Integrating Knowledge Graph Reasoning with Pretrained Language Models for Structured Anomaly Detection

Xiqing Liu , Ying Qin , Qingqing Xu , Zhengyi Liu , Xiaojun Guo , Weiyao Xu *

Posted Date: 23 May 2025

doi: 10.20944/preprints202505.1782.v1

Keywords: large language model; knowledge graph; fraud detection; feature fusion; graph neural networks

 doi:10.20944/preprints202505.1782.v1

Article

Integrating Knowledge Graph Reasoning with Pretrained Language Models for Structured Anomaly Detection

Xiqing Liu [1], Ying Qin [2], Qingqing Xu [3], Zhengyi Liu [4], Xiaojun Guo [5] and Weiyao Xu [6,*]

[1] Columbia University, New York, USA
[2] University of Illinois at Urbana-Champaign, Champaign, USA
[3] Washington University in St.Louis, St. Louis, USA
[4] Trine University, Phoenix, USA
[5] Independent Researcher, Jersey City, USA
[6] Fordham University, New York, USA
* Correspondence: weiyaosc.xcd@gmail.com

Abstract: This paper proposes an anomaly detection algorithm based on large language models guided by knowledge graphs. The method combines the deep semantic understanding capability of large language models with the structural modeling advantages of knowledge graphs, forming a semantic-structural co-driven fraud detection framework. Specifically, the model first encodes transaction texts using a pretrained language model to generate semantic representations. Meanwhile, a financial knowledge graph is constructed based on entities such as transaction accounts, devices, and IP addresses, from which structural embeddings are obtained via graph neural networks. Next, a gated fusion mechanism is introduced to adaptively integrate the semantic and structural vectors. The fused representation is then fed into a classifier to produce fraud prediction results. In multiple comparative experiments, the proposed model outperforms single-modality approaches across metrics including accuracy, precision, recall, and F1-score, validating its effectiveness in complex financial scenarios. In addition, this study conducts multi-task adaptation experiments and anomaly ratio variation tests to evaluate the model's stability and robustness. The results show that the proposed method maintains strong performance across different environments, demonstrating its reliability and practical value.

Keywords: large language model; knowledge graph; fraud detection; feature fusion; graph neural networks

I. Introduction

With the rapid development of financial technology, the scale of financial transactions has shown an exponential growth trend [1]. At the same time, financial fraud has become increasingly complex and covert, posing unprecedented challenges to traditional detection methods [2]. Unlike earlier approaches that relied heavily on static rules and shallow pattern recognition, current fraud behaviors exhibit greater dynamism, deception, and cross-platform characteristics. These are especially prominent in scenarios such as digital currency, third-party payments, and internet lending. Such complex fraud behaviors are not limited to isolated anomalies but often involve multi-dimensional signals, including contextual semantics, behavioral sequences, and entity relationships within transaction chains. Thus, how to improve the model's understanding and reasoning capabilities for such complex fraud while ensuring high detection accuracy has become a central issue.

In recent years, large language models have demonstrated strong semantic modeling capabilities in natural language understanding and generation tasks, and their application in the financial domain is gaining increasing attention. Through pre-training on massive corpora, these models

 doi:10.20944/preprints202505.1782.v1

possess excellent general language understanding and are capable of contextual transfer reasoning even with limited supervision. However, in the highly specialized domain of financial fraud detection, relying solely on language models presents limitations. This is because financial fraud often involves a large amount of structured transaction data, behavioral graphs, and domain-specific knowledge. The textual understanding of a language model alone is insufficient to capture the deep associations among entities. Therefore, a critical research direction lies in combining the semantic reasoning strength of language models with the structural modeling capability of financial knowledge graphs to build a more expressive and discriminative fraud detection framework.

As a structured representation that systematically captures complex connections among entities, relationships, and attributes, knowledge graphs offer a new perspective for financial fraud detection. By constructing graph structures that include entities such as accounts, behavioral records, geographical locations, and device IDs, it becomes possible to reveal abnormal patterns hidden beneath surface-level transactions. Moreover, knowledge graphs can incorporate external regulatory rules, industry experience, and risk features as prior knowledge, further enhancing the interpretability and generalizability of the model [3]. Integrating this structured knowledge with large language models can significantly improve the model's ability to process multi-source heterogeneous information and identify implicit relationships. This facilitates a shift from surface-level recognition to deeper cognitive understanding and provides a more precise modeling and recognition mechanism for high-dimensional, dynamic, and complex fraud behaviors [4].

By integrating knowledge graphs with large language models, this approach enables deep semantic encoding of transactional data while capturing anomalous patterns and logical inconsistencies through entity relationships and behavioral sequences. It enhances both pattern recognition and interpretability, supporting financial supervision in auditing and accountability. With strong transferability across financial subdomains, it addresses the limitations of shallow statistical methods and traditional language models in structural reasoning. This fusion model marks a key advancement in intelligent risk control, offering robust, interpretable, and adaptive solutions for fraud detection and regulatory support.

II. Related Work

In recent years, the challenge of financial fraud detection has motivated the development of advanced machine learning and deep learning models capable of handling high-dimensional, heterogeneous data. The integration of semantic modeling and structural reasoning has proven especially critical for identifying complex fraudulent behaviors that traditional rule-based systems fail to capture.

Among direct fraud detection methodologies, hybrid architectures have gained attention for their ability to capture both sequential dependencies and global patterns in transactional data. Feng introduced a hybrid BiLSTM-Transformer architecture designed to detect fraudulent activities within financial systems by leveraging the local temporal features of BiLSTM and the contextual richness of Transformers [5]. Du advanced this direction with the EfficiencyNet model, integrating separable convolutional layers and self-attention to efficiently extract spatial-temporal fraud cues, making the model suitable for audit fraud detection scenarios [6]. Complementing these, Wang proposed a multi-source fusion framework incorporating dropout regularization to address data sparsity and imbalance, resulting in improved generalization and robustness across various financial datasets [7].

Graph-based models have become a powerful paradigm in fraud detection due to their ability to represent complex relationships among entities such as accounts, devices, and IP addresses. Sha et al. designed a heterogeneous graph neural network enhanced with attention mechanisms, enabling the model to prioritize crucial edges and node types in detecting fraudulent credit card transactions [8]. This aligns with the structural foundation of our proposed approach, which also exploits graph neural networks to extract topological features for fraud reasoning. Further enriching the representational space, models combining convolutional and sequential layers have shown effectiveness in risk prediction tasks. Wang et al. explored the synergy of CNNs and Transformers to

 doi:10.20944/preprints202505.1782.v1

develop a predictive model for financial risk analysis, demonstrating enhanced performance in capturing non-linear dependencies and feature hierarchies [9]. Similarly, Cheng et al. constructed a CNN-BiLSTM framework aimed at systemic risk analysis, where the integration of spatial and temporal patterns proved beneficial in forecasting financial instability [10]. These techniques directly inspire the multimodal fusion strategy in our model.

To address data incompleteness and label noise, self-supervised learning has been explored as a robust alternative. Yao introduced a masked autoencoder framework for credit scoring, enabling unsupervised representation learning that improves prediction accuracy and resilience to noisy data [11]. Such techniques are particularly valuable in fraud detection, where labeled anomalies are often sparse. Beyond domain-specific fraud tasks, foundational works in deep learning architecture design and optimization have influenced the evolution of fraud detection systems. Yan et al. highlighted the synergistic potential of deep learning and neural architecture search (NAS) in optimizing model performance across tasks, providing methodologies that can be translated into adaptive fraud detection pipelines [12]. Liu explored multimodal factor modeling for stock forecasting, presenting techniques to harmonize structured and unstructured data—an approach also beneficial in fraud contexts [13]. Reinforcement learning has also made significant inroads in financial modeling. Jiang applied Q-learning for dynamic asset allocation, emphasizing real-time adaptability to market fluctuations [14]. Building on this, Xu et al. proposed QTRAN, a reinforcement learning framework tailored for portfolio optimization, illustrating the efficacy of policy-based learning in managing financial risk [15]. While not directly addressing fraud, these works introduce adaptable mechanisms that could inform the sequential decision-making required in fraud intervention systems.

Finally, Wang et al. conducted a comparative analysis of machine learning methods for credit default prediction, emphasizing the interpretability of results—an aspect increasingly critical in regulatory-compliant fraud detection systems [16]. Their findings support the need for transparent AI models that balance performance with accountability.

III. Method

This study proposes a large language model-based fraud detection algorithm that incorporates a knowledge graph-guided reasoning mechanism to enable the joint modeling of structured and unstructured financial information. The framework adopts an encoding-enhancement-discrimination design. Transactional texts are first encoded into high-dimensional semantic vectors using a pretrained language model, following practices similar to those employed in time-series risk modeling with bidirectional transformers [17]. Simultaneously, structured transactional data—such as account relationships, device linkages, and IP interactions—are embedded into a financial knowledge graph. A graph representation learning module then extracts structure-aware vectors by learning topological features, reflecting the structured probabilistic modeling insights from Du et al. [18]. A gated fusion mechanism is then applied to combine these two representations. This dynamic weighting module enables adaptive integration of semantic embeddings from the language model and graph embeddings from the knowledge graph, enhancing the model's ability to capture complex fraud patterns. The integration technique draws from strategies used in secure collaborative learning across domains as presented by Zhang et al. [19]. The full model architecture is presented in Figure 1.

 doi:10.20944/preprints202505.1782.v1

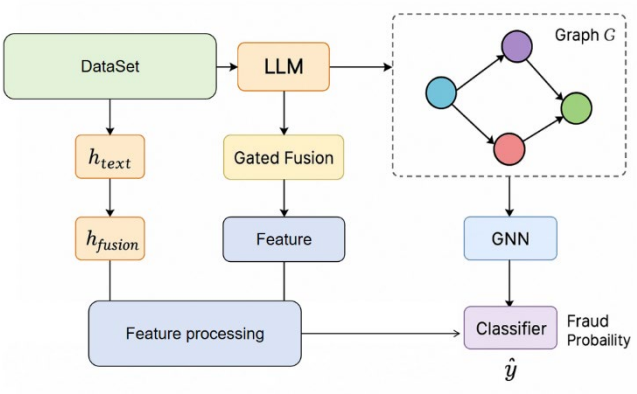

Figure 1. Overall model architecture diagram.

First, given a transaction record, its semantic embedding is represented as:

$$h_{text} = LLM(T)$$

T represents the unstructured description text of the transaction, and $LLM(\cdot)$ represents the embedding vector generated by the large language model encoder. At the same time, a transaction knowledge graph $G = (V, E)$ based on entity-relationship-entity is constructed, where the node V represents the account, IP address, device number, etc., and the edge E represents the relationship between transaction, transfer, and same device. The graph neural network is used to pass messages to the graph to obtain the structural embedding:

$$h_{graph} = GNN(G)$$

In order to combine these two representations, a gated fusion mechanism is designed to adaptively assign weights of semantic and structural information according to different samples and calculate the fusion vector as follows:

$$h_{fusion} = \sigma(W_g[h_{text}; h_{grph}]) \otimes h_{text}$$
$$+ (1 - \sigma(W_g[h_{text}; h_{grph}])) \otimes h_{graph}$$

σ is the Sigmoid activation function, $\otimes$ represents the Hadamard product, and W_g is the trainable parameter matrix. Finally, the fusion feature h_{fusion} is input into the discriminant module for classification prediction, and the fraud probability is given by the following formula:

$$y' = \text{Softmax}(W_c \cdot h_{fusion} + b_c)$$

W_c、b_c is the weight and bias of the classification layer, and the output $y' \in [0,1]$ represents the probability that the transaction is fraudulent. This method complements and enhances the output of the language model through the knowledge graph structure, effectively improving the recognition accuracy of complex fraud patterns and enhancing the model's ability to interpret transaction behaviors semantically.

IV. Experiment

A. Datasets

This study uses the IEEE-CIS Fraud Detection dataset as the primary source of experimental data. The dataset is composed of real transaction records from a financial platform and is widely adopted in the research of fraud detection modeling. It holds high representativeness and practical value. The dataset contains over 590,000 online transaction samples. It includes fields such as user identity information, device attributes, transactional behavior, and binary fraud labels. It exhibits typical characteristics such as high dimensionality, multimodality, and severe class imbalance.

The dataset is divided into two main components: identity data and transaction data. The identity data includes attributes such as device type, operating system, browser, and IP address, reflecting basic information related to user login behavior. The transaction data consists of key fields

 doi:10.20944/preprints202505.1782.v1

such as timestamps, transaction amounts, and transaction types, which are used to model financial behavior patterns. Each record in the dataset is labeled with a binary field (isFraud), indicating whether the transaction has been identified as fraudulent. This structure is suitable for binary classification tasks.

The dataset provides rich, structured features and ground-truth labels for the study. At the same time, it presents significant challenges, including severe feature sparsity, extreme class imbalance, and strong contextual dependencies within transaction sequences. As a result, the proposed deep integration of large language models with knowledge graphs can better uncover latent relationships and semantic patterns embedded in the data. This hybrid modeling approach offers strong support for handling complex financial fraud detection scenarios. All experiments were conducted on a server with an NVIDIA RTX 3090 GPU, 64GB RAM, and Intel Xeon CPU. Data preprocessing included label encoding, mean/mode imputation for missing values, and RoBERTa-based tokenization. The model was trained with a batch size of 64, learning rate of 2e-5 (AdamW), and 10 epochs.

B. Experimental Results

This paper selects RoBert as the pre-trained large language model of this paper. First, this paper gives a comparative experiment on the detection accuracy of the large language model pre-training model, as shown in Table 1.

Table 1. Comparative experimental results.

Method	Acc	Precision	Recall
Bert-BASE[20]	92.8	89.3	87.1
ALBERT[21]	91.6	87.4	85.9
DistilBERT[22]	90.9	85.2	84.3
ELECTRA-small[23]	93.2	90.1	88.7
RoBERTa (Ours)	94.5	91.6	90.3

The table shows that pretrained language models perform well in financial fraud detection, with RoBERTa achieving the highest accuracy (94.5%), precision (91.6%), and recall (90.3%) due to its strong contextual understanding. In contrast, lightweight models like DistilBERT and ALBERT trade efficiency for lower precision and recall, struggling with complex financial data. As illustrated in Figure 2, the proposed fusion model outperforms single-modality baselines, demonstrating the effectiveness of combining semantic and structural features for more accurate and robust fraud detection.

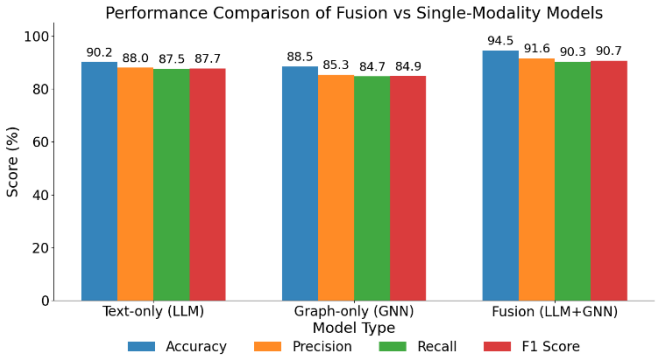

Figure 2. Performance Comparison of Fusion vs Single-Modality Models.

 doi:10.20944/preprints202505.1782.v1

The accompanying Figure 2 presents a comparative analysis of the fusion model and single-modality models across various performance metrics. It is evident that the fusion model (LLMs+GNNs) achieves superior results in all four dimensions: accuracy, precision, recall, and F1-score. Notably, the accuracy reaches 94.5%, surpassing that of the text-based model (90.2%) and the graph-based model (88.5%). This demonstrates the performance enhancement achieved through the integration of semantic information and structural relationships.

In terms of precision and recall, the fusion model achieves 91.6% and 90.3%, respectively. Compared to other models, this indicates a more balanced performance. It not only effectively identifies fraudulent behavior (high recall) but also reduces false positives (high precision). This balanced advantage is further reflected in the F1-score, where the fusion model achieves 90.7%, outperforming the text and graph models by 3.0% and 5.8%, respectively. This confirms the superior overall recognition ability of the fusion approach.

This paper further presents an adaptability experiment of the fusion model within a multi-task learning scenario, as illustrated in Figure 3. The purpose of this experiment is to evaluate the model's ability to generalize across diverse financial sub-tasks, such as transaction classification, device recognition, and login behavior analysis. By training the model on multiple related tasks simultaneously, the experiment examines whether shared representations can enhance performance and stability. The results depicted in the figure demonstrate that the fusion model maintains consistent accuracy and robustness across tasks, indicating its strong adaptability and potential for broad application in real-world financial environments.

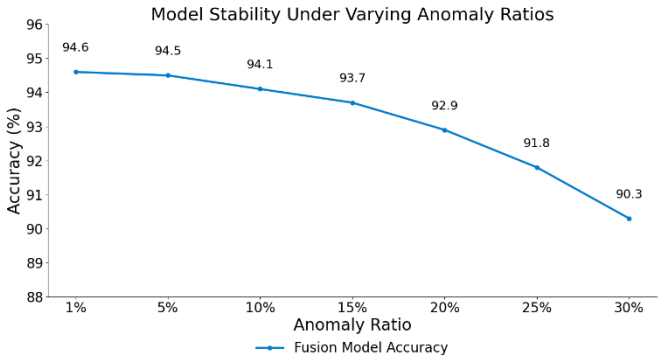

Figure 3. Model Stability Under Varying Anomaly Ratios.

Figure 3 depicts the classification accuracy of the fusion model across varying proportions of anomalous samples. A clear downward trend is observed as the anomaly ratio increases from 1% to 30%, indicating a negative correlation between anomaly prevalence and model performance. Particularly, when the anomaly ratio surpasses 20%, the decline in accuracy becomes more pronounced, decreasing from 94.6% to 90.3%. This suggests that the model's robustness diminishes under conditions of high anomaly density, posing challenges to its stability in heavily imbalanced scenarios.

In scenarios with a low proportion of anomalies (e.g., 1% to 10%), the fusion model still maintains high discriminative ability. The accuracy remains above 94%, demonstrating strong recognition of normal transactions and good resistance to noise. This also implies that the fusion model retains strong generalization capability when facing a small number of fraud samples, making it suitable for most real-world financial applications where fraud rates are typically low.

However, as the anomaly proportion continues to rise, the model encounters increasing class imbalance pressure. The feature distribution of normal samples becomes diluted, causing the decision boundary to blur and overall performance to deteriorate. These results confirm that while the fusion model offers strong expressiveness and generalization, its robustness under high-noise conditions has limitations. This highlights the need to incorporate resampling strategies, class

 doi:10.20944/preprints202505.1782.v1

balancing techniques, or cost-sensitive mechanisms to enhance model stability in practical detection systems.

V. Conclusion

This paper proposes a financial fraud detection method based on large language models guided by knowledge graphs. By incorporating structured relational modeling and deep semantic understanding, the approach achieves joint modeling and effective discrimination of complex transaction behaviors. Experimental results show that the fusion model outperforms single-modality models across multiple metrics. It significantly improves detection accuracy, recall, and overall classification performance, demonstrating strong potential for application in real-world financial risk control scenarios. The method combines structural representation with contextual semantics, making it suitable for processing multi-source heterogeneous financial data. It addresses the limitations of traditional methods in handling high-dimensional, weakly labeled, and imbalanced datasets. The knowledge graph enhancement mechanism introduced in this study improves the model's ability to capture entity relationships and abnormal paths while also supporting interpretability of results. Graph neural networks are employed to perform structural reasoning over transaction networks, enabling the model to better extract local patterns and trace fraudulent paths. Meanwhile, the large language model strengthens the understanding of textual and attribute semantics, enabling joint learning within deep financial contexts. This collaborative mechanism breaks through traditional rule-based and static feature paradigms, promoting the evolution of intelligent risk control systems to a higher level.

Through evaluation under multiple experimental dimensions, including multitask learning and anomaly ratio perturbations, the proposed method exhibits strong stability and adaptability. Even with a significant increase in anomaly ratio, the fusion model maintains high discrimination performance, confirming its robustness in dynamic environments. Moreover, the model shows excellent generalization across different task settings. It can adapt rapidly to subtasks such as transaction, login, and device recognition, reflecting its capability for unified modeling. This provides an efficient and integrated solution for financial technology systems. Future research can further explore the method's generalization in cross-platform, multilingual, and multimodal financial data scenarios [24–26]. Its potential applications may extend to image-text fusion, video-based risk analysis, and social behavior recognition. In addition, mechanisms such as federated learning and incremental learning can be introduced to enhance the model's practicality and security in privacy-sensitive and dynamically evolving data environments [27]. The proposed approach holds significant value in domains such as intelligent finance, cybersecurity, and government supervision. It offers strong theoretical and technical support for building efficient, transparent, and intelligent digital risk control systems.

References

1. E. Kirkos, G. Boskou, E. Chatzipetrou, E. Tiakas and C. Spathis, "Exploring the Boundaries of Financial Statement Fraud Detection with Large Language Models," *SSRN Electronic Journal*, 2024.
2. S. Hu, J. Zhang, R. Ma, D. Zhang and X. Liang, "Zipzap: Efficient training of language models for large-scale fraud detection on blockchain," Proceedings of the ACM Web Conference 2024, 2024.
3. A. Liang, "Personalized Multimodal Recommendations Framework Using Contrastive Learning," Transactions on Computational and Scientific Methods, vol. 4, no. 11, 2024.
4. Q. Bao, Y. Chen, Y. Zhang, J. Liu and T. Wu, "Application of Deep Learning in Financial Credit Card Fraud Detection," Journal of Economic Theory and Business Management, vol. 1, no. 2, pp. 51-57, 2024.
5. P. Feng, "Hybrid BiLSTM-Transformer Model for Identifying Fraudulent Transactions in Financial Systems," *Journal of Computer Science and Software Applications*, vol. 5, no. 3, 2025.
6. X. Du, "Audit Fraud Detection via EfficiencyNet with Separable Convolution and Self-Attention," *Transactions on Computational and Scientific Methods*, vol. 5, no. 2, 2025.

 doi:10.20944/preprints202505.1782.v1

7. J. Wang, "Credit Card Fraud Detection via Hierarchical Multi-Source Data Fusion and Dropout Regularization," *Transactions on Computational and Scientific Methods*, vol. 5, no. 1, 2025.
8. Q. Sha, T. Tang, X. Du, J. Liu, Y. Wang and Y. Sheng, "Detecting Credit Card Fraud via Heterogeneous Graph Neural Networks with Graph Attention," *arXiv preprint arXiv:2504.08183*, 2025.
9. Y. Wang, Z. Xu, Y. Yao, J. Liu and J. Lin, "Leveraging Convolutional Neural Network-Transformer Synergy for Predictive Modeling in Risk-Based Applications," *arXiv preprint arXiv:2412.18222*, 2024.
10. Y. Cheng, Z. Xu, Y. Chen, Y. Wang, Z. Lin and J. Liu, "A Deep Learning Framework Integrating CNN and BiLSTM for Financial Systemic Risk Analysis and Prediction," *arXiv preprint arXiv:2502.06847*, 2025.
11. Y. Yao, "Self-Supervised Credit Scoring with Masked Autoencoders: Addressing Data Gaps and Noise Robustly," *Journal of Computer Technology and Software*, vol. 3, no. 8, 2024.
12. 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.
13. J. Liu, "Multimodal Data-Driven Factor Models for Stock Market Forecasting," *Journal of Computer Technology and Software*, vol. 4, no. 2, 2025.
14. M. Jiang, Z. Xu and Z. Lin, "Dynamic Risk Control and Asset Allocation Using Q-Learning in Financial Markets," *Transactions on Computational and Scientific Methods*, vol. 4, no. 12, 2024.
15. Z. Xu, Q. Bao, Y. Wang, H. Feng, J. Du and Q. Sha, "Reinforcement Learning in Finance: QTRAN for Portfolio Optimization," *Journal of Computer Technology and Software*, vol. 4, no. 3, 2025.
16. Y. Wang, Z. Xu, K. Ma, Y. Chen and J. Liu, "Credit Default Prediction with Machine Learning: A Comparative Study and Interpretability Insights", 2024.
17. Y. Wang, "Time-Series Premium Risk Prediction via Bidirectional Transformer," Trans. Comput. Sci. Methods, vol. 5, no. 2, 2025.
18. J. Du, S. Dou, B. Yang, J. Hu and T. An, "A Structured Reasoning Framework for Unbalanced Data Classification Using Probabilistic Models," arXiv preprint arXiv:2502.03386, 2025.
19. Y. Zhang, J. Liu, J. Wang, L. Dai, F. Guo and G. Cai, "Federated Learning for Cross-Domain Data Privacy: A Distributed Approach to Secure Collaboration," arXiv preprint arXiv:2504.00282, 2025.
20. J. de Brito Brás de Oliveira, "Decoding the numbers and language behind financial statement fraud," M.S. thesis, 2024.
21. X. Yang, Y. Li, Z. Wang, L. Zhao and Y. Feng, "FinChain-BERT: A high-accuracy automatic fraud detection model based on NLP methods for financial scenarios," Information, vol. 14, no. 9, pp. 499, 2023.
22. J.-W. Chang, N. Yen and J. C. Hung, "Design of a NLP-empowered finance fraud awareness model: the anti-fraud chatbot for fraud detection and fraud classification as an instance," Journal of Ambient Intelligence and Humanized Computing, vol. 13, no. 10, pp. 4663-4679, 2022.
23. A. Gupta, "Attacking ELECTRA-Small: Universal Adversarial Triggers for Reading Comprehension", 2023.
24. R. Sawhney, P. Mathur, A. Mangal, P. Khanna, R. R. Shah and R. Zimmermann, "Multimodal multi-task financial risk forecasting," Proceedings of the 28th ACM International Conference on Multimedia, pp. 456-465, 2020.
25. S. Xue, T. Chen, F. Zhou, Q. Dai, Z. Chu and H. Mei, "FAMMA: A Benchmark for Financial Multilingual Multimodal Question Answering,", 2025.
26. G. Bhatia, E. M. B. Nagoudi, H. Cavusoglu and M. Abdul-Mageed, "Fintral: A family of GPT-4 level multimodal financial large language models," arXiv preprint arXiv:2402.10986, 2024.
27. H. Kaur, V. Rani, M. Kumar, M. Sachdeva, A. Mittal and K. Kumar, "Federated learning: a comprehensive review of recent advances and applications," Multimedia Tools and Applications, vol. 83, no. 18, pp. 54165-54188, 2024.