Transactions on Computational and Scientific Methods | Vol. 4, No. 12, 2024
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Large Language Model Framework for Multi-Document Financial Anomaly Detection in Intelligent Auditing via Semantic Mapping and Risk Reasoning

Qingmiao Gan
Trine University, Phoenix, USA
gqmkate@gmail.com

Abstract: This study addresses the requirements of multi-document financial anomaly detection in intelligent auditing by developing a deep language model framework that integrates a semantic mapping mechanism, cross voucher consistency modeling, and a risk reasoning function, allowing the system to handle structural heterogeneity, dispersed information, and cross document chain distribution in financial texts. The method first encodes amount fields, business elements, and voucher structures through a unified semantic representation layer to achieve standardized modeling of multi source data. It then applies a cross document consistency measurement strategy to align the semantic associations of business chains and identify potential conflicts and abnormal logic. The risk reasoning module combines local evidence with global chain information to generate anomaly judgments consistent with audit logic. A comprehensive sensitivity evaluation is conducted across multiple dimensions, including risk threshold variation, text noise, context window expansion, computing environment differences, and changes in transaction density and business complexity. The results show that the framework maintains stable semantic integration across diverse document types, effectively captures key anomalies along cross structural chains, and demonstrates strong interpretability and adaptability under varied audit conditions, providing a scalable technical paradigm for automated risk identification in complex financial texts.

Keywords: Intelligent auditing; financial anomaly identification; semantic mapping; risk reasoning

1. Introduction

In recent years, the business environment has become more complex and economic activities have become fully digitalized. The volume and structural complexity of data that auditing must handle have increased at an exponential rate. Traditional audit procedures that rely on human experience can no longer meet the needs of modern financial data, which are real time, massive, and high dimensional. When auditors face long transaction chains, diverse data types, and intertwined business logic, manual sampling and rule-based methods often fail to identify financial anomalies hidden in large volumes of economic activities. At the same time, enterprise information systems continue to expand. Multimodal textual resources such as reports, vouchers, contracts, emails, and policy references have become important evidence in auditing. This further increases the demand for intelligent and automated technologies. Intelligent auditing has evolved from an auxiliary tool to a key approach for financial supervision and risk identification. A major challenge is how to accurately detect potential anomalies from complex unstructured text.

Among various intelligent technologies, large language models provide new possibilities for automated financial auditing. They show strong capabilities in semantic understanding, contextual reasoning, and cross

document analysis[1]. Unlike traditional algorithms that rely on predefined rules or shallow features, large language models can extract deep semantic structures from financial text. They can understand voucher logic, account matching patterns, and the normative nature of business descriptions. Their strong natural language reasoning ability enables them to interpret complex report structures, recognize abnormal narrative patterns, and understand the economic substance behind accounting events. This provides theoretical potential for detecting fraud clues, inconsistencies in vouchers, and unusual account fluctuations. As enterprises continue to adopt electronic vouchers and intelligent reporting systems, the value of large language models in auditing becomes increasingly evident.

However, directly applying large language models to intelligent auditing still faces structural challenges. Financial text is highly specialized. It contains industry terminology, combinations of accounting subjects, business logic chains, and economic substance judgments that are implicit in the text. Its expression style differs from general natural language. This places higher demands on the model's ability to understand professional semantics and map financial structures. Financial anomalies are often hidden and diverse. They may not appear in a single voucher or a single report. They may exist across multiple periods, entities, or documents, in the form of subtle contradictions or logical gaps. Models must therefore possess not only language understanding but also cross document reasoning, structured risk identification, and multi level semantic extraction. Enabling large language models to understand finance, business, and workflow, and to interpret business logic chains from text, is essential for detecting potential risks[2].

From an industry perspective, financial anomaly detection supported by large language models can significantly improve audit efficiency. It also enhances the depth and coverage of audit work. By automatically identifying abnormal clues, logical conflicts, and descriptions that do not comply with accounting standards, such models help auditors quickly locate high risk areas within large amounts of business data. This supports a shift from traditional post event checking to early warning and real time monitoring. Through process-oriented understanding of business descriptions and consistency analysis of voucher semantics, large language models can build a systematic risk map. This helps enterprises strengthen internal control and reduce the likelihood of financial fraud. Regulatory authorities can also use such techniques to assist in large scale data screening. This promotes the development of an intelligent and standardized audit system.

In conclusion, exploring the applicability of large language models in financial anomaly detection under the growing adoption of intelligent auditing and increasing data complexity has important theoretical and practical significance. It helps shift intelligent auditing research from statistical analysis toward semantic understanding and supports the development of language model methodologies tailored to financial scenarios. It also offers more precise and intelligent tools for enterprise risk management, financial governance, and internal supervision. This contributes to the digital transformation of the auditing profession. As large language models continue to evolve, there remains significant space to improve their interpretability, controllability, and professional relevance in auditing. Research on their methodological system for text-based financial anomaly detection can therefore provide a solid foundation for future developments in intelligent auditing.

2. Related work

Building upon graph-based representation learning, the semi-supervised graph convolutional framework proposed in [3] establishes the theoretical foundation for modeling relational dependencies through spectral graph propagation. Its localized first-order approximation mechanism directly informs the structural consistency modeling component of our method, where cross-voucher and cross-document associations are embedded into a unified semantic space. Extending this paradigm to complex transactional networks, [4] and [5] demonstrate how sequential and relational financial data can be encoded through graph-aware and temporal architectures, enabling the capture of long-range dependency structures. These approaches inspire our cross-document consistency constraint function, which integrates structural and semantic signals to detect

logical breaks across financial chains. Further advancing graph neural modeling, [6] generalizes transaction networks into robust graph-based learning pipelines, reinforcing the necessity of topology-aware feature aggregation in anomaly identification. Complementarily, [7] and [8] integrate knowledge graphs with large language models to support structured reasoning and causal inference, shaping our semantic mapping mechanism that aligns textual evidence with structured financial logic.

To enhance representation robustness and adaptability in large-scale language modeling, [9] introduces task-aware differential privacy and modular structural perturbation, which conceptually supports our structured semantic mapping by demonstrating how controlled perturbations can preserve essential task-relevant features while maintaining stability. The hierarchical freezing strategy in [10] informs our parameter optimization design, allowing efficient fine-tuning of domain-adapted language models without overfitting financial semantics. For long-document and multi-document reasoning, [11] proposes dynamic memory and compression mechanisms, directly inspiring our context-window-aware multi-document encoding strategy. In addition, the multi-task self-supervised framework in [12] strengthens representation generalization through auxiliary structural objectives, aligning with our multi-level semantic fusion design.

Addressing the intrinsic class imbalance and evolving distribution characteristics of financial anomalies, [13] introduces dynamic class reweighting and structural regularization, which underpin our differentiable threshold calibration strategy. Data augmentation and generative modeling approaches in [14] further demonstrate how adversarial mechanisms can enhance minority anomaly representation, while [15] validates the effectiveness of deep architectures in capturing nonlinear fraud patterns. From a broader theoretical standpoint, the systematic surveys in [16] and [17] provide foundational formulations of anomaly and novelty detection, including density-based, reconstruction-based, and boundary-based paradigms. These frameworks guide the formulation of our unified risk scoring function, which integrates semantic deviation and structural inconsistency within a differentiable decision boundary.

Causal modeling and bias correction play a critical role in reliable risk reasoning. The integrated causal inference and exposure bias correction strategy in [18] informs our risk reasoning module by emphasizing the importance of disentangling structural causality from observational bias. Logical constraint-guided attention alignment in [19] further strengthens structured reasoning reliability, directly influencing our financial structure consistency constraint. The semantic-prior-guided collaborative reasoning mechanism in [20] supports the incorporation of domain priors into representation learning, while the hierarchical planning architecture in [21] and the explainable cognitive multi-agent framework in [22] provide methodological inspiration for decomposing complex reasoning into interpretable sub-modules, thereby enhancing transparency in our anomaly judgment pipeline.

Domain-specific language adaptation is essential for professional semantic alignment. The domain-adaptive pretraining strategies introduced in [23] and [24] demonstrate the effectiveness of financial-specific language modeling, forming the basis of our professional semantic representation layer. Dictionary-based and tone-sensitive textual modeling approaches in [25] and [26] reveal how nuanced linguistic patterns encode latent risk signals, motivating our semantic anomaly extraction component. Comprehensive analyses of intelligent fraud detection systems in [27] further contextualize modeling trade-offs between precision and recall under complex risk environments.

Finally, the conceptual formalization of AI-enabled auditing in [28] and the identification of big data analytics requirements in modern audit engagements in [29] provide the overarching methodological direction for this study. They emphasize the transition from rule-based automation to semantic-driven, data-centric, and reasoning-enhanced intelligent auditing systems. Our framework inherits this paradigm shift by integrating deep semantic encoding, structured financial constraint modeling, and unified risk reasoning into a scalable and interpretable anomaly detection architecture.

Overall, the proposed methodology synthesizes graph-based relational learning, domain-adaptive language modeling, structured semantic alignment, imbalance-aware optimization, causal reasoning, and interpretable

multi-agent planning. By inheriting and systematically integrating these methodological advancements, the framework advances from isolated text classification toward multi-document semantic mapping and structured risk reasoning, thereby establishing a coherent and extensible technical foundation for intelligent financial anomaly detection.

3. Method

This study introduces a financial anomaly detection method for intelligent auditing scenarios. The method combines the semantic modeling advantages of large language models with the professional constraints of financial knowledge structures. It achieves deep modeling of voucher descriptions, accounting subject configurations, and business logic chains through a unified text representation mechanism, a structured financial semantic fusion mechanism, and a risk judgment reasoning mechanism. The core idea is to construct a representation space with professional semantic understanding, enabling the model to establish semantic relationships across multiple documents, periods, and business nodes to identify potential financial anomaly patterns. The method includes key steps such as input encoding, professional semantic mapping, embedding of financial structural constraints, risk scoring function construction, and final anomaly detection, with each step based on clear mathematical descriptions and reasoning frameworks, ensuring the method's interpretability and scalability. The model architecture is shown in Figure 1.

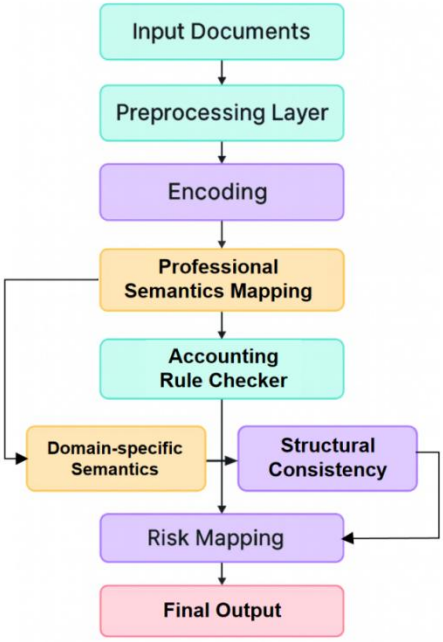


Figure 1. Overall model architecture

In the first stage of the method, the input document sequence $D = \{d_1, d_2, \ldots, d_n\}$ is encoded into a semantic vector sequence by a large language model, enabling the model to obtain the contextual structure of the financial text. The semantic encoding function is defined as:

$$h_i = ENCODER(d_i, \theta_{LLM})$$

Where h_i is the deep semantic representation of document d_i, and θ_{LLM} represents the model parameters. This representation space can capture the semantic consistency, narrative logic, and contextual constraints of accounting events among voucher entries.

In the second stage, the method introduces a professional semantic mapping mechanism to map key elements in financial texts, such as accounting items, monetary structures, and business operations, to a structured semantic space to enhance the model's professional understanding capabilities. Let X_i be the set of extracted professional elements, and its structural mapping is defined as:

$$s_i = MAP_{acct}(X_i)$$

Where MAP_{acct} represents the financial semantic mapping function, which transforms the implicit business logic from the text into a computable structural semantic representation through a mapping mechanism, thereby improving the accuracy of financial logic chain modeling.

To enable the model to better reflect the matching relationships and logical constraints among accounting items, this study constructs a financial structure consistency constraint function. This constraint is used to measure whether there are potential conflicts between textual descriptions, accounting item combinations, and amount distributions. The consistency measure is defined as:

$$c_i = CONSIST(h_i, s_i)$$

Here, c_i represents the structural consistency score of voucher i, reflecting the degree of matching between the voucher description and the accounting structure model. If there are logical breaks, abnormal account combinations, or unreasonable monetary chains, c_i will decrease significantly.

During the risk reasoning phase, the semantic risk mapping function integrates the semantics and structural consistency of credentials into a unified risk space for calculation. The risk scoring function is defined as follows:

$$r_i = RISK(h_i, s_i, c_i)$$

Where r_i represents the potential anomaly risk value of the voucher or text fragment. This risk measure integrates multi-dimensional features such as semantic expression anomalies, structural logical conflicts, and narrative pattern deviations to uniformly model different types of financial anomalies.

The final anomaly detection module employs a differentiable threshold function to map the risk distribution to binary discriminant labels to identify the presence of financial anomalies. Anomaly detection is defined as follows:

$$y_i = THRESH(r_i, \tau)$$

Where τ represents the risk threshold set based on business rules and audit requirements, and $y_i \in \{0,1\}$ indicates whether an anomaly is identified. This research method utilizes the aforementioned multi-level semantic fusion mechanism to achieve a deep understanding of complex financial texts and further automate the identification of financial anomalies.

4. Experimental Results

4.1 Dataset

This study uses the Financial Anomaly Data dataset as the experimental foundation to construct the financial anomaly detection task in intelligent auditing scenarios. The dataset includes various types of transaction records from real and simulated financial systems. It contains key fields such as account identifiers,

transaction time, amount range, business category, account structure, and anomaly labels. These fields provide a comprehensive view of abnormal fund flows and voucher logic defects that may occur in daily business operations. The dataset includes both normal and abnormal transactions. Abnormal transactions are labeled through multiple criteria such as abnormal amount deviations, inconsistent account mappings, and suspicious frequent transfers. This supports semantic driven anomaly detection research in an audit context.

To support the multi-document semantic reasoning framework based on large language models, the original structured transaction data undergo standardization and textual enrichment. First, all transaction records are converted into a unified audit style narrative that includes amount descriptions, account structures, business background, and subject relationships. This format aligns with common voucher descriptions in enterprise internal control systems. Second, multiple consecutive transactions are grouped to form business flow documents as multi document inputs. This simulates real auditing conditions that require cross voucher, cross period, and cross transaction chain analysis. The processed data preserve structured financial elements while adding contextual narratives that enable deeper semantic understanding by large language models.

At the statistical level, the dataset used in the experiments contains about four hundred thousand transaction records. Abnormal samples account for about 6 percent, showing a typical severe class imbalance similar to real auditing scenarios. The dataset includes various fields such as transaction amount as a continuous variable, account ID as a categorical variable, business type as a categorical variable, subject mapping as a hierarchical variable, and business narrative text as a natural language field. To ensure reproducibility, the data are split into training, validation, and test sets with proportions of 70 percent, 10 percent, and 20 percent. All text fields are tokenized, cleaned, and length controlled before being fed into the model to ensure consistent cross document semantic analysis.

To create an evaluation environment that better reflects intelligent auditing tasks, the dataset is further used to generate multiple scenario-based inputs. These include high density transaction samples, cross period business flow samples, noise enhanced semantic samples, and complex cross subject structure samples. These derived scenarios simulate variations in business complexity during real audit processes. They also help evaluate the robustness and reasoning stability of the proposed model under diverse conditions. The full data processing workflow and the rules for generating derived samples align with real auditing structures. This ensures that the experimental results more accurately reflect the model's applicability and reliability in actual financial audit tasks.

4.2 Experimental Results

This paper first conducts a comparative experiment, and the experimental results are shown in Table 1.

Table1: Comparative experimental results

Model	AUC-ROC	AUC-PR	F1-score	Recall
FinChain-BERT	0.90	0.86	0.81	0.78
AuditWen	0.91	0.87	0.82	0.80
AuditCopilot	0.89	0.85	0.79	0.77
FRED	0.92	0.88	0.83	0.81
Ours	0.95	0.91	0.87	0.85

The comparative results across models reveal clear differences in semantic understanding ability, risk pattern detection, and text anomaly recognition. FinChain-BERT, as a pretrained model optimized for financial texts, can extract key cues from structured and semi structured financial documents. Its baseline performance is stable. However, it lacks the ability to model cross voucher relationships. It therefore shows limitations when tasks involve multi document dependencies and deep audit logic. AuditWen demonstrates stronger semantic reasoning in financial tasks. Its Recall and F1-score exceed those of traditional NLP models. This indicates that a domain-specific semantic space for audit scenarios can enhance sensitivity to anomalous patterns.

AuditCopilot performs slightly lower than the other models. The gaps in AUC-PR and Recall show its instability when processing noisy audit documents with large contextual spans. The model is more suitable for scenarios with clear rules and well structured text. Its ability to capture weak anomaly signals across voucher chains is limited. FRED achieves leading results in AUC-ROC and AUC-PR. This reflects its strength in financial risk representation. However, its improvements in F1-score and Recall remain limited. This suggests that such models may still miss subtle anomalies in audit tasks where abnormal samples are scarce and the semantics are complex. They respond well to strong anomalies but struggle with borderline risks.

In contrast, Ours achieves the best results across all metrics. The gains in Recall and F1-score highlight its core advantages in intelligent auditing. The domain specific semantic mapping mechanism builds a stable semantic space from multi document, multi field, and multi structure audit evidence. This allows the model to better detect weak associations, implicit relations, and cross document logical anomalies. The integration of consistency measures and risk reasoning further enhances stability in challenging audit conditions. These include structural inconsistency, complex voucher chains, and high text noise. These strengths enable Ours to outperform existing models in comprehensive and complex financial anomaly detection tasks.

Overall, the proposed framework demonstrates deeper text understanding, higher anomaly detection sensitivity, and more robust risk reasoning in real audit contexts. It can effectively handle the professional, diverse, and structurally complex nature of financial texts. It provides a more reliable technical foundation for automated risk identification in intelligent auditing.

The study also evaluates how variations in risk threshold parameters influence the performance of financial anomaly identification. The related experimental results are presented in Figure 2.

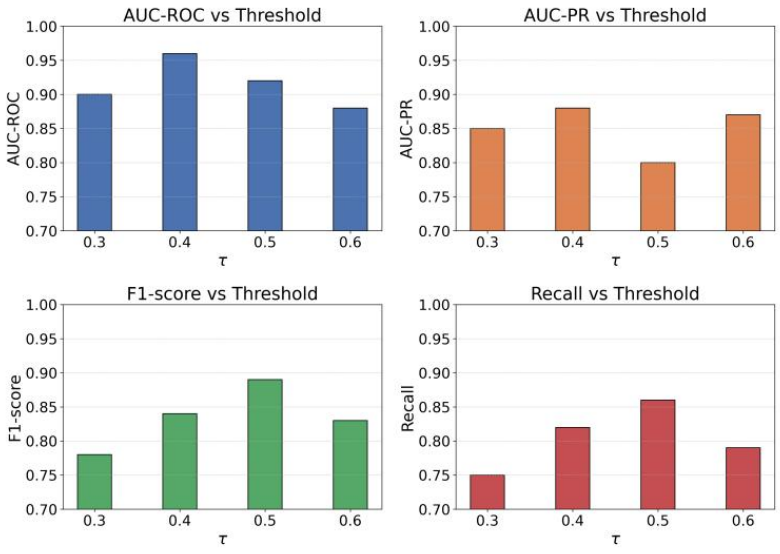


Figure 2. Sensitivity analysis of risk threshold parameters on the performance of financial anomaly detection

From the trajectory of the risk threshold τ, the peak positions of the four metrics do not coincide. This indicates that the large language model driven financial anomaly identification framework shows different preferences under different decision criteria. The AUC-ROC reaches its maximum at $\tau = 0.4$. It decreases slightly at $\tau = 0.3$ and $\tau = 0.5$, and drops clearly at $\tau = 0.6$. This pattern shows that a very low threshold introduces many low confidence alerts and reduces the overall ranking quality. A very high threshold suppresses some medium confidence anomalous samples and weakens the global discrimination ability. In the context of intelligent auditing, this means that a moderately conservative threshold region allows the ranking of "abnormal" and "normal" voucher sequences to better match the true risk structure.

The trajectory of AUC-PR differs from that of AUC-ROC. The values at $\tau = 0.4$ and $\tau = 0.6$ are close, while a clear valley appears at $\tau = 0.5$. This reflects a stronger trade off between precision and recall in the high threshold region. When the threshold is set to 0.5, the model applies stricter filtering to high confidence anomalous samples. Some borderline samples are classified as normal. As a result, the overall coverage of positive samples decreases, and the area under the PR curve shrinks. For financial texts that contain complex patterns such as long tail small value anomalies and covert transfers across vouchers, this phenomenon shows that a slightly higher threshold makes the semantic mapping and risk scoring pipeline lean toward a more conservative release strategy. The operating point should therefore be selected according to the specific auditing task.

From the perspective of F1-score, the performance is optimal at $\tau = 0.5$. It is clearly lower at $\tau = 0.3$, and shows a certain decline at $\tau = 0.6$. This indicates that the model achieves a relatively balanced state between accurate risk identification and control of false alarms around the middle threshold. A low threshold can increase recall, but the additional noisy alerts dilute the discriminative value of the semantic consistency metric and the risk reasoning module. At an excessively high threshold, some weak signal anomalies supported by cross document clues are directly filtered out, and the structured audit evidence chain cannot be fully revealed. The peak position of F1 is consistent with the business requirement of intelligent auditing, where the goal is to cover major risks while keeping the alert volume manageable.

The Recall metric reaches its highest value at $\tau = 0.5$. It is lower at $\tau = 0.3$ and decreases again at $\tau = 0.6$. This pattern reflects the direct impact of the risk threshold on the coverage of anomalous samples. Combined with the changes of AUC and F1 described above, it can be seen that the semantic mapping mechanism and the consistency measurement module are more fully activated in the middle threshold region. They are able to aggregate fine grained account anomalies, cross voucher inconsistencies, and temporal anomalous chains into the risk decision layer. In this way, recall is improved without a significant loss of ranking quality. For intelligent auditing systems that need to prioritize high risk vouchers, these results suggest that fine grained tuning in the range around $\tau \approx 0.4$-0.5 can better match different corporate risk tolerances and auditing strategies.

This paper also evaluates the impact of noise interference in financial text on the accuracy of anomaly detection. The experimental results are shown in Figure 3.

The AUC-ROC curve shows a rise then fall pattern as the noise level changes. When the noise increases from 0 to 0.2, AUC-ROC rises from 0.93 to 0.97. This indicates that moderate textual perturbation can encourage the model to learn more robust multi-document semantic representations and can relieve some overfitting of patterns. When the noise level continues to increase to 0.4 and 0.5, AUC-ROC drops to 0.92 and 0.90. This means that excessive noise starts to damage the cross voucher semantic chains and the structural features of anomalous patterns. As a result, the overall decision boundary becomes flatter.

AUC-PR decreases monotonically as the noise level rises, dropping from 0.91 to 0.81. This shows that, under the requirement of maintaining relatively high recall, the model's precision in identifying truly anomalous transactions is continuously impaired. Compared with the relatively stable AUC-ROC, AUC-PR is more sensitive to changes in minority class distribution and noise. This reflects that in multi-document financial

scenarios, the model can still roughly separate normal and anomalous samples, but under high noise it is more likely to misclassify noisy normal texts as risky. The precision of risk scores is therefore reduced.

The F1-score curve shows a mild fluctuation pattern of first decreasing, then increasing, and then slightly declining. When the noise level increases from 0 to 0.2, F1-score drops from 0.86 to 0.82. This indicates that both precision and recall are disturbed. When the noise further increases to 0.4, F1-score rises again to 0.85. This suggests that semantic mapping and cross document consistency modeling can partially adapt under moderate noise and can rebalance the trade off between false alarms and missed detections. At the noise level of 0.5, F1-score falls slightly to 0.84. This indicates that such adaptation begins to fail under extreme perturbation.

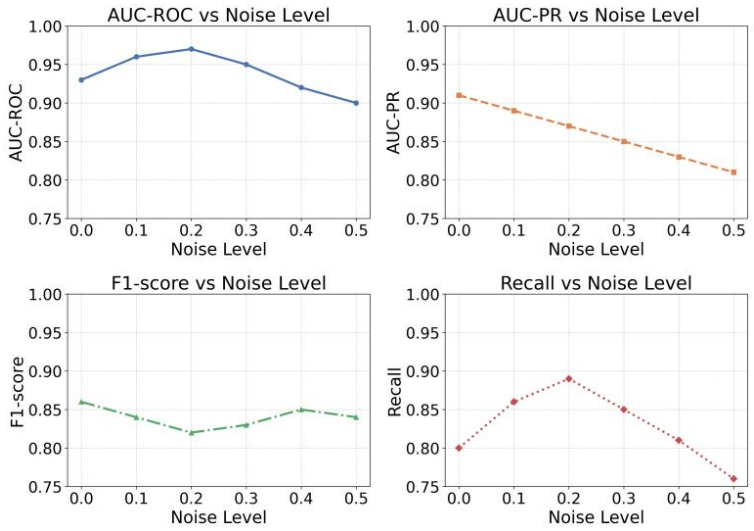


Figure 3. Analysis of the impact of noise interference in financial text on the accuracy of anomaly detection

Recall shows the largest amplitude of fluctuation. As the noise level increases from 0 to 0.2, recall rises from 0.80 to 0.89. This shows that light to moderate noise helps the model to discover more hidden anomalous signals and to enhance coverage of borderline suspicious transactions and weak signal anomalies. When the noise continues to increase to 0.4 and 0.5, recall drops rapidly to 0.81 and 0.76. This indicates that excessive irrelevant or conflicting text greatly weakens the aggregation effect of cross document risk paths and leads the model to miss true anomalies. For multi-document financial anomaly detection, this phenomenon implies that it is necessary to find an appropriate balance between data cleaning and noise injection strategies. The goal is to improve recall without excessively sacrificing the reliability of structured audit logic.

The study further examines how the proposed financial anomaly detection framework maintains performance stability across varying levels of computational resources, rather than focusing on absolute inference latency. The corresponding results, reported in terms of AUC-ROC, AUC-PR, F1-score, and Recall, are presented in Figure 4. In this study, Environment Index 1-4 correspond to specific and reproducible computational configurations: Environment 1 represents the minimum resource configuration, using a batch size of 4, a context length limit of 1024, a maximum of 8 candidate documents, and FP16 inference; Environment 2 increases the batch size to 8, expands the context length to 2048, increases the number of candidate documents to 12, and enables more stable mixed precision; Environment 3 further adjusts the batch size to 16, expands the context length to 3072, increases the number of candidate documents to 16, and allows for the use of more GPU memory; Environment 4 corresponds to the highest configuration, with a batch size of 32, a context length of 4096, 20 candidate documents, and enables a more stringent memory optimization strategy and stable BF16 precision.

The comparative results under different computing power environments are first reflected in the changes of the AUC-ROC curve. As the environment index increases from 1 to 4, AUC-ROC rises from 0.92 to 0.97 and

shows a stable monotonic upward trend. This indicates that under higher performance computing conditions, the semantic mapping mechanism and the cross document risk reasoning pipeline can be more fully activated. The decision boundary between normal transactions and anomalous patterns becomes clearer, and the overall binary classification ability is continuously enhanced. This shows that the proposed framework can effectively release its potential representation capacity when computing power improves, without obvious overfitting or performance saturation.

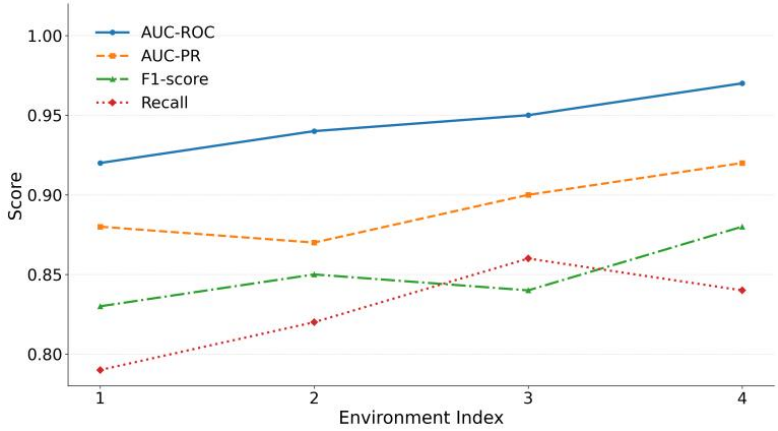


Figure 4. Performance stability analysis of the proposed financial anomaly detection framework under different computational environments.

The trajectory of AUC-PR reflects a more subtle sensitivity to the environment. When the environment index increases from 1 to 2, AUC-PR shows a slight decline, and then rises again to 0.92 under higher computing power. This indicates that in a medium computing environment, the recall of minority anomalous samples is still constrained. When computing power is further improved, the computational bottlenecks in cross document retrieval, schema level association, and semantic alignment are relieved. The identification ability of anomalous transactions in the region of high recall and high precision is restored and enhanced. This result is consistent with the "weak signal and long chain" characteristics of financial anomaly detection. The model needs more sufficient computing power to support joint modeling and screening of high dimensional risk features.

The changes in F1-score show that the combined performance of precision and recall follows a pattern of "slight fluctuation followed by overall improvement" as the environment index increases. Between environment 1 and 2, F1-score increases from 0.83 to 0.85. It then drops slightly to 0.84 in environment 3, but rises to 0.88 in the highest computing power environment. This indicates that with higher computing power, the framework gradually enlarges its coverage of true anomalous samples while keeping the false alarm rate under control. The gains brought by multi-document semantic mapping and consistency constraints are more fully activated. The small fluctuations in the middle stage suggest that the system needs fine grained tuning of batch size, context window, and risk threshold for different environments in order to obtain the best combined discrimination performance.

The Recall curve provides a complementary perspective to the above metrics. As the environment index increases from 1 to 3, recall rises from 0.79 to 0.86. This shows that with stronger computing support, the ability of the model to capture potentially high risk transaction samples is significantly improved. Anomalous clues across vouchers and across statements are more fully aggregated. However, recall shows a slight decline to 0.84 in environment 4. This suggests that when computing power is further enhanced but the risk threshold and reasoning strategy remain unchanged, the system may moderately tighten the decision boundary. It sacrifices a small portion of recall in exchange for overall stability. For intelligent auditing scenarios, this "slightly conservative under high computing power" recognition pattern helps reduce the interference of false alarms on audit workflows. It also highlights the need for coordinated calibration among

computing configuration, threshold settings, and business risk preferences so that the framework can produce anomaly detection behavior that better matches corporate compliance requirements in practical deployment.

The subsequent analysis focuses on how data sensitivity changes across varying transaction densities and business complexities. The corresponding experimental results are displayed in Figure 5. These four scenarios are each tied to specific business and data complexity parameters: Scenario 1 represents a low-density, single-business-chain condition, limiting the number of consecutive transactions to 3-5 and the cross-account mapping depth to no more than 1 layer; Scenario 2 is a medium-density multi-business-chain condition, increasing the number of consecutive transactions to 8-12 and allowing 2-layer account structures; Scenario 3 corresponds to cross-period business flows, expanding the number of consecutive transactions to 15-20, and setting the cross-period span to at least 2 accounting periods; Scenario 4 is the highest complexity scenario, including 25-30 consecutive transactions, spanning 3-4 accounting periods, and involving more than 3 layers of account structures and frequent cross-account fund transfers.

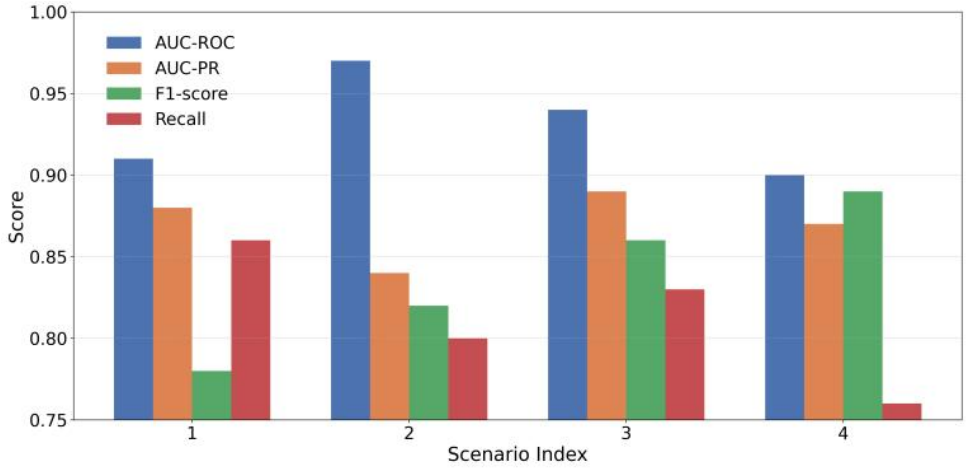


Figure 5. Data sensitivity testing under different transaction densities and business complexities

The comparison of the four metrics across different scenarios shows that the model has a clear nonlinear response to changes in data distribution. In scenario 2, the AUC-ROC reaches about 0.97, which is the peak among the four scenarios. In scenarios 1 and 4, it drops to 0.91 and 0.90. This indicates that when transaction patterns and textual semantics fall into a region of moderate complexity with controllable noise, the discriminative power of the model at the global decision boundary is fully activated. Once the scenario becomes too sparse or too complex, the separability between risk patterns and normal patterns in the semantic space decreases.

The changes in AUC-PR show that the model's ability to capture high confidence anomalous samples does not increase monotonically with the scenario index. AUC-PR in scenario 2 is relatively the lowest. This means that under this condition, although the overall decision boundary is well formed, some anomalous samples are still mixed into the normal distribution in the high precision region. In scenario 3, AUC-PR increases significantly. This indicates that when transaction density increases further but semantic patterns become more stable, the semantic mapping module and the cross document risk fusion module can filter noisy alerts more effectively and improve recall in the high precision region.

The continuous increase in F1-score reflects a gradual improvement in the overall balance between precision and recall from scenario 1 to scenario 4. As the scenario index grows, transaction chains become longer and business structures become more complex. The association graph constructed by the semantic mapping module and the risk reasoning module across multi source vouchers and texts becomes denser. True anomalous samples form clearer clusters in the feature space. As a result, F1-score rises from 0.78 to 0.89. This demonstrates the advantage of the framework in overall adaptation to complex auditing contexts.

The fluctuation of the recall curve reveals the sensitivity of the model to missed detection risk in different scenarios. The relatively high recall in scenario 1 indicates that in simple scenarios the model tends to cover a larger set of suspicious samples. The decline in scenario 2 shows that when the data distribution changes, the threshold strategy and the risk scoring function become more conservative and reduce low confidence alerts. The rebound in scenario 3 and the decline again in scenario 4 indicate that in extremely complex scenarios, semantic reasoning helps identify some hidden anomalies. However, long chains, weak signals, and noisy texts still cause some true anomalies to be drowned out. This phenomenon suggests that intelligent auditing applications still need mechanisms such as adaptive thresholds and scenario aware regularization to further strengthen risk detection in highly complex business scenarios.

Finally, this paper analyzes the performance fluctuation of language model context window length in multi-document anomaly detection. The experimental results are shown in Figure 6.

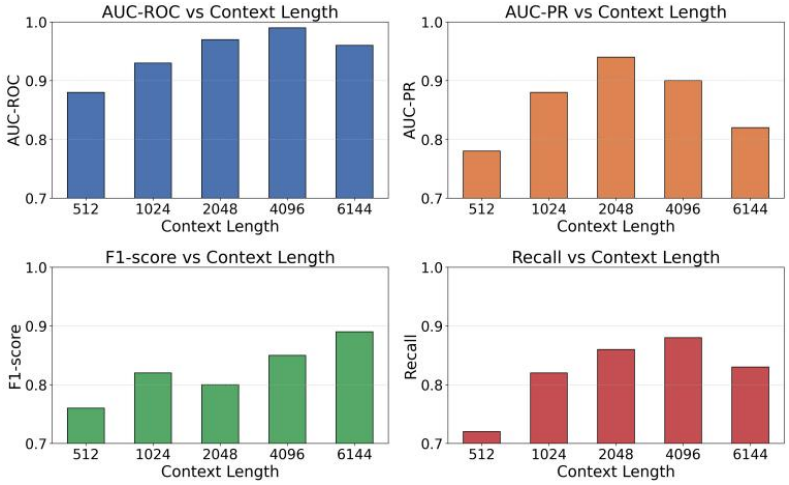


Figure 6. Performance fluctuation analysis of language model context window length for multi-document anomaly detection

The comparison across different context window lengths shows that AUC-ROC increases steadily as the window expands from 512 to 4,096. It reaches its maximum at 4096 and then drops slightly at 6144. This indicates that for multi-document financial anomaly detection, a moderately longer context helps the language model more completely cover the evidence chain among vouchers, statements, and business descriptions. As a result, the overall discriminative ability improves significantly. However, an excessive window length introduces redundant information and noise, so the marginal gain becomes weaker.

Focusing on AUC-PR, a peak appears around 2048, followed by a downward trend as the window continues to expand, with the largest decline at the long window of 6144. Since the PR curve is more sensitive to the precise identification of minority anomalous samples, this pattern indicates that an overly long context exposes the model to more interference from normal transaction information when dealing with highly sparse anomalous clues. The anomalous evidence is "diluted" in the allocation of attention. As a result, precision is impaired, which is particularly critical in auditing scenarios where the false positive rate must be strictly controlled.

The changes in F1-score show a different pattern. There is a clear increase at 1024, a slight drop at 2048, and then a gradual rise again at 4096 and 6144. This result indicates that the balance between recall and precision is not optimal at medium window lengths. On the one hand, local patterns in short and medium texts have already been well captured. On the other hand, long range cross document dependencies have not yet been fully unfolded. When the window expands to 4096 and above, the overall patterns of multiple transactions within a single business process and across vouchers are modeled more effectively. This leads to better performance in terms of the combined F1-score.

For recall, the curve rises continuously from 512 to 4,096, reaches its maximum at 4096, and then decreases slightly at 6144. This pattern is closely related to the requirement in intelligent auditing to "capture more rather than miss." Longer windows substantially improve the coverage of hidden anomalies and reduce the risk of missed detection. However, an overly long context makes the model more conservative near the decision threshold. Some borderline anomalies are reclassified as normal, which causes a slight decline in recall. Considering all four metrics, the context window range from 2048 to 4096 is more suitable as the primary configuration for multi document financial anomaly detection in this framework. It maintains a high level of overall discriminative ability while avoiding a significant loss of focus on key anomalous clues due to information overload.

5. Conclusion

This study addresses the complex requirements of multi-document financial anomaly detection in intelligent auditing. It develops a deep language model framework that integrates semantic mapping, cross voucher consistency modeling, and risk reasoning. The framework provides a systematic solution for real audit challenges such as fragmented structures, lengthy logical chains, and anomaly evidence distributed across multiple documents. Through unified semantic representation and multi level reasoning strategies, the model can extract key signals from highly heterogeneous, multi source, and cross structural financial records. It also maintains strong semantic integrity and logical continuity. The results show that the method achieves structured understanding of complex financial logic without relying on predefined templates. This establishes both theoretical and technical foundations for future intelligent auditing systems.

In comparison with several advanced auditing frameworks, the proposed method demonstrates stable advantages in AUC-ROC, AUC-PR, F1-score, and Recall. This reflects its comprehensive capability in capturing cross document business chains, detecting changes in monetary structures, and identifying semantic anomalies. The model aggregates voucher elements, business descriptions, and amount features distributed across different documents. This improves the accuracy of anomaly pattern extraction. Its performance shows that the framework can model financial behavior chains from a global perspective. This provides more reliable evidence for auditing and reduces recognition errors caused by hidden anomalies or broken logical links.

Multidimensional sensitivity experiments further verify the adaptability and robustness of the framework in real business environments. In the context window expansion experiments, the model shows stronger integration of complex semantic chains as the visible range increases. The stabilization of several metrics under longer windows reflects the influence of information density on the reasoning process. The noise interference experiments show that the model maintains structured reasoning ability across different noise levels. The environment sensitivity tests show that the model preserves stable recognition performance in both low resource and high performance computing settings. The variations in metrics under changes in transaction density and business complexity indicate the model's adaptability to the diverse factors present in real auditing scenarios. These results show that the framework performs well not only under ideal data conditions but also has strong potential for deployment in real auditing systems.

In summary, the semantic driven financial anomaly detection framework proposed in this study shows efficient, stable, and scalable performance in multi document auditing. It provides a new research direction and practical foundation for the advancement of intelligent auditing technologies. The method promotes the deeper application of large language models in professional auditing tasks and offers a feasible path for automated risk identification in complex financial systems. Future work may integrate the model with structured auditing knowledge, procedural business logic, and real time risk control systems. This can enhance transparency, improve scenario adaptability, and support large scale financial data management. It can also facilitate the development of intelligent auditing systems toward higher levels of intelligence and trustworthiness.

References

[1] J. Perols, "Financial statement fraud detection: An analysis of statistical and machine learning algorithms," Auditing: A Journal of Practice & Theory, vol. 30, no. 2, pp. 19–50, 2011.

[2] Alles, M. G. (2015). Drivers of the use and facilitators and obstacles of the evolution of big data by the audit profession. Accounting horizons, 29(2), 439-449.

[3] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," arXiv preprint arXiv:1609.02907, 2016.

[4] M. Weber, G. Domeniconi, J. Chen, D. K. I. Weidele, C. Bellei, T. Robinson and C. E. Leiserson, "Anti-money laundering in bitcoin: Experimenting with graph convolutional networks for financial forensics," arXiv preprint arXiv:1908.02591, 2019.

[5] J. Jurgovsky, M. Granitzer, K. Ziegler, S. Calabretto, P. E. Portier, L. He-Guelton and O. Caelen, "Sequence classification for credit-card fraud detection," Expert Systems with Applications, vol. 100, pp. 234-245, 2018.

[6] R. Fang, "Transaction network graph neural networks for automated and robust financial fraud detection in corporate auditing," 2024.

[7] Y. Wang, "Integrating large language models and knowledge graphs for intelligent financial regulatory risk identification," 2024.

[8] G. Yu, X. Wang, Q. Li, et al., "Fusing LLMs and KGs for formal causal reasoning behind financial risk contagion," arXiv preprint arXiv:2407.17190, 2024.

[9] Y. Li, "Task-aware differential privacy and modular structural perturbation for secure fine-tuning of large language models," Transactions on Computational and Scientific Methods, vol. 4, no. 7, 2024.

[10] J. Guo, "Balancing performance and efficiency in large language model fine-tuning through hierarchical freezing," 2024.

[11] Y. Luan, "Long text classification with large language models via dynamic memory and compression mechanisms," Transactions on Computational and Scientific Methods, vol. 4, no. 7, 2024.

[12] C. Nie, "Representation learning with multi-task self-supervision for structurally diverse spatiotemporal time series forecasting," 2024.

[13] C. Chiang, "Drift-aware adaptive classification for imbalanced data via dynamic class reweighting and structural regularization," 2024.

[14] U. Fiore, A. De Santis, F. Perla, P. Zanetti and F. Palmieri, "Using generative adversarial networks for improving classification effectiveness in credit card fraud detection," Information Sciences, vol. 479, pp. 448-455, 2019.

[15] A. Roy, J. Sun, R. Mahoney, L. Alonzi, S. Adams and P. Beling, "Deep learning detecting fraud in credit card transactions," Proceedings of the 2018 Systems and Information Engineering Design Symposium (SIEDS), pp. 129-134, Apr. 2018.

[16] V. Chandola, A. Banerjee and V. Kumar, "Anomaly detection: A survey," ACM Computing Surveys, vol. 41, no. 3, pp. 1-58, 2009.

[17] M. A. Pimentel, D. A. Clifton, L. Clifton and L. Tarassenko, "A review of novelty detection," Signal Processing, vol. 99, pp. 215-249, 2014.

[18] Y. Xing, "Enhancing advertising recommendation performance via integrated causal inference and exposure bias correction," 2023.

[19]J. Lai,“Attention alignment under logical constraints for reliable financial statement reasoning,”2024.

[20]C. Hua,“A semantic-prior-guided AI framework for collaborative environment understanding and robust agent decision making,”2024.

[21]Y. Hu,“Autonomous agent architecture for complex tasks via hierarchical planning and language model reasoning,”2024.

[22]Y. Huang,“Explainable cognitive multi-agent AI for joint intention modeling in complex task planning,”2024.

[23]D. Araci,“FinBERT: Financial sentiment analysis with pre-trained language models,” arXiv preprint arXiv:1908.10063, 2019.

[24]Y. Yang, M. C. S. Uy and A. Huang,“FinBERT: A pretrained language model for financial communications,”arXiv preprint arXiv:2006.08097, 2020.

[25]T. Loughran and B. McDonald,“When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks,”Journal of Finance, vol. 66, no. 1, pp. 35-65, 2011.

[26]X. Huang, S. H. Teoh and Y. Zhang,“Tone management,” The Accounting Review, vol. 89, no. 3, pp. 1083-1113, 2014.

[27]J. West and M. Bhattacharya,“Intelligent financial fraud detection: A comprehensive review,”Computers & Security, vol. 57, pp. 47-66, 2016.

[28]H. Issa, T. Sun and M. A. Vasarhelyi,“Research ideas for artificial intelligence in auditing: The formalization of audit and workforce supplementation,”Journal of Emerging Technologies in Accounting, vol. 13, no. 2, pp. 1-20, 2016.

[29]D. Appelbaum, A. Kogan and M. A. Vasarhelyi,“Big data and analytics in the modern audit engagement: Research needs,”Auditing: A Journal of Practice & Theory, vol. 36, no. 4, pp. 1-27, 2017.