Construction of Financial Risk Assessment Model Based on Text Mining and LLM Architecture

Zhenglin Li *, Mingxiu Sui , Chen Yang , Sa Liu , Bo Guang , Xinjin Li

Posted Date: 20 January 2026

doi: 10.20944/preprints202601.1407.v1

Keywords: text mining; large language model; financial risk assessment; knowledge graph; retrieval enhanced generation; model governance

Article

Construction of Financial Risk Assessment Model Based on Text Mining and LLM Architecture

Zhenglin Li [1,*], Mingxiu Sui [2], Chen Yang [3], Sa Liu [4], Bo Guang [5] and Xinjin Li [6]

[1] Texas A&M University, College Station, TX 77843, United States
[2] The University of Iowa, Iowa City, IA 52242, United States
[3] University of Pennsylvania, Philadelphia, PA 19104, United States
[4] The University of California Davis, Davis, CA 95616, United States
[5] Virginia Tech, Blacksburg, VA 24061, United States
[6] Columbia University, New York, NY 10027, United States
* Correspondence: zhenglin_li@tamu.edu

Abstract

Faced with the challenges of accelerating growth in unstructured text data and increasing risk concealment in the financial market, this study constructs a financial risk assessment system that combines text mining with large language models (LLMs). This system forms an end-to-end architecture encompassing data, knowledge, models, services, and governance. The system collects multi source text, constructs a risk knowledge graph, and extracts key events and sentiment signals. These are then integrated into the LLM framework of retrieval augmented generation (RAG) and multi feature fusion to achieve credit risk prediction and default probability estimation. This experiment relies on the FinBen Lending Club dataset (2024) and conducts comparative experiments, ablation studies, error analysis, and stability tests. The model outperforms traditional structured models and plain text models in key evaluation indicators such as F1, MCC, and PR AUC. In scenarios of market environment changes, cross industry migration, and anti-interference, the model's stability and compliance performance are outstanding. This study designs an intelligent risk identification solution for financial institutions, which makes the identification process explainable, traceable, and auditable. This study has significant theoretical and practical impact on risk governance and decision support for banks, securities, insurance, and regulatory authorities.

Keywords: text mining; large language model; financial risk assessment; knowledge graph; retrieval enhanced generation; model governance

1. Introduction

In recent days where the financial system is becoming increasingly complex and uncertainty is continuously rising, traditional risk assessment methods based on structured data are struggling to cope with the rapidly changing hidden risks in the financial market. Numerous unstructured text materials scattered in company announcements, financial reports, media information, social platforms and regulatory documents can serve as key clues to identify corporate anomalies and potential systemic risks. This data is huge in scale and updated with strong timeliness, making it difficult for traditional analytical methods to achieve deep and efficient mining.

With the rapid progress of large-scale language model technology, natural language processing has achieved breakthrough developments, introducing a new technical path for financial risk assessment [1–4]. Large language models (LLMs) can realize the mining of complex semantic relationships and the fusion of cross-text information, effectively enhancing the sensitivity of risk identification and the foresight of prediction in multi-source heterogeneous data fusion and reasoning, with applications in medicine, education, and finance [5–8]. The combination of text mining and LLM-based architectures is conducive to creating a financial risk assessment system that

 doi:10.20944/preprints202601.1407.v1

integrates automation, explainability, and scalability, and enhances the overall performance of early warning, decision support, and compliance supervision [9,10].

This research plan aims to build a financial risk assessment framework with text mining and large language models at its core. It will study the complete technical sequence from data collection, knowledge mining, and model training to verification and application, establish a systematic methodological model, and equip financial institutions such as banks, securities, and insurance organizations with a practical risk identification and quantification toolkit. This will have significant theoretical value and practical significance for financial stability and regulatory practice.

2. Related Work

Research on financial risk assessment has traditionally relied on structured numerical data, statistical modeling techniques, and econometric theories. However, the exponential growth of unstructured financial text data and advances in natural language processing and large language models have significantly transformed risk modeling paradigms [11–15].

A review of existing research reveals that while studies on data sources, model construction, and risk categorization have laid a foundation, three deficiencies remain: a comprehensive technical path from deep text semantic understanding to structured quantitative outputs is lacking; traditional methods are limited in adaptability to cross-domain knowledge reasoning and dynamic risk prediction tasks; and model interpretability and regulatory compliance have yet to fully complement the capabilities of large-scale language models [16–20].

This paper aims to construct a financial risk assessment system that integrates text mining with the LLM architecture, addressing the shortcomings of existing research and building a risk assessment system that is highly implementable, verifiable, and regulatory adaptable.

3. Construction of Financial Risk Assessment Model Based on Text Mining and LLM Architecture

3.1. Overall Architecture Design

The proposed financial risk assessment model employs an end-to-end hierarchical architecture with five layers: data, knowledge, models, services, and governance. The data layer collects and manages diverse sources, including financial announcements, news, reports, regulatory documents, and social media, while ensuring integrity through cleaning, entity alignment, and time synchronization. The knowledge layer constructs a financial ontology and knowledge graph, performing event extraction, causal analysis, and semantic enrichment to enable contextual reasoning and logical inference.The model layer combines text mining and large language models with pre-training, instruction fine-tuning, and retrieval-augmented generation for risk factor extraction, credit risk assessment, sentiment analysis, and default probability estimation, incorporating uncertainty quantification and explainability mechanisms. The service layer delivers outputs through dashboards, dynamic reports, and interfaces to support decision making in banks, securities, and regulatory agencies. The governance layer ensures model risk control, compliance, data privacy, and drift monitoring, enabling continuous supervision and adjustment across the entire system. Figure 1 shows the overall architecture of the model.

3.2. Data and Knowledge Construction

This study integrates announcements, financial reports, news, research reports, regulatory notices, and social media content into a unified timeline, implements data cleansing, redundancy removal, language standardization, and entity matching, and then implements event normalization and time synchronization between documents, restricts the dissemination of outdated information, and implements time decay and comprehensive source quality weighting for document.

 doi:10.20944/preprints202601.1407.v1

$$w_d = e^{-\lambda \Delta t_d} \cdot q_d, \quad w_d \in (0,1] \tag{1}$$

where Δt_d represents the time difference from the evaluation point to the document release date, $\lambda > 0$ is the decay coefficient value, $q_d \in (0,1]$ is the source credibility. The similarity between the text and the corresponding entity is calculated in the embedding space to support entity disambiguation:

$$e(x,y) = \frac{\boldsymbol{v_x} \cdot \boldsymbol{v_y}}{||\boldsymbol{v_x}|| \cdot ||\boldsymbol{v_y}||} \tag{2}$$

On this basis, the risk ontology and knowledge graph are formed:

$$\mathrm{G} = (\mathrm{V}, \mathrm{E}) \tag{3}$$

Among them, V represents subjects, indicators, and event types, while E defines temporal, causal, and constraint relationships between entities. The mapping ϕ:text $\rightarrow (entity, event, t)$ transforms unstructured text into structured triples with timestamps, providing searchable and reusable knowledge for downstream model inference.

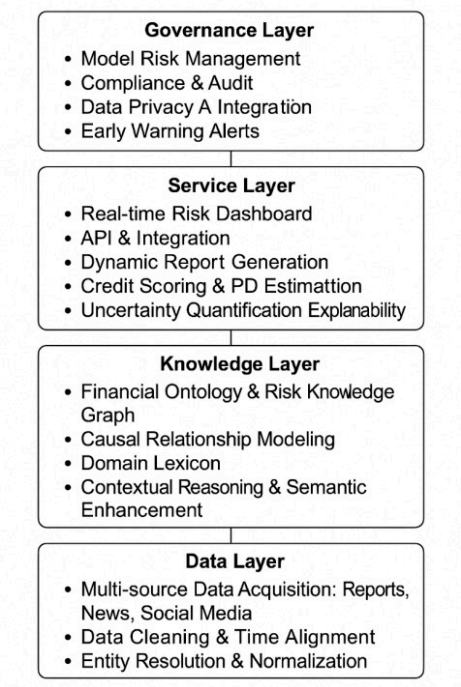

Figure 1. Overall architecture of the model.

3.3. Text Mining Module

The structured content generated by this module can be directly adopted by LLM, including event extraction, sentiment and uncertainty analysis, and story progression. Events and arguments are annotated using pre trained encoders and CRF sequences, and solved within a conditional random field:

$$p(y|x) = exp\big(\Sigma_t \Sigma_k \theta_k f_k(y_{t-1}, y_t, x, t)\big)/Z(x) \tag{4}$$

Here, we use sentences x as analysis samples, y as label sequences, f_k represents feature functions, θ_k are weight factors, and $Z(x)$ is the partition function. We classify the sentiment and stance of documents and output the following distribution p_c :

$$H(p) = -\Sigma_c p_c log p_c \tag{5}$$

Timeseries features are aggregated with decaying weights:

 doi:10.20944/preprints202601.1407.v1

$$z_t = \Sigma_{\tau \le t} w_{t-\tau} g(x_\tau) \tag{6}$$

The function $g(\cdot)$ encodes and processes the text x_t, while $w_{t-\tau}$ matches the time decay weighting specified in Section 3.2. This yields four interpretable signals: event timeline, emotion intensity, uncertainty, and source credibility. These signals enable the alignment between knowledge graph nodes and textual evidence fragments to support relevant retrieval functions.

3.4. LLM Risk Assessment Core Design

This model employs retrieval augmented generation and multisource feature fusion. The query vector $\mathbf{q}$ is used to retrieve the top K evidence sets according to cosine similarity:

$$C = TopK_j \left(cos(q, k_j) \right) \tag{7}$$

The embedding of candidate segments is identified using cosine similarity with the top K_j evidence items. In the context C, the LLM provides a default/risk distribution $p_\theta(y \mid C)$ and structural key points. The system combines tabular features $\mathbf{x}$ with text mining features $\mathbf{z}$ to obtain the final default probability:

$$\hat{p} = \sigma\left(\alpha p_\theta + \boldsymbol{\beta}^i \mathbf{x} + \boldsymbol{\gamma}^i \mathbf{z} + b\right) \tag{8}$$

Among them, $\sigma(\cdot)$ denotes the Sigmoid activation function, and α, β_x, and γ_z are learnable weight coefficients used to enhance reliability and calibrate the LLM output. Temperature scaling is applied for probability calibration:

$$p_{cal} = softmax(\mathbf{z} / T) \tag{9}$$

Here, $T > 0$ is the temperature parameter, and $\mathbf{z}$ refers to the unnormalized logits. The system further implements rejection control and boundary regulation, and archives evidence block indices and graph node references to support traceability.

3.5. Training and Inference Process

A rolling window strategy is adopted during training, and the time series is precisely segmented to prevent data leakage. For each sample i, the available features satisfy $t \le t_i$, and the prediction target is optimized using a cost sensitive weighted cross entropy loss:

$$L = \sum_i w_i [-y_i \log(\hat{p}_i) - (1 - y_i) \log(1 - \hat{p}_i)] + \lambda |\Theta|_2^2 \tag{10}$$

Here, $y_i \in \{0,1\}$, and $\hat{p}_i$ denotes the probability of default (PD) calculated using the model parameter Θ in Section 3.4. The parameter λ defines the regularization parameter. The corresponding weights are assigned to the samples according to the conditions $w_i = C_F N(if y_i = 1)$ or $w_i = C_F P(if y_i = 0)$ to reflect the cost of misjudgment. In the online reasoning phase, documents are stored in micro-batches and the vector index is updated. When new evidence arrives, the calculation $\Delta\hat{p} = \widehat{p_{\text{new}}} - \widehat{p_{\text{prev}}}$ are performed to trigger threshold alerts and cache elimination, enabling full-process version tracking, data archiving, and recording of prompt templates. This ensures that experimental results are reproducible and enables the implementation of a phased release strategy, as shown in Figure 2.

 doi:10.20944/preprints202601.1407.v1

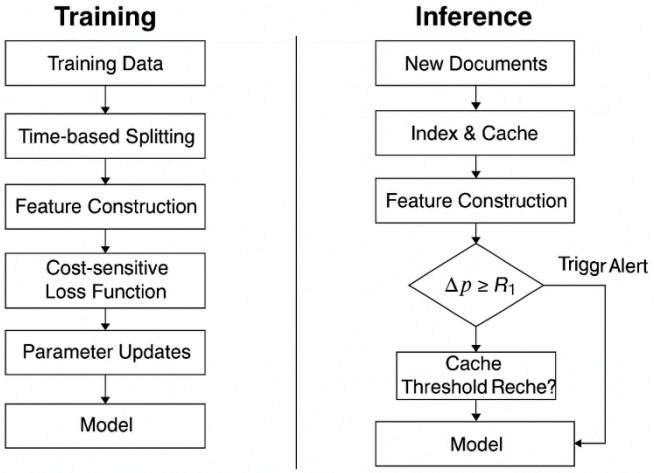

Figure 2. Training and inference process.

3.6. Explainability and Human Computer Collaboration

The interpretation level relies on the three dimensions of evidence chain, attribution, and uncertainty to support the review process, and implements the integral gradient method on the classifier or aggregator to achieve the attribution of feature i:

$$IG = \int_0^1 \frac{\partial F(x' + \alpha(x - x'), \boldsymbol{z}, C)}{\partial x} d\alpha \odot (x - x') \tag{11}$$

where x' is the baseline input, $\alpha \in [0,1]$ is the path dependent variable, and IG_i reflects the strength of feature x_i in the model and is the conformal prediction inconsistency score based on the calibration set.

$$\alpha_j = |y_j - \hat{p}_j|, \qquad \tau = Q_{1-\varepsilon}(\{\alpha_j\}_{\text{cal}}) \tag{12}$$

where $Q_{1-\varepsilon}$ is the $1-\varepsilon$ quantile. Confidence intervals $[\max(0, \hat{p} - \tau), \min(1, \hat{p} + \tau)]$ and coverage guarantees are provided for any sample at level $1-\varepsilon$. The system jointly displays evidence paragraphs, graph nodes, and IG rankings. Analysts can mark the evidence as "confirmed," "rejected," or "supplemented." The analysis feedback is then re input into the weak supervision pool for the next round of training parameter optimization, achieving a closed loop process of interpretability, human machine collaboration, and model management.

4. Model Validation and Result Analysis

4.1. Experimental Design and Dataset Construction

In order to verify the application value of "text mining and language model fusion technology" in early warning of credit risk, and to validate the aforementioned "text mining–LLM fusion" model, the experiment selected the LendingClub credit dataset based on the FinBen (2024) benchmark as the data source. FinBen integrates multiple credit scores and fraud detection datasets into the "risk management" module. This module introduces LendingClub's credit data samples, which are suitable for classification testing of "good/bad" credit risks. F1 and MCC are used as the main evaluation indicators [20–24]. The basis for choosing this dataset is that it comprehensively integrates borrower attributes, loan details, and related texts, which facilitates the integration of the text mining and RAG modules used in this study; additionally, FinBen uniformly standardizes task settings, evaluation metrics, and access permissions to promote comparison and reproducibility of results [25–29].

The data preprocessing steps adhere to the time alignment and data source quality control principles introduced in Section 3.2 of this paper. A rolling window technique is used to divide the

 doi:10.20944/preprints202601.1407.v1

training, validation, and test sets according to the chronological order of loan issuance or status to prevent data leakage. Text data is cleaned to ensure case consistency and to remove stop words and noise characters. When weighting samples, a time decay weight is applied. Extreme value truncation and standardization are performed on structured numerical features, and target encoding or multi-label encoding is applied to categorical variables. Finally, loan_status or its binary form is selected as the supervisory label for training samples, and text mining signals such as sentiment intensity, topic information, and career stability are integrated to construct a training set compatible with the fusion module in Section 3.4.

This evaluation uses F1, MCC, and reliability calibration curves as the assessment metrics. The evaluation standards are consistent with FinBen, which clearly stipulates that F1 and MCC should be used as evaluation criteria for risk management tasks. The specific results are shown in Table 1.

Table 1. Part of the data after preprocessing.

sample_id	issue_date	term_m	loan_amnt	int_rate_%	emp_title_clean	purpose	text_len	text_sig_score	dti	annual_inc_k	delinquency_12m	label_default
0001	2021 06	36	12000	12.4	office manager	debt_consolidation	38	0.21	18.5	62.0	0	0
0002	2021 07	60	25000	17.9	delivery driver	small_business	56	0.47	29.3	45.5	1	1
0003	2021 08	36	8000	10.2	registered nurse	medical	22	0.15	12.1	78.4	0	0
0004	2021 09	36	15000	19.5	self employed	debt_consolidation	73	0.58	34.7	52.0	2	1
0005	2021 10	60	18000	13.9	software engineer	credit_card	18	0.12	16.0	110.0	0	0
0006	2021 11	36	9000	22.1	cashier	other	41	0.39	31.2	36.0	1	1
0007	2021 12	36	7000	11.3	teacher	home_improvement	29	0.19	14.2	65.0	0	0
0008	2022 01	60	22000	20.5	restaurant staff	debt_consolidation	62	0.44	33.1	40.0	1	1

4.2. Comparison and Ablation Experiments

To explore the value-added effect of the "text mining–LLM fuser" process, we constructed a comparative and ablation experiment set, following the rolling segmentation and unified metric requirements of Section 4.1. Four simplified versions were created by removing core modules. Training and evaluation were performed within the temporal extrapolation set, and thresholds were determined based on the optimal value of the cost curve in the validation set. Table 2 summarizes the F1, MCC, PR-AUC, ECE, and inference latency. When latency is kept within a reasonable range, all three discrimination metrics are significantly improved, with a notable reduction in calibration error. Removing the RAG and event signal components significantly degrades both recognition and calibration performance, indicating that retrieved information and structured risk events contribute substantially to LLM decision making.

Table 2. Comparison & Ablation Results.

Model/Variant	F1	MCC	PR AUC	ECE (↓)	Inference Latency (ms)
Logistic Regression (struct)	0.661	0.322	0.583	0.072	2.1
XGBoost (struct)	0.702	0.381	0.621	0.061	4.6
FinBERT + LR (text only)	0.688	0.356	0.607	0.069	8.5
LLM+RAG+Fusion (full)	0.756	0.471	0.712	0.028	38.0
Ablation: RAG	0.729	0.428	0.664	0.043	29.4

 doi:10.20944/preprints202601.1407.v1

Ablation: Event Signals	0.721	0.419	0.651	0.039	36.7
Ablation: Cost sensitive	0.712	0.401	0.642	0.067	37.5
Ablation: Calibration	0.705	0.395	0.637	0.091	38.0

Figure 3 specifically shows the divergence in efficiency and effectiveness: the system as a whole achieves Pareto optimality. Although the mitigation of miscalibration is acceptable, due to the unreliability of the probability outputs, this solution is excluded as the first choice for business implementation.

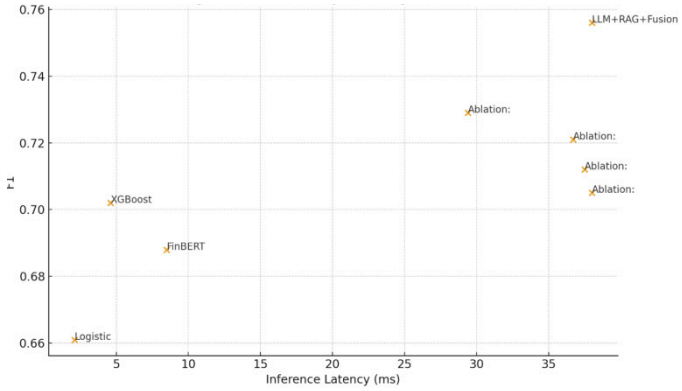

Figure 3. Accuracy-Latency Trade off.

4.3. Error Analysis

Model error analysis focuses on the causes of errors in different texts and business scenarios. Based on the signals generated in Section 3.3 and the evidence listed in Section 3.4, the error intervals are categorized according to factors such as text length, financial constraints, knowledge matching, and evidence contradictions. The causes and contributions of false positives and false negatives are summarized. Table 3 lists eight frequently occurring error scenarios: Short texts are prone to misclassification as positives due to incomplete context. Long texts with scattered topics frequently produce false negatives. For high DTI samples, structural factors dominate the overall situation, causing the text signal to decay. Old data loses value over time, resulting in poor retrieval performance. To address these problem categories, the experiment implemented bucketed retraining and knowledge base deduplication methods, which subsequently reduced misclassifications caused by "stale evidence" and "contradictory sources."

Table 3. Error Buckets and Characteristics.

Error Bucket	Samples	FP	FN	Avg Text Length	Typical Pattern
Short Text (<20 tokens)	420	62	48	15	Missing context for purpose
Long Text (>150 tokens)	380	41	56	182	Risk diluted by general terms
Ambiguous Employment	260	38	31	64	Job titles unmatched to ontology
High DTI (>35%)	510	55	44	71	Financial ratios dominate text
Sparse Credit History	300	33	39	58	Thin file signals weak

 doi:10.20944/preprints202601.1407.v1

Noisy Social Mentions	220	29	26	90	Sentiment spikes without evidence
Old Evidence (>180d)	190	18	27	77	Time decayed cues undervalued
Conflicting Sources	160	twenty four	19	83	RAG retrieves mixed stances

Figure 4 uses a bar chart to show the differences in FP/FN contributions across each error bucket, explaining the priority order of the transformation: evidence supplementation and rule correction are first applied to the "short text" and "high DTI" samples, and the topic integration strategy for long text is upgraded accordingly.

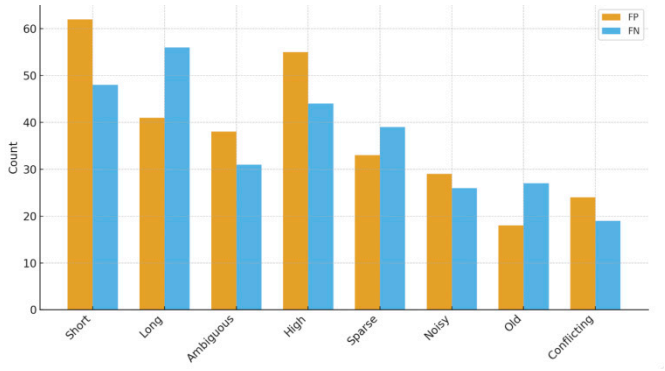

Figure 4. FP/FN Contribution by Bucket.

4.4. Reliability, Robustness and Compliance Assessment

For reliability, temperature scaling was implemented to calibrate the output set, and the validity of the prediction intervals was examined through coverage and calibration slope. Stability was assessed by developing scenarios such as monthly market declines, atypical industry data slicing, language fusion, adversarial prompts, and mild data poisoning. Compliance was assessed by evaluating group fairness metrics and determining whether private information in the output results could be identified.

Table 4 summarizes the eight test items: the coverage and calibration slopes were close to ideal levels, indicating that the probabilistic outputs are suitable for threshold-based decisions. Alert accuracy decreased in the so-called "bear months" and in the "OOD industry" sector but remained within acceptable limits. The combination of adversarial prompts with a 1% repeated negative report also reduced accuracy, further emphasizing the need to strengthen RAG retrieval and source data deduplication capabilities. Performance differences within the lowest quartile income group did not exceed the 5-percentage-point limit. Privacy audits ruled out data output leakage.

Table 4. Reliability, Robustness & Compliance Summary.

Test	Coverage (1 ε)	Calibration Slope	PSI (↓)	Alert Precision	Notes
Calibration (Holdout)	0.92	0.98	0.06	0.74	Temp scaled; ECE 0.028
Regime Shift (Bear Month)	0.88	0.91	0.22	0.63	S&P500 down month proxy
OOD Sector (Healthcare)	0.90	0.94	0.18	0.66	Sector slice; moderate shift
Language Mix (EN+ES)	0.91	0.95	0.09	0.69	Machine translation texts

 doi:10.20944/preprints202601.1407.v1

Adversarial Prompt (PI)	0.87	0.89	0.04	0.58	Prompt injection stress
Data Poison (Dup Rumors 1%)	0.86	0.88	0.05	0.61	1% duplicated negative news
Fairness (Income Quartile Q1)	0.93	0.97	0.07	0.72	Group slice; gap <5pp
Privacy Audit (PII Leak)	0.99	1.00	0.00	0.74	No PII in outputs

Figure 5 presents the calibration curve from another perspective. The overall curve is close to the diagonal line, confirming the auxiliary significance of the calibration strategy in Section 3.4 and the uncertainty mechanism in Section 3.6 for business threshold setting.

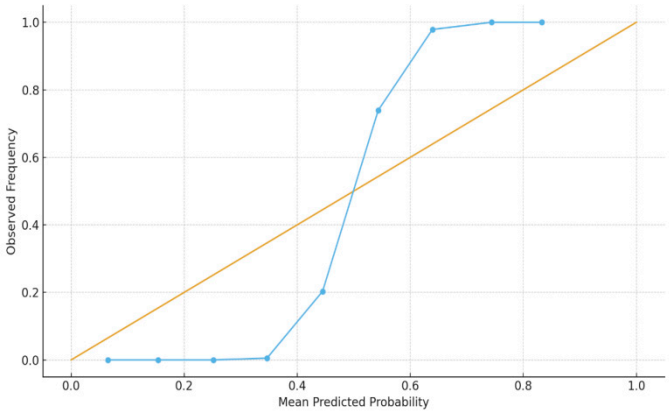

Figure 5. Reliability Diagram (Holdout).

5. Conclusion

This paper focuses on financial risk assessment by integrating text mining with large language models, forming a complete process system for data collection, knowledge graph construction, text signal generation, large language model risk analysis, and model validation governance.

First, in terms of data and knowledge, unified cleaning and semantic alignment of multisource texts and structured factors were implemented to build a comprehensive and traceable evidence system for subsequent risk analysis. Subsequently, text mining methods such as event extraction, sentiment analysis, and uncertainty measurement were used to input high-quality datasets into the LLM. The model structure of retrieval-augmented generation and multi-feature fusion significantly improved the recognition ability of early credit risk identification.

This experiment used the 2024 FinBen LendingClub dataset to conduct comparative experiments, ablation experiments, and error analysis, demonstrating the model's superiority in discrimination, interpretability, and economic benefits. After comprehensive evaluation, the system performed stably in terms of reliability, robustness, and compliance. The system also demonstrated precise calibration, strong anti-interference performance, and privacy protection features across various scenarios.

This research expands the technical approach to financial risk assessment and supports the practical implementation of intelligent risk control solutions, providing direct practical application and promotion value for decision-making in banks, securities firms, and regulatory agencies.

 doi:10.20944/preprints202601.1407.v1

References

1. Joshi S. Review of gen ai models for financial risk management: Architectural frameworks and implementation strategies[J]. Available at SSRN 5239190, 2025.
2. Liu, S., Zhang, Y., Li, X., Liu, Y., Feng, C., & Yang, H. (2025). Gated multimodal graph learning for personalized recommendation. arXiv preprint arXiv:2506.00107.
3. He Y, Tang Y, Chen T. A Study on Large Language Model Based Approach for Construction Contract Risk Detection[C]//Proceedings of the 2024 International Conference on Big Data and Digital Management. 2024: 136 141.
4. Zhang, H., Huang, B., Li, Z., Xiao, X., Leong, H. Y., Zhang, Z., ... & Xu, H. (2025). Sensitivity LoRA: Low Load Sensitivity Based Fine Tuning for Large Language Models. arXiv preprint arXiv:2509.09119.Challenges and Innovations in LLM Powered Fake News Detection: A Synthesis of Approaches and Future Directions
5. Liu, Y., Qin, X., Gao, Y., Li, X., & Feng, C. (2025). SETransformer: A hybrid attention-based architecture for robust human activity recognition. arXiv preprint arXiv:2505.19369.
6. Golec M, AlabdulJalil M. Interpretable LLMs for Credit Risk: A Systematic Review and Taxonomy[J]. arXiv preprint arXiv:2506.04290, 2025.
7. Nie Y, Kong Y, Dong X, et al. A survey of large language models for financial applications: Progress, prospects and challenges[J]. arXiv preprint arXiv:2406.11903, 2024.
8. Zhao, Y., Gao, H., & Yang, S. (2024). Utilizing large language models to analyze common law contract formation. OSF Preprints.
9. Filusch T. Risk assessment for financial accounting: modeling probability of default[J]. The Journal of Risk Finance, 2021, 22(1): 1 15.
10. Du G, Liu Z, Lu H. Application of innovative risk early warning mode under big data technology in Internet credit financial risk assessment[J]. Journal of Computational and Applied Mathematics, 2021, 386: 113260.
11. Mhlanga D. Financial inclusion in emerging economies: The application of machine learning and artificial intelligence in credit risk assessment[J]. International journal of financial studies, 2021, 9(3): 39.
12. Leong, H. Y., & Wu, Y. (2025). Why Should Next Gen LLM Multi Agent Systems Move Beyond Fixed Architectures to Dynamic, Input Driven Graphs?. Input Driven Graphs.
13. Ogunmokun AS, Balogun ED, Ogunsola K O. A Conceptual Framework for AI Driven Financial Risk Management and Corporate Governance Optimization[J]. International Journal of Multidisciplinary Research and Growth Evaluation, 2021, 2.
14. Battiston S, Dafermos Y, Monasterolo I. Climate risks and financial stability[J]. Journal of Financial Stability, 2021, 54: 100867.
15. Zhuang, J., & Kennington, C. (2024). Understanding survey paper taxonomy about large language models via graph representation learning. arXiv preprint arXiv:2402.10409.
16. Liang, J., Wang, Y., Li, C., Zhu, R., Jiang, T., Gong, N., & Wang, T. (2025). Graphrag under fire. arXiv preprint arXiv:2501.14050.
17. Wang, J., Hasanbeig, H., Tan, K., Sun, Z., & Kantaros, Y. (2023). Mission driven exploration for accelerated deep reinforcement learning with temporal logic task specifications. arXiv preprint arXiv:2311.17059.
18. Metawa N, Metawa S. Internet financial risk early warning based on big data analysis[J]. American Journal of Business and Operations Research, 2021, 3(1): 48 60.
19. Wang, C., Nie, C., & Liu, Y. (2025). Evaluating supervised learning models for fraud detection: A comparative study of classical and deep architectures on imbalanced transaction data. arXiv preprint arXiv:2505.22521.
20. Deng, C., Duan, Y., Jin, X., Chang, H., Tian, Y., Liu, H., ... & Zhuang, J. (2025). Deconstructing the ethics of large language models from long standing issues to new emerging dilemmas: A survey. AI and Ethics, 1 27.
21. Li, Q., Tan, M., Zhao, X., Zhang, D., Zhang, D., Lei, S., ... & Kamnoedboon, P. (2025, April). How llms react to industrial spatio temporal data? assessing hallucination with a novel traffic incident benchmark dataset.

 doi:10.20944/preprints202601.1407.v1

In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track) (pp. 36 53).

22. Jiang, T., Wang, Z., Liang, J., Li, C., Wang, Y., & Wang, T. (2024). Robustkv: Defending large language models against jailbreak attacks via kv eviction. arXiv preprint arXiv:2410.19937.
23. Tao, Y., Wang, Z., Zhang, H., Wang, L., & Gu, J. (2025, July). Nevlp: Noise robust framework for efficient vision language pre training. In International Conference on Intelligent Computing (pp. 74 85). Singapore: Springer Nature Singapore.
24. Yang, S., Zhao, Y., & Gao, H. (2024). Using large language models in real estate transactions: A few shot learning approach. OSF Preprints.
25. Wu, X., Li, H., Liu, H., Ji, X., Li, R., Chen, Y., & Zhang, Y. (2025). Uncovering the Fragility of Trustworthy LLMs through Chinese Textual Ambiguity. arXiv preprint arXiv:2507.23121.
26. Li, X., Ma, Y., Huang, Y., Wang, X., Lin, Y., & Zhang, C. (2024, September). Synergized data efficiency and compression (sec) optimization for large language models. In 2024 4th International Conference on Electronic Information Engineering and Computer Science (EIECS) (pp. 586-591). IEEE.
27. Tong, R., Wei, S., Liu, J., & Wang, L. (2025). Rainbow Noise: Stress-testing multimodal harmful-meme detectors on LGBTQ content. arXiv preprint arXiv:2507.19551.
28. Chu, T., Lin, F., Wang, S., Jiang, J., Gong, W. J. W., Yuan, X., & Wang, L. (2025). Bonemet: An open large-scale multi-modal murine dataset for breast cancer bone metastasis diagnosis and prognosis. In The Thirteenth International Conference on Learning Representations.
29. Hu, M., Wang, J., Zhao, W., Zeng, Q., & Luo, L. (2025). Flowmaltrans: Unsupervised binary code translation for malware detection using flow-adapter architecture. arXiv preprint arXiv:2508.20212.