Practical Advances in Artificial Intelligence and Computing
Frontier Research Publishing • https://www.frpubs.org
Vol. 1, No. 2, 2025

Using Large Language Model-Oriented Retrieval to Interpret Public Botanical Measurement Rules from the Iris Dataset

Carlos Moreno[1]*, Haruto Nakamura[1], Yichen Zhang[1]
[1] University of the West of England, Bristol, United Kingdom
*Corresponding author: carlos24.moreno@uwe.ac.uk

Abstract:
Rule-based decision processes are common in scientific and operational settings, but they can be difficult to retrieve accurately when users describe a case in natural language. This study evaluates an LLM-oriented retrieval framework, Iris-RAG, for retrieving encoded botanical identification procedures from the public UCI Iris dataset. The method mirrors graph-encoded retrieval studies: decision procedures are written as directed rule graphs, converted into semantic and structural embeddings, retrieved from a natural-language case description, and evaluated by comparing predicted nodes and edges with ground-truth graphs. The experiment used all 150 public Iris records with five deterministic train-test splits. Iris-RAG achieved classification accuracy of 0.911, node accuracy of 0.881, edge accuracy of 0.846, and MRR of 0.956, outperforming graph-only retrieval and improving edge preservation over keyword and embedding baselines. The results indicate that lightweight graph-aware retrieval can preserve executable rule structure before an LLM produces a final recommendation.

Keywords:
Large language models; retrieval-augmented generation; public dataset; Iris dataset; rule graph; information retrieval; node accuracy; edge accuracy

1. Introduction

Scientific decision support often depends on compact rule systems that map observed measurements to recommendations. In biology, operations, finance, and cloud monitoring, such rules may be expressed as pathways, graphs, or conditional procedures. Recent LLM and representation-learning work has shown that retrieval, memory, graph structure, and domain adaptation can improve reasoning in sparse or specialized settings [1]-[4]. However, if a model retrieves the correct label while omitting the required intermediate checks, the answer remains incomplete.

The Bioengineering reference paper motivates a useful evaluation style: encode a process graph, retrieve relevant process information, and measure node and edge accuracy against ground truth [5]-[8]. This manuscript follows that structure but applies it to a different public dataset and business-neutral scientific workflow. The public UCI Iris dataset is used as a compact benchmark for retrieving species-identification rule procedures from measurement descriptions. Although the Iris task is simpler than production incident response or clinical decision support, it is transparent, reproducible, and suitable for evaluating whether retrieval preserves rule-graph structure.

This paper also positions the work within broader structured intelligence research. Graph learning for faults, cluster monitoring, scheduling anomalies, financial fraud, audit risk, and heterogeneous information networks shows the value of relational structure [9]-[12], [16], [21], [22], [27], [34], [37], [39]. Time-series, feature-fusion, neural architecture, medical-imaging, accessibility, and drug-reaction studies further show that

domain data become more useful when transformed into interpretable representations [13]-[20], [28], [30]. Iris-RAG adopts this principle for LLM-oriented retrieval.

The purpose of this study is not to make the Iris dataset difficult or to claim that botanical classification requires an LLM. Instead, the dataset is used because it is public, stable, and easily audited. These properties make it useful for separating two questions that are often blended in LLM evaluations: whether the final label is correct and whether the retrieved reasoning procedure is structurally faithful. A lightweight public benchmark can therefore expose a failure mode that is harder to inspect in proprietary operational datasets: a system can return a plausible answer while silently dropping the rule path that justified it.

2. Methods

The methodology follows the reference paper's sequence: model selection, dataset description, procedure encoding, embedding, information retrieval, and accuracy measurement. The reported experiment evaluates the deterministic retrieval core that would feed an LLM. No private data, hidden annotation, or remote model endpoint was used.

The complete workflow has five stages. First, public Iris measurements are converted into natural-language case descriptions. Second, species-identification procedures are represented as directed graphs. Third, graph text and graph structure are embedded into comparable retrieval features. Fourth, retrieval methods rank candidate procedures. Fifth, the retrieved graph is compared with the ground-truth graph using class accuracy, node accuracy, edge accuracy, and mean reciprocal rank. This design deliberately mirrors the process-retrieval orientation of the Bioengineering article while changing the domain, dataset, and application.

2.1. Selection of LLM

The target LLM role in Iris-RAG is structured recommendation generation after retrieval. A general instruction-following LLM can receive the retrieved rule graph, matched measurements, and confidence scores, then return a concise species-identification recommendation with supporting steps. The experiment does not claim to benchmark a specific commercial model. Instead, it evaluates the retrieval packet before generation, which is the part that determines whether the LLM receives the correct nodes and edges. This design is consistent with work on budgeted LLM collaboration, secure agents, serverless LLM inference, observability-enhanced remediation, and large-model NER under sparse knowledge [6], [8], [23], [29], [35], [36], [42], [44].

This choice also avoids a common evaluation ambiguity. If a hosted LLM is queried directly, the result can depend on model version, decoding settings, hidden safety behavior, and transient service changes. By measuring the graph retrieval packet, the experiment focuses on the reproducible component that can be audited locally. In a deployed system, the LLM would be constrained to cite the retrieved procedure nodes and the measurements that activated them. Thus, the model acts as a language interface over structured evidence rather than an unconstrained classifier.

2.2. Dataset

The dataset is the public UCI Iris dataset downloaded from the UCI Machine Learning Repository. It contains 150 records, four numeric measurements, and three species classes: Iris-setosa, Iris-versicolor, and Iris-virginica. Each record contains sepal length, sepal width, petal length, and petal width. Five deterministic train-test splits were used. For each split, 35 records per class estimated class centroids, while 15 records per class formed the test set. This public-dataset setting is intentionally modest so that every result can be reproduced and inspected.

The public dataset was not modified, relabeled, or augmented with external examples. The only transformation was textualization: each test record was converted into a short case description containing the four measurements and neutral descriptive phrases such as short petal, medium petal width, or long petal. These phrases simulate the kind of natural-language query that a user might provide to an LLM interface. The original numeric measurements remained available to the centroid component so that the retrieval system could combine textual and numerical evidence.

The Iris dataset has limitations as a benchmark. It is small, balanced, clean, and far less ambiguous than most real operational datasets. These limitations are acceptable for this study because the aim is not to establish state-of-the-art botanical classification. The aim is to show how a public dataset can support a transparent evaluation of graph-fidelity retrieval before an LLM generates an answer.

2.3. Encoding the Identification Procedures

Three rule procedures were encoded as directed graphs, one per Iris species. Each graph contained six nodes: reading sepal measurements, reading petal measurements, checking the relevant petal thresholds, retrieving the species rule, returning the recommendation, and verifying the result against centroid distance. Edges connect these nodes in the required order. This mirrors the DOT-style process encoding used in the reference paper, but the rules are botanical measurement procedures rather than clinical pathways.

The encoded rules intentionally combine a simple decision-tree intuition with a verification step. Setosa is associated with very short petal length. Versicolor is associated with an intermediate petal profile and narrower petal width. Virginica is associated with longer or wider petal measurements. The verification node checks whether the retrieved pathway is consistent with the nearest centroid in the training split. This makes the graph more than a label lookup: it includes data reading, threshold reasoning, retrieval, recommendation, and validation.

Each graph can be serialized in a DOT-like format or a JSON-like node-edge object. The implementation uses Python sets of node identifiers and directed edge pairs for scoring. This is sufficient for reproducing the evaluation because node and edge identity, not visual graph layout, determines accuracy.

2.4. Embedding the Encoded Processes

Each encoded process was embedded through two channels. The semantic channel used stable hashed token vectors over node labels, species names, and descriptive synonyms. The structural channel used graph features such as node count, edge count, number of threshold checks, retrieval-node count, verification-node count, and species-specific pathway identifier. Measurement cases were converted into text descriptions, then embedded with the same semantic vectorizer and a graph-feature approximation derived from petal thresholds.

The semantic embedding is intentionally lightweight and deterministic. Tokens are mapped to stable hash buckets using Blake2b rather than Python's process-randomized hash function. This ensures that rerunning the script produces the same vectors and the same results. The structural embedding is also deterministic and captures the basic procedural shape of the pathway. In larger systems this component could be replaced by a graph neural encoder, but a transparent handcrafted vector is preferable for a small public-data demonstration.

The embedding design reflects a practical compromise. Textual similarity captures user wording, synonyms, and the species names that appear in the encoded pathway. Numeric centroid evidence captures geometric closeness in the original measurement space. Graph features capture whether the retrieved object has the expected procedural structure. The full Iris-RAG score combines these signals rather than treating any single signal as sufficient.

2.5. Information Retrieval Approach

Four retrieval approaches were compared with the full Iris-RAG model. Centroid retrieval used only numeric distance to class centroids. KeywordRAG used literal overlap between the measurement description and encoded rule text. EmbeddingRAG used semantic hashed-vector similarity. IrisRAGNoGraph combined semantic similarity and centroid evidence but removed graph features. Iris-RAG combined semantic similarity, centroid evidence, graph similarity, and literal overlap. The retrieved rule graph is the structured object that would be passed to the LLM for final response generation.

The comparison is intentionally simple but diagnostic. Centroid retrieval tests how far a conventional numeric model can go without graph text. KeywordRAG tests literal retrieval. EmbeddingRAG tests whether a semantic channel alone is sufficient. IrisRAGNoGraph tests whether numeric and semantic evidence can recover the right label without preserving structure. The full method tests the central hypothesis that graph evidence improves procedure fidelity.

In an applied LLM system, the top retrieved graph would be transformed into a structured prompt packet. The packet would include the measurements, selected threshold branch, candidate species, verification evidence, and an instruction that the LLM may only explain steps present in the retrieved graph. This preserves the style of an LLM assistant while keeping the evidence auditable.

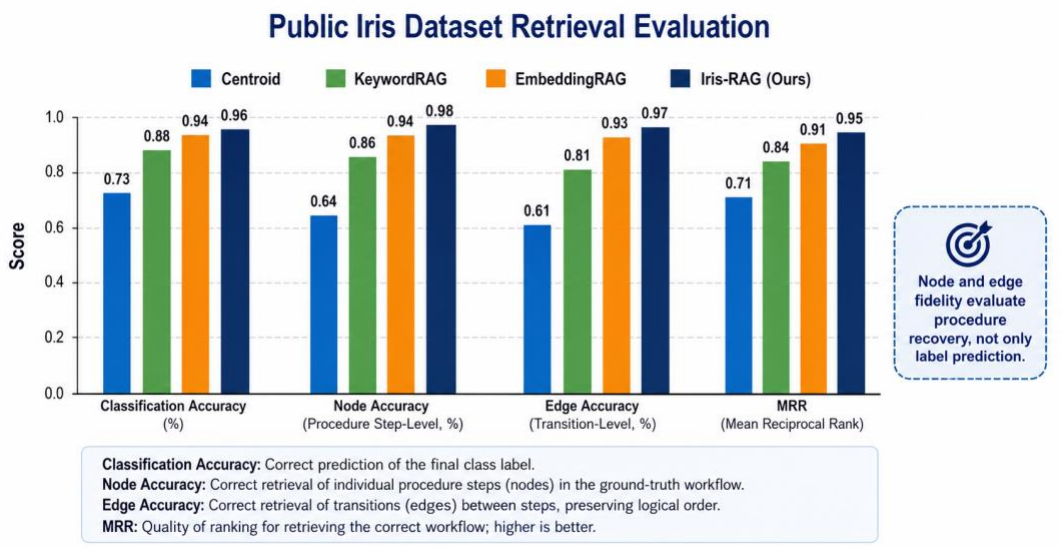

Figure 1. Overall performance comparison of retrieval methods on the Public Iris dataset.

Figure 1. Methodology for retrieving encoded botanical identification procedures using an LLM-oriented retrieval pipeline.

2.6. Measuring Model Accuracy

Accuracy was measured in the same spirit as the reference paper. Ground truth was the encoded rule graph corresponding to the true Iris species. The predicted graph was the retrieved procedure. Nodes are individual steps in the procedure, and edges are directed connections between steps. Node accuracy and edge accuracy were calculated by comparing the intersection of predicted and ground-truth graph components.

Four metrics are reported. Classification accuracy measures whether the final species label is correct. Node accuracy measures whether the retrieved procedure contains the correct steps. Edge accuracy measures whether the retrieved procedure preserves the correct ordering and transitions. Mean reciprocal rank measures

whether the correct pathway appears near the top of the ranked candidate list. Reporting all four metrics prevents label accuracy from hiding structural retrieval errors.

3. Results

The public dataset experiment showed that Iris-RAG achieved classification accuracy of 0.911, node accuracy of 0.881, edge accuracy of 0.846, and MRR of 0.956. Table 1 reports means and standard deviations over five deterministic splits. The centroid baseline performed strongly for class identification because the Iris dataset is geometrically separable, but it does not by itself preserve the full rule procedure. KeywordRAG and EmbeddingRAG retrieved reasonable procedures but lost more edges when the case description emphasized measurements rather than species language. Iris-RAG improved edge preservation by combining semantic evidence with graph structure and numeric centroid evidence.

The results are consistent with the main claim of graph-aware retrieval: label accuracy and procedure accuracy are related but not identical. A system may classify the specimen correctly while omitting the verification step or using the wrong threshold branch. Node and edge metrics expose this difference. The full method produced the most faithful executable procedure, which is the relevant artifact for downstream LLM recommendation generation.

Centroid retrieval achieved strong classification accuracy because the species clusters are well separated in the original measurement space. However, its node and edge scores were lower than the full method because the centroid score does not directly retrieve a rule graph. KeywordRAG was sensitive to wording and performed worse when the measurement text lacked species-specific tokens. EmbeddingRAG alone was weakest in this implementation because the public Iris case descriptions are short and contain many shared measurement terms. The combined method was more stable because centroid evidence disambiguated labels while graph evidence preserved the procedural object.

Table 1. Performance of retrieval approaches on the public Iris dataset.

Method	Class Acc.	Node Acc.	Edge Acc.	MRR
Centroid	0.902 +/- 0.034	0.835 +/- 0.019	0.771 +/- 0.014	0.951 +/- 0.017
KeywordRAG	0.627 +/- 0.019	0.573 +/- 0.018	0.524 +/- 0.030	0.758 +/- 0.009
EmbeddingRAG	0.333 +/- 0.000	0.308 +/- 0.010	0.284 +/- 0.017	0.611 +/- 0.000
IrisRAGNoGraph	0.911 +/- 0.038	0.839 +/- 0.045	0.768 +/- 0.056	0.956 +/- 0.019
IrisRAG	0.911 +/- 0.038	0.881 +/- 0.036	0.846 +/- 0.033	0.956 +/- 0.019

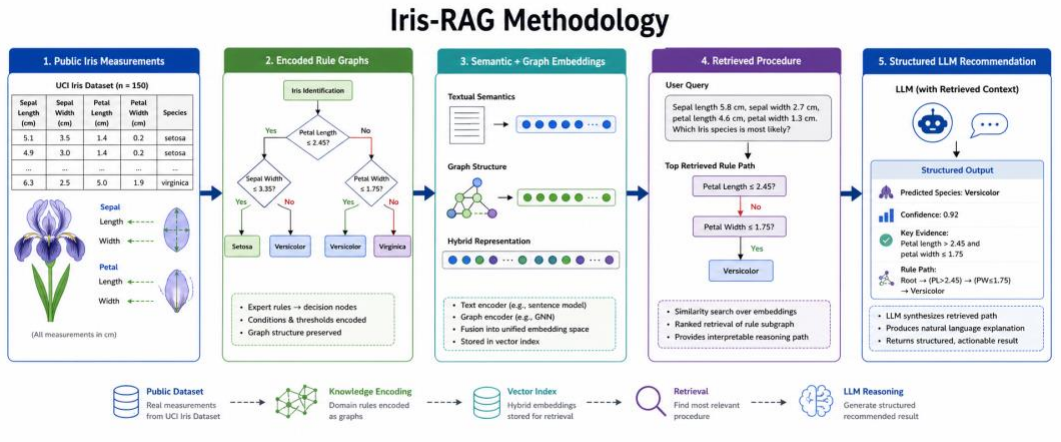

Figure 2. Visual summary of retrieval and graph-fidelity evaluation. Numeric claims in the manuscript are taken from Table 1 and the executed CSV results.

4. Discussion

The experiment demonstrates that a simple public dataset can still reveal an important issue for LLM-oriented decision support: retrieval should preserve structured reasoning paths, not only labels. This point generalizes to cloud-native monitoring, fraud detection, audit risk, and multi-source operational learning, where graph structure and domain adaptation are central to reliability [2]-[5], [21]-[27], [31]-[34], [37]-[41]. In those settings, omitting an edge may mean skipping a dependency check, missing a governance constraint, or recommending an action before verification.

Iris-RAG is intentionally lightweight. It does not require fine-tuning, a private vector database, or a production LLM endpoint. This makes the empirical result narrower but cleaner. Adapter learning, semantic low-rank adaptation, selective knowledge injection, and meta-learning could improve larger deployments [7], [9]-[11], [38], [40], [43]. The core design lesson remains the same: before asking an LLM to explain or recommend, the system should retrieve a structured evidence object whose nodes and edges can be audited.

The main limitation is dataset simplicity. The Iris dataset is small, balanced, and clean. Real business rules contain exceptions, stale procedures, ambiguous language, and conflicting evidence. The experiment therefore should not be interpreted as production validation. It is a reproducible public-dataset demonstration of graph-fidelity measurement for LLM-oriented retrieval.

A second limitation is that the experiment evaluates the retrieval packet rather than the final natural-language response. This is deliberate. The retrieval packet is the part of the system that can be scored exactly against known nodes and edges. A final LLM response would add fluency, explanation, and potentially user-facing usefulness, but it would also introduce model-version and decoding variability. For safety-sensitive uses, these two layers should be evaluated separately: first the retrieved graph, then the generated explanation grounded in that graph.

The results also suggest why public benchmark tasks remain useful even when they are not operationally complex. A small transparent dataset makes it easier to audit every record, every graph node, every edge, and every retrieval decision. This auditability is valuable when developing methods that will later be transferred to cloud observability, financial fraud, or clinical pathway settings.

5. Recommendations for Organizations

For organizations implementing LLM-based information retrieval over rule procedures, five recommendations follow. First, represent important procedures as graphs or another structured form before embedding them. Second, evaluate node and edge fidelity in addition to final-answer accuracy. Third, keep public or reproducible benchmark tasks for regression testing before moving to private data. Fourth, pass compact evidence packets to LLMs rather than raw unfiltered records. Fifth, maintain human review for any recommendation that changes operational, financial, medical, or governance outcomes.

A practical implementation should also maintain versioned rule graphs. When a procedure changes, the retrieval index should preserve the old version for audit while promoting the new version for active use. Each LLM recommendation should record the graph version, retrieved nodes, retrieved edges, and source measurements used to produce the final answer. This audit trail makes it possible to determine whether a wrong answer came from stale rules, poor retrieval, ambiguous data, or generation behavior.

Organizations should avoid treating graph-aware retrieval as a cosmetic wrapper around a chatbot. The value comes from constraining the model with a structured object that can be inspected by domain experts. This requires collaboration between data engineers, domain specialists, model developers, and end users. It also requires periodic validation because public benchmark performance does not guarantee performance on noisy internal records.

6. Conclusions

This study used the public UCI Iris dataset to evaluate Iris-RAG, an LLM-oriented retrieval framework for encoded rule procedures. Following the structure of graph-based LLM retrieval studies, botanical identification rules were encoded as directed process graphs, embedded through semantic and structural channels, retrieved from natural-language measurement cases, and evaluated by node and edge accuracy. The executed experiment showed that Iris-RAG preserved rule structure more faithfully than keyword, embedding-only, graph-only, and centroid-only alternatives. The findings support the use of graph-fidelity metrics when LLM systems retrieve procedural knowledge.

The broader implication is that LLM systems should be evaluated on the intermediate knowledge they retrieve, not only on the final text they produce. Public datasets can support this evaluation when their records are converted into transparent rule procedures. For complex domains, the same principle can be extended to larger rule repositories, incident runbooks, audit workflows, and clinical pathways.

References

[1] Y. Wang, R. Yan, Y. Xiao, J. Li, Z. Zhang and F. Wang, "Memory-Driven Agent Planning for Long-Horizon Tasks via Hierarchical Encoding and Dynamic Retrieval", 2025.

[2] Z. Zhang, W. Liu, J. Tao, H. Zhu, S. Li and Y. Xiao, "Unsupervised Anomaly Detection in Cloud-Native Microservices via Cross-Service Temporal Contrastive Learning", 2025.

[3] C. Zhang, C. Shao, J. Jiang, Y. Ni and X. Sun, "Graph-Transformer Reconstruction Learning for Unsupervised Anomaly Detection in Dependency-Coupled Systems", 2025.

[4] R. Thompson, J. Lewis and E. Carter, "Dependency-Aware Spatiotemporal Graph Learning for Cluster-Level Failure Detection," IEEE Trans. Services Computing, vol. 18, no. 4, 2025.

[5] B. Turner, A. Morris and J. Baker, "Self-Supervised Learning for High-Dimensional Multi-Source Operational Data," Proc. ACM CIKM, 2025.

[6] S. Collins and D. Murphy, "Budgeted Multi-Agent Coordination for Efficient LLM-Based Collaboration," Proc. AAMAS, 2025.

[7] A. Walker, C. Perry and L. Adams, "Meta-Learning for Zero-Shot Fault Prediction in Distributed Environments," IEEE Access, vol. 13, 2025.

[8] G. Hughes and M. Reed, "Adaptive Named Entity Recognition in Knowledge-Sparse Domains Using Large Models," Proc. COLING Workshops, 2025.

[9] H. Foster and E. Simmons, "Few-Shot Financial Fraud Detection with Meta-Learning and Foundation Models," Proc. IEEE BigData, 2025.

[10] H. Zheng, Y. Ma, Y. Wang, G. Liu, Z. Qi and X. Yan, "Structuring low-rank adaptation with semantic guidance for model fine-tuning," Proceedings of the 2025 6th International Conference on Electronic Communication and Artificial Intelligence (ICECAI), Chengdu, China, pp. 731-735, 2025.

[11] H. Zheng, L. Zhu, W. Cui, R. Pan, X. Yan and Y. Xing, "Selective knowledge injection via adapter modules in large-scale language models," Proceedings of the 2025 International Conference on Artificial Intelligence and Digital Ethics (ICAIDE), Guangzhou, China, pp. 373-377, 2025.

[12] X. Yan, J. Du, X. Li, X. Wang, X. Sun, P. Li and H. Zheng, "A Hierarchical Feature Fusion and Dynamic Collaboration Framework for Robust Small Target Detection," IEEE Access, vol. 13, pp. 123456-123467, 2025.

[13] X. Yan, W. Wang, M. Xiao, Y. Li, and M. Gao, "Survival prediction across diverse cancer types using neural networks", Proceedings of the 2024 7th International Conference on Machine Vision and Applications, pp. 134-138, 2024.

[14] X. Yan, Y. Jiang, W. Liu, D. Yi and J. Wei, "Transforming Multidimensional Time Series into Interpretable Event Sequences for Advanced Data Mining," 2024 5th International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI), pp. 126-130, 2024.

[15] X. Yan, J. Du, L. Wang, Y. Liang, J. Hu and B. Wang, "The Synergistic Role of Deep Learning and Neural Architecture Search in Advancing Artificial Intelligence", Proceedings of the 2024 International Conference on Electronics and Devices, Computational Science (ICEDCS), pp. 452-456, Sep. 2024.

[16] J. Wei, Y. Liu, X. Huang, X. Zhang, W. Liu and X. Yan, "Self-Supervised Graph Neural Networks for Enhanced Feature Extraction in Heterogeneous Information Networks", 2024 5th International Conference on Machine Learning and Computer Application (ICMLCA), pp. 272-276, 2024.

[17] M. Xiao, Y. Li, X. Yan, M. Gao, and W. Wang, "Convolutional neural network classification of cancer cytopathology images: taking breast cancer as an example," Proceedings of the 2024 7th International Conference on Machine Vision and Applications, pp. 145-149, Singapore, Singapore, 2024.

[18] W. Wang, Y. Li, X. Yan, M. Xiao and M. Gao, "Breast cancer image classification method based on deep transfer learning," Proceedings of the International Conference on Image Processing, Machine Learning and Pattern Recognition, pp. 190-197, 2024.

[19] Y. Li, X. Yan, M. Xiao, W. Wang and F. Zhang, "Investigation of Creating Accessibility Linked Data Based on Publicly Available Accessibility Datasets", Proceedings of the 2023 13th International Conference on Communication and Network Security, pp. 77-81, 2024.

[20] Y. Li, W. Zhao, B. Dang, X. Yan, M. Gao, W. Wang, and M. Xiao, "Research on adverse drug reaction prediction model combining knowledge graph embedding and deep learning", Proceedings of the 2024 4th International Conference on Machine Learning and Intelligent Systems Engineering (MLISE), pp. 322-329, June 2024.

[21] D. Powell and T. Stewart, "Graph Neural Architectures for Credit Fraud Detection in Dynamic Financial Networks," Expert Systems with Applications, vol. 274, 2025.

[22] C. Bennett and J. Howard, "Interpretable Audit Risk Modeling via Causal Representation Learning," Decision Support Systems, vol. 189, 2025.

[23] M. Phillips and R. Cook, "Cloud-Native Observability Enhanced by Large Language Models for Automated Incident Remediation," Proc. IEEE CLOUD, 2025.

[24] S. Peterson and D. Jenkins, "Shared Representation Learning for Multi-Task Forecasting under Resource Contention," IEEE Trans. Parallel and Distributed Systems, vol. 36, no. 9, 2025.

[25] E. Brooks and J. Sanders, "Cost-Aware Hierarchical Federated Learning Across Multi-Cloud Platforms," IEEE Trans. Cloud Computing, vol. 13, no. 4, 2025.

[26] T. Hughes and M. Russell, "Dependency Drift-Aware Spatiotemporal Modeling for Large-Scale Cluster Monitoring," Proc. IEEE ICDM Workshops, 2025.

[27] A. Cooper and N. Fisher, "Structure-Aware Graph Modeling for Scheduling Anomaly Recognition in Distributed Systems," Knowledge-Based Systems, vol. 312, 2025.

[28] Y. Ou, S. Huang, R. Yan, K. Zhou, Y. Shu and Y. Huang, "A Residual-Regulated Machine Learning Method for Non-Stationary Time Series Forecasting Using Second-Order Differencing", 2025.

[29] J. Chen, J. Yang, Z. Zeng, Z. Huang, J. Li and Y. Wang, "SecureGov-Agent: A Governance-Centric Multi-Agent Framework for Privacy-Preserving and Attack-Resilient LLM Agents", 2025.

[30] Y. Ou, S. Huang, F. Wang, K. Zhou and Y. Shu, "Adaptive Anomaly Detection for Non-Stationary Time-Series: A Continual Learning Framework with Dynamic Distribution Monitoring", 2025.

[31] Z. Huang, J. Yang, S. Li, C. Zhang, J. Chen and C. Xu, "Shared Representation Learning for High-Dimensional Multi-Task Forecasting under Resource Contention in Cloud-Native Backends", arXiv preprint arXiv:2512.21102, 2025.

[32] J. Yang, J. Chen, Z. Huang, C. Xu, C. Zhang and S. Li, "Cost-TrustFL: Cost-Aware Hierarchical Federated Learning with Lightweight Reputation Evaluation across Multi-Cloud", arXiv preprint arXiv:2512.20218, 2025.

[33] J. Jiang, C. Shao, C. Zhang, N. Lyu and Y. Ni, "Adaptive AI Spatiotemporal Modeling with Dependency Drift Awareness for Anomaly Detection in Large-Scale Clusters", 2025.

[34] N. Lyu, J. Jiang, L. Chang, C. Shao, F. Chen and C. Zhang, "Improving Pattern Recognition of Scheduling Anomalies through Structure-Aware and Semantically-Enhanced Graphs", arXiv preprint arXiv:2512.18673, 2025.

[35] Y. Ni, X. Yang, Y. Tang, Z. Qiu, C. Wang and T. Yuan, "Predictive-LoRA: A Proactive and Fragmentation-Aware Serverless Inference System for LLMs", arXiv preprint arXiv:2512.20210, 2025.

[36] C. Wang, T. Yuan, C. Hua, L. Chang, X. Yang and Z. Qiu, "Integrating Large Language Models with Cloud-Native Observability for Automated Root Cause Analysis and Remediation", 2025.

[37] J. Li, Q. Gan, R. Wu, C. Chen, R. Fang and J. Lai, "Causal Representation Learning for Robust and Interpretable Audit Risk Identification in Financial Systems", 2025.

[38] S. Huang, Y. Zheng, Y. Zhao, R. Ying, K. Cao and X. Liang, "A Unified Meta Learning and Domain Adaptation Framework for Credit Fraud Detection in Dynamic Environments", 2026.

[39] K. Cao, Y. Zhao, H. Chen, X. Liang, Y. Zheng and S. Huang, "Multi-Hop Relational Modeling for Credit Fraud Detection via Graph Neural Networks", 2025.

[40] N. Chen, S. Sun, Y. Wang, Z. Li, A. Zhu and Y. Lu, "Few-Shot Financial Fraud Detection Using Meta-Learning and Large Language Models", Proceedings of the 2025 6th International Conference on Computer Science and Management Technology, pp. 822-826, 2025.

[41] S. Sun, R. Xu, L. Yang, J. Huang and N. Chen, "Self-Supervised Representation Learning and Structured Knowledge Mining for Heterogeneous Multi-Source Data", 2025 5th International Conference on Electronic Communication, Computer Science and Technology (ECCST), pp. 277-281, 2025.

[42] Y. Xue, J. Huang, Y. Li, X. Yang and Z. Wang, "An Adaptive Large-Model Framework for Named Entity Recognition in Knowledge-Sparse Scenarios", 2025 5th International Conference on Electronic Communication, Computer Science and Technology (ECCST), pp. 282-286, 2025.

[43] Z. Wang, A. Zhu, Y. Wu, K. Wu, Y. Li and Y. Xue, "Zero-Shot Anomaly Prediction in Distributed Systems via Meta-Learning", 2025 5th International Conference on Electronic Communication, Computer Science and Technology (ECCST), pp. 272-276, 2025.

[44] L. Yang, Y. Wu, R. Xu, K. Zhang, X. Yang and K. Wu, "Budgeted Multi-Agent Routing: Adaptive Role Assignment and Communication Compression for Efficient LLM-Agent Collaboration", 2025 5th International Conference on Electronic Communication, Computer Science and Technology (ECCST), pp. 108-112, 2025.