

Research on Automatic Generation of Financial Research Reports Based on Large Model Integration and Small Discrimination Models

Zuoning Shi

Division of Mathematics and Artificial Intelligence, Qilu University of Technology, Jinan,
Shandong, 250000, China
Shi060116@yeah.ne

ABSTRACT

Although large language models have made significant progress, automatic generation of financial research reports is still limited by quality fluctuations of single-model architectures and insufficient domain expertise. Existing research mainly focuses on isolated model applications, while multi-model collaborative mechanisms and modular quality assessment systems for the rigor of financial text structure remain lacking.

To address these issues, this paper proposes a dual-layer architecture of “multi-generation models + lightweight discriminative models.” Specifically, heterogeneous large models like DeepSeek, Kimi, and Gemini are used to generate multiple versions of research reports, and prompt design is optimized through retrieval-augmented generation (RAG); Establishing a five-dimensional quality assessment framework, fine-tuning GPT-5-mini based on high-scoring Guosheng Securities research reports and training data from 50 industry research reports, achieving professional alignment across fields; Furthermore, a modular optimization strategy and a cross-model consistency validation algorithm are designed to integrate the strengths of each generator.

Main innovations: Establishing a three-layer paradigm of “generation—evaluation—integration,” extending large model integration theory to professional financial text production; A fine-tuning method for domain alignment discrimination models for multidimensional quality assessment is proposed; A modular preferential fusion mechanism is designed to tap into the complementary advantages of heterogeneous models while ensuring structural and logical consistency.

This study deepens the theoretical understanding of multi-model collaboration in domain-specific text generation, providing a scalable framework for automated production of financial research reports. The proposed “manual control + intelligent generation” human-machine collaboration protocol clarifies the intervention trigger mechanism under extreme market conditions, balancing automation efficiency with professional prudence.

Keywords: Large Language Model; Model Integration; Financial Research Reports; Text Generation; Quality Assessment

1. INTRODUCTION

Financial research reports, abbreviated as research reports, are professional research reports in the financial industry. These reports are usually written by financial institutions or academic research organizations and are used for in-depth analysis of financial markets, industries, companies, and economic trends, providing decision-making references for investors, businesses, and governments.

Financial research reports, as highly specialized, structurally complex, and information-dense texts, have a growing demand for automated processing.

Currently, generative artificial intelligence and large language models (LLMs) demonstrate powerful capabilities in natural language understanding, reasoning, and generation, enabling automation and intelligence in financial research reports.

Each large model generates different language styles and depth of thought chains, allowing small models to better compare the generated research reports and produce high-quality reports.

2. LITERATURE REVIEW

2.1 Related Cases

2.1.1 Relevant foreign cases

In earlier papers, Enhancing Credit Risk Reports Generation using LLMs: An Integration of Bayesian Networks and Labeled Guide Prompting ^[1]. A new method combining Bayesian networks and “tag-guided prompts” was proposed, using GPT-4 to generate high-quality, credible credit risk reports and outperform human analysts in experiments. This has subsequently led to the trend of large models influencing the financial industry.

2.1.2 Relevant domestic cases

In recent years, domestic large models have emerged, launching three financial large models:

Huaxia Bank has independently developed the Longzhi Kangxi quantum foundational large model ^[2].

By applying improved algorithms, mechanisms, computing resources, computing power, and models to enhance performance, it builds low-cost, high-efficiency AI, but the released version has only 14B at its maximum and has relatively few parameters.

Liu Shu (2025) ^[3] used KIMI for financial auditing.

The document provides the main differences, but the list of differences is incomplete.

Renmin University “Yulan · Integrated Sight” (Yulan-FinSight) ^[4].

Code-driven Variable Memory (CAVM) architecture and iterative visual enhancements.

The “Kuangshi” financial large model from SCCAI and FinAgent ^[5].

Build a high-quality financial corpus and knowledge graph, integrating human feedback.

The above case shows that the development of domestic financial large models has deeply penetrated from general capabilities to vertical fields, but in multi-model collaboration,

Key aspects such as generative quality assessment and modular integration still lack a systematic research framework.

2.2 Text readability

Research on the readability of financial texts can be traced back to the pioneering work of Li (2022) ^[6], which was the first large-scale test of the readability of 10-K annual reports. Research indexed by Web of Science shows that traditional metrics such as the Gunning Fog Index and Flesch-Kincaid Grade Level are widely used to measure the complexity of financial documents. Specifically, the Fog Index estimates the required years of education for reading comprehension by comparing sentence length to complex words, with the formula:

$$\text{Fog Index} = 0.4 \times \left[\frac{\text{sentences}}{\text{words}} + 100 \times \frac{\text{complex words}}{\text{words}} \right] \quad (1)$$

Loughran & McDonald (2014) ^[7] proposed in a highly cited Web of Science paper that traditional metrics suffer from the problem of “misinterpretation of financial terms”—complex financial terms (such as “amortization” and “derivatives”) are mistakenly classified as “difficult terms,” leading to an underestimation of readability. To this end, they developed alternative metrics based on financial vocabulary lists, such as LM dictionary coverage and Campbell Harvey financial term density.

2.3 Research Progress on Hall Detection in Large Language Models

Current detection methods can be summarized into three categories, which are systematically explained in the review review of the Journal of Software ^[8] (CNKI index, CCF Chinese Category A):

Model uncertainty-based methods: detecting hallucinations through internal signals such as token probability and entropy value. For example, SelfCheckGPT calculates BERTScore consistency through multiple samplings.

External knowledge validation method: Using a retrieval-augmented generative (RAG) architecture, model outputs are compared with knowledge bases. Web of Science research shows that external knowledge can significantly improve detection accuracy, but search quality remains a key variable.

Large model-based agent-based approach: Using LLMs as judges (LLM-as-a-judge), directly assessing hallucination through template design.

In financial research report generation scenarios, the above three methods each have their own advantages and disadvantages: uncertainty methods have high computational efficiency but may miss logical errors; External verification is highly reliable but relies on the quality of the knowledge base; Agents have good flexibility in evaluation but can have bias. This study integrates two strategies: RAG external validation and discriminant model evaluation, aiming to balance efficiency and accuracy.

3. METHODS AND EXPERIMENTAL DESIGN

3.1 Research Objectives

This study focuses on the reliability and readability of research reports, using GPT-5-mini as a small discrimination model to conduct multidimensional evaluation and horizontal comparison of research reports generated by the three major mainstream large models—DeepSeek, Kimi, and Gemini—to verify the differences in capabilities and stability performance of different large models in financial research report generation tasks. On this basis, this study attempts to integrate the strengths of each model through paragraph-level merit selection and text fusion strategies, generating comprehensive research reports with better quality and stronger consistency, providing a reusable technical path for automated production of financial texts.

3.2 Research Content

This article selects BYD Company in the second quarter of 2025 as a research case, integrating data around core indicators such as total revenue, net profit attributable to shareholders, gross margin, net margin, R&D investment, and industry competitive landscape. The experimental procedure is as follows:

First, the integrated structured data was fed into three major models: DeepSeek, Kimi, and Gemini. Each model generated two independent drafts, totaling six initial research reports. Next, the six drafts were broken down by modules (Abstract, Industry Analysis, Company Analysis, Financial Analysis, Investment Risk & Recommendations) and submitted to the fine-tuned GPT-5-mini discrimination model for paragraph-level scoring. Third, based on the scoring results, the top two paragraphs in each module are selected as candidates, and text fusion algorithms are used for integration and consistency checks to generate a new comprehensive research report. Finally, the quality of the merged research reports is evaluated again, and the best outputs of a single model are compared horizontally to verify the effectiveness of the fusion strategy.

3.3 Technical Route

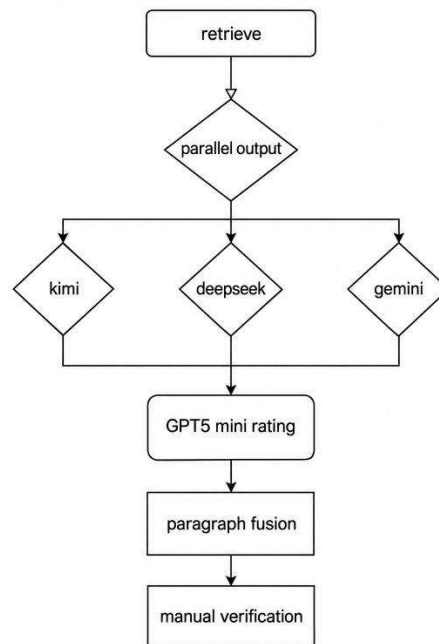


Figure 1. Technical Roadmap.

3.3.1 Large Model Generation Layer

Prompt Engineering:

Character Setting: “You are a financial industry analyst, skilled at writing high-quality research reports.”

Style constraints: “The text content is logically rigorous, insightful, and focuses on narration over viewpoints; The language style is professional and cautious, with clear layers.”

Structure Template: Summary → Industry Analysis → Company Analysis → Financial Analysis → Investment Recommendations → Risk Alerts RAG (Retrieval-Enhanced Generation).

Debugging:

Integrate BYD’s Q2 2025 financial data and market environment information, vector search for historical research reports and industry data.

3.3.2 Evaluation Layer for Small Discriminative Models

After the large model completed the generation of research reports, six reports were submitted to the small model. This study established a five-dimensional quality assessment system that meets financial industry standards, using a 1-10 point scoring scale.

You are a professional expert in financial text quality assessment, with a dual background in CFA and NLP analysis.

1. Language Clarity (1-10 points): Whether terminology is explained sufficiently and whether the sentence structure is concise and easy to understand.
2. Structural Logic (1-10 points): Whether the paragraph organization is reasonable, and whether the argument chain is clear and closed.
3. Information density (1-10 points): The amount of effective information per unit length.
4. Professional threshold (1-10 points): Difficulty for non-professional readers to understand, and whether it is friendly.
5. Logical depth (1-10 points): Whether the causal reasoning is rigorous and whether the predictions are sufficiently based.

3.3.3 Prompt Analysis Framework

In addition to the above five-dimensional scoring, the discrimination model also needs to be analyzed in depth from the following three perspectives:

Argument completeness: Whether the assumptions → data analysis → conclusion derivation is closed.

Causal logic: whether correlations are mistaken for causality, and whether variables are controlled by confusion.

Risk disclosure: Whether potential risks have been adequately discussed.

Model Training Strategy:

High-score sample study: Selected Guosheng Securities' "Fixed Income Review: How to Understand the Deviation Between Social Financing and Monetary Growth" as a high-scoring reference (9,9,8,8,9).

Data fine-tuning: Using data from 50 recent industry research reports as the training set, the average score is calibrated to a range of 7.5-8.0 through supervised fine-tuning, aligning with actual industry standards to avoid overly lenient or harsh ratings.

3.3.4 Text Fusion Layer (Fusion Layer)

This layer is used to integrate and merge the high-scoring paragraphs after scoring to generate a new research report. The specific design is as follows:

Integrated prompt design:

JSON {"instruction": "

Each paragraph in the paragraphs is merged into a draft chapter according to its original meaning;

Each revision listed in corrections is executed sequentially (replace/insert/delete, etc.),

"output_format":

{"merged_text": {"section_id": "text"},

"change_log":

[{"action": "replace_text", "paragraph_id": "A3", "detail": "..."}], "pending_verifications": [{"target": "...", "reason": "Unverified"}]},

"quality_control":

"If the verifications mark as unverified or missing, please add the '[verification needed]' mark at the corresponding position in the final text"

The core advantages of this design are: on one hand, it preserves the complete fused text through merged_text fields; On the other hand, by change_log recording detailed changes of each modification, process traceability is achieved; More importantly, through pending_verifications mechanisms, explicit annotation of data or conclusions that cannot be automatically verified not only increases output transparency but also effectively resists information illusions, providing a clear entry point for subsequent manual verification.

3.3.5 Evaluation and Validation

3.3.5.1 Comparison of Model Texts with Fusion Texts

Control group: single model (DeepSeek/Kimi/Gemini) generated in full.

Experimental Group: Module selection + text fusion generation.

3.3.5.2 Fusion of long text and reduced text

Used for word count control, polishing 2,808 long texts to 1,550 simplified texts to optimize information density.

A secondary evaluation of the simplified text was conducted.

3.3.5.3 Manual sampling inspection

Researchers with financial industry backgrounds were invited to manually review the automated scoring results, focusing on checking paragraphs with abnormally high or low scores to verify the rationality of the automatic scoring and the reliability of the discrimination model.

3.4 Feasibility Analysis

Based on computers and large language models, large models generate databases with strong critical capabilities for both large and small models, both mature and persuasive.

3.5 Key issues to be addressed

How can the performance differences of different large language models in the generation of financial research reports be quantified?

How can an effective evaluation system be designed to assess the quality of generated content?

How can optimal fusion of multi-model generation results and text consistency processing be achieved?

4. EXPERIMENTAL PROCESS

4.1 Research Methods

Combining experimental research methods with case analysis:

The experimental subjects were BYD (2025Q2) research report generation tasks.

Data sources include company financial reports, market environment data, and industry research reports.

The model selection adopted three mainstream models: DeepSeek, Kimi, and Gemini (each generating two versions) as the generative models.

The reason for choosing BYD is that, as a leading company in the new energy vehicle industry, BYD's business spans automobiles, batteries, electronics, and other sectors. Its financial structure is complex, and the competitive landscape is ever-changing, which places high demands on the logical depth and information density of research reports.

4.2 Evaluation and Selection

Table 2. Five-dimensional comparison diagram.

Model	Abstract	Industry analysis	Company analysis	Financial analysis	Investment risks and recommendations
kimi-1	7.2	7.8	7.6	8.2	8.0
deepseek-1	7.2	7.4	7.2	7.8	7.4
gemini-1	7.8	7.6	7.4	7.8	7.2
kimi-2	7.6	7.8	7.8	8.0	7.8
deepseek-2	7.8	6.8	7.3	7.8	7.0
gemini-2	7.8	7.6	8.0	8.2	7.4

This table is a score table for research reports generated by large models and evaluated by small models. It is used to select the top two scores in each segment for integration. It was observed that DeeEPSEEK2 showed significant fluctuations in industry analysis (7.4 to 6.8), while Kimi's performance was more stable, so the stable scoring paragraph was selected. For the summary section, deepseek2 and Gemini2 were chosen. Since the models were duplicated, the Gemini-2 after secondary training was chosen.

4.3 Examples of Integrated Text Design

P2: "Fangchengbao delivers 100,000 yuan in 18 months + Yangwang over 10,000 yuan + fifth-generation DM fuel consumption 2.6L + Tiandi's Eye daily data 44 million kilometers"

P3: “Brand synergy (Dynasty Stable / Fangchengbao Kuai / Denza D9 / Yangwang over 10,000 units) + globalization (3 countries building factories + roll-on/roll-off) + technology (2.6L/10C fast charging/silicon carbide)”

Post-integration output: BYD has built a clear brand gradient matrix: “Dynasty/Ocean” firmly holds its foundation, “Fangchengbao” delivers over 100,000 units in 18 months, verifying its youth-oriented strategy, “Denza D9” becomes the world’s first new energy MPV to accumulate 250,000 units, and the “Yangwang” high-end series surpasses 10,000 units, demonstrating brand elevation potential. On the technical side, the fifth-generation DM technology reduces fuel consumption to 2.6L when battery depleted (the world’s lowest), and the Super e platform features 10C flash-charging batteries and silicon carbide power devices; In terms of intelligence, the “Eye of the God” advanced intelligent driving system has been extended to 100,000-yuan models, with the cumulative number of intelligent driving models expected to exceed 710,000 by the end of 2025, generating 44 million kilometers of data daily, providing a solid foundation for algorithm iteration [verification needed]. Globalization has entered a harvest period simultaneously: factories in Cambodia, Hungary, and Brazil have successively started production, six self-owned roll-on/roll-off ships have been put into operation, supporting products in over 110 countries, and overseas gross margin (19.82%) surpassed domestic sales by 2.85 percentage points, reflecting localized operational effectiveness.

During the fusion phase, the merged text is generated according to the JSON format, along with the update log and the fields to be verified. This improves visibility and resists information illusions, making manual verification easier.

4.4 Comparison of Model Texts and Fusion Texts

According to reviews, the fused text contains 2,808 characters, classified as long texts. Its scores are: Language clarity: 9, structural logic: 9, information density: 10, professional threshold: 9, logical depth: 9, and average: 9.4. You can see that the overall five dimensions have improved, making the research report nearly perfect.

4.5 Reduction and Re-scoring

The model kimi was used for polishing and reduction. After re-grading, the scores for the 1550-word text are: Language clarity: 9, Structural logic: 10, Information density: 10, Professional threshold: 8, Logical depth: 9, Average: 9.4. The average score remains unchanged, indicating that the article’s professionalism has not decreased, and the word count has decreased, making the text more dense and easier to read.

5. CONCLUSION

5.1 Regarding large models

Large model strategies are effective. By comparing the generation performance of the three major models—DeepSeek, Kimi, and Gemini—in each module of BYD’s Q2 2025 research report, it is found that a single model struggles to maintain optimal performance across all dimensions. On the DeepSeek model, its logical reasoning is deep, but the generated text quality fluctuates and is not suitable for mass generation. The advantages of Kimi are precise financial analysis, strong data extraction capabilities, and fluent expression; its disadvantages are weak summarizing ability and lengthy articles. Gemini, on the other hand, has strong extraction capabilities and no obvious disadvantages. By adopting an integrated strategy of “selecting the best by module + text fusion,” the overall quality of research reports can be improved from an average of 7.8 for a single model to 9.4.

5.2 Regarding small models

After high-scoring research report training and fine-tuning data from 50 reports, the GPT-5-mini evaluation model demonstrates professionally aligned scoring capabilities:

Before training: Scores are generally high (8.5+), with insufficient differentiation.

After training: Scores are concentrated between 7.5 and 8.0, aligning with industry benchmarks.

The evaluation dimension covers five key indicators and can accurately identify professional issues such as inconsistent data usage and marketing presentations.

However, small models still have limitations. The evaluation ability of small models is limited by the timeliness of the training data;

High cost for multi-model API calls;

Under extreme market volatility, the rigor of forecast logic still requires manual scrutiny.

Its evaluation capabilities are limited by the timeliness of training data, making it difficult to capture the latest market trends and industry changes in real time; API costs are higher for multi-model collaborative calls, especially in large-scale, high-frequency scenarios;

Moreover, in the face of extreme market fluctuations, the rigor of model prediction logic may be insufficient, requiring manual review and final oversight by professionals.

5.3 Innovation of the Project

Applies a dual-layer architecture of “large model integration + small discrimination model” to financial research report generation.

Design a five-dimensional evaluation system that meets financial industry standards, solving the problem of traditional indicators being unapplicable.

Achieve modular generation and dynamic fusion, breaking through the capability limits of a single model.

REFERENCES

- [1] Teixeira, A. C., Marar, V., Yazdanpanah, H., et al. (2023). Enhancing credit risk reports generation using LLMs: An integration of Bayesian networks and labeled guide prompting. In *Proceedings of the 4th ACM International Conference on AI in Finance* (pp. 1-8). ACM. <https://doi.org/10.1145/3604237.3626902>
- [2] Gong Weihua. Huaxia Bank has accumulated 240 large model scenario requirements and has built an integrated AI integration team for technology lines and subsidiaries [EB/OL]. Mobile Payment Network, 2025-12-03. <https://www.mpaypass.com.cn/news/202512/03113846.html>
- [3] Liu Shu. Application of Generative AI in Financial Auditing—A Case Study of Kimi [J]. Auditing Research, 2025, (01): 42-50
- [4] Renmin University of China Hillhouse School of Artificial Intelligence. Yulan · Rongguan: The award-winning financial research report generation system is now open source! [EB/OL]. (2025-12-19)
- [5] Shanghai University of Finance and Economics. Shanghai University of Finance and Economics releases Kuangshi Finance and Education Large Model [EB/OL]. (2025-06-07)
- [6] Li F. A new measure of financial texts' readability[J]. Journal of Financial Economics, 2022. (SSCI Q1, Web of Science)
- [7] Loughran, T., & McDonald, B. (2014). Measuring readability in financial disclosures. The Journal of Finance, *69*(4), 1643–1671. <https://doi.org/10.1111/jofi.12162>
- [8] Wang Fang, et al. Review of Large Language Model Evaluation [J]. Journal of Software, 2024, 35(6): 2789-2812. (CCF Chinese Category A)