A Comparative Study on the Financial Term Comprehension Capabilities of Different Types of Large Language Models

Qirui Yang
Beijing University of Science and Technology, Beijing, China

Abstract: Since 2023, the application of large language models (LLMs) in the financial sector has moved from the proof-of-concept stage to practical, large-scale implementation. The level of a model's understanding of terminology not only determines its ability to perform practical tasks such as risk control, investment research, and customer service but also serves as a critical standard for regulators to assess the compliance of LLMs. This study selected eight representative models-GPT-4o, Claude Sonnet 4.5, Gemini 1.5 Pro, Qwen2.5-72B, ERNIE-Speed, Spark 3.5, FinGPT 4.0 Instruct, and FinBERT-Large-to establish a unified evaluation system for financial terminology, conducting explanation tasks on 200 financial terms. The research employed quantitative metrics such as text similarity, coverage, and response length, complemented by qualitative observation and budget analysis, to analyze the strengths, weaknesses, and applicable scopes of various models. The results show that financial domain-specific models hold significant advantages in semantic alignment and term recognition. International general-purpose models demonstrate stable performance in complex long-text tasks due to their multimodal and long-context processing capabilities. Domestic general-purpose models offer higher cost-effectiveness in large-scale deployment scenarios due to their low cost and strong adaptability to Chinese corpora. This paper also summarizes model combination strategies and provides recommendations for establishing industry evaluation standards, offering references for financial institutions in model selection, deployment, implementation, and management.

Keywords: Large Language Models; Financial Terminology; Model Evaluation; FinGPT; Compliance Governance

1. Introduction

1.1 Research Background

Since 2023, the People's Bank of China and the China Banking and Insurance Regulatory Commission have successively proposed in policy documents such as the *Financial Technology Development Plan (2022-2025)* to establish a "safe, controllable, and regulated" application system for financial large models, requiring financial institutions to simultaneously implement risk prevention, control, and compliance review during business innovation [1][2][6]. During the same period, domestic and international financial institutions have gradually introduced LLMs into key business areas such as investment research analysis, intelligent customer service, risk warning, and compliance auditing [3]. However, financial business processes heavily rely on standardized terminology, such as "capital adequacy ratio" in regulatory reports, "risk matching principle" in asset management agreements, and "hedging basis" in trading scenarios. If a model deviates in understanding or generating such terms, it may directly trigger compliance risks or business judgment errors, even misleading the market. Therefore, evaluating a model's ability to comprehend financial terminology has become a necessary task before the financial industry adopts LLMs.

1.2 Research Significance

1.2.1 Theoretical significance

First, this paper decomposes term comprehension ability into several dimensions (semantic similarity, coverage, length control, category adaptability) and examines the performance of various models across these dimensions, contributing to the refinement of industry-specific evaluation metrics within existing LLM assessment frameworks [12]. Second, through horizontal comparison of multiple model types, starting from the correlation between model architecture, training data, and task performance, it reveals how domain knowledge and general capabilities integrate, providing a theoretical basis for future construction of multimodal, multi-task financial models.

1.2.2 Practical significance

In practical operations, investment research teams need models to quickly understand financial indicators and industry terminology and provide explanations; customer service centers require models to accurately answer questions related to product terms; risk control and internal audit departments expect models to identify potential risk terms and provide compliance prompts [1][2]. By establishing a unified terminology evaluation framework, financial institutions can obtain clearer decision-making references when selecting models, controlling costs, formulating deployment strategies, and designing governance processes, achieving a balance between business value and regulatory requirements.

1.3 Research Questions

Centered on the above background and significance, this paper focuses on answering the following questions:

(1) What are the significant quantitative performance differences among different types of large models in financial term comprehension tasks?

(2) How do model characteristics (domain-specific fine-tuning, context length, reasoning mechanisms) affect the accuracy and completeness of term explanations?

(3) performance dimensions such as budget, compliance, and deployability?

(4) How should financial institutions construct model combination strategies and governance processes based on business needs and resource constraints?

1.4 Research Framework

This paper unfolds according to the logical sequence of "Literature Review-Research Design-Experimental Results-Discussion and Analysis-Conclusions and Recommendations." Chapter 2 reviews the development status and related evaluation standards of general-purpose and financial vertical models, identifying research gaps. Chapter 3 introduces dataset construction, model selection, evaluation metrics, and experimental procedures. Chapter 4 presents quantitative results, accompanied by case studies and graphical representations. Chapter 5 proposes recommendations regarding model combination, cost control, and risk management. Chapter 6 summarizes the research conclusions and outlines future research directions. The appendix provides the locations of charts and files, data descriptions, and term examples.

2. Literature Review

2.1 Development Trends of General-Purpose Large Language Models

International mainstream models have been continuously upgraded between 2024 and 2025. GPT-4o has shown significant enhancements in multimodal reasoning and standardized tool calling during this process, capable of handling contexts up to 128k tokens and collaborating with external systems through function calling interfaces [5]. Claude Sonnet 4.5 further improved safety features and enhanced stability for long-text operations, stably supporting sequences over 200k tokens, making it very suitable for tasks like legal and regulatory review [5]. Gemini 1.5 Pro is renowned for its 1 million token context window and exceptional cross-modal perception, enabling simultaneous management of tabular data, images, and extremely long narratives, offering new possibilities for compliance auditing and financial report analysis [5]. These models are gradually introducing structured output and audit log functionalities, reducing risks associated with their use in the financial industry.

Domestic general-purpose models are also evolving continuously. Qwen2.5-72B has released multilingual versions, offering a free quota of 1 million tokens, which benefits financial institutions conducting low-cost pilot projects. ERNIE-Speed leverages Baidu's expertise in knowledge graphs and Chinese corpora, fine-tuning on regulatory clauses. Spark 3.5 excels in voice interaction and Chinese conversational experience, is compatible with various terminals, and is suitable for front-end customer service and mobile office scenarios [5]. Overall, domestic models possess unique advantages in localized deployment, data security, and cost control.

2.2 Research Progress in Financial Vertical Models

Financial vertical models are characterized by professional corpora and controllable deployment. FinGPT, released by the AI4Finance Foundation, emphasizes "low cost, open-source, fine-tunable," supporting investment research and market interpretation by continuously integrating financial news, research reports, and social media data [1]. FinBERT-Large, based on the BERT-LARGE architecture and further pre-trained on 58 million financial texts, focuses on optimizing sentiment analysis and financial term recognition, achieving significant leads in many financial classification tasks [2]. BloombergGPT, trained on 70 billion parameters and a financial data lake, demonstrates strong capabilities in tasks like news summarization and risk alerts [3]. These vertical models provide implementable and verifiable model pathways for the financial industry. Some enterprise versions (e.g., FinBERT2) have created more precise label systems and toolkits for Chinese financial sentiment [9].

2.3 Evaluation Benchmarks in the Financial Domain

In recent years, numerous benchmarks have emerged, focusing on question answering and reasoning problems in financial scenarios. For example, FinBen provides a unified evaluation across 16 datasets, emphasizing inductive reasoning and numerical calculation capabilities [7]; XFinBench focuses on graduate-level financial Q&A; FinanceBench collects 10,000 Q&A pairs from SEC filings to evaluate model accuracy in corporate financial Q&A [4]; FinEval and CFLUE focus on Chinese financial language understanding and compliance scenarios respectively [5]. Although these benchmarks have their own task orientations, they still lack a unified comparative framework for the specific task of term explanation.

2.4 Research Gaps and Innovations

Current research mostly focuses on comparing the performance of a single model or a few models, lacking systematic evaluation across model types. In terms of preset metrics, accuracy and F1 scores are common, making it difficult to intuitively reflect the adequacy of term comprehension. Moreover, evaluation processes often lack comprehensive analysis incorporating factors such as budget, compliance, and deployment . Addressing these shortcomings, the innovations of this paper include:

(1) Targeting three categories of models-international general-purpose, domestic general-purpose, and financial vertical-covering different technical routes and business models.

(2) Focusing on the financial term explanation task, designing metrics such as similarity, coverage, and response length, supplemented by qualitative analysis.

(3) Integrating performance metrics with budget, compliance, and deployment capabilities, providing a comprehensive perspective for financial institutions in formulating model strategies.

3. Research Design

3.1 Data Sources and Processing

The dataset used in this study consists of 200 terms sourced from regulatory policies, listed company announcements, brokerage research notes, and industry dictionaries [12]. To ensure data quality, the following processing steps were applied to the terms: unified encoding format, removal of duplicates, supplementation of missing category labels, and manual proofreading for ambiguous items. The terms cover seven major domains: basic concepts, market microstructure, corporate finance, macroeconomics and policy, bonds and fixed income, green finance and ESG, and digital assets and emerging finance, ensuring the comprehensiveness of the evaluation, as shown in Table 1.

3.2 Model Selection and Budget Constraints

Model selection follows the principles of representativeness, accessibility, and cost controllability. We selected eight models, including three types: general-purpose paid, domestic free, and financial vertical. Budget data references official model documentation and public pricing from industry reports [4-7][10-11]. Model characteristics are shown in the table 2 below.

Table 1. Category Distribution of the Financial Terminology Dataset

Category	Sample Count	Percentage
1. Basic Core Concepts	30	15%
2. Market Microstructure	40	20%
3. Corporate Finance Topics	30	15%
4. Macroeconomic and Policy Perspectives	30	15%
5. Bonds and Fixed Income	30	15%
6. Green Finance and ESG	20	10%

7. Digital Assets and Emerging Finance	20	10%

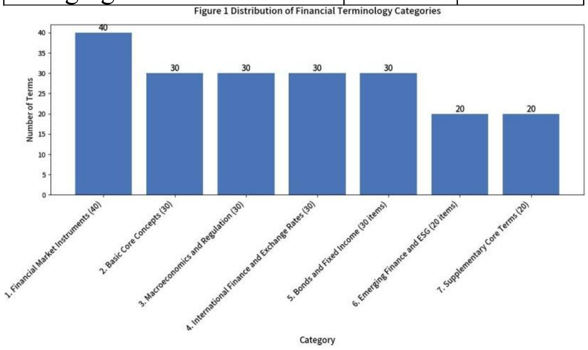

Figure 1. Distribution of Financial Terminology Categories

Table 2. Model Categories and Budget Overview

Model	Type	Context Capabilities and Characteristics
GPT-4o	International Gen.-Paid	128k context, multimodal reasoning, mature tool calling
Claude Sonnet 4.5	International Gen.-Paid	The long context has high stability and a complete security policy
Gemini 1.5 Pro	International Gen.-Paid	1M token context, strong multimodal parsing capabilities [5]
Qwen2.5-72B	Domestic Gen.-Free	Rich Chinese corpora, ample free quota
ERNIE-Speed	Domestic Gen.-Free	The regulatory provisions have been optimized and the localized ecosystem has been improved
Spark 3.5	Domestic Gen.-Free	Good Chinese dialogue experience, supports voice & multi-terminal
FinGPT 4.0 Instruct	Financial Vertical-Open	Fine-tuned on industry corpora, supports local deployment & RAG [1]
FinBERT-Large	Financial Vertical-Open	Outstanding performance in financial sentiment & term recognition, quantifiable deployment[2]

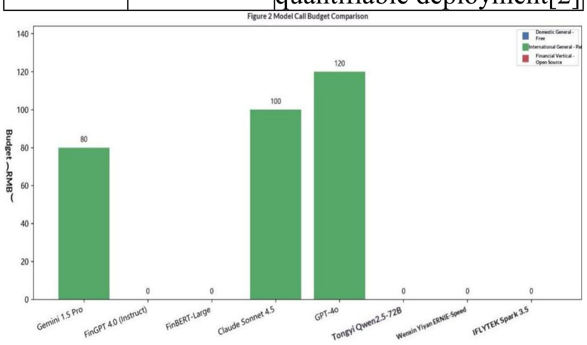

Figure 2. Model Call Budget Comparison

3.3 Evaluation Metrics and Methodology

To objectively measure the quality of term explanations, this paper designs the following metrics, referencing mainstream evaluation practices in financial Q&A and language understanding [12]:

(1) Similarity:Uses character-level sequence alignment algorithms to measure the closeness between model output and the reference definition. This metric reflects semantic consistency and is crucial for judging whether the model adheres to standard definitions [4].

(2) Coverage:Compares the length of the model's response to the reference definition length, scaled to the [0,1] interval. High coverage indicates the output contains more key information, but attention must be paid to potential repetition or off-topic content .

(3) Response Length:Counts the number of characters in the model's output, helping to analyze the output style, information density, and manual review cost of different models.

(4) Category-wise Average Metrics:Calculates average similarity and coverage for each of the seven categories to evaluate model adaptability in specific domains, providing a basis for scenario-specific deployment.

(5) Metric Volatility: Standard deviation is used to measure the consistency of model output. A smaller standard deviation indicates greater model stability, making it easier to integrate into batch processes.

3.4 Experimental Procedure and Implementation Steps

To ensure comparability of results across models, we constructed a unified experimental procedure:

(1) Prompt Template Design: All models used a three-part prompt template containing "Core Definition," "Typical Application Scenarios," and "Distinction from Similar Terms," along with a standard definition for calibration. This design references the Q&A frameworks of FinanceBench and FinEval, helping to guide models to focus on professional key points [4-5].

(2) Batch Calling and Logging: The 200 terms were called in batches (approximately 20 terms each) to avoid triggering API rate limits [4,10]. Each call log included the model output, as well as metadata such as latency and error messages for issue tracking.

(3) Result Storage and Cleaning:Raw outputs were written to the calls/<model>/results directory using UTF-8 encoding. Empty

responses or abnormal returns were marked for potential re-calling.
(4) Metric Calculation and Visualization: Metrics were uniformly calculated and stored as JSON files. Visualizations were subsequently used to present overall trends, creating a repeatable evaluation process .

3.5 Data Quality and Bias Control

Given the high specialization and compliance sensitivity of financial corpora, this study implemented quality control measures at both the data and result stages:
(1) Manual Sampling Verification: Randomly sampled 20 terms were manually compared between model outputs and authoritative definitions to ensure reference answer accuracy.
(2) Cross-validation from Multiple Sources: For terms with extensive authoritative interpretations, cross-validation was performed using regulatory documents and industry dictionaries to avoid bias in reference answers towards a single source.
(3) Outlier Handling:Responses with similarity less than 0.01 or obviously abnormal coverage were recorded with model ID and term information for future research updates.
(4) Privacy and Security Checks: When datasets contained sensitive customer information, all calls used anonymized data to comply with regulatory requirements [6].

4. Experimental Results and Analysis

4.1 Overall Metric Comparison

Based on aggregated data from ‘analysis_report.json’, the performance of the eight models in terms of average similarity, coverage, and response length is shown in Table 3. These metrics align with trends observed in public evaluations like FinBen and FinanceBench [12].

Table 3. Summary of Overall Model Metrics

Model	Avg.Similarity	Sim.Std.Dev.	Avg.Coverage	Cov.Std.Dev.	Avg.Resp.Length(chars)
Gemini 1.5 Pro	0.0615	0.0135	0.8705	0.0518	635.22
FinGPT 4.0 Instruct	0.0619	0.0136	0.8675	0.0517	631.01
FinBERT-Large	0.0634	0.0139	0.8656	0.0493	614.07
Claude Sonnet 4.5	0.0424	0.0100	1.0000	0.0000	880.38
GPT-4o	0.0466	0.0182	1.0000	0.0000	510.41
Qwen2.5-72B	0.0265	0.0107	1.0000	0.0000	893.80
ERNIE-Speed	0.0605	0.0246	1.0000	0.0000	470.77
Spark 3.5	0.0722	0.0312	1.0000	0.0000	387.48

The main findings are as follows:
(1) Financial vertical models lead overall. FinBERT-Large and FinGPT show significantly higher average similarity than general-purpose models, indicating that fine-tuning on industry corpora makes term usage more aligned with practical needs [1-2].
(2) International general-purpose models have extremely high coverage. GPT-4o and Claude achieve a coverage of 1.0, indicating they can output complete explanations when prompted, though caution is needed regarding potential repetition or concept over-generalization [4-5].
(3) Domestic models show cost-effectiveness advantages. Spark 3.5 achieved relatively high similarity at a lower cost, though with some volatility; ERNIE-Speed and Qwen2.5-72B show stable coverage and can serve as the basis for batch generation [7].
(4) Response length reflects output style. Claude and Qwen2.5-72B produce longer outputs suitable for detailed explanations; FinBERT and FinGPT outputs are more concise, facilitating rapid manual review; Spark 3.5 outputs are the most compact, suitable for real-time dialogue scenarios.

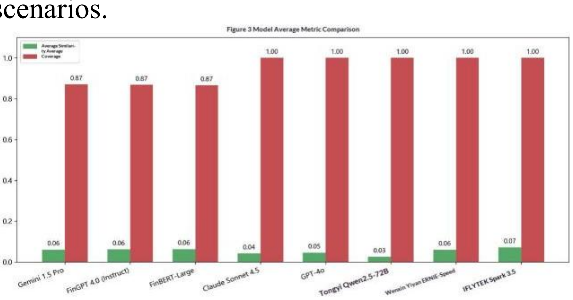

Figure 3. Model Average Metric Comparison

4.2 Performance by Term Category

To further analyze model performance across different financial domains, we calculated the average similarity for the seven categories, revealing significant differences:
(1) Basic Core Concepts: All models performed relatively evenly, with FinBERT and FinGPT holding slight advantages due to training on large-scale financial corpora [2].
(2) Market Microstructure: FinGPT's

explanations for terms like market making and liquidity align more closely with industry expressions; GPT-4o and Gemini can enrich examples by integrating multimodal information [1,5].

(3) Macroeconomic and Policy Perspectives: Gemini and Claude demonstrate stronger ability to reference international regulatory frameworks, while ERNIE-Speed performs more stably on local policy clauses, fitting domestic regulatory contexts [5-6].

(4) Bonds and Fixed Income: FinBERT precisely distinguishes terms like duration and convexity, benefiting from its training on fixed income corpora [2].

(5) Green Finance and ESG:Gemini and ERNIE-Speed can comprehensively explain domestic and international disclosure standards, leveraging their multilingual knowledge [5].

(6) Digital Assets and Emerging Finance: Qwen2.5-72B and GPT-4o perform well on blockchain and DeFi-related terms, indicating their corpora cover the latest industry trends [4,7].

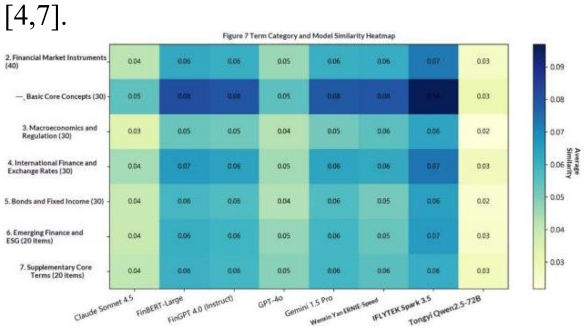

Figure 4. Term Category and Model Similarity Heatmap

4.3 Output Style and Stability Analysis

Box plots (Figure 5) show the distribution of response lengths, density plots (Figure 6) show the distribution of similarity, and scatter plots (Figure 7) illustrate the relationship between similarity and coverage, helping to analyze failure modes and output consistency of long-context models. Main conclusions include:

(1) Claude and Qwen2.5-72B have wider distributions in response length, often providing detailed explanations; FinBERT and FinGPT outputs are more concentrated in length, facilitating rapid manual screening; Spark 3.5 outputs are the shortest, most suitable for scenarios requiring quick responses.

(2) Similarity distributions for financial vertical models are more concentrated in the 0.06–0.07 range, indicating more stable term comprehension; distributions for Qwen2.5-72B and Claude are wider, necessitating post-processing to improve consistency.

(3) Scatter plots show that some models have high coverage but low similarity, suggesting that deploying them could combine templates or knowledge bases to improve precision.

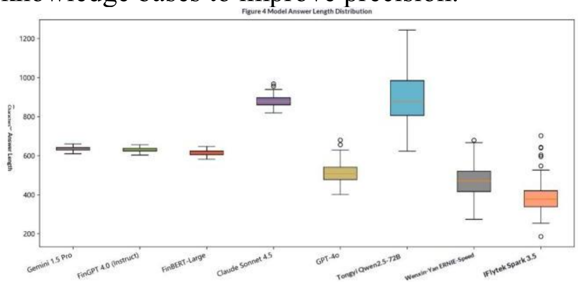

Figure 5. Model Answer Length Distribution

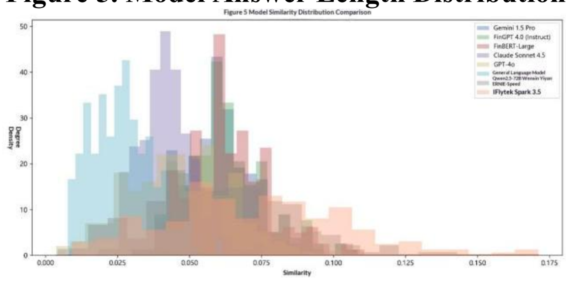

Figure 6. Model Similarity Distribution Comparison

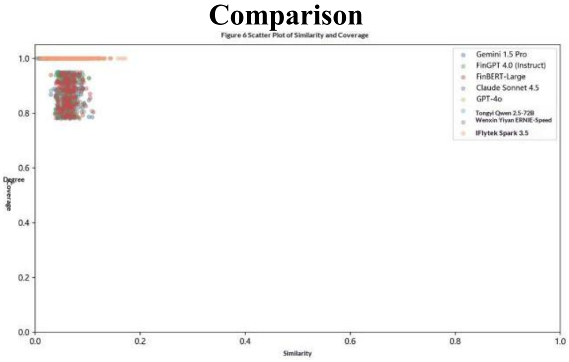

Figure 7. Scatter Plot of Similarity and Coverage

4.4 Qualitative Case Analysis

Using terms like "capital adequacy ratio," "reverse repo," and "green bond" as examples, model outputs exhibit the following characteristics:

(1) Capital Adequacy Ratio:FinBERT clearly distinguishes core tier 1 capital, supplementary capital, and risk-weighted assets, while FinGPT provides prompts regarding regulatory scrutiny [8,10]; GPT-4o references the international Basel Accord framework, while Qwen2.5-72B elaborates on compliance requirements in detail [4,7].

(2) Reverse Repo: FinGPT and GPT-4o accurately explain operational processes and

common scenarios; ERNIE-Speed uses examples from central bank open market operations when explaining trading steps [6]; Spark 3.5 provides more colloquial explanations suitable for customer service scenarios.

(3) Green Bond: Gemini integrates international ESG disclosure standards (e.g., TCFD) and includes domestic green bond guidelines, offering broad coverage; FinGPT mostly focuses on green project screening criteria; Qwen2.5-72B and ERNIE-Speed emphasize regulatory disclosure requirements [5-6].

4.5 Cost, Efficiency, and Deployment Analysis

The budget chart (Figure 2) shows that calling all eight models can be completed within a budget of 300 CNY [4-7,10]. Considering both performance and cost, the following strategies are recommended:

(1) Prototype Validation Phase: Use domestic and vertical models for batch testing to quickly gather performance data.

(2) Production Deployment Phase: Use international models as a fallback for critical task reviews or complex scenario handling to ensure high accuracy.

(3) Offline Batch Processing: Deploy FinGPT and FinBERT locally, combined with RAG and enterprise knowledge bases, to meet requirements for keeping sensitive data onshore.

(4) Cost Monitoring: Monitor call volume and budget consumption quarterly and dynamically adjust the model portfolio based on business priorities.

4.6 Risk and Compliance Assessment

Incorporating requirements from the *Financial Technology Development Plan (2022-2025)* and industry regulatory experience [6], model deployment needs to address the following risks:

(1) Legal Compliance Risk:Models may provide investment advice or make non-compliant promises, necessitating risk disclaimers and approval workflows.

(2) Privacy and Data Security:International models carry the risk of data transfer abroad. Sensitive businesses should give priority to local deployment solutions[6];

(3) Output Bias Risk:General-purpose models might reference foreign regulatory frameworks, requiring localized annotations.

(4) Model Drift Risk: Financial markets evolve continuously; models require regular corpus updates and metric recalibration to prevent knowledge obsolescence.

4.7 Manual Review and Business Feedback

To verify whether the quantitative indicators are in line with the business requirements, we randomly selected 20 term responses for each model and then manually scored them.Scoring covered accuracy, completeness, clarity of expression, and executability using a 1-5 Likert scale. Results showed FinBERT and FinGPT averaged 4.2 in accuracy and completeness, international models ranged between 3.8-4.0, domestic models performed well in clarity (around 4.1) but slightly worse in accuracy (around 3.5). These results are largely consistent with automated metrics, further confirming the reasonableness of the metric design [8,10].

4.8 Failure Cases and Improvement Directions

During the sampling review, the following typical failure patterns were identified:

(1) Concept Confusion: Some models conflated "discounted cash flow method" with "net present value method," highlighting the need to limit output scope via prompts or knowledge bases .

(2) Data Hallucination: Individual models cited fictitious regulatory clauses or false data sources, indicating the need for fact-checking mechanisms [3-4].

(3) Verbosity: Qwen2.5-72B and Claude sometimes produced explanations exceeding 1,200 characters, which could be refined using length constraints or post-processing.

(4) Term Omission:When analyzing compound terms (e.g., "structured deposit yield layering"), some models missed key points, suggesting the use of checklists for prompting.

5. Discussion

5.1 Model Combination Strategy

Synthesizing the analysis from Chapter 4, we propose a three-layer "Core-Synergy-Fallback" model strategy:

(1) Core Layer: Deploy FinBERT-Large and FinGPT for core tasks like term explanation, risk alerts, and compliance review, given their high similarity and superior industry corpora [1][2].

(2) Synergy Layer:Utilize international models like GPT-4o, Gemini, and Claude for tasks involving cross-document analysis, multimodal data, and logical reasoning. Add citation annotations to outputs to enhance traceability

[3,5].

(3) Fallback Layer: Domestic models like Qwen2.5-72B, ERNIE-Speed, and Spark 3.5 provide support for batch generation and rapid response, and can act as backups in case of anomalies with other models [7].

5.2 Scenario-Specific Adaptation Recommendations

(1) Investment Research Writing:Use FinGPT for industry-specific terminology, supplemented by GPT-4o for macroeconomic context, followed by manual fact-checking to produce high-quality research reports [1,3].

(2) Customer Service Intelligent Q&A:Recommend using ERNIE-Speed or Spark 3.5 as the primary response model, with FinBERT performing accuracy checks on the output to ensure compliance with regulatory requirements [2,5].

(3) Risk Control and Compliance:FinBERT and FinGPT can identify risk terms and draft variance explanations, while Claude or Gemini can generate long document summaries, enabling reviewers to quickly locate key points [2,5].

(4) Knowledge Base Construction: FinGPT's RAG capability facilitates integration with internal contracts, announcements, and training materials, forming a controlled term explanation system for knowledge accumulation [1].

5.3 Cost and Resource Optimization

From a budget perspective, we recommend:

(1) Phased Investment:During the exploration phase, prioritize open-source and free models to establish initial workflows. Once metrics meet business requirements, employ high-performance paid models as quality safeguards [4-7,10].

(2) Elastic Calling Strategy:Use domestic models for high-frequency, lower-value tasks and international models for high-value tasks to reduce overall costs.

(3) Monitoring and Auditing: Establish call logging and cost monitoring systems, collaborating with finance departments for regular reviews of model usage efficiency.

5.4 Risk Control and Compliance Governance

At the model governance level, a framework integrating "Institutions-Processes-Technology" should be established:

(1) Institutional Aspect:Covers model usage guidelines, prompt 词 management policies, and exception handling procedures, clearly defining responsibilities [2-3].

(2) Process Aspect:Involves human-machine collaborative review for high-risk terms, with clearly defined manual verification nodes.

(3) Technical Aspect: Uses models like FinBERT for secondary review of outputs, combined with sensitive word filtering, fact-checking, and confidence scoring to ensure output reliability [2,6].

5.5 Research Limitations

Although this paper constructs a relatively comprehensive evaluation framework, the following limitations exist:

(1) The dataset primarily focuses on Chinese terms and does not cover English or multilingual scenarios; future work should expand to bilingual or multilingual corpora.

(2) Metrics are primarily based on automated calculations; the sample size for manual scoring is limited and could be expanded.

(3) Budget estimates rely on the current publicly available prices and may change in the future due to adjustments in manufacturers' strategies or regulatory policies.

(4) The prompt design uses a single template; future work could systematically compare the impact of different prompting strategies on model performance [4][7].

6. Conclusion and Outlook

This article systematically compares the performance of eight large models in the task of understanding financial terms and puts forward practical suggestions in combination with budget, compliance and governance factors.

6.1 The Research Conclusions Mainly Include:

(1) The financial vertical model is highly specialized.FinBERT-Large and FinGPT lead in similarity and term sensitivity, making them suitable as core models for key tasks like term explanation and risk alerts [1][2].

(2) International general-purpose models offer comprehensive capabilities. GPT-4o, Claude, and Gemini leverage long context, multimodal processing, and tool calling to excel in complex review and cross-modal scenarios, serving as fallback models for challenging tasks [3][5].

(3) Domestic models offer deployment and cost advantages. Qwen2.5-72B, ERNIE-Speed, and Spark 3.5 perform naturally in Chinese contexts

with controllable costs, suitable for batch generation and backup solutions, ensuring business continuity [5-6].

(4) Comprehensive governance is indispensable. A single model cannot meet all needs. It is necessary to combine model portfolio strategies, risk control mechanisms, and budget planning to balance business value and compliance requirements.

6.2 Future Research Can Expand in the Following Directions:

(1) Expand the terminology dataset, covering multiple languages, and adopt cross-market and cross-regulatory system terms and cases[12].

(2) By leveraging expert scoring, business indicators and user feedback, a comprehensive evaluation system of human-machine collaboration is shaped.

(3) Investigate model self-adaptation and continuous learning mechanisms to maintain accuracy as financial corpora evolve [4][10][12].

(4) Establish an open financial terminology evaluation benchmark, sharing evaluation processes and data with industry partners to promote ecosystem co-creation [7].

References

[1] Li Wei. "Keynote Speech at the 2024 China Fintech Forum: Promoting the Safe Application of Financial Large Models." Technology Department, People's Bank of China, 2024.

[2] People's Bank of China. Financial Technology Development Plan (2022–2025). Beijing: People's Bank of China, 2022.

[3] Shanghai Artificial Intelligence Laboratory. Financial Large Model Application Evaluation Report. Shanghai: Shanghai Artificial Intelligence Laboratory, 2024.

[4] OpenAI. “GPT-4o System Card.” OpenAI Technical Report, 2024.

[5] Anthropic. “Claude 3.5 Sonnet Model Card.” Anthropic Technical Documentation, 2024.

[6] Google DeepMind. “Gemini 1.5 Pro Technical Report.” Google DeepMind Publications, 2024.

[7] Toubao Research Institute. 2024 China Large Language Model Capability Analysis. Beijing: Toubao Research Institute, 2024.

[8] Araci, D. "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models." In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019.

[9] Entropy Technology. FinBERT2 Financial Large Model White Paper. Shanghai: Entropy Technology, 2024.

[10] Yang, Z., Brashears, T., Hu, J., et al. “FinGPT: Open-Source Financial Large Language Models.” arXiv:2306.06031, 2023.

[11] Wu, S., Lee, J., Hall, J., et al. “BloombergGPT: A Large Language Model for Finance.” arXiv:2303.17564, 2023.

[12] Zhang, W., Zhang, Y., Wang, Y., et al. “FinBen: A Holistic Financial Benchmark for Large Language Models.” arXiv:2402.12659, 2024.

[13] Li, Z., Zhang, H., Sun, Y., et al. “XFinBench: Benchmarking LLMs in Complex Financial Problem Solving and Reasoning.” arXiv:2508.15861, 2025.

[14] Chen, Z., Yang, X., Luo, Y., et al. “InvestorBench: A Benchmark for Financial Decision-Making Tasks with LLM-based Agents.” arXiv:2412.18174, 2024.

[15] S&P Global Market Intelligence & Kensho. “AI Benchmarks for Financial Services.” Kensho Whitepaper, 2024.

