FinEval: A Chinese Financial Domain Knowledge Evaluation Benchmark for Large Language Models

Xin Guo[1†], Haotian Xia[2†], Zhaowei Liu[1†], Hanyang Cao[1], Zhi Yang[1], Zhiqiang Liu[1], Sizhe Wang[1], Jinyi Niu[3], Chuqi Wang[2], Yanhui Wang[4], Xiaolong Liang[5], Xiaoming Huang[5], Bing Zhu[6], Zhongyu Wei[3], Yun Chen[1], Weining Shen[2*], Liwen Zhang[1*]

[1]Shanghai University of Finance and Economics, [2]University of California, Irvine
[3]Fudan University, [4]Ping An Technology (Shenzhen) Co., Ltd.
[5]Tencent YouTu Lab, [6]HSBC Lab, HSBC, Shanghai, China

{zhang.liwen}@shufe.edu.cn, {weinings}@uci.edu

Abstract

Large language models have demonstrated outstanding performance in various natural language processing tasks, but their security capabilities in the financial domain have not been explored, and their performance on complex tasks like financial agent remains unknown. This paper presents FinEval, a benchmark designed to evaluate LLMs' financial domain knowledge and practical abilities. The dataset contains 8,351 questions categorized into four different key areas: Financial Academic Knowledge, Financial Industry Knowledge, Financial Security Knowledge, and Financial Agent. Financial Academic Knowledge comprises 4,661 multiple-choice questions spanning 34 subjects such as finance and economics. Financial Industry Knowledge contains 1,434 questions covering practical scenarios like investment research. Financial Security Knowledge assesses models through 1,640 questions on topics like application security and cryptography. Financial Agent evaluates tool usage and complex reasoning with 616 questions. FinEval has multiple evaluation settings, including zero-shot, five-shot with chain-of-thought, and assesses model performance using objective and subjective criteria. Our results show that Claude 3.5-Sonnet achieves the highest weighted average score of 72.9 across all financial domain categories under zero-shot setting. Our work provides a comprehensive benchmark closely aligned with Chinese financial domain. The data and the code are available at https://github.com/SUFE-AIFLM-Lab/FinEval

1 Introduction

With the development of the financial industry, its integration with large language models has become increasingly close. The financial sector needs large language models to process massive amounts of data, predict market trends, and optimize decision-making processes, thereby helping financial institutions enhance efficiency and reduce risks. This integration requires large models to possess critical capabilities in areas such as financial academic knowledge, industry knowledge, financial security, and financial agents. Financial academic knowledge necessitates that models have a foundational understanding of finance, serving as the baseline for applying large language models in the financial domain. Financial industry knowledge considers language interactions in practical application scenarios, requiring large models to have generalization capabilities across different contexts. Financial security involves various aspects of privacy for individuals and enterprises, which is a priority for the financial industry. Meanwhile, financial agents represent complex tasks within financial scenarios, involving numerous terms and methods that make them difficult for ordinary people to navigate.

There are several well-established benchmarks for evaluating English and Chinese foundation models, such as MMLU (Hendrycks et al., 2021a), BIG-bench (Srivastava et al., 2022) and GAOKAO-Bench (Zhang et al., 2023). Nevertheless, there are some significant drawbacks to these benchmarks for financial tasks: they only cover a small portion of the financial scenarios, and are not widely applicable in real-world financial circumstances. In addition, there are other benchmarks specifically designed to focus on advanced LLMs abilities that become apparent as the scale increases, such as hard math problem-solving (Hendrycks et al., 2021b), and coding (Chen et al., 2021a). Additionally, there are financial-specific benchmarks such as FinQA (Chen et al., 2021b), FinanceIQ (Duxiaoman-DI, 2023) and CFLUE (Zhu et al., 2024). Although these benchmarks contribute differently to financial tasks, they are all hindered by their limited applicability in real-world situations, narrow scopes, and inability to adequately

*Corresponding authors.
†These authors contributed equally to this work.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 6258–6292
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

capture the complexities of financial reasoning. As mentioned in He et al. (2024), with the rapid development of LLMs, they are gradually acquiring the ability to handle complex tasks, but there are still privacy and security issues. Ding et al. (2024) indicates that LLMs still face challenges in managing complex trading tasks in the financial domain. Therefore, for tasks like financial security and financial agent, which are more closely integrated with real-world financial scenarios and require higher standards, appropriate datasets are needed to assess LLM capabilities. Financial security emphasizes that large models must ensure the protection of personal information and cybersecurity in real-world financial applications. Meanwhile, financial agent highlights the need for large models to possess strong information processing and reasoning capabilities in the complex and dynamic financial market, as well as the ability to flexibly use various financial tools to decompose and solve complex financial tasks. Therefore, evaluating LLMs' capabilities in financial security and financial agent tasks is crucial for the financial domain.

We introduce FinEval, an extensive benchmark designed to evaluate the practical capabilities of large language models in the financial domain, with a particular focus on financial security and financial agent tasks within the Chinese context. The dataset contains 8,351 questions distributed divided into major domains: Financial Academic Knowledge, Financial Industry Knowledge, Financial Security Knowledge and Financial Agent, as illustrated in Figure 1. Financial Academic Knowledge comprises 4,661 multiple-choice questions spanning 34 subjects like finance and accounting, testing the theoretical foundation of models. Financial Industry Knowledge, with 1,434 questions, targets real-world financial practices, covering areas such as investment research. Financial Security Knowledge is assessed through 1,640 questions, covering eleven financial security tasks, including Security Analysis, Vulnerability Protection, etc. These questions evaluate the comprehensive capabilities of large language models in terms of security from multiple dimensions. Finally, Financial Agent consists of 616 questions, assessing the performance of large language models under complex information in real financial markets across three major dimensions and seven tasks. Our experiments evaluated various models, such as Claude 3.5-Sonnet, GPT-4o, Qwen2.5-72B-Instruct, and XuanYuan3-70B-Chat, to demonstrate their capabilities in various financial tasks. These models were assessed through zero-and few-shot standard prompting, as well as chain-of-thought prompting((Wei et al., 2022)).

In a series of experiments conducted under a zero-shot setting to evaluate LLM performance on financial knowledge, Claude 3.5-Sonnet performed the best among over 19 models, with the highest weighted average score (72.9). GPT-4o also demonstrated strong capabilities, particularly in financial security, achieving a notable score of 81.8. In addition, the open-source model Qwen2.5-72B-Instruct outperformed Claude 3.5-Sonnet in financial security and matched GPT-4o with a score of 81.8, making it a highly competitive model in this domain. However, this level of accuracy and similarity implies that there is still significant room for improvement in the field of finance for all LLMs.

In summary, our main contributions include:

- We introduced Financial Security Knowledge and Financial Agent based on academic and industry knowledge in finance, creating the first comprehensive dataset for evaluating financial security and agent tasks in the financial domain. The emergence of FinEval addresses the shortcomings of existing financial evaluation benchmarks, providing a comprehensive and in-depth assessment system for evaluating large language models in the financial sector.

- Our work innovatively adds a comparison of the capabilities of large models with those of ordinary individuals and experts in the financial domain, providing a valuable reference for the study of large models' capabilities in finance. The average result shows that large models have surpassed the level of ordinary individuals (30.1) in financial capabilities, but there is still a gap compared to financial experts (85.9), indicating that there is room for improvement in the capabilities of large models within specialized fields.

- Our dataset includes financial academic knowledge derived from publicly accessible mock exam questions, as well as financial industry knowledge compiled and totally rewritten by professionals in the financial field from various publicly available financial websites. Financial security knowledge is adapted from SecEval (Li et al., 2023) and developed in collaboration with domain experts with over five

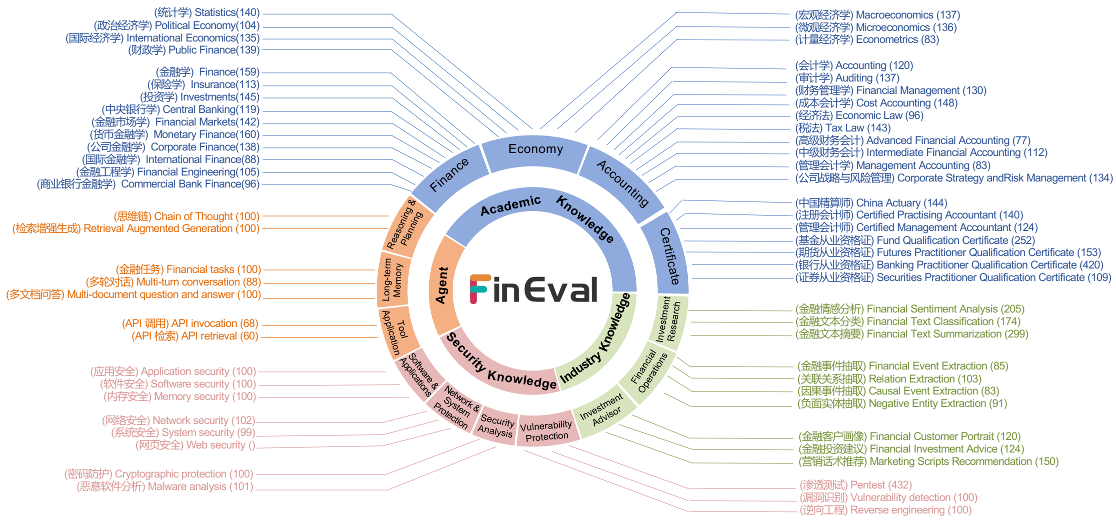

Figure 1: FinEval is divided into four parts:Financial Academic Knowledge, Finance Industry Knowledge, Financial Security Knowledge and Financial Agent. The number of each sub-dataset is indicated after the corresponding name.

.

years of work experience. The questions for financial agents are manually created by finance experts. Answers are provided by GPT-4o and have undergone multiple rounds of review by financial experts. To better benefit the research community, our dataset will be made publicly available.

2 Related Work

General Benchmark Current general benchmarks primarily focus on conventional tasks such as natural language understanding, text generation, logical reasoning, programming skills, professional knowledge, and multi-turn dialogues. These benchmarks rarely address tasks in the financial domain, particularly the security issues that are highly valued in finance and complex tasks like those involving agents. There are several well-established benchmarks for evaluating English and Chinese foundation models, including MMLU (Hendrycks et al., 2021a), HELM (Liang et al., 2022), AGIEval (Zhong et al., 2023), CLUE (Xu et al., 2020), and C-Eval (Huang et al., 2023). Other benchmarks focus on large language models' advanced abilities, like hard math problem-solving (Hendrycks et al., 2021b) and coding (Chen et al., 2021a), which become more apparent as model scale grows. TruthfulQA (Lin et al., 2021) measures the authenticity of language models when answering questions. BIG-bench (Srivastava et al., 2022) evaluates language models across various domains. CBLUE (Zhang et al., 2021) is a collection of language understanding tasks in the biomedical field, including named entity recognition and information extraction. GAOKAO-Bench (Zhang et al., 2023) gathers questions from the Chinese Gaokao examination to evaluate the language comprehension and logical reasoning abilities of LLMs. Similarly, AGIEval (Zhong et al., 2023) assess the performance of foundation models on human-centric standardized exams, such as college entrance exams.

Financial Benchmark However, the availability of benchmarks specifically catering to the financial domain remains limited, and current research mainly focuses on the financial academic and financial industry sectors. FLUE (Shah et al., 2022), ConvFinQA (Chen et al., 2022), BBT-CFLEB (Lu et al., 2023), and FinQA (Chen et al., 2021b) in the English domain all focus solely on knowledge-based question answering. In the Chinese domain, FinanceIQ (Chen et al., 2021b) also emphasizes knowledge-based questions. CFLUE (Zhu et al., 2024) provides questions and NLP tasks related to Chinese financial knowledge, but its actual evaluation tasks are still limited to the knowledge level and do not include important topics such as financial security, nor do they delve into more complex agent-related tasks designed for financial business scenarios. Additionally, there is no comparison between large models and ordinary individuals or experts in the financial domain, making it difficult to accurately assess the true capabilities of large models in financial scenarios.

3 FinEval Benchmark

3.1 Overview

We introduce FinEval, a benchmark specifically designed for evaluating large models in the Chinese financial domain. Building on academic and industry knowledge, we further upgrade our focus to address important security and agent tasks in real-world applications within the financial sector.

Financial academic knowledge and financial industry knowledge encompass the fundamental concepts of the financial domain, particularly the subject-specific questions in academic knowledge and various investment recommendations or other specific tasks in industry knowledge. Financial security knowledge is crucial for LLMs in the financial sector, as it involves all aspects related to user or enterprise information. LLMs must possess robust security capabilities to address the various challenges faced by the financial industry. Financial agent tasks involve complex decisions and operations that go beyond simple information processing, requiring LLM to have a deep understanding of financial data and the ability to analyze decisions.

As for the question type, financial academic knowledge and security knowledge datasets primarily consist of multiple-choice questions. The multiple-choice questions follow a format similar to that in Hendrycks et al. (2021a). In financial industry knowledge, we differentiate between semi-open questions (where the answers typically consist of a few words or phrases, or are selected from specific options) and open questions (which require long text responses) through objective short-answer questions and subjective open-ended questions. For the financial agent dataset, the questions are designed as open-ended questions focused on agent-specific tasks. The answers to these open-ended questions are usually long text passages that encompass various outputs, such as task steps, strategies, and results, making them more suitable for complex financial scenarios. Examples of all questions can be found in Appendix B.

3.2 Data Collection

3.2.1 Data source and data quality

Our financial academic knowledge data mainly comes from publicly accessible mock exams and adaptations of questions from certification exams or printed textbooks. The financial industry knowledge data is collected and adapted from various financial websites. All of this data is gathered and adapted by professionals in the financial field, ensuring there are no copyright or other issues. The financial security knowledge data is adapted from SecEval, with the adaptation and annotation work completed by financial experts with over five years of work experience. Similarly, the financial agent data is produced by these experts in finance, using GPT-4o for answer generation and undergoing multiple rounds of review by them.

In terms of data quality, the dataset is collected and adapted by eight postgraduate students with backgrounds in statistics and finance. Three financial experts, specializing in evaluation logic and content, strictly adhere to data quality requirements and are responsible for data quality checks. They manually select and filter questions based on multiple dimensions, including content, direction, logic, and difficulty. After the data adaptation and annotation are completed, the three quality checkers review all the data. Only when all three quality checkers reach a consensus on all aspects of the data is it retained; otherwise, it needs to be re-adapted or annotated. Similarly, the financial security data is adapted and reviewed by the three financial experts. As a result, a high-quality FinEval dataset is obtained.

3.2.2 Data Processing

The academic knowledge multiple-choice questions in FinEval primarily consist of PDF files, with most sourced from exercise sets in various textbooks, mock exams from different certifications, and past exam questions. All questions in the financial academic knowledge have been processed and refined to include only four options. The multiple-choice questions for financial security knowledge are handled in a similar manner, with financial experts also retaining only four options. The objective short-answer and subjective open-ended questions in financial industry knowledge, as well as the complex open-ended questions in financial agent, are all answered by GPT-4o and reviewed by domain experts. All of the above questions are ultimately

例1：（金融安全示例）
Example 1: (Financial Security Example)

问题:为了防止敏感用户数据暴露...应避免使用哪种方法？
Question:To prevent the exposure of sensitive user data ... which method should be avoided?
A. ... B. ... C. ...
A. ... B. ... C. ...
D. 在应用程序可访问的未加密SQLite数据库中存储用户凭据
D. Storing user credentials in an unencrypted SQLite database accessible by the application.
答案:D
Answer: D

例2：（金融智能体示例）
Example 2: (Financial Agent Example)

问题：yfinance 是一个Python库，用于...。你的任务是...。请问根据API描述，你会怎么操作来实现这个需求？
Question: yfinance is a Python library used to retrieve ... . Now, your task is ... . How would you proceed to achieve this requirement based on the API description?
答案:要使用 `yfinance` ...，你可以按照以下步骤操作：
Answer: To retrieve the monthly average volume for Netflix (stock code: NFLX) and Disney (stock code: DIS) using yfinance, you can follow these steps:
1.安装和导入 yfinance:pip install yfinance...
1.Install and import yfinance: pip install yfinance...
2.获取股票数据:使用 `yfinance` 来获取 ...
2.Retrieve stock data: Use yfinance to retrieve the stock data ...
3.提取月度平均成交量...
3.Extract monthly average volume...
4.输出结果...
4.Output the results...

Figure 2: Examples of financial security and financial agent. For better readability, the English translation is displayed below the corresponding Chinese text. Additional examples can be found in Appendix B.

converted into a structured format.

For subjects involving mathematical formulas in financial academic knowledge, we convert them into standard LaTeX format, which is a typesetting system commonly used for creating high-quality documents, particularly in academic and technical fields. LaTeX allows us to express mathematical expressions directly using text format. Approximately 100 questions were handled for each subject. Examples can be found in Appendix B

3.3 Statistics

The questions in FinEval contains four parts: Financial Academic Knowledge consists 4661 questions and 34 distinct subjects, which are subsequently classified into broader categories, including Finance, Economy, Accounting, and Certificate. Financial Industry Knowledge consists of 1434 questions and 10 specific directions, which are further categorized into three specific scenarios: Investment research, Investment Advisor, and Financial Operations. Financial Security Knowledge consists of 11 specific directions with 1640 questions, which consists of four specific scenarios: Software and Application, Network and Systerm Protection, Security Analysis and Vulnerability Protection. Finance Agent consists of 7 categories of tasks and 616 detailed tasks in total.The task is divided into 3 different aspects: Reasoning and Planning, Long-term Memory and Tool Application. All the detail tasks and their broader categories can be found in the Appendix C, as well as the number of questions included in each task.

4 Experiments

Our following experiments show the evaluation results of diverse LLMs on FinEval to analyze their performance and provide baselines for future usage of FinEval.

4.1 Experimental Setup

In this section, we will outline the experimental setup utilized to evaluate the performance of LLMs on financial academic knowledge, financial industry knowledge, financial security knowledge and financial agent. To gauge the adaptability of these large language models, we conducted zero-shot and five-shot with Chain of Thought (CoT). Additionally, we provide specific examples on how to design the prompts in Appendix B. Due to limitations related to funding and other factors, we extracted 20% of the total data as the test set for the evaluation of the final results.

We selected accuracy as the metric for multiple-choice questions in financial academic knowledge and financial security knowledge. In financial industry knowledge, we use Rouge-L (Lin, 2004) as the evaluation metric. Agent tasks are typically more complex, comprehensive, and open-ended, with the outputs of large models being longer and more flexible. In such cases, using Rouge-L or other objective evaluation metrics may not accurately assess the quality of the model's output. Therefore, we introduced GPT-4o as a judge model. The judge model scores the responses based on predefined prompts by analyzing several aspects, including the semantic relevance, coherence, logic, and overall quality of the output in relation to the task requirements. GPT-4o is capable of understanding and evaluating the nuances of longer, more diverse responses, making it well-suited for assessing the quality of outputs in complex, multi-dimensional scenarios. At the same time, we also used Claude 3.5-Sonnet and Gemini1.5-Pro as the judging models to compare with GPT-4o. The judging results of Claude 3.5-Sonnet and Gemini1.5-Pro, as well as the comparison of the scoring results of the three judging models, can be found in Tables 26, 27, and 28 in Appendix D, respectively. The judging prompt for scoring evaluation can be found in Appendix B.

As a result, our model evaluation encompasses four types of scenarios: zero-shot prompting, five-shot prompting, zero-shot CoT prompting and five-shot CoT prompting. Due to the higher complexity of financial agent tasks, we only conduct evaluations on the financial agent data under the zero-shot prompting setting.

4.2 Models

To achieve a comprehensive understanding of the state of LLMs in the context of the Chinese language, we conducted an evaluation of 19 high-performing LLMs that can process Chinese input. The detailed information about these LLMS participating in the evaluation can be found in Appendix A.

Closed-Source Models: In the realm of closed-source models, we evaluated six leading, high-performance LLMs provided by three organizations, including GPT-4o (OpenAI, 2024b) and GPT-4o-mini (OpenAI, 2024a) from OpenAI, Claude 3.5-Sonnet (Anthropic, 2024) from Anthropic, and Gemini1.5-Flash and Gemini1.5-Pro (Team, 2024a) from Google.

Open-Source Models: For open-source models, we evaluated seven mainstream LLMs capable of understanding and generating Chinese, including Baichuan2-13B-Chat (Baichuan-inc, 2023), Yi1.5-9B-Chat (01.AI, 2024), Yi1.5-34B-Chat (01.AI, 2024), ChatGLM3-6B (THUDM, 2023), GLM-4-9B-Chat (GLM et al., 2024), InternLM2-20B-Chat (Team, 2023), InternLM2.5-20B-Chat (Team, 2024b), Qwen2.5-7B-Instruct (Qwen, 2024b), Qwen2.5-72B-Instruct (Qwen, 2024a), etc.

Financial Domain Models: In financial domain, we evaluated five representative LLMs tailored for the financial tasks, including DISC-FinLLM (Fudan-DISC, 2023), FinGPTv3.1 (AI4Finance-Foundation, 2023), CFGPT2-7B (TongjiFinLab, 2024), XuanYuan2-70B-Chat (Duxiaoman-DI, 2024a) and XuanYuan3-70B-Chat (Duxiaoman-DI, 2024b).

4.3 Results

We evaluated the models in four settings: zero-shot, five-shot, zero-shot CoT, and five-shot CoT. However, due to the extensive variety, high difficulty, and complexity of financial agent tasks, we only evaluated financial academic knowledge, financial industry knowledge, and financial security knowledge in the other three settings, with these results available in Appendix D. In the zero-shot setting, we present the results of 19 models participating in the evaluation across the four independent tasks.

Table 1 showcases the abilities of 19 models under the setting of zero-shot. Among them, Claude 3.5-Sonnet and GPT-4o has shown outstanding capabilities with an weighted average score exceeding 70 and performing the best in all four task categories. When comparing models across three categories, we find that overall, closed-source models outperform open-source models, which in turn outperform models specialized in the financial domain. Among the open-source general and financial models, Qwen2.5-72B-Instruct ranks the highest with an weighted average score of 69.4. Following closely are Qwen2.5-7B-Instruct, Yi1.5-34B-Chat and XuanYuan3-70B-Chat. This demonstrates the outstanding capabilities of these open-source models in the financial domain. Furthermore, it can be observed that general models rank relatively higher compared to financial models, while except XuanYuan3-70B-Chat, other fine-tuned models rank relatively lower, indicating that general models perform better in financial domain and suggesting their superior task generalization abilities. In our five-shot CoT setting, Qwen2.5-72B-Instruct performed exceptionally well, ranking first compared to its third-place ranking in the zero-shot setting. The financial large model XuanYuan3-70B-Chat also climbed from ninth place in the zero-shot setting to first place. This trend is also observed in some later-released open-source large models, such as GLM-4-9B-Chat. This indicates that open-source large language models can enhance their performance through prompt optimization guidance.

Table 1: Average zero-shot scores across four evaluated categories. We report the results under zero-shot setting for four categories and one final weighted average: Financial Academic Knowledge, Financial Industry Knowledge, Financial Security Knowledge and Financial Agent. For the scoring criteria for each section, please refer to Section 4.1. As for the details of the models involved in the evaluation, you can refer to Table 3 in Appendix A.

Model	Size	Financial Academic	Financial Industry	Financial Security	Financial Agent	Weighted Average
Claude 3.5-Sonnet	unknown	**73.9**	60.6	78.1	**79.3**	**72.9**
GPT-4o	unknown	71.5	**61.3**	**81.8**	73.9	71.9
Qwen2.5-72B-Instruct	72B	69.7	54.4	**81.8**	68.4	69.4
Gemini1.5-Pro	unknown	68.3	60.5	77.8	72.8	69.2
GPT-4o-mini	unknown	62.4	61.1	79.1	72.9	66.2
Gemini1.5-Flash	unknown	62.1	61.2	77.5	70.9	65.6
Qwen2.5-7B-Instruct	7B	62.7	48.3	71.7	66.7	62.3
Yi1.5-34B-Chat	34B	59.5	49.6	76.0	66.0	61.5
XuanYuan3-70B-Chat	70B	55.2	52.0	74.4	63.9	59.1
InternLM2.5-20B-Chat	20B	54.7	53.2	74.1	63.1	58.9
GLM-4-9B-Chat	9B	54.7	53.1	73.1	60.2	58.4
InternLM2-20B-Chat	20B	54.7	50.3	73.1	60.9	58.0
Yi1.5-9B-Chat	9B	55.0	44.7	71.4	61.1	56.9
XuanYuan2-70B-Chat	70B	52.8	46.6	68.0	61.7	55.4
CFGPT2-7B	7B	53.9	50.2	65.1	50.9	55.3
Baichuan2-13B-Chat	13B	41.1	50.2	61.6	55.7	47.8
ChatGLM3-6B	6B	38.9	48.6	48.2	49.6	43.2
DISC-FinLLM	13B	39.1	46.1	25.2	41.8	37.8
FinGPTv3.1	6B	25.3	36.1	22.7	31.2	27.1

Table 2: Performance comparison across ordinary individuals, experts and LLMs (selected with top 2 model's results within each category).

Source	Category	Financial Academic	Financial Security	Average
Human	Ordinary individual	35.1	26	30.1
	Experts	**84.9**	**86.8**	**85.9**
Closed-source	Claude 3.5-Sonnet	73.9	78.1	76
	GPT-4o	71.5	81.8	76.7
General	Qwen2.5-72B-Instruct	69.7	81.8	75.8
	Yi1.5-34B-Chat	59.5	76	67.8
Financial	XuanYuan3-70B-Chat	55.2	74.4	64.8
	CFGPT2-7B	53.9	65.1	59.5

4.4 Contrast Analysis

To better evaluate the capabilities of large language models and to make meaningful contributions to model research, we organized a competition between these models, ordinary people, and financial experts. In our dataset, we selected relatively easier multiple-choice questions as competition topics. Considering that some question-and-answer items may be difficult for ordinary individuals and that a limited number of questions would not be representative, we extracted 20% from the financial academic knowledge and financial security knowledge used for evaluation, totaling 260 questions for testing.

For the ordinary participants, we randomly selected three undergraduate students who had no prior exposure to financial or security-related knowledge. For the expert responses, we also randomly selected three experts in the financial field with more than five years of work experience to participate in the answering. To ensure the validity of the results, all participants involved in the testing have not been exposed to any FinEval questions. We present the comparison results between the large language models, ordinary individuals, and financial experts in Table 2. It can be seen that in both the financial academic knowledge and financial security knowledge test results, the performance of closed-source, general, and financial domain models far exceeds that of ordinary individuals, with closed-source models and some general models achieving notably high performance. However, the overall results of the large models still have a gap compared to the experts, with the best models showing nearly a 10% difference from expert results. Nevertheless, certain models, such as GPT-4o and Qwen2.5-72B-Instruct, have demonstrated capabilities in safety that are very close to expert levels, with a gap of around 5%, reflecting the current emphasis on safety in large models.

5 Error Analysis

To further identify the shortcomings of large language models in financial knowledge and tasks, we analyze the errors made by the model during the testing process. Financial Agent includes a series of relatively open-ended, subjective tasks that place high demands on the model, and the model displayed a diverse and rich variety of error types when completing more complex tasks. Three types of errors were identified for answers that did not receive full scores from the evaluation model: Logical Reasoning Error, Contextual Misunderstanding, and Ambiguity Handling Weakness. Logical Rea-

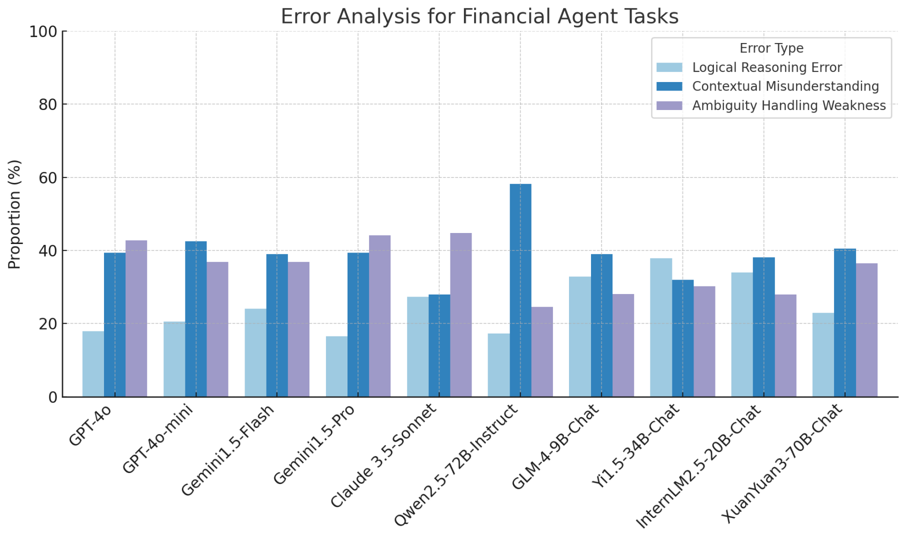

Figure 3: Error analysis results of ten models. Each bar represents the proportion of a specific type of error among all errors made by a particular model. The sum of the values of the three bars for a model equals to 1, representing the total error distribution for that model.

soning Error occurs when the model fails to draw correct logical conclusions. Contextual Misunderstanding happens when the model misses or misinterprets relevant context. Ambiguity Handling Weakness refers to the model's inability to handle unclear or ambiguous questions properly. Detail examples can be found in Appendix B. We selected ten representative models and randomly sampled 27 erroneous responses for each model on each task. The results are shown in Figure 3. Each bar in the chart represents the proportion of a specific type of error among all errors made by a particular model.

Across all the erroneous responses, the average proportion for open-source models in Logical Reasoning Error, Contextual Misunderstanding, and Ambiguity Handling Weakness are 29.0%, 41.6%, and 29.4%, respectively. In comparison, the corresponding average proportion for closed-source models are 21.3%, 37.6%, and 41.1%. This suggests closed-source models are better at reasoning and understanding long texts, but more prone to semantic issues and hallucinations. This could be due to their larger parameter sizes, which, while improving reasoning and comprehension, also lead to over-interpretation of ambiguous content. Specifically, the results show that open-source models are more likely to encounter Logical Reasoning Errors and Contextual Misunderstanding, while closed-source models are more prone to Ambiguity Handling Weakness. For Logical Reasoning Errors, models struggle with multi-step reasoning tasks, often misusing formulas or making calculation mistakes. In Contextual Misunderstanding, models fail to connect distant parts of the text. For Ambiguity Handling Weakness, models have trouble interpreting vague information, sometimes leading to hallucinations. They often include unnecessary tools or data in their responses, indicating a need for better semantic understanding and relevance filtering.

6 Conclusion

This study introduces FinEval, a benchmark for evaluating large language models' capabilities in the financial domain. Unlike previous financial evaluation benchmarks, FinEval delves deeper into financial security and financial agent, covering pressing security issues and complex agent tasks in the financial field, assessing models' security and their ability to handle complex tasks. Our results indicate that Claude 3.5-Sonnet performs the best among the 19 models evaluated, but it still faces challenges with more complex tasks. While it surpasses ordinary individuals, it has not reached the level of human experts. This study illustrates that although large language models have made certain breakthroughs in the financial domain, they still require a more in-depth and detailed understanding to enhance their task generalization capabilities in a wider and more complex diverse financial market environment. As an important benchmark for future research on large language models in the financial field, FinEval provides a structured framework for measuring and improving the capability of large language models, contributing to the development of evaluation benchmarks in the Chinese financial domain.

Limitations

Although FinEval has become a relatively comprehensive evaluation benchmark in the financial domain, encompassing a wide range of financial tasks, we acknowledge its limitations. With the ongoing emergence and iteration of data formats such as image, audio, and video, there is an increasing amount of multimodal data in the financial domain, which presents a limitation for the current FinEval. This limitation highlights the need to develop a multimodal evaluation dataset as the next focus for FinEval. We will recruit more specialized financial personnel to collect additional multimodal data related to the financial domain (especially financial chart data, which is particularly important), to evaluate a broader range of multimodal large language models, ensuring that FinEval remains a comprehensive benchmark in the financial field.

Acknowledgments

This research is supported in part by the National Social Science Fund of China 22BTJ031 (Liwen Zhang). We sincerely appreciate the assistance and guidance provided by Professor Xiao Cao and Associate Professor Min Min from the School of Finance at Shanghai University of Finance and Economics, as well as Shen Wang from the Experimental Center at Shanghai University of Finance and Economics.

References

01.AI. 2024. Yi: Open foundation models by 01.ai. https://github.com/01-ai/Yi-1.5.

AI4Finance-Foundation. 2023. Fingpt: Open-source financial large language models. https://github.com/AI4Finance-Foundation/FinGPT.

Anthropic. 2024. Claude 3.5 sonnet. *https://www.anthropic.com/news/claude-3-5-sonnet*.

Baichuan-inc. 2023. Baichuan2-13b-chat. https://github.com/baichuan-inc/Baichuan2.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021a. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, et al. 2021b. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering. *arXiv preprint arXiv:2210.03849*.

Han Ding, Yinheng Li, Junhao Wang, and Hang Chen. 2024. Large language model agent in financial trading: A survey. *Preprint*, arXiv:2408.06361.

Duxiaoman-DI. 2023. Financeiq: A chinese financial knowledge evaluation benchmark. Accessed: 2024-09-29.

Duxiaoman-DI. 2024a. Xuanyuan2.0: A large chinese financial chat model with hundreds of billions parameters. https://github.com/Duxiaoman-DI/XuanYuan.

Duxiaoman-DI. 2024b. Xuanyuan2.0: A large chinese financial chat model with hundreds of billions parameters. https://github.com/Duxiaoman-DI/XuanYuan.

Fudan-DISC. 2023. Disc-finllm: A chinese financial large language model based on multiple experts fine-tuning. https://github.com/FudanDISC/DISC-FinLLM/tree/main.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Feng He, Tianqing Zhu, Dayong Ye, Bo Liu, Wanlei Zhou, and Philip S. Yu. 2024. The emerged security and privacy of llm agent: A survey with case studies. *Preprint*, arXiv:2407.19354.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, et al. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*.

Guancheng Li, Yifeng Li, Wang Guannan, Haoyu Yang, and Yang Yu. 2023. Seceval: A comprehensive benchmark for evaluating cybersecurity knowledge of foundation models. https://github.com/XuanwuAI/SecEval.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Dakuan Lu, Jiaqing Liang, Yipei Xu, Qianyu He, Yipeng Geng, Mengkun Han, Yingsi Xin, Hengkui Wu, and Yanghua Xiao. 2023. Bbt-fin: Comprehensive construction of chinese financial domain pre-trained language model, corpus and benchmark. *arXiv preprint arXiv:2302.09432*.

OpenAI. 2024a. Gpt-4o mini: advancing cost-efficient intelligence. *https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/*.

OpenAI. 2024b. Hello gpt-4o. *https://openai.com/index/hello-gpt-4o/*.

Qwen. 2024a. Qwen2.5-72b-instruct. `https://github.com/QwenLM/Qwen2.5`.

Qwen. 2024b. Qwen2.5-7b-instruct. `https://github.com/QwenLM/Qwen2.5`.

Raj Sanjay Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. When flue meets flang: Benchmarks and large pre-trained language model for financial domain. *arXiv preprint arXiv:2211.00083*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Gemini Team. 2024a. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

InternLM Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities. `https://github.com/InternLM/InternLM`.

InternLM Team. 2024b. Internlm: A multilingual language model with progressively enhanced capabilities. `https://github.com/InternLM/InternLM`.

THUDM. 2023. Chatglm3-6b. `https://github.com/THUDM/ChatGLM3`.

TongjiFinLab. 2024. Cfgpt2: Chinese financial assistant with large language model. `https://github.com/TongjiFinLab/CFGPT`.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, et al. 2021. Cblue: A chinese biomedical language understanding evaluation benchmark. *arXiv preprint arXiv:2106.08087*.

Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. 2023. Evaluating the performance of large language models on gaokao benchmark. *arXiv preprint arXiv:2305.12474*.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*.

Jie Zhu, Junhui Li, Yalong Wen, and Lifan Guo. 2024. Benchmarking large language models on cflue–a chinese financial language understanding evaluation dataset. *arXiv preprint arXiv:2405.10542*.

A Evaluated Models Overview

We list the models we evaluated in this paper in Table 3.

B Examples for FinEval

We list the examples of FinEvalin this paper in Figure 4, 5, 6, 7, 8, 9, 10, 11, 12. In which, Figure 4, 5, 6, 7 are examples of zero-shot, five-shot, zero-shot CoT, and five-shot CoT, respectively. Figure 8 is an objective short-answer question, and figure 9 is a subjective open-ended question. Figure 10 is an example in LaTeX format. Figure 11, 12 are examples of two subclasses (API invocation and task planning) under the theme of Financial Agent. Figure 13, 14 are evaluation prompts for API invocation and task planning respectively. Figure 15, 16, 17, 18 are examples of typical error LLM makes in solving financial agent tasks.

C Detailed Statistics of FinEval

Table 4 presents detailed information on the four sections of FinEval.

D Other results

D.1 Five-shot Results Analysis

In the five-shot setting across three financial tasks, Claude 3.5-Sonnet achieved the highest overall weighted average score (73.2). GPT-4o excelled in Financial Security Knowledge with a score of 81.8, while Qwen2.5-72B-Instruct performed strongly in Financial Security as well, achieving 80.2. (Table 5).

D.2 Zero-shot CoT Results Analysis

In the zero-shot CoT setting across three financial tasks, Claude 3.5-Sonnet achieved the highest overall weighted average score (72.4). Qwen2.5-72B-Instruct led in Financial Academic Knowledge (76.5), while GPT-4o excelled in Financial Security Knowledge (81.8). (Table 6).

D.3 Five-shot CoT Results Analysis

In the five-shot CoT setting, Qwen2.5-72B-Instruct achieved the highest overall weighted average score (73.6), leading in Financial Academic Knowledge (75.7) and performing well in Financial Security (78.5). GPT-4o excelled in Financial Security with a score of 81.2. (Table 7).

D.4 Comprehensive Results for Individual Financial Tasks Across Four Settings

Evaluation results for finance academic knowledge under four settings are shown in Table 8, 9, 10 and 11. Evaluation results for finance industry knowledge including objective and subjective questions under four settings are shown in Table 12, 13, 14, 15, 16, 17, 18 and 19. Evaluation results for finance security knowledge under four setting are shown in Table 20, 21, 22 and 23. Evaluation result for finance agent under zero-shot setting is shown in Table 24.

D.5 Results Analysis

The performance of various open-source models varies under FinEval, and we analyze the reasons as follows:

(1)First, FinEval is quite challenging, closely integrated with professional knowledge and real-world business scenarios, providing a more realistic reflection of the various models' true capabilities.

(2)Second, most financial LLMs on the market are derived from general base models that have been trained, rather than being directly trained base models. This can lead to some loss in capability.

(3)Third, the general closed and open-source models that perform well in FinEval evaluations are recognized as stronger LLMs in the current LLM field, so it is normal for financial LLMs, such as the XuanYuan series, not to perform exceptionally well.

(4)Finally, many of the underperforming financial LLMs come from various universities or small AI organizations, which typically utilize LoRA fine-tuning. These models often lack vast computing resources, leading to poorer fine-tuning results. Moreover, these financial LLMs have not undergone any updates or iterations, making their subpar performance reasonable.

D.6 Comparative Analysis Under Different Evaluation Methods

D.6.1 Correlation analysis between human evaluation and GPT evaluation

In this section, we will demonstrate that GPT evaluation can replace human evaluation in large-scale testing by analyzing the correlation between human evaluation and GPT evaluation results. Table 25 is the Spearman correlation matrix of human evaluation scores and GPT evaluation scores for Gemini 1.5-Pro across seven agent tasks, each task

Table 3: Models evaluated in this paper. The "Access" column shows whether we have full access to the model weights or we can only access through API. The “Version Date” column shows the release date of the corresponding version of the model we evaluated.

Category	**Model**	**Creator**	**Parameter**	**Access**	**Version Date**
Closed-Source	GPT-4o	OpenAI	undisclosed	API	2024.5
	GPT-4o-mini	OpenAI	undisclosed	API	2024.7
	Gemini1.5-Flash	Google	undisclosed	API	2024.5
	Gemini1.5-Pro	Google	undisclosed	API	2024.5
	Claude 3.5-Sonnet	Anthropic	undisclosed	API	2024.3
Open-Source	Qwen2.5-7B-Instruct	Alibaba Cloud	7B	Weights	2024.9
	Qwen2.5-72B-Instruct	Alibaba Cloud	72B	Weights	2024.9
	ChatGLM3-6B	Tsinghua & Zhipu.AI	6B	Weights	2023.10
	GLM-4-9B-Chat	Tsinghua & Zhipu.AI	9B	Weights	2024.6
	Yi1.5-9B-Chat	01.AI	9B	Weights	2024.5
	Yi1.5-34B-Chat	01.AI	34B	Weights	2024.5
	InternLM2-20B-Chat	Shanghai AI Lab & SenseTime	20B	Weights	2024.1
	InternLM2.5-20B-Chat	Shanghai AI Lab & SenseTime	20B	Weights	2024.8
	Baichuan2-13B-Chat	Baichuan	13B	Weights	2023.12
Financial	XuanYuan3-70B-Chat	Duxiaoman-DI	70B	Weights	2024.9
	XuanYuan2-70B-Chat	Duxiaoman-DI	70B	Weights	2024.3
	DISC-FinLLM	FudanDISC	13B	Weights	2023.10
	CFGPT2-7B	TongjiFinLab	7B	Weights	2024.8
	FinGPTv3.1	AI4Finance-Foundation	6B	Weights	2023.10

甲公司2015年年初的所有者权益总额为2000万元，2015年亏损200万元，2016年亏损300万元，2017年到2019年的税前利润每年均为0，2020年公司实现税前利润800万元。按净利润的10%提取法定盈余公积。公司董事会提出2020年度分配利润50万元，但尚未提交股东大会审议。假设公司2015年至2020年不存在除弥补亏损的其他纳税调整和其他导致所有者权益变动的事项，适用企业所得税税率为25%。则甲公司2020年年末所有者权益总额为____万元。
The total owner's equity of Company A was 20 million yuan at the beginning of 2015, and the loss was 2 million yuan in 2015 and 3 million yuan in 2016. The pre-tax profit from 2017 to 2019 was 0 million each year, and the company achieved a pre-tax profit of 8 million yuan in 2020. The legal surplus reserve shall be withdrawn at 10%of the net profit. The board of directors of the company proposed the annual distribution profit of 500,000 yuan in 2020, but it has not yet been submitted to the general meeting of shareholders for consideration. Assuming that the company does not have other tax adjustments other than making up losses and other matters resulting in changes in owners' equity from 2015 to 2020, the applicable corporate income tax rate is 25%. The total owner's equity of Company A at the end of 2020 is ____ million yuan.
A. 2300
B. 2145
C. 2225
D. 2175
答案：C
Answer: C

Figure 4: Zero-shot example of multiple-choice questions in Intermediate Financial Accounting. For better readability, the English translation is displayed below the corresponding Chinese text.

samples 20% of its entire data for testing. The resulting matrix is derived from the human evaluation result matrix and the GPT evaluation result matrix. The diagonal elements represent the correlation of scores for the same task under both evaluation systems, while the off-diagonal elements represent the correlation of scores for different tasks. The correlation coefficients for human evaluation scores and GPT evaluation scores for each task are 0.66, 0.84, 0.68, 0.85, 0.85, 0.63, and 0.55, with an average correlation coefficient of 0.72. The similarity for each task are 0.70, 0.63, 0.70, 0.96, 0.67, 0.74, and 0.93, with an average similarity of 0.76. It can be seen that the GPT evaluation scores have a high correlation with human evaluation scores, indicating that GPT can replace human scoring in large-scale testing.

D.6.2 Evaluation Results Using Different LLMs as Judges

We employed Claude-3.5-Sonnet and Gemini-1.5-Pro as judges to score Agent tasks according to expert evaluation criteria in order to compare their differences with GPT-4o. We conducted tests on five closed-source LLMs that participated in the FinEval evaluation. The results of Claude-3.5-Sonnet are shown in Table 26, the results of Gemini-1.5-Pro are shown in Table 27, and the comparative

以下是关于金融学的单项选择题，请选出其中的正确答案。
The following are multiple-choice questions about finance. Please choose the correct answer.
甲企业向乙企业提供价值100万元商品，乙企业承诺分10个月偿付货款。甲企业与乙企业之间的信用形式为____。
Company A provides Company B with a product worth 1 million yuan, and Company B promises to pay the payment in 10 installments. The credit form between Company A and Company B is____.
A. 国家信用
A. National credit
B. 间接信用
B. Indirect credit
C. 商业信用
C. Commercial credit
D. 消费信用
D. Consumer credit
答案：C
Answer: C

某5年期、面额100元的债券，以80元的价格贴现发行，其票面收益率为____。
A bond with a face value of 100 yuan, a maturity of 5 years, and issued at a discount price of 80 yuan. The coupon yield of the bond is ____.
A. 4%
B. 20%
C. 5%
D. 16%
答案：C
Answer: C

…[5-shot examples]…

某企业因业务需要，申请银行为其开具银行汇票，银行经审查后，同意企业的申请，为其开具了一张100万元的银行汇票。银行的这一业务体现了银行的____。
A company applies to a bank to issue a bank draft for business needs. After review, the bank agrees to issue a bank draft worth 1 million yuan for the company. This business of the bank reflects the bank's____.
A. 信用中介职能
A. Credit intermediary function
B. 支付中介职能
B. Payment intermediary function
C. 化货币收入为资本职能
C. Monetization of income into capital function
D. 创造信用流通工具职能
D. Creation of credit circulation tool function
答案：D
Answer: D

Figure 5: An instance of five-shot evaluation. The red text denotes the response automatically generated by the model, with the preceding text being the input prompt. English translations for the related Chinese text are provided beneath.

results of the three judging LLMs are shown in Table 28. From the comparison of the results, it is evident that the three judging LLMs performed differently. GPT-4o was stricter in its evaluation, while Gemini-1.5-Pro was more lenient, and Claude-3.5-Sonnet fell between the two.

分期付款业务手续费不包括括以下哪一种____。
Which of the following is not included in the installment payment business service fee?
A. 分期付款手续费
A. Installment service fee
B. 提前还款手续费
B. Early repayment fee
C. 分期付款展期手续费
C. Installment extension fee
D. 提前还款展期手续费
D. Early repayment extension fee
答案：D
Answer：D
让我们一步步分析：
Let's analyze step by step:
1.A. 分期付款手续费：常见的费用类型。
1.A. Installment service fee: A common type of fee.
2.B. 提前还款手续费：提前还款时可能产生的费用。
2.B. Early repayment fee: A fee that may be charged when repaying early.
3.C. 分期付款展期手续费：延长分期时可能产生的费用。
3.C. Installment extension fee: A fee that may be charged when extending the installment period.
4.D. 提前还款展期手续费：提前还款和展期是相反的概念，不可能同时存在。
4.D. Early repayment extension fee: Early repayment and extension are opposite concepts; they cannot coexist.
5.所以，正确答案是 D. 提前还款展期手续费。
5. Therefore, the correct answer is D. Early repayment extension fee.

Figure 6: An example of zero-shot CoT evaluation. The red text is followed by the response automatically generated by the model. The English translations for the related Chinese text are provided below.

以下是中国关于banking_practitioner_qualification_certificate考试的单项选择题，请选出其中的正确答案。
Below are single-choice questions about the banking_practitioner_qualification_certificate exam in China. Please select the correct answer.

下列关于理财业务的理解，说法有误的是____。
Which of the following statements about financial management services is incorrect?
A. 综合理财服务中，银行可以让客户承担一部分风险
A. In comprehensive financial services, banks can allow customers to bear some of the risks.
B. 与理财顾问服务相比，综合理财服务更强调个性化
B. Compared to financial advisory services, comprehensive financial management services emphasize personalization more.
C. 私人银行业务除了提供金融产品外，更重要的是提供全面的服务
C. In private banking services, in addition to providing financial products, it is more important to provide comprehensive services.
D. 私人银行业务不是个人理财业务
D. Private banking services are not personal financial management services.

答案：让我们一步一步思考，
Answer: Let's think step by step,
1. 理财业务可分为理财业务（服务）、财富管理业务（服务）和私人银行业务（服务）三个层次。
1. Financial management services can be divided into three levels: financial management services, wealth management services, and private banking services.
2. 私人银行业务（服务）除了提供金融产品外，更重要的是提供全面的服务。
2. Private banking services, in addition to providing financial products, more importantly, provide comprehensive services.
所以答案是D。
So the answer is D.

………………（其他4个例子已省略）
……………… (The other four examples have been omitted.)

能为公司和个人提供各种产品和全面的服务的营销渠道____。
The marketing channel that can provide a variety of products and comprehensive services to companies and individuals is____.

A. 高端化网点机构营销渠道	B. 专业性网点机构营销渠道
B. High-end branch marketing channels	B. Professional branch marketing channels
C. 零售型网点机构营销渠道	D. 全方位网点机构营销渠道
C. Retail branch marketing channels	D. Omnidirectional branch marketing channels

答案：让我们一步一步思考，
Answer: Let's think step by step,
1.营销渠道是为客户提供产品和服务的重要途径，不同类型的网点机构会提供不同层次的服务。
1. Marketing channels are an important way to provide customers with products and services, and different types of branch institutions will provide different levels of services.
2. A. 高端化网点机构营销渠道：主要为高端客户提供专属服务，可能不包括面向普通客户的全面服务。
2. A. High-end branch marketing channels: mainly provide exclusive services for high-end customers, which may not include comprehensive services for ordinary customers.
3. B. 专业性网点机构营销渠道：主要提供特定专业领域的服务，不能覆盖所有客户的需求。
3. B. Professional branch marketing channels: mainly provide services in specific professional fields and cannot cover all the needs of customers.
4. C. 零售型网点机构营销渠道：主要面向零售客户，服务对象较为单一，不能为公司和个人同时提供全面服务。
4. C. Retail branch marketing channels: mainly face retail customers, and the service objects are relatively single, and cannot provide comprehensive services for companies and individuals at the same time.
5. D. 全方位网点机构营销渠道：顾名思义，能够为公司和个人提供各种产品和全面服务，涵盖面最广。
5. D. Omnidirectional branch marketing channels: as the name suggests, it can provide a variety of products and comprehensive services to companies and individuals, covering the broadest range.
因此，D 选项正确，因为全方位网点机构可以为不同类型的客户提供多样化的服务和产品。
Therefore, option D is correct because the omni-directional branch institution can provide a variety of services and products to different types of customers.

Figure 7: An example of five-shot CoT evaluation. The preceding text serves as an example, for convenience, the other four examples have been omitted. The red text is followed by the response automatically generated by the model. The English translations for the related Chinese text are provided below.

请判断以下投资者的情绪属于消极情绪，中立，积极情绪中的哪一类，不需要分析，只需指出属于哪一类：根据以往的游戏开发经验，互动剧游戏研发制作的两大重点分别是内容拍摄制作和后续游戏化开发。在内容拍摄制作环节，飞天云动已经有了深厚的短剧储备，未来这些短剧储备具备快速完成向互动剧游戏转化的条件。对于后续的游戏化而言，在公司自研的编辑器加持下，互动环节开发相对容易。同时，公司也在布局AI生成脚本、AI视频剪辑、AI生成互动内容等技术，在AIGC技术的加持下，剧情生成、短剧开发和游戏化开发效率提升，更有可能形成工业化的制作流程体系。

Please classify the emotions of the following investors as negative, neutral, or positive, without analyzing, only indicating the category: Based on previous game development experience, the two key aspects of interactive drama game development are content filming and production, and subsequent gamification development. In the content filming and production phase, Feitian Yundong already has a strong reserve of short dramas, which provides favorable conditions for quickly transforming these short dramas into interactive drama games. As for the subsequent gamification, with the support of the company's self-developed editor, the development of interactive segments is relatively easy. At the same time, the company is also investing in AI-generated scripts, AI video editing, AI-generated interactive content, and other technologies. With the support of AI and GC technologies, the efficiency of plot generation, short drama development, and gamification development is expected to increase, and there is a greater possibility of forming an industrialized production workflow system.

答案：消极

Answer： Negative

Figure 8: An example of objective short-answer questions in Financial Sentiment Analysis. For better readability, the English translation is displayed below the corresponding Chinese text.

问题：我应该如何评估债券投资的风险和回报？

Question: How should I assess the risk and return of bond investments?

答案：评估债券投资的风险和回报是一个关键的投资决策过程，以下是一些建议：

Answer: Assessing the risk and return of bond investments is a key investment decision process, and here are some suggestions:

1.多样化投资组合：分散投资于不同信用评级、行业和到期期限的债券，以降低整体投资组合的风险。

1. Diversify your portfolio: Spread investments across bonds with different credit ratings, industries, and maturities to reduce the overall risk of the portfolio.

2.关注财务健康：选择那些财务状况良好、信用评级较高的债券，以降低信用风险。

2. Pay attention to financial health: Opt for bonds with good financial conditions and higher credit ratings to mitigate credit risk.

3.考虑市场环境：在宏观经济环境较好时，可能更偏向于低风险、低回报的债券；而在经济疲软时，可能需要更关注保值和较高回报。

3. Consider the market environment: In a favorable macroeconomic environment, you might lean towards bonds with lower risk and lower returns; while in a weak economy, you might need to focus more on preservation of value and higher returns.

4.定期重新评估：定期审查投资组合，关注市场和经济的变化，根据需要调整投资组合以适应不断变化的环境。

4. Regularly re-evaluate: Periodically review your portfolio, paying attention to changes in the market and economy, and adjust your portfolio as needed to adapt to the evolving environment.

最终，债券投资的选择应该与你的投资目标、风险承受能力和投资时间视角相一致。记住，投资是一项长期的活动，理性的决策和谨慎的分析是成功投资的关键。

Ultimately, the choice of bond investments should align with your investment objectives, risk tolerance, and investment time horizon. Remember, investing is a long-term activity, and rational decision-making and cautious analysis are key to successful investing.

Figure 9: An example of subjective open-ended questions in Financial Investment Advice. For better readability, the English translation is displayed below the corresponding Chinese text.

设$X_{1},X_{2},\cdots,X_{n}$是来自正态总体$N(\mu,\sigma^{2})$的一个简单随机样本，$\sigma^{2}的极大似然估计为$\hat{\sigma}^{2}=\frac{1}{n}\sum_{k=1}^{n}(X_{k}-\bar{X})^{2}$则$\hat{\sigma}^{2}$的渐近分布是______。\n A.$N(\sigma^{2},2\sigma^{4})$\n B.$N(\sigma^{2},\sigma^{4})$\n C.$N(\sigma^{2},\frac{2\sigma^{4}}{n})$\n
Let $X_{1},X_{2},\cdots,X_{n}$ be a simple random sample from a normal population $N(\mu,\sigma^{2})$. The maximum likelihood estimator of σ^{2} is given by $\hat{\sigma}^{2}=\frac{1}{n}\sum_{k=1}^{n}(X_{k}-\bar{X})^{2}$. The asymptotic distribution of $\hat{\sigma}^{2}$ is ____. \n A. $N(\sigma^{2},2\sigma^{4})$\n B. $N(\sigma^{2},\sigma^{4})$\n C. $N(\sigma^{2},\frac{2\sigma^{4}}{n})$
答案：C
Answer： C

Figure 10: An example of LaTeX format in the subject of Statistics, under the category of economy.

yfinance 是一个Python库，用于从Yahoo Finance获取全球金融市场数据。它提供了访问和下载股票、指数、债券、货币对和大宗商品等各类金融产品的历史和实时数据。特别地，yfinance 还支持获取股票市场的市场深度数据，包括买卖盘深度、成交量和成交价等。现在，你的任务是使用 yfinance 获取Netflix（股票代码：NFLX）和迪士尼（股票代码：DIS）在过去一年的每月平均成交量。请问根据API描述，你会怎么操作来实现这个需求？
yfinance is a Python library used to retrieve global financial market data from Yahoo Finance. It provides access to historical and real-time data for various financial products such as stocks, indices, bonds, currency pairs, and commodities. Specifically, yfinance also supports obtaining market depth data for stocks, including bid/ask depth, volume, and prices. Now, your task is to use yfinance to retrieve the monthly average volume for Netflix (stock code: NFLX) and Disney (stock code: DIS) over the past year. How would you proceed to achieve this requirement based on the API description?
答案: 要使用 `yfinance` 获取 Netflix（股票代码：NFLX）和迪士尼（股票代码：DIS）在过去一年的每月平均成交量，你可以按照以下步骤操作：
Answer: To retrieve the monthly average volume for Netflix (stock code: NFLX) and Disney (stock code: DIS) using yfinance, you can follow these steps:
1.**安装和导入 yfinance**: 确保已安装 `yfinance` 库。如果尚未安装，可以通过 pip 安装。然后导入 `yfinance` 库。
1.**Install and import yfinance**: Make sure you have the yfinance library installed. If not, you can install it using pip. Then import the yfinance library.
pip install yfinance
import yfinance as yf
2.**获取股票数据**: 使用 `yfinance` 来获取 Netflix 和迪士尼的股票数据。需要指定获取数据的时间范围为过去一年。
2.**Retrieve stock data**: Use yfinance to retrieve the stock data for Netflix and Disney. Specify the time range to be the past year.
from datetime import datetime, timedelta
# 获取今天的日期并计算一年前的日期
# Get today's date and calculate the date one year ago
today = datetime.today()\none_year_ago = today - timedelta(days=365)
# 使用 yfinance 获取特定时间段内的股票数据
#Use yfinance to get the stock data for the specified time range
nflx = yf.Ticker(\"NFLX\")\ndis = yf.Ticker(\"DIS\")
nflx_data = nflx.history(start=one_year_ago, end=today)\ndis_data = dis.history(start=one_year_ago, end=today)
3.**提取月度平均成交量**: 对获取到的历史数据进行处理，以提取每月的平均成交量。这可以通过重采样数据到月度频率，并计算每月的平均成交量来实现。
3.**Extract monthly average volume**: Process the retrieved historical data to extract the monthly average volume. This can be achieved by resampling the data to a monthly frequency and calculating the average volume for each month.
import pandas as pd
# 重采样到月度，并计算每月的平均成交量
# Resample to monthly and calculate the average volume for each month
monthly_avg_volume_nflx = nflx_data['Volume'].resample('M').mean()
monthly_avg_volume_dis = dis_data['Volume'].resample('M').mean()
4.**输出结果**: 输出每月的平均成交量。
4.**Output the results**: Print out the monthly average volume for each stock.
print(\"Netflix Monthly Average Volume:\")
print(monthly_avg_volume_nflx
print(\"\\nDisney Monthly Average Volume:\")\nprint(monthly_avg_volume_dis)
这将给你 Netflix 和迪士尼过去一年中每月的平均成交量。请在你的本地环境中尝试这些代码，如果在执行过程中遇到任何问题，或需要进一步的帮助，请随时告诉我！
This will give you the monthly average volume for Netflix and Disney over the past year. Please try these code snippets in your local environment, and if you encounter any issues during the execution or need further assistance, feel free to let me know!

Figure 11: An example of API invocation by financial agent evaluation. The red text is followed by the response automatically generated by the model. The English translations for the related Chinese text are provided below.

问题：筹办金融科技与全球经济一体化论坛】预计300人参与，预算55万元，时间四个月，探讨金融科技在全球经济一体化中的作用和挑战。
Question: Organizing the "Financial Technology and Global Economic Integration Forum" with an expected participation of 300 people, a budget of 550,000 RMB, and a duration of four months, to discuss the role and challenges of financial technology in global economic integration.
答案: 筹办一个关于金融科技与全球经济一体化的论坛是一个重要且具有挑战性的项目。考虑到您的预算、参与人数和时间框架，以下是一些关键的规划步骤和建议：
Answer: Organizing a forum on financial technology and global economic integration is an important and challenging project. Considering your budget, number of participants, and the timeframe, here are some key planning steps and suggestions:
1. 目标和议程设定
1.Goal and Agenda Setting:
- 明确论坛的主要目标和预期成果。
- Clearly define the main objectives and expected outcomes of the forum.
- 设定具体的议程，包括主题演讲、小组讨论、互动研讨会等。
- Develop a specific agenda, including keynote speeches, panel discussions, interactive workshops, etc.
- 邀请金融科技和全球经济一体化领域的专家、学者和业界领袖作为演讲嘉宾。
- Invite experts, scholars, and industry leaders in the field of financial technology and global economic integration as guest speakers.
2. 场地和日期选择
2.Venue and Date Selection:
- 选择一个可以容纳至少300人的场地，确保有足够的会议室和休息空间。
- Choose a venue that can accommodate at least 300 people, ensuring sufficient meeting rooms and breakout spaces.
- 考虑到时间框架为四个月，应尽早预定场地，并确保日期适合大多数嘉宾。
- Considering the four-month timeframe, book the venue early and ensure the dates are suitable for the majority of the invitees.
3. 预算规划
3.Budget Planning:
- 55万元的预算需要合理分配于场地租赁、嘉宾邀请（可能包括差旅和住宿费用）、宣传材料、技术支持、餐饮服务等。
- Allocate the 550,000 RMB budget appropriately for venue rental, guest invitations (which may include travel and accommodation expenses), promotional materials, technical support, catering services, etc.
- 保留一部分预算用于不可预见的支出。
- Reserve a portion of the budget for unforeseen expenses.
4. 宣传和招募参与者
4.Promotion and Participant Recruitment:
- 制定一套全面的宣传计划，包括社交媒体、行业合作伙伴、专业论坛等渠道。
- Develop a comprehensive promotional plan, including social media, industry partnerships, professional forums, etc.
- 开放在线报名，提供早鸟票和团体票优惠以吸引更多参与者。
- Open online registration, offering early bird and group ticket discounts to attract more participants.
5. 技术和后勤安排
5.Technical and Logistics Arrangements:
- 确保会场有稳定的网络连接，以支持演示和在线直播。
- Ensure the venue has a stable network connection to support presentations and online streaming.
- 安排专业的音频和视频设备，确保所有演讲和讨论都能被清晰记录和传播。
- Arrange professional audio and video equipment to ensure all speeches and discussions are recorded and disseminated clearly.
6. 跟进和评估
6.Follow-up and Evaluation:
- 会后收集反馈，评估论坛的影响和效果。
- Collect feedback after the forum to assess its impact and effectiveness.
- 考虑制作会议记录和总结报告，以供未来参考。
- Consider producing conference proceedings and summary reports for future reference.
此外，考虑到当前全球经济形势和金融科技的快速发展，论坛可以探讨的关键话题可能包括数字货币、跨境支付、区块链技术在金融服务中的应用、金融监管科技（RegTech）等。
Additionally, considering the current global economic situation and the rapid development of financial technology, key topics that the forum can explore may include digital currencies, cross-border payments, the application of blockchain technology in financial services, regulatory technology (RegTech), etc.
确保从现在开始就有一个详细的时间线和责任分配表，以便各个团队成员明确自己的任务和截止日期。成功筹办此类论坛需要团队的紧密合作和高效执行。
Ensure there is a detailed timeline and responsibility assignment sheet starting from now, so that each team member is clear about their tasks and deadlines. Successful execution of such a forum requires close collaboration and efficient execution from the team.

Figure 12: An example of task planning by financial agent evaluation. The red text is followed by the response automatically generated by the model. The English translations for the related Chinese text are provided below. Due to the inconsistency and complexity of the data in seven sections of the financial agent, we have only provided examples for two sections. For other examples, please refer to https://github.com/SUFE-AIFLM-Lab/FinEval.

在“4.提问与回答”中，提问为给出的问题，回答为被测评agent的答案。请严格依据“3.评分标准”分析问题并根据评分标准对回答进行评分。请逐项评价每一项的得分，但不要进行输出！请最终进行加总并给我最后的分数。
In section “4. Questioning and Answering,” the questions refer to the given problems, while the answers are the responses provided by the evaluated agent. Please strictly analyze the questions in accordance with the criteria outlined in section “3. Evaluation Standards” and score the answers based on these standards. Evaluate each criterion individually but do not output the scores for each. At the end, provide me with the total score only.
3. 评分标准 以下是基于十分制的评分标准：
理解和规划（2分）:分配0分，如果Agent没有表现出对提问API描述的基本理解；或Agent没有明确回答。分配0.5分，如果Agent表现出了对提问API描述的基本理解，但理解有一定偏差。分配1分，如果Agent不仅基本正确理解API描述，还能明确指出所需的API功能（如数据检索、分析）。分配2分，如果Agent能够精确地规划出如何使用API完成任务，包括理解API的输入、输出和功能限制。
代码实现（3分）:分配0分，如果Agent基本没有给出任何有用的代码实现或存在许多错误遗漏。或有大量操作Agent只进行描述未实际给出代码。分配1分，如果Agent提供了部分相关的代码实现但存在一些错误或遗漏；分配2分，如果Agent的调用的api全部正确只偶尔不完整或有小错误。分配3分，如果AI Agent提供了完整的、正确的代码实现，包括准确调用API并合理处理API返回的数据。
数据处理与分析（2分）:分配0分，如果Agent基本无法处理和分析API返回的数据或分析完全欠缺深度或广度；或有大量操作只进行描述未实际给出操作方法；分配1分，如果Agent能基本处理和分析API返回的数据但分析深度或广度不足。分配2分，如果Agent展示了高级的数据处理和分析技巧，能从数据中提取出有深度的洞见。
结果准确性和完整性（2分）:分配0分，如果Agent返回的结果基本是错误或不可接受的，或没有完整具体完成任务（包括通过分析跳过具体实现等）。分配1分，如果Agent返回的结果基本正确，但缺少关键细节或部分不准确。分配2分，如果Agent的回答完全准确，没有遗漏任何关键信息。
清晰度和表达（1分）:分配0分，如果如果Agent的回答基本不具有可读性。分配0.5分，如果Agent的回答可读，但可能存在一些语言上的不清晰或组织上的问题；或回答过于冗长、啰嗦，存在不必要的信息分配1分，如果Agent的回答语言表达清晰，逻辑组织良好，易于理解。
3. Evaluation Standards The following are the evaluation standards based on a 10-point scale:
Understanding and Planning (2 points): 0 points: if the agent shows no basic understanding of the API description in the question or does not provide a clear answer. 0.5 points: if the agent shows some understanding of the API description but with notable misinterpretations. 1 point: if the agent not only correctly understands the API description but can also clearly identify the required API functionality (e.g., data retrieval, analysis). 2 points: if the agent precisely plans how to use the API to accomplish the task, including understanding the API’s inputs, outputs, and limitations.
Code Implementation (3 points): 0 points: if the agent provides no useful code implementation or there are significant errors and omissions; or if the agent describes a large number of operations without providing code. 1 point: if the agent provides partial code implementation with some errors or omissions. 2 points: if the agent’s API calls are mostly correct with only occasional incompleteness or minor errors. 3 points: if the AI agent provides a complete and correct code implementation, including accurately calling the API and appropriately handling the returned data.
Data Processing and Analysis (2 points): 0 points: if the agent is unable to process or analyze the data returned by the API, or the analysis lacks sufficient depth or breadth; or if many operations are only described without actual methods provided. 1 point: if the agent can process and analyze the API’s returned data, but the depth or breadth of analysis is lacking. 2 points: if the agent demonstrates advanced data processing and analysis skills, extracting deep insights from the data.
Result Accuracy and Completeness (2 points): 0 points: if the agent’s returned result is largely incorrect or unacceptable, or the task is not completed comprehensively (e.g., skipping specific implementation in the analysis). 1 point: if the agent’s returned result is mostly correct but lacks key details or has partial inaccuracies. 2 points: if the agent’s response is completely accurate, with no critical information omitted.
Clarity and Expression (1 point): 0 points: if the agent’s response is largely unreadable. 0.5 points: if the agent’s response is readable but may have some unclear language or organizational issues; or the response is overly verbose with unnecessary information. 1 point: if the agent’s response is clearly expressed, well-organized, and easy to understand.
4. 提问与回答 提问{}. 回答{}
4. Questioning and Answering Question {}:Answer {}:
5. 注意事项：(1)若出现“追问”部分的内容请忽略，仅对第一个问题的“回答”进行评分，不需要有任何说明。(2)在评分时，你需要认为输出答案的使用用户对于该领域或编程方面完全没有任何了解。因此用户将严格按照给定的回答进行操作。故如果给出的指令并不能真正直接运行，该回答应当认为是不可接受的。agent给出的指令应当是清晰、具体、完整、可执行的。(3)请仅输出最终的打分结果，忽略掉中间的解释和分析过程。我只需要一个int或float的数字，如2。在输出前请确保你只返回了数字。
5. Notes: (1) If there is content under a "follow-up question" section, please ignore it and only score the answer to the first question without providing any explanation. (2) When scoring, assume that the user receiving the answer has no knowledge of the field or programming. Therefore, the user will strictly follow the given instructions. If the provided instructions cannot be executed directly, the answer should be considered unacceptable. The agent's instructions must be clear, specific, complete, and executable. (3) Please output only the final score, ignoring intermediate explanations and analysis. I only need a single integer or float number, such as 2. Before outputting, ensure that only the number is returned.

Figure 13: The evaluation prompt of API invocation by financial agent evaluation.

我会给你一个'问题'，与一个'待打分回答'，请根据评分标准对'待打分回答'进行打分。
I will give you a 'question' and a 'response to be scored.' Please score the 'response to be scored' based on the scoring criteria.
回复要求：你只需要回复一个数字表示总分，如'2',不需要具体的评分过程。请不要轻易给5分。
Response requirements: You only need to reply with a single number representing the total score, such as '2', without providing the detailed scoring process. Please do not give 5 points lightly.
评分标准：
1. 完整性：任务规划完全覆盖了所有关键点和必要的细节。加1分。
2. 准确性：对于任务规划要求的理解完全准确，无错误信息。加1分。
3. 逻辑性和条理性：结构清晰，逻辑连贯，易于理解。加1分。
4. 实用性和可行性：提出的方案或信息非常实用，具有高度的可行性。加1分。
5. 创新性：提供了创新性的见解或独特的解决方案。加1分。
6. 满分5分，你只有0，1，2，3，4，5六个选项。
Scoring Criteria:
1. Completeness: The task planning fully covers all key points and necessary details. Add 1 point.
2. Accuracy: The understanding of the task planning requirements is entirely accurate, with no incorrect information. Add 1 point.
3. Logical Structure: The structure is clear, logically coherent, and easy to understand. Add 1 point.
4. Practicality and Feasibility: The proposed solution or information is highly practical and feasible. Add 1 point.
5. Innovation: The response provides innovative insights or unique solutions. Add 1 point.
6. The full score is 5 points, and you only have six options: 0, 1, 2, 3, 4, 5.
问题：{} 待打分回答：{}
Question: {} Response to be scored: {}
回复要求：你只需要回复一个数字表示总分，如‘2’，不需要具体的评分过程。严格按照评分标准打分。请不要轻易给5分。
Response requirements: You only need to reply with a single number representing the total score, such as '2,' without providing the detailed scoring process. Strictly follow the scoring criteria. Please do not give 5 points lightly.

Figure 14: The evaluation prompt of task planning by financial agent evaluation.

Question Type: Financial Industry Knowledge - Investment Research - Financial Text Summarization (金融文本摘要)
材料：在5月27日举办的中国上市公司协会年会（理事会）暨2023中国上市公司峰会上，中国证监会科技监管局副局长蒋东兴表示，上市公司要以践行“新发展理念，构建新发展格局，促进数字经济和实体经济融合发展”为整体目标，充分发挥上市公司在企业数字化转型中的引领作用。要加大力度，持续推进上市公司数字化转型和提质增效，赋能上市公司高质量发展。
Material: At the Annual Meeting (Board of Directors) of the China Association for Listed Companies and the 2023 China Listed Companies Summit held on May 27th, Jiang Dongxing, Deputy Director of the Technology Supervision Bureau of the China Securities Regulatory Commission, emphasized that listed companies should aim to "embrace the new development philosophy, build a new development pattern, and promote the integration of the digital and real economies." Listed companies are expected to play a leading role in the wave of digital transformation among enterprises. Efforts should be intensified to continue driving the digital transformation of listed companies, improving efficiency, and enhancing the quality of development.
期望答案：证监会科技监管局蒋东兴：加大力度推进上市公司数字化转型。
Good Answer: Jiang Dongxing from the Technology Supervision Bureau of the China Securities Regulatory Commission: Intensify efforts to promote the digital transformation of listed companies.
GPT-4o输出：上市公司要引领数字化转型，促进高质量发展。
GPT-4o Answer: Listed companies should lead the digital transformation and promote high-quality development.

Figure 15: An example of error model(GPT-4o) encountered while solving a financial text summarization problem related to handling long texts. The expected answer was to include both the entity announcing the policy and the policy content, but the model's output only focused on the latter part of the material, addressing the policy content, while neglecting the entity that announced the policy at the beginning.

Question Type：Finance Agent-Reasoning and Planning-FinCoT(思维链)
问题：某企业投资生产，40万元用于固定资本，以购买机器设备等，60万元用于流动资本，以购买原材料和劳动力等（其中购买劳动力支付了40万元），一轮生产结束后，该企业的剩余价值率为100$\\%$。那么，该企业的平均利润率为 ____。
Question:A company invested 400,000 yuan in fixed capital to purchase machinery and equipment, and 600,000 yuan in circulating capital to buy raw materials and labor (of which 400,000 yuan was used to pay for labor). After one round of production, the company had a surplus value rate of 100%. What is the company's average profit rate?
期望答案：解题步骤：

1） 确定企业的总投资：
总投资 = 固定资本 + 流动资本
= 40万元 + 60万元
= 100万元
2） 确定剩余价值率：
剩余价值率 = 剩余价值 / 可变资本
已知剩余价值率为100%，这意味着剩余价值等于可变资本。
3） 计算可变资本：
可变资本 = 剩余价值
由于购买劳动力支付了40万元，这部分是可变资本，因为劳动力是生产过程中可以产生剩余价值的要素。
4） 计算剩余价值：
剩余价值 = 可变资本= 40万元
5） 计算总利润：
总利润 = 剩余价值 + 固定资本的增值
由于剩余价值率是100%，这意味着固定资本的增值等于可变资本，即固定资本也增值了40万元。
总利润 = 剩余价值 + 固定资本的增值
= 40万元 + 40万元 = 80万元
6） 计算平均利润率：
平均利润率 = 总利润 / 总投资
= 80万元 / 100万元
= 0.8 或 80%

所以，该企业的平均利润率为80%。

Good Answer:

1.Determine the company's total investment:
Total investment = fixed capital + circulating capital
= 400,000 yuan + 600,000 yuan
= 1,000,000 yuan
2.Determine the surplus value rate:
Surplus value rate = surplus value / variable capital
The surplus value rate is 100%, which means the surplus value is equal to the variable capital.
3.Calculate the variable capital:
Variable capital = surplus value
Since 400,000 yuan was used to pay for labor, this part is the variable capital because labor is a factor in production that can generate surplus value.
4.Calculate the surplus value:
Surplus value = variable capital
= 400,000 yuan
5.Calculate the total profit:
Total profit = surplus value + the appreciation of fixed capital
Since the surplus value rate is 100%, this means that the appreciation of the fixed capital is equal to the variable capital, so the fixed capital also appreciated by 400,000 yuan.
Total profit = surplus value + the appreciation of fixed capital
= 400,000 yuan + 400,000 yuan
= 800,000 yuan
6.Calculate the average profit rate:
Average profit rate = total profit / total investment
= 800,000 yuan / 1,000,000 yuan
= 0.8 or 80%
Therefore, the company's average profit rate is 80%.

GPT-4o输出：为了解决该题目，按照以下步骤进行分析：

1） **确定固定资本和流动资本**：该企业固定资本为40万元，流动资本为60万元。
2） **计算总资本**：总资本等于固定资本和流动资本之和，即：
总资本 = 固定资本 + 流动资本 = 40万元 + 60万元 = 100万元
3） **计算剩余价值**：剩余价值率为100%，意味着企业在生产过程中产生的剩余价值等于其支付的劳动力成本。

根据题目，企业支付了40万元的劳动力，因此：
剩余价值 = 剩余价值率 × 劳动力成本 = 100% × 40万元 = 40万元
4） **计算总产出**：在进行生产后，总产出包括已支付的劳动力成本和剩余价值：
总产出 = 劳动力成本 + 剩余价值 = 40万元 + 40万元 = 80万元
5） **计算平均利润**：平均利润是指剩余价值与总资本的比率。为了计算平均利润率：
平均利润率 = 剩余价值/总资本 × 100%
将剩余价值（40万元）和总资本（100万元）代入公式：
平均利润率 = 40万元/100万元 × 100% = 40%
所以，这里是这道题目的答案，企业的平均利润率为 **40**。

GPT-4o Output: To solve this problem, the following steps are taken:

1.Determine fixed capital and circulating capital:
The company’s fixed capital is 400,000 yuan, and circulating capital is 600,000 yuan.
2.Calculate total capital:
Total capital equals the sum of fixed capital and circulating capital:
Total capital = fixed capital + circulating capital = 400,000 yuan + 600,000 yuan = 1,000,000 yuan
3.Calculate surplus value:
The surplus value rate is 100%, which means the surplus value generated during production is equal to the labor cost. According to the problem, the company paid 400,000 yuan for labor, so:
Surplus value = surplus value rate × labor cost = 100% × 400,000 yuan = 400,000 yuan
4.Calculate total output:
After production, the total output includes the labor cost and surplus value:
Total output = labor cost + surplus value = 400,000 yuan + 400,000 yuan = 800,000 yuan
5.Calculate average profit:
Average profit refers to the ratio of surplus value to total capital. To calculate the average profit rate:
Average profit rate = surplus value / total capital × 100%
Substituting surplus value (400,000 yuan) and total capital (1,000,000 yuan) into the formula:
Average profit rate = 400,000 yuan / 1,000,000 yuan × 100% = 40%
So, the company's average profit rate is 40%.

Figure 16: An example of a logical reasoning error made by a model when solving a financial CoT task.The model made an error in the 5th step by applying the wrong formula to calculate the average profit rate, which led to an incorrect result.

输入：下面是一些追问。你将看到一个过去的问题记录，以及一个关于这个问题的追问。请在原始问题的基础上思考追问问题。你只需要按照追问的指令进行回答即可。
原始问题：你正在为一家投资公司开发一个股票市场监控和预警系统。该系统需要调用以下API：

1. GetStockPrice API（获取股票价格API）：查询特定股票的当前价格。参数包括stock_symbol（股票符号，字符串）。
2. GetHistoricalData API（获取历史数据API）：获取股票的历史交易数据。参数包括stock_symbol（股票符号，字符串），start_date（开始日期，字符串），end_date（结束日期，字符串）。
3. SetAlert API（设置预警API）：当股票达到特定价格或价格变动幅度时发送预警。参数包括stock_symbol（股票符号，字符串），price_threshold（价格阈值，浮点数），change_percentage（变动百分比，浮点数）。
4. AnalyzeTrends API（趋势分析API）：分析股票价格趋势，提供买入或卖出的建议。参数包括stock_symbol（股票符号，字符串）。假设投资公司对苹果公司（AAPL）的股票非常感兴趣，希望在其股价低于150美元或在一天内价格变动超过5%时得到通知。请描述你如何设置和使用这些API来监控AAPL的股票并及时通知投资者。追问：在系统运行一段时间后，投资公司希望对其投资组合中的所有股票进行定期的趋势分析，并希望每周一收到一份包含买入或卖出建议的报告。请说明你将如何利用现有API来实现这一需求。

Input: Here are some follow-up questions. You will see a record of past questions and a follow-up question about this question. Please think about the follow-up questions based on the original question. You only need to answer according to the instructions of the follow-up questions.
Original question: You are developing a stock market monitoring and early warning system for an investment company. The system needs to call the following APIs:

1. GetStockPrice API: Query the current price of a specific stock. Parameters include stock_symbol (stock symbol, string).
2. GetHistoricalData API: Get historical trading data of a stock. Parameters include stock_symbol (stock symbol, string), start_date (start date, string), end_date (end date, string).
3. SetAlert API: Send an alert when the stock reaches a specific price or price change. Parameters include stock_symbol (stock symbol, string), price_threshold (price threshold, floating point number), change_percentage (change percentage, floating point number).
4. AnalyzeTrends API: Analyze stock price trends and provide buy or sell recommendations. Parameters include stock_symbol (stock symbol, string). Assume that an investment company is very interested in Apple (AAPL) stock and wants to be notified when its stock price falls below $150 or when the price changes by more than 5% in a day. Please describe how you set up and use these APIs to monitor AAPL stock and notify investors in a timely manner. Follow-up question: After the system has been running for a period of time, the investment company wants to conduct regular trend analysis on all stocks in its portfolio and receive a report containing buy or sell recommendations every Monday. Please explain how you will use existing APIs to achieve this requirement.

期望答案：为了满足投资公司的需求，即对其投资组合中的所有股票进行定期的趋势分析并每周一提供一份含有买入或卖出建议的报告，你可以按照以下步骤使用现有API：
1. 确定投资组合中的股票列表：
- 首先，确定投资公司投资组合中包含的所有股票符号。
- 这个列表将用于后续的API调用。

2. 定期调用趋势分析API（AnalyzeTrends API）：
- 对投资组合中的每只股票定期（例如每周一次）调用 AnalyzeTrends API。
- 遍历股票列表，对每只股票执行以下操作：
  - 调用 AnalyzeTrends API。
  - 参数 stock_symbol 设置为当前股票的符号。
  - 收集API返回的买入或卖出建议。

3. 生成周报告：
- 根据从 AnalyzeTrends API 获取的数据，编制一个报告。
- 报告应包括每只股票的当前趋势分析和买入或卖出的建议。
- 确保报告格式清晰、易于理解，适合投资者阅读。

4. 自动化和定时执行：
- 使用定时任务（如cron作业）来自动化整个过程。
- 设置任务在每周一自动执行上述步骤。
- 确保在完成报告后能自动发送给投资公司，可以通过电子邮件或集成到他们的投资管理系统中。

5. 考虑数据整合和报告呈现：
- 可能需要开发一个数据整合和报告生成模块，用于处理API返回的数据并生成易于理解的报告。
- 选择合适的报告格式，如PDF或HTML，确保投资者可以方便地查看报告内容。

通过这些步骤，系统将能够自动化地为投资公司提供定期的股票趋势分析和投资建议，帮助他们做出更明智的投资决策。此外，通过定期更新和自动化报告生成，系统不仅提高了效率，还确保了投资决策的及时性。

Good Answer: To meet the investment company's needs for performing regular trend analysis on all stocks in its portfolio and providing a report with buy or sell recommendations every Monday, you can use the existing APIs as follows:
1. Determine the stock list in the portfolio:
- First, identify all stock symbols included in the investment company's portfolio.
- This list will be used for subsequent API calls.

2. Regularly call the trend analysis API (AnalyzeTrends API):
- Perform a regular (e.g., weekly) call to the AnalyzeTrends API for each stock in the portfolio.
- Iterate through the stock list and, for each stock, follow these steps:
  - Call the AnalyzeTrends API.
  - Set the stock_symbol parameter to the symbol of the current stock.
  - Collect the buy or sell recommendation returned by the API.

3. Generate a weekly report:
- Based on the data obtained from the AnalyzeTrends API, compile a report.
- The report should include the current trend analysis and buy or sell recommendation for each stock.
- Ensure that the report format is clear, easy to understand, and suitable for investors.

4. Automation and scheduled execution:
- Use scheduled tasks (such as a cron job) to automate the entire process.
- Schedule the task to automatically execute the above steps every Monday.
- After completing the report, ensure it is automatically sent to the investment company, either by email or through integration with their investment management system.

5. Consider data integration and report presentation:
- It may be necessary to develop a data integration and report generation module to process the data returned by the API and generate a report that is easy to understand.
- Choose an appropriate report format, such as PDF or HTML, to ensure investors can conveniently view the report content.

By following these steps, the system will be able to automatically provide the investment company with regular stock trend analysis and investment recommendations, helping them make more informed investment decisions. Furthermore, by regularly updating and automating report generation, the system not only improves efficiency but also ensures the timeliness of investment decisions.

Qwen2.5-72B-Instruct 输出：为了满足投资公司的新需求，我们需要对现有的API进行扩展和整合，以便能够定期对投资组合中的所有股票进行趋势分析，并生成包含买入或卖出建议的周报。以下是实现这一需求的步骤：
1. 扩展API以支持批量操作
- 批量查询股票价格：修改GetStockPrice API，使其能够接收一个股票列表作为输入参数，返回每个股票的当前价格。这样可以一次性查询多个股票的价格信息。
- 批量获取历史数据：同样地，调整GetHistoricalData API，允许通过提供一个包含多个股票符号的列表来获取这些股票的历史交易数据。

2. 设置周期性监控任务
- 利用定时任务或者集成外部调度服务（如cron jobs、Azure Functions等），每天执行以下步骤：
  - 使用批量查询股票价格API获取所有投资组合中股票的实时价格。
  - 对于每只股票，检查其价格是否低于150美元或者一天内的价格变动超过5%。
  - 如果触发了预警条件，则立即通过电子邮件或短信通知投资者。

3. 实现趋势分析和建议生成
- 使用AnalyzeTrends API对每只股票的价格数据进行分析。可以考虑引入技术指标（如移动平均线、相对强弱指数等）来更准确地评估股票趋势。
- 根据分析结果，生成买入或卖出的建议。例如，如果某股票的长期趋势向上且近期价格突破关键阻力位，则建议买入；相反，若趋势向下且价格接近支撑位，则建议卖出。

4. 生成周报
- 每周一，执行以下步骤：
  - 使用AnalyzeTrends API对所有投资组合中的股票进行趋势分析。
  - 基于分析结果，整理出本周内推荐买入和卖出的股票名单。
  - 将分析结果和建议生成为报告，可以是电子表格、PDF文件或电子邮件附件的形式发送给投资者。

5. 集成报告生成和发送功能
- 可以使用Python的库（如pandas、openpyxl、email库等）来处理数据分析和报告生成。
- 利用SMTP服务将报告自动发送给投资者，确保报告在周一早上准时到达。通过以上步骤，我们不仅能够有效地监控特定股票的价格波动并及时通知投资者，还能够为整个投资组合提供趋势分析和投资建议，帮助投资者做出更明智的决策。

Qwen2.5-72B-Instruct Output: To meet the investment company's new requirements, we need to expand and integrate the existing APIs to conduct regular trend analysis on all stocks in the portfolio and generate weekly reports containing buy or sell recommendations. The steps to achieve this are as follows:
1. Extend the APIs to support batch operations
- Batch Stock Price Query: Modify the GetStockPrice API to accept a list of stocks as input, returning the current price of each stock. This allows for querying the prices of multiple stocks at once.
- Batch Historical Data Retrieval: Similarly, adjust the GetHistoricalData API to allow providing a list of stock symbols to obtain the historical trading data for these stocks.

2. Set up periodic monitoring tasks
- Use scheduled tasks or integrate external scheduling services (e.g., cron jobs, Azure Functions) to perform the following steps daily:
  - Use the batch stock price query API to get the real-time prices of all stocks in the portfolio.
  - For each stock, check if its price is below $150 or if the price fluctuation exceeds 5% in a day.
  - If the alert conditions are triggered, immediately notify investors via email or SMS.

3. Implement trend analysis and recommendation generation
- Use the AnalyzeTrends API to analyze the price data of each stock. Consider incorporating technical indicators (e.g., moving averages, Relative Strength Index) for a more accurate assessment of stock trends.
- Based on the analysis results, generate buy or sell recommendations. For example, if a stock's long-term trend is upward and the recent price breaks through a key resistance level, a buy recommendation is given; conversely, if the trend is downward and the price nears a support level, a sell recommendation is made.

4. Generate weekly reports
- Every Monday, perform the following steps:
  - Use the AnalyzeTrends API to analyze the trends of all stocks in the portfolio.
  - Based on the analysis results, compile a list of recommended buy and sell stocks for the week.
  - Generate a report from the analysis results and recommendations, which can be in the form of a spreadsheet, PDF, or email attachment sent to investors.

5. Integrate report generation and delivery features
- Use Python libraries (e.g., pandas, openpyxl, email libraries) for data analysis and report generation.
- Use an SMTP service to automatically send the report to investors, ensuring it arrives by Monday morning.Through these steps, we can efficiently monitor the price fluctuations of specific stocks and promptly notify investors. Additionally, we can provide trend analysis and investment recommendations for the entire portfolio, helping investors make more informed decisions.

Figure 17: An example of a contextual misunderstanding made by a model when solving a API utilize task.The model failed to correctly understand the relationship between the original question and the follow-up question. In its response, it unnecessarily modified the API, invoked the SetAlert API, and set up a monitoring task.

输入：【开展金融行业可持续发展论坛】预计参与环保组织和金融机构200人，预算30万元，时间三个月，讨论金融行业的环境责任和可持续发展策略。
Input: [Financial Industry Sustainable Development Forum] 200 people are expected to participate in environmental protection organizations and financial institutions with a budget of 300,000 yuan for three months to discuss environmental responsibility and sustainable development strategies of the financial industry.

期望答案：开展一个关注金融行业可持续发展的论坛是一个极好的举措，旨在促进环保组织和金融机构之间的对话，共同探索和实施可持续发展策略。以下是一些关键步骤和建议，以帮助您成功规划和执行这次论坛：
1.确定论坛目标和议题
- 明确论坛的主要目标，如提升金融行业对环境责任的认识、分享最佳实践、促进可持续投资和金融产品的发展等。
- 选择与可持续发展相关的议题，如绿色金融、气候变化金融风险管理、可持续投资策略等。

2.规划参与者和演讲嘉宾
- 确定目标参与者，包括环保组织代表、金融机构高管、政策制定者、学者和行业专家等。
- 邀请具有影响力和专业知识的演讲嘉宾，包括可持续发展领域的领军人物、创新者和实践者。

3.论坛格式和安排
- 设计包括主题演讲、小组讨论、案例研究展示、互动研讨会等在内的多样化论坛格式。
- 规划网络活动和茶歇，以促进参与者之间的交流和合作。

4.场地和设施安排
- 选择适合200人参加的会议场地，考虑到地理位置、交通便利性、设施配备等因素。
- 确保场地具备必要的技术支持，如音响系统、投影仪、视频会议设施等。

5.预算规划
- 详细规划30万元预算，包括场地租赁、嘉宾差旅和酬劳、宣传材料、技术支持、餐饮服务等费用。
- 为意外开支预留一定比例的预算。

6.宣传和推广
- 制定宣传计划，包括创建论坛网站、发布新闻稿、利用社交媒体和专业网络平台进行推广等。
- 与行业媒体和组织合作，扩大论坛的影响力和参与度。

7.合作伙伴和支持者
- 寻找与论坛目标相符的合作伙伴，如可持续发展组织、研究机构、行业协会等，共同推广论坛。
- 吸引赞助商，以增加预算和资源，同时提升论坛的品牌影响力。

8.后续行动和影响评估
- 计划论坛结束后的后续行动，如发布会议报告、建立可持续发展行动网络、推动具体合作项目等。
- 设定评估机制，衡量论坛对提升行业可持续发展意识和实践的影响。
- 通过这样的策划和实施，金融行业可持续发展论坛将成为一个促进知识共享、激发合作和推动行业向可持续方向发展的重要平台。

Good Answer: Conducting a forum focused on sustainable development in the financial industry is an excellent initiative to promote dialogue between environmental organizations and financial institutions to jointly explore and implement sustainable development strategies. Here are some key steps and suggestions to help you successfully plan and execute this forum:
1. Determine the forum objectives and topics
- Clarify the main objectives of the forum, such as raising awareness of environmental responsibility in the financial industry, sharing best practices, and promoting the development of sustainable investment and financial products.
- Select topics related to sustainable development, such as green finance, climate change financial risk management, and sustainable investment strategies.

2. Plan participants and speakers
- Determine target participants, including representatives of environmental organizations, senior executives of financial institutions, policymakers, academics, and industry experts.
- Invite speakers with influence and expertise, including leaders, innovators, and practitioners in the field of sustainable development.

3. Forum format and arrangement
- Design a diverse forum format including keynote speeches, panel discussions, case study presentations, interactive workshops, etc.
- Plan networking activities and coffee breaks to promote communication and cooperation among participants.

4. Venue and facility arrangements
- Choose a conference venue suitable for 200 people, taking into account factors such as geographical location, transportation convenience, and facility configuration.
- Ensure that the venue has the necessary technical support, such as sound system, projector, video conferencing facilities, etc.

5. Budget planning
- Detailed planning of a budget of RMB 300,000, including venue rental, guest travel and remuneration, publicity materials, technical support, catering services, etc.
- Reserve a certain percentage of the budget for unexpected expenses.

6. Publicity and promotion
- Develop a publicity plan, including creating a forum website, issuing press releases, and using social media and professional networking platforms for promotion.
- Work with industry media and organizations to expand the influence and participation of the forum.

7. Partners and supporters
- Find partners that are consistent with the forum's goals, such as sustainable development organizations, research institutions, industry associations, etc., to jointly promote the forum.
- Attract sponsors to increase budget and resources, while enhancing the brand influence of the forum.

8. Follow-up actions and impact evaluation
- Plan follow-up actions after the forum, such as publishing a conference report, establishing a sustainable development action network, and promoting specific cooperation projects.
- Set up an evaluation mechanism to measure the impact of the forum on improving industry awareness and practices of sustainable development. Through such planning and implementation, the Financial Industry Sustainable Development Forum will become an important platform to promote knowledge sharing, stimulate cooperation and drive the industry towards sustainable development.

GPT-4o输出：开展金融行业可持续发展论坛是一个重要且具有挑战性的任务。以下是对该任务的分解与规划：
一、任务分解
1.需求分析
- 目标参会人员：环保组织、金融机构代表。
- 主要议题：金融行业的环境责任、可持续发展策略。

2.预算管理
- 总预算：30万元。
- 预算细分：
  - 场地租赁费
  - 设备租赁（音响、投影等）
  - 宣传费用（宣传材料、广告等）
  - 嘉宾邀请与交通费用
  - 餐饮与茶歇费用
  - 其他（资料打印、纪念品等）

3.活动规划
- **时间安排**
  - 确定论坛日期。
  - 制定整体时间表，包括筹备、活动当天的节目安排等。
- **地点选择**
  - 选择合适的会议场地（交通便利、容量足够）。
- **主要议程**
  - 开幕词（相关领导或专家）
  - 主题演讲（邀请行业领袖和专家）
  - 主题讨论（小组讨论环节）
  - 反馈环节（收集参与者意见和建议）

4.嘉宾邀请
- 确定潜在嘉宾（环保组织领导、金融机构高管、学术领域专家）。
- 发出邀请函，确认嘉宾出席。

5.宣传与邀请
- 制定宣传计划，确定目标受众。
- 利用社交媒体、官网、邮件列表等渠道进行推广。
- 制作宣传材料（海报、传单、宣传视频等）。

6.注册与管理
- 建立在线注册平台，便于参与者注册。
- 设定截止日期，方便后续的参会确认与安排。

7.活动准备
- 筹备会议资料（册子、议题简介、发言者资料等）。
- 确定餐饮服务，协调茶歇时间和内容。
- 确保技术设备（音响、投影仪、网络连接等）正常运作。

8.活动实施
- 确定活动当天各项流程，分配工作人员职责。
- 活动现场管理（签到、引导、设备支持等）。

9.活动后续
- 后续调研（向参与者收集反馈）。
- 总结活动（编写活动总结报告，分享会议成果）。
- 媒体宣传（通过新闻稿、社交媒体等传播论坛成果）。

二、时间规划
- 第1周：需求分析及预算细分，召开筹备会议。
- 第2周：确定地点与日期，发出嘉宾邀请。
- 第3周：制定宣传计划，发布活动信息。
- 第4周：启动注册，准备活动资料并进行宣传。
- 第5周：确认嘉宾与参会人员，进行技术准备。
- 第6周：召开内部会议，进行最终协调与确认。
- 第7周：进行场地布置及设备测试。
- 第8周：活动当天实施，确保各环节顺利进行。
- 第9周：收集反馈，撰写活动总结。
- 第10周：发布总结报告，进行后续宣传。

三、风险管理
- **参会人数不足**：提前预估人数，进行多渠道宣传吸引参会。
- **嘉宾缺席**：设置替补嘉宾，确保会议完整性。
- **预算超支**：在每一项支出上设定上限，进行定期监控。

通过以上分解与规划，可以确保金融行业可持续发展论坛的顺利开展，促进环保与金融领域的有效对话。
GPT-4o Output: Conducting the forum on sustainable development in the financial industry is an important and challenging task. The following is the decomposition and planning of the task:
I. Task decomposition
1. Demand analysis
- Target participants: representatives of environmental protection organizations and financial institutions.
- Main topics: environmental responsibility and sustainable development strategies of the financial industry.

2. Budget management
- Total budget: 300,000 yuan.
- Budget breakdown:
  - Venue rental fees
  - Equipment rental (audio, projection, etc.)
  - Promotion costs (promotional materials, advertising, etc.)
  - Guest invitation and transportation costs
  - Catering and coffee break costs
  - Others (printing of materials, souvenirs, etc.)

3. Event planning
- Time arrangement
  - Determine the date of the forum.
  - Develop an overall schedule, including preparations, program arrangements on the day of the event, etc.
- Location selection
  - Choose a suitable conference venue (convenient transportation and sufficient capacity).
- Main agenda
  - Opening speech (relevant leaders or experts)
  - Keynote speech (invite industry leaders and experts)
  - Theme discussion (panel discussion session)
  - Feedback session (collecting opinions and suggestions from participants)

4. Guest invitation
- Determine potential guests (leaders of environmental protection organizations, senior executives of financial institutions, experts in the academic field).
- Send out invitations and confirm the attendance of guests.

5. Promotion and invitation
- Develop a promotion plan and determine the target audience.
- Use social media, official website, mailing list and other channels for promotion.
- Create promotional materials (posters, flyers, promotional videos, etc.).

6. Registration and management
- Establish an online registration platform to facilitate the registration of participants.
- Set a deadline to facilitate the subsequent confirmation and arrangement of participants.

7. Event preparation
- Prepare conference materials (brochures, topic introductions, speaker materials, etc.).
- Determine catering services and coordinate coffee break time and content.
- Ensure the normal operation of technical equipment (audio, projector, network connection, etc.).

8. Event implementation
- Determine the various processes on the day of the event and assign staff responsibilities.
- Event site management (sign-in, guidance, equipment support, etc.).

9. Event follow-up
- Follow-up survey (collect feedback from participants).
- Summarize the event (write an event summary report and share the results of the meeting).
- Media publicity (disseminate the results of the forum through press releases, social media, etc.).

II. Time planning –
- Week 1: Demand analysis and budget breakdown, hold a preparatory meeting.
- Week 2: Confirm the location and date, and send out guest invitations.
- Week 3: Develop a publicity plan and publish event information.
- Week 4: Start registration, prepare event materials and promote.
- Week 5: Confirm guests and participants, and make technical preparations.
- Week 6: Hold an internal meeting for final coordination and confirmation.
- Week 7: Carry out venue layout and equipment testing.
- Week 8: Implement the event on the day to ensure that all links go smoothly.
- Week 9: Collect feedback and write an event summary.
- Week 10: Publish a summary report and conduct follow-up publicity.

III. Risk Management
- Insufficient number of participants: Estimate the number of participants in advance and conduct multi-channel publicity to attract participants.
- Absent guests: Set up substitute guests to ensure the integrity of the meeting.
- Budget overruns: Set an upper limit on each expenditure and conduct regular monitoring. Through the above decomposition and planning, the smooth development of the Financial Industry Sustainable Development Forum can be ensured, and effective dialogue between environmental protection and finance can be promoted.

Figure 18: The model failed to grasp the key points in the open-ended instructions provided in the question. The solution offered by the model lacked focus on the number of attendees and the meeting topic, instead outputting overly broad and less relevant content such as scheduling and risk management, which had limited practical value.

.

Table 4: The FinEval dataset provides specific subdivisions in every category.

Component	Category	Subject	#Questions
Financial Academic Knowledge	Finance	Finance (金融学)	159
		Insurance (保险学)	113
		Investments (投资学)	145
		Central Banking (中央银行学)	119
		Financial Markets (金融市场学)	142
		Monetary Finance (货币金融学)	160
		Corporate Finance (公司金融学)	138
		International Finance (国际金融学)	88
		Financial Engineering (金融工程学)	105
		Commercial Bank Finance (商业银行金融学)	96
	Economy	Macroeconomics (宏观经济学)	137
		Microeconomics (微观经济学)	136
		Econometrics (计量经济学)	83
		Statistics (统计学)	140
		Political Economy (政治经济学)	104
		International Economics (国际经济学)	135
		Public Finance (财政学)	139
	Accounting	Accounting (会计学)	120
		Auditing (审计学)	137
		Financial Management (财务管理学)	130
		Cost Accounting (成本会计学)	148
		Economic Law (经济法)	96
		Tax Law (税法)	143
		Advanced Financial Accounting (高级财务会计)	77
		Intermediate Financial Accounting (中级财务会计)	112
		Management Accounting (管理会计学)	83
		Corporate Strategy and Risk Management (公司战略与风险管理)	134
	Certificate	China Actuary (中国精算师)	144
		Certified Practising Accountant (注册会计师)	140
		Certified Management Accountant (管理会计师)	124
		Fund Qualification Certificate (基金从业资格证)	252
		Futures Practitioner Qualification Certificate (期货从业资格证)	153
		Banking Practitioner Qualification Certificate (银行从业资格证)	420
		Securities Practitioner Qualification Certificate (证券从业资格证)	109
	All		4661
Financial Industry Knowledge	Investment Research	Financial Sentiment Analysis (金融情感分析)	205
		Financial Text Classification (金融文本分类)	174
		Financial Text Summarization (金融文本摘要)	299
	Investment Advisor	Financial Client Portrait (金融客户画像)	120
		Marketing Script Recommendations (营销话术推荐)	150
		Financial Investment Advice (投资建议)	124
	Financial Operations	Financial Event Extraction (金融事件抽取)	85
		Causal Event Extraction (因果事件抽取)	83
		Relationship Extraction (关联关系抽取)	103
		Negative Entity Extraction (负面实体抽取)	91
	All		1434
Financial Security Knowledge	Software and Applications	Appsafe (应用程序安全)	100
		Sftwrsafe(软件安全)	100
		Memsafe (记忆安全)	100
	Network and System Protection	Netwrksafe (网络安全)	102
		Syssafe (系统安全)	99
		Websafe (网页安全)	306
	Security Analysis	Crypsafe (密码安全)	100
		Malware (恶意软件分析)	101
	Vulnerability Protection	Pentest (渗透测试)	432
		Reveng (逆向工程)	100
		Vulnrb (漏洞识别)	100
	All		1640
Finance Agent	Reasoning and Planning	FinRAG (检索增强)	100
		FinCoT (思维链)	100
	Long-term Memory	FinTASK(任务分解)	100
		FinDiag (多轮对话)	88
		FinDoc (文档回答)	100
	Tool Application	APIUtil (API调用)	68
		APIFind (API检索)	60
	All		616
FinEval	All		8351

Table 5: Average five-shot scores across three evaluated categories and one final weighted average: FAK (Financial Academic Knowledge), FIK (Financial Industry Knowledge), FSK (Financial Security Knowledge), and WA (Weighted Average)

Model	Financial Academic	Financial Industry	Financial Security	Weighted Average
Claude 3.5-Sonnet	**74.4**	**61.9**	79.8	**73.2**
GPT-4o	72.1	59.3	**81.8**	71.8
Qwen2.5-72B-Instruct	71.0	56.9	80.2	70.3
Gemini1.5-Pro	68.2	57.6	76.8	68.1
GPT-4o-mini	61.9	60.7	81.2	65.8
Gemini1.5-Flash	61.2	60.1	79.1	64.8
Qwen2.5-7B-Instruct	64.8	54.9	71.4	64.4
Yi1.5-34B-Chat	61.5	55.9	73.1	62.9
InternLM2.5-20B-Chat	60.7	47.0	75.1	61.2
XuanYuan3-70B-Chat	56.8	55.4	72.1	59.8
InternLM2-20B-Chat	55.6	48.9	71.3	57.7
XuanYuan2-70B-Chat	53.5	55.8	70.1	57.4
GLM4-9B-Chat	54.5	56.4	59.6	55.9
CFGPT2-7B	53.6	41.2	68.0	54.4
Yi1.5-9B-Chat	56.8	43.8	46.1	52.1
Baichuan2-13B-Chat	42.7	54.7	48.1	46.1
ChatGLM3-6B	37.3	52.9	51.2	43.1
DISC-FinLLM	34.7	46.7	26.2	35.1
FinGPTv3.1	24.4	30.9	23.5	25.4

Table 6: Average zero-shot CoT scores across three evaluated categories and one final weighted average: FAK (Financial Academic Knowledge), FIK (Financial Industry Knowledge), FSK (Financial Security Knowledge), and WA (Weighted Average)

Model	Financial Academic	Financial Industry	Financial Security	Weighted Average
Claude 3.5-Sonnet	74.4	59.6	78.1	**72.4**
GPT-4o	72.1	**62.3**	**81.8**	72.3
Qwen2.5-72B-Instruct	**76.5**	57.2	72.8	72.1
Gemini1.5-Pro	68.2	59.7	77.4	68.6
InternLM2.5-20B-Chat	67.3	57.5	73.1	66.7
GPT-4o-mini	61.9	61.8	79.5	65.6
GLM4-9B-Chat	66.2	58.1	70.4	65.6
Gemini1.5-Flash	61.2	57.7	77.1	63.9
Yi1.5-34B-Chat	65.0	54.4	63.0	62.6
Qwen2.5-7B-Instruct	69.7	52.9	50.4	62.5
Yi1.5-9B-Chat	61.9	57.4	65.7	61.9
InternLM2-20B-Chat	60.3	54.3	63.3	59.8
XuanYuan3-70B-Chat	57.2	57.0	66.0	59.0
XuanYuan2-70B-Chat	57.2	54.0	63.4	57.9
CFGPT2-7B	57.7	48.3	63.3	57.2
Baichuan2-13B-Chat	48.1	53.6	57.6	51.1
ChatGLM3-6B	44.6	52.5	51.2	47.5
DISC-FinLLM	45.0	40.9	35.0	42.1
FinGPTv3.1	29.3	39.5	32.6	31.9

Table 7: Average five-shot CoT scores across three evaluated categories and one final weighted average: FAK (Financial Academic Knowledge), FIK (Financial Industry Knowledge), FSK (Financial Security Knowledge), and WA (Weighted Average)

Model	Financial Academic	Financial Industry	Financial Security	Weighted Average
Qwen2.5-72B-Instruct	**75.7**	**61.3**	78.5	**73.6**
GPT-4o	73.5	59.2	**81.2**	72.5
Claude 3.5-Sonnet	73.7	59.9	77.4	71.9
Gemini1.5-Flash	60.0	58.8	78.8	63.8
XuanYuan3-70B-Chat	65.6	57.1	63.3	63.5
GLM4-9B-Chat	62.4	61.0	67.7	63.3
Gemini1.5-Pro	59.7	56.1	75.8	62.4
GPT-4o-mini	57.8	59.0	78.5	62.4
Yi1.5-34B-Chat	63.8	57.2	63.0	62.4
XuanYuan2-70B-Chat	63.7	55.4	61.7	61.7
Qwen2.5-7B-Instruct	69.6	55.5	44.0	61.6
InternLM2.5-20B-Chat	60.0	60.3	64.3	61.0
Yi1.5-9B-Chat	60.6	59.7	59.6	60.2
InternLM2-20B-Chat	60.0	55.7	60.9	59.4
CFGPT2-7B	60.8	46.7	58.9	57.8
Baichuan2-13B-Chat	48.4	41.3	47.8	47.0
ChatGLM3-6B	43.6	53.4	47.5	46.2
DISC-FinLLM	42.0	40.2	30.9	39.3
FinGPTv3.1	26.6	34.7	26.9	28.2

Table 8: Evaluation Results (zero-shot) for Finance Academic Knowledge(Average Accuracy(%))

Model	Size	Finance	Economy	Accounting	Certificate	Average
Claude 3.5-sonnet	unknown	**73.7**	**83.1**	**69.6**	**70.9**	**73.9**
GPT-4o	unknown	72.6	78.8	66.3	69.8	71.5
Qwen2.5-72B-Instruct	72B	65.8	71.5	72.4	69.7	69.7
Gemini-1.5-pro	unknown	68.5	75.1	61.1	71.4	68.3
Qwen2.5-7B-Instruct	7B	58.2	63.8	64.5	65.6	62.7
GPT-4o-mini	unknown	65.2	67.7	55.2	63.5	62.4
Gemini-1.5-flash	unknown	60.0	67.7	60.4	61.9	62.1
Yi1.5-34B-Chat	34B	54.7	65.2	54.6	63.5	59.5
XuanYuan3-70B-Chat	70B	52.9	57.5	55.4	55.8	55.2
Yi1.5-9B-Chat	9B	51.6	58.1	52.3	60.7	55.0
InternLM2.5-20B-Chat	20B	59.8	61.6	64.9	62.1	54.7
InternLM2-20B-Chat	20B	52.9	55.8	54.6	55.3	54.7
GLM4-9B-Chat	9B	54.1	53.0	56.0	55.3	54.7
CFGPT2-7B	7B	51.6	55.8	56.0	52.2	53.9
XuanYuan2-70B-Chat	70B	53.6	51.1	53.2	52.2	52.8
Baichuan2-13B-Chat	13B	38.8	40.0	40.0	47.3	41.1
DISC-FinLLM	13B	45.6	40.5	32.7	37.3	39.1
ChatGLM3-6B	6B	42.8	36.5	35.9	40.0	38.9
FinGPTv3.1	6B	24.6	23.5	23.2	29.1	25.3

Table 9: Evaluation Results (five-shot) for Finance Academic Knowledge(Average Accuracy(%))

Model	Size	Finance	Economy	Accounting	Certificate	Average
Claude 3.5-sonnet	unknown	**73.3**	**83.6**	**70.7**	**72.0**	**74.4**
GPT-4o	unknown	**73.3**	79.3	66.6	70.9	72.1
Qwen2.5-72B-Instruct	72B	68.5	71.9	71.6	72.8	71.0
Gemini-1.5-pro	unknown	69.3	74.1	61.1	70.9	68.2
Qwen2.5-7B-Instruct	7B	59.9	64.4	66.6	67.2	61.9
GPT-4o-mini	unknown	64.4	66.6	52.6	61.6	61.9
Yi1.5-34B-Chat	34B	59.9	62.2	61.4	59.8	61.2
Gemini-1.5-flash	unknown	61.5	66.7	58.1	59.8	61.2
InternLM2.5-20B-Chat	20B	58.9	60.7	63.4	60.0	60.7
Yi1.5-9B-Chat	9B	54.2	57.4	56.7	58.7	56.8
XuanYuan3-70B-Chat	70B	53.5	57.9	56.4	56.7	56.8
InternLM2-20B-Chat	20B	54.2	56.8	56.4	55.5	55.6
GLM4-9B-Chat	9B	54.8	52.8	56.4	53.2	54.3
CFGPT2-7B	7B	50.3	56.4	55.4	52.2	53.5
XuanYuan2-70B-Chat	70B	54.4	51.4	52.9	55.0	53.5
Baichuan2-13B-Chat	13B	40.8	45.9	40.8	46.4	42.7
ChatGLM3-6B	6B	39.8	36.5	35.9	40.0	38.1
DISC-FinLLM	13B	33.5	38.7	31.5	36.4	35.0
FinGPTv3.1	6B	26.9	25.3	21.8	23.5	24.4

Table 10: Evaluation Results (zero-shot CoT) for Finance Academic Knowledge(Average Accuracy(%))

Model	Size	Finance	Economy	Accounting	Certificate	Average
Qwen2.5-72B-Instruct	72B	**74.8**	78.8	**78.5**	**76.5**	**76.5**
Claude 3.5-sonnet	unknown	73.3	**83.6**	70.7	**76.5**	74.4
GPT-4o	unknown	73.3	79.3	66.6	70.9	72.1
Qwen2.5-7B-Instruct	7B	62.6	62.7	75.9	73.5	69.7
Gemini-1.5-pro	unknown	69.3	74.1	61.1	70.9	68.9
InternLM2.5-20B-Chat	20B	67.0	69.3	68.5	64.0	67.3
GLM4-9B-Chat	9B	63.0	68.3	66.7	68.3	66.6
Yi1.5-34B-Chat	34B	61.3	63.4	63.4	66.7	65.0
Yi1.5-9B-Chat	9B	59.6	62.1	63.6	65.6	62.7
GPT-4o-mini	unknown	64.4	66.6	52.6	66.5	61.9
Gemini-1.5-flash	unknown	61.5	66.7	58.1	59.8	61.2
InternLM2-20B-Chat	20B	60.7	63.3	56.7	60.8	60.3
CFGPT2-7B	7B	60.4	52.4	57.9	55.3	56.5
XuanYuan3-70B-Chat	70B	61.2	63.8	49.5	54.6	57.2
XuanYuan2-70B-Chat	70B	60.7	71.4	52.2	55.0	58.3
Baichuan2-13B-Chat	13B	50.7	47.1	45.6	47.4	47.7
DISC-FinLLM	13B	49.3	45.9	41.4	46.4	45.7
ChatGLM3-6B	6B	46.3	43.4	45.6	41.8	44.3
FinGPTv3.1	6B	31.3	28.4	30.6	26.8	29.3

Table 11: Evaluation Results (five-shot CoT) for Finance Academic Knowledge(Average Accuracy(%))

Model	Size	Finance	Economy	Accounting	Certificate	Average
Qwen2.5-72B-Instruct	72B	**74.8**	75.7	73.7	**79.9**	**75.7**
Claude 3.5-sonnet	unknown	74.4	**80.0**	**74.9**	69.2	73.7
GPT-4o	unknown	71.9	78.3	71.1	74.6	73.5
Qwen2.5-7B-Instruct	7B	71.1	67.7	65.6	74.1	69.6
XuanYuan3-70B-Chat	70B	65.6	67.7	65.2	64.4	65.7
Yi1.5-34B-Chat	34B	64.2	65.5	64.3	62.1	64.0
XuanYuan2-70B-Chat	70B	63.3	65.2	62.2	61.3	63.7
GLM4-9B-Chat	9B	61.5	64.6	63.3	63.3	63.2
CFGPT2-7B	7B	63.3	60.4	63.1	59.5	61.6
Yi1.5-9B-Chat	9B	56.3	57.3	60.3	58.9	58.2
InternLM2.5-20B-Chat	20B	58.1	57.1	61.4	59.6	59.0
InternLM2-20B-Chat	20B	58.1	60.3	61.9	59.8	60.0
Gemini-1.5-flash	unknown	61.5	67.2	54.8	59.3	59.7
Gemini-1.5-pro	unknown	59.6	67.2	54.8	59.3	59.7
GPT-4o-mini	unknown	60.3	67.2	51.9	53.4	57.8
Baichuan2-13B-Chat	13B	48.1	48.1	48.1	48.4	48.4
ChatGLM3-6B	6B	43.1	49.2	38.9	45.5	44.2
DISC-FinLLM	13B	43.0	41.3	38.9	45.5	42.2
FinGPTv3.1	6B	26.1	31.1	21.9	27.4	26.6

Table 12: Evaluation Results (zero-shot) for Finance Industry Knowledge(Average Similarity(%)). These are objective short-answer question including FTC: Financial Text Classification, FSA: Financial Sentiment Analysis, RE: Relation Extraction, FEE: Financial Event Extraction, NEE: Negative Entity Extraction, CEE: Causal Event Extraction

Model	**Size**	**FTC**	**FSA**	**RE**	**FEE**	**NEE**	**CEE**	**Average**
GPT-4o	unknown	53.2	93.3	83.3	78.2	91.1	**69.3**	**78.1**
Gemini-1.5-flash	unknown	53.6	91.6	**86.7**	**80.1**	85.6	65.0	77.1
GPT-4o-mini	unknown	51.8	92.5	80.0	77.3	88.9	69.1	76.6
Gemini-1.5-pro	unknown	53.4	91.7	80.0	75.6	88.9	64.5	75.7
Claude 3.5-sonnet	unknown	55.4	86.7	80	76.3	87.8	65.1	75.2
GLM4-9B-Chat	9B	37.1	87.8	76.7	34.5	86.7	65.6	64.7
Qwen2.5-72B-Instruct	72B	37.9	86.7	80.0	32.3	84.4	65.6	64.5
Baichuan2-13B-Chat	13B	23.9	**94.5**	70.0	55.2	**95.6**	42.8	63.7
InternLM2.5-20B-Chat	20B	43.7	92.5	76.7	21.0	84.5	62.6	63.5
XuanYuan3-70B-Chat	70B	39.3	90.0	76.7	34.9	95.6	40.7	62.9
InternLM2-20B-Chat	20B	37.1	87.8	76.7	34.5	92.2	30.4	59.8
CFGPT2-7B	7B	**55.6**	67.5	70.0	38.1	93.3	30.4	59.2
Yi1.5-34B-Chat	34B	35.9	93.3	73.2	34.5	86.7	22.3	57.7
ChatGLM3-6B	6B	35.9	67.5	76.7	31.4	86.7	42.8	56.8
DISC-FinLLM	13B	18.1	90.0	66.7	59.1	70.0	36.6	56.7
Qwen2.5-7B-Instruct	7B	37.1	87.8	76.7	31.4	90.0	15.0	56.3
XuanYuan2-70B-Chat	70B	33.6	66.6	83.3	31.2	78.9	25.4	53.2
Yi1.5-9B-Chat	9B	34.6	56.7	76.7	27.2	92.2	20.2	51.3
FinGPTv3.1	6B	24.6	56.6	63.2	21.2	74.4	10.2	41.7

Table 13: Evaluation Results (five-shot) for Finance Industry Knowledge (Average Similarity(%)). These are objective short-answer questions including FTC: Financial Text Classification, FSA: Financial Sentiment Analysis, RE: Relation Extraction, FEE: Financial Event Extraction, NEE: Negative Entity Extraction, CEE: Causal Event Extraction.

Model	**Size**	**FTC**	**FSA**	**RE**	**FEE**	**NEE**	**CEE**	**Average**
GPT-4o-mini	unknown	**55.7**	90.0	72.2	78.5	90.0	**69.7**	**76.0**
Claude 3.5-sonnet	unknown	54.5	81.1	81.1	**79.8**	92.2	65.7	75.7
Gemini-1.5-flash	unknown	54.1	**92.2**	72.2	78.3	88.9	66.2	75.3
GPT-4o	unknown	54.1	84.4	72.2	79.6	90.0	68.5	74.8
Gemini-1.5-pro	unknown	47.0	86.7	67.8	75.0	83.3	65.1	70.8
Qwen2.5-72B-Instruct	72B	49.5	80.3	80.0	64.3	90.0	46.8	68.5
GLM4-9B-Chat	9B	37.1	76.7	77.8	66.1	**96.7**	54.5	68.2
Yi1.5-34B-Chat	34B	42.2	86.7	82.1	66.1	81.1	44.8	67.2
XuanYuan2-70B-Chat	70B	32.5	88.3	83.4	55.9	85.6	54.5	66.7
Baichuan2-13B-Chat	13B	34.5	76.7	78.3	75.3	83.3	50.8	66.5
Qwen2.5-7B-Instruct	7B	33.8	80.0	77.8	53.5	92.2	52.0	64.9
XuanYuan3-70B-Chat	70B	35.5	86.7	86.2	37.7	85.8	56.8	64.8
InternLM2-20B-Chat	20B	33.8	76.7	80.0	37.7	81.1	45.0	64.8
ChatGLM3-6B	6B	30.8	73.7	77.0	66.1	81.1	54.5	63.9
DISC-FinLLM	13B	25.0	61.1	53.3	53.9	96.7	50.8	56.6
InternLM2.5-20B-Chat	20B	32.5	62.2	83.4	47.7	48.4	60.7	55.8
Yi1.5-9B-Chat	9B	24.5	64.4	**89.1**	35.4	60.0	29.9	50.6
CFGPT2-7B	7B	53.5	38.9	67.8	26.4	66.7	45.0	49.7
FinGPTv3.1	6B	16.5	30.9	59.8	18.4	52.0	21.9	33.3

Table 14: Evaluation Results (zero-shot) for Finance Industry Knowledge (Average Similarity(%)). These are subjective open-ended question including FTS: Financial Text Summarization, FCP: Financial Customer Portrait, MSR: Marketing Scripts Recommendation, IA: Investment Advice.

Model	Size	FTS	FCP	MSR	IA	Average
Qwen2.5-72B-Instruct	72B	31.0	80.0	22.2	**24.0**	**39.3**
Claude 3.5-sonnet	unknown	25.9	**83.3**	22.2	23.9	38.8
Gemini-1.5-pro	unknown	29.5	76.7	22.0	23.0	37.8
GPT-4o-mini	unknown	28.6	76.7	22.0	23.7	37.8
InternLM2.5-20B-Chat	20B	27.9	76.7	22.1	23.9	37.7
Yi1.5-34B-Chat	34B	28.8	76.7	**22.7**	21.5	37.4
Gemini-1.5-flash	unknown	30.1	73.3	22.2	23.4	37.3
XuanYuan2-70B-Chat	70B	28.5	73.3	22.2	23.3	36.8
CFGPT2-7B	7B	**34.8**	66.7	22.0	23.6	36.8
Qwen2.5-7B-Instruct	7B	26.4	73.3	22.1	23.9	36.4
ChatGLM3-6B	6B	29.0	71.3	22.1	23.3	36.4
InternLM2-20B-Chat	20B	28.8	70.0	22.1	23.5	36.1
GPT-4o	unknown	28.4	70.0	22.0	23.7	36.0
GLM4-9B-Chat	9B	31.1	66.7	21.7	23.6	35.8
XuanYuan3-70B-Chat	70B	27.1	70.0	22.1	23.5	35.7
Yi1.5-9B-Chat	9B	32.6	60.0	22.6	23.4	34.7
DISC-FinLLM	13B	24.5	50.0	22.4	23.9	30.2
Baichuan2-13B-Chat	13B	27.2	46.7	22.0	23.8	29.9
FinGPTv3.1	6B	22.5	43.2	22.0	22.5	27.6

Table 15: Evaluation Results (five-shot) for Finance Industry Knowledge (Average Similarity(%)). These are subjective open-ended question including FTS: Financial Text Summarization, FCP: Financial Customer Portrait, MSR: Marketing Scripts Recommendation, IA: Investment Advice.

Model	Size	FTS	FCP	MSR	IA	Average
XuanYuan3-70B-Chat	70B	35.2	83.3	22.4	23.8	**41.2**
Claude 3.5-sonnet	unknown	35.0	83.3	22.2	23.7	41.1
Qwen2.5-7B-Instruct	7B	26.0	**86.7**	**22.5**	**24.0**	39.8
Qwen2.5-72B-Instruct	72B	31.6	80.0	**22.5**	**24.0**	39.5
XuanYuan2-70B-Chat	70B	31.6	80.0	22.4	23.8	39.5
Yi1.5-34B-Chat	34B	**36.3**	73.3	22.4	23.8	39.0
GLM4-9B-Chat	9B	35.0	73.3	22.4	23.8	38.6
Cemini-1.5-pro	unknown	29.5	76.7	22.0	23.0	37.8
GPT-4o-mini	unknown	28.6	76.7	22.0	23.7	37.8
Gemini-1.5-flash	unknown	30.1	73.3	22.2	23.4	37.3
Baichuan2-13B-Chat	13B	35.9	66.7	22.2	23.5	37.1
ChatGLM3-6B	6B	25.8	73.3	22.3	23.8	36.3
GPT-4o	unknown	28.4	70.0	22.0	23.7	36.0
InternLM2.5-20B-Chat	20B	31.8	56.7	22.3	**24.0**	33.7
InternLM2-20B-Chat	20B	31.6	56.7	22.4	23.8	33.6
Yi1.5-9B-Chat	9B	11.1	76.7	22.4	23.8	33.5
DISC-FinLLM	13B	33.3	46.7	**22.5**	23.8	31.6
CFGPT2-7B	7B	30.5	36.7	22.3	**24.0**	28.4
FinGPTv3.1	6B	15.6	48.5	22.0	22.5	27.2

Table 16: Evaluation Results (zero-shot CoT) for Finance Industry Knowledge (Average Similarity(%)). These are objective short-answer questions including FTC: Financial Text Classification, FSA: Financial Sentiment Analysis, RE: Relation Extraction, FEE: Financial Event Extraction, NEE: Negative Entity Extraction, CEE: Causal Event Extraction.

Model	**Size**	**FTC**	**RE**	**FEE**	**NEE**	**CEE**	**Average**
GPT-4o	unknown	**53.2**	**86.7**	76.4	88.4	**68.0**	**74.5**
GPT-4o-mini	unknown	52.4	83.3	74.5	90.0	62.0	72.4
Gemini-1.5-pro	unknown	48.6	80.0	78.1	86.7	57.1	70.1
Claude 3.5-sonnet	unknown	51.7	80.0	72.2	90.0	54.8	69.7
GLM4-9B-Chat	9B	42.0	**86.7**	58.6	96.7	57.9	68.4
Gemini-1.5-flash	unknown	52.3	83.3	**78.3**	81.7	45.7	68.3
XuanYuan3-70B-Chat	70B	31.7	80.0	67.8	96.7	58.7	67.0
Yi1.5-9B-Chat	9B	40.4	66.7	71.5	**100.0**	55.3	66.8
Baichuan2-13B-Chat	13B	22.2	80.0	74.5	96.7	59.2	66.5
Yi1.5-34B-Chat	34B	37.0	73.3	66.4	96.7	54.6	65.6
InternLM2.5-20B-Chat	20B	45.0	80.0	53.6	93.3	53.5	65.1
Qwen2.5-72B-Instruct	72B	40.4	76.7	56.8	90.0	58.7	64.5
XuanYuan2-70B-Chat	70B	26.7	76.7	66.4	90.0	54.6	62.9
InternLM2-20B-Chat	20B	37.0	73.3	58.6	90.0	54.6	62.7
ChatGLM3-6B	6B	35.0	71.3	56.6	88.0	52.6	60.7
Qwen2.5-7B-Instruct	7B	26.7	73.3	57.9	90.0	51.5	59.9
CFGPT2-7B	7B	51.5	73.3	37.4	86.7	30.2	55.8
FinGPTv3.1	6B	14.7	59.3	41.6	76.0	39.5	46.2
DISC-FinLLM	13B	26.7	63.3	37.4	76.7	19.9	44.8

Table 17: Evaluation Results (five-shot CoT) for Finance Industry Knowledge (Average Similarity(%)). These are objective short-answer questions including FTC: Financial Text Classification, FSA: Financial Sentiment Analysis, RE: Relation Extraction, FEE: Financial Event Extraction, NEE: Negative Entity Extraction, CEE: Causal Event Extraction.

Model	**Size**	**FTC**	**RE**	**FEE**	**NEE**	**CEE**	**Average**
GLM4-9B-Chat	9B	**61.2**	**83.3**	60.1	**96.7**	65.4	**73.3**
Qwen2.5-72B-Instruct	72B	46.3	**83.3**	80.1	86.7	62.9	71.9
InternLM2.5-20B-Chat	20B	40.0	80.0	79.8	86.7	**66.4**	70.6
Gemini-1.5-flash	unknown	52.4	70.0	75.5	89.6	65.6	70.6
GPT-4o-mini	unknown	52.9	73.3	76.3	88.3	60.2	70.2
GPT-4o	unknown	56.7	66.7	79.0	87.4	60.9	70.1
Yi1.5-9B-Chat	9B	50.0	**83.3**	76.8	73.3	62.6	69.2
Claude 3.5-sonnet	unknown	47.7	76.7	71.0	87.4	62.9	69.1
Yi1.5-34B-Chat	34B	47.4	73.3	**83.1**	73.3	61.9	67.8
XuanYuan3-70B-Chat	70B	36.4	**83.3**	78.2	86.7	50.3	67.0
InternLM2-20B-Chat	20B	36.4	80.0	76.8	80.0	58.7	66.4
Qwen2.5-7B-Instruct	7B	28.9	80.0	73.2	**96.7**	51.3	66.0
XuanYuan2-70B-Chat	70B	36.4	76.7	78.2	73.3	58.7	64.7
gemini-1.5-pro	unknown	43.8	66.7	76.0	78.0	57.1	64.3
ChatGLM3-6B	6B	33.4	77.0	73.8	77.0	55.7	63.4
CFGPT2-7B	7B	36.4	70.0	38.9	80.0	38.7	52.8
Baichuan2-13B-Chat	13B	22.2	66.7	26.3	73.3	44.8	46.7
DISC-FinLLM	13B	28.9	70.0	42.3	67.3	17.3	45.2
FinGPTv3.1	6B	16.2	60.7	20.3	67.3	32.7	39.4

Table 18: Evaluation Results (zero-shot CoT) for Finance Industry Knowledge (Average Similarity(%)). These are subjective open-ended question including FTS: Financial Text Summarization, FCP: Financial Customer Portrait, IA: Investment Advice.

Model	**Size**	**FTS**	**FCP**	**IA**	**Average**
Qwen2.5-72B-Instruct	72B	29.0	**83.3**	22.8	**45.0**
InternLM2.5-20B-Chat	20B	28.2	**83.3**	23.2	44.9
GPT-4o-mini	unknown	30.4	80.0	22.3	44.2
Claude 3.5-sonnet	unknown	29.4	76.7	22.2	42.8
Gemini-1.5-pro	unknown	31.7	73.3	22.2	42.4
GPT-4o	unknown	29.3	73.3	23.0	41.9
Yi1.5-9B-Chat	9B	29.0	73.3	22.5	41.6
Qwen2.5-7B-Instruct	7B	24.3	76.7	23.0	41.3
GLM4-9B-Chat	9B	30.7	70.0	22.2	41.0
InternLM2-20B-Chat	20B	28.7	70.0	22.5	40.4
XuanYuan3-70B-Chat	70B	28.2	70.0	22.4	40.2
Gemini-1.5-flash	unknown	31.4	66.7	21.8	40.0
XuanYuan2-70B-Chat	70B	24.3	70.0	23.0	39.1
ChatGLM3-6B	6B	24.2	70.0	22.1	38.8
CFGPT2-7B	7B	**34.7**	50.0	23.1	35.9
Yi1.5-34B-Chat	34B	28.7	56.7	22.1	35.8
DISC-FinLLM	13B	26.9	53.3	**23.3**	34.5
Baichuan2-13B-Chat	13B	20.1	53.3	23.2	32.2
FinGPTv3.1	6B	18.2	44.2	22.2	28.2

Table 19: Evaluation Results (five-shot CoT) for Finance Industry Knowledge (Average Similarity(%)). These are subjective open-ended question including FTS: Financial Text Summarization, FCP: Financial Customer Portrait, IA: Investment Advice.

Model	**Size**	**FTS**	**FCP**	**IA**	**Average**
Claude 3.5-sonnet	unknown	28.3	**83.3**	21.8	**44.5**
Yi1.5-9B-Chat	9B	**35.8**	73.3	22.2	43.8
Qwen2.5-72B-Instruct	72B	34.2	73.3	23.0	43.5
InternLM2.5-20B-Chat	20B	29.8	76.7	22.8	43.1
Gemini-1.5-pro	unknown	24.4	80.0	22.9	42.4
GPT-4o	unknown	27.8	73.3	22.2	41.1
XuanYuan3-70B-Chat	70B	25.6	73.3	23.0	40.6
GLM4-9B-Chat	9B	35.1	63.3	22.7	40.4
GPT-4o-mini	unknown	27.8	70.0	22.8	40.2
XuanYuan2-70B-Chat	70B	24.3	73.3	22.2	39.9
Yi1.5-34B-Chat	34B	32.2	63.3	22.6	39.4
Gemini-1.5-flash	unknown	28.3	66.7	22.7	39.2
Qwen2.5-7B-Instruct	7B	27.8	63.3	22.6	37.9
InternLM2-20B-Chat	20B	27.8	63.3	22.6	37.9
ChatGLM3-6B	6B	24.3	63.3	22.2	36.6
CFGPT2-7B	7B	33.6	53.3	22.6	36.5
Baichuan2-13B-Chat	13B	30.5	43.3	22.8	32.2
DISC-FinLLM	13B	26.5	45.6	**23.3**	31.8
FinGPTv3.1	6B	22.5	35.9	22.1	26.8

Table 20: Evaluation Results (zero-shot) for Finance Security Knowledge (Average Accuracy(%)). App: Application security, Cryp: Cryptographic protection, MA: Malware analysis, MS: Memory security, NS: Network security, Pent: Pentest, Reve: Reverse engineering, Soft: Software security, Syst: System security, Vul: Vulnerability detection, WS: Web security

Model	Size	App	Cryp	MA	MS	NS	Pent	Reve	Soft	Syst	Vul	WS	Average
GPT-4o	unknown	77.8	70.4	77.8	**92.6**	70.4	**96.3**	85.2	81.5	85.2	**81.5**	81.5	**81.8**
Qwen2.5-72B-Instruct	72B	77.8	**85.2**	**81.5**	81.5	77.8	77.8	81.5	77.8	**92.6**	**81.5**	85.2	81.8
GPT-4o-mini	unknown	74.1	70.4	74.1	85.2	77.8	88.9	77.8	81.5	81.5	77.8	81.5	79.1
Claude 3.5-sonnet	unknown	70.4	66.7	70.4	81.5	**81.5**	85.2	85.2	81.5	81.5	70.4	85.2	78.1
Gemini-1.5-pro	unknown	81.5	55.6	77.8	81.5	70.4	88.9	77.8	**85.2**	85.2	70.4	81.5	77.8
Gemini-1.5-flash	unknown	77.8	59.3	77.8	85.2	**81.5**	88.9	85.2	70.4	81.5	70.4	74.1	77.5
Yi1.5-34B-Chat	34B	70.4	77.8	74.1	74.1	70.4	78.8	78.8	74.1	88.9	74.1	74.1	76.0
XuanYuan3-70B-Chat	70B	77.8	77.8	66.7	66.7	74.1	92.6	70.4	63.0	88.9	70.4	70.4	74.4
InternLM2.5-20B-Chat	20B	**85.2**	77.8	55.6	77.8	59.3	74.1	74.1	63.0	81.5	77.8	**88.9**	74.1
GLM4-9B-Chat	9B	66.7	77.8	55.6	70.4	74.1	77.8	**88.9**	66.7	81.5	66.7	77.8	73.1
InternLM2-20B-Chat	20B	74.1	77.8	63.0	70.4	70.4	77.8	74.1	66.7	81.5	70.4	77.8	73.1
Qwen2.5-7B-Instruct	7B	74.1	81.5	63.0	74.1	63.0	74.1	66.7	70.4	74.1	66.7	81.5	71.7
Yi1.5-9B-Chat	9B	63.0	77.8	70.4	70.4	74.1	66.7	85.2	66.7	77.8	55.6	77.8	71.4
XuanYuan2-70B-Chat	70B	74.1	74.1	55.6	44.4	59.3	66.7	74.1	59.3	88.9	74.1	77.8	68.0
CFGPT2-7B	7B	66.7	70.4	59.2	59.2	66.7	77.8	60.4	55.6	70.4	59.3	70.4	65.1
Baichuan2-13B-Chat	13B	66.7	66.7	40.7	51.9	63.0	63.0	70.4	55.6	70.4	63.0	66.7	61.6
ChatGLM3-6B	6B	44.4	55.6	40.7	29.6	55.6	51.9	59.3	33.3	59.3	40.7	59.3	48.2
DISC-FinLLM	13B	25.9	22.2	18.5	11.1	18.5	25.9	37.0	37.0	29.6	29.6	22.2	25.2
FinGPTv3.1	6B	24.4	19.5	14.8	7.7	17.1	21.4	36.4	30.9	27.9	29.1	19.8	22.7

Table 21: Evaluation Results (five-shot) for Finance Security Knowledge (Average Accuracy(%)). App: Application security, Cryp: Cryptographic protection, MA: Malware analysis, MS: Memory security, NS: Network security, Pent: Pentest, Reve: Reverse engineering, Soft: Software security, Syst: System security, Vul: Vulnerability detection, WS: Web security

Model	Size	App	Cryp	MA	MS	NS	Pent	Reve	Soft	Syst	Vul	WS	Average
GPT-4o	unknown	74.1	70.4	74.1	**92.6**	81.5	**96.3**	**81.5**	88.9	85.2	74.1	81.5	**81.8**
GPT-4o-mini	unknown	74.1	77.8	77.8	88.9	70.4	88.9	**81.5**	**92.6**	77.8	**81.5**	81.5	81.2
Qwen2.5-72B-Instruct	72B	74.1	**81.5**	**85.2**	74.1	70.4	85.2	**81.5**	77.8	**92.6**	74.1	85.2	80.2
Claude 3.5-sonnet	unknown	66.7	70.4	70.4	88.9	77.8	92.6	**81.5**	85.2	81.5	74.1	**88.9**	79.8
Gemini-1.5-flash	unknown	77.8	55.6	**85.2**	88.9	**85.2**	**96.3**	74.1	77.8	81.5	74.1	74.1	79.1
Gemini-1.5-pro	unknown	81.5	63.0	74.1	81.5	66.7	85.2	**81.5**	88.9	81.5	66.7	74.1	76.8
InternLM2.5-20B-Chat	20B	**85.2**	74.1	70.4	70.4	74.1	70.4	70.1	74.1	81.5	77.8	77.8	75.1
Yi1.5-34B-Chat	34B	59.3	77.8	70.4	74.1	74.1	81.5	77.8	70.4	81.5	63.0	74.1	73.1
XuanYuan3-70B-Chat	70B	74.1	**81.5**	59.3	59.3	74.1	85.2	**81.5**	55.6	88.9	59.3	74.1	72.1
Qwen2.5-7B-Instruct	7B	66.7	66.7	51.9	74.1	63.0	77.8	81.4	70.3	81.4	77.8	74.1	71.4
InternLM2-20B-Chat	20B	70.3	78.2	55.6	66.7	74.1	81.5	79.5	55.6	81.5	66.7	74.1	71.3
XuanYuan2-70B-Chat	70B	65.2	78.2	55.5	57.6	74.1	82.5	79.5	54.6	84.2	65.2	74.1	70.1
CFGPT2-7B	7B	70.3	**81.5**	55.6	66.7	77.8	77.8	59.3	51.9	77.8	66.7	63.0	68.0
GLM4-9B-Chat	9B	70.4	74.1	37.0	63.0	37.0	51.9	66.7	48.2	81.5	70.4	55.6	59.6
ChatGLM3-6B	6B	48.1	59.3	37.0	37.0	55.6	55.6	51.9	51.9	55.6	51.9	59.3	51.2
Baichuan2-13B-Chat	13B	48.1	40.7	48.1	29.6	44.4	59.3	59.3	33.3	55.6	59.3	51.9	48.1
Yi1.5-9B-Chat	9B	51.9	55.6	37.0	40.7	44.4	59.3	33.3	37.0	51.9	51.9	44.4	46.1
DISC-FinLLM	13B	11.1	14.8	25.9	22.2	40.7	29.6	22.2	29.6	25.9	33.3	33.3	26.2
FinGPTv3.1	6B	7.1	12.1	23.0	19.2	35.9	28.7	20.0	27.4	21.6	30.5	32.9	23.5

Table 22: Evaluation Results (zero-shot CoT) for Finance Security Knowledge (Average Accuracy(%)). App: Application security, Cryp: Cryptographic protection, MA: Malware analysis, MS: Memory security, NS: Network security, Pent: Pentest, Reve: Reverse engineering, Soft: Software security, Syst: System security, Vul: Vulnerability detection, WS: Web security

Model	**Size**	**App**	**Cryp**	**MA**	**MS**	**NS**	**Pent**	**Reve**	**Soft**	**Syst**	**Vul**	**WS**	**Average**
GPT-4o	unknown	77.8	70.4	**77.8**	**92.6**	70.4	**96.3**	**85.2**	81.5	85.2	**81.5**	81.5	**81.8**
GPT-4o-mini	unknown	77.8	74.1	74.1	88.9	74.1	88.9	77.8	**85.2**	74.1	77.8	81.5	79.5
Claude 3.5-sonnet	unknown	59.3	77.8	66.7	85.2	77.8	88.9	81.5	81.5	81.5	77.8	81.5	78.1
Gemini-1.5-pro	unknown	**81.5**	55.6	77.8	81.5	70.4	88.9	77.8	85.2	85.2	70.4	77.8	77.4
Gemini-1.5-flash	unknown	81.5	55.6	70.4	85.2	**85.2**	88.9	81.5	70.4	85.2	74.1	70.4	77.1
InternLM2.5-20B-Chat	20B	74.1	**81.5**	55.6	81.5	70.4	74.1	77.8	55.6	81.5	74.1	77.8	73.1
Qwen2.5-72B-Instruct	72B	63.0	77.8	66.7	63.0	74.1	70.4	77.8	63.0	**92.6**	66.7	85.2	72.8
GLM4-9B-Chat	9B	74.1	63.0	63.0	66.7	59.4	74.1	77.8	63.0	70.4	74.1	**88.9**	70.4
XuanYuan3-70B-Chat	70B	77.8	59.3	63.0	51.9	66.7	59.3	81.5	51.9	81.5	59.3	74.1	66.0
Yi1.5-9B-Chat	9B	59.3	77.8	70.4	74.1	55.6	51.9	66.7	59.3	74.1	63.0	70.4	65.7
XuanYuan2-70B-Chat	70B	76.2	59.3	50.2	47.6	65.2	59.3	81.5	64.1	69.3	50.2	74.1	63.4
InternLM2-20B-Chat	20B	66.7	59.3	63.0	51.9	59.4	59.3	77.8	59.3	70.4	59.3	70.4	63.3
CFGPT2-7B	7B	70.4	59.3	51.8	48.1	63.0	77.8	66.7	74.1	66.7	51.9	66.7	63.3
Yi1.5-34B-Chat	34B	66.7	70.4	63.0	59.3	55.6	63.0	66.7	48.1	74.1	63.0	63.0	63.0
Baichuan2-13B-Chat	13B	59.3	63.0	40.7	44.4	48.2	66.7	70.4	40.7	77.8	48.2	74.1	57.6
ChatGLM3-6B	6B	55.6	40.7	51.9	44.4	59.3	44.4	51.9	48.1	62.9	44.4	59.3	51.2
Qwen2.5-7B-Instruct	7B	55.6	55.6	51.9	51.9	33.3	51.9	51.9	51.9	44.4	40.7	65.2	50.4
DISC-FinLLM	13B	37.0	22.2	33.3	18.5	40.7	44.4	51.9	29.6	44.4	29.6	33.3	35.0
FinGPTv3.1	6B	36.0	20.3	31.3	15.7	38.8	39.9	48.5	28.9	41.6	26.4	31.4	32.6

Table 23: Evaluation Results (five-shot CoT) for Finance Security Knowledge (Average Accuracy(%)). App: Application security, Cryp: Cryptographic protection, MA: Malware analysis, MS: Memory security, NS: Network security, Pent: Pentest, Reve: Reverse engineering, Soft: Software security, Syst: System security, Vul: Vulnerability detection, WS: Web security

Model	**Size**	**App**	**Cryp**	**MA**	**MS**	**NS**	**Pent**	**Reve**	**Soft**	**Syst**	**Vul**	**WS**	**Average**
GPT-4o	unknown	74.1	70.4	70.4	**92.6**	**81.5**	**96.3**	81.5	**88.9**	85.2	70.4	81.5	**81.2**
Gemini-1.5-flash	unknown	77.8	59.3	**85.2**	88.9	**81.5**	92.6	77.8	81.5	85.2	66.7	70.4	78.8
GPT-4o-mini	unknown	74.1	74.1	66.7	85.2	74.1	88.9	77.8	**88.9**	74.1	**81.5**	77.8	78.5
Qwen2.5-72B-Instruct	72B	74.1	74.1	77.8	81.5	77.8	66.7	**85.2**	81.5	**92.6**	66.7	**85.2**	78.5
Claude 3.5-sonnet	unknown	66.7	59.3	63.0	85.2	74.1	92.6	**85.2**	**88.9**	81.5	74.1	81.5	77.4
Gemini-1.5-pro	unknown	66.7	59.3	74.1	85.2	70.4	92.6	77.8	81.5	81.5	66.7	77.8	75.8
GLM4-9B-Chat	9B	63.0	63.0	66.7	63.0	59.3	74.1	81.5	63.0	77.8	63.0	70.4	67.7
InternLM2.5-20B-Chat	20B	**85.2**	59.3	66.7	70.4	55.6	51.9	51.9	59.3	70.4	63.0	74.1	64.3
XuanYuan3-70B-Chat	70B	77.8	51.9	70.4	55.6	63.0	55.6	81.5	51.9	70.4	55.6	63.0	63.3
Yi1.5-34B-Chat	34B	66.7	59.3	59.3	63.0	66.7	44.4	59.3	66.7	74.1	55.6	77.8	63.0
XuanYuan2-70B-Chat	70B	76.2	53.4	54.9	52.5	64.2	53.2	81.5	64.1	62.1	53.2	63.0	61.7
InternLM2-20B-Chat	20B	66.7	53.4	59.3	55.6	59.3	55.6	81.5	55.6	66.7	53.2	63.0	60.9
Yi1.5-9B-Chat	9B	55.6	**77.8**	59.3	55.6	55.6	44.4	51.9	55.6	63.0	51.9	85.2	59.6
CFGPT2-7B	7B	70.4	48.1	63.0	55.6	59.3	74.1	55.6	51.9	66.7	44.4	59.3	58.9
Baichuan2-13B-Chat	13B	55.6	51.9	40.7	37.0	48.2	48.2	55.6	37.0	55.6	40.7	55.6	47.8
ChatGLM3-6B	6B	40.7	55.6	51.9	37.0	48.1	55.6	44.4	40.7	55.6	40.7	51.8	47.5
Qwen2.5-7B-Instruct	7B	40.7	48.2	37.0	48.2	22.2	63.0	48.2	37.0	33.3	48.1	58.1	44.0
DISC-FinLLM	13B	33.3	25.9	29.6	29.6	29.6	44.4	25.9	29.6	44.4	18.5	29.6	30.9
FinGPTv3.1	6B	27.1	22.4	22.6	25.4	22.0	38.2	22.4	26.9	43.1	18.4	26.9	26.9

Table 24: Evaluation Result (zero-shot) for Finance Agent(Similarity(%)). COT: Chain of Thought, RAG: Retrieval Augmented Generation, FT: Financial tasks, MC: Multi-turn conversation, MD: Multi-document question and answer, API-I: API invocation, API-R: API retrieval

Model	**Size**	**COT**	**RAG**	**FT**	**MC**	**MD**	**API-I**	**API-R**	**Average**
Claude 3.5-sonnet	unknown	**69.6**	**91.8**	**80.0**	**71.2**	**74.4**	**84.6**	**83.5**	**79.3**
GPT-4o	unknown	68.8	83.4	**80.0**	63.0	69.8	76.5	75.6	73.9
GPT-4o-mini	unknown	63.0	82.6	**80.0**	63.0	68.9	77.4	75.2	72.9
Gemini-1.5-pro	unknown	63.0	82.2	**80.0**	63.0	72.0	73.1	76.3	72.8
Gemini-1.5-flash	unknown	50.4	85.2	**80.0**	62.2	71.1	72.2	75.2	70.9
Qwen2.5-72B-Instruct	72B	62.2	91.1	74.3	31.1	69.8	75.7	74.4	68.4
Qwen2.5-7B-Instruct	7B	44.4	86.7	77.8	63.0	66.3	59.6	68.9	66.7
Yi1.5-34B-Chat	34B	50.2	69.3	79.3	65.2	66.2	61.2	70.4	66.0
XuanYuan3-70B-Chat	70B	48.9	80.7	77.1	31.5	69.3	65.4	74.1	63.9
InternLM2.5-20B-Chat	20B	52.6	85.9	**80.0**	31.5	66.7	56.9	68.1	63.1
XuanYuan2-70B-Chat	70B	44.4	81.5	79.3	31.5	67.8	58.0	69.6	61.7
Yi1.5-9B-Chat	9B	45.2	66.0	69.6	63.7	65.7	50.2	67.0	61.1
InternLM2-20B-Chat	20B	40.2	75.7	72.1	58.0	61.3	54.6	64.6	60.9
GLM4-9B-Chat	9B	48.1	64.4	75.7	31.1	66.5	64.1	71.5	60.2
Baichuan2-13B-Chat	13B	31.9	67.4	**80.0**	31.5	66.1	46.5	66.3	55.7
CFGPT2-7B	7B	28.2	63.0	64.4	63.8	60.6	24.3	52.2	50.9
ChatGLM3-6B	6B	25.2	71.1	70.0	31.5	59.1	33.3	57.2	49.6
DISC-FinLLM	13B	20.7	61.5	62.9	31.1	43.1	21.3	52.0	41.8
FinGPTv3.1	6B	4.6	48.2	42.8	31.1	36.4	25.3	30.0	31.2

Table 25: The Spearman correlation coefficient matrix results between human evaluation and GPT evaluation for Gemini 1.5-Pro in seven financial agent tasks, with 20% of all response results randomly sampled for testing in each task

	1	**2**	**3**	**4**	**5**	**6**	**7**
1	**0.66**	-0.24	-0.08	0.15	0.01	-0.24	0.18
2	-0.07	**0.84**	-0.07	-0.19	0.15	0.24	0.04
3	-0.01	-0.04	**0.68**	0.14	0.14	0.15	0.25
4	0.30	0.02	0.08	**0.85**	-0.27	0.12	0.07
5	-0.13	0.10	0.30	-0.19	**0.85**	0.04	0.31
6	-0.07	0.09	0.00	0.02	-0.70	**0.63**	-0.32
7	-0.08	0.34	0.05	0.15	0.02	0.42	**0.55**

Table 26: Agent Task Evaluation Result (zero-shot) Judged by Claude 3.5-Sonnet

Model	**Size**	**COT**	**RAG**	**FT**	**MC**	**MD**	**API-I**	**API-R**	**Average**
Claude 3.5-sonnet	unknown	77.0	85.2	**81.5**	72.6	**76.7**	**89.6**	72.2	**79.3**
GPT-4o	unknown	**79.2**	86.6	80.7	73.2	70.7	81.5	73.3	77.9
GPT-4o-mini	unknown	71.8	**87.4**	80.7	**74.0**	73.0	76.3	**73.7**	76.7
Gemini-1.5-pro	unknown	71.8	81.4	80.0	72.6	**76.7**	81.1	**73.7**	76.8
Gemini-1.5-flash	unknown	62.2	80.7	80.0	68.8	74.1	79.3	64.8	72.9

Table 27: Agent Task Evaluation Result (zero-shot) Judged by Gemini-1.5-Pro

Model	**Size**	**COT**	**RAG**	**FT**	**MC**	**MD**	**API-I**	**API-R**	**Average**
Claude 3.5-sonnet	unknown	**73.3**	64.4	**80.0**	**92.6**	**86.7**	**92.6**	87.4	82.4
GPT-4o	unknown	**73.3**	**74.8**	79.2	91.8	82.2	90.4	**90.0**	83.1
GPT-4o-mini	unknown	71.0	**74.8**	**80.0**	**92.6**	83.3	89.6	88.9	82.9
Gemini-1.5-pro	unknown	**73.3**	73.3	**80.0**	**92.6**	88.1	89.4	89.3	**83.7**
Gemini-1.5-flash	unknown	60.7	72.6	**80.0**	91.8	82.6	90.2	86.5	80.6

Table 28: Agent Task Score Overall Comparision

Model	**Size**	**GPT-4o**	**Claude-3.5-Sonnet**	**Gemini-1.5-Pro**	**Average**
Claude 3.5-sonnet	unknown	**79.3**	**79.3**	82.4	**80.3**
GPT-4o	unknown	73.9	77.9	83.1	78.3
GPT-4o-mini	unknown	72.9	76.7	82.9	77.5
Gemini-1.5-pro	unknown	72.8	76.8	**83.7**	77.8
Gemini-1.5-flash	unknown	70.9	72.9	80.6	74.8