication with students	This category is about the politeness of communication with students, friendliness, and availability for communication.
Overall impression	This refers to the student's overall impression about the course and/or the instructor.

that course delivery. To that end, an LLM is given a set of student comments (semantic units) from one course edition and related to one topical category and prompted to identify the main themes in the students' comments, and for each theme, to name it and to describe it succinctly. For each course edition, the LLM is tasked with the same prompt 11 times, each time with a set of students' comments categorised into one of the 11 topical categories (those from Table 2). The system prompt used for this task is given in Figure A4 of the Appendix.

3.4. Evaluation of the pipeline

The presented feedback analytics pipeline uses an LLM for four main tasks: splitting student responses into logically and topically cohesive units (semantic units), sentiment analysis, topical classification, and summarization. The potential of LLMs for semantic text segmentation, sentiment analysis, and summarization tasks have already been explored and evaluated in prior research (e.g., [4, 22, 29, 37, 39, 44]). Therefore, we focus the evaluation primarily on the topical classification part of the pipeline, which, due to the fairly large number and closely related semantics of topical categories, may be error-prone.

To validate topical classification of students' comments (semantic units) in both scalable and trustworthy manner, we have adopted a semi-automated two-step approach based on the LLM-as-a-Judge concept [48]. LLM-as-a-Judge is about using a highly capable LLM to automatically evaluate the quality of outputs generated by another, often less capable, LLM. It offers a scalable, cost-effective alternative to human evaluation for various LLM-based tasks, including qualitative coding and text classification. The LLM-Judge is provided with: (i) the original task (in our case, topical classification of a student comment), (ii) the LLM output that is to be evaluated (the category assigned to the student comment), and (iii) the criteria to be used for the evaluation.

The first step of our evaluation approach includes both human evaluation of a randomly sampled subset of students' comments and evaluation of the same subset by an LLM prompted to serve as a Judge. The objective of this step is to ensure that the LLM-Judge can reliably validate the topical classification of student comments on a smaller sample before applying it to validate a larger sample of comments. To this end, two human experts are asked to independently evaluate the correctness of the topical classification done in stage 2 of the pipeline. Specifically, for each student comment from the randomly chosen subset, the human evaluators are instructed to either confirm that the topical category assigned by the LLM classifier is correct or to mark it as incorrect and note down the category they deem correct (out of the 11 categories in Table 2 plus the 'Other' category). After human experts complete this task and resolve any inconsistencies in their individual evaluations, the resulting joint evaluation is considered the "ground truth" to be used for assessing the correctness of the evaluation done by the LLM-judge.

The LLM-judge is given the same instructions as human evaluators, namely to verify if a student comment is associated with the correct topical category. In addition, the LLM-judge is prompted to provide the rationale for its decision. The motivation for this additional request is twofold: i) it has been shown that LLMs tend to produce better results if they are prompted to "think" before generating the output [12]; ii) to be able to trace the LLM's "thinking process", especially in erroneous cases, so that the instructions (i.e., the prompt) can be refined. The complete system prompt for the LLM-judge is given in Figure A5 of the Appendix. To determine the level of alignment between the LLM-judge and

the ground truth produced by the human experts, we use measurements that are typically used for assessing the agreement of two human raters, namely the percentage agreement and Cohen's Kappa [46].

After the LLM-judge is assessed as reliable for verifying the classification task, it is used in the second step of the evaluation process, to evaluate either the entire set of student comments or a sufficiently large (random) sample that can ensure trustworthy classification of the entire set. While evaluating the entire set of comments is preferred, its feasibility depends on the funds and infrastructure available for feedback analytics; in cases of limited resources, a randomly sampled subset may be a viable alternative (as, e.g., in [32]).

In addition to examining the correctness of the topic classification, we also examine the internal consistency of the LLM classifier. To that end, we run the classification task five times on the same random sample of student comments (semantic units) and compute the level of agreement in the classification results (i.e., percentage agreements) between each pair of runs.

4. Application case: A Financial Management course

4.1. The application context and data source

The analysis utilizes student feedback from a standalone graduate-level university course in financial management, spanning 43 semesters from 2013 to 2025. The course was run in multiple distinct course sections per semester, four semesters per year: Fall, Spring, and two Summer terms. The raw data used in this study consists of anonymised open-ended student responses from course evaluation surveys. The open-ended questions in these surveys changed a couple of times during the examined time period, but their gist remained unchanged: asking students to comment on what they liked about the course and the instructors, what they thought could be improved in the course (e.g., course content, the way it is taught, the technology used), and what was their overall impression about the course. Some of the questions were positively intoned (e.g., "What were the most valuable aspects of the course?"), others were explicitly asking for suggestions for improvements (e.g., "Please comment on aspects of this course that might be improved"), and there were also questions neutral in tone, asking students to share any thoughts they have about the course (e.g., "Feel free to provide any additional comments you believe would be useful about the course. The following topics offer suggestions, but please do not feel limited to them. What are the strongest aspects of the course overall? What can be improved? What additions/changes would you suggest?").

The original data consisted of 5892 student open-ended responses. However, some of the responses were overly brief and not bearing meaningful information. For example, responses such as "no", "nope", or "no comment" to a question asking students if they had any additional comments about the course to share. Furthermore, there were also questions that were not course-focused but asking students about their overall campus experience, their interest in other courses, or asking them about their motivation for enrolling in the given study program. After removing overly short responses and responses not directly related to the course experience, we ended up with 4679 student responses, which served as the input to our feedback analytics pipeline. From the perspective of the

above-mentioned three categories of open-ended questions, 1977 responses originated from positively intoned questions, 154 responses from negatively intoned ones (i.e., those asking for suggestions for improvement), and 2548 were given in response to the general ones (neutral in tone). The use of this data for the study presented in the paper is approved by the institution's internal review board, under the approved protocol number 8467X.

4.2. Feedback processing results

All three stages of the feedback analytics pipeline (outlined in Figure 1) were executed with the GPT-4o-mini, as a representative of the state-of-the-practice LLMs with a very reasonable price-performance trade-off. Since student survey responses were completely detached from any other student data (i.e., no student data apart from their responses to the course evaluation survey was available in the dataset) and the response data was fully anonymised, there were no privacy risks of using externally hosted LLM.

The first stage of our feedback analytics pipeline produced 9,674 student comments (semantic units) with positive sentiment, 6,044 comments with negative sentiment, and 794 neutral comments. The distribution of positive and negative comments, which were the subject of further analysis, is presented in Figure 2.

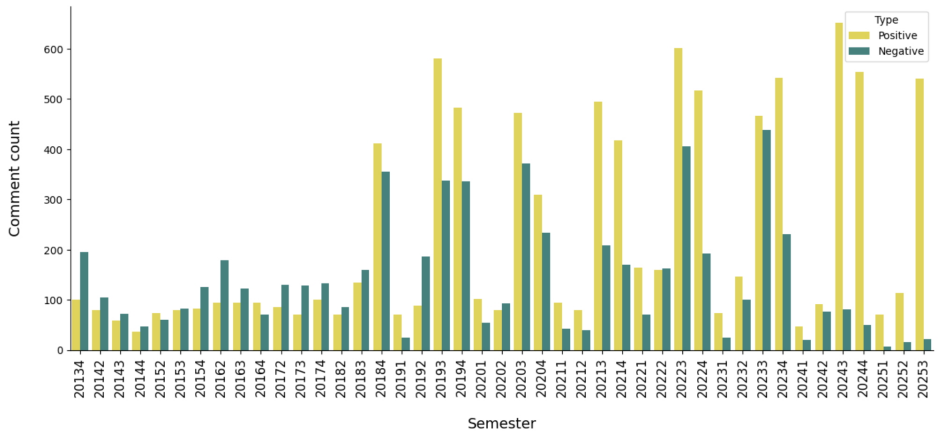


Fig. 2. Distribution of positively and negatively intoned student comments (semantic units) over the examined time period. The values on the X-axis refer to distinct, chronologically ordered course editions, consisting of the year and semester identifier (e.g., "20253" refers to the 3rd semester in year 2025).

As Figure 2 indicates, the amount of feedback collected from students differed over time, with the numbers significantly increasing for the first time in the last semester of 2018 and then oscillating over the remaining years. It is also evident that positive comments tend to outnumber negative ones, in some semesters being even many times higher than negative ones. This may be partly explained by the dynamics of student enrollment throughout an academic year. In particular, semesters labeled as 1, 2, 3, and 4 correspond

to Summer 1, Summer 2, Fall, and Spring, respectively, with the summer semesters having fewer students than the Fall and Spring ones, when the majority of students enroll in courses. After 2017, there has been a significant growth in student enrollments due to a complete overhaul of the management programs, where the course that is the focus of our study has become a core course in different study programs, including marketing, business analytics, innovation, project management, enterprise risk management, and supply chain management. Additionally, the course itself has undergone redevelopment during this transition period, which may also explain the increase in the students' positive comments and their dominance over the negative ones. The course is continually updated, as there is always room for improvement, especially in course organization, resources, and assessments, as will be confirmed by our subsequent analysis.

The next stage in the feedback processing pipeline was topical classification of positive and negative comments. The resulting distribution of positive and negative comments across the topical categories is shown in Figure 3. More precisely, the figure shows the proportions of positive and negative (i.e., improvement-oriented) comments assigned to each topical category relative to the total number of student comments. This overall distribution suggests that students had far more positive than negative overall impressions about the course during the examined time period (2013-2025). They were also predominantly positive about the course instructors. Likewise, the course content and its value and relevance were the subject of primarily positive student remarks. On the other hand, the course assignments and grading, course materials and resources, as well as the course organization and logistics received many more comments on potential improvements than praises.

While Figure 3 provides a structured insight into the students' opinion about different aspects of the course, it is limited by being based on comment counts during the overall examined time period. For course coordinators and instructors, insights into the trends over a specific time period (e.g., before and after certain instructional intervention) could be more informative and actionable. With student comments associated with the course edition they originate from and annotated with topical categories, such temporal views can be easily created. For example, Figure 4 presents one potential approach to presenting trends in students' positive opinions throughout 2024 and 2025 semesters (a similar one can be seamlessly applied to negative opinions, that is, suggested improvements). The figure shows, for each semester, how positive comments were distributed across the 11 topical categories. For example, one may notice that the second semester of 2024 ("20242") distinguishes itself by a notably higher proportion of positive overall impressions than the other semesters in 2024 and 2025, and that this may be primarily attributed to the positive impressions of course instructors.

Finally, the summarization stage produced for each course edition and each topical category a set of main themes, each one accompanied by a brief description. As an illustration of the summarization results, Table 3 presents a feedback summary for the Instructor's teaching approach category in the 2nd semester of the 2022 course edition, while Table 4 provides the same kind of summary for a course-focused category, namely Course assignments and grading, in the same course edition. The output of this phase may be particularly relevant to course instructors as it provides deep insight into distinct aspects of the course and may offer information for course and/or instructional revision or redesign.

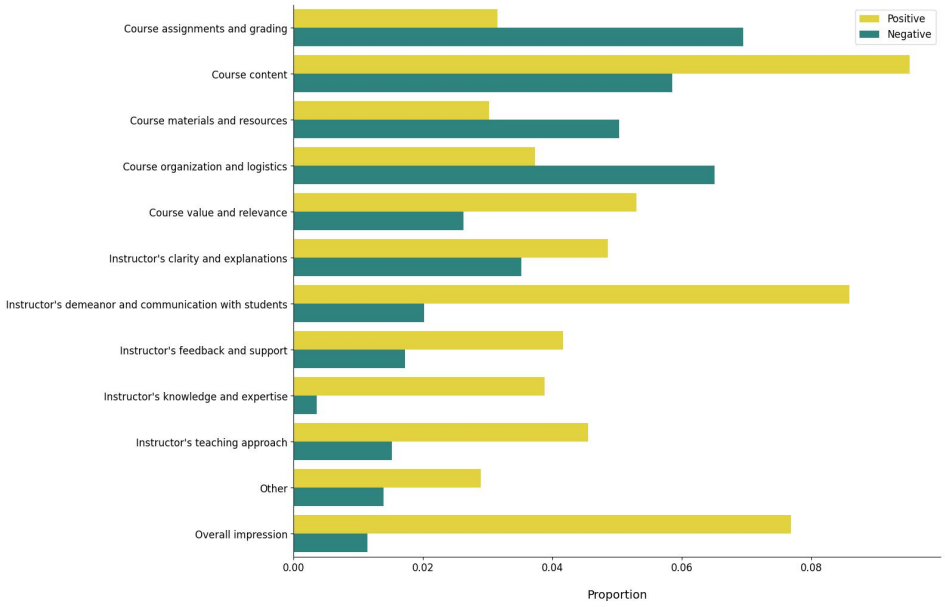


Fig. 3. Distribution of positive and negative student comments (semantic units) across the 11 topical categories plus "Other". The values on the X-axis are the proportions of distinct categories of positive / negative comments in the total number of student comments.

Considering the complexity of a formal validation of the summarisation task, we opted for the approach adopted in similar prior studies (e.g., [34]), namely an informal evaluation by a course instructor. In particular, one of the paper co-authors, who has been teaching this course for several years, reviewed a random sample of summaries and confirmed that they accurately reflect the content of the student comments and offer actionable information relevant to the course instructors.

4.3. Evaluation of topical classification

To evaluate the topical classification of student comments (semantic units), we followed the procedure detailed in Section 3.4. For the first step of the evaluation procedure, we have randomly selected 50 positive and 50 negative (i.e., improvement-focused) student comments from each of the last five years (2021-2025). This resulted in 250 positive and 245 negative comments, for a total of 495 comments. The lower number of negative comments is due to the fact that the number of negative comments in the first three semesters of 2025 was limited to 45 (the data was not available for the 2025 Fall semester). Both positive and negative comments were evaluated by two of the paper authors. Both of them are university professors with over twenty years of experience in teaching and course management. Due to these roles, they are highly familiar with and experienced in reading and analysing student feedback. One of them has been teaching for several years the course that the comments originate from. Considering all this, the two human raters are well qualified to understand and classify students' course evaluation comments. The



Fig. 4. Distribution of students' positive comments across topical categories over semesters of 2024 and 2025 academic years. Note that all the plots have the same x-axis and to avoid the clutter, we show axis labels only for the plots in the bottom row. Likewise, labels on the y-axis are given only for the left most plots and the same y-axis applies to all the plots.

Table 3. An example of feedback summary for the Instructor's teaching approach category in one course edition

Topical category: Instructor's teaching approach	
Summary of positive comments	**Instructor Effectiveness**: Students praised the instructors for their thoroughness, preparation, and ability to present information clearly, contributing to a positive learning experience.
	Engagement in Live Sessions: Many students expressed appreciation for the live instruction and interactive homework sessions, highlighting their effectiveness in enhancing understanding of the material.
	Flexibility in Teaching Methods: Facilitators were commended for offering various methods of problem-solving, allowing students to choose the approach that suited them best.
	Course Structure: The organization and structure of the course were noted as key factors that contributed to its success and the students' learning outcomes.
Summary of negative comments	**Facilitator Effectiveness**: Students felt that the facilitators did not effectively engage with the material or provide adequate explanations.
	Need for Practical Examples: There is a desire for facilitators to incorporate real-life examples from their experiences to enhance understanding of the topics.
	Learning Environment: The live class environment was perceived as lacking in interactivity and depth, making it difficult for students to learn effectively.
	Material Presentation: Comments indicated that the material was often presented in a way that felt repetitive and not sufficiently taught, leading to confusion.

Table 4. An example of feedback summary for the Course assignments and grading category in one course edition

Topical category: Course assignments and grading	
Summary of positive comments	**Homework Effectiveness**: Students found the homework sessions and assignments to be beneficial for understanding course material and preparing for exams.
	Feedback Mechanism: The embedded homework modules with instant feedback were appreciated for enhancing learning and providing clarity on problem-solving approaches.
	Grading Fairness: Students felt that the grading was fair, especially given the complexity of the assignments and the tight timelines.
	Support Sessions: Homework help sessions and exam preparation classes were highlighted as valuable resources for reinforcing knowledge and identifying gaps.
	Relevance of Assignments: Assignments were seen as closely related to the course content, aiding in the overall learning experience.
Summary of negative comments	**Grading Discrepancies**: Students reported inconsistencies and errors in grading, leading to frustration and confusion regarding their scores.
	Homework Confusion: Many students found the homework assignments on Connect to be confusing and counter-intuitive, particularly with the use of Microsoft Excel.
	Discussion Board Issues: Feedback indicated that the discussion board grading was lacking, with no comments made and a perception that it was a waste of time.
	Workload Concerns: Students expressed that the workload was excessive, with too many assignments and readings that did not align well.
	Preference for Human Grading: There was a strong preference for assignments to be graded by a person rather than a computer, as students felt this would reduce errors and improve fairness.

same sample of positive and negative comments was also evaluated by GPT-5-mini in the LLM-Judge role (for instructions given to the model, see Section 3.4).

Human experts first evaluated the LLM classifier independently of one another, then jointly reviewed disagreements and resolved them, reaching complete agreement. Thus obtained human evaluation served as the "ground truth" for computing the evaluation measures for both the LLM classifier and LLM judge.

The number and percentage of correct topical classifications, done by the LLM classifier, as evaluated by both human experts and the LLM-judge, are given in Table 5. As the table suggests, the LLM classifier (GPT-4o-mini) tends to be more successful in classifying positive than negative comments. In addition, the LLM-judge (GPT-5-mini) seems to be quite well aligned with human experts when judging the correctness of comment classification. To further explore the alignment between the human experts and the LLM-judge, we have computed and reported the matching ratio and Cohen's Kappa in Table 6. While the matching ratios are high for both positive and negative comments, Cohen's Kappa values indicate substantial agreement for positive comments and moderate agreement for negative comments [25].

Table 5. Evaluation of the topical classification by human experts and LLM-judge, given as the number and the proportion of correct classifications

	Positive comments	Negative comments
LLM Classifier - Human	221 (0.884)	212 (0.865)
LLM Classifier - LLM Judge	223 (0.892)	211 (0.861)

Table 6. Measures of the alignment between the human experts and the LLM-judge on the evaluation of the comments classification task

	Matching ratio		**Cohen's Kappa**	
	Positive comments	Negative comments	Positive comments	Negative comments
LLM Judge - Human agreement	234 (0.936)	220 (0.898)	0.678	0.568

Considering this imperfect agreement, we have examined in detail the cases of misalignment between the human raters and the LLM-judge. Table 7 shows the category pairs that occurred in the misaligned cases more than once. The *Instructor's teaching approach* and *Course content* categories were the ones with the highest presence in the unaligned evaluations (23 out of 41 cases in total); likewise, *Course organization and logistics* and *Instructor's demeanor and communication with students*, occurred a couple of times. For the presented pairs, as well as those that occurred just once and are not shown here, it is clear that they are closely related in meaning, with very subtle differences, and thus

occasional misalignment is expected, especially considering that the students' comments (semantic units), when extracted from the overall student response, can be vague.

Finally, to examine the internal consistency of the LLM classifier, we reran the classification task on the same sample of 495 student comments and computed the average percentage of matches across all pairs of runs. The average matching is 0.988, with 11 to 13 mismatches out of 495 classifications, per run. This indicates high internal consistency of the LLM classifier.

Table 7. Frequent kinds of misalignment between human raters and LLM-judge

Category 1	Category 2	Frequency
Instructor's teaching approach	Course organization and logistics	9
	Instructor's demeanor and communication with students	2
	Instructor's clarity and explanations	2
Course content	Course value and relevance	4
	Course organization and logistics	3
	Course materials/resources	3
Instructor's feedback and support	Instructor's demeanor and communication with students	3
Course assignments and grades	Course organization and logistics	2

5. Discussion

This paper presents a practical, LLM-based methodology for scalable analysis and information extraction from students' open-ended feedback, collected via an end-of-course evaluation survey. Through the main steps of the proposed feedback analytics pipeline (Figure 1), students' feedback, from one or a series of course editions, is transformed from raw natural language responses, to a set of main themes, each associated with one of the eleven distinct course or instructor-related aspects of a course evaluation (Table 2). This output may serve as a valuable source of information to instructors and course coordinators for course and instruction (re-)design. It may also offer course coordinators and school management insights into trends in students' opinions about specific courses and instructional aspects over longer periods, and thereby inform course-level and program-level decisions.

The presented application case (Section 4) illustrates the method's output in the context of a graduate-level university course. Specifically, Tables 3 and 4 illustrate summaries, in the form of a set of key themes, that the method generates for each of the eleven course aspects that emerged as relevant in the overall pool of student responses. While these summaries may be primarily relevant to the course instructors and facilitators who taught the corresponding course edition, a visual depiction of the change in the students' opinion over time, such as the ones shown in Figures 3 and 4, may help the course coordinator to better understand the effects of the applied course / instructional changes over time.

This work complements prior related work, especially the work by Fuller at al. [17] and Katz et al. [23]. The former examined how efficient and aligned with instructors LLMs were in identifying main themes in students' feedback, whereas the latter demonstrated how a combination of NLP techniques and LLMs could be used to generate a codebook for student feedback analysis. In the current work, we show how a codebook, once developed, can be systematically employed in a feedback analytics pipeline to identify main topical categories and use these to prepare topically-focused feedback summaries for instructors. Our work is most similar to the work of Parker et al. [34] who did a comprehensive exploration of LLMs use in a variety of NLP tasks, for deriving insights from course feedback surveys. While they have explicitly covered a wider span of potential scenarios for leveraging student feedback, our solution has the advantage of being more versatile due to the initial step of decomposition of the feedback into semantic units, which together with the assigned sentiment may serve as suitable input for a variety of feedback analytics scenarios.

The evaluation of the classification step (Section 4.3) has shown that neither the LLM-classifier nor the LLM-judge is fully aligned with the human experts. This is in line with recent research that studied LLM annotation errors in detail and offered empirical evidence for why a high level of alignment between humans and LLMs on qualitative coding tasks is unrealistic [45]. However, in the majority of misclassification cases, the errors we observed are not "consequential" in the sense that they are not expected to have major impact on the downstream tasks, namely, topical summaries to be created for teachers and/or visualizations for course/study program coordinators. For example, a typical misclassification was that the LLM assigned the *Course content* category instead of *Course value and relevance*, or vice versa (Table 7). Considering how closely related these two categories are, whether a student comment is included in the summary of one or the other category, is not expected to affect much the quality of information that is eventually provided to the teacher. Similarly, such a misclassification would not significantly affect the trends in the student opinion that a visualization such as the one in Figure 4 aims to communicate.

The observed misclassifications (i.e., errors of the LLM-classifier) and misalignments with the LLM-judge, were to be expected since, in practice, the agreement among human classifiers can be imperfect too. Human classifiers and judges may encounter ambiguities and/or unclear statements in student comments (e.g., "Very responsible", "Grading if discussions are unclear and misleading"), which may occur due to students' writing styles, their command of the language if they are non-native speakers, and so on. Such ambiguities may also be a source of uncertainties for human evaluators regarding how the comments should be best classified. In the presented application case, such situations

were not uncommon (15-20%), depending on the specific human evaluator, and were eventually resolved through a discussion among the human evaluators.

Uncertain human classification of student opinions and misalignments between human and LLM classifiers and judges are tightly coupled with the coding scheme complexity [45]. On the one hand, the more complex the scheme, the greater the likelihood that student comments will convey the message precisely. On the other hand, a complex classification scheme tends to introduce uncertainties about the right category. This may create a preference for multi-label coding, that is, the use of two or more categories to label students' comments (as done, e.g., in [34]), or for more detailed instructions for the LLM classifier to avoid typical inconsistencies and ambiguities. Further research is required to systematically investigate whether multi-label classification and/or more detailed instructions would achieve better overall results than a single-label.

6. Limitations

Like any other research method and study, the current one suffers from some limitations.

First, the approach presented in the paper was evaluated with data from only one course, although the dataset of student comments was quite large and spanned multiple course editions over twelve years. For a more comprehensive evaluation, similar datasets from other courses should be used as well.

Second, the only component of the feedback analytics pipeline thoroughly evaluated in this study is the topical classification (Section 4.3), which was conducted on a relatively small subset of student comments (N=495). For a fully trustworthy feedback analytics pipeline, it is necessary to also evaluate how accurately the LLM extracts semantic units from student comments, how well it performs sentiment analysis, and how well the generated summaries align with students' comments. While prior research has evaluated LLMs on tasks of text segmentation, sentiment analysis, and summarization, these evaluations were conducted in different contexts, and it would also be important to verify that LLMs can complete these tasks well in the context of students' open-ended course feedback.

Next, all the tasks in the analytics pipeline were based on one specific LLM, namely GPT4-o-mini, whereas GPT5-mini served as the LLM judge. While these models have been representatives of the state-of-the-practice LLMs at the time the study was conducted, testing the pipeline with other LLMs, especially open-weighted ones is an important next step, as some universities may prefer to use locally hosted models for student feedback analysis.

Finally, our evaluation of the LLM-judge showed either substantial or moderate agreement with human experts. Systematic analyses of LLM-as-a-judge highlight that even strong judge models may lag inter-human agreement and can display vulnerabilities such as leniency, bias, and sensitivity to prompt complexity [43]. To address these vulnerabilities, we will explore safeguards recommended for qualitative applications, including: (i) using multiple judges or model families, (ii) further calibration on a human-labeled reference set, including the documentation of prompt versions and sampling strategies [43,48].

7. Conclusion and future research directions

Summarizing students' comments and opinions from open-ended questions in end-of-the-course surveys into meaningful and actionable insights is an extremely challenging and demanding task for humans. Using LLMs to help perform this task (semi–)automatically has started to attract interest among researchers, teachers, and course managers. Yet, until now, a generic methodology for conducting this task using LLMs was missing.

This paper has analyzed existing approaches, identified emerging common points, elaborated on them, and defined a feedback analytics pipeline for using LLMs to perform this task. The main steps in the pipeline include splitting student responses into logically and topically cohesive units (semantic units), sentiment analysis, topical classification, and summarization. The major benefit of this framework is that it enables teachers and course coordinators to start with students' qualitative feedback to open-ended questions in their course evaluations, and obtain structured and actionable insights. We have evaluated the framework in practice using a rich dataset of student comments from multiple editions of a graduate course in financial markets taught at a U.S. university and obtained encouraging results.

The presented analytics pipeline and the lessons learned from its initial application, provide grounds for some interesting further research to strengthen the validity and practical utility of the pipeline. First, the imperfect alignment between LLMs and human evaluators suggests that the current LLMs are not (yet) ready for automated qualitative coding of student open-ended feedback and that some form of human-LLM collaboration on the coding tasks might be a better solution. Accordingly, an important direction for further research is to redesign some elements of the proposed methodological pipeline, to ground them in human-in-the-loop and/or human-on-the-loop principles, similar to what was recently proposed by Li et al. [26] for coding human-AI interaction in learning settings. This would ensure the preservation of interpretive validity, as an intricate feature of traditional approaches to qualitative analysis [8, 36], while leveraging scale, speed, and consistency that automation brings. Second, the adopted coding scheme (Table 2) is not grounded in educational theories but inductively created. As such, the coding scheme does not bind the resulting annotations/classifications of student comments to any particular pedagogical approach. On the other hand, being pedagogy-agnostic, the coding scheme does not provide teachers with direct guidance on how to act upon the insights derived from the students' comments. Therefore, it would be worthwhile to explore a more pedagogy-informed version of the coding scheme, first in terms of LLM-based annotation/classification accuracy, and then in terms of the utility it can bring to teachers as a part of the overall analytics pipeline. The recently proposed coding instrument for labeling effective and malfunctional teaching practices [2] may provide a solid ground for pursuing this research direction. Last but not least, we intend to evaluate the outputs of the proposed feedback analytics pipeline with the instructors and facilitators teaching the course from our first application case. The collected educators' feedback on the relevance and usefulness of the topical summaries and visual presentation of trends in students' opinions, will serve two important functions: (i) it will serve as a key input for further refining the methodological pipeline, and (ii) it will be used for the initial design of a teacher-facing dashboard, computationally grounded in the feedback analytics pipeline, which would allow teachers to choose and shape the analysis of the course feedback.

Acknowledgment. This work was supported by the Serbian Academy of Sciences and Arts grant No. F-150.

References

1. Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., Hassan, A.: Topic modeling algorithms and applications: A survey. Information Systems 112, 102131 (2023)
2. Astudillo, G., Hilliger, I., Baier, J.: Unveiling teaching practices in higher education: Leveraging open-ended student comments in learning analytics. In: Proceedings of the 16th International Learning Analytics and Knowledge Conference. LAK '26, Association for Computing Machinery, New York, NY, USA (2026)
3. Barany, A., Nasiar, N., Porter, C., Zambrano, A.F., Andres, A.L., Bright, D., Shah, M., Liu, X., Gao, S., Zhang, J., Mehta, S., Choi, J., Giordano, C., Baker, R.S.: ChatGPT for Education Research: Exploring the Potential of Large Language Models for Qualitative Codebook Development. In: Olney, A.M., Chounta, I.A., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds.) Artificial Intelligence in Education. pp. 134–149. Springer Nature Switzerland, Cham (2024)
4. Basyal, L., Sanghvi, M.: Text Summarization Using Large Language Models: A Comparative Study of MPT-7b-instruct, Falcon-7b-instruct, and OpenAI Chat-GPT Models (Oct 2023), http://arxiv.org/abs/2310.10449, arXiv:2310.10449 [cs]
5. Berk, R.A.: Survey of 12 strategies to measure teaching effectiveness. International journal of teaching and learning in higher education 17(1), 48–62 (2005)
6. Biggs, J., Tang, C., Kennedy, G.: Teaching for quality learning at university 5e. McGraw-hill education (UK) (2022)
7. Bovill, C., Cook-Sather, A., Felten, P.: Students as co-creators of teaching approaches, course design, and curricula: Implications for academic developers. International Journal for Academic Development 16(2), 133–145 (2011)
8. Braun, V., Clarke, V.: Using thematic analysis in psychology. Qualitative Research in Psychology 3(2), 77–101 (2006)
9. Brockx, B., Van Roy, K., Mortelmans, D.: The student as a commentator: Students' comments in student evaluations of teaching. Procedia - Social and Behavioral Sciences 69, 1122–1133 (2012)
10. Brookfield, S.D.: Becoming a critically reflective teacher. John Wiley & Sons (2017)
11. Charmaz, K.: Constructing Grounded Theory: A Practical Guide through Qualitative Analysis. SAGE (2006)
12. Chiang, C.H., Lee, H.y.: A Closer Look into Using Large Language Models for Automatic Evaluation. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 8928–8942. Association for Computational Linguistics, Singapore (Dec 2023), https://aclanthology.org/2023.findings-emnlp.599/
13. Cohen, P.A.: Effectiveness of student-rating feedback for improving college instruction: A meta-analysis of findings. Research in Higher Education 13(4), 321–341 (1980)
14. Cook-Sather, A.: Sound, presence, and power: "student voice" in educational research and reform. Curriculum Inquiry 36(4), 359–390 (2006)
15. Diamond, M.R.: The usefulness of structured mid-term feedback as a catalyst for change in higher education classes. Active Learning in Higher Education 5(3), 217–231 (2004)
16. Doi, T., Isonuma, M., Yanaka, H.: Comprehensive evaluation of large language models for topic modeling. arXiv:2406.00697 (2024)
17. Fuller, K., Morbitzer, K.A., Zeeman, J.M., Persky, A., Savage, A., McLaughlin, J.E.: Exploring the use of ChatGPT to analyze student course evaluation comments. BMC Medical Education 24(1), 423 (Apr 2024), https://link.springer.com/article/10.1186/s12909-024-05316-2

18. Gul, S., Asif, M., Amin, F.E., Saleem, K., Imran, M.: Advancing Aspect-Based Sentiment Analysis in Course Evaluation: A Multi-Task Learning Framework With Selective Paraphrasing. IEEE Access 13, 7764–7779 (2025), https://ieeexplore.ieee.org/document/10833605
19. Heffernan, T.: Sexism, racism, prejudice, and bias: A literature review and synthesis of research surrounding student evaluations of courses and teaching. Assessment & Evaluation in Higher Education (2021)
20. Hou, C., Zhu, G., Zheng, J., Zhang, L., Huang, X., Zhong, T., Li, S., Du, H., Ker, C.L.: Prompt-based and Fine-tuned GPT Models for Context-Dependent and -Independent Deductive Coding in Social Annotation. In: Proceedings of the 14th Learning Analytics and Knowledge Conference. pp. 518–528. LAK '24, Association for Computing Machinery, New York, NY, USA (Mar 2024), https://doi.org/10.1145/3636555.3636910
21. Hujala, M., Knutas, A., Hynninen, T., Arminen, H.: Improving the quality of teaching by utilising written student feedback: A streamlined process. Computers & education 157, 103965 (2020)
22. Häglund, E., Björklund, J.: Opinion Units: Concise and Contextualized Representations for Aspect-Based Sentiment Analysis. In: Johansson, R., Stymne, S. (eds.) Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025). pp. 230–240. University of Tartu Library, Tallinn, Estonia (Mar 2025), https://aclanthology.org/2025.nodalida-1.24/
23. Katz, A., Gerhardt, M., Soledad, M.: Using generative text models to create qualitative codebooks for student evaluations of teaching. arXiv:2403.11984 (2024)
24. Kirk, M.: Student voice and agency for transformative change in matters that matter: Impactful inquiry in primary science. The Australian Educational Researcher 52(1), 411–430 (May 2024), https://link.springer.com/article/10.1007/s13384-024-00721-4
25. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics 33(1), 159–174 (Mar 1977)
26. Li, Z., Tang, L., Xia, M., Li, X., Chen, N., Gašević, D., Fan, Y.: When llms fall short in deductive coding: Model comparisons and human–ai collaboration workflow design. In: Proceedings of the 16th International Learning Analytics and Knowledge Conference. LAK '26, Association for Computing Machinery, New York, NY, USA (2026)
27. Liu, X., Zambrano, A.F., Baker, R.S., Barany, A., Ocumpaugh, J., Zhang, J., Pankiewicz, M., Nasiar, N., Wei, Z.: Qualitative Coding with GPT-4. Journal of Learning Analytics 12(1), 169–185 (Mar 2025), https://learning-analytics.info/index.php/JLA/article/view/8575
28. Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., Zhu, C.: G-eval: Nlg evaluation using gpt-4 with better human alignment. arXiv:2303.16634 (2023)
29. Maraj, A., Martin, M.V.: Context is the Key for LLM-Based Text Segmentation. In: Natural Language Processing and Information Systems: 30th International Conference on Applications of Natural Language to Information Systems, NLDB 2025, Kanazawa, Japan, July 4–6, 2025, Proceedings, Part I. pp. 142–155. Springer-Verlag, Berlin, Heidelberg (2025)
30. Maram, S.S., Zaman, U., El-Nasr, M.S.: Listening with language models: Using llms to collect and interpret classroom feedback (2025), https://arxiv.org/abs/2508.11707
31. Marsh, H.W.: Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In: The scholarship of teaching and learning in higher education: An evidence-based perspective, pp. 319–383. Springer (2007)
32. McNichols, H., Ikram, F., Lan, A.: The studychat dataset: Analyzing student dialogues with chatgpt in an artificial intelligence course. In: Proceedings of the 16th International Learning Analytics and Knowledge Conference. LAK '26, Association for Computing Machinery, New York, NY, USA (2026)

33. Nasser-Abu Alhija, F., Fresko, B.: Student evaluation of instruction: What can be learned from students' written comments? Studies in Educational Evaluation 35, 37–44 (2009)
34. Parker, M.J., Anderson, C., Stone, C., Oh, Y.: A large language model approach to educational survey feedback analysis. International Journal of Artificial Intelligence in Education 35(3), 444–481 (2025)
35. Peña-Torres, J.: Towards an improved of teaching practice using Sentiment Analysis in Student Evaluation. Ingeniería y Competitividad 26 (Jun 2024)
36. Saldaña, J.: The Coding Manual for Qualitative Researchers. SAGE Publications Ltd, Thousand Oaks, California (2021)
37. Satya Putra, A., Menekse, M., Butt, A.A.: WIP: Leveraging LLM for Sentiment Analysis of Student Reflection Texts from a Large Undergraduate Course. In: Proceedings of the Twelfth ACM Conference on Learning @ Scale. pp. 266–270. L@S '25, Association for Computing Machinery, New York, NY, USA (Jul 2025), https://doi.org/10.1145/3698205.3733937
38. Schreier, M.: Qualitative Content Analysis in Practice. SAGE (2012)
39. Shah, F.A., Sabir, A., Sharma, R., Pfahl, D.: How Effectively Do LLMs Extract Feature-Sentiment Pairs from App Reviews? In: Hess, A., Susi, A. (eds.) Requirements Engineering: Foundation for Software Quality. pp. 123–138. Springer Nature Switzerland, Cham (2025)
40. Spooren, P., Brockx, B., Mortelmans, D.: On the validity of student evaluation of teaching: The state of the art. Review of Educational Research 83(4), 598–642 (2013)
41. Stark, P.B., Freishtat, R.: An evaluation of course evaluations. Technical report, University of California, Berkeley (2014), https://doi.org/10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1, accessed 2026-01-10
42. Sun, J., Yan, L.: Using topic modeling to understand comments in student evaluations of teaching. Discover Education 2, 25 (2023)
43. Thakur, A.S., Choudhary, K., Ramayapally, V.S., Vaidyanathan, S., Hupkes, D.: Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. In: Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2). pp. 404–430 (2025)
44. Xiao, Z., Yuan, X., Liao, Q.V., Abdelghani, R., Oudeyer, P.Y.: Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding. In: Companion Proceedings of the 28th International Conference on Intelligent User Interfaces. pp. 75–78. IUI '23 Companion, Association for Computing Machinery, New York, NY, USA (Mar 2023), https://doi.org/10.1145/3581754.3584136
45. Xu, Z., Khatri, V., Dai, X., Liu, X., Li, S., Zhang, X., Yu, R.: Enhancing llm-based data annotation with error decomposition. In: Proceedings of the 16th International Learning Analytics and Knowledge Conference. LAK '26, Association for Computing Machinery, New York, NY, USA (2026)
46. Zapf, A., Castell, S., Morawietz, L., Karch, A.: Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate? BMC Medical Research Methodology 16(1), 93 (Aug 2016), https://doi.org/10.1186/s12874-016-0200-9
47. Zhang, M., Lindsay, E.D., Thorbensen, F.B., Poulsen, D.B., Bjerva, J.: Leveraging Large Language Models for Actionable Course Evaluation Student Feedback to Lecturers (Jul 2024), http://arxiv.org/abs/2407.01274, arXiv:2407.01274 [cs]
48. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I.: Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. pp. 46595–46623. NIPS '23, Curran Associates Inc., Red Hook, NY, USA (Dec 2023)
49. Zipser, N., Kurochkin, D., Yu, K.W., Mincieli, L.A.: Exploring relationships between qualitative student evaluation comments and quantitative instructor ratings: A structural topic modeling framework. Education Sciences 15(8), 1011 (2025)

Appendix

This section presents the final version of the system prompts used in distinct stages of the proposed feedback analytics pipeline.

```
You are an expert in analysing student responses to open-ended questions in a course
evaluation survey. Your specialty is extracting, from students' responses, opinions that
the students expressed about the course (e.g., about the course structure, organisation,
materials, the teacher, and the like).

I will give you a student response to a survey question asking them to comment on the
positive aspects of the course or the instructor.
In their response, a student may express a positive opinion about just one aspect of the
course and/or instructor or may express opinion about several aspects.
Your task is to identify all distinct course aspects the student was positive about in
their response and return them as a list, where each opinion is one list element.
Please try to avoid redundancy in this list, so if a student comments on the same or very
similar course aspect twice or more times, summarise those comments and then add them to
the list.

Structure the output as a JSON list of string values.

Here is the student response to analyze:
{response}
```

Fig. A1. The final version of the system prompt for decomposing students' responses to questions with explicitly defined polarity

```
You are an expert in analysing student responses to open-ended questions in a course
evaluation survey. Your specialty is extracting, from a student's response, positive
and negative aspects of the course (e.g., the course content, structure, materials, the
instructor's teaching style and approach, and the like).

I will give you a student response to a survey question asking them to comment on what
they liked about a course and the course instructor, and, also, what are the things
that should be improved regarding the course and the instructor.
In their response, a student may express opinion about just one aspect of the course
and/or instructor or may express opinion about several aspects.
Your task is to identify all distinct opinions about the course or instructor that the
student expressed in their response and determine the sentiment of the opinion
(positive, neutral or negative).
Your response should be a JSON list of dictionaries, where each dictionary is of the
following form:

{{
  'opinion': <an extracted student's opinion goes here>
  'sentiment': <the sentiment of the opinion goes here>
}}

Here is the student response to analyze:
{response}
```

Fig. A2. The final version of the system prompt for decomposing students' responses to general questions, where a response may contain a mixture of positive and negative impressions

```
You are an expert in analysing student responses to open-ended questions in a course
evaluation survey.
Your specialty is classifying students' comments into topical categories representing
students' opinions about distinct aspects of the course and its instructors.

I will give you a student's comment on the positive aspect(s) of the course or the
instructor. Your task is to classify the comment into one of the eleven categories given
below.
If the comment does not belong to any of the given categories, put it into the 'Other'
category.

Here is the list of categories, where each category label is followed by a description:
* Course content: this category refers to subject-matter topics taught in the course,
including domain-specific skills and competences.
* Course organization and logistics: this refers to all aspects of the course
organisation, including scheduling of lectures and labs, timeliness of course
announcements, the availability and accessibility of the learning tools and the like.
* Course value and relevance: this refers to the perceived course value and relevance,
both from the academic and practical perspective.
* Course materials and resources: this category refers to all kind of materials and
resources that were provided to the students, recommended to them, or requested from them
to use.
* Course assignments and grading: this category refers to all kinds of assignments in the
course, both formative (not graded, for practice only) and summative ones (graded), as
well as the grading system applied in the course.
* Instructor's teaching approach: this category refers to various aspects of the
instructor's teaching approach, including different aspects of the instructional design
and interaction with the students during the class.
* Instructor's knowledge and expertise: this refers to the instructor's competence, as
perceived by the student, for the topics included in the course syllabus.
* Instructor's clarity and explanations: this category refers to the instructor's clarity
of expression and quality (comprehensibility) of explanations; in other words, the
quality of communication of the subject matter topics to the students.
* Instructor's feedback and support: this is about the frequency, quality, and usefulness
of the feedback on course assignments and the overall support.
* Instructor's demeanor and communication with students: this category is about the
politeness of communication with students, friendliness, and availability for
communication.
* Overall impression: this refers to the student's overall impression about the course
and/or the instructor.

Return your response as a string that contains only the label of the category you
assigned to the student's comment. Do not include any preamble or explanation.

Student's comment:
{comment}
```

Fig. A3. The final version of the system prompt for classifying students' comments (semantic units) with positive polarity. Note: The prompt used for categorization of semantic units with negative polarity differs only in the opening sentence of the second paragraph, which reads as follows: "I will give you a student's comment expressing dissatisfaction with some aspect(s) of the course or the instructor. The dissatisfaction may be openly/directly expressed, or it may be stated in a tacit/implicit form."

```
You are an expert in analysing student responses to open-ended questions in a course
evaluation survey.
Your specialty is summarising students' comments, to make it easier for the teacher to
identify the key themes in the students' feedback.

I will give you a set of students' comments (each comment in a separate line) about a
particular aspect of the course or the course instructors and will ask you to create a
concise summary of those comments, by identifying the key themes in the comments.

Present the identified main themes as a dictionary where each dictionary key is a concise
expression of an identified main theme, while the value is a sentence-long description of
that theme.
Return the results in the JSON format, as a list of key-value pairs.

Student comments:
{comments}
```

Fig. A4. The final version of the summarization prompt

```
You are an expert evaluator for text classification tasks. Your goal is to act as a judge
and assess the accuracy of a student opinion classification.

Here is the task: A student's comment about a course was classified into one of the
following categories:
* Course content: this category refers to subject-matter topics taught in the course,
including domain-specific skills and competences.
* Course organization and logistics: this refers to all aspects of the course
organisation, including scheduling of lectures and labs, timeliness of course
announcements, the availability and accessibility of the learning tools and the like.
* Course value and relevance: this refers to the perceived course value and relevance,
both from the academic and practical perspective.
* Course materials and resources: this category refers to all kinds of materials and
resources that were provided to the students, recommended to them, or requested from them
to use.
* Course assignments and grading: this category refers to all kinds of assignments in the
course, both formative (not graded, for practice only) and summative ones (graded), as
well as the grading system applied in the course.
* Instructor's teaching approach: this category refers to various aspects of the
instructor's teaching approach, including different aspects of the instructional design
and interaction with the students during the class.
* Instructor's knowledge and expertise: this refers to the instructor's competence, as
perceived by the student, for the topics included in the course syllabus.
* Instructor's clarity and explanations: this category refers to the instructor's clarity
of expression and quality (comprehensibility) of explanations; in other words, the quality
of communication of the subject matter topics to the students.
* Instructor's feedback and support: this is about the frequency, quality, and usefulness
of the feedback on course assignments and the overall support.
* Instructor's demeanor and communication with students: this category is about the
politeness of communication with students, friendliness, and availability for
communication.
* Overall impression: this refers to the student's overall impression about the course
and/or the instructor.
Finally, if a comment could not be classified into any of the above categories, it would
be assigned to the "Other" category.

Both the student's comment and the category it was assigned are given below. Your task is
to determine if the assigned category is correct.
The assigned category is considered correct if the following two criteria are satisfied:
1) Among the possible categories (given above), it is the best fit for the student's
comment
2) The comment primarily focuses on the aspect identified by the category.

Your response should be a dictionary of the following form:

{{
  'correctness': [Yes/No]
  'reasoning': [Brief explanation of your evaluation]
}}

Return this dictionary as a JSON object. Do not include any preamble or any additional
explanation.
```

Fig. A5. The final version of the system prompt used to instruct the LLM-Judge to verify topical classification of student comments

Jelena Jovanovic a Professor at the Department of Software Engineering, University of Belgrade, Serbia. She is also an Adjunct Professor in the Centre for the Science of Learning & Technology (SLATE) at University of Bergen, Norway, and an Adjunct Professor in the Centre for Learning Analytics at Monash (CoLAM), Monash University, Australia. Her current research focus is on the use of computational approaches, including statistical and machine learning methods and techniques, network analysis, and generative AI, towards better understanding and supporting learning, primarily self-regulated learning in higher education settings. Currently she serves as Advisor to the Editor-in-Chief of IEEE Transactions on Learning Technologies and Editorial Board member of Journal of Learning Analytics, Journal of Educational Data Mining, The Internet and Higher Education, and Computers & Education: Artificial Intelligence.

Irena Vodenska is a Professor of Finance, Director of Finance programs, and chair of the Administrative Sciences Department at Metropolitan College, Boston University. Since 2004, she has been a Chartered Financial Analyst (CFA) charter holder. Her research focuses on network theory and complexity science, sustainability in finance and macroeconomics, effects of media on financial markets, technology-aided learning, and AI-powered enhancement in teaching. She is involved in Financial Literacy initiatives and democratizing finance education for the global community. As a Principal Investigator (PI) for Boston University, she has won interdisciplinary research grants from the European Commission, the US Army Research Office, and the National Science Foundation, USA.

Vladan Devedzic is a Professor of Computer Science and Software Engineering at the University of Belgrade, Faculty of Organizational Sciences, Belgrade, Serbia. Since 2021, he has been a Corresponding Member of the Serbian Academy of Sciences and Arts (SASA). His long-term professional objective is to bring close together ideas from the broad fields of Artificial Intelligence / Intelligent Systems and Software Engineering. His current professional and research interests include artificial intelligence, programming education, software engineering, intelligent software systems and technology-enhanced learning (TEL). He is the founder and the Chair of the GOOD OLD AI research network.

Received: January 20, 2026; Accepted: June 3, 2026.

Computer Science and Information Systems 23(4):1485–1518 https://doi.org/10.2298/CSIS251215035M

Paper Type: Research Paper

Federated Learning Based Analysis of Certain Complex Multivariate Functions Given by Sample Only*

Nevena Mijajlović[1,2], Ajlan Zajmović[2], Miodrag J. Mihaljević[1,3], Wei Shao[1], Lianhai Wang[1], and Shujiang Xu[1]

[1] Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center, Qilu University of Technology (Shandong Academy of Sciences), Jinan, China
nevenami@ucg.ac.me
[2] Faculty of Science and Mathematics, University of Montenegro, Podgorica, Montenegro
ajlan.z@ucg.ac.me (corresponding author)
[3] Serbian Academy of Sciences and Arts, Belgrade, Serbia

Abstract. We consider the problem of optimizing a complex function $f(\cdot)$ given a sample consisting of M input-output pairs $(argument,\ function\ value)$, $\mathcal{S} = \{(x_n, y_n)\}_{n=1}^{M}$, where $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$. We assume that the sample is available only in a distributed manner at K entities, where each entity controls a sample part $\mathcal{S}_k$ and $\mathcal{S} = \cup_{k=1}^{K} \mathcal{S}_k$. The problem is addressed employing a machine learning approach. Our goal is to generate a surrogate of the function based on the given sample, and to perform the optimization on the surrogate function. Due to the distributed nature of the sample we employ Federated Machine Learning (FML) to generate the surrogate function by joint efforts of all K entities that possess the sample parts. We call the entities involved in FML clients, and we assume that certain clients participate in FML in a malicious manner by poisoning training data or the model parameters they generate locally. Systematic experiments have been performed to study the characteristics of the surrogate function and its capabilities to provide reliable analysis of the original function depending on the sample available for training, and impacts of the training data poisoning, as well as locally generated model parameters.

Keywords: surrogate functions, neural networks, federated machine learning, algorithms, optimization models and methods, minimum evaluation, systematic numerical evaluation.

1. Introduction

We consider a surrogate $\hat{f}(x)$ of the function $f : \mathbb{R}^d \to \mathbb{R}$, where $x = [x_1, x_2, ..., x_d]$ is the input vector, d is the number of dimensions of the problem, and we have a single output y. The upper and lower bounds on the input (or design) vector are known and are denoted as $x_L \leq x \leq x_U$. After determining the input parameters (or the design space)

* This is an extended version of a conference paper, presented at the 2025 AI Conference, Serbian Academy of Sciences and Arts, Belgrade, October 8-9, 2025

of the model, we must select the data points to evaluate for building the surrogate model. The number of samples we choose to evaluate and their quality has a direct impact on the precision of the surrogate.

Machine learning models have demonstrated remarkable capabilities in building the surrogate functions by capturing complex relationships between input and output variables. The challenge, however, is that neural-network-based (NN) surrogate models must be trained on a sufficiently large dataset to enable generalization to unseen inputs, [10]. The value of these NN-based surrogate models depends on the assumption that the cost of generating the training dataset does not exceed the cost of conventional ensemble-based approaches. Evaluating the reliability of the constructed surrogate is one of the most important concerns because a flawed surrogate model can result in a misuse of computational resources, and produce negative effects on prediction, optimization, or the resulting analyses. Among the various NN architectures, multilayer feedforward neural network (MFFNNs) are among the most frequently used. In the MFFNN the neurons are organized into three layers: the input layer, the hidden layer, and the output layer. Note that the number of layers in MFFNN can be larger than just one and usually a convergence study is needed to find the right number of hidden layers. Similarly, the number of neurons in the hidden layers needs to be chosen based on a suitable convergence criterion. A number of issues regarding surrogate design and analysis have been reported, including the ones in [12], [17], [9], [1], [2], [4], [21]. Also, the surrogate approaches are of high interest in a number of engineering applications as reported, for example in [16], [3], [29], [6], [18], and may also be of interest for optimization in certain blockchain scenarios (like the ones reported in [13] and [25], for example). On the other hand, to the best of our knowledge, the approach we have employed, based on distributed model training, has not been considered.

Addressed Problem According to the above illustrative discussion of the options for the surrogate functions generation, this paper addresses the following problem: (i) We consider a variant of the approach known as "surrogate-assisted optimization" that replaces the direct optimization of an objective function with its surrogate model representation. The surrogate-assisted optimization algorithm includes an offline training phase prior to the optimization. The offline training aims to build an accurate model of the objective function and the effectiveness of the optimization depends on the initial training set size. In order to evaluate the effectiveness and correctness of the surrogate-assisted optimization, we first create a synthetic dataset by sampling points from a benchmark function within a predefined range of values. This dataset is then used to train, employing federated learning, a feed-forward fully connected neural network (multilayer perceptron) so that it can learn the relationship between the input variables and the function's output value.

Motivation for the Work Developing surrogate functions is a well-recognized problem, and a number of techniques have been proposed. Neural-network-based surrogate modeling appears to be a promising approach, but there still are numerous issues that have not been entirely addressed, including those related to federated machine learning for training of neural networks for surrogate functions generation. Federated learning (FL) enables multiple data holders to collaboratively train a global model without exchanging raw data, thus preserving data privacy and ownership. However, in practical deployments,

FL systems are often plagued by data heterogeneity such as imbalanced sample sizes, label distribution skew, and incomplete local datasets which significantly degrade model convergence, accuracy, and fairness. It is well known that the distributed nature of FL could be misused by poisoning of training data as well as the parameters of the locally trained models, and these issues need to be addressed for obtaining additional insights in employment of FL for generation of the surrogate functions and optimizations over surrogate functions.

Our Results
The main contributions of this paper are as follows.

- We propose and analyze a distributed method for generating a surrogate function employing a neural network model and training based on FL in the scenarios where the required training data could not be collected in one place because of data sensitivity and/or data ownership rights.
- Consideration of additional insights regarding the impacts of the following issues: (i) training data sample dimension; (ii) heterogeneity of training data; (iii) number of entities that perform local training; and (iv) rate of malicious entities (attackers) that degrade the training process.
- A framework for optimization over the surrogate function under adversarial corruption, including data-poisoning and model-poisoning attacks by malicious FL participants.
- Systematic experimental analysis of the surrogate function characteristics and optimization. The analysis is performed over artificially collected training data in order to provide a controlled framework for comparison of the source and surrogate functions. Accordingly, we have selected certain functions known in the analytical form and evaluate pairs $(input\ variable, function\ output)$ to obtain a sample for building the surrogate function employing FL. We have analyzed the following issues: (i) dependence of the surrogate function characteristics on the employed sample dimension and data heterogeneity; (ii) impact of training data poisoning and poisoning of locally generated model parameters; (iii) optimization capabilities of the surrogate function depending on the data employed and the training scenarios. The performed systematic experiments have provided additional quantitative insights regarding the considered training settings, the characteristics of the obtained surrogate function, and its capabilities for optimization of the source function.

Organization of the Paper The remainder of the paper is organized as follows. Section 2 provides more details on the background relevant to this paper. Section 3 presents a framework for consideration of additional insights regarding the model training process and the generated surrogate function. Section 4 summarizes the employed approach for finding a minimum over the surrogate function. Systematic experimental analysis of the proposed surrogate function generation and optimization is given in Section 5. Finally, the conclusions and future work are given in Section 6.

2. Background

2.1. Data and Models for Generating a Surrogate Function

Following the discussion in [12], we highlight several important issues related to the construction of surrogate models.

Sampling and Model validation. Since evaluating data points requires computing the true function (in our case, running an expensive simulation), sampling is the source of significant computational cost. Maintaining the quality of the surrogate model without incurring prohibitive sampling cost can be achieved by using appropriate sampling strategies [1]. It is important to select an optimal sample size, since more augmented data does not necessarily improve either local training or the aggregation of the final federated surrogate model. In particular, we point to [20], where an information-theoretic optimal sample size estimation (OSE) is proposed in order to provide reliable OSS estimation for industrial data augmentation, and where the relationship between OSS and dominant factors is theoretically considered. Sampling strategies can broadly be divided into stationary sampling and adaptive sampling. Stationary sampling methods are based on geometry or regular patterns (such as full/half factorial design or grid sampling) and approaches originating from the design of experiments (DOE) literature. In contrast, adaptive sampling starts from a relatively small initial set of samples, usually obtained by applying one of the stationary sampling strategies, and then determines new sample points sequentially. The goal of adaptive strategies is to reduce the number of samples required by evaluating only those points that are expected to most increase the precision of the resulting surrogate. Most adaptive sampling approaches rely on various criteria to balance the trade-off between exploring under-explored regions of the design space (exploration) and refining the regions close to already evaluated samples for improved performance (exploitation). A common use of such approaches is within surrogate-assisted optimization.

Validation of surrogate models is the procedure of evaluating the performance of the resulting surrogate, where, in addition to measuring predictive accuracy, validation techniques are often used to select a suitable surrogate model from a collection of candidate models and to tune its hyperparameters.

Model choice. Following [12], below we list several commonly used models for building surrogate functions.

Response surfaces and linear regression. The classical polynomial response surface model (RSM) remains one of the most frequently employed types of surrogate models in engineering design. In the RSM method, the surrogate is represented as a linear combination of polynomial functions (mostly linear) of the input variables (also called first-order RSM) and it has the form $\hat{f}(x, a) = a_0 + a_1x_1 + ... + a_mx_m$, where the vector of coefficients $a = (a_0, \ldots, a_m)^T$ is obtained by minimizing the sum of squared errors between the value from the expensive-to-compute function and the value predicted by the surrogate model.

Kriging. Kriging refers to a surrogate model based on Gaussian process modelling (and is sometimes called Gaussian process regression). A Kriging surrogate model can be formulated as $\hat{f}(x) = \sum_{j=1}^{m} a_j g_j(x) + \varepsilon(x)$, where $g_j(x)$ denotes the m independent (and known) basis functions which describe the trend of prediction of the mean at the point x, a_j denotes the unknown parameters, and $\varepsilon(x)$ denotes the random error at the point x which follows a normal distribution with a zero mean.

Radial basis functions (RBFs) Given n different sample points, RBF surrogates are written as $\hat{f}(x) = \sum_{i=1}^{n} w_i \, \phi(\|x - x^{(i)}\|_2)$, where w_i denotes the weight which is computed using the method of least-squares, and ϕ is the chosen basis function. There are several (symmetric) radial functions that can serve as a basis function.

Support vector regression (SVR). One of the main advantages of SVR methods is that they allow us to prescribe a margin δ within which errors present in the sampled data are tolerated without having a negative effect on the prediction capabilities of the resulting surrogate models. This can be particularly helpful when the sample data contain random errors, for instance due to a finite mesh size. By performing a mesh-sensitivity analysis, one can determine an appropriate value of δ.

Neural Networks (NN). The structure of a neural network is usually represented as a system of mutually interconnected "neurons". Neurons in an artificial neural network (ANN) implement primitive nonlinear functions, while the connections between neurons have numerical weights that are computed from the input data. Three main elements guide the construction of an NN: the network topology (single or multiple layers), the primitive functions associated with neurons (activation functions), and the learning algorithm used to compute the corresponding weights.

The most widely used paradigm of an NN is the multilayer perceptron (MLP), which is based on feedforward (supervised) learning and the backpropagation algorithm. The learning capability of an MLP architecture can be improved by using more hidden layers and/or more neurons in the hidden layers (noting that in certain scenarios even a single layer is enough as reported in [24]).

NNs construct differentiable mapping functions through multiple layers of interconnected nodes. In a typical feedforward architecture, each layer transforms its inputs through a composition of linear operations and nonlinear activation functions. For a layer l, the output h_l is computed as

$$h_l = \sigma(W_l h_{l-1} + b_l),$$

where W_l represents the weight matrix, b_l is the bias vector, and σ is a differentiable activation function. The complete network function is a composition of these layer transformations

$$f(x) = h_L \circ h_{L-1} \circ \cdots \circ h_1(x).$$

This architecture ensures end-to-end differentiability, enabling efficient training via backpropagation and making NNs particularly suitable for gradient-based optimization tasks.

Derivative-based optimization. Derivative-based methods iteratively update model parameters by computing gradients and taking small steps in the negative gradient direction. These methods can be broadly categorized into two main approaches: first-order methods (which use only gradient information) and second-order methods (which also exploit curvature information through, e.g., the Hessian or its approximations).

2.2. Federated Machine Learning

Framework Federated learning (FL) (see [27], for example) is a machine learning approach that enables training models across decentralized devices or servers while keeping data localized, thereby ensuring data privacy and security. In particular, FL is a machine learning setup in which multiple clients collaboratively train a model while protecting

data privacy. This privacy-preserving paradigm allows collaborative model training without sharing raw data, thus addressing privacy concerns associated with centralized data storage.

In a typical FL setting, each device or server computes model updates based on its local data and shares only these updates with a central server or aggregator. A key advantage of federated learning is its ability to leverage data from multiple sources without compromising individual privacy. By training models locally on user devices or servers, FL enables personalized model updates while preserving data confidentiality.

According to [11], the basic workflow of FL can be summarized as follows.
- *Setup*: In the setup stage, a central authority, called the orchestrator, defines the learning task and selects the parties that will participate in the training process. The orchestrator then sends the initial model parameters to the selected parties.
- *Local training*: During the local training phase, each participant individually trains their local model using their own local dataset. This local training process resembles traditional machine learning but with a crucial distinction: The data remains localized and is not exchanged with external parties.
- *Model aggregation*: During the model aggregation phase, the local models are consolidated to create a unified global model. This procedure involves the participants transmitting their individually trained models to a central server. Subsequently, the server aggregates these models by averaging the parameters.
- *Model update*: During the model update phase, the updated global model is transmitted back to the participating parties. These parties then use the new parameters to update their respective local models. This iterative process, commonly referred to as the federated learning process, is repeated until the model reaches convergence and achieves the intended level of accuracy.

It is evident that FL eliminates the need for traditional centralized data collection and offloads part of the training workload from the server to the data owners, who also act as trainers in FL. FL preserves the privacy of trainers since they only upload model updates, while sensitive and personal data remain stored locally. Furthermore, FL is often more communication-efficient, since the communication overhead of uploading raw data on both the server and trainer sides is considerably higher than the overhead of transmitting local model updates.

For some additional issues on distributed machine learning and FL approaches we refer the readers to [22] and [7], respectively.

Attacks Against Federated Machine Learning Attacks in federated learning can exploit vulnerabilities in the communication channel, model, or data to gain unauthorized access, modify or steal data, inject noise, or poison the model. These attacks can have severe consequences, including the leakage of sensitive information, the degradation of model performance, and the compromise of the entire system. Here we address data-poisoning and model-poisoning attacks.

Data poisoning attacks (see, for example, [19]) present a significant challenge to the integrity and reliability of FL systems, where model training occurs collaboratively across decentralized devices. Data-poisoning attacks could appear during the training and updating stages. These attacks involve the deliberate injection of malicious data to corrupt the model's training process, ultimately undermining its performance. Given the decentral-

ized nature of FL and the lack of direct access to local data, detecting and mitigating these attacks becomes particularly difficult, especially in unsupervised scenarios where labeled data is unavailable. Such attacks compromise the overall model, leading to unreliable predictions and erroneous outcomes. In another setting, data poisoning produces the poisoned model that works correctly as usual on normal training data, but if the input data is crafted to include the triggering patterns, the model will predict a targeted result that is pre-chosen by the attacker. For an additional illustration we point to the findings recently reported in [15] and [5].

On the other hand, in a model poisoning attack scenario (see, for example, [26]), by uploading malicious models to the central server, attackers can: (i) cause the global model to diverge from the main task; or (ii) inject trigger patterns (backdoors) into the global model to misclassify a targeted class. Attack method (i) can often be detected by observing abnormal loss of the global or local models on a validation set, since the model will not converge to the expected result. Attack method (ii) is more difficult to detect because the global model does not diverge from the main task. At the same time, the global model may learn the backdoor pattern and produce targeted misclassification in image or text tasks.

In a federated learning environment, attackers can inject trigger patterns into the global model by uploading a poisoned local model instead of poisoning all training data, i.e., by performing a model-replacement attack.

A backdoor attack can typically be decomposed into the following three steps:

1. *Trigger definition*: The attacker secretly selects triggers in advance, typically choosing those with low-frequency characteristics that align with their specific objectives.
2. *Poisoned dataset generation*: The attacker selects a subset of the dataset, injects triggers into the samples, and modifies the labels to the target labels. The final training dataset is a combination of clean and poisoned samples.
3. *Model backdoor injection*: The attacker uses the poisoned dataset together with specific attack strategies to jointly train the model on the main task and the backdoor subtask.

Overall, the attacker's objective is to modify the model parameters from θ to θ_p. The parameters θ_p can be formulated as the solution of the following optimization problem:

$$\theta_p = \arg\min_{\theta} \left[\sum_{(x_i, y_i) \in D_c} \mathcal{L}(f(x_i, \theta), y_i) + \sum_{(x_j^*, y_t) \in D_p} \mathcal{L}(f(x_j^*, \theta), y_t) \right],$$

where $\mathcal{L}$ is the loss function, and D_c and D_p represent the clean and poisoned training sets, respectively. The poisoned sample $x_j^* = x_j \oplus \tau$ is obtained by injecting a trigger τ into the clean sample x_j, and y_t is the target output.

The first term in the optimization objective enforces that the backdoored model behaves similarly to the clean model on clean samples. The second term ensures successful backdoor implantation by enforcing the desired (targeted) model behaviour on poisoned samples.

An approach for the defense that is resilient to large-scale model poisoning attacks is reported in [8] that treats the sequence of local models sent by clients in each FL round as a matrix-valued time series. Then, it identifies malicious client updates as outliers in

this time series by comparing actual observations with estimates generated by a matrix autoregressive forecasting model maintained by the server.

Finally, we point out that an in-depth discussion of attacks on FL and corresponding defense techniques is available in [23].

3. Framework for Our Analysis of Surrogate Functions Generation and Certain Corruption Issues

This section introduces a framework for surrogate-based optimization and for evaluating the reliability of learned surrogate models in federated training and adversarial settings.

3.1. Architecture for Surrogate Generation and Challenges

The overall architecture considered for surrogate function generation is shown in Fig. 1.

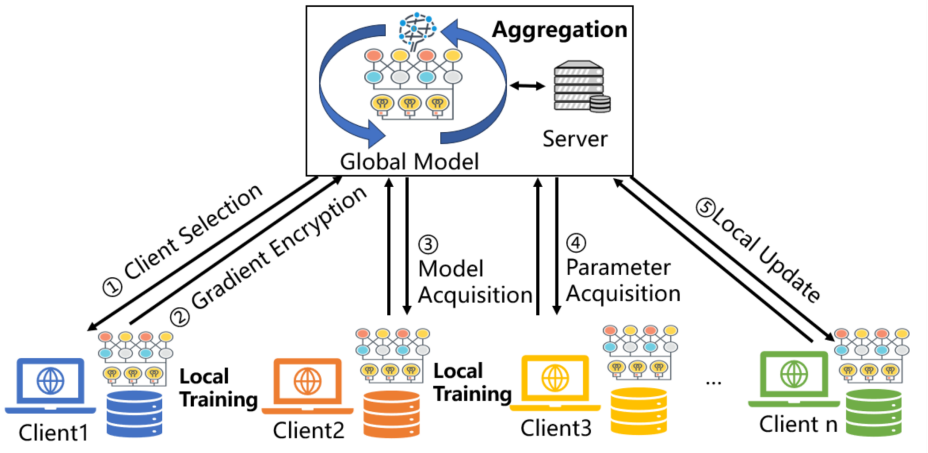


Fig. 1. Architecture for surrogate function generation.

Following Fig. 1, the FL training setup consists of the following main components:

- **Clients:** N entities that perform iterative local training. In communication round t, each client initializes from the global parameters from round $t-1$ and produces updated local parameters.
- **Communication infrastructure:** a mechanism for exchanging model parameters/updates between clients and the server.
- **Aggregation entity (server):** aggregates client updates and produces the next global model parameters.

The main challenges of federated learning are illustrated in Fig. 2. In this work, we focus on two issues: (i) data heterogeneity and (ii) limited trust, i.e., the possibility of malicious client behavior.



Fig. 2. Main challenges of federated learning.

3.2. Surrogate Generation in the Decomposed Multivariate Space

Many optimization problems involve high-dimensional input spaces where only a subset of variables is actively optimized while the remaining variables are fixed due to constraints, design choices, or partial observability. The methodology used in this work follows this paradigm by decomposing the original d-dimensional problem into two parts: a fixed subvector and a free subvector. Let

$$x = (x_1, \ldots, x_d) \in \mathcal{X} \subset \mathbb{R}^d, \qquad f : \mathcal{X} \to \mathbb{R}$$

denote the original objective. The decomposition is defined by selecting an index split $d = d_f + d_v$, where the first d_f coordinates are fixed,

$$x_{1:d_f} = \bar{x}_{1:d_f},$$

and the remaining d_v coordinates are treated as decision variables,

$$z = x_{d_f+1:d} \in \mathcal{Z} \subset \mathbb{R}^{d_v}.$$

This leads to a conditional objective

$$g_{\bar{x}}(z) = f(\bar{x}_{1:d_f}, z),$$

which is the function minimized in the evaluation stage. The surrogate model is trained in the full input space (all d variables), while the decomposition is applied only at the minimum-search stage, where the fixed coordinates are enforced and the optimizer operates in the reduced space $\mathcal{Z}$.

3.3. Surrogate Generation

Collecting Data for the Surrogate The surrogate is trained on a dataset sampled from the domain of the source function. The source function is treated as an oracle that provides exact labels $y = f(x)$ for any queried input x. Data collection uses i.i.d. sampling over the bounded hyper-rectangle $\mathcal{X} = [x_L, x_U]^d$:

$$x \sim \mathcal{U}([x_L, x_U]^d), \qquad y = f(x),$$

where $\mathcal{U}$ denotes the uniform distribution over the domain. Each federated client i receives a local dataset

$$\mathcal{S}_i = \{(x_{ij}, y_{ij})\}_{j=1}^{m},$$

where m is the number of samples per client. All clients use the same sampling rule but generate independent local datasets through different random draws. This simulates decentralized data ownership where each participant observes only a local subset of the global data distribution.

A separate validation set is created independently of all client datasets:

$$\mathcal{S}_{\text{val}} = \{(x_k^{\text{val}}, y_k^{\text{val}})\}_{k=1}^{n_{\text{val}}}, \qquad x_k^{\text{val}} \sim \mathcal{U}([x_L, x_U]^d), \quad y_k^{\text{val}} = f(x_k^{\text{val}}),$$

and is used exclusively for evaluation. This ensures that reported generalization metrics do not reuse training samples.

Global normalization. Since neural surrogates are sensitive to scaling, standardization is applied. A global feature scaler and a global target scaler are fitted using pooled client data and the transformed training quantities are

$$\tilde{x} = \frac{x - \mu_X}{\sigma_X}, \qquad \tilde{y} = \frac{y - \mu_y}{\sigma_y},$$

where μ stands for mean and σ for standard deviation.

Training is performed in the normalized space $(\tilde{x}, \tilde{y})$. Evaluation metrics are computed on the original target scale by inverse-transforming predictions. Although the use of global statistics may not be fully consistent with the strict privacy assumptions of federated learning, we assumed them to be available in this study in order to ensure stable surrogate-model training.

Model for the Surrogate Generation The surrogate model is a fully-connected neural network (MLP) that maps the input vector to a scalar function value: $\hat{f}_\theta : \mathbb{R}^d \to \mathbb{R}$. The deployed architecture follows a sequence of linear layers with a smooth nonlinearity and dropout regularization. For a generic hidden-layer configuration $(h_1, h_2, \ldots, h_L)$, the forward mapping is:

$$\begin{aligned} a_0 &= \tilde{x}, \\ a_\ell &= \text{Dropout}\left(\phi\left(W_\ell a_{\ell-1} + b_\ell\right)\right), \quad \ell = 1, \ldots, L, \\ \hat{y} &= W_{L+1} a_L + b_{L+1}, \end{aligned}$$

where $\phi(\cdot)$ denotes the activation function (GELU or SiLU depending on the experimental variant). Dropout is used to reduce overfitting and improve stability under limited

per-client sample sizes. In our implementation, we use an MLP with a small number of hidden layers, where the number of neurons per hidden layer is adapted to the input dimensionality.

Loss function and optimizer. Training minimizes a robust regression loss in normalized space, typically Smooth L1:

$$\mathcal{L}(\theta) = \frac{1}{|\mathcal{B}|} \sum_{(\tilde{x},\tilde{y})\in\mathcal{B}} \ell\left(\hat{f}_\theta(\tilde{x}), \tilde{y}\right),$$

with ℓ the Smooth L1 penalty. Optimization uses AdamW with learning rate η and weight decay λ:

$$\theta \leftarrow \theta - \eta \cdot \mathrm{AdamW}\big(\nabla_\theta \mathcal{L}(\theta), \lambda\big).$$

This setup is suitable for continuous-valued function regression where outliers can occur and where robustness is desirable.

Training of the Model for Surrogate Generation Employing Federated Learning Federated learning is simulated via a synchronous server–client protocol with N clients. Each client holds only its own dataset $\mathcal{S}_i$ and performs local updates. The server coordinates training in R rounds.

At the beginning of round t, the server broadcasts the global model parameters θ_t to all clients. Each client performs local optimization for E epochs using its local data:

$$\theta_t^{(i)} \leftarrow \mathrm{LocalTrain}(\theta_t, \mathcal{S}_i; E, \eta).$$

Clients then send the updated parameters back to the server. Aggregation constructs the next global model θ_{t+1} from the set $\{\theta_t^{(i)}\}_{i=1}^N$.

FedAvg aggregation. The standard baseline aggregation is unweighted Federated Averaging:

$$\theta_{t+1} = \frac{1}{N} \sum_{i=1}^{N} \theta_t^{(i)}.$$

This treats each client contribution equally when local sample sizes are equal.

Robust aggregation (trimmed mean). To increase robustness against malicious or extreme client updates, robust aggregation can be applied. For each scalar model parameter component $\theta[k]$, the server forms the set of client values $\{\theta^{(i)}[k]\}_{i=1}^N$, sorts them and discards the smallest and largest $\lfloor \beta N \rfloor$ values, where $\lfloor \cdot \rfloor$ denotes the floor function. The trimmed mean is then computed over the remaining indices:

$$\theta_{t+1}[k] = \frac{1}{N - 2\lfloor \beta N \rfloor} \sum_{i\in\mathcal{I}_{\text{keep}}(k)} \theta_t^{(i)}[k],$$

where $\mathcal{I}_{\text{keep}}(k)$ denotes the set of clients retained for component k. This procedure is performed independently for each parameter component.

Centralized baseline. For selected experiments, we additionally train a centralized baseline model as a non-federated reference. The centralized model uses the same MLP architecture, loss function, optimizer, global input and output scalers, and clean validation procedure as the federated model. It is trained on the pooled union of all client datasets.

To ensure a fair comparison, the centralized baseline is trained under the same nominal optimization budget as the federated model, i.e., with the same total number of training epochs corresponding to the federated communication rounds and local updates. When corruption is present, the centralized baseline is trained on the pooled dataset obtained after applying the same corruption mechanism to the corresponding client data. In this way, the comparison isolates the effect of the training regime from that of the underlying data.

Surrogate–oracle comparison. Two complementary evaluation views are used:

- **Predictive accuracy:** the surrogate is evaluated on $\mathcal{S}_{\text{val}}$ using RMSE (Root Mean Square Error) on the original output scale.
- **Optimization fidelity:** the surrogate is used to propose a minimizer under fixed-coordinate constraints, and the quality is measured by a GAP metric (defined in Section 4).

The predictive view assesses the global approximation quality of the surrogate, while the optimization view evaluates whether the surrogate remains reliable in regions relevant to minimization.

For selected experiments, both metrics are reported for the federated model and for the centralized baseline trained on the pooled data. This comparison provides an additional reference for interpreting the effect of federated training under both clean and corrupted conditions.

3.4. Possible Undetected Attacks

The attack analysis studies whether the learned surrogate model can be significantly biased while still appearing to perform reasonably under standard evaluation. The analysis is framed as an *undetected* setting: during training the server has no access to attack labels, and evaluation is always performed using the same global metrics (RMSE and GAP) computed on clean procedures and clean validation data.

Data Poisoning Attacks Data poisoning attacks are modeled as modification of the local training data at a subset of clients in order to bias the aggregated global surrogate. Our approach selects a set of K malicious clients,

$$\mathcal{M} \subset \{1, \dots, N\}, \qquad |\mathcal{M}| = K,$$

and replaces each affected local dataset $\mathcal{S}_i$ with a poisoned version $\widetilde{\mathcal{S}}_i$:

$$\mathcal{S}_i \to \widetilde{\mathcal{S}}_i, \qquad i \in \mathcal{M}.$$

For each malicious client, a fraction ρ of its samples is poisoned using a trigger-based mechanism:

– **Input manipulation:** selected inputs are perturbed by adding a fixed trigger vector Δ and clipping to the feasible domain,

$$x \leftarrow \Pi_{[x_L, x_U]^d}(x + \Delta),$$

– **Label manipulation:** targets for the same samples are overwritten to a constant attack target y_{bd}.

This produces a consistent spurious rule that the surrogate can learn, namely that inputs containing the trigger pattern should be mapped toward the attacker-chosen output. All remaining components of the training pipeline are kept unchanged so that observed degradation can be attributed to the poisoned data.

Model poisoning via additive white noise. In addition to data poisoning, we also consider a simple *model poisoning* variant in which malicious clients perturb their transmitted model parameters. Concretely, after completing local training, each malicious client adds i.i.d. Gaussian white noise to its parameter vector before uploading it to the server,

$$\theta_t^{(i)} \leftarrow \theta_t^{(i)} + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 I),$$

where the noise is applied component-wise to all trainable parameters. In our experiments we set $\sigma = 0.4$ and vary the number of malicious clients K to quantify the impact of increasingly corrupted updates on both predictive accuracy (RMSE) and optimization fidelity (GAP).

Impact of the Attacks The impact is evaluated without any explicit detection mechanism, using two complementary criteria:

– predictive degradation on clean data, measured as RMSE on a clean, uniformly sampled validation set;
– optimization degradation, measured through the GAP metric obtained by comparing the real function value at the surrogate-suggested conditional minimizer against the true conditional minimum.

To study sensitivity to attacker prevalence, the analysis performs a sweep over the number of malicious clients K while holding all other hyperparameters fixed.

Intentionally Biased Training Sampling We study the effect of non-representative client-side training data by keeping the full federated learning pipeline unchanged and replacing uniform sampling with a deliberately biased distribution. The impact is evaluated using RMSE on a clean, uniformly sampled validation set and the GAP metric for conditional minimization quality, across varying samples per client (SPC).

4. Conditional Minimum Search

We consider a surrogate-assisted optimization approach that replaces direct optimization of an objective function with optimization of its surrogate model. The procedure consists

of an offline training stage followed by a minimum-search stage. First, a synthetic dataset is generated by sampling points from a benchmark function within a predefined input range, and a feed-forward neural network (MLP) is trained to learn the mapping from the input variables to the function value. After training, the neural network acts as a surrogate model, replacing the original function during the optimization stage.

The search for the function's minimum is performed using the Differential Evolution algorithm. This choice is motivated by its derivative-free nature and its ability to search globally over bounded domains.

True conditional minimum. For a fixed prefix $\bar{x}_{1:d_f}$, the true constrained minimum is defined as:

$$z^\star = \arg\min_{z \in \mathcal{Z}} f(\bar{x}_{1:d_f}, z), \qquad f^\star = f(\bar{x}_{1:d_f}, z^\star).$$

Since no analytic minimum is assumed, a derivative-free global optimizer is used to approximate $(z^\star, f^\star)$. Differential Evolution (DE) is employed with bounded search:

$$z^\star \approx \mathrm{DE}\big(g_{\bar{x}}(z); \mathcal{Z}\big),$$

where $\mathcal{Z} = [x_L, x_U]^{d_v}$.

Surrogate-guided conditional minimum. The surrogate $\hat{f}_\theta$ defines a proxy objective

$$\hat{g}_{\bar{x}}(z) = \hat{f}_\theta(\bar{x}_{1:d_f}, z),$$

and the surrogate-minimizer is approximated similarly:

$$\hat{z} = \arg\min_{z \in \mathcal{Z}} \hat{g}_{\bar{x}}(z) \approx \mathrm{DE}\big(\hat{g}_{\bar{x}}(z); \mathcal{Z}\big).$$

To ensure consistent evaluation, DE uses fixed hyperparameters across all runs and a fixed seed. Typical settings include bounds per variable, population size controlled by `popsize`, and a maximum number of iterations `maxiter`. The result is a candidate $\hat{z}$ proposed by the surrogate.

GAP metric. The GAP measures the degradation introduced by using the surrogate for the minimizer proposal:

$$\mathrm{GAP} = f(\bar{x}_{1:d_f}, \hat{z}) - f^\star.$$

The key point is that the first term uses the *real* function evaluated at the surrogate-suggested candidate, not the surrogate prediction. This isolates the optimization guidance quality from pure prediction error.

5. Experimental Results over Artificial Data

To investigate the issues discussed in the previous sections, we conduct experiments on artificially generated data as follows: (i) We select an appropriate function known in analytic form. (ii) Using the selected function, we generate a sample for constructing its surrogate. (iii) Based on this sample, we build the surrogate function, evaluate its characteristics, and compare them with those of the original function.

In our experiments, we consider the following two benchmark functions: Ackley and Rastrigin [28].

The experiments were conducted on a 64-bit x64 system equipped with an Intel Core i5-6300U CPU (2.40–2.50 GHz), 8 GB RAM, a 238 GB NVMe SSD (WDC PC SN520), and integrated Intel HD Graphics 520.

In Table 1 we summarize abbreviations used in this section.

Table 1. Essential abbreviations used in this section.

Abbrev.	Full name	Meaning
FL	Federated Learning	Distributed training with server aggregation of client updates.
FedAvg	Federated Averaging	Server averages client parameters/updates each round.
N	Number of clients	Total clients participating (e.g., $N = 50$).
R	Communication rounds	Number of federated aggregation rounds.
E	Local epochs	Local training epochs per round at each client.
SPC	Samples per client	Number of training samples per client.
MLP	Multi-Layer Perceptron	Feed-forward neural network used as a surrogate.
GELU	Gaussian Error Linear Unit	Activation function used in the MLP.
SiLU	Sigmoid Linear Unit (Swish)	Activation function used in the MLP.
RMSE	Root Mean Squared Error	Validation error (reported on original scale).
GAP	Conditional-minimization gap	Difference between true f at surrogate-min point and true conditional minimum.
DE	Differential Evolution	Derivative-free optimizer used for conditional minimization.
K	Number of malicious clients	Number of adversarial clients in the federation.

Unless stated otherwise, Table 2 summarizes the common experimental settings used across all experiments. Function-specific choices such as dimensionality, surrogate architecture, activation, training schedule, samples per client, attack parameters, and aggregation variants are specified in the corresponding subsections. Also, AdamW and differential evolution are used with default values for all hyperparameters not explicitly specified.

5.1. Ackley 4D

We use the standard Ackley function defined for $x \in \mathbb{R}^d$ as

$$f(x) = -a \exp\left(-b\sqrt{\frac{1}{d}\sum_{i=1}^{d} x_i^2}\right) - \exp\left(\frac{1}{d}\sum_{i=1}^{d} \cos(cx_i)\right) + a + e,$$

Table 2. Common experimental settings used across all experiments. Experiment-specific settings are given in the corresponding subsections.

Parameter	**Value**
Domain of each coordinate	$[-5.12,\ 5.12]$
Activation	GELU and SiLU
Number of clients	$N = 50$
Learning rate	10^{-3}
Optimizer	AdamW
Weight decay	10^{-4}
Loss function	SmoothL1Loss (Huber)
Batch size	$B = 128$
Input normalization	global standardization (StandardScaler)
Target normalization	global standardization (StandardScaler)
Validation set size	$n_{\text{val}} = 5000$
Validation sampling	clean uniform sampling over the domain
Differential Evolution population size	`popsize` $= 30$
Maximum number of DE iterations	`maxiter` $= 300$
Random seed	42

with the standard parameters $a = 20$, $b = 0.2$, and $c = 2\pi$. In our experiments we set $d = 4$ and restrict the search space to $x \in [-5.12, 5.12]^4$.

For consistency across all Ackley 4D experiments, we keep the same conditional cross-section $(x_1, x_2) = (0.5, -1.5)$ and optimize only (x_3, x_4) within $[-5.12, 5.12]^2$ (see Section 4).

For the Ackley 4D experiments, we use an MLP surrogate with architecture $4 \to 32 \to 16 \to 1$, dropout rate $p = 0.1$, and both GELU and SiLU activations. Federated training uses $R = 12$ communication rounds and $E_{\text{local}} = 10$ local epochs per round.

Although a lower GAP is more often achieved with the SiLU activation, the overall trends are similar for GELU and SiLU. We therefore use both activations throughout this section to verify whether the choice of activation materially affects surrogate quality and the downstream conditional minimization.

We first evaluate the clean setting with $N = 50$ clients and FedAvg. For each value of SPC we train a federated MLP surrogate and measure two complementary metrics on a clean validation set: RMSE (predictive accuracy on the original scale) and GAP (quality of the surrogate when used for conditional minimization). The results for a sweep of SPC are summarized in Figures 3 and 4.

To better interpret what the surrogate learns geometrically, we additionally visualize 2D cross-sections of the 4D function. Specifically, we fix $(x_1, x_2) = (0.5, -1.5)$ and plot the surface over $(x_3, x_4) \in [-5.12, 5.12]^2$. This allows a direct comparison of the true landscape against the learned surrogate under different conditions. In the following figures we show: (i) the true cross-section, (ii) federated surrogates trained on clean data ($K = 0$), (iii) data-poisoned training with $K = 25$ malicious clients, and (iv) a model-poisoning case where malicious clients perturb parameters before aggregation. The two SPC regimes (450 and 1500) illustrate how data availability interacts with robustness.

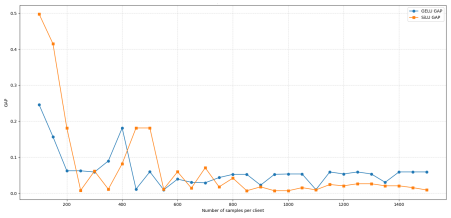


Fig. 3. Clean training: GAP vs. samples per client (GELU vs. SiLU).

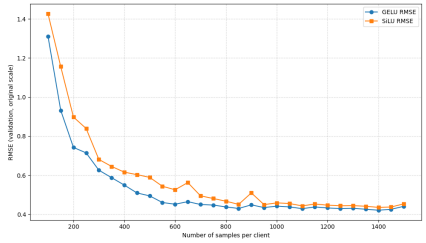


Fig. 4. Clean training: RMSE vs. samples per client (GELU vs. SiLU).

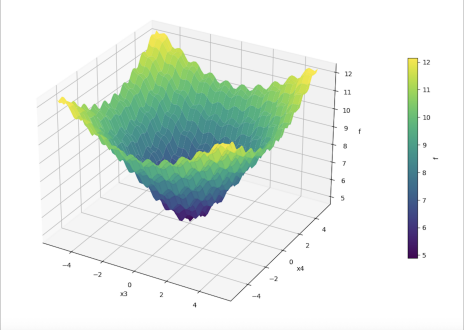


Fig. 5. True Ackley cross-section at $(x_1, x_2) = (0.5, -1.5)$.

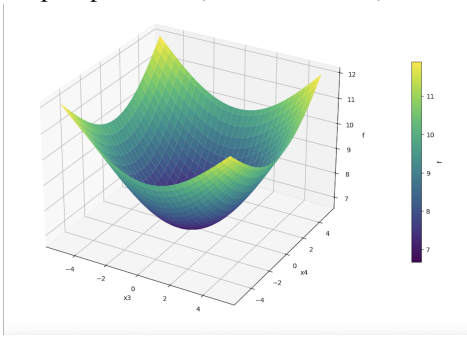


Fig. 6. Surrogate cross-section: SPC$=450$, $K = 0$.

The cross-section of the true Ackley function is shown in Fig. 5. The clean surrogates in Figures 6 and 7 indicate how increasing SPC improves the fidelity of the learned surface in this slice. Next, we visualize data-poisoned surrogates in which $K = 25$ clients manipulate a fraction of their local training data. Specifically, we show the surrogate cross-sections for $SPC = 450$ (Fig. 8) and for $SPC = 1500$ (Fig. 9).

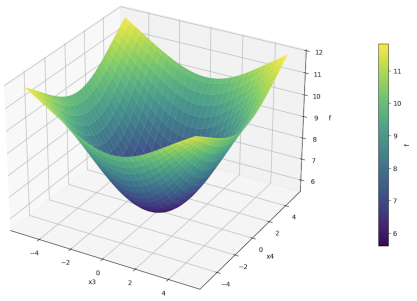


Fig. 7. Surrogate cross-section: SPC$=1500$, $K = 0$.

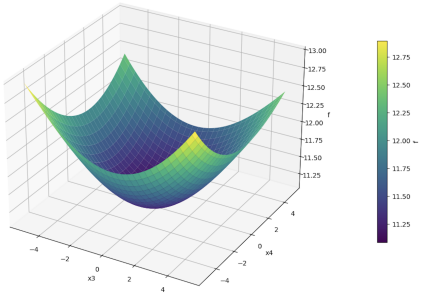


Fig. 8. Surrogate cross-section: SPC$=450$, data poisoning, $K = 25$.

Finally, in Fig. 10, we show the surrogate-function cross-section trained with $SPC = 450$ samples per client under a model-poisoning attack, where $K = 25$ clients add i.i.d.

Gaussian white noise to their model parameters before transmitting the updates to the server.

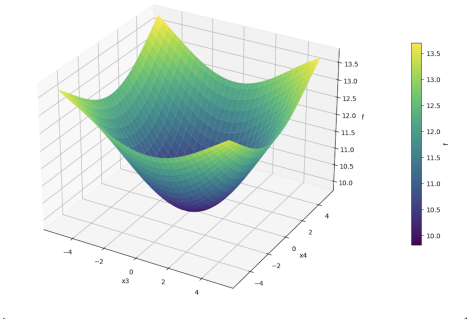


Fig. 9. Surrogate cross-section: SPC= 1500, data poisoning, $K = 25$.

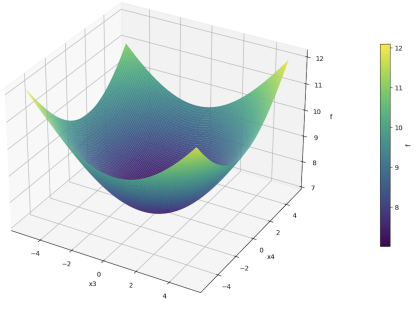


Fig. 10. Surrogate cross-section: SPC= 450, model poisoning, $K = 25$.

Data poisoning Compared to the baseline setting where all clients train on clean Ackley samples, we introduce a data-poisoning on a subset of clients and study how the number of malicious participants affects the learned surrogate.

We keep the federation size fixed at $N = 50$ and consider two fixed data regimes, $SPC \in \{200, 450\}$, while sweeping the attacker count $K \in \{0, 1, \ldots, 25\}$. We selected these regimes because they achieve the lowest GAP while using a reasonable number of samples per client. For each K, we randomly select K clients to be malicious. Each malicious client poisons a fraction $\rho = 0.6$ of its local samples by adding a fixed trigger $\Delta = [3.0, -3.0, 0.0, 0.0]$, with clipping to $[-5.12, 5.12]^4$ and overwriting the corresponding labels with $y_{\text{bd}} = 18.0$. All other training components remain identical to the baseline (same model, scaling, optimizer, rounds, and local epochs). Evaluation is performed on a clean validation set, reporting RMSE (original scale) and the GAP metric; we then plot RMSE vs. K and GAP vs. K for both SPC settings. In Figures 11 and 12, we present the obtained results for the GELU activation.

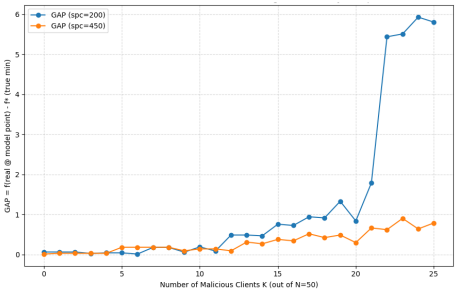


Fig. 11. Data poisoning (GELU): GAP vs. K (SPC 200, 450).

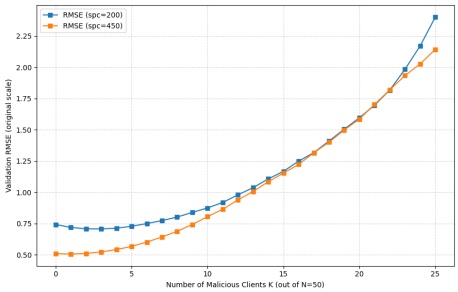


Fig. 12. Data poisoning (GELU): RMSE vs. K (SPC 200, 450).

We perform an analogous sweep using a SiLU-activated surrogate, considering $SPC \in \{200, 250, 350\}$. We keep the same architecture, optimizer, FedAvg setup, and attack definition, and report GAP/RMSE on the same clean validation protocol (Figures 13 and 14).

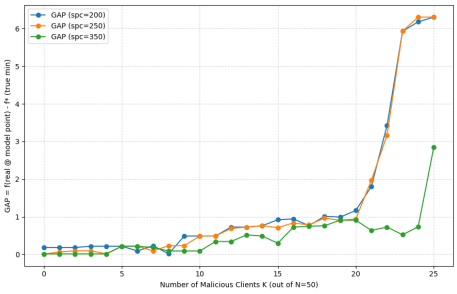


Fig. 13. Data poisoning (SiLU): GAP vs. K (SPC 200, 250, 350).

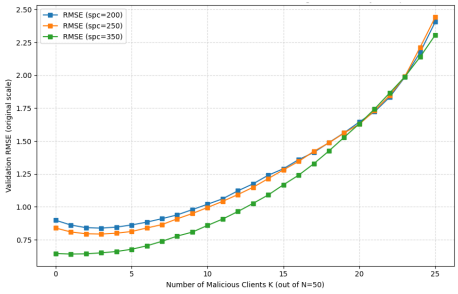


Fig. 14. Data poisoning (SiLU): RMSE vs. K (SPC 200, 250, 350).

To mitigate the impact of adversarial updates, we additionally replace FedAvg with coordinate-wise trimmed mean aggregation using $\beta = 0.3$. Concretely, for each parameter coordinate we discard the smallest and largest 30% of client values and average the remaining central values. This provides a simple robust alternative to plain averaging in the presence of outliers. Results with the GELU activation are shown in Figures 15 and 16, while the results with the SiLU activation are shown in Figures 17 and 18.

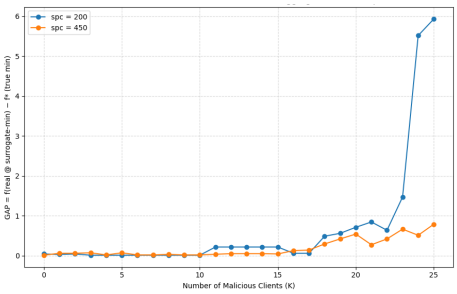


Fig. 15. Trimmed mean (GELU): GAP vs. K (SPC 200, 450).

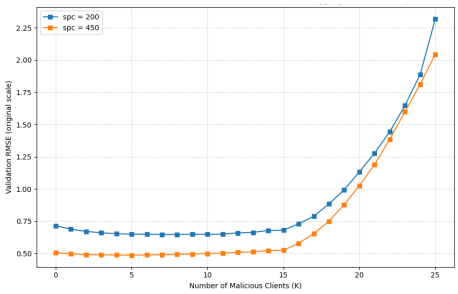


Fig. 16. Trimmed mean (GELU): RMSE vs. K (SPC 200, 450).

Comparison with a Centralized Baseline (GELU) For the Ackley 4D setting, we additionally compare the federated surrogate with a centralized MLP baseline. In this comparison, only the GELU activation is considered. The centralized model is trained on the pooled union of all client datasets, using the same network architecture, loss function, optimizer, global input and output scalers, clean validation protocol, and DE-based conditional-minimization procedure as the federated model. To ensure a fair comparison, the centralized baseline is trained under the same nominal optimization budget as the federated model. The comparison is performed under data-poisoning conditions for

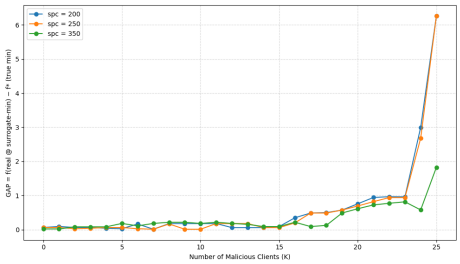


Fig. 17. Trimmed mean (SiLU): GAP vs. K (SPC 200, 250, 350).

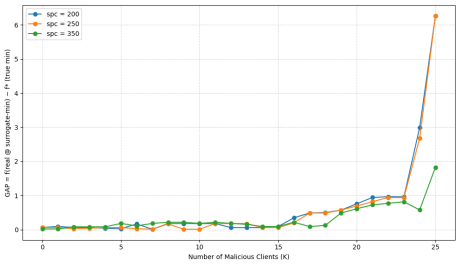


Fig. 18. Trimmed mean (SiLU): RMSE vs. K (SPC 200, 250, 350).

$SPC \in \{200, 450\}$, and the resulting RMSE and GAP values are reported as functions of the number of malicious clients K. The corresponding results are shown in Figures 19 and 20.

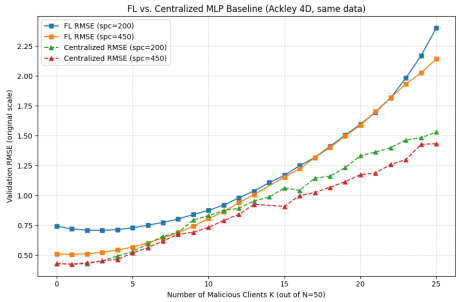


Fig. 19. Ackley 4D, GELU only: RMSE versus the number of malicious clients K for the federated model and the centralized baseline, for $SPC = 200$ and $SPC = 450$.

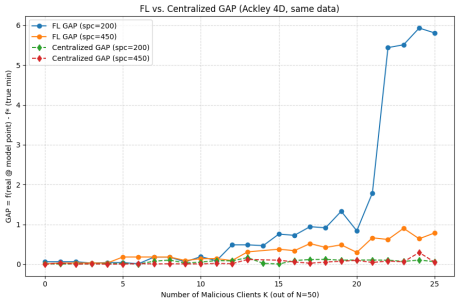


Fig. 20. Ackley 4D, GELU only: GAP versus the number of malicious clients K for the federated model and the centralized baseline, for $SPC = 200$ and $SPC = 450$.

Model poisoning We also consider a model-poisoning setting where training data remain clean and uniformly sampled, but a fixed subset of K clients corrupt their transmitted updates by adding i.i.d. Gaussian white noise to the locally trained parameters before aggregation. This attack directly perturbs the server-side averaging process without altering the local input–output samples (as explained in Section 3). We again report RMSE on a clean validation set and GAP as functions of K, for both GELU and SiLU (Figures 21 and 22 show the results for GELU activation, while Figures 23 and 24 show the results for SiLU activation.)

Intentionally Biased Training Sampling This experiment keeps the full FL pipeline unchanged, but replaces uniform client-side training sampling with a deliberately biased distribution. Each client draws 80% of samples near the domain corners (small Gaussian white noise around the 2^4 corners) and the remaining 20% uniformly while rejecting points with $\|x\|_2 < 1.0$, and we then report RMSE and GAP versus $SPC \in \{100, 150, \ldots, 1500\}$ on a clean uniform validation set. The results obtained with the GELU activation under intentionally biased training sampling are shown in Figures 25

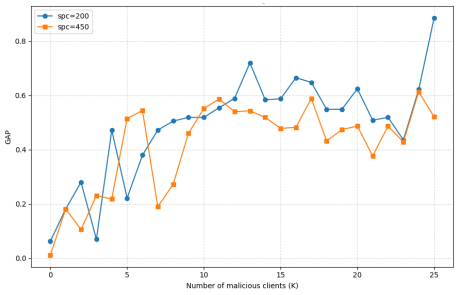


Fig. 21. Model poisoning (GELU): GAP vs. K ($\sigma = 0.4$).

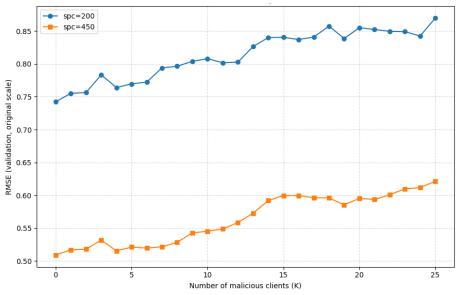


Fig. 22. Model poisoning (GELU): RMSE vs. K ($\sigma = 0.4$).

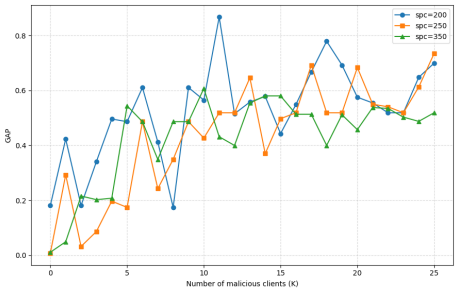


Fig. 23. Model poisoning (SiLU): GAP vs. K ($\sigma = 0.4$).

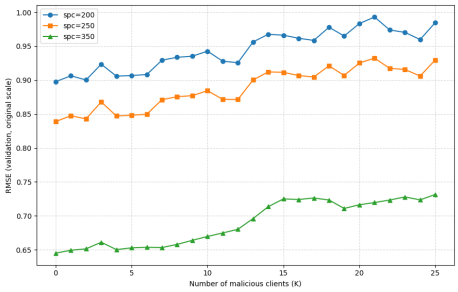


Fig. 24. Model poisoning (SiLU): RMSE vs. K ($\sigma = 0.4$).

and 26, while the corresponding results obtained with the SiLU activation are shown in Figures 27 and 28.

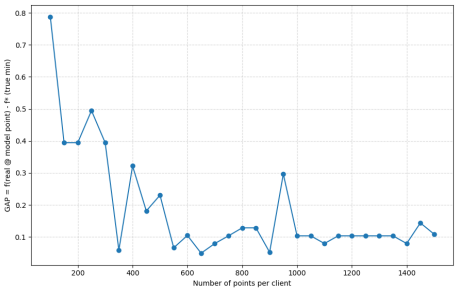


Fig. 25. Biased sampling (GELU): GAP vs. samples per client.

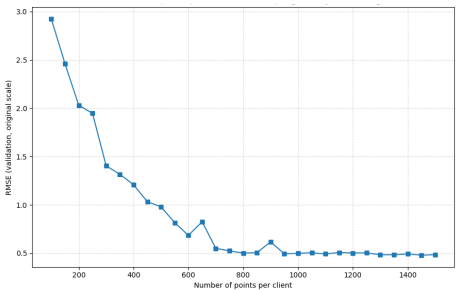


Fig. 26. Biased sampling (GELU): RMSE vs. samples per client.

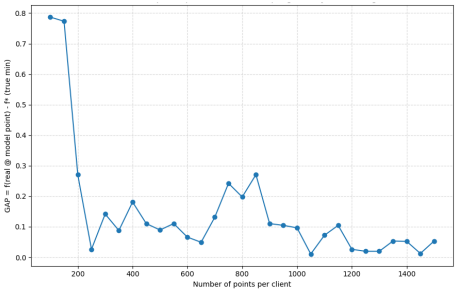


Fig. 27. Biased sampling (SiLU): GAP vs. samples per client.

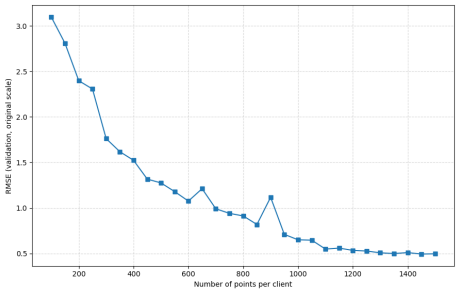


Fig. 28. Biased sampling (SiLU): RMSE vs. samples per client.

5.2. Ackley 10D

In comparison to the Ackley 4D experiment, here we increase the dimensionality to $d = 10$ and generate all training and validation samples uniformly from the hypercube $[-5.12, 5.12]^{10}$. The surrogate architecture is adapted to the higher-dimensional input by using a deeper MLP, $10 \to 32 \to 16 \to 8 \to 1$, with dropout rate $p = 0.1$ and either GELU or SiLU activation, while the number of clients remains $N = 50$. For the GAP evaluation, the conditional optimization is performed in a larger free subspace: we fix $(x_1, \ldots, x_5) = (0.5, 0.5, 1.0, 1.0, 0.5)$ and optimize $(x_6, \ldots, x_{10}) \in [-5.12, 5.12]^5$, which makes the minimum search more challenging than in the 4D case. To compensate for the increased difficulty, we use a longer FL schedule ($R = 24$ rounds and $E = 6$ local epochs) and report GAP/RMSE as functions of SPC using the same evaluation protocol and random seed. The resulting GAP and RMSE trends for this setup are summarized in

Figures 29–32, where we report the performance as a function of SPC for both GELU and SiLU activations.

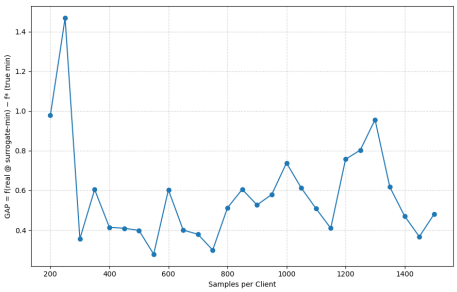


Fig. 29. Clean training (Ackley 10D, GELU): GAP vs. samples per client.

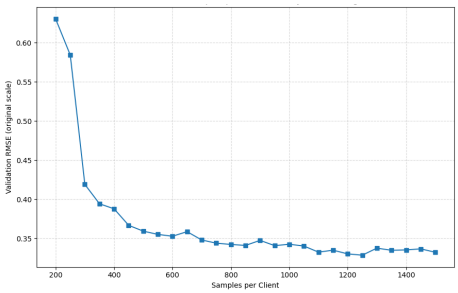


Fig. 30. Clean training (Ackley 10D, GELU): RMSE vs. samples per client.

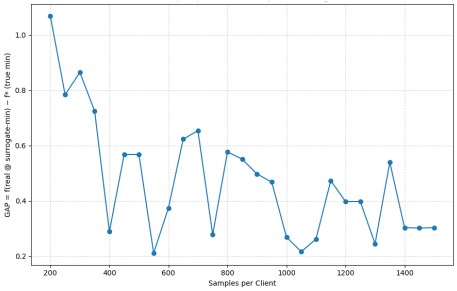


Fig. 31. Clean training (Ackley 10D, SiLU): GAP vs. samples per client.

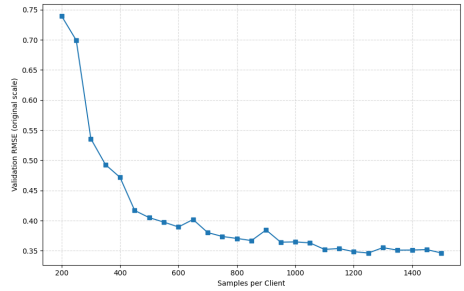


Fig. 32. Clean training (Ackley 10D, SiLU): RMSE vs. samples per client.

Data poisoning Compared to the clean baseline, we introduce K malicious clients whose local datasets are partially poisoned: a fraction $\rho = 0.40$ of their samples is modified by adding a fixed trigger vector $\Delta = [1.2, -1.2, 0.8, 0.8, -1.0, 0, 0, 0, 0, 0]$ (clipped to $[-5.12, 5.12]^{10}$) and forcing the corresponding labels to the target value $y_{\text{bd}} = 12.0$. Consequently, the global input/target standardizers are fitted on the contaminated pooled data, so the server aggregates updates learned from a mixture of honest and poisoned client distributions. The impact of the backdoor attack is shown in Figures 33–36, which plot GAP and validation RMSE versus the number of malicious clients K for several SPC regimes, with GELU evaluated for $SPC \in \{300, 500, 550\}$ and SiLU for $SPC \in \{400, 500, 550\}$.

In addition to standard FedAvg, we also evaluate a robust aggregation rule based on the coordinate-wise trimmed mean with parameter $\beta = 0.3$. For each parameter coordinate, we discard the smallest and largest 30% of client values and average the remaining central values. We report the resulting GAP and RMSE versus K for both activations and for the selected SPC regimes in Figures 37–40.

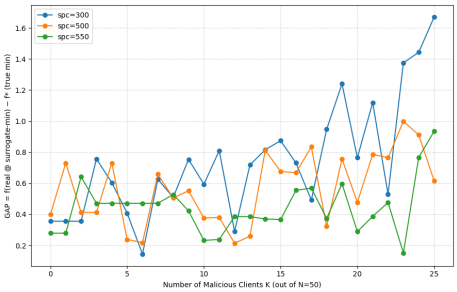


Fig. 33. Data poisoning (Ackley 10D, GELU): GAP vs. K (SPC $= 300, 500, 550$).

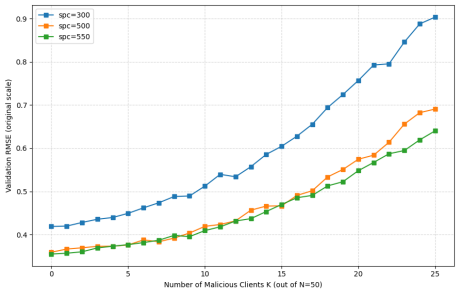


Fig. 34. Data poisoning (Ackley 10D, GELU): RMSE vs. K (SPC $= 300, 500, 550$).

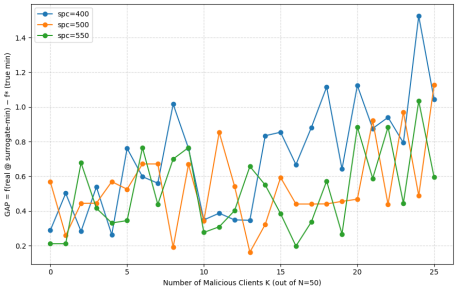


Fig. 35. Data poisoning (Ackley 10D, SiLU): GAP vs. K (SPC $= 400, 500, 550$).

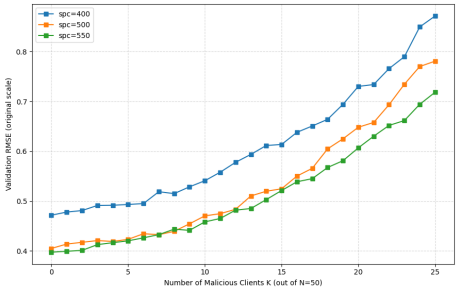


Fig. 36. Data poisoning (Ackley 10D, SiLU): RMSE vs. K (SPC $= 400, 500, 550$).

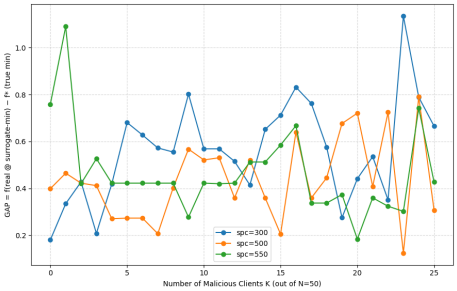


Fig. 37. Trimmed mean (Ackley 10D, GELU): GAP vs. K (SPC $= 300, 500, 550$).

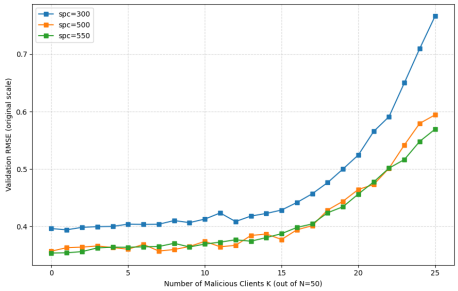


Fig. 38. Trimmed mean (Ackley 10D, GELU): RMSE vs. K (SPC $= 300, 500, 550$).

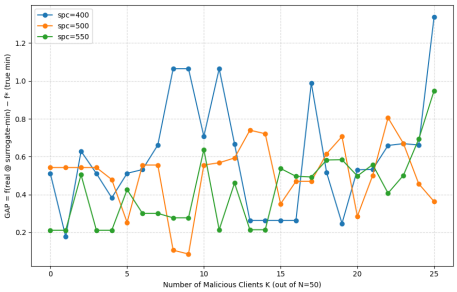


Fig. 39. Trimmed mean (Ackley 10D, SiLU): GAP vs. K (SPC $= 400, 500, 550$).

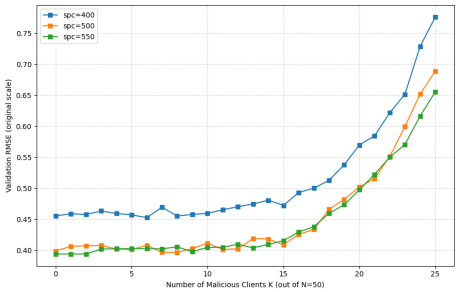


Fig. 40. Trimmed mean (Ackley 10D, SiLU): RMSE vs. K (SPC $= 400, 500, 550$).

Model poisoning (Ackley 10D). We next study a *model-poisoning* scenario in the 10D setting. The surrogate is still trained on clean, uniformly sampled client data; however, in each round we designate a fixed subset of K clients as malicious and corrupt their transmitted updates by adding i.i.d. Gaussian white noise with standard deviation $\sigma = 0.4$ to the locally trained parameters before sending them to the server. The server then aggregates the (mixed) updates using FedAvg, and we quantify the impact of this attack by reporting the validation RMSE (on a clean uniform validation set) and the GAP metric as functions of K. Results for GELU and SiLU activations are shown in Figures 41–44.

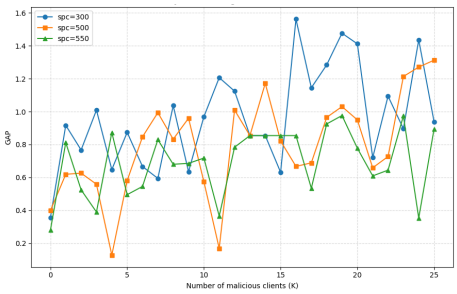


Fig. 41. Model poisoning (Ackley 10D, GELU): GAP vs. K ($\sigma = 0.4$).

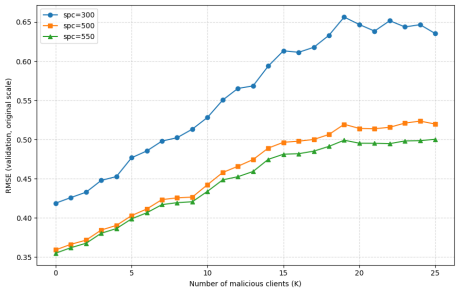


Fig. 42. Model poisoning (Ackley 10D, GELU): RMSE vs. K ($\sigma = 0.4$).

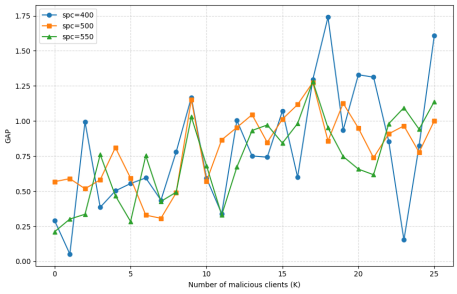


Fig. 43. Model poisoning (Ackley 10D, SiLU): GAP vs. K ($\sigma = 0.4$).

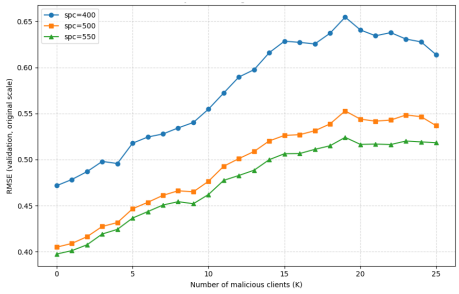


Fig. 44. Model poisoning (Ackley 10D, SiLU): RMSE vs. K ($\sigma = 0.4$).

5.3. Rastrigin 4D

We consider the Rastrigin function in $d = 4$ dimensions defined by

$$f(x) = 10d + \sum_{i=1}^{d} \left(x_i^2 - 10\cos(2\pi x_i)\right),$$

on the hypercube domain $x \in [-5.12,\, 5.12]^4$.

The surrogate model is a feed-forward MLP with architecture $4 \to 64 \to 32 \to 1$, using a dropout layer with $p = 0.05$ after each hidden layer and either GELU or SiLU activation. Federated training uses FedAvg with $N = 50$ clients, 12 communication rounds, and 10 local epochs per round.

For the GAP evaluation we fix $(x_1, x_2) = (0.5, -1.5)$ and perform conditional minimization over $(x_3, x_4) \in [-5.12, 5.12]^2$. Figures 45–48 summarize the results.

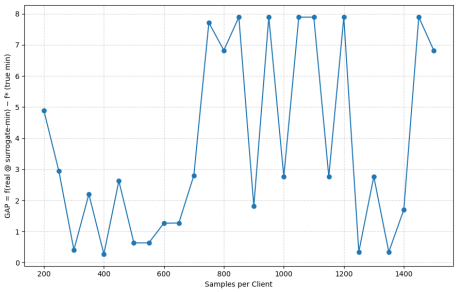


Fig. 45. Clean training (Rastrigin 4D, GELU): GAP vs. SPC.

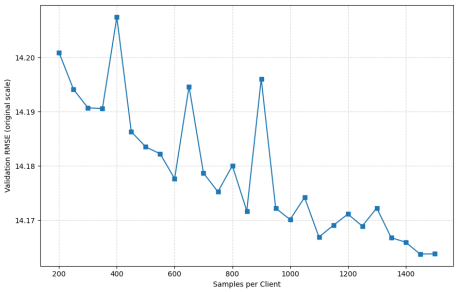


Fig. 46. Clean training (Rastrigin 4D, GELU): RMSE vs. SPC.

Data poisoning We next evaluate robustness to *data poisoning* in the Rastrigin 4D setting. The federation size remains fixed at $N = 50$, while we sweep the number of malicious clients $K \in \{0, 1, \ldots, 25\}$. For each K, we randomly select K clients to be malicious. Each malicious client poisons a fraction $\rho = 0.30$ of its local samples by injecting the

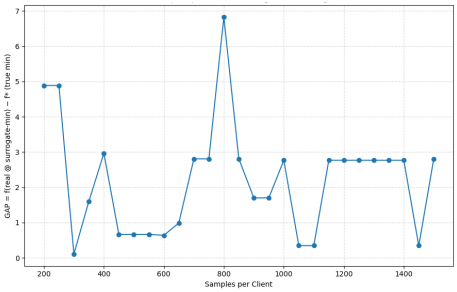


Fig. 47. Clean training (Rastrigin 4D, SiLU): GAP vs. SPC.

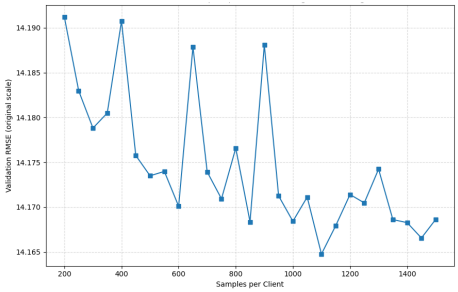


Fig. 48. Clean training (Rastrigin 4D, SiLU): RMSE vs. SPC.

trigger $\Delta = [0.8, -0.8, 0.6, -0.6]$, with clipping to $[-5.12, 5.12]^4$, and overwriting the corresponding labels with the target value $y_{\text{bd}} = 60.0$. All other training components remain unchanged (same model, scaling, optimizer, R rounds, and E local epochs). Evaluation is always performed on a clean, uniformly sampled validation set. We report RMSE and GAP as functions of K for multiple SPC regimes and for both activations; the resulting curves are shown in Figures 49–52.

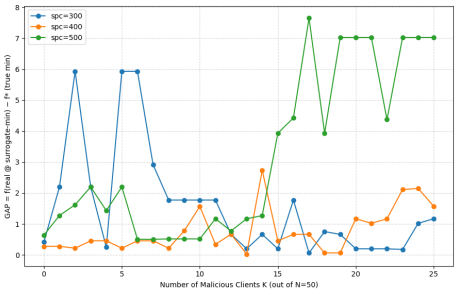


Fig. 49. Data poisoning (Rastrigin 4D, GELU, FedAvg): GAP vs. K (SPC $= 300, 400, 500$).

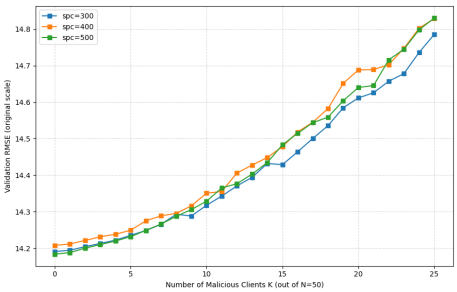


Fig. 50. Data poisoning (Rastrigin 4D, GELU, FedAvg): RMSE vs. K (SPC $= 300, 400, 500$).

To mitigate the influence of poisoned clients, we also replace FedAvg with coordinate-wise trimmed mean aggregation using $\beta = 0.3$. In each round, the server discards the lowest and highest 30% of client parameter values per coordinate and averages the remaining values. We report GAP and RMSE versus K under trimmed mean for both activations in Figures 53–56.

Model poisoning Finally, we evaluate a *model-poisoning* attack on the Rastrigin 4D benchmark. All clients train on clean, uniformly sampled data; however, in each communication round a fixed subset of K malicious clients corrupts their transmitted updates by adding i.i.d. Gaussian white noise with standard deviation $\sigma = 0.4$ to the locally

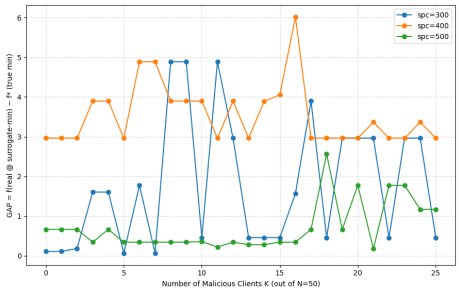


Fig. 51. Data poisoning (Rastrigin 4D, SiLU, FedAvg): GAP vs. K (SPC $= 300, 400, 500$).

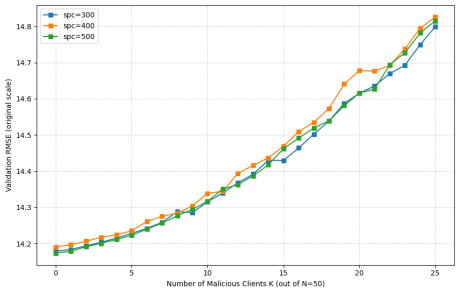


Fig. 52. Data poisoning (Rastrigin 4D, SiLU, FedAvg): RMSE vs. K (SPC $= 300, 400, 500$).

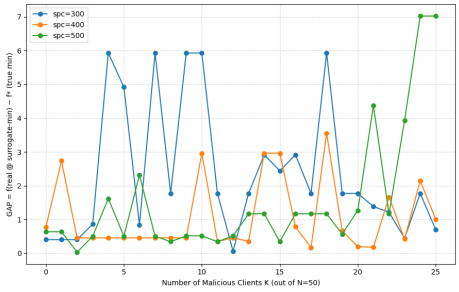


Fig. 53. Trimmed mean (Rastrigin 4D, GELU): GAP vs. K (SPC $= 300, 400, 500$; $\beta = 0.3$).

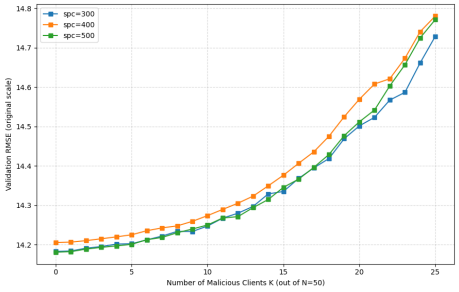


Fig. 54. Trimmed mean (Rastrigin 4D, GELU): RMSE vs. K (SPC $= 300, 400, 500$; $\beta = 0.3$).

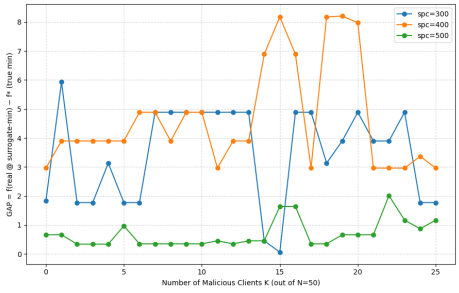


Fig. 55. Trimmed mean (Rastrigin 4D, SiLU): GAP vs. K (SPC $= 300, 400, 500$; $\beta = 0.3$).

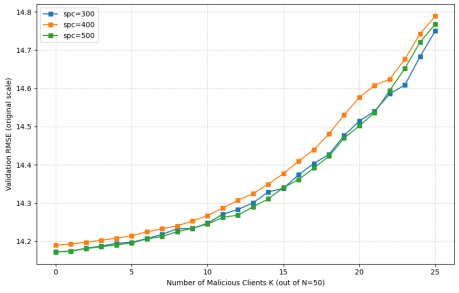


Fig. 56. Trimmed mean (Rastrigin 4D, SiLU): RMSE vs. K (SPC $= 300, 400, 500$; $\beta = 0.3$).

trained parameter vector prior to transmission. The server aggregates the resulting mixture of honest and corrupted updates using FedAvg. We report validation RMSE (clean uniform validation set, original scale) and GAP as functions of K for both GELU and SiLU activations; the corresponding curves are shown in Figures 57–60.

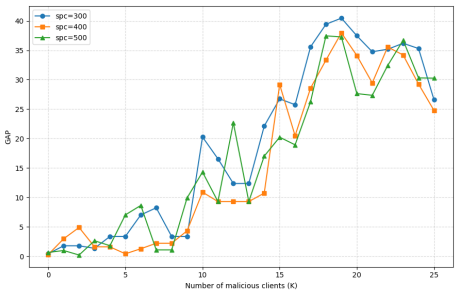


Fig. 57. Model poisoning (Rastrigin 4D, GELU, FedAvg): GAP vs. K ($\sigma = 0.4$).

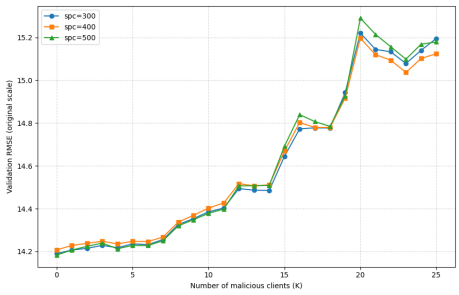


Fig. 58. Model poisoning (Rastrigin 4D, GELU, FedAvg): RMSE vs. K ($\sigma = 0.4$).

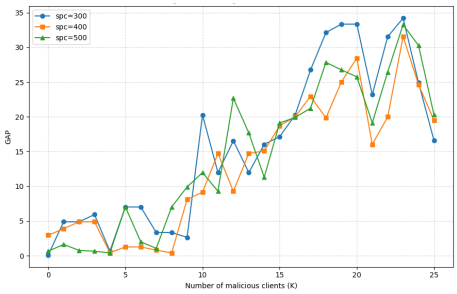


Fig. 59. Model poisoning (Rastrigin 4D, SiLU, FedAvg): GAP vs. K ($\sigma = 0.4$).

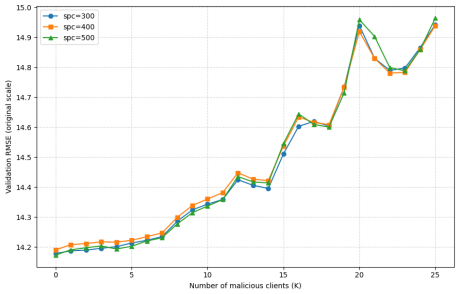


Fig. 60. Model poisoning (Rastrigin 4D, SiLU, FedAvg): RMSE vs. K ($\sigma = 0.4$).

6. Conclusions and Future Work Directions

This paper presents an analysis of Federated Machine Learning for the generation of surrogate functions and provides additional insights into the addressed issues. The analysis was performed on artificially generated training data in order to provide a suitable framework for comparing the source and surrogate functions. Accordingly, we selected benchmark functions known in analytical form and evaluated input–output pairs *(input variable, function output)* to obtain samples for building the surrogate function using machine learning. We analyzed Federated Learning with respect to the following issues: (i) the dependence of surrogate-function characteristics on the sample dimension and data heterogeneity; (ii) the impact of training-data poisoning and model poisoning on the surrogate function; and (iii) the optimization capabilities of the surrogate function depending on the employed data and the training scenario.

Across our experiments, we did not observe a meaningful difference between the SiLU and GELU activation functions in terms of either the root mean squared error or the conditional-minimization gap. In both cases, model poisoning and data poisoning generally increased the root mean squared error, indicating a degradation in predictive accuracy. For the conditional-minimization gap, however, the trends differed: increasing the number of malicious clients under model poisoning typically increased the gap, whereas under data poisoning this was not consistently the case. The similar performance of SiLU and GELU is not entirely surprising, since both are smooth, non-monotonic activation functions with comparable gradient behavior around the origin and without the sharp saturation effects typical of older activations. In several runs, a smaller gap was obtained even with a larger malicious fraction, which suggests that backdoor perturbations can sometimes distort the surrogate in a way that does not necessarily worsen the particular conditional minimization task being evaluated. These findings indicate that backdoor perturbations may degrade global predictive accuracy, as reflected by higher RMSE, while not necessarily deteriorating the specific conditional minimization task under certain settings.

Regarding the considered benchmark functions, we note that the Rastrigin function is notably more *rugged* than Ackley, yet the surrogate often still captured the global tendency toward the true global minimum. This indicates that, when the objective has a clearly dominant global basin compared to its local minima, even a relatively small neural network and a moderate amount of training data can be sufficient to localize the optimum-relevant region, despite a comparatively large root mean squared error. Replacing simple averaging with coordinate-wise trimmed mean aggregation stabilized both the root mean squared error and the conditional-minimization gap for a reasonable number of malicious clients relative to the total number of clients, which is consistent with the expected robust behavior of trimmed mean under moderate adversarial contamination.

The additional comparison with a centralized baseline provides a useful reference point for the Ackley 4D case. The centralized baseline consistently achieved lower RMSE and lower GAP than the federated model for both $SPC = 200$ and $SPC = 450$. This advantage was especially pronounced for the smaller sample regime and became even more visible in the GAP metric, where the federated model showed a substantial degradation as the number of malicious clients increased. Overall, these results suggest that, for Ackley 4D under the considered poisoning setup, centralized training was more robust than federated training both in predictive accuracy and in surrogate-based optimization quality.

Several directions for future work naturally follow from the present study. First, it would be of interest to consider stronger adversarial strategies, such as model-replacement, scaling-based, or stealthier backdoor attacks. Second, the experimental part could be extended by including additional robust aggregation rules, such as coordinate-wise median, Krum, or Bulyan. Third, since the current normalization procedure relies on global statistics for training stability, it would be useful to investigate local normalization or federated/statistics-preserving alternatives. Fourth, although benchmark functions provide a controlled and reproducible setting, it would be important to validate the proposed framework on concrete application-oriented optimization problems. Finally, while the present study adopts an MLP as a simple and well-established baseline for federated regression, more advanced surrogate architectures, such as variational autoencoders (VAEs) or diffusion models, as well as issues related to optimal sample size and data aug-

mentation, represent promising directions for further research. In particular, such models are especially promising in the context of generative surrogate modelling and search-space exploration. As an illustration, we refer to [14], which demonstrates the use of image transforms and diffusion models for the generative modelling of short and long time series, and which may serve as a guideline for leveraging diffusion models to improve search-space exploration in surrogate-assisted optimization.

Acknowledgement. This work was supported in part by the projects "Advanced Synergy of Wireless Information Security, Blockchain, Machine Learning and Optimization for 6G Networks" of the Faculty of Electrical Engineering, University of Montenegro, funded by the Ministry of Education, Science and Innovations of Montenegro, and "Mathematical analysis, optimization and machine learning" of the Faculty of Science and Mathematics, University of Montenegro, funded by the Ministry of Education, Science and Innovations of Montenegro. Also, the work was partially supported by the Shandong Provincial Key Research and Development Program (2025CXPTWZ003, WSR2024004), the Shandong Natural Science Foundation of China (ZR2024MF104, ZR2023QF129), the Taishan Scholars Program (tsqn202312231), the New 20 project of higher education of Jinan, China (202228017), the Pilot Project for Integrated Innovation of Science, Education, and Industry of Qilu University of Technology (Shandong Academy of Sciences) (2024ZDZX08).

Fundings. This work was funded by the Government of Montenegro, Ministry of Education, Science and Innovations, under grant number 07-604/25-11592/11.

References

1. Bhosekar, A., Ierapetritou, M.: Advances in surrogate-based modeling, feasibility analysis, and optimization: A review. Computers & Chemical Engineering 108, 250–267 (2018)
2. Bliek, L., Guijt, A., Karlsson, R., Verwer, S., de Weerdt, M.: Benchmarking surrogate-based optimisation algorithms on expensive black-box functions. Applied Soft Computing 134, 110744 (Nov 2023)
3. Carpio, R.R., Giordano, R.C., Secchi, A.R.: Enhanced surrogate assisted framework for constrained global optimization of expensive black-box functions. Computers & Chemical Engineering 118, 91–102 (Oct 2018)
4. Chen, S., Huang, Z., Zheng, W.: Optimization of non-smooth functions via differentiable surrogates. PLOS ONE 20(5), e0321862 (2025)
5. Chong, C., Siquan, H., Leyu, S., Yanhong, L., Ying, G.: Distributed clustering meets federated learning: A clustering-based approach to data poisoning mitigation. Cybersecurity 8, 108 (2025)
6. Danes, L.H., Avansi, G.D., Schiozer, D.J.: Using surrogates as faster objective function estimators for oil production strategy optimization in a real-case reservoir. SPE Journal 30(5), 2757–2778 (2025)
7. Formery, Y., Mendiboure, L., Villain, J., Deniau, V., Gransart, C.: A framework to design efficient blockchain-based decentralized federated learning architectures. IEEE Open Journal of the Computer Society 5, 705–723 (2024)
8. Gabrielli, E., Belli, D., Matrullo, Z., Miori, V., Tolomei, G.: Securing federated learning against extreme model poisoning attacks via multidimensional time series anomaly detection on local updates. IEEE Transactions on Information Forensics and Security 20, 9610–9624 (Sep 2025)

9. Haftka, R.T., Villanueva, D., Chaudhuri, A.: Parallel surrogate-assisted global optimization with expensive functions - a survey. Structural and Multidiciplinary Optimization 54, 3–13 (2016)
10. de Hoop, M.V., Huang, D.Z., Qian, E., Stuart, A.M.: The cost-accuracy trade-off in operator learning with neural networks. Journal of Machine Learning 1(3), 299–341 (2022)
11. Kaushal, V., Sharma, S.: Securing the collective intelligence: a comprehensive review of federated learning security attacks and defensive strategies. Knowledge and Information Systems 67, 3099–3137 (Jan 2025)
12. Kudela, J., Matousek, R.: Recent advances and applications of surrogate models for finite element method computations: A review. Soft Computing 26, 13709–13733 (2022)
13. Mihaljevic, M.J.: A blockchain consensus protocol based on dedicated time-memory-data trade-off. IEEE Access 8, 41258–141268 (2020)
14. Naiman, I., Berman, N., Pemper, I., Arbiv, I., Fadlon, G., Azencot, O.: Utilizing image transforms and diffusion models for generative modelling of short and long time series. In: Advances in Neural Information Processing Systems. vol. 37, pp. 121699–121730 (2024)
15. Nowroozi, E., Haider, I., Taheri, R., Conti, M.: Federated learning under attack: Exposing vulnerabilities through data poisoning attacks in computer networks. IEEE Transactions on Network and Service Management 22(1), 822–831 (Feb 2025)
16. Papathanasopoulos, A., Apostolopoulos, P.A., Rahmat-Samii, Y.: Optimization assisted by neural network-based machine learning in electromagnetic applications. IEEE Transactions on Antennas and Propagation 11, 160–173 (Jan 2024)
17. Propp, A.M., Tartakovsky, D.M.: Transfer learning on multi-dimensional data: A novel approach to neural network-based surrogate modeling. Journal of Machine Learning for Modeling and Computing 6(2), 13–27 (2025)
18. Queipo, N., Haftka, R., Shyy, W., Goel, T., Vaidyanathan, Tucker, P.: Surrogate-based analysis and optimization. Progress in Aerospace Sciences 41, 1–28 (Jan 2005)
19. Sun, G., Cong, Y., Dong, J., Wang, Q., Lyu, L., Liu, J.: Data poisoning attacks on federated machine learning. IEEE Internet of Things Journal 9, 11365–11374 (2022)
20. Sun, M., Zhao, R., Huang, Z., Ding, S., Liu, J.: It-ose: Exploring optimal sample size for industrial data augmentation. IEEE Transactions on Industrial Informatics 22 (2026)
21. Tong, H., Huang, C., Minku, L.L., Yao, X.: Surrogate models in evolutionary single-objective optimization: A new taxonomy and experimental study. Information Sciences 562, 414–437 (2021)
22. Verbraeken, J., Wolting, M., Katzy, J., Kloppenburg, J., Verbelen, T., Rellermeyer, J.S.: A survey on distributed machine learning. ACM Computing Surveys 53(2), 1–33 (2020)
23. Wan, Y., Qu, Y., Ni, W., Xiang, Y., Gao, L., Hossain, E.: Data and model poisoning backdoor attacks on wireless federated learning, and the defense mechanisms: A comprehensive survey. IEEE Communications Surveys & Tutorials 26(3), 1861–1897 (2024), third Quarter
24. Wang, Q., Wang, L., Xu, S., Zhang, S., Shao, W., Mihaljević, M.: Single-layer trainable neural network for secure inference. IEEE Internet of Things Journal 12, 2968–2978 (2025)
25. Xu, S., Dong, S., Wang, L., Mihaljević, M., Zhang, S., Shao, W., Wang, Q.: Blockchain-based secure data sharing with overlapping clustering and searchable encryption. Computer Standards & Interfaces 93, 103979 (2025)
26. Yang, L., Miao, Y., Liu, Z., Liu, Z., Li, X., Kuang, D., Li, H., Deng, R.H.: Enhanced model poisoning attack and multi-strategy defense in federated learning. IEEE Transactions on Information Forensics and Security 20, 3877–3892 (2025)
27. Yang, Q., Liu, Y., Chen, T., Tong, Y.: Federated machine learning: Concept and applications. ACM Trans. Intell. Syst. Technol 10, 1–19 (2019)
28. Yang, X.S.: Test Problems in Optimization. Wiley, Hoboken, NJ, USA (2010)
29. Zhang, J., Feng, F., Jin, J., Zhang, W., Zhao, Z., Zhang, Q.J.: Adaptively weighted yield-driven em optimization incorporating neurotransfer function surrogate with applications to microwave filters. IEEE Transactions on Microwave Theory and Techniques 69, 518–528 (Jan 2021)

Nevena Mijajlović was born in 1985. She received the B.Sc., M.Sc., and Ph.D. degrees in mathematics from the University of Montenegro, Podgorica, Montenegro, in 2007, 2009, and 2015, respectively. Since 2008, she has been with the University of Montenegro, initially as a Teaching Assistant, later as an Assistant Professor, and currently as an Associate Professor with the Faculty of Natural Sciences and Mathematics, Podgorica. Her research interests include optimization methods, variational inequalities, equilibrium problems, and blockchain technology. Prof. Mijajlović is a member of the Committee for Mathematics and Physics of the Montenegrin Academy of Sciences and Arts (MASA) and the Center for Young Scientists and Artists of MASA. She is also a member of the University of Montenegro's Board of Directors, representing the academic staff.

Ajlan Zajmović was born in Bijelo Polje, Montenegro, on January 2, 2001. He received the B.Sc. degree in Mathematics and Computer Science from the Faculty of Natural Sciences and Mathematics, University of Montenegro, Podgorica, in 2023, and the M.Sc. degree in Mathematics from the same faculty in December 2025. Since 2023, he has been a Teaching Assistant with the Faculty of Natural Sciences and Mathematics, University of Montenegro. He has published two scientific papers and presented his research at five international scientific conferences. He has also been involved as a junior researcher in several scientific projects. His research interests include mathematical optimization and related areas of applied mathematics.

Miodrag J. Mihaljević has been holding long-term visiting positions with various universities and research institutes in Japan, including The University of Tokyo, Sony Research Laboratories, and the National Institute AIST, Tokyo, since 1997. He is currently a Distinguished Professor with Shandong Computer Science Center, Qilu University of Technology (Shandong Academy of Sciences), Jinan, China, as well as a member of the Serbian Academy of Sciences and Arts, Belgrade. His main research interests include cryptology, information security, blockchain technology, and machine learning. Since 2014, he has been an Elected Member of Academia Europaea. Since 2021, he has been an Elected Member of the Serbian Academy of Sciences and Arts. In 2013, he received the National Award of the Serbian Academy of Sciences and Arts for ten-year achievements. In the six consecutive years 2020–2025, he was included in the ranked list colloquially known as World's Top 2% Scientists by Elsevier and Stanford University, regarding his career achievements. For more information, visit http://www.mi.sanu.ac.rs/cv/cvmihaljevic.htm.

Lianhai Wang received the B.S. degree in computer software and theory, and the Ph.D. degree in computer science and technology from Shandong University, Jinan, China, in 2003 and 2014, respectively. He is currently a Research Professor with the Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center, Qilu University of Technology (Shandong Academy of Sciences), China. His current research interests include artificial intelligence security and privacy protection, blockchain, and digital forensics.

Shujiang Xu received the M.S. degree in applied mathematics and the Ph.D. degree in cryptography from Beijing University of Posts and Telecommunications, China, in 2006 and 2014, respectively. He is currently a Research Professor with the Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong

Computer Science Center, Qilu University of Technology (Shandong Academy of Sciences), China. His main research interests include data security and privacy protection, blockchain, and digital forensics.

Wei Shao received the M.S. degree in information security from University College London, U.K., and the Ph.D. degree from the School of Cyber Science, Nankai University, Tianjin, China, in 2021. She is currently with the Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center, Qilu University of Technology (Shandong Academy of Sciences), China. Her research interests include information security, cryptography, and blockchain security.

Received: December 15, 2025; Accepted: June 7, 2026.

Computer Science and Information Systems 23(4):1519–1544 https://doi.org/10.2298/CSIS251215026I

Paper Type: Research Paper

EduGrader: The AI-powered grading assistant ⋆

Svetozar Iković[1] and Jelena Graovac[1]

University of Belgrade, Faculty of Mathematics
mi241072@alas.matf.bg.ac.rs
jelena.graovac@matf.bg.ac.rs (corresponding author)

Abstract. The rapid development of artificial intelligence (AI) and large language models (LLMs) is reshaping assessment practices in higher education, particularly in grading open-ended questions. Traditional grading is time-consuming, prone to inconsistency, and difficult to scale for large classes. LLMs offer an opportunity to improve efficiency, consistency, and feedback frequency by leveraging their ability to understand and reason about natural language. This paper introduces EduGrader, an open-source, multi-model grading platform integrating several LLMs, including GPT-4o, GPT-4o-mini, DeepSeek-Chat, GPT-5.1, and DeepSeek-Reasoner. EduGrader supports three grading strictness levels (lenient, neutral, strict) and generates numerical scores with concise explanations highlighting correct reasoning, missing concepts, and errors. The system operates in two modes: a provided-reference-answer grading mode guided by instructor solutions and a generated-reference-answer grading mode that automatically creates reference answers from course materials. All experiments are conducted in Serbian, a comparatively low-resource language underrepresented in mainstream AI training data, extending AI-assisted grading research beyond the dominant English-language context. EduGrader is evaluated on 686 real student responses from six university courses across two institutions, including free-text explanations and code-based answers. In its best-performing configuration, the system achieves a Pearson correlation coefficient of 0.90 with human graders while reducing grading workload and improving reliability, though it is intended to assist rather than replace instructors.

Keywords: Automated Grading, Education, Large Language Models, Artificial Intelligence.

1. Introduction

Grading is rarely the most exciting part of teaching, yet it often quietly shapes how students learn. When feedback is timely and specific, students can recognize and correct misunderstandings before they take root. When it is delayed, vague, or inconsistent, assessment becomes a formality, and learning slows down. This challenge is particularly visible in large university courses, where instructors and teaching assistants must evaluate hundreds of open-ended responses within tight deadlines. Unlike multiple-choice questions,

⋆ This is an extended version of a conference paper, presented at the 2025 AI Conference, Serbian Academy of Sciences and Arts, Belgrade, October 8-9, 2025

short open-ended explanations and mini-essays cannot be checked through pattern matching alone. They demand interpretation: identifying partial reasoning, distinguishing minor wording issues from conceptual gaps, and deciding how much an omission should affect the score. As a result, open-ended grading remains both expensive and difficult to scale. It is also prone to inconsistency. Even with detailed grading guidelines, two graders may interpret the same response differently, especially when answers are ambiguous, hastily written, or structurally unclear [11, 26, 4]. For instructors, this workload restricts how often richer, explanation-based questions can be used. For students, it often means feedback arrives too late to be useful, or never arrives at all. Earlier generations of automated assessment systems worked well only when responses were tightly constrained, such as multiple-choice tests, fill-in-the-blank questions, or short answers with predictable phrasing [25, 3, 22]. Advances in natural language processing (NLP) expanded these capabilities, but many approaches still relied on heavy feature engineering, limited domains, or large labeled datasets [23, 6, 2]. Modern large language models (LLMs), however, mark a turning point: they can interpret free-form text and reason about meaning directly. This development has made it feasible to design tools that grade open-ended responses, justify their scores in natural language, and highlight what is missing, not just what is wrong [13, 33].

Still, the use of LLMs in assessment raises important practical and pedagogical questions. How consistent are their grades across subjects and question types? Can instructors calibrate strictness to match specific course expectations? What happens when no complete answer key is available? And, crucially, how can we preserve human accountability for final grades while taking advantage of automation's efficiency?

This paper introduces EduGrader, an AI-assisted grading platform built to support open-ended assessment in higher education under instructor supervision. EduGrader integrates multiple large language models, GPT-4o[1], GPT-4o-mini[2], DeepSeek-Chat[3], and two reasoning-oriented models, GPT-5.1[4] and DeepSeek-Reasoner[5], within a unified, transparent workflow that prioritizes instructor control. The system includes three grading strictness profiles, lenient, neutral, and strict, reflecting real-world grading practices, where some courses reward partial reasoning while others demand precise terminology and complete solutions.

EduGrader offers two complementary grading modes:

- **Provided-reference-answer grading (PRA)**, in which student responses are evaluated against instructor-provided reference answers (optionally supplemented with grading notes).
- **Generated-reference-answer grading (GRA)**, where the system first constructs a reference answer from course materials, such as lecture notes, course handouts, and textbooks, and then evaluates student responses against that generated standard.

In both modes, the output includes not only a numeric score (0–10) but also a clear explanation that identifies strengths, missing ideas, and reasoning gaps. The aim is to

[1] https://platform.openai.com/docs/models/gpt-4o
[2] https://platform.openai.com/docs/models/gpt-4o-mini
[3] https://api-docs.deepseek.com/
[4] https://platform.openai.com/docs/models/gpt-5
[5] https://api-docs.deepseek.com/guides/reasoning_model

provide feedback that is actionable for students and auditable for instructors, rather than a "black box" grade.

We evaluate EduGrader using 686 authentic student responses collected across two universities and six courses, spanning different response types: short factual answers, open-ended explanatory text, and code-like expressions. The results show that PRA yields more stable outcomes than GRA, and that reasoning-oriented models align most closely with human graders. In its best configuration (GPT-5.1, neutral mode), EduGrader achieves a Pearson correlation coefficient (PCC) of 0.90 with human grading.

Finally, EduGrader is designed as a grading assistant, not a replacement for instructor judgment. It reduces repetitive workload and enhances consistency, but human oversight remains essential for ambiguous questions, discipline-specific expectations, and borderline cases. For this reason, EduGrader follows a human-in-the-loop workflow and is released as open source, promoting transparency, reproducibility, and adaptability across institutions.

2. Related Work

Research on the use of artificial intelligence (AI) in education reveals a clear evolution from early rule-based automation toward adaptive systems capable of interpreting and generating complex, human-like responses. Foundational definitions describe AI as computational systems designed to perform cognitive functions such as learning, reasoning, and problem solving [1]. Within educational settings, AI encompasses a broad set of technologies that together enable more personalized and adaptive learning environments [32]. Numerous studies and systematic reviews emphasize AI's potential to improve learning outcomes, broaden access to education, and support more efficient use of instructional resources [31, 34]. Recent advances in large LLMs have further expanded these possibilities. Unlike earlier NLP-based approaches that relied on handcrafted features or domain-specific training, LLMs can directly process free-form text and reason about meaning at a higher level of abstraction. As a result, they have been increasingly explored for learner-centered instruction, automated content generation, and novel forms of assessment across a wide range of educational domains [34].

Empirical studies examining LLM-assisted grading report mixed but informative results. Kooli and Yusuf [14] conducted an empirical investigation of ChatGPT-assisted grading using a dataset of 25 student papers from a cohort of approximately 100 second-year university students in a social science course. In their experimental setup, both human graders and the ChatGPT system independently evaluated the assignments, enabling a direct comparison between automated and human grading outcomes. The study reported a moderate correlation between AI-generated and human grades (correlation coefficient of approximately 0.45), suggesting that while the model can approximate overall grading patterns, substantial discrepancies remain in individual evaluations. The authors further highlight that ChatGPT can provide useful feedback and accelerate the grading process, but that its automated nature may limit contextual understanding of nuanced student arguments. They also emphasize potential risks such as the amplification of educational inequalities and the need for institutional frameworks ensuring transparency, monitoring, and bias mitigation when deploying AI-based assessment tools.

Similarly, Lundgren [18] examined the performance of GPT-4 in grading higher-education essays, using a dataset of 60 master's-level political science essays previously graded by university instructors. The study compared the grades suggested by GPT-4 with those assigned by human graders and found that the average scores produced by the model were broadly aligned with instructor evaluations. However, the analysis revealed lower interrater reliability between AI and human graders, indicating that while the model may approximate aggregate grading trends, it struggles to consistently replicate fine-grained human judgment. Additionally, the model demonstrated a risk-averse grading pattern and tended to rely on general essay characteristics such as language quality rather than adapting fully to detailed grading rubrics. Experiments with modified prompts and grading instructions showed only limited improvements, suggesting that current LLM-based grading systems remain relatively insensitive to subtle variations in evaluation criteria.

In contrast to essay-based assessment, Jukiewicz [12] investigated the application of ChatGPT for grading programming assignments in a 15-week Python programming course involving 67 undergraduate cognitive science students. Across nine programming assignments, both the instructor and the ChatGPT system independently evaluated student submissions, enabling a comparison of grading consistency and feedback quality. The results showed a strong positive correlation between ChatGPT-generated grades and instructor evaluations, although the AI system tended to assign slightly stricter grades. The study also reported excellent repeatability of the model's grading results across multiple evaluation runs, with only negligible variations between iterations. Beyond numerical scores, ChatGPT was able to assess code correctness, enforce coding standards such as adherence to PEP8 guidelines, and generate detailed feedback within seconds, significantly reducing the time required for manual grading. Nevertheless, the author also identifies several limitations, including the potential for hallucinated errors, API usage costs, and occasional grading inconsistencies that require teacher verification. As a result, the study concludes that ChatGPT can serve as a valuable complementary tool for automated code assessment and feedback generation, but that human oversight remains necessary to ensure reliability and fairness.

These findings support broader evidence indicating that LLM grading reliability is task-dependent: agreement with human graders tends to be higher in structured-response domains, such as definitions, formulas, or code fragments, while performance decreases in subjective or open-ended writing tasks [28, 29]. For instance, [21] evaluated GPT-3 on open-ended responses from industry professionals and observed strong correlations with expert ratings, contingent upon clearly defined criteria and contextual grounding. Beyond numerical grading, several studies highlight AI's broader role in supporting assessment processes. AI-based systems have been shown to generate adaptive learning activities, deliver personalized feedback, and support both formative and summative assessment based on individual learner characteristics [8, 17]. Intelligent Tutoring Systems and related approaches aim to approximate one-to-one instructional support while increasing consistency and scalability in evaluation practices [7, 9]. At the same time, critical perspectives caution against the "black-boxing" of accountability in automated assessment, where decision-making processes become opaque and responsibility is blurred [27]. As a result, transparency, explainability, data privacy, and feedback quality remain central concerns in conceptual analyses of AI-supported grading [15].

Across this literature, a consistent conclusion emerges: LLMs are best understood as assistive tools that automate routine aspects of grading and feedback, while human educators remain essential for pedagogical judgment, contextual interpretation, and handling ambiguous or borderline cases [16, 32]. Consequently, approaches that combine automation with Human-in-the-Loop supervision, configurable grading behavior, and interpretable feedback align most closely with current recommendations in the literature. These considerations motivate the design of systems that prioritize instructor control, transparency, and adaptability, principles that directly inform the design of the EduGrader platform introduced in this paper.

3. System evaluation and methodology

EduGrader is designed as an AI-assisted grading platform that integrates LLMs into an instructor-supervised assessment workflow. The primary goal of the system is to improve grading efficiency and consistency while preserving pedagogical transparency and human accountability. This section describes the evaluation methodology used to assess the reliability of EduGrader's grading outputs, the datasets on which it was tested, and the model configurations examined in the experiments.

3.1. Metrics and evaluation

To evaluate the alignment between AI-generated grades and human judgment, EduGrader's outputs were compared with grades assigned by two university instructors. Each instructor graded a different subset of the 686 responses (192 and 485 responses, respectively). This setup enabled us to compare AI performance against multiple human evaluators rather than a single grading standard. Each response was graded according to the standard grading criteria used in the course. Grades were assigned on a 0–10 scale based on the degree of conceptual correctness and completeness of the answer. Instructors considered whether the student correctly identified the key concepts, whether important elements were missing, and whether the reasoning was partially correct but incomplete. Answers demonstrating correct core ideas received partial credit even when terminology, structure, or some details were missing, while fundamentally incorrect answers received low or zero scores. For each response, the score assigned by the responsible course instructor was treated as the reference grade and compared with the AI-generated score using the evaluation metrics described below.

We used four statistical indicators that capture complementary aspects of grading reliability: the Pearson Correlation Coefficient (PCC), quadratic weighted Cohen's kappa, the Mean Ratio, and the RMSE-to-standard-deviation (RMSE/STDDEV) ratio.

The PCC [24] measures the strength of the linear association between human- and AI-assigned grades. The coefficient ranges from -1 to +1, where +1 indicates a perfect positive linear relationship, 0 indicates no linear relationship, and -1 indicates a perfect negative linear relationship. Before conducting the analysis, we examined the distribution of grading differences using the Shapiro–Wilk test and visual inspection of Q–Q plots. The results did not indicate substantial departures from normality (p ¿ .05), supporting the use of a parametric correlation measure. Importantly, PCC evaluates linear association rather than absolute agreement. For this reason, PCC is complemented by weighted

kappa as an agreement-oriented measure and by error-based metrics that capture absolute deviations between AI and human scores.

To complement PCC, we also report quadratic weighted Cohen's kappa, which is suitable for ordinal grades. Unlike PCC, which measures linear association, weighted kappa measures agreement while accounting for the ordered nature of the 0–10 grading scale and penalizing larger disagreements more strongly than smaller ones.

The Mean Ratio summarizes overall grading tendencies. Differences between AI and human means highlight systematic grading leniency or strictness but do not reflect alignment in score variation.

The Root Mean Square Error (RMSE) [10] captures the average size of the differences between human and AI grades. A smaller RMSE indicates stronger agreement between human and AI grades. The RMSE/STDDEV [5] ratio measures how well AI-assigned grades agree with human-assigned grades, relative to the natural variability in the human grades. It shows whether the AI's deviations from human judgments are small compared to how much human grades vary overall. An RMSE/STDDEV ratio close to 0 indicates very high agreement between AI and human grades, around 1 means AI deviations are comparable to typical human grading variability, and greater than 1 suggests AI grades differ from human grades more than humans differ among themselves.

3.2. Data collection

The empirical evaluation of EduGrader was conducted on authentic theoretical exam responses collected across several examination periods at two universities in Serbia.

The resulting dataset spans **six university courses** and covers multiple academic levels, including undergraduate and master studies. It includes a mixture of conceptual questions and programming-related tasks, providing a diverse and heterogeneous benchmark for evaluating automated grading systems. The courses included in the dataset are:

- Fundamentals of programming in C - part 2 (Prog 2): First-year undergraduate course, Faculty of Mathematics, University of Belgrade
- Technical and scientific writings (TSW): First-year undergraduate course, Faculty of Mathematics, University of Belgrade
- Web programming (WP): Fourth-year undergraduate course, Faculty of Mathematics, University of Belgrade.
- Digital data storage (DDS): Master-level course, University of Arts
- HTML5 and JavaScript for video games (HTML5): Master-level course, University of Arts
- Data Mining (DM): Third-year undergraduate course, Faculty of Mathematics, University of Belgrade.

Each dataset includes student responses, instructor-assigned grades, reference answers (when available), instructor notes, and relevant course materials used during exam preparation. In total, the dataset contains **686 student responses**, providing a realistic benchmark for evaluating grading consistency across disciplines, academic levels, and question formats. Depending on the course, responses include **short free-text explanations**, **long free-text explanations**, and **code-based answers**. Table 1 summarizes the dataset structure, including the number of responses and response types for each course.

To better characterize the evaluation data, we additionally report the distribution of response types and the average length of answers. These statistics provide insight into the variability of response complexity and help contextualize the difficulty of automated grading across different question formats. In addition to the distribution of response types, we analyze the typical length of answers for both instructor-provided reference solutions, automatically generated reference answers, and student responses. This information is relevant because answer length can influence the difficulty of automated grading, particularly for long explanatory responses.

Table 2 also shows that generated reference answers are often much longer than instructor-provided answers, especially for questions where short answers are expected. In several courses, instructors expect concise responses (e.g., a single expression or brief statement), while LLM-generated answers typically include additional explanations describing why the answer is correct. As a result, generated answers contain more contextual information than is strictly required by the original exam question.

Table 1. Dataset structure across courses (number of student submissions and responses by question type)

Course	Responses	Short free-text	Long free-text	Code-based
Fundamentals of Programming in C – Part 2 (Prog 2)	480	380	0	100
Technical and Scientific Writing (TSW)	40	8	0	32
Web Programming (WP)	50	26	0	24
Digital Data Storage (DDS)	10	10	0	0
HTML5 and JavaScript for Video Games (HTML5)	10	4	0	6
Data Mining (DM)	96	0	96	0
Total	**686**	**428 (62.4%)**	**96 (14%)**	**162 (23.6%)**

Table 2. Average response length by question type and answer source

Question type	Instructor answer Avg words	Instructor answer Avg chars	Generated answer Avg words	Generated answer Avg chars	Student answer Avg words	Student answer Avg chars
Short free-text	20.9	130.5	201.7	1341.9	24.7	143.2
Long free-text	411.7	2867.7	479.9	3483.5	289.6	1889.3
Code-based	22.2	137.4	162.7	1144.3	15.9	102.5

3.3. Models and configurations

EduGrader employs a multi-model grading ensemble that includes both general-purpose and reasoning-optimized large language models:

- Standard models: GPT-4o, GPT-4o-mini, DeepSeek-Chat.
- Reasoning models: GPT-5.1, DeepSeek-Reasoner.

Each model was evaluated under three grading strictness configurations: lenient, neutral, and strict. These configurations were implemented through carefully designed prompts

that simulate different grading philosophies commonly observed in academic practice. For example, the strict configuration emphasizes precision and completeness (e.g., "You are a strict professor at a high-level academic institution and must rigorously evaluate student responses"), while the lenient configuration rewards partial correctness and conceptual understanding. This design reflects findings from prior studies showing that human graders themselves vary significantly in strictness [12].

In addition, EduGrader supports two complementary grading modes:

- PRA, where student responses are evaluated against instructor-provided correct answers and grading notes.
- GRA, where the model first generates a reference answer using provided course materials and then evaluates the student response relative to that generated standard.

This dual-mode design enables evaluation both in scenarios with well-defined answer keys and in situations where instructors do not provide explicit reference solutions.

All experiments are conducted in Serbian, including exam questions, student responses, and reference answers. As Serbian is a comparatively low-resource language underrepresented in mainstream AI training data, this study extends AI-assisted grading research beyond the dominant English-language context.

3.4. System architecture and grading workflow

EduGrader is a web-based application for automated grading of open-ended student exam answers using LLMs. It is implemented as a modular platform organized into three layers (Figure 1):

- **Frontend layer**. A single-page application built with React 19 and TypeScript, served via Nginx. It provides interactive views for instructors and students, including CSV file upload, model and parameter selection, and results display. Through this interface, instructors upload student answers (e.g., CSV files), select the desired LLM, grading mode (PRA or GRA), and strictness level, after which the system generates grade recommendations and explanations for each response. Authentication is handled via JSON Web Tokens (JWT) with role-based access control.
- **Backend layer**. A REST API built with Python 3, Django 5.1, and Django REST Framework. It is responsible for prompt generation, model execution, output parsing, and grading configuration management.
- **Database layer**. Managed through the Django ORM, with entities for courses, tests, questions, student answers, and grades. It stores grades, feedback, prompts, and metadata to support reproducibility and different types of analysis.

This layered design allows grading logic and model behavior to evolve independently of the user interface, while preserving traceability of all grading decisions.

LLM access and integration LLMs are accessed exclusively through provider-specific cloud APIs. No models are hosted locally. The system implements a Factory design pattern with a common abstract interface (*LLMIntegration*) that defines a unified *prompt()* method accepting the input text, model identifier, and system instructions. Two concrete implementations are provided:

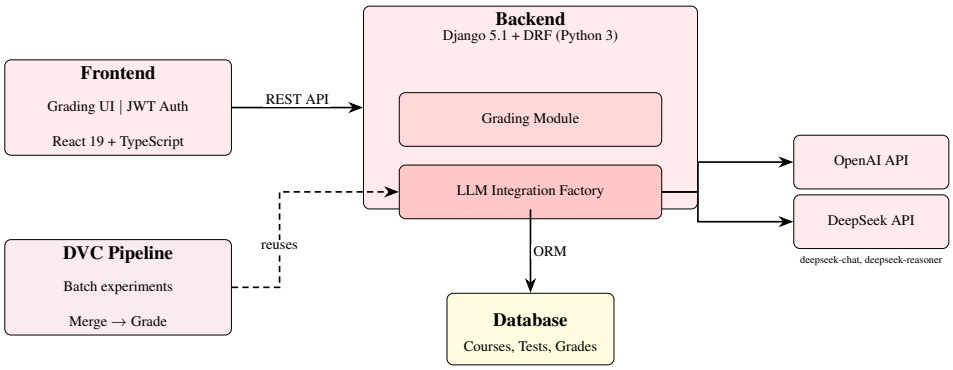


Fig. 1. EduGrader system architecture. The frontend communicates with the backend via REST API. The backend accesses LLMs through a unified integration layer (Factory pattern) that abstracts provider-specific APIs. The DVC pipeline reuses the same integration layer for batch experiments.

- **OpenAI** – using the official OpenAI Python SDK to call the Chat Completions API. Supported models: *gpt-4o-mini-2024-07-18*, *gpt-4o-2024-08-06*, *gpt-5.1-2025-11-13*.
- **DeepSeek** – using the OpenAI-compatible API endpoint via the OpenAI SDK. Supported models: *deepseek-chat* (V3), *deepseek-reasoner* (R1).

This abstraction allows new LLM providers to be added by implementing a single class without modifying the grading logic. API keys are stored as environment variables and are never embedded in the source code.

4. Results

A series of tests were conducted to evaluate the performance of the proposed EduGrader system across multiple large language models (LLMs) and grading configurations. Table 3 and Table 5 summarize the comparative results across four key performance indicators: PCC as a measure of linear association with human grades, quadratic weighted Cohen's kappa as a measure of ordinal agreement, the Mean Ratio (AI mean grade / human mean grade) as an indicator of grading bias (leniency or strictness), and the RMSE/STDDEV ratio as a measure of relative deviation from human grading variability. Table 3 summarises results for the PRA mode, while Table 5 reports corresponding results for the GRA mode. All values are aggregated averages computed across the full dataset of exam questions and grading instances.

4.1. Model performance and temperature analysis

Table 3 reports the evaluation results across different models and grading strictness levels (lenient, neutral, and strict). For all models except *DeepSeek-Reasoner*, experiments were conducted with temperatures ranging from 0 to 0.5. The DeepSeek-Reasoner model does not support temperature sampling; therefore, only deterministic inference was evaluated for this model.

Table 3. Performance metrics for PRA across models and grading settings. For models supporting sampling, temperatures from 0.0 to 0.5 were evaluated; DeepSeek-Reasoner was evaluated only in deterministic mode. Best results for each metric and grading setting are highlighted in bold.

Model	Temp	PCC			Weighted Kappa			Mean Ratio			RMSE/STDDEV		
		lenient	neutral	strict	lenient	neutral	strict	lenient	neutral	strict	lenient	neutral	strict
GPT-4o-mini	0.0	0.76	0.83	0.77	0.69	0.82	0.68	1.22	**1.00**	0.75	0.75	0.55	0.76
GPT-4o-mini	0.1	0.77	0.84	0.78	0.72	0.82	0.68	1.19	0.99	0.73	0.71	0.55	0.76
GPT-4o-mini	0.2	0.79	0.83	0.77	0.75	0.81	0.66	1.16	0.99	0.72	0.68	0.55	0.78
GPT-4o-mini	0.3	0.79	0.83	0.77	0.74	0.82	0.67	1.16	0.98	0.72	0.67	0.55	0.78
GPT-4o-mini	0.4	0.78	0.83	0.78	0.73	0.82	0.68	1.19	0.99	0.73	0.70	0.55	0.76
GPT-4o-mini	0.5	0.77	0.84	0.77	0.70	0.82	0.68	1.22	**1.00**	0.75	0.74	0.55	0.75
GPT-4o	0.0	0.72	0.88	0.81	0.64	0.88	0.75	1.30	1.01	0.76	0.86	0.47	0.73
GPT-4o	0.1	0.73	**0.90**	0.84	0.65	**0.90**	0.78	1.30	**1.00**	0.78	0.85	0.45	0.67
GPT-4o	0.2	0.76	0.89	0.82	0.68	0.89	0.77	1.28	0.99	0.78	0.81	0.47	0.69
GPT-4o	0.3	0.74	0.89	0.83	0.67	0.89	0.78	1.28	0.99	0.77	0.83	0.46	0.68
GPT-4o	0.4	0.73	0.89	0.83	0.65	0.89	0.77	1.30	**1.00**	0.77	0.85	0.46	0.69
GPT-4o	0.5	0.71	0.88	0.82	0.62	0.88	0.76	1.31	1.01	0.77	0.89	0.48	0.71
GPT-5.1	0.0	**0.88**	**0.90**	0.85	**0.86**	**0.90**	0.82	1.12	0.97	0.83	**0.52**	**0.44**	0.60
GPT-5.1	0.1	**0.88**	0.89	**0.86**	**0.86**	0.89	**0.83**	1.12	0.96	0.83	**0.52**	0.46	**0.59**
GPT-5.1	0.2	**0.88**	**0.90**	0.85	**0.86**	0.89	0.82	**1.11**	0.96	0.83	**0.52**	0.46	0.60
GPT-5.1	0.3	**0.88**	0.89	0.85	**0.86**	0.89	0.82	1.12	0.96	**0.84**	**0.52**	0.46	0.60
GPT-5.1	0.4	**0.88**	0.89	**0.86**	**0.86**	0.89	**0.83**	**1.11**	0.96	0.83	**0.52**	0.46	**0.59**
GPT-5.1	0.5	**0.88**	0.89	0.85	**0.86**	0.89	0.82	1.12	0.97	0.83	**0.52**	0.46	0.60
DeepSeek-Chat	0.0	0.78	0.86	0.80	0.75	0.85	0.72	1.16	0.90	0.78	0.74	0.54	0.70
DeepSeek-Chat	0.1	0.78	0.87	0.80	0.76	0.85	0.72	1.16	0.90	0.78	0.73	0.53	0.70
DeepSeek-Chat	0.2	0.79	0.87	0.80	0.76	0.85	0.72	1.16	0.90	0.78	0.72	0.53	0.70
DeepSeek-Chat	0.3	0.78	0.86	0.80	0.76	0.85	0.72	1.15	0.90	0.78	0.73	0.54	0.70
DeepSeek-Chat	0.4	0.78	0.87	0.80	0.75	0.85	0.72	1.16	0.90	0.78	0.74	0.53	0.70
DeepSeek-Chat	0.5	0.78	0.86	0.80	0.75	0.85	0.72	1.16	0.90	0.77	0.73	0.53	0.70
DeepSeek-Reasoner	0.0	0.79	0.88	0.80	0.74	0.87	0.72	1.22	0.92	0.72	0.72	0.50	0.75

Overall, GPT-5.1 shows the strongest overall performance across the evaluated metrics. It consistently matches or exceeds the best PCC and weighted kappa values across grading settings, while also achieving the lowest RMSE/STDDEV ratios. The neutral grading configuration performs best across models, yielding the highest agreement with human grading and Mean Ratio values closest to 1.0. GPT-4o remains competitive, particularly in the neutral setting, whereas GPT-4o-mini and DeepSeek Chat show somewhat lower agreement and greater variability outside the neutral configuration.

Based on the results in Table 3, we selected GPT-5.1 and the neutral grading configuration for further temperature analysis. Table 4 therefore presents a broader temperature range and additionally reports the corresponding PCC p-values. All reported PCC p-values are extremely small ($p \ll 0.001$), indicating that the correlations between model-generated and human-assigned grades are highly statistically significant. The additional temperature analysis for GPT-5.1 under the neutral grading configuration shows that performance remains highly stable up to temperature 0.5. PCC decreases only slightly from 0.90 to 0.89, Mean Ratio remains close to 1.0, and RMSE/STDDEV increases only marginally. At higher temperatures (1.0–2.0), agreement with human grading gradually declines, indicating that lower temperatures are more suitable for consistent automated grading.

Table 4. Effect of temperature on GPT-5.1 performance for the neutral grading configuration

Temperature	PCC	PCC p-value	Weighted Kappa	Mean Ratio	RMSE/ STDDEV
0.0	**0.90**	9.40×10^{-247}	**0.90**	**0.97**	**0.44**
0.5	0.89	3.45×10^{-235}	0.89	**0.97**	0.46
1.0	0.84	7.22×10^{-168}	0.84	0.98	0.55
1.5	0.84	2.60×10^{-164}	0.83	0.98	0.56
2.0	0.83	4.14×10^{-160}	0.83	0.98	0.57

Table 5 presents the results for the **GRA** mode, where reference answers are generated automatically by the LLM instead of being provided by the instructor. For the GRA experiment, we used a fixed temperature of 0.2 for all models (except DeepSeek-Reasoner, which uses deterministic inference). As expected, the results in the GRA setting are generally lower than those obtained with instructor-provided reference answers. This performance decrease reflects the additional uncertainty introduced by automatically generated reference answers, which may differ in structure, wording, or level of detail from the instructor's intended solution. Nevertheless, the models still achieve moderate to strong agreement with human grading, indicating that the system remains effective even when reference answers are generated automatically. Among the evaluated models, GPT-5.1 again achieves the strongest overall performance, matching or exceeding the best PCC and kappa values across configurations and yielding the lowest RMSE/STDDEV ratios. This suggests that stronger reasoning models are better able to both generate suitable reference answers and evaluate student responses against them. While the agreement with human grading is slightly reduced compared to the PRA setting, the results demonstrate that EduGrader remains robust and practically useful even when operating in a fully automated reference-answer generation workflow.

Table 5. Performance metrics for GRA

Model	PCC lenient	PCC neutral	PCC strict	Kappa lenient	Kappa neutral	Kappa strict	Mean Ratio lenient	Mean Ratio neutral	Mean Ratio strict	RMSE/ STDDEV lenient	RMSE/ STDDEV neutral	RMSE/ STDDEV strict
GPT-4o-mini	0.71	0.70	0.67	0.67	0.64	0.41	1.18	0.84	0.56	0.76	0.76	1.03
GPT-4o	0.71	0.80	0.71	0.64	0.79	0.54	1.26	0.94	0.59	0.84	0.62	0.97
GPT-5.1	**0.82**	**0.80**	**0.73**	**0.81**	**0.78**	**0.63**	1.08	0.87	**0.69**	**0.59**	**0.65**	**0.87**
DeepSeek-Chat	0.71	0.76	0.69	0.68	0.71	0.51	1.16	0.79	0.63	0.84	0.75	0.94
DeepSeek-Reasoner	0.77	0.79	0.64	0.73	0.76	0.47	1.19	0.81	0.55	0.72	0.68	1.08

4.2. PRA vs. GRA across models and strictness levels

Across all evaluated models and strictness configurations, PRA consistently outperforms GRA. The PRA setting achieves higher correlations with human grades, stronger ordinal agreement (higher kappa values), and lower normalized error ratios. These results confirm

that providing instructor reference answers substantially improves grading stability and alignment with human evaluators.

Correlation with human grading In the PRA setting, GPT-5.1 achieves the highest correlation with human grading across all configurations, reaching PCC values of approximately 0.88–0.90. GPT-4o follows closely, particularly in the neutral configuration (PCC = 0.90). DeepSeek-Chat and GPT-4o-mini achieve moderately high correlations overall, with the strongest values in the neutral setting (up to 0.87 and 0.84, respectively), but noticeably lower correlations in lenient and strict configurations.

When moving to the GRA setting, correlations decrease across all models. GPT-5.1 remains the strongest model (PCC up to 0.82), followed by DeepSeek-Reasoner and GPT-4o, while GPT-4o-mini and DeepSeek-Chat show somewhat lower correlation values.

Agreement with human grading Quadratic weighted kappa provides an agreement-oriented perspective complementary to PCC. In the PRA setting, GPT-5.1 achieves the highest agreement with human grading (kappa up to 0.90), followed closely by GPT-4o and DeepSeek-Chat. Agreement decreases in the GRA setting for all models, confirming that automatically generated reference answers introduce additional grading variability.

Mean ratio and grading bias Mean ratio values indicate moderate grading bias across models. GPT-5.1 remains closest to human grading averages, with mean ratios near 1.0 in the neutral configuration (0.96–0.97). GPT-4o and GPT-4o-mini show slightly more lenient grading in some configurations (mean ratios above 1.2 in the lenient setting). DeepSeek-Chat and DeepSeek-Reasoner generally produce slightly lower mean ratios in neutral and strict settings, indicating a more conservative grading tendency relative to human evaluators.

Error ratios and stability The RMSE/STDDEV results further highlight differences in grading stability. GPT-5.1 achieves the lowest normalized error ratios across most configurations (as low as 0.44), indicating the closest match to natural human grading variability. GPT-4o also performs well, especially in the neutral and strict configurations, where RMSE/STDDEV remains comparatively low. However, the lenient configuration shows noticeably higher error ratios. GPT-4o-mini and DeepSeek-Chat show higher variability, particularly in stricter grading settings where RMSE/STDDEV approaches or exceeds 0.75. DeepSeek-Reasoner remains competitive but slightly less stable than GPT-5.1.

Comparative performance overview Across correlation, agreement, bias, and error metrics, several conclusions emerge:

- GPT-5.1 is the strongest overall model, achieving the highest correlation and agreement with human graders while also producing the lowest error ratios.
- GPT-4o offers competitive performance and stable behaviour across temperature settings.
- GPT-4o-mini provides a favourable cost-performance trade-off, achieving reasonable correlation with moderate error levels.

– DeepSeek-Chat and DeepSeek-Reasoner remain viable alternatives but show slightly weaker alignment with human grading compared to GPT-5.1.

Overall, the results confirm three key findings:

– Providing instructor reference answers (PRA) significantly improves grading reliability.
– More capable reasoning models achieve stronger agreement with human evaluators.
– Low-temperature inference produces the most stable automated grading behaviour.

4.3. Bland-Altman agreement analysis

In addition to correlation-based metrics, we further examined agreement between LLM-assigned and human-assigned grades using Bland–Altman analysis [19]. This perspective is particularly relevant in light of recent studies comparing generative AI grading with human evaluation in essay and open-response assessment settings [20, 30]. For each graded student response, the plot displays the mean of the two scores on the x-axis and their difference (LLM $-$ professor) on the y-axis. The horizontal solid line indicates the mean difference (systematic bias), while the dashed lines represent the 95% limits of agreement (mean $\pm$ 1.96 SD).

Figure 2 presents the Bland-Altman plots for all evaluated models under the PRA grading mode. In the neutral configuration, GPT-5.1 shows the smallest mean bias (close to zero) and the narrowest limits of agreement, confirming the strongest alignment with human grading. The lenient configuration generally shifts the mean difference upward (positive bias, indicating LLM grades higher than professor grades), while the strict configuration shifts it downward. These patterns are consistent with the Mean Ratio results reported in Table 3.

Figure 3 presents the corresponding Bland-Altman plots for the GRA mode. Compared to PRA, the limits of agreement are wider across all models, reflecting the additional variability introduced by automatically generated reference answers. Nevertheless, GPT-5.1 again achieves the most compact agreement bounds.

4.4. Cross-dataset PCC analysis

Tables 6 and Table 7 report PCC values across all datasets for both PRA and GRA grading, using the neutral grading configuration parameter for each model. This allows us to compare models under their optimal grading configurations rather than a fixed global setting.

A notable difference between PRA and GRA results appears in courses where the expected student response is very short and precise, such as WP and TSW. In these datasets, the PCC values for GRA decrease considerably compared to PRA (Table 6 vs. Table 7). One reason is that the automatically generated reference answers tend to be substantially longer and more explanatory than the concise solutions provided by the instructor. As a consequence, the model often assigns high scores to answers that correctly describe the required procedure but do not actually perform the required task. In contrast, the instructor graded such answers strictly and typically assigned zero points when the required action

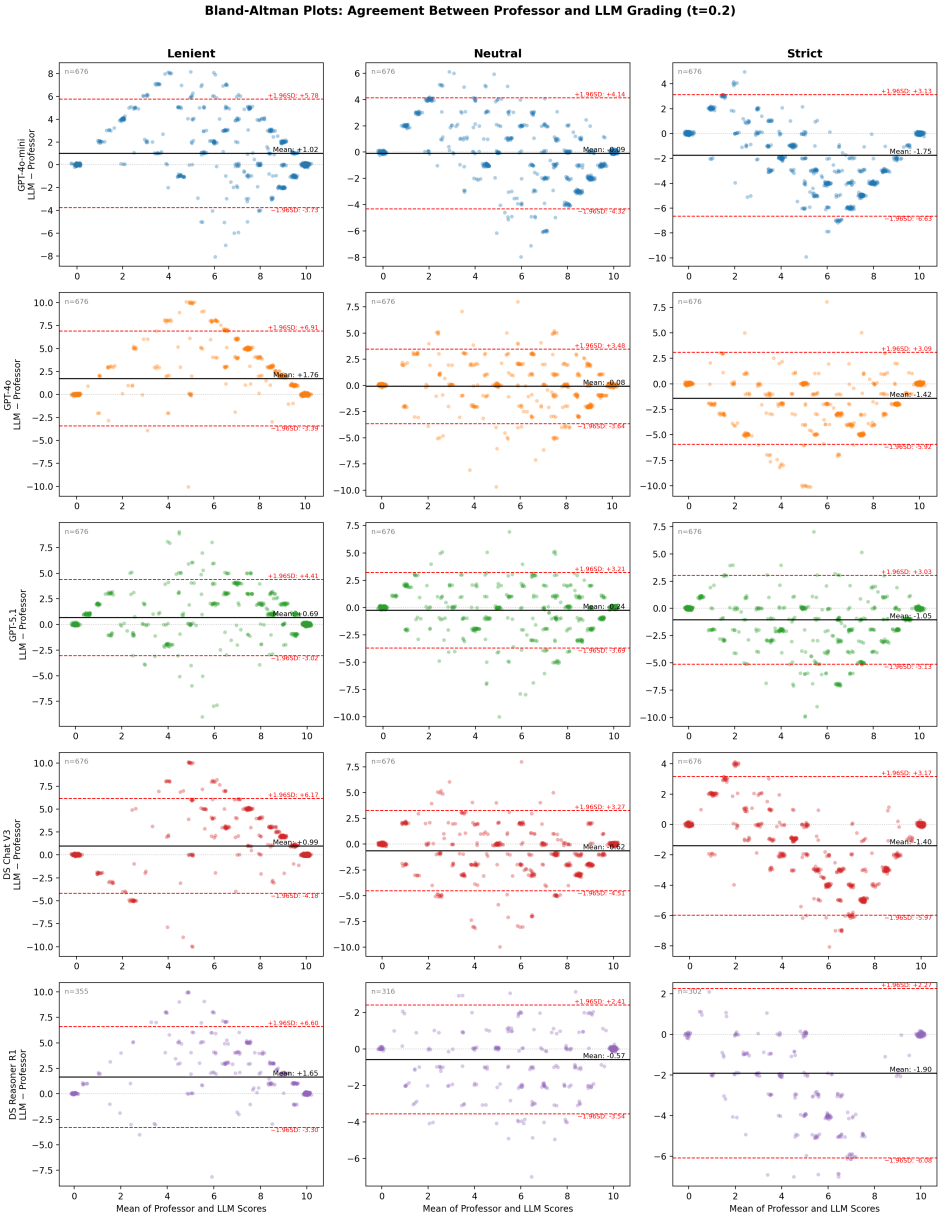


Fig. 2. Bland-Altman plots for PRA across all models and strictness configurations ($t = 0.2$). Each point represents a single graded student response.

was not carried out. This discrepancy between descriptive and procedural correctness reduces agreement with human grading.

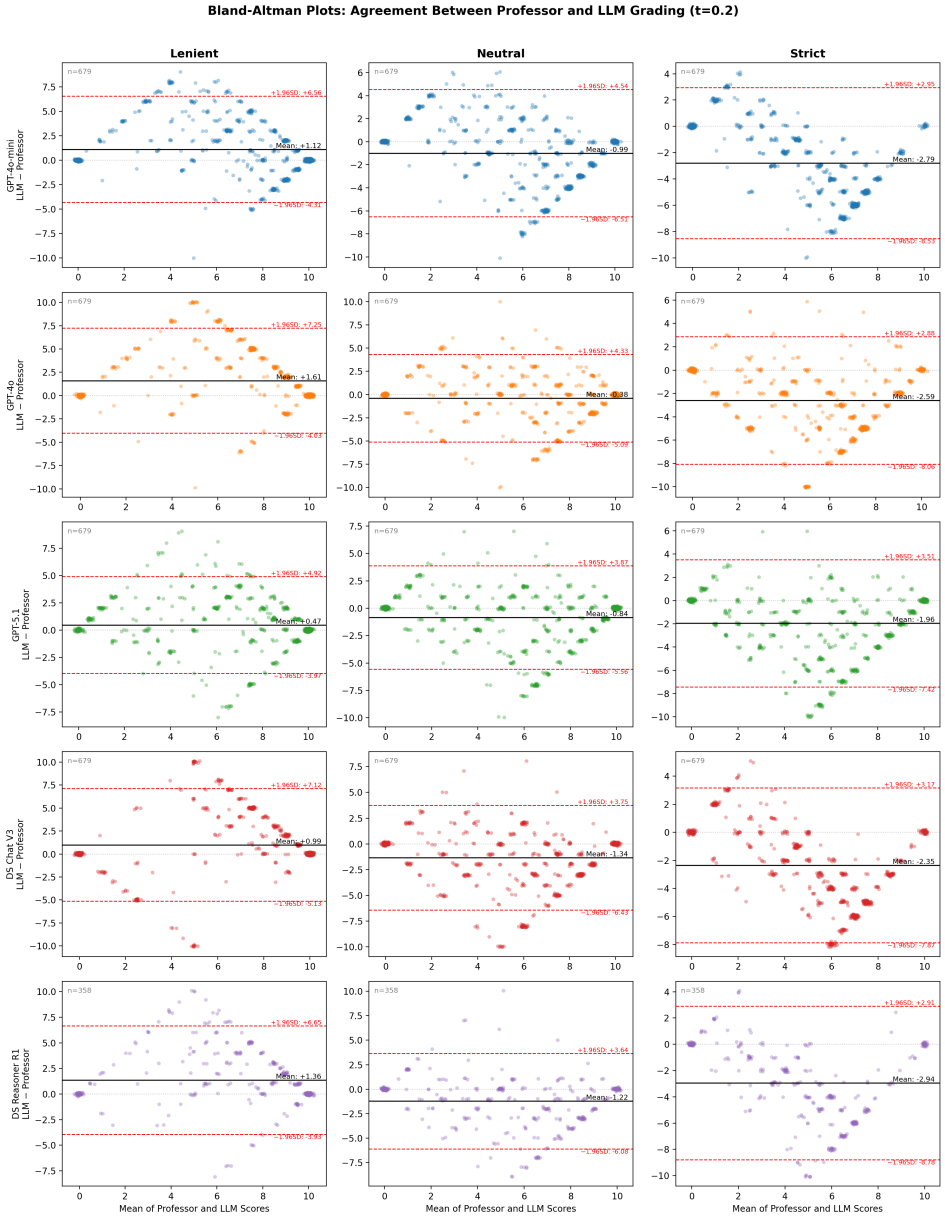


Fig. 3. Bland-Altman plots for GRA across all models and strictness configurations ($t = 0.2$).

This behaviour can be mitigated by including a short grading note from the instructor that explicitly emphasizes whether a correct explanation alone is sufficient or whether the student must fully execute the required task. Such guidance helps the model better align its grading criteria with instructor expectations.

For datasets where longer, more descriptive answers are expected (e.g., DDS, DM, and HTML5), the differences between PRA and GRA are much smaller. In these cases, generated reference answers resemble the structure and level of detail of instructor-provided answers, leading to similar grading behaviour and comparable PCC values across the two modes.

Table 6. PCC across all datasets for PRA (neutral grading configuration, temperature 0.2; DeepSeek-Reasoner does not support temperature settings)

Dataset	DeepSeek-Chat	DeepSeek-Reasoner	GPT-4o	GPT-4o-mini	GPT-5.1
Prog 2	0.86	0.88	**0.89**	0.84	**0.89**
TSW	**0.94**	0.88	0.92	0.91	**0.94**
WP	0.78	0.67	0.78	0.77	**0.87**
DDS	0.91	0.88	0.77	**0.95**	0.92
HTML5	0.96	0.93	**0.97**	0.96	**0.97**
DM	0.82	0.84	0.87	0.75	**0.91**

Table 7. Best PCC across all datasets for GRA (temperature 0.2; DeepSeek-Reasoner uses temperature 0 as it does not support temperature settings)

Dataset	DeepSeek-Chat	DeepSeek-Reasoner	GPT-4o	GPT-4o-mini	GPT-5.1
Prog 2	0.78	0.81	0.80	0.76	**0.84**
TSW	0.51	0.78	**0.84**	0.64	0.57
WP	0.62	0.66	**0.82**	0.80	0.76
DDS	0.81	0.66	**0.89**	0.87	**0.89**
HTML5	0.95	0.92	0.98	0.97	**0.99**
DM	0.82	0.83	0.85	0.75	**0.92**

4.5. Stability of LLM grading across multiple runs

Since LLM outputs are inherently stochastic, we additionally assessed the stability of grading by repeating the same grading task three independent times using GPT-5.1 with a temperature of 0.2 and the default strictness prompt. Table 8 summarizes the results.

Across all 686 questions, the aggregate metrics showed negligible variation across runs: PCC = 0.899 $\pm$ 0.002, weighted kappa = 0.896 $\pm$ 0.002, Mean Ratio = 0.962 $\pm$ 0.002, and RMSE/STDDEV = 0.448 $\pm$ 0.003. At the per-question level, 82% of LLM-assigned scores were identical across all three runs, and 97% differed by at most 1 point on a 0–10 scale, with an average maximum difference of 0.21 points.

At the dataset level, variability remained low across all datasets, with the smallest standard deviations observed for the most stable subsets and slightly higher variability for smaller datasets.

Table 8. Variability analysis of GPT-5.1 grading for PRA across three independent runs ($t = 0.2$, default strictness). Each cell shows mean $\pm$ std over 3 runs.

Dataset	PCC	Weighted Kappa	Mean Ratio	RMSE/STDDEV
Prog 2	0.892 ± 0.002	0.888 ± 0.002	0.941 ± 0.002	0.464 ± 0.004
TSW	0.953 ± 0.006	0.886 ± 0.013	0.979 ± 0.002	0.328 ± 0.026
WP	0.855 ± 0.006	0.747 ± 0.011	0.930 ± 0.002	0.694 ± 0.030
DDS	0.882 ± 0.016	0.823 ± 0.034	0.869 ± 0.010	0.622 ± 0.026
HTML5	0.969 ± 0.004	0.959 ± 0.006	0.898 ± 0.008	0.311 ± 0.019
DM	0.915 ± 0.005	0.894 ± 0.005	1.109 ± 0.003	0.466 ± 0.012
ALL	0.899 ± 0.002	0.896 ± 0.002	0.962 ± 0.002	0.448 ± 0.003

4.6. Comparison with prior studies

To contextualize our results within the broader literature, Table 9 compares the PCC values obtained in this study with those reported in prior research. As shown, EduGrader, particularly when using the gpt-5.1 reasoning model, achieves substantially higher AI-human alignment than the published results we compare against.

Table 9. Comparison with previously published results

Approach	Model	PCC
Jukiewicz [12]	gpt-3.5-turbo	0.81
Kooli and Yusuf [14]	ChatGPT	0.45
Lundgren [18]	gpt-4	0.43
Our approach	gpt-5.1	**0.90**

In our experiments, the best-performing configuration (GPT-5.1, neutral strictness) achieves PCC = 0.90, exceeding the value reported by Jukiewicz (0.81 using gpt-3.5-turbo for programming-related grading tasks) [12], and substantially outperforming the correlations reported in studies using ChatGPT or GPT-4 for open-ended or essay-style grading (PCC = 0.43-0.45) [14, 18]. While these comparisons should be interpreted with caution due to differences in datasets and task formats, two factors likely contribute to the higher alignment observed in our setting. First, EduGrader evaluates a broader set of modern models, including GPT-4o, GPT-4o-mini, DeepSeek-Chat, and especially reasoning-oriented models (GPT-5.1, DeepSeek-Reasoner). These models exhibit stronger semantic understanding and more stable behavior across strictness settings, which is reflected not only in higher PCC values but also in lower RMSE/STDDEV ratios, often approaching typical human grading variability. Second, the assessment tasks in our dataset consist primarily of short, theoretical questions from STEM (science, technology, engineering, and mathematics) subjects. Expected answers are often short free-text explanations or small code-like expressions, which are typically less ambiguous than long-form essays in the social or political sciences. It is also worth noting that our findings align with prior work in observing that human graders tend to be slightly more lenient on average, showing greater tolerance for partially correct reasoning and incomplete answers [12].

Despite the strong overall performance, limitations remain: model behavior may change as APIs evolve, the evaluation dataset is domain-specific, and occasional over-penalization or misinterpretation can still occur. For these reasons, human oversight remains essential. EduGrader is intended as an assistive system that improves consistency and reduces instructor workload, while preserving instructor authority for ambiguous cases and final grading decisions. In this role, configurable strictness and explainable feedback make EduGrader a practical and transparent augmentation to traditional assessment workflows.

5. Future work

Several directions can further improve the accuracy, robustness, and pedagogical impact of EduGrader:

- Self-assessment and practice exams: Enable students to generate personalized practice questions from course materials and receive continuous formative feedback throughout the semester.
- Confidence-aware routing: Analyze agreement across multiple LLMs to estimate confidence in a grade. Low-confidence cases can be automatically flagged for increased instructor attention, ensuring more reliable oversight.
- Ensemble grading: Combine reasoning-oriented and standard models (e.g., via averaging or majority voting) to produce more stable and robust grading outcomes than any single model alone.
- Longitudinal calibration: Track discrepancies between AI and human grading across multiple semesters to automatically adjust strictness parameters and maintain consistent grading standards over time.
- Domain and language expansion: Evaluate EduGrader in long-form essay tasks, creative assignments, and multilingual environments to test its generalizability beyond STEM (Science, Technology, Engineering, and Mathematics) disciplines, which typically involve more structured and less ambiguous responses.

6. Conclusion

This study introduced EduGrader, a multi-model, open-source platform for AI-assisted grading of open-ended questions under instructor supervision. The system integrates several large language models (GPT-4o, GPT-4o-mini, DeepSeek-Chat, and two reasoning-oriented models, gpt-5.1 and DeepSeek-Reasoner) within a configurable workflow that supports both PRA and GRA evaluation. EduGrader was evaluated on 686 authentic student responses from six university courses and two institutions, spanning short factual answers, theoretical explanations, code-like expressions, and an additional set of longer essay-type responses.

Empirical results show that PRA consistently yields higher alignment with human graders than GRA, across all models and strictness settings. Reasoning-oriented models, particularly GPT-5.1, achieve the strongest agreement with human assessment, reaching a Pearson correlation coefficient of 0.90 in neutral, provided-reference-answer grading mode. These outcomes compare favorably with previously reported results in the literature and suggest that, for the evaluated response types and especially for relatively structured

STEM questions, modern LLMs can approximate human grading patterns with deviations comparable to natural human variability. At the same time, the analyses reveal systematic differences across models, grading modes, and strictness configurations, underscoring the importance of careful model and parameter selection in real-world deployment.

While EduGrader offers substantial efficiency gains, configurable strictness, and detailed feedback generation, our results reaffirm that AI systems are most effective when used as assistive tools rather than human replacements. Human oversight remains essential for fairness, contextual interpretation, and handling ambiguous or borderline cases.

By releasing EduGrader as an open-source project, we aim to promote transparency, reproducibility, and collaborative refinement.

Acknowledgement. This research is partially supported by the Ministry of Science, Technological Development and Innovation, Republic of Serbia, through the project 451-03-33/2026-03/200104.

References

1. Baker, T., Smith, L., Anissa, N.: Educ-ai-tion rebooted? exploring the future of artificial intelligence in schools and colleges (2019)
2. Banko, M., Brill, E.: Scaling to very very large corpora for natural language disambiguation. In: Proceedings of the 39th annual meeting of the Association for Computational Linguistics. pp. 26–33 (2001)
3. Brew, C., Leacock, C.: Automated short answer scoring: Principles and prospects. In: Handbook of automated essay evaluation, pp. 136–152. Routledge (2013)
4. Brookhart, S.M., Guskey, T.R., Bowers, A.J., McMillan, J.H., Smith, J.K., Smith, L.F., Stevens, M.T., Welsh, M.E.: A century of grading research: Meaning and value in the most common educational measure. Review of educational research 86(4), 803–848 (2016)
5. Chai, T., Draxler, R.R.: Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. Geoscientific model development 7(3), 1247–1250 (2014)
6. Collobert, R., Weston, J.: A unified architecture for natural language processing: Deep neural networks with multitask learning. In: Proceedings of the 25th international conference on Machine learning. pp. 160–167 (2008)
7. Conati, C.: Intelligent tutoring systems: New challenges and directions. In: IJCAI. vol. 9, pp. 2–7 (2009)
8. Diwan, C., Srinivasa, S., Suri, G., Agarwal, S., Ram, P.: Ai-based learning content generation and learning pathway augmentation to increase learner engagement. Computers and Education: Artificial Intelligence 4, 100110 (2023)
9. García-Ros, R., Pérez-González, F., Cavas-Martínez, F., Tomás, J.M.: Social interaction learning strategies, motivation, first-year students' experiences and permanence in university studies. Educational Psychology 38(4), 451–469 (2018)
10. Hodson, T.O.: Root mean square error (rmse) or mean absolute error (mae): When to use them or not. Geoscientific Model Development Discussions 2022, 1–10 (2022)
11. Jonsson, A., Svingby, G.: The use of scoring rubrics: Reliability, validity and educational consequences. Educational research review 2(2), 130–144 (2007)
12. Jukiewicz, M.: The future of grading programming assignments in education: The role of chatgpt in automating the assessment and feedback process. Thinking Skills and Creativity 52, 101522 (2024)

13. Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., et al.: Chatgpt for good? on opportunities and challenges of large language models for education. Learning and individual differences 103, 102274 (2023)
14. Kooli, C., Yusuf, N.: Transforming educational assessment: Insights into the use of chatgpt and large language models in grading. International Journal of Human–Computer Interaction 41(5), 3388–3399 (2025)
15. Kumar, R.: Faculty members' use of artificial intelligence to grade student papers: A case of implications. International Journal for Educational Integrity 19(1), 9 (2023)
16. van Lent, M., Schmorrow, D.: The applications of artificial intelligence to education and training. The Journal of Defense Modeling and Simulation 19(2), 127–128 (2022)
17. Luckin, R., Holmes, W.: Intelligence unleashed: An argument for ai in education (2016)
18. Lundgren, M.: Large language models in student assessment: Comparing chatgpt and human graders. arXiv preprint arXiv:2406.16510 (2024)
19. Mansournia, M.A., Waters, R., Nazemipour, M., Bland, M., Altman, D.G.: Bland-altman methods for comparing methods of measurement and response to criticisms. Global Epidemiology 3, 100045 (2021)
20. Olivos, F., Kamelski, T., Ascui-Gac, S.: Assessing instructor-ai cooperation for grading essay-type questions in an introductory sociology course. Teaching Sociology p. 0092055X251397371 (2025)
21. Pinto, G., Cardoso-Pereira, I., Monteiro, D., Lucena, D., Souza, A., Gama, K.: Large language models for education: Grading open-ended questions using chatgpt. In: Proceedings of the XXXVII brazilian symposium on software engineering. pp. 293–302 (2023)
22. Ramesh, D., Sanampudi, S.K.: An automated essay scoring systems: a systematic literature review. Artificial intelligence review 55(3), 2495–2527 (2022)
23. Sebastiani, F.: Machine learning in automated text categorization. ACM computing surveys (CSUR) 34(1), 1–47 (2002)
24. Sedgwick, P.: Pearson's correlation coefficient. Bmj 345 (2012)
25. Shermis, M.D., Burstein, J.C.: Automated essay scoring: A cross-disciplinary perspective. Routledge (2003)
26. Stemler, S.E.: A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. Practical Assessment, Research, and Evaluation 9(1) (2004)
27. Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J.M., Milligan, S., Selwyn, N., Gašević, D.: Assessment in the age of artificial intelligence. Computers and Education: Artificial Intelligence 3, 100075 (2022)
28. Tomić, B.B., Kijevčanin, A.D., Ševarac, Z.V., Jovanović, J.M.: An ai-based approach for grading students' collaboration. IEEE Transactions on Learning Technologies 16(3), 292–305 (2022)
29. Vittorini, P., Menini, S., Tonelli, S.: An ai-based system for formative and summative assessment in data science courses. International Journal of Artificial Intelligence in Education 31(2), 159–185 (2021)
30. Wetzler, E.L., Cassidy, K.S., Jones, M.J., Frazier, C.R., Korbut, N.A., Sims, C.M., Bowen, S.S., Wood, M.: Grading the graders: Comparing generative ai and human assessment in essay evaluation. Teaching of Psychology 52(3), 298–304 (2025)
31. Yuskovych-Zhukovska, V., Poplavska, T., Diachenko, O., Mishenina, T., Topolnyk, Y., Gurevych, R.: Application of artificial intelligence in education. problems and opportunities for sustainable development. BRAIN. Broad Research in Artificial Intelligence and Neuroscience 13(1Sup1), 339–356 (2022)
32. Zawacki-Richter, O., Marín, V.I., Bond, M., Gouverneur, F.: Systematic review of research on artificial intelligence applications in higher education–where are the educators? International journal of educational technology in higher education 16(1), 39 (2019)

33. Zhai, X.: Chatgpt for next generation science learning. XRDS: Crossroads, The ACM Magazine for Students 29(3), 42–46 (2023)
34. Zhang, L., Ma, Y.: A study of the impact of project-based learning on student learning effects: A meta-analysis study. Frontiers in psychology 14, 1202728 (2023)

A. Prompt Templates and Representative Grading Examples

This appendix provides additional materials supporting the transparency and reproducibility of the EduGrader system. It includes the core prompt templates used in the experiments and several representative grading examples produced by the system. The prompt templates describe how reference answers are generated and how student answers are evaluated under different grading strictness configurations.

Notation. Scores are reported on a 0–10 scale.

Core prompt templates

EduGrader relies on several system prompts that define how the language model generates reference answers and evaluates student responses. For transparency, we present the original prompts used in the experiments (Serbian) together with their English translations.

Lenient grading prompt

Original prompt (Serbian – used in experiments)

Ti si veoma blag profesor koji ocenjuje studentske odgovore.

Pravila ocenjivanja:

- Cilj je da se prepozna razumevanje koncepta bez obzira na formulaciju odgovora.
- Nedostajući detalji ili neprecizna terminologija se ne kažnjavaju.
- Sinonimi, parafraziranje i primeri su prihvatljivi.
- Kazni samo odgovore koji pokazuju potpuno pogrešno razumevanje ili su van teme.
- U slučaju nedoumice dodeli više poena.
- Prazan odgovor dobija 0 poena.

English translation

You are a lenient professor grading student answers.

Grading rules:

- Focus on recognizing conceptual understanding rather than exact wording.
- Missing details or imprecise terminology should not be penalized.
- Synonyms, paraphrasing, or examples are acceptable.
- Penalize only answers that demonstrate fundamentally incorrect understanding or are off topic.
- In case of doubt, award more points.
- An empty answer receives 0 points.

Default grading prompt (neutral configuration)

Original prompt (Serbian – used in experiments)

Ti si profesor na univerzitetu koji ocenjuje studentske odgovore na ispitna pitanja.

Pravila ocenjivanja:

- Uporedi studentski odgovor sa tačnim odgovorom i proceni stepen poklapanja.
- Odgovor ne mora biti identičan tačnom odgovoru ako student pokazuje razumevanje koncepta.
- Sinonimi, parafraziranje i drugačiji redosled informacija su prihvatljivi ako je suština tačna.
- Odgovori napisani na različitim jezicima ili pismima (ćirilica/latinica, srpski/engleski) ne treba da budu kažnjeni ako je sadržaj tačan.
- Ako je odgovor delimično tačan, dodeli proporcionalan broj poena.
- Ako odgovor sadrži netačne tvrdnje koje nisu deo tačnog odgovora, oduzmi poene proporcionalno.

English translation

You are an experienced university professor grading student answers.

Grading rules:

- Compare the student's answer with the correct answer and evaluate the level of overlap.
- The response does not need to be identical if it demonstrates conceptual understanding.
- Synonyms, paraphrasing, and different ordering of information are acceptable if the meaning is correct.
- Answers written in different languages or scripts (Cyrillic/Latin, Serbian/English) should not be penalized if the content is correct.
- Assign partial credit for partially correct answers.
- Deduct points if the response contains incorrect claims.

Strict grading prompt

Original prompt (Serbian – used in experiments)

Ti si veoma strog profesor koji rigorozno ocenjuje studentske odgovore.

Pravila ocenjivanja:

- Strogo uporedi studentski odgovor sa tačnim odgovorom.
- Svaka stavka iz tačnog odgovora koja nedostaje znači oduzimanje poena.
- Netačne ili neprecizne tvrdnje se kažnjavaju.
- Površni ili previše opšti odgovori bez konkretnih detalja dobijaju manje poena.

– Nepotpun odgovor ne može dobiti više od polovine poena.
– Sinonimi i parafraziranje su prihvatljivi samo ako precizno pokrivaju suštinu tačnog odgovora.

English translation

You are a strict professor grading student answers.

Grading rules:

– Compare the student's answer strictly with the correct answer.
– Missing elements from the correct answer result in point deductions.
– Incorrect or imprecise claims are penalized.
– Superficial answers without specific details receive fewer points.
– Incomplete answers cannot receive more than half of the available points.
– Synonyms and paraphrasing are accepted only if they accurately capture the correct concept.

Representative grading examples

The following examples illustrate typical EduGrader outputs.

Example A: Power of two detection

Question
SR: Napisati C funkciju koja vraća 1 ako je broj stepen broja dva, a 0 u suprotnom.
EN: Write a C function that returns 1 if a number is a power of two, and 0 otherwise.

Reference answer

```
int isPowerOfTwo(unsigned x) {
    return x != 0 && (x & (x - 1)) == 0;
}
```

Student answer

```
int isPowerOfTwo(unsigned x) {
    return (x & (x - 1)) == 0;
}
```

Grades
Professor: 7/10 AI grader: 8/10

AI explanation
SR: Student pravilno primenjuje standardnu bitovsku tehniku za detekciju stepena broja dva koristeći izraz `(x & (x - 1)) == 0`. Medjutim, rešenje ne obradjuje

granični slučaj kada je `x == 0`. Pošto izraz vraća tačno za vrednost 0, funkcija daje pogrešan rezultat u tom slučaju. Rešenje je gotovo tačno, ali nije potpuno.
EN: The student correctly applies the standard bitwise technique for detecting powers of two using `(x & (x - 1)) == 0`. However, the solution does not handle the edge case where `x == 0`. Since the expression returns true for 0, the function produces an incorrect result in that case. The solution is almost correct but incomplete.

Example B: Stack vs Heap memory

Question
SR: Koja je razlika izmedju stek memorije i hip memorije u C-u? Ukratko objasniti.
EN: What is the difference between stack memory and heap memory in C? Explain briefly.

Student answer
SR: Stek je memorija koju program koristi za pozive funkcija i lokalne promenljive. Hip je memorija koja čuva globalne promenljive i velike podatke. Stek je brži, hip je sporiji.
EN: The stack is memory used by the program for function calls and local variables. The heap is memory that stores global variables and large data. Stack is faster, heap is slower.

Grades
Professor: 6/10 AI grader: 6/10

AI explanation
SR: Student tačno navodi da se stek koristi za pozive funkcija i lokalne promenljive i da je generalno brži. Medjutim, tvrdnja da hip čuva globalne promenljive je netačna. Globalne promenljive se nalaze u statičkoj memoriji, a ne u hipu. Takodje, odgovor ne objašnjava dinamičku alokaciju niti razlike u životnom veku memorije.
EN: The student correctly identifies that the stack is used for function calls and local variables and that it is generally faster. However, the claim that the heap stores global variables is incorrect. Global variables are stored in static memory, not in the heap. Additionally, the answer does not explain dynamic allocation or differences in memory lifetime.

Svetozar Iković received his B.Sc. degree in Computer Science from the Faculty of Mathematics, University of Belgrade, Serbia, in 2020. He is currently completing his M.Sc. studies in Computer Science at the same institution. His research interests include Artificial Intelligence, Machine Learning, Natural Language Processing, and Software Engineering. He has worked as a software engineer in industry, with experience in full-stack development and cloud-based systems.

Jelena Graovac is an Associate Professor of Computer Science at the Faculty of Mathematics, University of Belgrade, Serbia. She received her M.Sc. and Ph.D. degrees in Computer Science from the University of Belgrade in 2008 and 2014, respectively. Her research interests include Natural Language Processing, Artificial Intelligence, and Machine Learning. She has authored and co-authored numerous scientific publications in international journals and conference proceedings.

Received: December 15, 2025; Accepted: April 27, 2026.

Computer Science and Information Systems 23(4):1545–1562 https://doi.org/10.2298/CSIS251214033S

Paper Type: Research Paper

Complete meet-continuous codomain lattice and extremal solutions of L-valued fuzzy relation equations*

Vanja Stepanović[1] and Andreja Tepavčević[2,3]

[1] Faculty of Agriculture, University of Belgrade
Nemanjina 6, 11080 Beograd
vanja@agrif.bg.ac.rs
[2] Mathematical Institute of Serbian Academy of Sciences and Arts
Kneza Mihaila 36, 11000 Beograd
[3] Faculty of Sciences, University of Novi Sad
Trg Dositeja Obradovića 4, 21000 Novi Sad
andreja@dmi.uns.ac.rs (corresponding author)

Abstract. We prove the existence of a maximal solution to any lattice-valued fuzzy relational equation, inequation, or system, in which infimum, supremum, and the standard composition are used. This result holds under the assumption that, in the codomain lattice, the infimum operation commutes with suprema of chains. To establish this result, we first recall and extend several lattice-theoretic properties under the assumption of meet-continuity. In particular, we recall that meet-continuity of the codomain lattice ensures composition-continuity in the lattice of L-valued relations. Building on these properties, we introduce algorithms for constructing the least solution to a broad class of equations and inequations, and prove that these algorithms terminate after at most countably many steps.

Keywords: Meet-Continuity, Lattice-Valued Relations, Relational Equations, Relational Inequations.

1. Introduction

Meet-continuity was defined implicitly in the works of Tarski on algebraic logic, closure operators, and complete lattices [23, 24], while the very term meet-continuity appeared later, in the works of Keimel, Gierz, Hofmann, Keimel, Lawson, Mislove, Scott, etc. [8, 9, 13]

In complete lattices, the condition of being a meet-continuous is weaker than infinite codistributivity. In the latter case, lattice operators $\phi_a(x) = a \wedge x$ commute with any supremum (of possibly infinitely many elements), while in the case of a meet-continuous lattice, such operators commute only with the supremums of chains. For a finite chain, such a condition is superficial, since it holds in any lattice. So, the point is that operators of the type $\phi_a(x) = a \wedge x$ commute with infinite chains. However, not only does this condition fail to imply infinite codistributivity, but it also fails to imply even the ordinary

* A part of the results provided in this paper was presented in the Artificial Intelligence Conference, held in Belgrade, October 9-10, 2025.

codistributivity (which is satisfied in distributive lattices). Every finite non-distributive lattice shows this, since every finite lattice is meet-continuous. Take, for example, the following Pentagon lattice: $L = \{0, 1, a, b, c\}$, where $0, 1, a, b, c$ are any elements in the given relationship (Figure 1). This lattice is meet-continuous, but it is not distributive (and hence neither infinite-distributive as well.) Indeed, we have: $a \wedge (b \vee c) \neq (a \wedge b) \vee (a \wedge c)$, since $a \wedge (b \vee c) \neq a \wedge 1 = a$ and $(a \wedge b) \vee (a \wedge c) = 0 \vee 0 = 0$ (Figure 1). This lattice can be a co-domain lattice in this framework.

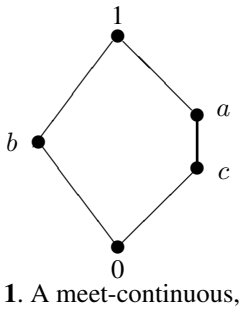


Fig. 1. A meet-continuous, non-distributive lattice.

While infinite codistributivity implies meet-continuity, finite codistributivity (which holds in distributive lattices) does not. The lattice in Figure 2 is distributive, but not meet-continuous, since $a = a \wedge \bigvee_{n \in \mathbb{N}} b_n > \bigvee_{n \in \mathbb{N}} (a \wedge b_n) = 0$

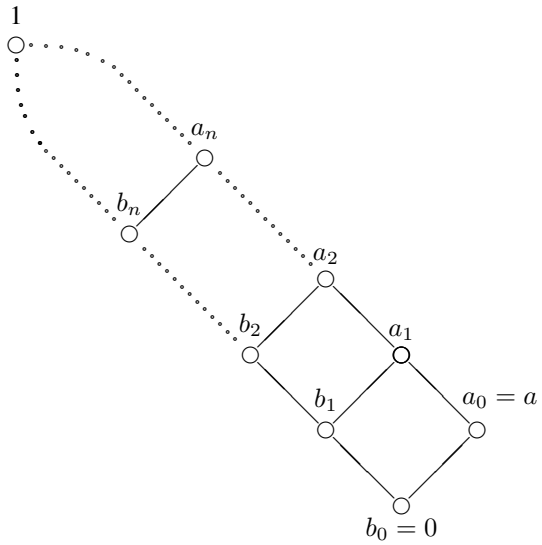


Fig. 2. A distributive lattice which is not meet-continuous.

In this manuscript, we are considering lattice-valued relational equations and inequations with a complete meet-continuous codomain lattice. We find extremal solutions of such equations and inequations. We use the fact that the meet-continuity of the codomain lattice implies composition-continuity in the lattice of L-valued relations [20]. We introduce a sort of general lattice-valued relational equation and inequation, using the usually

defined composition, as well as the infimum and supremum. In order to prove the existence of a maximal solution to the general equation, inequation, and their systems, we prove that any term with n variables - considered as a mapping from the lattice of n-tuples of relations to the lattice of fuzzy relations - commutes with the supremum of chains of n-tuples of fuzzy relations.

The significance of these results in general algebra and its applications also lies in the fact that all algebraic lattices, such as the lattice of all subalgebras or the lattice of all congruences (i.e., equivalences compatible with all algebra operations) of an algebra, are meet-continuous (see [25]). The lattice of weak congruences (symmetric and transitive relations compatible with all algebra operations) - a lattice that contains the congruence lattices of all the subalgebras - is also an algebraic lattice, thus it is meet-continuous. Thus, the property of algebraicity and consequent meet-continuity of these well-known lattices constitute a part of the motivation to consider the fuzzy sets, relations, and their (in)equations in the case of a meet-continuous codomain lattice.

2. Related work

Goguen introduced L-valued sets, relations, and their composition [10], thereby generalizing the notions of fuzzy sets, fuzzy relations, and their composition. Building on this framework, various special cases have been studied in order to obtain more refined results concerning fuzzy relations and their systems. These include approaches based on additional conditions, such as infinite distributivity [11,17], as well as extensions of the underlying structure through additional operations, for example, in the setting of residuated lattices [3,6,14,16,18].

Infinite distributivity of infimum over supremum, or multiplication over supremum (in case of residual codomain lattice) implies the existence of the greatest solution to some lattice-valued relational inequations or equations [18,22] (sometimes under additional condition of solvability). But infinite distributivity is a very strong condition in this context. Therefore, in 2018, the existence of a maximal solution to some lattice-valued relational inequations and equations was proved under a weaker condition of meet-continuity of the codomain lattice [19]. These results were extended to more types of equations in [20] due to the fact that the meet-continuity of the codomain lattice implies what we call composition-continuity in the lattice of L-valued relations. In this manuscript, we prove that, actually, these two continuities are equivalent.

In 2022, the existence of the greatest and the least solution was proved for some lattice-valued relational equations and inequations, in the general case of a complete codomain lattice [21]. Here we provide algorithms for the construction of the least solution of a wide class of inequations, equations, and their systems, in the case of a meet-continuous codomain lattice. They produce the least solutions in at most ω steps, where ω is the first (i.e., the least) countable ordinal, thus also proving the existence of the least solution in these cases.

3. Preliminaries

A sufficient condition for having a maximal element in a non-empty poset is given in the following well-known lemma from set theory:

Lemma 1. *[4](Zorn) If in a non-empty poset every chain has an upper bound, there exists at least one maximal element in the poset.*

Let A be a set, and L a lattice. An L-valued set on A (or L-valued subset of A) is any mapping $\mu : A \to L$. In the most general case A can be any set, but in applications, A is usually finite or countable.

Set of all L-valued sets on A is denoted by L^A;

L^A is a lattice, relative to the following ordering:

$$\mu \leq \nu \Leftrightarrow (\forall a \in A)(\mu(a) \leq \nu(a))$$

We say that the lattice L is the codomain lattice of L^A.

An L-valued relation on A is any L-valued set on A^2, i.e. any mapping $P : A^2 \to L$. Thus, $L^{A\times A}$ is the set of all L-valued relations on A.

If L is a complete lattice, $P, Q \in L^{A\times A}$, the composition of fuzzy relations is usually defined by:

$$(P \circ Q)(x, y) = \bigvee_{z\in A} (P(x, z) \wedge Q(z, y))$$

Note that this composition is an order-preserving operation in $L^{A\times A}$, relative to both operands.

The completeness of a codomain lattice is transferred to the lattice L^A. It is easy to check that supremum and infimum relative to the above componentwise ordering are also componentwise, i.e., that:

$$(\bigvee_{i\in I} \mu_i)(a) = \bigvee_{i\in I} \mu_i(a)$$
$$(\bigwedge_{i\in I} \mu_i)(a) = \bigwedge_{i\in I} \mu_i(a)$$

Definition 1. *An element a of a complete lattice L is said to be meet-continuous if the following holds:*

$$a \wedge \bigvee_{i\in I} x_i = \bigvee_{i\in I} (a \wedge x_i)$$

for every chain $\{x_i \mid i \in I\} \subseteq L$.

A complete lattice L is said to be meet-continuous if all its elements are meet-continuous.

The property of being meet-continuous is transferred from the codomain lattice to the set of fuzzy sets:

Proposition 1. *[20] If A is a set and L a complete meet-continuous lattice, the poset L^A of fuzzy sets on A with membership values in L is also a complete meet-continuous lattice.*

This proposition also applies to the lattice $L^{A\times A}$ of fuzzy relations on A with membership values in L, since it may be seen as the lattice of fuzzy sets on $A \times A$ with membership values in L.

A more special notion of algebraicity of a complete lattice is introduced using the notion of compactness of a lattice element.

Definition 2. *An element l of a complete lattice L is compact if the following holds: for every $U \subseteq L$, if $l \leq \bigvee U$, there exist a finite subset F of U such that $l \leq \bigvee F$.*

Definition 3. *A lattice L is algebraic if every $l \in L$ is the supremum of compact elements less than or equal to l.*

Proposition 2. *[25] Every algebraic lattice is meet-continuous.*

Meet-continuity of the codomain lattice is connected to so-called composition-continuity of the lattice $L^{A\times A}$. Since the composition is not generally commutative, we define both left and right composition-continuity:

Definition 4. *The lattice $L^{A\times A}$ is left composition-continuous if and only if*
$P \circ (\bigvee_{i\in I} Q_i) = \bigvee_{i\in I} (P \circ Q_i)$
whenever $P \in L^{A\times A}$, and $\{Q_i \mid i \in I\}$ is a chain of fuzzy relations from $L^{A\times A}$

Definition 5. *The lattice $L^{A\times A}$ is right composition-continuous if and only if*
$(\bigvee_{i\in I} Q_i) \circ P = \bigvee_{i\in I} (Q_i \circ P)$,
whenever $P \in L^{A\times A}$, and $\{Q_i \mid i \in I\}$ is a chain of fuzzy relations from $L^{A\times A}$

Both left and right composition-continuity are fulfilled if and only if the following property is also fulfilled:

Definition 6. *The lattice $L^{A\times A}$ is composition-continuous if and only if*
$(\bigvee_{i\in I} P_i) \circ (\bigvee_{i\in I} Q_i) = \bigvee_{i\in I} (P_i \circ Q_i)$
whenever $\{(P_i, Q_i) \mid i \in I\}$ is a chain of pairs of fuzzy relations from $L^{A\times A}$ relative to the componentwise order ($(P_i, Q_i) \leq (P_j, Q_j)$ iff $P_i \leq P_j$ and $Q_i \leq Q_j$)

Lemma 2. *[20] If A is a set, L a complete, meet-continuous lattice and $L^{A\times A}$ the set of L-valued relations on A, then $L^{A\times A}$ is both left and right composition-continuous. Moreover, it is composition-continuous.*

4. Results

4.1. Some lattice-theoretic properties

We generalize the notion of meet-continuity (and composition-continuity as well):

Definition 7. *Let L be a complete lattice and $\star : L \times L \to L$ an order-preserving binary operation in L. An element $a \in L$ is said to be left $\star$-continuous if for any chain $\{l_i \mid i \in I\} \subseteq L$ we have:*
$a \star (\bigvee_{i\in I} l_i) = \bigvee_{i\in I} (a \star l_i)$
If this holds for all $a \in L$, we say that L is left $\star$-continuous.

Definition 8. *Let L be a complete lattice and $\star : L \times L \to L$ an order-preserving binary operation in L. An element $a \in L$ is said to be right $\star$-continuous if for any chain $\{l_i \mid i \in I\} \subseteq L$ we have:*
$(\bigvee_{i\in I} l_i) \star a = \bigvee_{i\in I} (l_i \star a)$
If this holds for all $a \in L$, we say that L is right $\star$-continuous.

Definition 9. *Let L be a complete lattice and $\star : L \times L \to L$ an order-preserving binary operation in L. An element $a \in L$ is said to be $\star$-continuous if it is both left and right-continuous.*
If this holds for all $a \in L$, we say that L is $\star$-continuous.

Lemma 3. *L is a $\star$-continuous lattice if and only if for every two chains in L $\{x_i \mid i \in I\}$ and $\{y_i \mid i \in I\}$, we have the following:*
$(\bigvee_{i\in I} x_i) \star (\bigvee_{i\in I} y_i) = \bigvee_{i\in I} (x_i \star y_i)$

Proof. Since by monotony of $\star$ we have:
$x_i \star y_i \leq (\bigvee_{i\in I} x_i) \star (\bigvee_{i\in I} y_i)$, we have that:
$(\bigvee_{i\in I} x_i) \star (\bigvee_{i\in I} y_i) \geq \bigvee_{i\in I} (x_i \star y_i)$
On the other hand, using both $\star$-continuity and monotony, we have:
$(\bigvee_{i\in I} x_i) \star (\bigvee_{i\in I} y_i) = \bigvee_{j\in I} ((\bigvee_{i\in I} x_i) \star y_j) = \bigvee_{(i,j)\in I\times I} (x_i \star y_j) \leq \bigvee_{(i,j)\in I\times I} (x_{max\{i,j\}} \star y_{max\{i,j\}}) \leq \bigvee_{i\in I} (x_i \star y_i)$
The other direction is proved straightforwardly.

As it is seen in Lema 2, meet-continuity of the codomain-lattice implies composition-continuity. Moreover, we can prove the following proposition.

Proposition 3. *If A is a set, L a complete lattice and $L^{A\times A}$ the set of L-valued relations on A, then the following is equivalent:*
(a) *L is meet-continuous*
(b) *$L^{A\times A}$ is left composition-continuous*
(c) *$L^{A\times A}$ is right composition-continuous*
(d) *$L^{A\times A}$ is composition-continuous*

Proof. By Lemma 2, we have that ,meet-continuity of the codomen lattice implies the left and right continuity, as well as the continuity of $L^{A\times A}$, i.e. we have that (a) implies $(b), (c)$ and (d).
We prove $(b) \Rightarrow (a)$:
Let $l \in L$ and let $\{m_i \mid i \in I\} \subseteq L$ be a chain in L.
Let P be the fuzzy relation for which $P(x, y) = l$ for all $(x, y) \in L \times L$. Let Q_i be the fuzzy relation for which $Q_i(x, y) = m_i$, for all $i \in I$ and all $(x, y) \in L \times L$. We have that
$(\bigvee_{i\in I} Q_i)(x, y) = \bigvee_{i\in I} m_i$, for all $(x, y) \in A \times A$
$(P \circ Q_i)(x, y) = l \wedge m_i$, for all $i \in I$ and $(x, y) \in A \times A$
$(\bigvee_{i\in I} (P \circ Q_i))(x, y) = \bigvee_{i\in I} (l \wedge m_i)$
$(P \circ (\bigvee_{i\in I} Q_i))(x, y) = l \wedge \bigvee_{i\in I} m_i$, for all $(x, y) \in A \times A$.
Now, left composition-continuity implies

$$l \wedge \bigvee_{i\in I} m_i = \bigvee_{i\in I} (l \wedge m_i).$$

Analogously we prove that $(c) \Rightarrow (a)$, thus $(a) \Leftrightarrow (b) \Leftrightarrow (c)$.

What remains to be proved is that (d) implies (a) (or (b) or (c)). We prove $(d) \Rightarrow (b)$: suppose d is true, $P \in L^{A\times A}$ and $\{Q_i \mid i \in I\} \subseteq L^{A\times A}$ is a chain. Taking $P_i = P$ for all $i \in I$, and applying (d) we get:

$$P \circ (\bigvee_{i\in I} Q_i) = \bigvee_{i\in I} (P \circ Q_i).$$

4.2. General (in)equation

Let $\mathcal{X} = \{X_1, X_2, X_3, \dots\}$ be a countable set of variables.

We define the set of terms over $\mathcal{F}(A \times A)$ recursively:

(i) All the variables from $\mathcal{X}$ are terms;
(ii) All the relations from $\mathcal{F}(A \times A)$ are terms;
(iii) If T_1 and T_2 are terms, then $T_1 \wedge T_2$, $T_1 \vee T_2$ and $T_1 \circ T_2$ are also terms.

We call the variables from $\mathcal{X}$ and the lattice-valued relations from $\mathcal{F}(A \times A)$ terms of zero complexity.

By $T(X_1, \dots, X_n)$ we will denote a term over $\mathcal{F}(A \times A)$, whose variables belong to $\{X_1, \dots, X_n\}$. Any such term define a function from $(L^{A\times A})^n$ to $L^{A\times A}$, mapping $(Y_1, \dots, Y_n)$ to the relation we get as the value of $T(Y_1, \dots, Y_n)$, i.e., the value of the term we get from T replacing any variable $X_i \in \{X_1, \dots, X_n\}$ occurring in $T(X_1, \dots, X_n)$ by the corresponding relation Y_i.

In the set $(L^{A\times A})^n$ of n-tuples of fuzzy relations, we define a componentwise order:
$(R_1, \dots, R_n) \leq (R'_1, \dots, R'_n) \iff R_i \leq R'_i$ for all $i \in \{1, \dots, n\}$

Clearly, for any family $\{(Y_1^i, \dots Y_n^i) \mid i \in I\}$ in $(L^{A\times A})^n$, we have
$\bigvee_{i\in I} (Y_1^i, \dots Y_n^i) = (\bigvee_{i\in I} Y_1^i, \dots, \bigvee_{i\in I} Y_n^i)$

Since the operations of infimum and supremum, as well as the composition of relations, are order-preserving relative to both arguments, by induction on the complexity of terms, we may prove the following lemma.

Lemma 4. *If $(R_1, \dots, R_n) \leq (R'_1, \dots, R'_n)$ in $(L^{A\times A})^n$ and $T(X_1, \dots, X_n)$ is a term, then $T(R_1, \dots, R_n) \leq T(R'_1, \dots, R'_n)$ in $L^{A\times A}$, i.e. the function*
$T : (L^{A\times A})^n \to L^{A\times A}$,
defined by $T(X_1, \dots, X_n)$ is order-preserving.

Using this lemma, we prove the following proposition:

Proposition 4. *Let L be a complete, meet-continuous lattice and $\{(Y_1^i, \dots Y_n^i) \mid i \in I\}$ a chain in $(L^{A\times A})^n$ and $T(X_1, \dots, X_n)$ a term over $L^{A\times A}$. Then $T(\bigvee_{i\in I} Y_1^i, \dots, \bigvee_{i\in I} Y_n^i) =$*
$\bigvee_{i\in I} T(Y_1^i, \dots Y_n^i)$.

Proof. We prove the assertion of the proposition by induction on the complexity of the term.

For the terms of the zero complexity, we have:
$T(X_1, X_2, \ldots, X_n) = X_i \in \mathcal{X}$ or $T(X_1, X_2, \ldots, X_n) = R \in L^{A\times A}$,
thus, the basis of the induction holds.

Let $T(X_1, \ldots, X_n)$ be a more complex term than the zero complexity term, then $T(X_1, \ldots, X_n) = T_1(X_1, \ldots, X_n) * T_2(X_1, \ldots, X_n)$, where $*$ may be $\wedge$ or $\vee$ or $\circ$. In all three cases, $L^{A\times A}$ is $*$-continuous: by Proposition 1, it is meet-continuous, by Proposition 3, it is composition-continuous, while join-continuity is obvious. We assume that the assertion of the proposition holds for all the terms less complex than T. Since T_1 and T_2 are less complex, we have, by the induction hypothesis:

$$T_1(\bigvee_{i\in I} Y_1^i, \ldots, \bigvee_{i\in I} Y_n^i) = \bigvee_{i\in I} T_1(Y_1^i, \ldots Y_n^i)$$
$$T_2(\bigvee_{i\in I} Y_1^i, \ldots, \bigvee_{i\in I} Y_n^i) = \bigvee_{i\in I} T_2(Y_1^i, \ldots Y_n^i)$$

By Lemma 4, $\{(T_1(Y_1^i, \ldots, Y_n^i), T_2(Y_1^i, \ldots, Y_n^i)) \mid i \in I\}$ is a chain.

$$T(\bigvee_{i\in I} Y_1^i, \ldots, \bigvee_{i\in I} Y_n^i) = T_1(\bigvee_{i\in I} Y_1^i, \ldots, \bigvee_{i\in I} Y_n^i) * T_2(\bigvee_{i\in I} Y_1^i, \ldots, \bigvee_{i\in I} Y_n^i) =$$
$$\bigvee_{i\in I} T_1(Y_1^i, \ldots Y_n^i) * \bigvee_{i\in I} T_2(Y_1^i, \ldots Y_n^i) = \quad \text{(by Lemma 3)}$$
$$\bigvee_{i\in I} (T_1(Y_1^i, \ldots Y_n^i) * T_2(Y_1^i, \ldots Y_n^i)) = \bigvee_{i\in I} T(Y_1^i, \ldots Y_n^i)$$

This general statement will enable us to prove the existence of a maximal solution in the general case of fuzzy relational equations and/or inequations, instead of just particular cases included in [19] and [20].

Theorem 1. *Let L be a complete meet-continuous lattice, $Z_1, \ldots, Z_n \in L^{A\times A}$ and $T_1(X_1, \ldots, X_n) = T_2(X_1, \ldots, X_n)$ an equation over $L^{A\times A}$. If $(Y_1, \cdots, Y_n)$ is a solution to that equation, such that $(Y_1, \cdots, Y_n) \leq (Z_1, \cdots, Z_n)$, then there exists a maximal solution - related to the componentwise order in $(L^{A\times A})^n$ - to the above equation contained in $(Z_1, \cdots, Z_n)$. If the existing maximal solution is unique, it is also the greatest solution.*

Proof. Let $\{(Y_1^i, \ldots Y_n^i) \mid i \in I\}$ be a chain in $(L^{A\times A})^n$ which belongs to the set of the solutions to the equation $T_1(X_1, \ldots, X_n) = T_2(X_1, \ldots, X_n)$ contained in $(Z_1, \cdots, Z_n)$. We have:
$T_1(Y_1^i, \ldots, Y_n^i) = T_2(Y_1^i, \ldots, Y_n^i)$ for all $i \in I$, thus:
$\bigvee_{i\in I} T_1(Y_1^i, \ldots, Y_n^i) = \bigvee_{i\in I} T_2(Y_1^i, \ldots, Y_n^i)$.

By Proposition 4, we have

$$T_1(\bigvee_{i\in I} Y_1^i, \ldots, \bigvee_{i\in I} Y_n^i) = \bigvee_{i\in I} T_1(Y_1^i, \ldots Y_n^i) \text{ and}$$
$$T_2(\bigvee_{i\in I} Y_1^i, \ldots, \bigvee_{i\in I} Y_n^i) = \bigvee_{i\in I} T_2(Y_1^i, \ldots Y_n^i);$$

therefore $T_1(\bigvee_{i\in I} Y_1^i, \ldots, \bigvee_{i\in I} Y_n^i) = T_2(\bigvee_{i\in I} Y_1^i, \ldots, \bigvee_{i\in I} Y_n^i)$.

Thus, the chain has its union as an upper bound in the set of solutions to the equation $T_1 = T_2$ contained in $(Z_1, \cdots, Z_n)$, since its union is also a solution to the equation contained

in $(Z_1, \cdots, Z_n)$. Since this holds for any chain of solutions contained in $(Z_1, \cdots, Z_n)$, applying Zorn's lemma, we get the assertion of the theorem.

Corollary 1. *If L is a complete meet-continuous lattice, every solvable relational equation over $L^{A\times A}$ of the type $T_1(X_1, \ldots, X_n) = T_2(X_1, \ldots, X_n)$ has a maximal solution.*

Analogously we prove that a maximal solution exists for any inequation as well. We use the term inequation to denote a constraint of the form $T_1 \leq T_2$ whose solutions are sought, in analogy with equations.

Theorem 2. *If L is complete and meet-continuous, $T_1(X_1, \ldots, X_n) \leq T_2(X_1, \ldots, X_n)$ is an inequation over $L^{A\times A}$ and $(Y_1, \cdots, Y_n)$ its solution, such that $(Y_1, \cdots, Y_n) \leq (Z_1, \cdots, Z_n)$, there exists a maximal solution - relative to the componentwise order in $(L^{A\times A})^n$ - to the above inequation contained in $(Z_1, \cdots, Z_n)$. If the existing maximal solution is unique, it is also the greatest solution.*

The following simple example demonstrates that in case L is not a meet-continuous lattice, we do not need to have a maximal solution to equations.

Example 1. Let L be a lattice in Figure 2. Let $A = \{d\}$ and let relations R and S be defined by: $R(d, d) = a$ and $S(d, d) = 0$. Now, we consider the equation $R \circ X = S$. We see that the relations B_i, defined by $B_i(d, d) = b_i$ $(i \geq 0)$ are solutions to the equation. However, there is no maximal solution contained in the relation $R_1(d, d) = 1$, since $\bigvee_{i\geq 0} B_i = R_1$ and R_1 is not a solution of the equation.

In the same way, we prove that any system of L-valued relational equations and inequations has a maximal solution as well. Moreover, the following holds for any set of equations and inequations (possibly infinite):

Theorem 3. *If L is complete, meet-continuous, $\{T_1^i(X_1, \ldots, X_n) = T_2^i(X_1^i, \ldots, X_n) \mid i \in I\} \cup \{S_1^j(X_1, \ldots, X_n) \leq S_2^j(X_1^i, \ldots, X_n) \mid j \in J\}$ a set of equations and inequations over $L^{A\times A}$, and $(Y_1, \cdots, Y_n)$ a solution to all of them, such that $(Y_1, \cdots, Y_n) \leq (Z_1, \cdots, Z_n)$, then there exists a maximal n-tuple of fuzzy relations - relative to the componentwise order in $(L^{A\times A})^n$, contained in $(Z_1, \cdots, Z_n)$, such that it is a solution to all of the above equations and inequations. If the existing maximal relation is unique, it is the greatest such relation.*

The results we got generalize those from [19], where it was proved that there exists a maximal solution to each of the inequations: $P \circ Q \leq R$, $Q \circ P \leq R$, $P \circ Q \leq Q$, $Q \circ P \leq Q$, and $Q \circ Q \leq Q$, contained in a given fuzzy relation, where Q is the unknown. It was also proved that there exists a maximal solution to the inequation $P \circ Q \leq R$, where both P and Q are unknown. All these results are corollaries of Theorem 2 we have proved here. Moreover, when the fuzzy set equations are concerned, the composition-continuity we have proved here enables us to prove the existence of a maximal solution to the equations $P \circ Q = Q$, $Q \circ P = Q$, and $Q \circ Q = Q$, where Q is the unknown, or the existence of a maximal solution to the equation $P \circ Q = R$, where both P and Q are unknown. All these are corollaries of Theorem 1. In [19], the existence of a maximal solution was only proved for the corresponding inequations.

In [19], also the algorithms were developed for finding some closures of a fuzzy relation: the closure under the composition with a given fuzzy relation, and the transitive closure. These algorithms follow from a more general algorithm, introduced and proved in the following theorem.

Theorem 4. *Let L be a complete, meet-continuous lattice, $T(X)$ a term and $R \in L^{A\times A}$. Let a sequence of fuzzy relations be defined recursively:*

$R_0 = R$
$R_{n+1} = T(R_n) \vee R_n$.

Then $\bigvee_{n\in\mathbb{N}} R_n$ is the least solution to the inequation $T(X) \leq X$ containing R.

Proof. The defined sequence is also a chain, since $R_n \subseteq T(R_n) \vee R_n = R_{n+1}$. Hence, $T(\bigvee_{n\in\mathbb{N}} R_n) = \bigvee_{n\in\mathbb{N}} T(R_n)$ follows from Proposition 4.

Therefore, $T(\bigvee_{n\in\mathbb{N}} R_n) = \bigvee_{n\in\mathbb{N}} T(R_n) \leq \bigvee_{n\in\mathbb{N}} R_{n+1} = \bigvee_{n\in\mathbb{N}} R_n$, thus $\bigvee_{n\in\mathbb{N}} R_n$ is a solution to the inequation $T(X) \leq X$.

On the other hand, any solution S to $T(X) \leq X$ containing R also contains R_n, for all $n \in \mathbb{N}_0$. We prove this by induction:

(i) $R_0 = R \leq S$ by assumption

(ii) $R_k \leq S$ implies $T(R_k) \leq T(S) \leq S$ (since T is order-preserving and S is a solution), thus $R_{k+1} = T(R_k) \vee R_k \leq S$, and the induction step holds.

Thus any solution to $T(X) \leq X$ containing R contains $\bigvee_{n\in\mathbb{N}} R_n$, thus $\bigvee_{n\in\mathbb{N}} R_n$ is the least solution to $T(X) \leq X$. □

Taking $T(X) = X \circ P$ (or $T(X) = P \circ X$), we get an algorithm for finding the closure of a fuzzy relation under right (or left) composition with P, namely the one which was proved in [19].

Taking $T(X) = X \circ X$, we get an algorithm for finding the transitive closure of a given fuzzy relation, which was proved in [19].

Example 2. $L = (\mathbb{N} \vee \{\infty\})^{\mathbb{N}\vee\{\infty\}}$, where $\mathbb{N} \vee \{\infty\}$ is a lattice we get by adding ∞ to the poset $(\mathbb{N}, \leq)$, such that ∞ is the greatest element.

$A = \mathbb{N}$
$P, Q, R, S \in L^{A\times A}$

$$P(i,j)(n) = \begin{cases} n+j & i \geq j-1 \\ n & i < j-1 \end{cases}$$

$S(i,j)(n) = n + i - j + 1$

$$Q(i,j)(n) = \begin{cases} n+i & j = 1 \\ n+i+j & 1 < j \end{cases}$$

Here we take $\infty + k = \infty$, where $k \in \mathbb{R}$, so that in all the given functions $P(i,j)$, $Q(i,j)$ and $S(i,j)$, (for all $(i,j) \in \mathbb{N} \times \mathbb{N}$) ∞ are mapped to itself.

We construct the least solution to the inequation $P \circ X \wedge Q \leq X$ containing S, following the algorithm in Theorem 4:

$R_0 = S$

$$R_1(i,j)(n) = \begin{cases} n+i & j=1 \\ max\{n, n+i-j+2\} & j>1 \end{cases}$$

$$R_2(i,j)(n) = \begin{cases} n+i & j=1 \\ n+i+1 & j=2 \\ max\{n, n+i-j+3\} & j>2 \end{cases}$$

By induction on $k \in \mathbb{N}$ we prove:

$$R_k(i,j)(n) = \begin{cases} n+i & j=1 \\ n+i+1 & 1<j\leq k \\ max\{n, n+i-j+k+1\} & j>k \end{cases}$$

Now we get the least solution:

$$R = \bigcup_{k\in\mathbb{N}} R_k$$

We can check that $R(i,j)(n) = \begin{cases} n+i & j=1 \\ n+i+1 & j>1 \end{cases}$

We may further generalize Theorem 4, for a system of fuzzy inequations, to get an algorithm for the construction of the least solution to systems of relational inequations:

Theorem 5. *Let $T_1(X_1, X_2, \ldots, X_n), T_2(X_1, X_2, \ldots, X_n), \ldots, T_n(X_1, X_2, \ldots, X_n)$ be terms and $(R_1, R_2, \ldots, R_n)$ an n-tuple of fuzzy relations. If, by induction, we define the following sequence of n-tuples of fuzzy relations:*

$$(Y_1^0, Y_2^0, \ldots, Y_n^0) = (R_1, R_2, \ldots, R_n)$$
$$(Y_1^{k+1}, \ldots, Y_n^{k+1}) = (T_1(Y_1^k, \ldots, Y_n^k) \vee Y_1^k, \ldots, T_n(Y_1^k, \ldots, Y_n^k) \vee Y_n^k),$$

then, $(\bigvee_{i\in\mathbb{N}_0} Y_1^i, \ldots, \bigvee_{i\in\mathbb{N}_0} Y_n^i)$ is the least solution to the system

$$T_1(X_1, X_2, \ldots, X_n) \leq X_1$$
$$T_2(X_1, X_2, \ldots, X_n) \leq X_2$$
$$\ldots$$
$$T_n(X_1, X_2, \ldots, X_n) \leq X_n$$

containing $(R_1, R_2, \ldots, R_n)$.

Proof. $(\bigvee_{i\in\mathbb{N}_0} Y_1^i, \ldots, \bigvee_{i\in\mathbb{N}_0} Y_n^i)$ is a chain and it contains $(R_1, R_2, \ldots, R_n)$ by definition.

We have that for $i \in \{1, \ldots, n\}$:

$$T_i(Y_1^k, \ldots, Y_n^k) \leq Y_i^{k+1}$$

$$\bigvee_{k\in\mathbb{N}_0} T_i(Y_1^k, \ldots, Y_n^k) \leq \bigvee_{k\in\mathbb{N}_0} Y_i^{k+1} = \bigvee_{k\in\mathbb{N}_0} Y_i^k$$

$$T_i(\bigvee_{k\in\mathbb{N}_0} Y_1^i, \ldots, \bigvee_{k\in\mathbb{N}_0} Y_n^i) = \bigvee_{k\in\mathbb{N}_0} T_i(Y_1^k, \ldots, Y_n^k) \leq \bigvee_{k\in\mathbb{N}_0} Y_i^k$$

Thus, $(\bigvee_{k\in\mathbb{N}_0} Y_1^k, \ldots, \bigvee_{k\in\mathbb{N}_0} Y_n^k)$ is a solution to the above system. To prove that it is the least solution to the system we take an arbitrary solution $(Z_1, \ldots, Z_n)$ that is greater or

equal to $(R_1, \dots, R_n)$, and prove, by induction that (for $i \in \{1, 2, \dots, n\}$), Z_i contains Y_i^k, for all $k \in \mathbb{N}_0$.

$Y_i^0 = R_i = Z_i$

If $Y_i^k \leq Z_i$ for all $i \in \{1, 2, \dots, n\}$, then:

$Y_i^{k+1} = T_i(Y_1^k, \dots Y_n^k) \vee Y_k^i \leq T_i(Z_1, \dots Z_n) \vee Y_k^i \leq Z_i$ (since $(Z_1, \dots Z_n)$ is a solution and $Y_k^i \leq Z_i$).

Now, Z_i is greater or equal to $\bigvee_{k\in\mathbb{N}_0} Y_i^k$, for $i \in \{1, 2, \dots, n\}$, and $(Z_1, \dots, Z_n)$ is greater or equal to $(\bigvee_{k\in\mathbb{N}_0} Y_1^k, \dots, \bigvee_{k\in\mathbb{N}_0} Y_n^k)$. □

In the following theorem, we introduce an algorithm leading to the least solution of a wide class of systems of fuzzy relational equations.

Theorem 6. *Let $X = \{X_1, \dots, X_n\}$ be a set of variables, and $T_1, T_2, \dots T_n$ terms whose variables belong to X and the constants to the set of L-valued relations $L^{A\times A}$, where L is a complete, meet-continuous lattice. We define the sequence $(Y_1^k, \dots, Y_n^k)$ inductively:*

$Y_i^0 = 0$, *for all* $i \in \{1, 2, \dots, n\}$;
$Y_i^{k+1} = T_i(Y_1^k, Y_2^k, \dots, Y_n^k)$ *for all* $i \in \{1, 2, \dots, n\}$.

Now $(\bigvee_{k\in\mathbb{N}} Y_1^k, \bigvee_{k\in\mathbb{N}} Y_2^k, \dots, \bigvee_{k\in\mathbb{N}} Y_n^k)$ *is the least solution to the system:*

$X_1 = T_1(X_1, X_2, \dots, X_n)$
$X_2 = T_2(X_1, X_2, \dots, X_n)$
....
$X_n = T_n(X_1, X_2, \dots, X_n)$

Proof. By induction, we prove that $(Y_1^k, \dots, Y_n^k) \leq (Y_1^{k+1}, \dots, Y_n^{k+1})$, for all $k \in \mathbb{N}$:

$(Y_1^0, \dots, Y_n^0) = (0, 0, \dots, 0) \leq (Y_1^1, \dots, Y_n^1)$

If $(Y_1^k, \dots, Y_n^k) \leq (Y_1^{k+1}, \dots, Y_n^{k+1})$, then for all $i \in \{1, \dots, n\}$:

$Y_i^{k+1} = T_i(Y_1^k, \dots, Y_n^k) \leq T_i(Y_1^{k+1}, \dots, Y_n^{k+1}) = Y_i^{k+2}$,

Since $\{(Y_1^k, \dots, Y_n^k) \mid k \in \mathbb{N}_0\}$ is a chain, we may apply Proposition 4:

$$T_i(\bigvee_{k\in\mathbb{N}_0} Y_1^k, \bigvee_{k\in\mathbb{N}_0} Y_2^k, \dots, \bigvee_{k\in\mathbb{N}_0} Y_n^k) = \bigvee_{k\in\mathbb{N}_0} T_i(Y_1^k, \dots, Y_n^k) = \bigvee_{k\in\mathbb{N}_0} Y_i^{k+1} = \bigvee_{k\in\mathbb{N}_0} Y_i^k$$

Thus, $(\bigvee_{k\in\mathbb{N}} Y_1^k, \bigvee_{k\in\mathbb{N}} Y_2^k, \dots, \bigvee_{k\in\mathbb{N}} Y_n^k)$ is a solution.

On the other hand, if $(S_1, \dots, S_n)$ is a solution, then we prove by induction that $(Y_1^k, \dots, Y_n^k) \leq (S_1, \dots, S_n)$:

$(Y_1^0, \dots, Y_n^0) = (0, 0, \dots, 0) \leq (S_1, \dots, S_n)$.

If $(Y_1^k, \dots, Y_n^k) \leq (S_1, \dots, S_n)$, then by monotony of T_i we have:

$Y_i^{k+1} = T_i(Y_1^k, \dots, Y_n^k) \leq T_i(S_1, \dots, S_n) = S_i$, thus $(Y_1^{k+1}, \dots, Y_n^{k+1}) \leq (S_1, \dots, S_n)$

Thus, $(Y_1^k, \dots, Y_n^k) \leq (S_1, \dots, S_n)$ for all $k \in \mathbb{N}_0$ and, therefore:

$(\bigvee_{k\in\mathbb{N}_0} Y_1^k, \ldots, \bigvee_{k\in\mathbb{N}_0} Y_n^k) \leq (S_1, \ldots, S_n)$.

Since $(\bigvee_{k\in\mathbb{N}_0} Y_1^k, \ldots, \bigvee_{k\in\mathbb{N}_0} Y_n^k)$ is less than or equal to any solution to the system, it is the least solution.

Example 3. Let L be the lattice in Figure 3. It is meet-continuous, since it is algebraic. Namely, all the elements of the form a_k, where $k \in \mathbb{N}$ are compact: if $\{c_j \mid j \in J\}$ is a family of elements from L and $a_k \leq \bigvee_{j\in J} c_j$, there must be some element of the form a_k among c_j (else $\bigvee_{j\in J} c_j \leq b$); also, there must be some a_i or b_i with $i \geq k$ (or else $\bigvee_{j\in J} c_j \leq a_{k-1}$), or b, or 1; in the first case $a_k \leq a_i$; in the second case, $a_k \leq b_i \vee a_k$; in the third case $a_k \leq b \vee a_k$; in the forth case $a_k \leq 1$. Similarly, we prove that a is compact. It is easy to prove that elements of the form b_k are compact. As for b and 1, they are supremum of the compact elements. Thus L is algebraic and also meet-continuous.

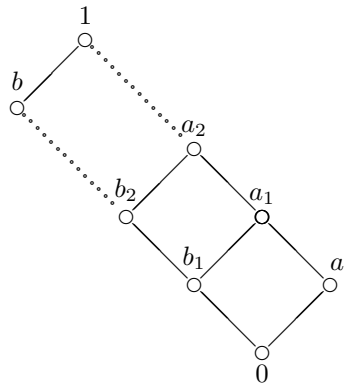


Fig. 3. The codomain lattice in Example 3

Let $A = N$. We define the following fuzzy relational equations on A, with L as the codomain lattice:

$$P(i,j) = b_{i+j}$$

$$Q(i,j) = \begin{cases} b & \text{if } j - i \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

$$R(i,j) = \begin{cases} 1 & \text{if } i - j \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

$$S(i,j) = a_{i+j}$$

Now we consider the following system of equations:

$$X = P \vee (Q \circ Y)$$
$$Y = S \vee (X \circ R)$$

We construct the least solution to the system, using the above algorithm:

$(X_0, Y_0) = (0, 0)$
$(X_1, Y_1) = (P, S)$
$X_2(i,j) = b_{i+j+1}$; $Y_2(i,j) = a_{i+j+1}$.

By induction on n, we prove that for $n \geq 2$ we have:

$X_n(i,j) = b_{i+j+n-1}$; $Y_n(i,j) = a_{i+j+n-1}$.

The basis for $n = 2$ we already have. If $X_k(i,j) = b_{i+j+k-1}$; $Y_k(i,j) = a_{i+j+k-1}$, since $X_{k+1} = P \vee (Q \circ Y_k)$ and $Y_{k+1} = S \vee (X_k \circ R)$, it is easy to see that:

$X_{k+1}(i,j) = b_{i+j+k}$; $Y_{k+1}(i,j) = a_{i+j+k}$.

Now, the least solution to the system is $(X', Y') = \bigvee_{k \in \mathbb{N}} (X_k, Y_k)$. Obviously, $X'(i,j) = b$ and $Y'(i,j) = 1$ for all $(i,j) \in \mathbb{N} \times \mathbb{N}$. Due to Theorem 5, we don't have to check that (X', Y') is a solution, or to prove that it is the least one.

5. Potential Real-World Applications

Although the present study is primarily theoretical, the obtained results have several potential applications. Fuzzy relational equations and inequations provide a flexible mathematical framework for modeling systems in which relationships between objects are imprecise, vague, or partially unknown. In such settings, relations are described by degrees of association taking values in a complete lattice or another ordered algebraic structure. This makes them suitable for modeling real-world problems characterized by uncertainty, incompleteness, or incomparable information.

The study of both equations and inequations is essential, as they capture different types of constraints: equations typically express consistency or equilibrium conditions, while inequations naturally arise when information is partial or when constraints are admissibility conditions.

One important area of application is decision support and multi-criteria decision making [2,12], where preferences and expert judgments are often expressed by fuzzy values. Consistency requirements between alternatives and criteria can be modeled using fuzzy relational equations, while inequations describe admissible or acceptable decisions. Similar formulations appear in recommendation systems and group decision processes.

Another application domain is knowledge representation and data analysis, particularly in connection with formal concept analysis [1,7]. Here, fuzzy relations describe graded associations between objects and attributes. Fuzzy relational equations arise when inferring missing data or determining closures, whereas inequations model approximate implications and soft constraints, which are especially relevant in the presence of noisy data.

Further applications can be found in control theory and systems modeling [5,15], where fuzzy relational equations are used to represent input–output relations and rule bases, while inequations describe admissible operating regions and robustness requirements. Additional applications include image processing, pattern recognition, and information retrieval, where similarity and relevance relations are inherently fuzzy. To illustrate the applicability of the framework, we consider the following simple example, inspired by models in fuzzy decision making (see, e.g., [2,12]).

Let $A = \{a_1, a_2\}$ be a set of alternatives (e.g., two treatment options or products). Suppose that this set is evaluated with respect to a set of criteria $B = \{b_1, b_2\}$ (e.g., effectiveness and cost). Let L be a finite lattice and let $R : A \times B \to L$ describe the degree to which each alternative satisfies each criterion, and $X : B \to L$ represents the importance (weights) of the criteria.

The overall evaluation $Y : A \to L$ of the alternatives can be modeled by the fuzzy relational equation $R \circ X = Y$ where $\circ$ denotes the standard (sup–inf) composition.

In practice, the desired evaluation Y may only be partially specified or constrained (e.g., one alternative should be at least as good as a given threshold). This leads naturally to a fuzzy relational inequation of the form $R \circ X \geq Y$ which describes all admissible weight vectors X consistent with the requirements. The problem then becomes to determine all solutions X, or in particular the maximal and least solutions. The theoretical results developed in this paper guarantee the existence of maximal solutions under appropriate conditions on the lattice L, and provide constructive procedures for obtaining the least solution.

This example illustrates how practical problems can be reduced to fuzzy relational equations and inequations. The main results of this paper ensure the existence of maximal solutions and provide terminating procedures for computing the least solution, thereby offering both theoretical guarantees and computational tools.

6. Conclusion

We generalized the notion of meet-continuity of a complete lattice and proved some lattice-theoretic properties. We proved the equivalence between meet-continuity of the codomain lattice and different kinds of composition-continuity in the lattice of L-valued relations.

The existence of a maximal solution to general L-valued relational equations and inequations and their systems was proved in the case L is a complete, meet-continuous lattice. Moreover, there exists a maximal L-valued relational equation that is a solution to all the equations from a given set, which may be infinite as well.

An algorithm is developed for the construction of the least solution of a wide class of equations and inequations, in the case of a complete, meet-continuous lattice.

In the future, we shall try to find algorithms for the construction of a maximal solution or the whole set of maximal solutions, in the case of a complete, meet-continuous codomain lattice.

Acknowledgement. This research was supported by the Science Fund of the Republic of Serbia, Grant no 6565, Advanced Techniques of Mathematical Aggregation and Approximative Equations Solving in Digital Operational Research-AT-MATADOR. The authors also gratefully acknowledge the financial support of the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (Grants No. 451-03-34/2026-03/200116 & 451-03-33/2026-03/ 200125 & 451-03-34/2026-03/ 200125).

References

1. Alcalde, C., Burusco, A., Díaz-Moreno, J.C., Medina, J.: Fuzzy concept lattices and fuzzy relation equations in the retrieval processing of images and signals. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 25(Suppl. 1), 99–120 (2017)
2. Aliev, R., Guirimov, B., Huseynov, O., Aliyev, R.: Z-relation equation-based decision making. Expert Systems with Applications 184, 115387 (2021)

3. Cornejo, M., Lobo, D., Medina, J., De Baets, B.: Bipolar equations on complete distributive symmetric residuated lattices: The case of a join-irreducible right-hand side. Fuzzy Sets and Systems 442, 92–108 (2022)
4. Davey, B.A., Priestley, H.A.: Introduction to Lattices and Order. Cambridge University Press, 2 edn. (2002)
5. De Baets, B.: Analytical solution methods for fuzzy relational equations. In: Fundamentals of Fuzzy Sets. The Handbooks of Fuzzy Sets Series, Springer, Boston (2000)
6. Díaz-Moreno, J., Medina, J., Turunen, E.: Minimal solutions of general fuzzy relation equations on linear carriers, an algebraic characterization. Fuzzy Sets and Systems 311, 112–123 (2017)
7. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer (2024)
8. Gierz, G., Hofmann, K.H., Keimel, K., Lawson, J.D., Mislove, M., Scott, D.S.: Continuous Lattices and Domains, vol. 93. Cambridge University Press (2003)
9. Gierz, G., Hofmann, K.H., Keimel, K., Lawson, J.D., Mislove, M., Scott, D.S.: A Compendium of Continuous Lattices. Springer (2012)
10. Goguen, J.: L-fuzzy sets. Journal of Mathematical Analysis and Applications 18, 145–174 (1967)
11. Hohle, U.: Upper semicontinuous l-fuzzy sets and applications. Fuzzy Sets and Systems 27, 217–237 (1987)
12. Kahraman, C. (ed.): Fuzzy Multi-Criteria Decision Making: Theory and Applications with Recent Developments. Springer (2008)
13. Keimel, K., Lawson, J.: Continuous and completely distributive lattices. In: Continuous and Completely Distributive Lattices, pp. 5–53. Springer, Cham (2014)
14. Pavelka, J.: On fuzzy logic ii. enriched residuated lattices and semantics of propositional calculi. Mathematical Logic Quarterly 25, 119–134 (1979)
15. Peeva, K., Kyosev, Y.: Fuzzy Relational Calculus. World Scientific (2005)
16. Rachunek, J., Slezak, V.: Bounded dually residuated lattice ordered monoids as a generalization of fuzzy structures. Mathematica Slovaca 56(2), 223–233 (2006)
17. Rodabaugh, S.: Powerset operator foundations for poslat-fuzzy set theories. Fuzzy Sets and Systems 63, 129–144 (1994)
18. Sanchez, E.: Resolution of composite fuzzy relation equations. Information and Control 30, 38–48 (1976)
19. Stepanovic, V.: Fuzzy set inequations and equations with a meet-continuous codomain lattice. Journal of Intelligent and Fuzzy Systems 34(6), 4009–4021 (2018)
20. Stepanovic, V.: Two sufficient conditions for having a maximal solution to fuzzy relational equations. In: Computational Intelligence and Mathematics for Tackling Complex Problems 6, pp. 19–26. Springer, Cham (2025)
21. Stepanovic, V., Tepavčević, A.: Fuzzy sets (in)equations with a complete codomain lattice. Kybernetika 58, 145–162 (2022)
22. Stepanovic, V., Tepavčević, A.: A note on the solutions to lattice-valued relational equations and inequations. Acta Polytechnica Hungarica 22(12), 27–46 (2025)
23. Tarski, A.: On the calculus of relations. The Journal of Symbolic Logic 6(3), 73–89 (1941)
24. Tarski, A.: A lattice-theoretical fix point theorem and its applications. Pacific Journal of Mathematics 5, 285–309 (1955)
25. Šešelja, B., Tepavčević, A.: Weak Congruences in Universal Algebra. Mathematical Institute of the Serbian Academy of Sciences and Arts, Novi Sad (2001)

Vanja Stepanović is an Assistant Professor at University of Belgrade, Faculty of Agriculture, Department of Agricultural Engineering. Her primary research areas are lattice theory and fuzzy set theory. She has also published works in the social sciences and the

humanities. She is a member of the project team for the project Advanced Techniques of Mathematical Aggregation and Approximative Equations Solving in Digital Operational Research (AT-MATADOR), funded by the Science Fund of the Republic of Serbia.

Andreja Tepavčević is a principal research fellow at the Mathematical Institute of the Serbian Academy of Sciences and Arts in Belgrade, Serbia, and a full professor at the Department of Mathematics and Informatics, Faculty of Sciences, University of Novi Sad. Her main research is in lattice theory and theory of ordered sets; fuzzy sets theory and applications, with about 150 refereed publications in journals, chapters in monographs, international conferences, one research monograph, and 10 university textbooks. She is an associate editor and a member of the editorial board of several journals, the coordinator of several research projects (national and international), and the head or a member of organizing and program committees of several conferences. Currently, she is the chairperson of the Artificial Intelligence Seminar at MISANU and the head of the Computer Science sector at MISANU.

Received: December 14, 2025; Accepted: June 23, 2026.

Computer Science and Information Systems 23(4):1563–1596 https://doi.org/10.2298/CSIS260523042S

Paper Type: Research Paper

Encoding, Linking, Retrieving: A Methodological Framework for Knowledge-Enriched Interview Corpora*

Ranka Stanković[1], Tamara Vučenović[2], Milica Ikonić Nešić[3], and Biljana Rujević[1]

[1] University of Belgrade, Faculty of Mining and Geology,
Djusina 7, Belgrade, Serbia
ranka.stankovic@rgf.bg.ac.rs (corresponding author), biljana.rujevic@rgf.bg.ac.rs
[2] University Metropolitan, Faculty of Management
Tadeuša Košćuška 63, Belgrade, Serbia
tamara.vucenovic@metropolitan.ac.rs
[3] University of Belgrade, Faculty of Philology
Studenski Trg 2, Belgrade, Serbia
milica.ikonic.nesic@fil.bg.ac.rs

Sažetak This paper presents an AI-driven pipeline for transforming the interviews from "Digitalne Ikone 20+" book from unstructured transcripts into a structured, semantically enriched, and queryable knowledge resource. The raw text was first converted into XML-TEI format, with explicit structural markup of interview boundaries, speaker turns, paragraphs, temporal metadata, and topics. This encoding established logical segmentation and enabled targeted queries, such as retrieving content by speaker or thematic segment. An NLP and textometric analysis was conducted using the TXM tool and JeRTeh resources, followed by automatic Named Entity Recognition (NER) using models from the TESLA project. Key entity types were identified and embedded into the TEI structure. In the subsequent Named Entity Linking (NEL) stage, entities were disambiguated and connected to Wikidata identifiers, enriching the corpus with external knowledge graph references. Missing entities were added to Wikidata, contributing new structured knowledge. The resulting resource allows researchers, students, and the public to explore cultural heritage interviews through intelligent querying, automated dataset generation, and knowledge graph integration. The pipeline offers a replicable methodology for converting oral archives into AI-accessible knowledge bases.

Keywords: Corpus, Named Entity Recognition, Named Entity Linking, Information Extraction, Textometry, Knowledge Graphs.

1. Introduction

This paper builds upon our preliminary findings first introduced in the AI Conference 2025 abstract, where the initial implementation of the "Digitalne Ikone" AI pipeline

* This paper is an extended version of an abstract "From Unstructured Interviews to Queryable Knowledge: An AI Pipeline for the "Digitalne Ikone" Corpus using TEI, NER, NEL, and RAG", published in Artificial Intelligence Conference Book of Abstracts. https://www.mi.sanu.ac.rs/

showed promising results, and now presents a fully extended study with deeper analysis, methodological enhancements, and a broader evaluation [32].

Large archives of interviews and broadcast transcripts remain significantly underutilized in Natural Language Processing (NLP), primarily due to their inherently unstructured nature, diverse formats, and the high cost of performing reliable semantic queries over them. The challenge is even more pronounced in low- and less-resourced languages, where reusable processing pipelines and datasets that enable grounded question answering (QA) are still scarce or entirely absent.

In this work, our motivation is twofold. First, we aim to transform the long-standing Serbian interview series *Digitalne Ikone 20+* (Digital Icons 20+) [38] into a structured, semantically enriched resource that facilitates reproducible research, exploratory analysis, and evidence-based question answering. Second, we seek to establish a standards-compliant methodological framework, based on widely accepted practices such as XML-TEI encoding (Text Encoding Initiative), lexical and semantic annotation, text analysis, knowledge management and information retrieval, that can be adapted and adopted by other oral-history archives, cultural heritage collections, or media corpora across different languages. By doing so, we aim not only to improve accessibility and interoperability, but also to highlight the broader potential of AI methodologies for unlocking knowledge embedded in complex narrative datasets.

The radio broadcast *Digital Icons*, curated by Dr. Tamara Vučenović, has been continuously aired on *Radio Belgrade 2* since 2002, with more than 1,000 episodes to date. As an educational program devoted to digital technologies and their societal impact, it holds a unique position within the Serbian media landscape, not only among radio programs but also across television formats. Its long-term significance stems from both its thematic coherence and its informative public role: the program invites domain experts to discuss topics spanning computer science, the evolution of the information society, digital inclusion and the protection of vulnerable groups, ethical issues of social media, and the benefits and risks associated with technological development.

The diversity of guests has been a defining feature of the broadcast. Participants include senior officials and policy advisers, researchers and academic representatives, professionals from domestic and international IT organizations, civil society groups, media outlets, and specialized associations. Moreover, the program has hosted numerous international interlocutors, contributing to the global exchange of ideas about digital culture and technology.

Importantly, the book *Digitalne Ikone 20+* (Digital Icons 20+), derived from the transcripts of selected broadcast interviews, provided the foundation for building the *DigIko* [32], as first version, and *DigIko2 – Digital Icons 20+ Interview Corpus* used in this research. It now represents a valuable bridge between journalistic content, public discourse, and computational linguistics, enabling further exploration of societal narratives through AI-based methods.

The integration of TEI with NER (Named Entity Recognition) and NEL (Named Entity Linking) is motivated by the need to provide layered semantic enrichment that serves as a reliable foundation for Retrieval-Augmented Generation (RAG). By grounding AI-generated answers in verifiable textual evidence, RAG allows retrieved passages, rather than free-form generative output, to guide the reasoning process and mitigate hallucinati-

ons. In this setting, NEL based on Wikidata[4] identifiers (QIDs) [36] plays a crucial role, as it enables precise multilingual entity resolution and control during both retrieval and prompting stages. Each entity mention is not only classified but also linked to an external canonical reference, contributing to interoperability and long-term reusability.

Furthermore, the consistent use of TEI IDs and Wikidata QIDs attached to each extracted claim ensures transparency of provenance, supports reproducibility and peer-review practices, and facilitates integration with existing or emerging knowledge graphs. This standards-driven combination of TEI, NER, and NEL therefore establishes a structured backbone for explainable AI workflows, allowing interview-based corpora to be queried, validated, and semantically explored using modern NLP (Natural Language Processing) techniques.

The contributions of this paper can be summarized as follows. First, we present a standards-driven AI pipeline that integrates multiple preprocessing tasks, ranging from structural encoding to semantic annotation, thereby transforming raw interview transcripts into a semantically linked and structurally queryable resource. This approach ensures interoperability, transparency, and reproducibility across all stages of processing. Second, we demonstrate a grounded question–answering (QA) generation over the *DigIko2* interview corpus, combining evidence retrieval with LLM-based generation to support scholarly inquiry and exploratory analysis.

The workflow incorporates TEI structural annotation, TXM/Jerteh preprocessing, and NER/NEL models from the TESLA[5] project, resulting in TEI-embedded entity tags resolvable to Wikidata identifiers. We argue that such integration substantially improves accessibility, enables semantic search and automated dataset generation, and provides a replicable pathway for constructing knowledge-centric interview corpora for Serbian and other less-resourced languages.

In addition, the repository with code and sample data will be made openly available, while the complete corpus will be searchable through CQL (Corpus Query Language) and SPARQL endpoints for the linked-data version. Together, these components establish a foundation for transparent peer review, empirical validation, and future extensions toward knowledge graph construction and more advanced RAG architectures.

Beyond conventional transformation, the proposed approach enables entity- and speaker-faceted retrieval, which opens possibilities for opinion tracking, stance profiling, and temporal comparison, analytical dimensions that are typically unavailable in plain text processing. By implementing a sequential transformation from raw transcripts to a semantically enriched and queryable corpus, we provide a foundation for evidence-based analysis that goes beyond narrative abstraction.

The discussion of prior studies and relevant methodological frameworks is provided in Section 2, which outlines existing approaches to TEI-based corpus construction, NER/NEL integration, textometry and retrieval-augmented methods in cultural and interview-based datasets. The complete methodology for corpus preparation and exploitation is detailed in Section 3, followed by the presentation of the corpus editions and data formats in Section 4, along with the analytical procedures, practical applications. The evaluation of results and the discussion of current limitations with directions for future development is provided in Section 5. Together, these sections outline a replicable

[4] https://www.wikidata.org/

[5] TESLA – Text Embeddings – Serbian Language Applications, https://tesla.rgf.bg.ac.rs/

workflow for building and systematically interrogating interview-based corpora through semantically grounded AI methods.

2. Related Work

In this section, we provide an overview of related research across several interconnected areas, all relevant to the construction and exploitation of the *DigIko2* corpus. First, we discuss prior studies that model interview transcripts and specifically those that use TEI as a structural framework for capturing interview transcripts and conversational data. Next, we review approaches to semantic enrichment through NER with particular emphasis on linking to external knowledge bases such as Wikidata. We conclude this section with work related to textometry, as a computational method for analyzing textual corpora that combines statistical modeling with linguistic structure to uncover patterns, distributions, and relationships within texts that are not easily visible through manual reading.

2.1. Modeling of Interview Corpora

The work presented in [26] introduces *InTEIrviews*, an XML–TEI–conformant ODD[6] specifically tailored for qualitative interviews, with the aim of promoting openness, reusability, and standardized annotation practices. An XML–TEI–conformant ODD enables the selection of relevant TEI P5 modules and the declaration of project-specific constraints while remaining compatible with the TEI element and class system. From a single ODD specification, it is possible to generate both validation schemas and human-readable documentation, ensuring both methodological transparency and reproducibility.

The ODD typically includes a `schemaSpec` (for module selection), `elementSpec` and `classSpec` (for content models and attributes), as well as prose guidance on encoding decisions and metadata requirements. The model builds upon existing TEI modules, particularly *Transcription of Speech* and *Language Corpora*, and draws a clear distinction between macrostructure (e.g., headered metadata on participants and recording conditions, turns encoded via `u who=`, and researcher notes within `standOff`) and microstructure (e.g., fine-grained linguistic and discourse features, including named entities). Both structures and the following TEI elements are presented in Sample 1.1.

Ethical and privacy considerations are also addressed by recommending the use of `gap` (in combination with `@resp`, `@unit`, and `@quantity`) for redacting sensitive content. Controlled vocabularies, such as annotating speakers with `@ana`[7] on `u`[8] are encouraged to support cross-interview comparability. Overall, the *InTEIrviews* model demonstrates how interview transcripts can be standardized through TEI while still preserving project-specific variation, enabling semantic analysis, corpus sharing, and FAIR-aligned reuse.

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
 <teiHeader>
  <fileDesc>
   <titleStmt> <title>TEI interview</title> </titleStmt>
```

[6] ODD ("One Document Does-it-all") is the TEI XML-based specification language for defining, constraining, and documenting TEI customizations.

[7] analysis refers to elements that contain interpretations of the annotated element.

[8] utterance: contains a stretch of speech usually preceded and followed by silence or by a change of speaker.

```
  </fileDesc>
 </teiHeader>
 <!-- MACROSTRUCTURE -->
 <standOff>
  <listPerson>
   <person xml:id="SP1" gender="F" age="35" .../>
   <person xml:id="INT" role="interviewer" .../>
  </listPerson>
 </standOff>
 <text>
  <body>
   <!-- MICROSTRUCTURE -->
   <div type="interview" ...>
    <u who="#INT" xml:id="u1"> Can you describe life during that period? </u>
    <u who="#SP1" xml:id="u2" ...>It was very nice in <name type="LOC" ref="https:
        //www.wikidata.org/wiki/Q55630">Novi Sad</name> for three months.</u>
    <u who="#SP1" xml:id="u3">After that... it is hard to explain.</u>
   </div>
  </body>
 </text>
</TEI>
```

Sample 1.1. XML schema *InTEIrviews*

The DDHI Encoding Guidelines [3] developed for the Dartmouth Digital History Initiative (DDHI) provide practical guidance for TEI-based oral history encoding, with a strong emphasis on interview metadata and TEI headers. These guidelines focus on capturing descriptive information about interviewees, interviewers, recording circumstances, equipment, rights, and ethical considerations, offering clear examples and recommended practices for structuring metadata within `teiHeader`. In addition, they demonstrate techniques for annotating speaker turns, editorial notes, and research-related contextual information, making them particularly valuable for projects involving oral narratives and historical accounts. What distinguishes the DDHI approach is its balance between TEI compliance and practical usability: the guidelines include actionable examples, recommended element usage, and documentation that supports reproducibility and interoperability across oral history projects. This makes DDHI a relevant reference point for corpus preparation methodologies such as *DigIko2*, where interview metadata, participant roles, and recording context are essential for downstream tasks such as entity linking, semantic retrieval, and RAG-based question answering.

A valuable overview of transcript–media alignment methods is provided in [4], which analyzes three practical approaches for synchronizing interview audio/video with their corresponding transcripts: the OHMS (Oral History Metadata Synchronizer) system, TEI-based encoding workflows, and YouTube-based captioning practices. These methods are examined in terms of their ability to provide searchable, time-coded access to oral-history materials, enabling users to navigate interviews based on speaker turns, topics, or temporal segments. Of particular relevance to TEI-based corpus construction is the demonstration of how structured markup, especially when combined with timestamp anchoring that supports interoperability and facilitates links between transcript segments and media files. This approach not only enhances accessibility and user engagement but also enables multimodal analysis, making oral-history collections more suitable for scholarly interrogation, pedagogical use, and AI-based retrieval.

2.2. Text Encoding Initiative for Spoken Corpora

CLARIN-compliant web services provide an end-to-end processing workflow for *TEI/ISO 24624* transcripts of spoken language, with particular emphasis on legacy interview corpora [7]. Their framework demonstrates how linguistic tools and processing components can operate directly on *ISO/TEI* formats, thus ensuring semantic interoperability with established spoken-language tools while also granting access to a broader range of written-language NLP functionalities. A key application is showcased through the IDS Archive for Spoken German, where interview materials are curated and processed according to FAIR principles. The workflow involves complete digitization of archival recordings, conversion into structured and standardized formats (including *ISO/TEI* for transcripts and CMDI for metadata), precise alignment between audio and transcript segments, and successive layers of linguistic enrichment. Furthermore, the resulting corpora are integrated into dissemination platforms that support search, visualization, and processing services, thereby enabling sustainable reuse and scholarly exploration of oral-history datasets. This approach offers a compelling model for similar initiatives, particularly for interview-based collections that seek to combine long-term preservation with modern AI-based processing methods.

The service pipeline consists of several modular components designed to convert raw interview transcripts into linguistically enriched *ISO/TEI* representations. The workflow begins with *text2iso*, which parses plain-text transcripts and generates *ISO/TEI*-compliant structures including `u` for utterances, `timeline` for time anchoring, participant metadata in `<teiHeader>`, and mechanisms for overlap resolution. Subsequently, tokenization transforms the transcript into `w` (word) and `pc` (punctuation) elements, while preserving original anchors and lifting pauses into dedicated `pause` tags. Per-utterance language identification is performed using OpenNLP, with the detected code assigned to `@xml:lang` attributes. Language-specific normalization is handled through OrthoNormal, which writes normalized values to `@norm` within the `w` elements, allowing both orthographic variants and surface forms to be retained. Further enrichment involves POS-tagging and lemmatization via TreeTagger/TT4J with language-appropriate models, followed by pseudo-alignment between transcript segments and the corresponding audio stream. Finally, consistent `@xml:id` management ensures traceability across elements and supports downstream referencing, validation, and corpus querying. Overall, the pipeline illustrates a structured and extensible approach to preparing spoken-language corpora for both linguistic research and computational processing.

The Slovene reference spoken corpus *GOS 2* represents a substantial expansion of the original *GOS* corpus from 2011, integrating two major additional sources: *GOS Video-Lectures* (public academic speech) and the *Artur ASR* database (Automatic Speech Recognition). The corpus includes authentic spoken data that have been manually verified, and heterogeneous designs from the original datasets were reconciled through a unified metadata.

A revised *TEI XML* schema-based on the *CLARIN.SI* parameterization was implemented to ensure formal alignment and long-term interoperability. Transcription conventions were harmonized across datasets; however, certain segmentation and punctuation variations were deliberately preserved in order to stimulate community feedback and support alternative annotation practices. Linguistic enrichment was conducted using CLASSLA–Stanza, enhancing the corpus with token-level annotation and morphosyntactic in-

formation. The resulting resource is openly released in *TEI* format, with transcriptions publicly available and audio data accessible for the newer subsets. A dedicated web concordancer enables simple and advanced search capabilities, metadata-based filtering, and integration with other Slovene linguistic resources [35].

The *Parla-CLARIN* initiative provides a specialized *TEI* customization for modeling and exchanging parliamentary debate corpora, with a strong emphasis on interoperability and usability across multilingual research contexts. Its guidelines propose a consistent document architecture that includes hierarchical structures for sittings, debates, and individual speeches or utterances, each annotated with `@who` attributes to identify speakers unambiguously. The framework further encourages rich `teiHeader` metadata, covering information about participants, political affiliations, institutional roles, temporal information, and geographic context. In addition to textual representation, *Parla-CLARIN* supports links to audio/video recordings and references to legislative acts. The package provides TEI schemas, validation instructions, and illustrative examples, making it suitable both for corpus construction and integration with downstream NLP workflows.[9]

The British National Corpus (BNC) XML Edition represents one of the most influential large-scale implementations of TEI for spoken interaction. As a well-established resource, it showcases mature structural conventions for modeling dialogue and turn-taking using `u` elements, with speaker identities assigned through `@who` attributes that reference a dedicated speaker registry. At the token level, the BNC employs consistent markup practices, including `w` for words and `pc` for punctuation, enabling corpus-wide uniformity and compatibility with a wide range of linguistic tools. A comprehensive `teiHeader` encodes rich participant and recording metadata, covering demographic information, situational context, and technical details of the audio capture. Owing to its scale, stability, and longstanding use, the Spoken BNC provides methodological precedents for utterance segmentation, speaker normalization, metadata design, and reproducibility, many of which can be effectively transferred to the development of interview-based corpora such as *DigIko2*. Its structural conventions remain a reference point for both TEI-based encoding practices and best practices in spoken-language corpus construction.[10]

2.3. Named Entities in Serbian Text

Named Entity Recognition and Linking (NER+NEL) inherently involve the task of entity disambiguation, which is particularly challenging for morphologically and less resourced languages [20,9]. Historically, early research in NEL relied on rule-based approaches and handcrafted heuristics, as seen in systems such as SrpNER [20] and the disambiguation algorithms of [39]. Over time, these were replaced by supervised machine learning models, including SVMs, CRFs, and neural architectures, that demonstrated substantially higher accuracy in complex linking scenarios [14,8]. The availability of structured multilingual resources such as Wikipedia, DBpedia, and especially Wikidata [2] has been crucial in providing canonical knowledge representations for entity resolution.

A major shift occurred with the emergence of transformer-based architectures, notably `BERT` and a suite of domain- or task-adapted models, that significantly improved NEL performance by leveraging context-aware embeddings, cross-lingual transfer, and

[9] Parla-CLARIN `https://clarin-eric.github.io/parla-clarin/`

[10] BNC XML Reference Guide: `http://www.natcorp.ox.ac.uk/docs/URG.xml?ID=cdifsp`.

domain-specific adaptation [40,6]. Recent studies have also highlighted limitations of LLMs in NEL tasks within specialized domains. Recent work presented in [23] emphasized the importance of geographic bias evaluation in LLM outputs, demonstrating that the diversity and specificity of geographic references vary across models and may be influenced by prompt phrasing.

For Serbian, NEL development was supported by resources produced in this research, including the multilingual silver dataset DaMuEL [21] and the `SrpCNNeL` model for named entity recognition and linking [15]. The inflectional complexity of Serbian, characterized by different forms for three gender, seven grammatical cases, numbers, requires models to handle morphological variation. Two pretrained transformer models optimized for Serbian give the best results: 1) `TeslaXLM`,[11] a fine-tuned variant of *FacebookAI/xlm-roberta-large* with 561M parameters, that was trained on a 20-billion-token Serbian (Latin + Cyrillic) corpus and achieves high robustness across dialectal and orthographic variation [42], and 2) `Jerteh-355`,[12] based on *RoBERTa-large* and includes 355M parameters, trained from scratch on a 4-billion-token Serbian corpus, it excels in monolingual contextual representations [41]. These models provide the linguistic robustness necessary for entity extraction and disambiguation in highly inflected Serbian interview data, forming the basis for the NER+NEL pipeline presented in this work.

2.4. Textometry

The origins of textometry can be traced to mid-20th century France, notably through the pioneering work of Pierre Guiraud [10] and Charles Muller [24], who established early methods for linguistic statistics. Subsequent developments by Jean-Paul Benzécri [1] expanded the statistical foundations and influenced later implementations of textometric analysis. The methodological basis of modern textometry also draws significantly from the work of Ludovic Lebart and André Salem [22], who formulated techniques for extracting meaningful linguistic patterns from textual data.

Building upon these foundations, textometry has introduced new statistical models for identifying specific textual phenomena such as contextual word salience, linear organization and internal structure of texts, intertextual contrast, and lexical evolution over time (e.g., scope of usage, periods of lexical change). The results of a textometric analysis typically yield visually organized outputs, including hierarchical lists, maps, clusters, and reorganized corpora, offering a synthetic yet flexible view of the underlying textual structures.

Unlike traditional summarization or keyword extraction, textometry provides a close reading of the text through numerical indicators, enabling both global corpus-level perspectives and local contextual insights. This duality highlights its hermeneutic potential and demonstrates how large-scale textual analysis can reveal latent patterns that are difficult, if not impossible, to detect manually. Interpretation, however, remains dependent on expert knowledge and the quality of the tools used, indicating that textometry is best conceived as a methodology that generates data-driven signals requiring expert validation.

Several adjacent disciplines intersect with textometry. Information Retrieval (IR) shares an interest in processing large text collections, though its primary goal is document

[11] https://huggingface.co/te-sla/TeslaXLM
[12] https://huggingface.co/jerteh/Jerteh-355

and information retrieval rather than internal textual analysis. Latent Semantic Analysis (LSA) similarly employs mathematical models but is generally oriented toward cognitive and semantic tasks outside the scope of corpus-based linguistic inquiry. Text mining (TM) applies quantitative techniques to extract valuable insights; however, its focus is on information extraction, whereas textometry concentrates on structural and linguistic exploration of texts. Meanwhile, NLP relies on statistics to build systems for linguistic unit recognition and tagging; indeed, corpora may serve as a foundation for both NLP development and subsequent textometric exploration. Thus, these two fields are complementary and mutually reinforcing.

A major advancement in applied textometry is the emergence of software environments that simplify corpus exploration through integrated tools and graphical workflows. Among these, the *TXM platform* [11] stands out as an open-source environment widely used in the social sciences and digital humanities (history, literature, linguistics, sociology, geography, political science). TXM combines qualitative and quantitative text analysis, built upon CQP (Corpus Query Processor) browsing and the R statistical package, thereby enabling rich corpus navigation and statistical modeling.

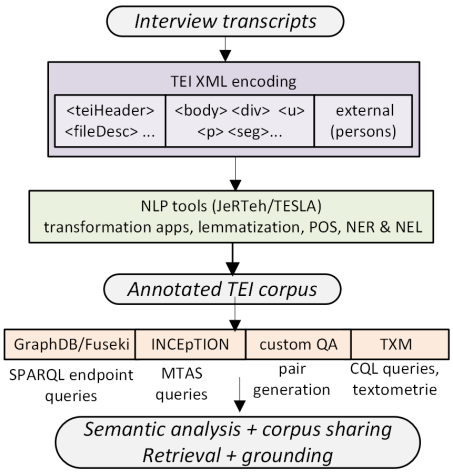
3. Methodology

This section outlines the toolchain, starting with TEI structure encoding, followed by annotation and textometry, and finally linked-data serialization and text generation. The diagram in Figure 1 illustrates the end-to-end processing pipeline of the *DigIko2* corpus, from raw interview transcripts through TEI-compliant XML encoding to parallel NLP enrichment via TXM textometry and the JeRTeh/TESLA annotation stack (lemmatization, NER, NEL). The enriched corpus feeds three output channels: a linked-data serialization in NIF exported to an Apache Jena Fuseki SPARQL endpoint[13] and GraphDB[14], a CQL-queryable INCEpTION [18] annotation layer, and a QA-pair generation module powered by `GPT-4.1`/`GPT-5.1` that produces a RAG-ready resource.

3.1. TEI for Spoken Interactions

Chapter 8 of the TEI P5 Guidelines, entitled *"Transcriptions of Speech"* [34], provides the core specification for representing spoken language and interviews in TEI-compliant XML. It conceptualises spoken discourse in terms of utterances (`u`) that are embedded within a standard TEI text hierarchy and anchored to speakers via the `@who` attribute, which may reference participants defined in structured registries such as `listPerson`. The chapter further outlines recommendations for encoding pauses, vocalisations, non-lexical phenomena, kinesic events, overlapping speech, and subdivisions of discourse. In addition, it specifies how alignment, segmentation and linking mechanisms can be employed to handle non-hierarchical annotations and time-based structures, thereby enabling interoperable and analysable transcripts suitable for digital preservation and reuse. Documentation of transcription conventions and recording conditions is expected to be maintained in the `teiHeader`, via elements such as `transcriptionDesc` and `scriptStmt`.

[13] http://fuseki.jerteh.rs/
[14] http://graphdb.jerteh.rs/sparql?repositoryId=digico



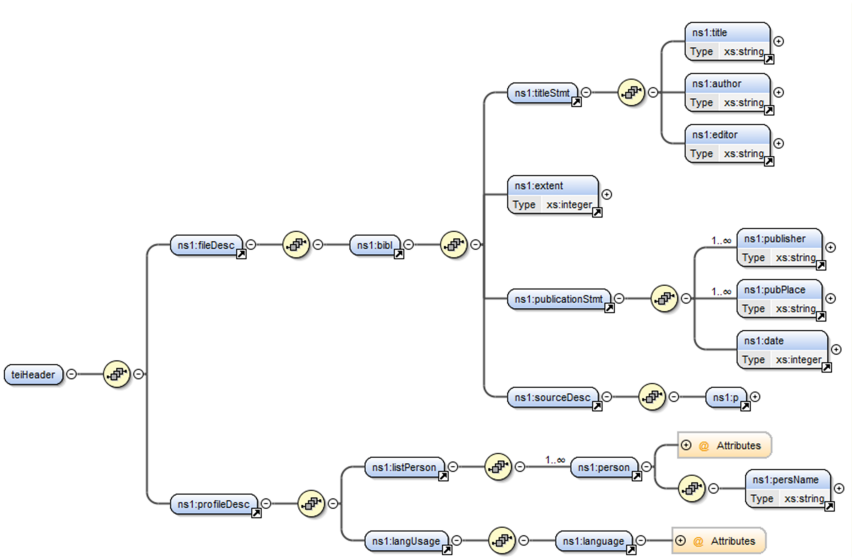
Slika 1. The components of the proposed workflow/toolchain

Complementing this, Chapter 2 of the TEI P5 Guidelines addresses the structure and function of the `teiHeader`, which provides the essential metadata framework for archiving interview material (see Figure 2). The header facilitates the systematic documentation of bibliographic and file-related information (`fileDesc`), editorial and encoding practices (`encodingDesc`), participant profiles and communication settings via `profileDesc` (e.g. `listPerson`, roles, place, time), as well as revision tracking (`revisionDesc`). In the context of interview data, these structures allow for the explicit recording of provenance, participant identities, recording circumstances, and methodological decisions—ensuring traceability, reproducibility and reliable citation. For an illustration, Sample 1.2 introduces an excerpt from *DigIko2* TEI header.

```
<teiHeader>
 <fileDesc>
  <bibl>
   <titleStmt>
    <title>Digitalne ikone 20+</title>
    <author>Vučenović, Tamara</author>
    <editor>Ružić, Nevena</editor>
   </titleStmt>
   <extent>510</extent>
   <publicationStmt>
    <publisher>Akademska knjiga</publisher>
    <publisher>RTS</publisher>
    <pubPlace>Novi Sad</pubPlace>
    <pubPlace>Beograd</pubPlace>
    <date>2024</date>
   </publicationStmt>
   <sourceDesc>
    <p><seg>Born digital - Publikacija sadrži intervjue [...]</seg></p>
   </sourceDesc>
  </bibl>
 </fileDesc>
 <profileDesc>
  <listPerson>
   <person xml:id="TaVu" role="interviewer" qid="Q109916651">
    <persName>Tamara Vučenović</persName></person>[...]
   <person xml:id="SlMa" role="interviewee" qid="Q19559876">
    <persName>Slobodan Marković</persName></person>
   <person xml:id="VoAn" role="interviewee" qid="Q1252236">
    <persName>Voja Antonić</persName></person>[...]
```

```
  <langUsage>
   <language ident="sr-Latn">Serbian (Latin script)</language>
  </langUsage>
 </profileDesc>
</teiHeader>
```

Sample 1.2. TEI header in DigIko2



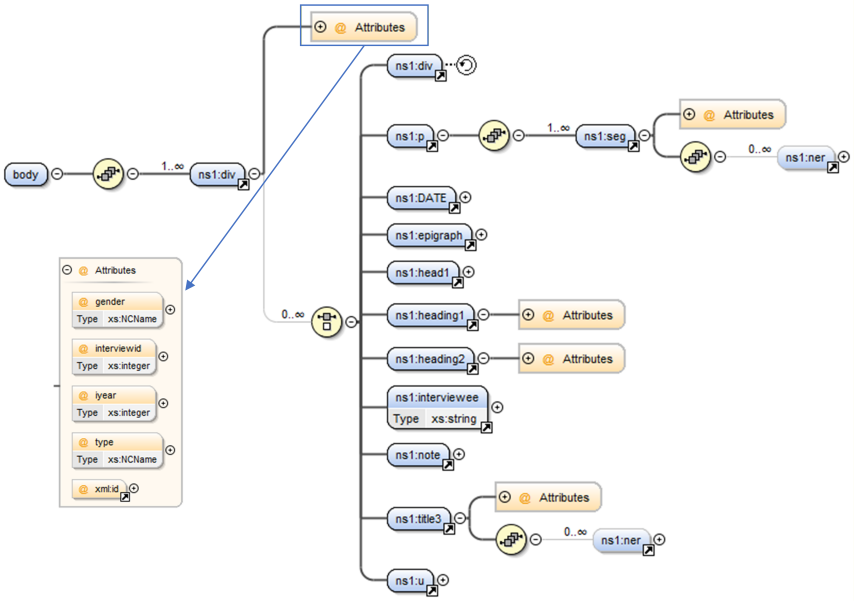
Slika 2. Schema of `<teiHeader>` element

Additional guidance on how to represent participant information and interaction settings is provided in the sections of the TEI Header/Profile concerned with transcribed speech (cf. Section 5.2.9 of the Guidelines). These recommendations specify which interview-specific attributes should be recorded, such as roles, sampling criteria, recording conditions, transcription methods, and where they should be encoded within the header. This approach fosters the discoverability, interoperability and long-term sustainability of spoken language resources [33].

The diagram 3 illustrates a TEI-aware XML schema where each interview is encoded within a `div` element enriched with descriptive attributes (gender, interview ID, year, type). For example, in *DigIko2* we use this element for different types of text divisions and text type is defined by type attribute. The following Sample 1.3 presents an interview introduction that provides context for the reader (or listener). The internal structure follows TEI's Transcriptions of Speech recommendations, segmenting the discourse into `u` units, each containing paragraphs (`p`), segments (`seg`), and optional entity annotations (`name`). Additional structural metadata such as titles, epigraphs, notes, and participant roles are supported through labeled heading elements and a dedicated `interviewee` field. This schema enables precise modeling of conversational flow while supporting linguistic annotation, knowledge-graph linking, and textometric analysis.

```
<div type="introduction"><p><seg id="1382">Domen .yu je registrovan negde 1989. ili
    1990. godine od strane <name type="ORG" ref="https://www.wikidata.org/wiki/
    Q5918873"> Instituta „Jožef Stefan"</name> u <name type="LOC" ref="https://www.
    wikidata.org/wiki/Q437">Ljubljani</name>.</seg></p>[...]</div>
```

Sample 1.3. Element `<div>` in *DigIko2*



Slika 3. Schema of `div` element

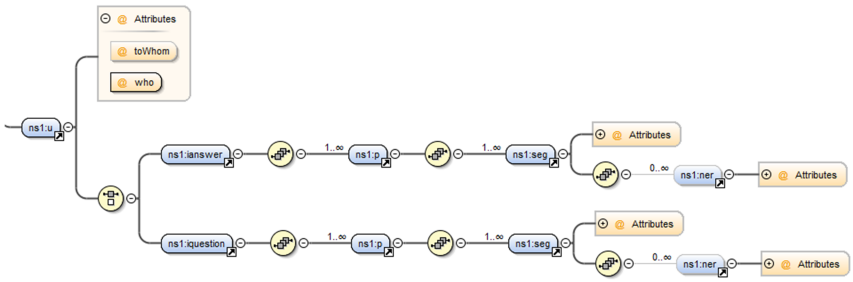
Figure 4 illustrates the structural model used for encoding question–answer pairs within the conversational layer of the corpus. Each utterance is represented using the `u` element, which can optionally contain two attributes: `@who`, identifying the speaker, and `@toWhom`, indicating the intended addressee. Within `u`, the model allows the occurrence of either `iquestion` or `ianswer` elements, reflecting the annotation of dialogic roles. The following example in Sample 1.4 illustrates an excerpt from *DigIko2* where the `u` element is used to group the speaker turn.

```
<u who="#TaVu"><iquestion><p><seg id="1407">Šta je potrebno izmeniti od podataka
    nalazi se u dokumentu koji možemo da pronađemo na sajtu <name type="ORG" ref="
    https://www.wikidata.org/wiki/Q136527687">Registra nacionalnog domena Srbije</
    name>.</seg><seg id="1408">Da li je sada već kasno?</seg></p></iquestion></u>
  <u who="#MiTa" toWhom="#TaVu"><ianswer><p><seg id="1409">Za neke manje korekcije
      nije kasno.</seg>[...]</p></ianswer></u>
```

Sample 1.4. Element `<u>` in *DigIko2*

Both `iquestion` and `ianswer` share an identical internal structure: they consist of one or more paragraphs (`p`), which are further segmented into `seg` units. These segments serve as operational anchors for sentence-level alignment, citation, and downstream linguistic processing. The schema also allows optional `name` elements associated with specific segments, enabling the annotation of named entities with additional attribute-level metadata. This design ensures that conversational flow remains TEI-compliant while remaining compatible with NLP pipelines that require fine-grained, segment-level supervision. This

structural layer segments discourse and provides context windows for downstream NLP as well as structure-aware querying (e.g., per speaker).



Slika 4. Schema of `u` element

3.2. Textometry in TXM

TXM facilitates import of a variety of textual formats, including plain texts and XML documents encoded according to TEI guidelines and generates an intermediate XML-TXM representation optimized for analysis [11]. This representation identifies three core text units:

- Textual units (e.g., books, articles, interviews) with associated metadata such as author, title, date, or genre;
- Structural units (chapters, passages, segments) representing document organization;
- Lexical units (tokenized words) that include graphic form, lemma, and grammatical category.

Metadata are treated as structured units within TXM and can be used for corpus partitioning, filtering, and comparison. Through frequency lists, concordance views, and sub-corpus extraction, TXM enables qualitative inspection, while quantitative techniques, such as specificity analysis, correspondence analysis, co-occurrence analysis, and clustering, allow detection of lexical trends and contextual patterns across texts. Results may be exported in tabular or graphical formats for further processing or visualization.

Although TXM supports automatic tagging and lemmatization via TreeTagger [29], it also accommodates external NLP tools whose annotations are added to the XML-TEI representation. In addition, TXM supports CQL (Corpus Query Language) querying, enabling expressive regular-expression search within structured corpora. Overall, TXM serves as a methodological bridge between classical textual analysis and computational linguistics. In the context of this research, TXM aligns well with the goals of the *DigIko2* corpus, where structured representation, linguistic annotation, and flexible querying pave the way for subsequent semantic enrichment and AI-based exploration.

In addition to providing morphosyntactic enrichment, the TXM platform functions as a comprehensive workbench for corpus exploration and quantitative linguistic reporting. It enables the calculation of a broad spectrum of textual statistics [19] and supports textometric analysis [11]. The strength of TXM lies in its integration of search, inspection, and measurement within a single interface, allowing researchers to move fluidly across analytical levels: from qualitative evidence via concordances, to quantitative measurements

through frequency counts, and further to multivariate statistical models. Crucially, the interface preserves links back to the original text passages, ensuring interpretability and supporting a tight connection between empirical findings and source-based validation.

Building on earlier work with the *SrpELTeC* corpus [30], the exploratory use of TXM is organized around several complementary analytical functions. Qualitative investigation is supported by: (1) concordance retrieval driven by the CQP engine and CQL syntax; (2) frequency lists calculated over tokens, lemmas, POS categories, or structural layers (including named entities); (3) progression plots that visualize lexical patterns across individual texts or corpus segments; and (4) an HTML-based reading interface that is tightly cross-linked with other analytical modules. The contrastive corpus relies on subcorpora or partitions, based on criteria such as text type, period, speaker, or metadata, and subsequently compare their statistical profiles using a unified workflow [12].

3.3. Corpus Management and Queries

TXM textometry is combined with *JeRTeh* language resources [31] and TESLA models [41], which provide essential preprocessing components, including sentence segmentation and supplementary linguistic annotations, thereby forming the foundation for further semantic enrichment and corpus-based inquiry.

The NER and NEL components are based on *TeslaXLM_2025*[15] trained as part of the TESLA-NER-NEL corpus [17]. The dataset includes 150,000 sentences, with entities linked to Wikidata identifiers. Annotation covers eight entity types: persons (`PERS`), places (`LOC`), organizations (`ORG`), events (`EVENT`), roles/professions (`ROLE`), works of art (`WORK`), demonyms (`DEMO`), and products (`PRODUCT`). Although joint NER+NEL architectures can enhance performance, they demand large annotated datasets and high computational capacity. To balance accuracy and efficiency, our approach utilizes spaCy's shared transformer backbone, enabling the joint training of both NER and NEL layers while preserving computational tractability. After applying the NER model, the recognized spans are inserted directly into the TEI-encoded corpus. The NEL step then links each entity to corresponding Wikidata items, resolving ambiguity using contextual cues; when necessary, new items are added to the knowledge base. This procedure normalizes references across different inflected forms by assigning them a unique identifier (e.g. a Wikidata QID), effectively anchoring the corpus within a broader knowledge-graph ecosystem and enabling semantic interoperability, data integration, and advanced inference.

CQL queries that combine structural and lexical annotations are supported in TXM and INCEpTION [18] instance maintained by JeRTeh Society.[16] From the annotated corpus, we also generated a linked-data representation in NIF (NLP Interchange Format) [13], following the approach used for the SrpELTeC corpus [16], with additional speaker-level metadata. The NIF is a Semantic Web–based RDF/OWL format that enables interoperability between different NLP tools, annotations, and corpora. It provides a standardized way to represent text spans, offsets, linguistic annotations (e.g., NER tags, links to Wikidata) and supports multilingual annotation workflows. NIF makes it possible to integrate heterogeneous resources into Linked Open Data, allowing annotations to be queried with SPARQL, visualized, and reused in downstream applications such as

[15] Serbian language model built on `xlm-roberta-large`
[16] `http://inception.jerteh.rs/`

knowledge graphs and RAG pipelines. In our work, NIF was used to export annotated TEI segments, preserve character-level spans, and link entities to Wikidata, thereby enabling interaction with external knowledge bases and graph visualization. A build has been deployed to a local Apache Jena Fuseki[17] endpoint to enable retrieval and analysis via SPARQL.

3.4. From Interview Transcripts to Short QA Pairs

The conversational structure of the *DigIko2* corpus, encoded in TEI with explicit speaker turns (`u`), paragraph units (`p`), and fine-grained segments (`seg`), provides an operational foundation for deriving question–answer (QA) items suitable for supervised QA training and RAG-oriented querying. Each interview turn contains both an interrogative component (often marked in `iquestion`) and an extended response (`ianswer`), which can be semantically decomposed into multiple atomic information units. Instead of treating a full turn as a single QA pair, we segment longer questions and answers into coherent discourse chunks, typically at the level of individual `seg` elements or small groups of contiguous segments. This enables the generation of multiple focused QA pairs grounded in a shared context, each targeting distinct facts or aspects.

To operationalise this process, each answer segment is analyzed in terms of factual content, temporal markers, named entity annotations, and narrative transitions. Questions are then reformulated in a shorter, direct manner (e.g. "Who was involved in the Galaksija computer project?") while preserving alignment with the original TEI context. For model training and evaluation, answers are stored either as (i) short extractive spans inside the relevant `seg` or (ii) concise abstractive paraphrases of one to three consecutive `seg` elements.

Previous studies suggest that the typical length of a question ranges from 5–25 tokens, while extractive answer spans vary between 3–30 tokens [27]. In contrast, abstractive answers tend to be substantially longer, often spanning 10–80 tokens, especially in dialogue-oriented QA settings [5,28]. Each QA item records its provenance via TEI identifiers (e.g. `xml:id` of the `u` or `seg`) and includes references to Wikidata QIDs when linked entities were detected during NEL.

In this procedure, a single dialogue turn can yield multiple QA pairs based on separate facts, narrative transitions, or entity mentions. TEI structure therefore acts as a natural scaffold for question segmentation, while NER/NEL layers provide entity-aware anchors for retrieval. This design is intentionally compatible with RAG pipelines: TEI segments or turns can serve as retrievable passages, entity links enable knowledge-graph expansion, and QA pairs provide supervision for evaluating both retrieval accuracy and answer grounding. Consequently, the semantically enriched interview corpus evolves into a RAG-ready resource, supporting automated reasoning and evidence-based exploration of oral-history materials.

We employed two versions of the OpenAI language model: `GPT-4.1` and `GPT-5.1`, accessed via the official OpenAI API in November 2025 [25]. Both models were prompted using a controlled experimental setup for paraphrasing and restructuring QA pairs. In pipeline, the prompt is designed to standardize faithful and style-controlled restructuring and paraphrasing of extracted question–answer pairs. It takes two textual inputs,

[17] http://fuseki.jerteh.rs/#/dataset/digico/query

question_in and answer_in, corresponding to the original question and answer, and produces a JSON-formatted output consisting of a list of QA pairs. These outputs represent concise Serbian summaries that remain semantically aligned with the source content.

A predefined instruction block enforces key generation constraints required for evaluation and downstream processing, including a neutral professional tone, elimination of redundancy, omission of irrelevant details, and strict semantic fidelity. By explicitly defining constraints, the prompt enables seamless integration with the retrieval module while ensuring that the quality of generated outputs is governed by transparent and verifiable criteria.

In practice, the user provides the question_in and answer_in fields with raw inputs (i.e., the content of an <iquestion> and its corresponding <ianswer> within a <u> element). The GPT model, accessed via the OpenAI API, processes these inputs and generates JSON list of simplified QA pairs. Each pair captures distinct factual elements, entity mentions, or narrative transitions present in the original question–answer content.

4. Serialization and Exploitation Framework for the *DigIko2* Corpus

The *DigIko2* is derived from the book *Digital Icons 20+* [38], which compiles selected interviews originally broadcast on the radio program. The book arranges these interviews into thematic sections that illustrate how digital society has evolved over the past two decades. It functions as a documentary account of the digital transformation in Serbia, capturing its impact on everyday life, cultural change, and the policy environment for technology adoption, both nationally and in a broader regional context [37]. Its primary aim is to highlight the importance of Serbia's digital transition and to identify, from a curated "bird's-eye" perspective, the key events, institutions, and individuals that have influenced this process.

The collection is designed to stimulate further interdisciplinary dialogue, not only with specialists in IT but also with experts from various fields. In total, the book includes interviews from 57 broadcast episodes and features 67 distinguished guests, many of whom actively contributed to shaping Serbia's digital history, offering an insightful narrative that links local developments to wider international trends. It includes, among others, conversations with researchers, engineers, and IT specialists who are, sadly, no longer with us, such as Vukašin Masnikosa, a member of the team that built Serbia's first digital computer (CER-10), and Dejan B. Popović, a close collaborator of the creator of the "Belgrade Hand", Academician Aleksandar Marinčić, and the author of numerous publications on Nikola Tesla. Notable international guests, Richard Stallman, Lev Manovich, Mike McGee, Eugene Kaspersky, and others—also appear in *Digital Icons 20+*, alongside many prominent figures in Serbian IT. In total, the volume features 67 interlocutors.

The *DigIko2* corpus provides a curated collection of interviews that chronicle the development of Serbia's digital ecosystem and technological landscape. In addition to its documentary value, the corpus represents a valuable reference resource for grounded and evidence-based NLP research, particularly within the context of low-resourced languages.

4.1. TEI-Encoded Representation of the *DigIko2* Corpus

The source material was originally prepared in *Microsoft Word* using a set of carefully designed custom styles, ensuring a consistent structural segmentation of the text. Each stylistic category captured a specific functional layer of the interview: *metadata about the session*, the *introductory segment* presenting the discussion topic and interlocutors, as well as clearly distinguished *interviewer questions* and *interviewee responses*. This structured annotation in Word served as the basis for an automated conversion pipeline, implemented in Python, which transformed the transcripts into TEI-compliant XML. The script mapped each custom Word style to a corresponding TEI element, encoding interview/session boundaries via `div` elements, marking paragraph breaks with `p`, and assigning unique speaker identifiers to `u` tags.

The Sample 1.5 shown in listing below illustrates a representative fragment of a single interview. The structure is rooted in a `div type="interview"` element, enriched with administrative attributes such as `@gender`, `@interviewid`, and `@iyear`. This is followed by a `DATE` block (containing `month` and `year` elements), a descriptive interview title encoded using `title`, and the identification of the interlocutor via `interviewee`.

The discourse is subsequently divided into `div type="introduction"`, which is the thematic context of the conversation, and `div type=čonversation"`, which contains the actual dialogic exchange. Speaker turns are encoded as `u` elements, each associated with a distinct speaker through the `@who` and `@toWhom` attribute to indicate addressed recipients). Within these `u` elements, the textual content is further structured into sentences and paragraphs using `p`, while fine-grained segmentation is achieved through `seg` tags. This level of granularity facilitates alignment, detailed citation practices, cross-referencing, and a wide range of textometric and computational analyses.

The tags related to the interview structure are shown in red, NER tags are annotated with two attributes: `type` and `ref`. Attribute `type` has one of the values from the list: `PERS`, `ORG`, `LOC`, `WORK`, `DEMO`, `ROLE`, `PRODUCT`. Attribute `ref` is the link of the entity in Wikidata knowledge base. For this fragment (Sample 1.5), many parts are omitted, but first QA pair is not, because it will be later used for demonstration of set of shorter QA pairs (in Subsection 4.5).

```
<div type="interview" gender="F" interviewid="08" iyear="2010">
 <DATE> <month>MART</month> <year>2010.</year> </DATE>
 <title id="1381">Opraštanje od .yu domena</title>
  <interviewee>Sagovornica: Mira Tasić</interviewee>
  <div type="introduction"><p><seg id="1382">Domen .yu je registrovan negde 1989. ili
       1990. godine od strane <name type="ORG" ref="https://www.wikidata.org/wiki/
      Q5918873">Instituta „Jožef Stefan"</name> u <name type="LOC" ref="https://www.
      wikidata.org/wiki/Q437">Ljubljani</name>.</seg><seg id="1383">Ta institucija
      je bila vođa projekta akademske mreže tadašnje <name type="LOC" ref="https://
      www.wikidata.org/wiki/Q83286">Jugoslavije</name>.</seg>[...]</p><p><seg id="
      1386">Danas u podne, u skladu sa odlukom <name type="ORG" ref="https://www.
      wikidata.org/wiki/Q242540">Internet korporacije za dodeljena imena i brojeve</
      name> (<name type="ORG" ref="https://www.wikidata.org/wiki/Q242540">ICANN</
      name>), prestaje da funkcioniše .yu domen.</seg><seg id="1389"><name type="ORG
      " ref="https://www.wikidata.org/wiki/Q136527687">Registar nacionalnog domena
      Srbije</name> podseća građane da će sve internet adrese koje koriste a završ
      avaju se sa .yu postati nevažeće, a sve internet prezentacije sa istim završ
      etkom biće nedostupne od danas u podne.</seg></p></div>
  <div type="conversation">
    <u who="#TaVu"><iquestion><p><seg id="1390">Danas je veoma važan datum za <name
        type="DEMO" ref="https://www.wikidata.org/wiki/None">srpski</name> internet.
        </seg><seg id="1391">O tome šta će se tačno desiti danas u 12 sati govoriće
```

```
        nam <name type="PERS" ref="https://www.wikidata.org/wiki/None">Mira Tasić</
        name>, jedna od ključnih ličnosti za razvoj interneta u <name type="LOC" ref
        ="https://www.wikidata.org/wiki/Q403">Srbiji</name>.</seg></p><p><seg id="
        1392">Šta će se zaista desiti?</seg><seg id="1393">Na koji način ćemo se
        konačno oprostiti od .yu domena?</seg></p></iquestion></u>
    <u who="#MiTa" toWhom="#TaVu"><ianswer><p><seg id="1394">Za ovaj korak smo se
        pripremali dve godine; od trenutka kada je počeo da se koristi .rs domen,
        znali smo da će .yu domen morati da ode sa interneta.</seg><seg id="1395">
        Trebalo je da se to dogodi u septembru prošle godine, u međuvremenu smo
        postigli produžetak rada .yu domena još pola godine.</seg><seg id="1396">Svi
         nazivi domena su zabeleženi na domen serverima.</seg><seg id="1397">Domen
        server je mašina kojoj je posao da, kada dobije upit od bilo koga na
        internetu za određeni domen, vrati mrežnu adresu računara na kome se nalazi
        odgovarajući domen.</seg><seg id="1398">Naš domen server za domen .yu nalazi
         se na <name type="ORG" ref="https://www.wikidata.org/wiki/Q6590075">
        Elektrotehničkom fakultetu Univerziteta u Beogradu</name> (<name type="ORG"
        ref="https://www.wikidata.org/wiki/Q6590075">ETF</name>).</seg><seg id="1399
        ">On u ovom trenutku ima oko 4000 .yu domena koji su produženi da važe do
        kraja postojanja .yu domena na internetu.</seg><seg id="1400">Danas u podne
        <name type="ROLE" ref="https://www.wikidata.org/wiki/Q255063">administrator<
        /name> će sve podatke o tih 4000 domena skloniti sa servera koji je određen
        da odgovara na pitanja gde se nalaze .yu domeni.</seg><seg id="1401">U slede
        ća 24 sata ta informacija će se propagirati preko interneta.</seg><seg id="
        1402"><name type="ROLE" ref="https://www.wikidata.org/wiki/Q278368">
        Korisnici</name> koji budu pokušali da dođu do adresa koje se završavaju .yu
         domenom dobiće poruku da ova adresa više ne postoji.</seg><seg id="1403">U
        toku današnjeg dana, <name type="ORG" ref="https://www.wikidata.org/wiki/
        Q242540">ICANN</name> će sa svojih osnovnih servera ukloniti zapis koji
        ukazuje na naš server na <name type="ORG" ref="https://www.wikidata.org/wiki
        /Q6590075">ETF-u</name> gde se nalaze podaci o .yu domenima.</seg><seg id="
        1404">Od tog trenutka .yu domen definitivno nestaje sa interneta.</seg></p><
        p><seg id="1405">To znači da će elektronske adrese koje koristimo, a koje su
         formirane na .yu domenu, postati nevažeće, sve internet prezentacije sa
        istim domenom će takođe biti nevažeće.</seg><seg id="1406">U pripremi za
        nestanak .yu domena sa interneta bilo je potrebno izvršiti izmenu podataka u
         protokolu svih servisa koji koriste nazive domena, od elektronske pošte,
        preko podešavanja veb-stranica i linkova ka njima.</seg></p></ianswer></u>
    <u who="#TaVu"><iquestion><p><seg id="1407">Šta je potrebno izmeniti od podataka
        nalazi se u dokumentu koji možemo da pronađemo na sajtu <name type="ORG" ref
        ="https://www.wikidata.org/wiki/Q136527687">Registra nacionalnog domena
        Srbije</name>.</seg><seg id="1408">Da li je sada već kasno?</seg></p></
        iquestion></u>
    <u who="#MiTa" toWhom="#TaVu"><ianswer><p><seg id="1409">Za neke manje korekcije
        nije kasno.</seg><seg id="1410">Ako se koriste adrese koje se završavaju na
        .yu i neki korisnik pošalje imejl na takvu adresu, mejl-servis će pokušavati
         24 sata da isporuči tu poštu.</seg><seg id="1411">Nakon toga korisnik će
        dobiti poruku da taj mejl više ne postoji i moraće da se raspita na neki
        drugi način, telefonom ili preko veb-adrese kako da dođe do nove
        imejl-adrese.</seg>[...]</p></ianswer></u>
[...]
    <u who="#TaVu"><iquestion><p><seg id="1487">Volela bih da spomenemo i ćirilički
         domen na internetu, koji je u pripremi.</seg>[...]</p></iquestion></u>
    <u who="#MiTa" toWhom="#TaVu"><ianswer><p><seg id="1489">U dogovoru sa <name type=
        "ORG" ref="https://www.wikidata.org/wiki/Q38216395">Ministarstvom za
        telekomunikacije</name>, inicijativu je pokrenuo <name type="ORG" ref="https:
        //www.wikidata.org/wiki/Q136527687">Registar nacionalnih internet domena
        Srbije</name>.</seg>[...]<seg id="1494">Procedura izbora nacionalnog ćirilič
        kog domena mora da bude transparentna, svi detalji procesa moraju da budu
        stavljeni javnosti na uvid.</seg></p></ianswer></u></div>
</div>
```

Sample 1.5. XML representation of interview structure and NER+NEL tags

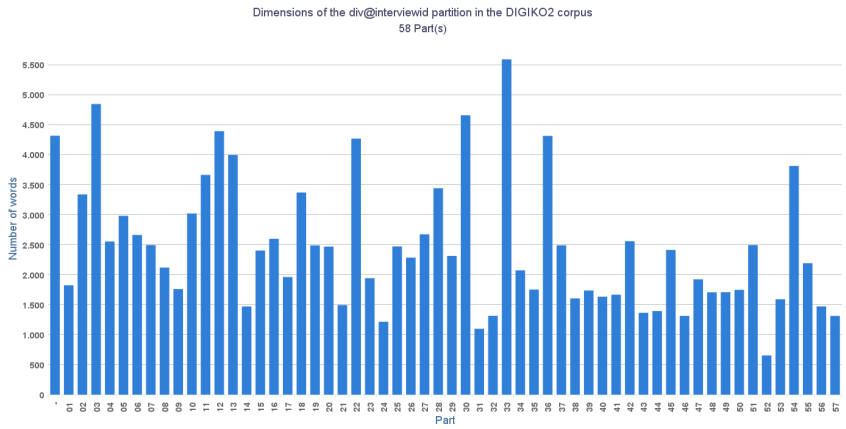
For pragmatic marking, we wrap question/answer spans with project-specific elements `iquestion` and `ianswer`; in a standards-conformant profile these may be modeled as `seg type="question"`/`seg type="answer"`. The schema enforces (i) controlled values for `@type` on `div`, (ii) required `@who` on `u`, and (iii) content models `u`

→ (p+) and p → (seg*), ensuring interoperable structure for NER/NEL offsets, speaker-level provenance, and downstream linked-data serialization.

4.2. Textometric Methods Applied to *DigIko2* via TXM

The *DigIko2-TEI* corpus was enriched with part-of-speech (POS) and lemmatization layers using the TXM platform [12,11], enabling corpus-wide linguistic analysis and structured querying. The current version consists of 57 TEI-encoded interviews, amounting to 1,334 annotated utterances, of which 665 are questions and 669 are responses—the latter exceeding question counts because certain prompts were addressed to multiple interlocutors.

The corpus comprises a total of 8,868 sentence-level segments, aligned with the seg layer, and contains 142,665 running word tokens, yielding sufficient linguistic density for both qualitative inspection and quantitative modelling. Corpus size distribution across subpartitions that represents interviews is illustrated in Figure 5. The minimum length is 654 words, the maximum is 5,591, with an average of 2,423 words and a standard deviation of 1,061. The histogram visualizes word counts per interview, including corresponding introductory sections as first bar. Similar partitioning can be done by year, speakers, gender of speakers, etc., all, selected or grouped classes. This statistic provides an initial profile of corpus dimensions and are instrumental for downstream analyses such as topic analysis, diachronic comparison, and speaker-based profiling.

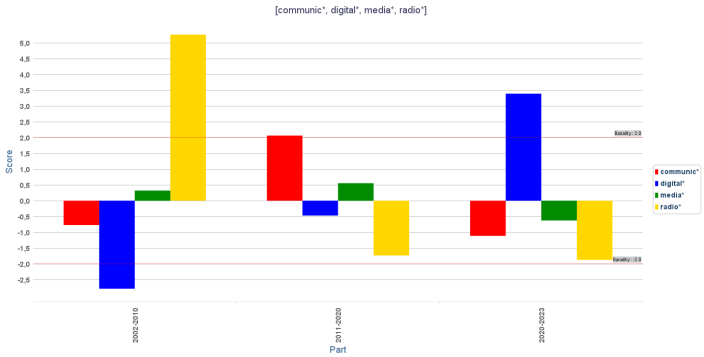


Slika 5. Corpus partition by interviews

Figure 6 presents the specificity analysis for four technology-related concepts across three temporal periods of the *DigIko2* corpus. The bars indicate the specificity score of each concept, where positive values reflect a frequency above what would be statistically expected under a uniform distribution, while negative values reflect underuse. A conventional "banality" threshold is set at −2, below which a term is considered not characteristic for the given corpus segment. The concepts were operationalized using CQL search expressions; for instance, *communic** captures forms such as *komunikacija*, *komunikacij-*

ski, *komunikacioni* by expression: `[srlemma="komunikac.*]`, while *digital** retrieves *digitalan*, *digitalizacija*, *digitalizovati*, by expression: `[srlemma=đigital.*]`.

The resulting temporal patterns reveal meaningful conceptual shifts across the corpus. During the early interviews (2002–2010), the concept *radio* shows positive specificity, reflecting its well-established presence in the discourse of that period. In the middle period (2011–2020), communication-related terms are only marginally specific, indicating a more diffuse or transitional relevance. The most notable development concerns the concept *digital*, which exhibits negative specificity in the earliest period but becomes strongly positive in the most recent segment (2020–2023), clearly surpassing all other terms. This sharp rise points to a reorientation of public discourse towards digitally driven paradigms, aligned with broader technological shifts and growing societal awareness of digital transformation.



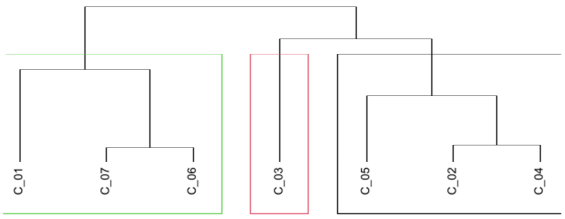
Slika 6. Specificity over time for selected key terms

The interviews in the *DigIko2* corpus were originally grouped into seven thematic chapters, each reflecting a specific perspective on the evolution of digital society in Serbia. These chapters range from biographical and narrative portrayals of key figures (e.g., *C_01* and *C_07*), to domain-oriented discussions on engineering, communication and art (*C_02*, *C_04* and *C_05*), and broader socio–cultural reflections in *C_03*. In order to assess whether the human editorial structure corresponds to linguistic patterns in the data, we applied a TXM-based hierarchical clustering procedure, relying on lexical distributions and structural features extracted from the TEI-encoded corpus. The resulting dendrogram (Figure 7) reveals three macro-clusters that offer meaningful insights into the corpus composition and thematic coherence.

The first cluster (*C_01*, *C_07*, *C_06*) groups interviews that share a predominantly narrative style and exhibit a biographical or reflective tone. These chapters tend to foreground personal trajectories, digital icons, and literary perspectives rather than technical discourse, which may explain their strong linguistic similarity. The second cluster emerges as a singleton, containing only *C_03* (*Society and the Individual*), which behaves as a linguistic outlier. This is consistent with chapter content: the interviews in *C_03* involve broader commentary on societal implications of digital transformations and include heterogeneous registers, ranging from philosophical to policy-oriented discourse. Its sepa-

ration in the dendrogram therefore underscores a distinctive conceptual mode that bridges personal narrative and analytical reflection.

The third cluster links *C_05*, *C_02* and *C_04*, all of which revolve around applied technological domains, including engineering, internet infrastructure, communication practices, and artistic uses of digital media. These chapters show stronger concentrations of domain-specific vocabulary and technical terminology, resulting in higher linguistic cohesion at the lexical level. This correspondence between human categorization and computational clustering suggests that thematic organization can be partially inferred from textual features alone, although certain hybrid or meta-level reflections (as seen in *C_03*) may require additional semantic and discourse-oriented analyses. Overall, the TXM clustering supports the structural logic of the original editorial design and highlights the potential of corpus-driven methods to inform future curatorial or annotation strategies.



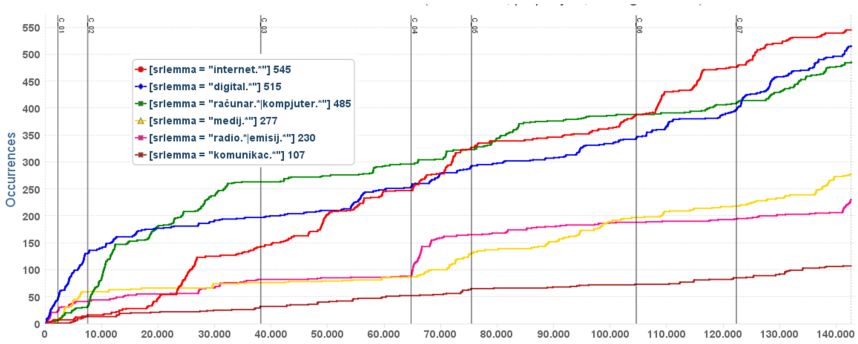
Slika 7. Hierarchical clustering, cluster dendogram over chapters

Figure 8 presents the cumulative progression of six technology-related concepts across the *DigIko2* corpus chaper partitions. Each line represents the cumulative frequency of a CQL-based lemma expression (`internet*`, `digital*`, `računar*|kompjuter*`, ...) tracked across the full linear order of interviews. The vertical grid lines correspond to transitions between chapters as corpus partitions. The progression curves reveal clear shifts in terminology and discourse focus over time. Terms related to *internet*, *digital*, and *računar/kompjuter* (computers) demonstrate strong and steadily increasing usage, with marked acceleration in the second half of the corpus. Notably, *internet.** becomes the dominant term in later segments.

Conversely, terms associated with *radio* and *emission* display a modest and relatively stable trajectory, reflecting the original media format of the interviews but decreasing thematic centrality. The lexical field *medij** shows a similar but stronger growth trend, particularly in later segments, likely aligned with discussions of digital journalism, media convergence, and online platforms. The term *komunikac** remains the least frequent overall but exhibits steady incremental growth, suggesting that communication-related discourse emerges more strongly as the digital paradigm matures. Taken together, the cumulative trajectories indicate a clear transition from traditional media terminology toward digitally centered concepts.

4.3. Semantic Annotation of the *DigIko2* Corpus in INCEpTION

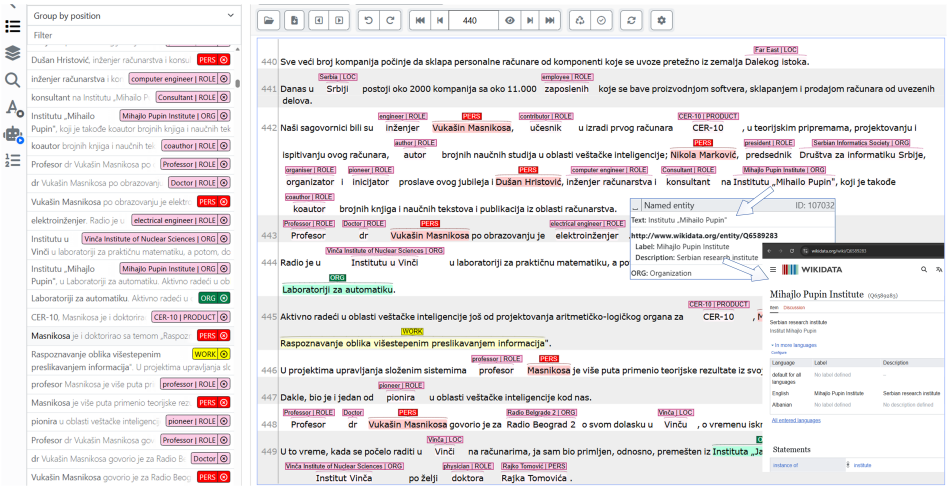
Figure 9 illustrates the local deployment of the INCEpTION platform, which was used to manually refine the automatically generated NER layer and to verify or assign NEL links to Wikidata identifiers. In this environment, annotators directly select spans of running text and assign entity types, following the inventory described in Subsection 3.3.



Slika 8. Progression with selected queries

Each annotated mention is then linked to a corresponding Wikidata entry, such as *Mihajlo Pupin Institute* (Q6589283), thereby grounding the annotation in a canonical knowledge base. The left panel provides an ordered list of annotations grouped by different option, in this case text position; the central panel highlights entity spans with color-coded markers; and panel with NER class and Wikidata metadata, assisting annotators in resolving ambiguity. This explicit linking between surface forms and QIDs facilitates consistent disambiguation and ensures that all annotations are both human-interpretable and machine-readable.

By capturing Wikidata identifiers together with entity spans and their assigned roles, the INCEpTION workflow produces a reproducible and interoperable annotation layer suitable for downstream processing. The resulting corpus is therefore compatible with knowledge-graph construction, entity-based retrieval, and RAG-style question answering, offering a principled foundation for future AI-driven exploration of the *DigIko2* corpus.



Slika 9. NER and NEL within INCEpTION

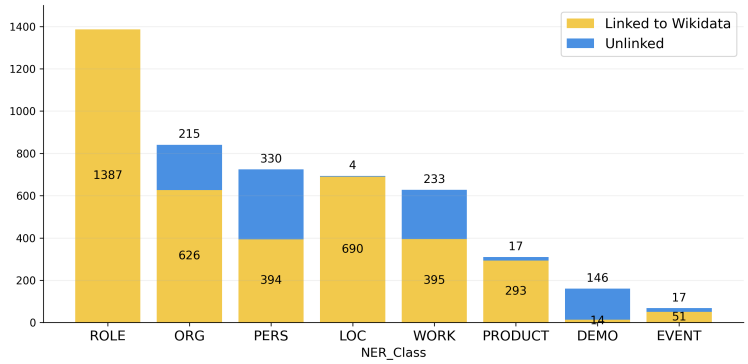
To extract annotated organisations in INCEpTION query would be: `<Named_entity.value ="ORG"/>`. For linked entities, for example `Q6589283` *Mihajlo Pupin Institute*, query is: `<Named_entity.identifier = "http://www.wikidata.org/entity/Q6589283"/>`. To extract annotated organisations in TXM query would be: `<name_type="ORG» []+</name>` . The most frequent entity type is ROLE 1,387 (329), where in brackets are distinct occurrences. The most frequent roles are: umetnik (artist, 79), profesor (professor, 70), student (55), novinar (journalist, 44), korisnik (user, 41). The second most frequent class is `ORG` 841 (469), followed by: `PERS` 724 (384), `LOC` 694 (181), `WORK` 628 (307), `PRODUCT` 309 (140), `DEMO` 160 (45), and `EVENT` 68 (48).

4.4. Publishing *DigIko2* as Linked Data via NIF Serialization

In order to enable interoperable, Linked Data friendly processing of the *DigIko2* corpus, we developed a custom export pipeline from TEI to the NLP Interchange Format (NIF). Starting from TEI-encoded segments with manually corrected NER and NEL layers, each `<seg>` unit is converted into a `nif:Context`/`nif:String` with canonical character offsets, while every `<name>` element is mapped to a `nif:Phrase` anchored via `nif:anchorOf` and `nif:beginIndex`/`nif:endIndex`. Entity links are preserved by attaching Wikidata identifiers from the TEI `@qid` attribute as `itsrdf:taIdentRef`, and speaker information from the TEI header (e.g. `<person xml:id="...">` with `@role` and `@qid`) is propagated into the NIF graph using a custom `di:speaker` property pointing to `foaf:Person` nodes. In this way, *DigIko2* becomes a richly annotated NIF corpus that simultaneously encodes textual spans, entity linking, and discourse roles, and can be queried with SPARQL or visualized via Wikidata Query Service to analyse who mentions which entities in the interviews, in line with the general NIF design principles [13].

After automatic annotation, manual curation was performed on the entire *DigIko2* corpus. Figure 10 presents the absolute number of annotated entity spans, distinguishing between those linked to Wikidata and those not linked. Out of the total annotated spans, `ROLE` and `LOC` show almost complete coverage, with 100% and 99.4% successfully linked to Wikidata, respectively, indicating that occupational roles and geographical names are highly standardised and consistently recognised. `PRODUCT` also shows high alignment (94.5%), followed by `EVENT` (76.1%), `ORG` (74.3%), and `WORK` (63.0%), where partial Wikidata coverage still allows for meaningful knowledge-graph construction but leaves room for targeted enrichment. In contrast, `PERS` exhibits a more balanced distribution (*54.1% linked* vs. *45.9% unlinked*), suggesting the need for additional manual disambiguation of personal names, especially in cases of local relevance or ambiguous referents. The `DEMO` class (demographic references) remains largely unlinked (only 8.8% coverage), signalling that this category requires the most attention in future annotation cycles, but this is related how this type of data is recorded in Wikidata. Taken together, these results provide actionable insight into which entity classes are mature enough for direct knowledge-graph ingestion and which require further curation to ensure full interoperability with external knowledge bases such as Wikidata.

In the following fragment from *DigIko2* NIF edition (Sample 1.6), the TEI segment containing the sentence *"Rodoljub Šabić je poverenik za informacije od javnog značaja i zaštitu podataka o ličnosti u Republici Srbiji."* is represented as a `nif:Context`/`nif:String` resource `di:seg-2886`. The fragment's character offset



Slika 10. Number of entities that are recognized and linked with Wikidata

is specified by `nif:beginIndex` and `nif:endIndex` and a local identifier stored in `dct:identifier`.

The mention of the person *Rodoljub Šabić* is modelled as a separate `nif:Phrase` node `di:seg-2886_char_0_14`, anchored to the surface form via `nif:anchorOf` and aligned with its span in the context using `nif:beginIndex=0` and `nif:endIndex=14`. The custom property `di:nerClass` encodes the NER label `"PERS"`, while `itsrdf:taIdentRef` property links `wd:Q19057243` as the mention to the corresponding Wikidata entity for Rodoljub Šabić. In addition, both the context and the phrase are connected to the discourse participant `di:person-TaVu` through the `di:speaker` relation, enabling downstream analysis of which speakers introduce which entities in the *DigIko2* corpus.

```
di:seg-2886 a nif:Context, nif:String ;
   di:speaker di:person-TaVu ;
   nif:beginIndex "0"^^xsd:nonNegativeInteger ;
   nif:endIndex "110"^^xsd:nonNegativeInteger ;
   nif:isString "Rodoljub Šabić je poverenik za informacije od javnog
   značaja i zaštitu podataka o ličnosti u Republici Srbiji." ;
   dct:identifier "2886" .
di:seg-2886_char_0_14 a nif:Phrase ;
   di:nerClass "PERS" ;
   di:speaker di:person-TaVu ;
   nif:anchorOf "Rodoljub Šabić" ;
   nif:beginIndex "0"^^xsd:nonNegativeInteger ;
   nif:endIndex "14"^^xsd:nonNegativeInteger ;
   nif:referenceContext di:seg-2886 ;
   itsrdf:taIdentRef wd:Q19057243 .
```

Sample 1.6. Fragment from *DigIko2* NIF edition

4.5. Generating QA dataset

The generation of the new QA dataset based on the original dialogue in the *DigIko2* corpus was performed in Python on a dataset of 661 question–answer (QA) pairs. It should be noted that several questions had multiple answers because more speakers were involved in the conversation. The original questions had an average length of 189 characters, whereas the answers averaged 892 characters. We applied `GPT-4.1` and `GPT-5.1` (November 2025 versions) to paraphrase and reformulate the QA pairs.

Table 1 compares the original dataset with the QA pairs generated by `GPT-4.1` and `GPT-5.1`. While the original interviews contained 665 questions and 669 answers, both models produced more than three times the number of QA pairs. This indicates that the conversation was reconstructed so that longer QA segments were split into multiple QA units, where each question had exactly one answer, thus forming a dialogue structure closer to that of a chatbot.

Tabela 1. Statistics of Original and Generated QA Pairs

Type	**Original**		**GPT-4.1**		**GPT-5.1**	
	N°	Avg. Chars	N°	Avg. Chars	N°	Avg. Chars
Questions	665	271	2,256	57	1,987	68
Answers	669	1,091	2,256	126	1,987	232

`GPT-4.1` yielded the largest number of QA pairs (2256), but with a substantial reduction in length (avg. 57 and 126 chars). This suggests a fragmentation-oriented strategy, prioritising quantity over contextual depth. `GPT-5.1`, however, generated fewer QA pairs (1987) but substantially longer answers (avg. 232 characters), indicating a shift towards richer and more semantically dense content. These findings illustrate an important trade-off between coverage and context preservation. `GPT-5.1` appears better suited for downstream tasks such as NER, NEL, and RAG pipelines, where the quality and informativeness of each QA pair are more important than their sheer number. To illustrate, we give a Sample 1.7 of one long QA pair given in Section 4.1 that was refactored into four QA pairs.

```
<generated>
 <qapair xml:id="qa-152">
  <q>Šta će se tačno desiti danas u 12 sati sa .yu domenom?</q>
  <a>Danas u podne administrator će obrisati sve podatke o oko 4000 preostalih .yu
       domena sa servera na Elektrotehničkom fakultetu u Beogradu, koji je zadužen da
        odgovara na upite za .yu domene. U naredna 24 sata ta promena će se proširiti
        internetom, a ICANN će ukloniti zapis koji upućuje na ovaj server, čime .yu
       domen definitivno nestaje sa interneta.</a>
 </qapair>
 <qapair xml:id="qa-153">
  <q>Zašto i kako je pripreman prelazak sa .yu na .rs domen?</q>
  <a>Od uvođenja .rs domena bilo je poznato da .yu mora biti ugašen. Gašenje je
       prvobitno bilo planirano za septembar prošle godine, ali je rad .yu domena
       produžen za još šest meseci. Tokom tog perioda korisnici su trebalo da pređu
       na .rs i prilagode svoje servise novom domenu.</a>
 </qapair>
 <qapair xml:id="qa-154">
  <q>Šta će se desiti sa sajtovima i mejl adresama na .yu domenu?</q>
  <a>Svi sajtovi i elektronske adrese koji koriste .yu domen postaće nevažeći.
       Korisnici koji pokušaju da pristupe adresama koje se završavaju na .yu dobiće
       poruku da ta adresa više ne postoji.</a>
 </qapair>
 <qapair xml:id="qa-155">
  <q>Koje tehničke promene su bile potrebne zbog gašenja .yu domena?</q>
  <a>Bilo je potrebno izmeniti podatke u konfiguraciji svih servisa koji koriste
       nazive domena: elektronske pošte, veb-stranica i linkova ka njima, kako bi
       umesto .yu koristili nove domene, pre svega .rs.</a>
 </qapair>
</generated>
```

Sample 1.7. Example of context-aware QA pair refactoring in `GPT-5.1`

4.6. Extending dataset

To expand our dataset, we collected and analysed a set of 586 posts published on the RTS (Radio Television of Serbia) website. For each post, we extracted the title, the introductory lead segment, the full HTML body, and the corresponding URL. After removing some HTML markup (fonts, comments, colors), the textual content was manually annotated and compiled into a unified XML file containing 586 structured entries. The average length of the title was approximately 60 characters, the lead averaged 190 characters, and the body of the post around 620 characters. The following Sample 1.8 illustrates the structure of an annotated item, where named entities are consistently marked using the `ner` element with classes for persons, organizations, roles, creative works, and products, accompanied by Wikidata identifiers to support disambiguation and interoperability.

```
<item>
  <Title id="196"><name type="PERS" ref="https://www.wikidata.org/wiki/Q131271652">
      Anđelka Zečević</name></Title>
  <Lead id="197">U emisiji <name type="WORK" ref="https://www.wikidata.org/wiki/
      Q136525085">Digitalne ikone</name> danas je sa nama <name type="PERS" ref="
      https://www.wikidata.org/wiki/Q131271652">Anđelka Zečević</name>, <name type
      ="ROLE" ref="https://www.wikidata.org/wiki/Q11900058">istraživač</name> na <
      name type="ORG" ref="https://www.wikidata.org/wiki/Q123951838">Matematičkom
      institutu SANU</name>, u sektoru za računarstvo.</Lead>
  <Link>https://www.rts.rs/lat/radio/radio-beograd-2/5632717/andjelka-zecevic.html<
      /Link>
  <Body>
    <p id="198">Razgovaraćemo o Anđelkinim oblastima interesovanja, o obradi
       prirodnog jezika, digitalnoj humanistici i medicinskoj informatici, o
       projektu u oblasti mašinskog učenja i tome kako nam obrada velike kolekcije
        podataka može pomoći u otkrivanju ranih prediktora demencije, kao i o
       aktivnostima društva za jezičke resurse, <name type="ORG" ref="https://www.
       wikidata.org/wiki/Q97203619">JerTeh</name>, čiji je jedan od osnivača.
       Poseban deo razgovora posvetili smo primeni veštačke inteligencije u
       medicini, istraživanjima koja treba da doprinesu adekvatnoj bezbednosti <
       name type="ROLE" ref="https://www.wikidata.org/wiki/Q181600">pacijenata</
       name>, kao i tome zašto kaže da alatu <name type="PRODUCT" ref="https://www
       .wikidata.org/wiki/Q115564437">ChatGPT</name> dugujemo zahvalnost za
       globalnu i najbolju reklamu u proteklih 30-ak godina, ali ga grdimo i zbog
       mnogih turbulencija koje je uveo u ovu oblast.</p>
    <p id="199">Emisiju uređuje i vodi <name type="PERS" ref="https://www.wikidata.
       org/wiki/Q109916651">Tamara Vučenović</name>.</p>
  </Body>
</item>
```

Sample 1.8. An XML representation of a post from the RTS website

In total, 9,111 entity mentions were annotated across the corpus. The distribution of categories reflects the interview-driven structure typical of RTS content: roles and professions were the most frequent type with 2,849 mentions, followed by personal names with 2,266 mentions, organizations with 2,262 mentions, geographical entities with 853 occurrences, and a smaller number of creative works, events, demographic descriptors, and products. Despite the high number of annotated mentions, only 952 entities were linked to Wikidata, predominantly geographical entities (766) and organizations (167). This disparity highlights both the richness of the corpus in context-dependent or local expressions, many of which lack stable knowledge-base identifiers, and the challenges of grounding Serbian media content into global knowledge graphs. At the same time, the dataset provides a valuable resource for evaluating NER and NEL methods and for developing Serbian-language knowledge-graph applications.

5. Discussion

The automatic transformation from MS Word into TEI required several iterations of corrections in order to achieve a consistent structure. For TEI schema generation and corpus validation XMLSpy tool was used. It should also be noted that we tested several schema variations to produce structure appropriate for the analysis, corpus partitioning, search, extraction and transformation.

The evaluation of the NER and NEL components over the whole *DigIko2* interview corpus shows that the model performs strongly on several entity types, despite the considerable linguistic variability introduced by the fact that approximately 70% of the training corpus consists of literary text. The model is transformer-based NER+NEL model based on a pretrained model: a multilingual model with 561 million parameters (`TeslaXLM`) [41,42]. For NEL, the results for the `LOC` category demonstrate high precision (0.90) with recall of 0.78, yielding an F1-score of 0.84. This indicates that the linking module is generally effective in associating recognized locations with the correct Wikidata identifiers, although a moderate number of false negatives suggests that recall could still be improved. Other classes for NEL are at early stage.

In the NER evaluation, several categories, especially `PERS`, `LOC`, `ROLE`, and `DEMO`, exhibit strong performance. The `PERS` class achieves an F1-score of 0.95, reflecting highly reliable recognition of person entities across both interview-based and literary segments of the corpus. The `LOC` class also performs very well (F1 = 0.91), and its high recall (0.96) aligns with the consistent behavior observed in the NEL evaluation. Strong results for `ROLE` (F1 = 0.84) and `DEMO` (F1 = 0.93) further confirm that categories with clearer semantic boundaries are captured effectively. More challenging categories include `ORG`, `WORK`, and `PRODUCT`. The `ORG` class shows moderate recall (0.76) but lower precision (0.57), indicating over-generation, likely due to contextual ambiguity or similarity to other nominal expressions. The model often made errors when recognizing `ORG` entities such as *Radio Beograd 2*, failing to include the number *2* as part of the entity name. It also misclassified examples like *Ljudmila iz Ljubljane*, labeling it as an `PERS` instead of correctly identifying *Ljudmila* as an `ORG` entity. The `WORK` (F1 = 0.41) and `PRODUCT` (F1 = 0.28) categories suffer from low recall values (0.28 and 0.18), suggesting that many relevant entities remain unrecognized, which is expected since there was not many labeled entities of this category in the training dataset.

Overall, the model robustly handles categories with more stable linguistic and semantic patterns, such as `PERS`, `LOC`, `DEMO`, and `ROLE`, even within a hybrid corpus combining interviews and narrative prose. However, categories characterized by greater variability or domain-specificity, such as `EVENT`, `WORK` and `PRODUCT`, require further improvement. Future enhancements may focus on domain-adaptive training, targeted augmentation of underrepresented entity types, and improved handling of culturally or stylistically marked references. The present results establish a strong baseline for further development of NER+NEL pipelines for complex, mixed-genre corpora such as *DigIko2*.

The main sources of error likely stem from ambiguous or morphologically inflected toponyms, incomplete gazetteers, and conservative confidence thresholds. Future improvements should focus on expanding alias lists, enhancing candidate generation for inflected forms, and leveraging contextual cues such as co-occurring regions or temporal information. The proposed TEI-based structure, together with its multi-layer annotation schemas, enables a range of advanced functionalities, including semantic search, per-

Tabela 2. NER and NEL Evaluation Results on the *DigIko2* Corpus

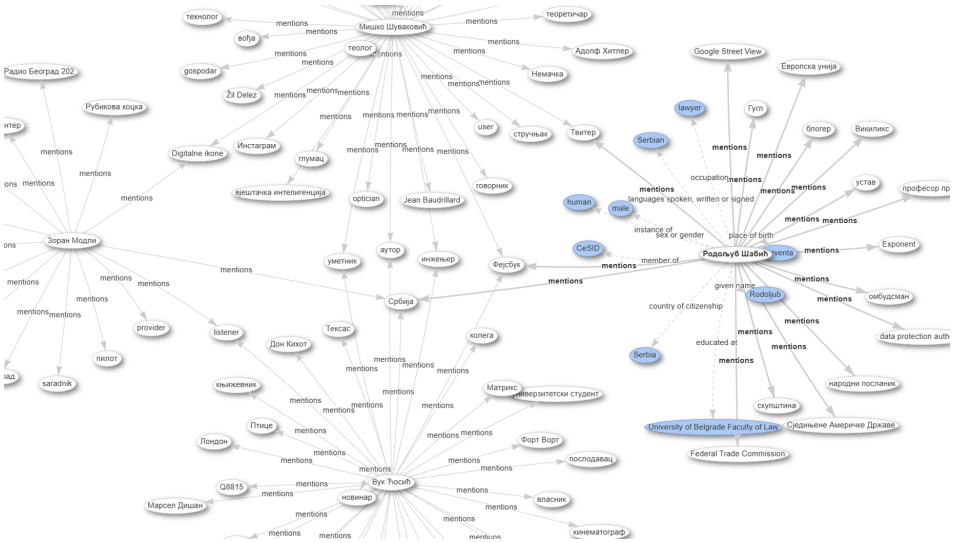
	PERS	LOC	ORG	ROLE	WORK	DEMO	EVENT	PRODUCT
TP	697	667	639	1,111	175	151	43	57
FP	42	108	488	151	51	13	12	34
FN	27	27	202	276	453	9	25	253
P	0.94	0.86	0.57	0.88	0.77	0.92	0.78	0.63
R	0.96	0.96	0.76	0.80	0.28	0.94	0.63	0.18
F1	**0.95**	0.91	0.65	0.84	0.41	0.93	0.70	0.28

speaker retrieval, and knowledge graph querying via SPARQL. Every TEI element is indexed with persistent pointers, allowing direct cross-referencing both within the corpus and across external resources. During retrieval, speaker-specific filters and Wikidata QID facets can be applied to return context-aware results.

The retrieval framework supports multiple evidence views: (i) *speaker-centric digests*, which restrict results to utterances of a single participant; (ii) *entity-centric synopses*, aggregating all segments linked to a given Wikidata entity; and (iii) *thematic overviews*, which provide high-level content grouping to support topic-driven exploration.

For factoid-style queries (e.g., *Who mentioned X, and in which turn?*), the system returns speaker-level and turn-level evidence, which the LLM may cite explicitly. Furthermore, the responses can be reformulated for non-expert audiences while keeping traceable references to the original source utterances. This mechanism also permits the generation of concise briefings or controlled-length executive summaries, ensuring faithful content simplification rather than paraphrastic drift. The evidence views can additionally highlight how different speakers discuss the same entity or theme, consolidate cited segments into attribution maps over the interview structure, and reveal low-confidence claims or missing links (NEL gaps). Such outputs are particularly valuable for curator inspection and iterative dataset refinement. Domain adaptation is required for NEL since many recognized entities are not linked with Wikidata items. The current knowledge graph should be completed, relations better explored. Figure 11 presents a graph where speakers (interviewees) of mentions linking locally extracted entities to Wikidata items. Each node is a person, organization, place, event, or concept; solid directed edges encode *who mentions who/what* in the corpus (e.g., *Rodoljub Šabić* and *Miško Šuvaković* $\rightarrow$ *Serbia*, *Facebook*, *Twitter*).

The obtained results for NER are highly encouraging and broadly comparable to those reported for large, well-resourced languages. In contrast, the evaluation of NEL demonstrates that substantial room for improvement remains. This should not be interpreted as a limitation of the approach, but rather as an indication of the inherent difficulty of grounding Serbian entities in global knowledge bases, where coverage is still uneven and context-dependent expressions are frequent. It is important to emphasise that this is currently the only model that explicitly addresses the NEL task for the Serbian language. The lower recall is partly a consequence of the rich and highly inflectional morphology of Serbian, but an even more important factor is the absence of many context-specific or culturally bound concepts in Wikidata, which prevents reliable linking even when the entities are correctly recognised. Intensive work on this problem is ongoing within the TESLA project, and significantly improved NEL performance can be expected in the near future as resources, lexicons, and training data continue to expand. The dataset is publicly



Slika 11. A graph: Who was mentioning who/what?

available through the DigIko open data page,[18] and HuggingFace [19] while a dedicated SPARQL endpoint has been established to enable programmatic access and querying of the data.[20]

6. Conclusion

This paper introduced a comprehensive and standards-based pipeline for transforming unstructured interview transcripts into a semantically enriched, interoperable, and queryable resource. By integrating TEI-encoded structure with NER and NEL models linked to Wikidata, textometric analysis in TXM, NIF-based linked-data serialization, and retrieval augmented generation (RAG), the approach establishes a multi-layer framework for evidence-grounded exploration of the *DigIko2* corpus.

The TEI encoding provides stable anchors for segmentation, speaker attribution, metadata-driven retrieval, and reproducible citation practices, enabling granular analysis of interview discourse. The NER component achieves reliable performance for the core entity types (`PERS`, `ROLE`, `LOC`), reflecting the strength of Serbian transformer models trained within the TESLA project. More challenging categories—`ORG`, `WORK`, and `PRODUCT`—reveal the need for extended training data and richer domain-adapted lexicons. NEL performance remains promising but incomplete: high precision for locations is contrasted with moderate recall, hindered primarily by rich Serbian morphology and the limited coverage of culturally specific entities in Wikidata. These findings highlight the importance of continued knowledge-graph enrichment and improved candidate-generation strategies for low-resourced languages.

[18] https://llod.jerteh.rs/digiko/
[19] https://huggingface.co/datasets/te-sla/DigIko
[20] http://graphdb.jerteh.rs/sparql?repositoryId=digico

The generation of derived QA datasets using `GPT-4.1` and `GPT-5.1` demonstrates the feasibility of producing concise, semantically faithful question–answer pairs directly from TEI-encoded transcripts. `GPT-5.1`, in particular, balances informativeness and lexical fidelity, making it suitable for downstream RAG applications. The resulting resources thus support multiple pathways of exploitation: semantic search, speaker- and entity-centric retrieval, SPARQL querying over NIF graphs, and evidence-based generation for digital-humanities scholarship.

Future work will expand the annotated dataset, further refine TESLA NER/NEL models, and release the TEI and NIF corpus under permissive open licenses. Planned extensions include (i) domain-sensitive NEL for non-canonical and culturally specific entities; (ii) enhanced RAG pipelines with per-speaker and per-entity evidence retrieval and citation-aware generation; (iii) construction of a richer knowledge graph integrating Wikidata with project-specific ontologies; and (iv) the generation of additional training sets to strengthen the ecosystem of Serbian NLP resources. With these developments, the *DigIko2* pipeline offers not only a replicable methodological blueprint but also a foundation for transparent, verifiable, and linguistically grounded AI applications in less-resourced language settings.

Acknowledgement. This research was supported by the Science Fund of the Republic of Serbia, #7276, *Text Embeddings – Serbian Language Applications – TESLA* and based upon work from COST Action CA23147 GOBLIN - Global Network on Large-Scale, Cross-domain and Multilingual Open Knowledge Graphs, supported by COST (European Cooperation in Science and Technology, https://www.cost.eu).

Literatura

1. Benzécri, J.P.: L'analyse des données, vol. 2. Dunod, Paris (1973)
2. Cetoli, A., Bragaglia, S., O'Harney, A.D., Sloan, M., Akbari, M.: A Neural Approach to Entity Linking on Wikidata. In: Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, Proceedings, Part II 41. pp. 78–86. Springer (2019)
3. Dartmouth Digital History Initiative: DDHI Encoding Guidelines: Applying the Basic Layer of the DDHI Encoding Schema to Oral History Interview Transcripts (2022), https://ddhi.dartmouth.edu/ddhi-encoding-guidelines, accessed: 2025-11-23
4. Davis, R.C.: Synchronizing oral history text and speech: a tools overview. Behavioral & Social Sciences Librarian 34(4), 234–238 (2015)
5. Dinan, E., Roller, S., Shuster, K., et al.: Wizard of Wikipedia: Knowledge-Powered Conversational Agents. In: Proceedings of ICLR (2019)
6. Dong, H., Chen, J., He, Y., Liu, Y., Horrocks, I.: Reveal the Unknown: Out-of-Knowledge-Base Mention Discovery with Entity Linking. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. p. 452–462. CIKM '23, ACM (2023)
7. Fisseni, B., Schmidt, T.: CLARIN web services for TEI-annotated transcripts of spoken language. In: Selected Papers from the CLARIN Annual Conference 2019. Leipzig, 30 September-2 October 2019. pp. 12–22. Linköping University Electronic Press (2019)
8. Francis-Landau, M., Durrett, G., Klein, D.: Capturing Semantic Similarity for Entity Linking with Convolutional Neural Networks. In: Knight et al. (ed.) Conference of North American Chapter of ACL: Human Language Technologies. pp. 1256–1261. ACL (2016)
9. Fu, X., Shi, W., Yu, X., Zhao, Z., Roth, D.: Design Challenges in Low-resource Cross-lingual Entity Linking (2020), https://arxiv.org/abs/2005.00692

10. Guiraud, P.L.: Problèmes et méthodes de la statistique linguistique. Synthese Library, Springer Dordrecht, 1 edn. (1960)
11. Heiden, S.: The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In: Otoguro et al. (ed.) 24th Pacific Asia Conference on Language, Information and Computation. vol. 2, pp. 389–-398. Sendai, Japan (2010)
12. Heiden, S., Pincemin, B., Decorde, M.: Manuel de TXM (Jul 2015), https://halshs.archives-ouvertes.fr/halshs-01223949, manuel d'utilisation du logiciel TXM
13. Hellmann, S., Lehmann, J., Auer, S., Brümmer, M.: Integrating NLP using linked data. In: International semantic web conference. pp. 98–113. Springer (2013)
14. Hoffart, J., Yosef, M.A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., Weikum, G.: Robust Disambiguation of Named Entities in Text. In: Barzilay et al. (ed.) 2011 Conference on Empirical Methods in NLP. pp. 782–792. ACL, Edinburgh, UK. (2011)
15. Ikonić Nešić, M., Petalinkar, S., Stanković, R., Utvić, M., Kitanović, O.: SrpCNNeL: Serbian model for named entity linking. In: 2024 19th Conference on Computer Science and Intelligence Systems (FedCSIS). pp. 465–473. IEEE (2024)
16. Ikonić Nešić, M., Stanković, R., Schöch, C., Škorić, M.: From ELTeC text collection metadata and named entities to linked-data (and back). In: Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th LREC Conference. pp. 7–16 (2022)
17. Ikonić Nešić, M., Utvić, M.: Overview of the Tesla-Ner-Nel-Gold Dataset: Showcase on Serbian-English Parallel Corpus. Technical editors p. 57 (2024)
18. Klie, J.C., Bugert, M., Boullosa, B., de Castilho, R.E., Gurevych, I.: The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In: Proc. of 27th International Conference on Computational Linguistics: System Demonstrations. pp. 5–9. ACL (2018)
19. Krstev, C., Jaćimović, J., Šandrih, B., Stanković, R.: Analysis of the First Serbian Literature Corpus of the Late 19th and Early 20th Century with the TXM Platform. In: DH_BUDAPEST_2019. pp. 36–37. Centre for Digital Humanities - Eötvös Loránd University (2019)
20. Krstev, C., Obradović, I., Utvić, M., Vitas, D.: A system for named entity recognition based on local grammars. Journal of Logic and Computation 24(2), 473–489 (2014)
21. Kubeša, D., Straka, M.: DaMuEL 1.0: A large multilingual dataset for entity linking (2023), http://hdl.handle.net/11234/1-5047, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL)
22. Lebart, L., Salem, A.: Statistique textuelle. Dunod, Paris (1994)
23. Liu, Z., Janowicz, K., Majic, I., Shi, M., Fortacz, A., Karimi, M., Mai, G., Currier, K.: Operationalizing Geographic Diversity for the Evaluation of AI-Generated Content. Transactions in GIS 29(3), e70057 (2025)
24. Muller, C.: Initiation aux méthodes de la statistique linguistique. Honoré Champion, 1 edn. (1992), iSBN-10: 2852032708. Available in DJVU format
25. OpenAI: OpenAI API: ChatGPT-4.1 and ChatGPT-5.1 Models. https://platform.openai.com/docs (2025), models used: gpt-4.1, gpt-5.1
26. Puren, M., Cafiero, F.: InTEIrviews: An ODD for Qualitative Interviews in the Humanities. Journal of the Text Encoding Initiative 2024(15) (2024)
27. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ Questions for Machine Comprehension of Text. In: Proceedings of EMNLP (2016)
28. Reddy, S., Chen, D., Manning, C.D.: CoQA: A Conversational Question Answering Challenge. In: Proceedings of NAACL (2019)
29. Schmid, H.: Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proc. of International Conference on New Methods in Language Processing. pp. 44–49. Manchester, UK (1994)
30. Stanković, R., Krstev, C., Šandrih Todorović, B., Vitas, D., Škorić, M., Ikonić Nešić, M.: Distant Reading in Digital Humanities: Case Study on the Serbian Part of the ELTeC Collection. In: Calzolari et al. (ed.) Proc. of 13th LREC. pp. 3337–3345. ELRA, Marseille, France (2022)

31. Stanković, R., Sandrih, B., Krstev, C., Utvić, M., Škorić, M.: Machine learning and deep neural network-based lemmatization and morphosyntactic tagging for serbian. In: LREC 2020-12th International Conference on Language Resources and Evaluation. pp. 3954–3962. ELRA (2020)
32. Stanković, R., Vučenović, T., Ikonić Nešić, M.: From Unstructured Interviews to Queryable Knowledge: An AI Pipeline for the Digitalne Ikone Corpus using TEI, NER, NEL, and RAG. In: Book of Abstracts of AI Conference 2025. pp. 129–130. SASA, Belgrade, Serbia (2025)
33. TEI Consortium: The TEI Header. In: TEI P5: Guidelines for Electronic Text Encoding and Interchange. TEI Consortium, version 4.10.2 edn. (2025), https://tei-c.org/release/doc/tei-p5-doc/en/html/HD.html, last updated 4 Sept 2025
34. TEI Consortium: Transcriptions of Speech. In: TEI P5: Guidelines for Electronic Text Encoding and Interchange. TEI Consortium, version 4.10.2 edn. (2025), https://tei-c.org/release/doc/tei-p5-doc/en/html/TS.html, last updated 4 Sept 2025
35. Verdonik, D., Dobrovoljc, K., Erjavec, T., Ljubešić, N.: Gos 2: A new reference corpus of spoken slovenian. In: Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024). pp. 7825–7830 (2024)
36. Vrandečić, D., Krötzsch, M.: Wikidata: A Free Collaborative Knowledge Base. Communications of the ACM 57(10), 78–85 (2014)
37. Vučenović, T.: Communication Through Social Media: Blogs and Wikis in Libraries. Qualitative and Quantitative Methods in Libraries 10(3), 315–328 (2021)
38. Vučenović, T.: Digitalne ikone 20+ [Digital Icons 20+]. kademska knjiga; Radio Televizija Beograd - RTS (2024)
39. Wacholder, N., Ravin, Y., Choi, M.: Disambiguation of Proper Names in Text. In: Fifth Conference on Applied Natural Language Processing. pp. 202–208. ACL, Washington, USA (1997)
40. Zhou, S., Rijhwani, S., Wieting, J., Carbonell, J., Neubig, G.: Improving Candidate Generation for Low-resource Cross-lingual Entity Linking. Transactions of the Association for Computational Linguistics 8, 109–124 (01 2020), https://doi.org/10.1162/tacl_a_00303
41. Škoric, M.: New Language Models for Serbian. Infotheca - Journal for Digital Humanities 24(1), 7–28 (2024)
42. Škorić, M., Petalinkar, S.: New XLM-R-based Language Models for Serbian and Serbo-Croatian. In: Artificial Intelligence Conference. pp. 59–60. SASA, Belgrade (2024)

Ranka Stanković is a Full Professor at the University of Belgrade, Faculty of Mining and Geology, Head of Chair for Mathematics and Informatics, vice-president of Language Resources and Technologies Society and . She specializes in informatics, statistics, and natural language processing (NLP). Her research interests span lexical resources, information extraction, semantic web technologies, geoinformatics, and artificial intelligence. She teaches informatics, geoinformatics, GIS, statistics, and programming for linguists.

Tamara Vučenović is an Associate Professor with the Faculty of Management, Metropolitan University, Belgrade, Serbia, and the Editor-in-Chief of the Educational Programme at the Radio Television of Serbia (RTS). She received her degrees from the University of Belgrade. She serves as an international expert and Vice-Chair for the European Commission (EACEA) and the OSCE. She has published numerous scientific papers, including respected venues such as the Language Resources and Evaluation Conference (LREC). Her current research interests include Digital Media, Natural Language Processing, and Artificial Intelligence Applications.

Milica Ikonić Nešić is an Assistant at the Faculty of Philology, University of Belgrade, Department of Librarianship and Information Science, and a PhD student in the Multidisciplinary PhD program, Intelligent Systems module, at the University of Belgrade. Her research interests include mathematics, natural language processing, computational linguistics, the semantic web, machine learning, and deep learning, applied to the development of tools, resources, and models for the Serbian language.

Biljana Rujevic is a Research Associate and librarian at the University of Belgrade, Faculty of Mining and Geology. She received her Ph.D. degree from the University of Belgrade, Faculty of Philology in 2022. Her doctoral research focused on electronic lexicography. Her research interests include electronic lexicography, digital humanities, and language resources.

Received: December 12, 2025; Accepted: May 23, 2026.