UCFE: A User-Centric Financial Expertise Benchmark for Large Language Models

Anonymous ACL submission

Abstract

This paper introduces the UCFE: User-Centric Financial Expertise benchmark, an innovative framework designed to evaluate the ability of large language models (LLMs) to handle complex real-world financial tasks. UCFE benchmark adopts a hybrid approach that combines human expert evaluations with dynamic, task-specific interactions to simulate the complexities of evolving financial scenarios. Firstly, we conducted a user study involving 804 participants, collecting their feedback on financial tasks. Secondly, based on this feedback, we created our dataset that encompasses a wide range of user intents and interactions. This dataset serves as the foundation for benchmarking 12 LLM services using the LLM-as-Judge methodology. Our results show a significant alignment between benchmark scores and human preferences, with a Pearson correlation coefficient of 0.78, confirming the effectiveness of the UCFE dataset and our evaluation approach. UCFE benchmark not only reveals the potential of LLMs in the financial sector but also provides a robust framework for assessing their performance and user satisfaction.The benchmark dataset and evaluation code are available [1].

1 Introduction

Recent advances in large language models (LLMs) have expanded their potential applications in finance (Wu et al., 2023; Huang et al., 2023; Kim et al., 2024). Finance professionals are increasingly using LLMs to solve specialized financial tasks (Li et al., 2023a; Xie et al., 2024; Yang et al., 2023; Zhang et al., 2023b). Financial tasks are inherently complex, involving specialized context, financial terminologies, legal intricacies, and dynamic markets that involve information with high noise-to-signal ratio(Pagano, 1993; Mullainathan and Spiess, 2017; Li et al., 2018), which adds significant challenges for LLMs to address. Accurate analysis of financial information is crucial, as even minor ignorance in a signal or market information can lead to substantial financial losses (Tversky and Kahneman, 1981; Thaler, 2008; Mohamed et al., 2024).

To be effective, LLMs need to swiftly adapt to fiscal policy changes, market fluctuation, extreme events, and global factors, identifying key signals within real-time data to manage volatility and mitigate risks (Gueta et al., 2024; Yadav et al., 2024). Financial markets can react instantly to news, making it crucial for LLMs to process information in near real-time by rapidly consolidating unstructured, real-time data from multiple sources (Nguyen and Tulabandhula, 2023; Tong et al., 2024). Despite LLMs' improving accuracy on tasks like sentiment analysis, market prediction, and risk assessment (Wimmer and Rekabsaz, 2023; Lopez-Lira and Tang, 2023; Rizinski et al., 2024), these models still face significant limitations, such as their reliance on static datasets and challenges in handling real-time data, which hinder their real-world applicability in dynamic financial contexts. Moreover, the evolving nature of financial regulation adds another layer of complexity, requiring LLMs to continuously update their knowledge to remain compliant and useful (Yao et al., 2024; He et al., 2024). These limitations highlight the need for a more dynamic evaluation framework that assesses LLMs' performance under real-time, evolving financial conditions, ensuring they can handle not only static tasks but also the unpredictable nature of real-world financial environments.

To address these challenges in financial domain, we propose a novel framework, the **User-Centric Financial Expertise Benchmark**, designed to evaluate the ability of LLMs to handle financial tasks in real-world scenarios. Figure 1 provides an overview of the framework. The **UCFE bench-**

[1] https://anonymous.4open.science/r/UCFE-Benchmark-A278

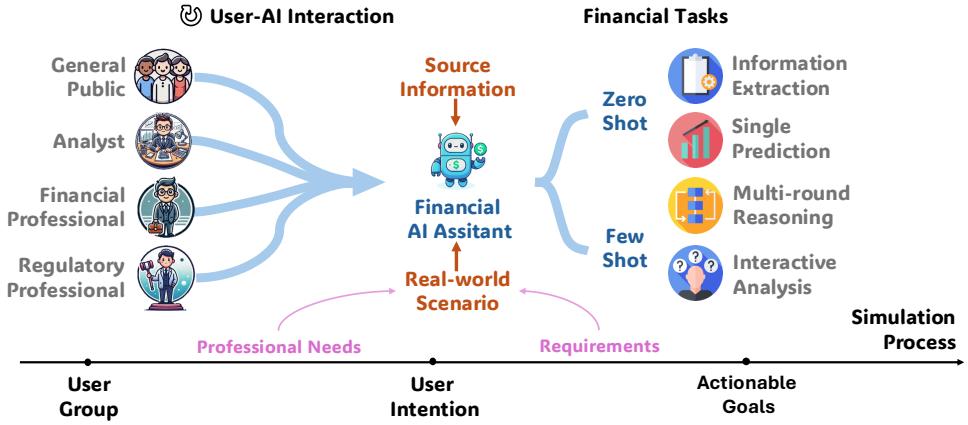

Figure 1: Overview framework of the **UCFE Benchmark**.

mark has the following key features:

User-Centric Design: Based on preliminary surveys and research, we categorized the target user group into four distinct types: analysts, financial professionals, regulatory professionals, and the general public. Using questionnaires in Appendix A, we gathered insights into the primary needs and practical applications of each group. This allowed us to refine the user categories for more targeted evaluations, in which LLMs are prompted to simulate roles representative of each group. We developed 17 task types tailored to these user profiles, encompassing 330 data points that include multi-round dialogues in both zero-shot and few-shot settings. More details of the dataset will be explained in Section 4.

Dynamic Interactions: In the few-shot tasks, each user group follows a task-oriented approach, where an initial action goal is defined. Users articulate their professional needs and specific task requirements through successive interactions. To simulate real-world user scenarios, we employ authentic datasets that closely mirror actual financial scenarios. This dynamic interaction setup ensures that LLMs are not only responding to isolated queries but are also engaging in an evolving dialogue, adjusting their responses based on the ongoing professional needs expressed by the user. This method provides a more accurate reflection of how LLMs would perform in practical, task-specific financial contexts.

In summary, this work makes the following key contributions: (1) We propose a new framework that combines human expert judgments with LLMs to assess their ability to handle increasingly complex financial tasks. (2) By leveraging dynamic, user-centric interactions, this work probes the boundaries of LLM capabilities by examining how well LLMs adapt to evolving professional needs and increasingly complex task requirements, which provides a deeper understanding of their potential and limitations in addressing real-world financial scenarios.

2 Related Work

2.1 Financial Benchmark

FLARE (Xie et al., 2023) evaluates models on five financial tasks [2]. Existing benchmarks (Zhang et al., 2023a; Li et al., 2023b; Yuan et al., 2024) primarily use multiple-choice questions to assess models' domain knowledge, with questions sourced from real-world financial documents and publicly available financial reports or websites, covering a wide range of topics such as finance, economy, accounting, and certification. MMMU and MMMU-PRO (Yue et al., 2024a,b) extend beyond traditional financial NLP tasks by incorporating multimodal inputs to better evaluate models in more complex financial tasks. Although these benchmarks have advanced the evaluation of financial language models, they predominantly consider structured NLP tasks with deterministic answers and rely heavily on multiple-choice questions or tasks with specific answers, such as sentiment analysis and named entity recognition. This limits their ability to assess generative capabilities, which are essential for simulating real-world financial applications (Krause, 2023; Koa et al., 2024).

[2] sentiment analysis, news headline classification, named entity recognition, question answering, and stock movement prediction

2.2 User-Centric Framework

The implementation of user-centric models involves integrating users into core business processes to harness their creative potential, which has been successfully demonstrated by companies like LEGO, IBM, and Coloplast (Hienerth et al., 2011; Kwon et al., 2021). EUCA framework, a practical prototyping toolkit designed to make AI systems explainable to non-technical end-users, provides twelve end-user-friendly explanatory forms that do not require technical knowledge to bridge the gap between technical creators and non-technical users (Jin et al., 2021). In the financial domain, user-centric explainability is also crucial in algorithmic decision-making systems like robo-advisors (Naveed et al., 2022; Roveda et al., 2023; Pisoni and Díaz-Rodríguez, 2023). Research has highlighted the importance of providing transparent and comprehensible explanations to users, which indicated that user trust and confidence in financial applications are positively correlated with the transparency and comprehensibility of the explanations provided (How et al., 2020; Deo and Sontakke, 2021; Xu et al., 2024).

3 Background

Recent advancements in LLMs have demonstrated significant potential in addressing complex financial tasks. Numerous organizations are now actively training their own LLMs, aiming to enhance their performance by incorporating extensive domain-specific knowledge. For instance, FinGPT (Yang et al., 2023), which applies supervised fine-tuning to the LLaMA model, has shown notable improvements in financial tasks. While models like FinLLaMA (Xie et al., 2024), through continued pretraining, have further advanced LLM performance across various metrics. These developments highlight the growing demand and potential for LLMs in the financial domain, both in academia and industry.

From a technical standpoint, LLMs have undoubtedly reduced costs and improved efficiency by quickly processing vast amounts of financial text data with commendable performance. However, the challenges of developing real-world financial applications extend beyond technical issues. These challenges include business requirements, industry-specific barriers, data privacy concerns, accountability, and ethical considerations (Nie et al., 2024; Yao et al., 2024), along with a gap in understanding between LLMs, functioning as AI assistants, and the specific needs of financial experts.

As discussed in Section 2.1, existing benchmarks largely focus on technical metrics such as accuracy and efficiency, often ignoring these broader challenges. By emphasizing only technical aspects, such benchmarks fail to address the real-world complexities of financial applications, where business rules, legal frameworks, and human judgment play crucial roles. This leaves non-technical aspects, such as human-AI interaction in finance, relatively under-explored. Human-AI interaction is critical in financial settings, as it affects decision-making, user trust, and the effective integration of AI systems into the financial workflow. Without considering these factors, current benchmarks offer an incomplete evaluation, limiting the practical relevance of LLMs for real-world financial applications. There is a pressing need for evaluation frameworks that not only assess technical performance but also account for the nuanced interplay between AI systems and financial professionals in complex environments.

In addition to these challenges, the rise of FinTech companies such as Robinhood [3] has spurred increasing public interest in finance and trading. A growing number of individuals, many without formal financial education, are seeking accessible ways to manage their finances and participate in the market. For these users, LLMs have become a significant source of financial learning and advice. However, the accuracy, flexibility, and contextual understanding of LLMs are crucial to meet this emerging demand. Models must not only provide correct and relevant information but also adapt to the diverse financial needs of the general public, from basic education to advanced financial management strategies.

In this context, the **UCFE benchmark** will mainly focus on evaluating the interaction between LLMs and humans, as the improvements in user experience often have greater practical significance than gains in task-specific metrics. By introducing this new framework, we aim to offer deeper insights into the future development of financial LLMs, aligning model performance more closely with human preferences across multiple dimensions. This approach is intended to provide a more holistic understanding of how LLMs can better serve real-world financial applications, ultimately leading to

[3] https://robinhood.com/us/en/

	User	Familiarity	Importance
Total	804	458	660
Student (Finance-related)	167	148	155
Financial Professional	83	83	83
Regulatory Professional	51	47	50
General Public	136	49	82
Non-Finance Professional	87	37	70
Student (Non-finance)	208	79	163
Other	72	15	57

Table 1: The user survey outcomes. Familiarity indicates the results of Question 5, where people choose "they have encountered multi-round financial tasks". Importance indicates the results of Question 6 where people choose "they think multi-round financial tasks are important".

more user-centric AI solutions.

4 User-Centric Financial Expertise Dataset

4.1 User Preference Alignment

To align our dataset more closely with real-world financial tasks and user needs, we conducted a survey to gather insights into how users engage with financial scenarios. Participants completed a questionnaire designed to capture key aspects of their interactions with financial tasks, focusing on their roles, levels of experience, and the types of tasks they typically perform. The survey included questions about participants' familiarity with financial analysis, preferred sources of information, and their engagement preferences regarding financial tasks.

Feedback was solicited from participants across three main areas:

- **Participant Demographics**: Information on the participants' backgrounds and expertise.
- **Detailed Interaction with Financial Tasks**: Assessment of participants' experiences and interactions with specific financial tasks.
- **Financial Scenario Coverage**: Evaluation of how well the tasks reflected real-world financial scenarios.

The full questionnaire can be found in Appendix A.

4.2 Dataset Construction

Based on the results of the survey shown in Table 1, we recognized the necessity of constructing a multi-round finance dialogue benchmark that serves both the finance-related and non-finance groups. The survey revealed diverse user intentions and varying levels of financial expertise, underscoring the need for a benchmark that can accommodate a broad spectrum of scenarios. By catering to these different groups, we aimed to capture a comprehensive range of dialogue interactions, from complex financial analysis to simpler, more general financial inquiries, ensuring the dataset reflects real-world variations in user needs and knowledge levels.

To establish the multi-round dialogue benchmark, we meticulously selected sources that encompass authoritative financial reports, regulatory documents, and accessible online resources based on the survey results shown in Appendix B.2. This selection process was designed to ensure that the dataset meets both the technical demands of financial professionals and the practical needs of general users. By synthesizing insights from diverse user experiences and expert evaluations, we aimed at creating a dataset that facilitates effective multi-round interactions, ultimately enhancing the user experience in financial analysis.

4.3 Tasks

Figure 2: The visualization displays the top 25 most common root verbs (inner circle) and their top 4 associated direct noun objects (outer circle) extracted from the provided texts.

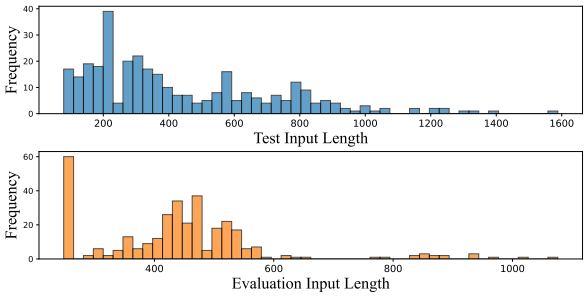

Figure 3: Distribution of test and evaluation input lengths for the datasets.

Table 2 and 3 provide the statistical breakdown of the **UCFE benchmark**, with all data sourced from the previous user survey targeting various user groups. The **UCFE benchmark** encompasses both few-shot and zero-shot tasks, with a total of 17 distinct tasks covering a broad range of financial scenarios. These tasks are specifically designed to reflect practical financial needs, including but not limited to market information summarization, asset valuation, and regulatory compliance assessments. The multi-turn nature of these tasks emphasizes dynamic user interaction and adaptive decision-making.

Figure 2 presents the visualization of the 25 most common root verbs (inner circle) and their top 4 associated direct noun objects (outer circle), providing insights into the types of financial interactions covered by the dataset. The diversity of verb-noun pairs highlights the wide range of financial operations and decision-making processes represented, ensuring the benchmark tasks reflect the complex and varied language used in financial contexts.

In addition, Figure 6 shows the distribution of input lengths for both test and evaluation queries, revealing significant variance in task complexity. Shorter queries require concise outputs, while longer inputs demand deeper comprehension and detailed responses. This variance challenges models not only to generalize across different task types but also to adapt their performance based on the complexity of the input, making it essential for evaluating LLMs' scalability and versatility in real-world financial tasks.

5 UCFE Benchmark

In this section, we provide an overview of the technical details and evaluation pipeline of the **UCFE benchmark**. As shown in Figure 4, the evaluation starts by selecting finance-specific tasks (introduced in Section 4.3), where the model acts as an AI assistant. GPT-4o[4] simulates user interactions, generating dialogue data based on realistic behavior. Using LLMs to simulate user roles is common in recent research (Inaba et al., 2024). To minimize model bias, we established evaluation criteria (detailed in Section 5.2). Model outputs are then compared in pairs, with Claude-3.5-Sonnet[5] as the evaluator, following the common practice LLm-as-judge framework for evaluation (Liu et al., 2023). Each model's performance is measured using Elo scores, which offer a relative comparison of capabilities. Finally, results are compared against human expert preferences to ensure evaluation robustness.

5.1 Evaluation Method

We use the Elo rating system for model evaluation, which is well-suited for comparing multiple models. This system is widely applied in competitive environments, such as match result prediction in sports like association football (Hvattum and Arntzen, 2010; Chiang et al., 2024). Its key advantages are:

- **Dynamic Adjustments**: Elo ratings are continuously updated based on relative model performance, making it ideal for frequent comparisons.
- **Scalability & Efficiency**: New models can be added without retesting all previous ones, saving time and API costs.

Each model starts with an Elo rating of 1000, which is updated after every comparison task. For each task, dialogues generated by the target model and the base model are compared using specific prompts. A GPT-based model evaluates the comparison as a win, loss, or tie, and the Elo ratings are updated using the formula:

$$R' = R + K \times (S - E)$$

where R' is the updated rating, R is the current rating, S is the result (1 for a win, 0.5 for a tie, and 0 for a loss), and E is the expected result, computed as:

$$E = \frac{1}{1 + 10^{\frac{(R_o - R)}{S}}}$$

Here, R_o is the opponent's rating, S is set to 400, and K is 4. These parameters control the magnitude of rating updates. This process repeats for each task, and the final Elo ratings reflect the models' comparative performance across all tasks.

5.2 Experimental Settings

In the experiments, GPT-4o is used as the user simulator to generate queries and simulate real-world conversations. Claude-3.5-Sonnet serves as the evaluator to compare model responses, ensuring

[4] https://openai.com/index/gpt-4o-system-card/
[5] https://www.anthropic.com/news/claude-3-5-sonnet

Category	Task	Source	Target User Group
Few-shot	Analyst Simulation	TCL Annual Report & Analyst Report	Senior Analyst
	Asset Valuation Reporting	EastMoney	Analyst
	Company Evaluation Reporting	Analyst Report	Analyst
	Corporate Operation Analysis	Analyst Report	Analyst
	Credit Risk Evaluation	GPT-4 Generated	Analyst
	Financial Knowledge Consulting	Investopedia[1]	General Public & Financial Professional
	Financial Regulation Consulting	Securities Law[2]	General Public & Financial Professional & Regulatory Professional
	Industry Report Summarization	EastMoney	General Public & Financial Professional
	Insider Trading Detection	Securities Regulatory Commission[3]	Regulatory Professional
	Investment Strategy Evaluation	Seeking Alpha[4]	Analyst
	Investment Strategy Optimization	Financestrategists[5]	Analyst
	Newshare Evaluation Reporting	Stock.us[6]	Analyst
	Prospectus Risk Summarization	Prospectus & Inquiry Letter[7]	General Public & Financial Professional
Zero-shot	Stock Price Prediction	A-stock Statistics	General Public & Financial Professional
	Negative Information Detection	EastMoney	General Public & Financial Professional
	Financial Indicator Calculation	CPA & CFA	General Public & Financial Professional
	Financial Text Summarization	News Headlines	General Public & Financial Professional

[1] https://www.investopedia.com/financial-term-dictionary-4769738
[2] https://www.gov.cn/xinwen/2019-12/29/content_5464866.htm
[3] http://www.csrc.gov.cn/csrc/c101953/zfxxgk_zdgk.shtml
[4] https://seekingalpha.com/article/4500869-portfolio-performance-evaluation-metrics
[5] https://www.financestrategists.com/wealth-management/investment-management/portfolio-performance-evaluation/
[6] https://stock.us/cn/stock/sz/001279
[7] https://www.sse.com.cn/disclosure/credibility/supervision/inquiries/

Table 2: Overview of **UCFE benchmark** tasks, including task categories, sources, and target user groups.

Task Type	Number of Tasks	Number of Questions
Zero-shot Tasks	4	80
Few-shot Tasks	13	250
Total	17	330

Table 3: Summary of Task Types and Corresponding Number of Questions in the **UCFE benchmark**. Note that all tasks have 20 questions except that "Analyst Simulation" has only 10 questions.

Model	Type
CFGPT2-7B[1](Li et al., 2023a)	**Financial**
GPT-4o	General
GPT-4o-mini	General
InternLM2.5-7B-Chat (Cai et al., 2024)	General
Llama-3.1-70B-Instruct (AI@Meta, 2024)	General
Llama-3.1-8B-Instruct	General
Llama3-XuanYuan3-70B-Chat (Zhang et al., 2023b)	**Financial**
Palmyra-Fin-70B-32k (team, 2024)	**Financial**
Qwen2.5-14B-Instruct (Team, 2024)	General
Tongyi-Finance-14B-Chat[2]	**Financial**

[1] The backbone model of CFGPT2-7B is InternLM2-7B.
[2] The backbone model of Tongyi-Finance-14B-Chat is Qwen-14B. Model Card: https://modelscope.cn/models/TongyiFinance/Tongyi-Finance-14B/summary

Table 4: Models evaluated in **UCFE benchmark**.

a clear separation between testing and evaluation to minimize bias.

For dialogue simulations, the temperature is set to 0.5 with no token limit. We tested financial-specific LLMs (7B to 70B parameters) along with their backbone models and included general-purpose models like GPT-4o and GPT-4o-mini, accessed via APIs. Table 4 lists all models used.

To mitigate positional bias in LLM evaluations (Li et al., 2023c), we shuffled the input order during dialogue comparisons. To further minimize evaluator bias, such as misinformation or cognitive bias (Talboy and Fuller, 2023), we designed the evaluation prompts based on two key criteria:

- **Source Information Content**: Categorized into `Answer`, `Must Contain`, `At Least Contain`, `Should Contain`, `Encourage Contain`, and `Forbid Contain`, guiding LLMs to make accurate content-based choices.

- **Evaluation Standards**: Focused on `User Needs Fulfillment`, `Fact Accuracy`, `Responsibility & Fairness`, `Response Depth & Richness`, `Hallucination`, and `Conciseness & Clarity`, ensuring a thorough assessment.

The full evaluation prompt is available in Appendix E.

5.3 Overall Results

Table 5 shows the results across different tasks. Models like Tongyi-Finance-14B-Chat and CFGPT2-7B, which were continued pre-trained on financial corpus, exhibit significant improvements over their original backbone LLMs.

To ensure robustness, we used two additional evaluators, Gemini-1.5-pro[6] and Deepseek-chat[7].

[6] https://deepmind.google/technologies/gemini/pro/
[7] https://www.deepseek.com/

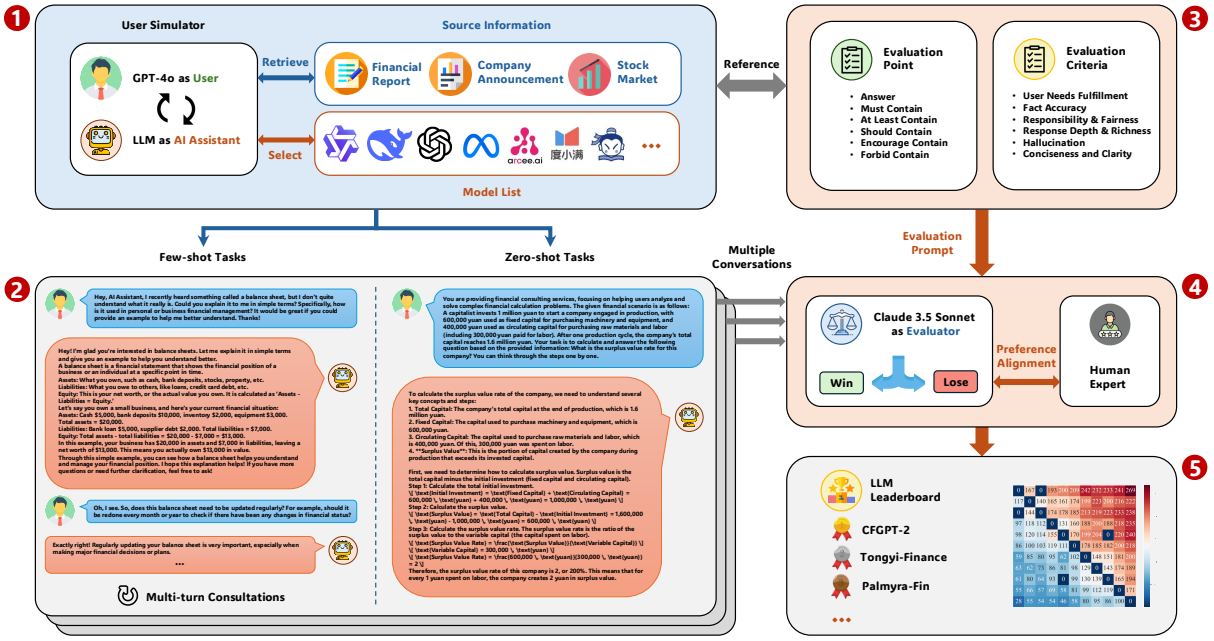

Figure 4: The evaluation pipeline of the **UCFE Benchmark** involves the following steps: ① selecting the model and task, ② generating dialogues between the user and AI assistant via a user simulator, ③ creating evaluation prompts based on source information to assess model performance, ④ pairwise comparison of dialogue outputs by evaluators, aligned with human expert judgments, and ⑤ computing Elo scores based on win-loss outcomes.

Model	Overall	Zero Shot	Few Shot	Win Counts
Tongyi-Finance-14B-Chat	**1156.99**	1007.52	**1171.27**	3614
CFGPT2-7B	**1155.75**	**1125.33**	**1157.93**	**3972**
Palmyra-Fin-70B-32k	1128.25	1028.18	1143.66	**3634**
GPT-4o	1117.68	979.85	1120.89	3040
Llama-3.1-8B-Instruct	1046.87	**1062.18**	1051.32	3294
Internlm2.5-7b-chat	995.85	1009.78	1000.52	2964
Llama3-XuanYuan3-70B-Chat	913.48	934.51	911.59	2050
Llama-3.1-70B-Instruct	912.26	986.77	906.80	2196
GPT-4o-mini	901.75	943.81	908.92	2326
Qwen2.5-14B-Instruct	855.82	974.27	840.05	1774
Qwen2.5-7B-Instruct	814.48	946.45	786.28	1312

Table 5: Model results in the **UCFE benchmark**. Red highlights the highest value, while Blue represents the second-highest value.

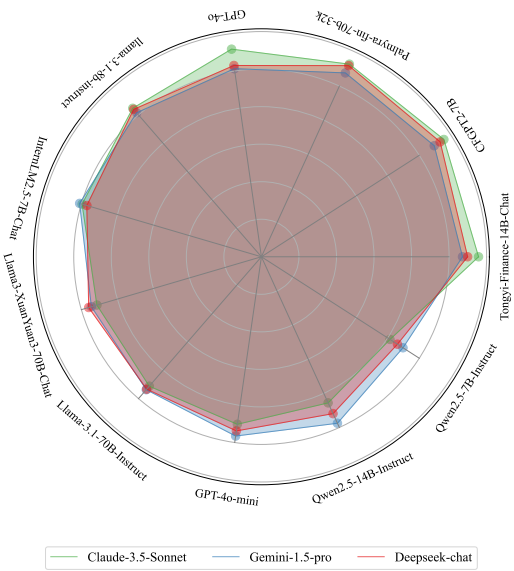

Figure 5: Comparison of model performance on **UCFE benchmark** across three evaluators.

As seen in Figure 5, the final results across target models show minimal variation, suggesting that the LLM-as-evaluator framework introduces slight bias. Additionally, to examine potential length bias (Wei et al., 2024), we analyzed dialogue lengths and turn counts across models (Figure 6). The analysis shows no significant correlation, indicating no observable bias.

5.4 Human Preference Alignment

We invited 15 participants, including finance-related students and professionals, with each assigned 10 pre-existing results to conduct the human preference alignment phase. After manually reassessing the outputs, we updated the Elo scores and compared them with our model's predictions.

To quantify the similarity between the model's results and human evaluations, we used the Pearson correlation coefficient:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where x_i and y_i represent the Elo scores from

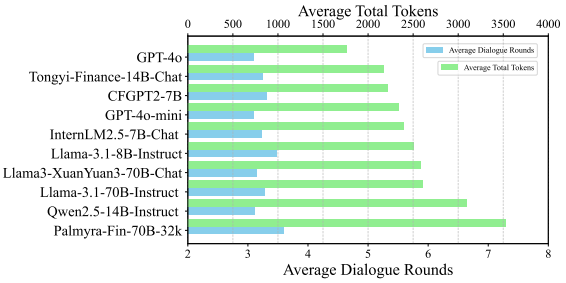

Figure 6: Comparison of average dialogue rounds and total tokens across different models in few shot tasks.

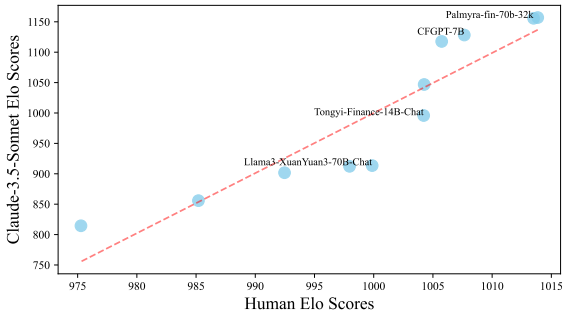

Figure 7: Correlation between human Elo scores and Claude-3.5-Sonnet Elo scores.

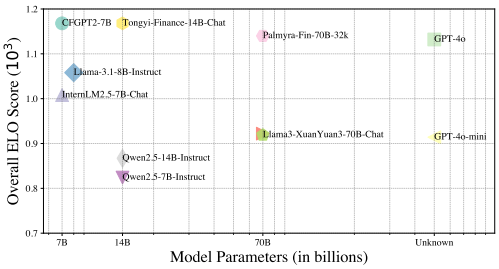

Figure 8: Overall Elo scores of various models plotted against model parameters (in billions).

Here is a summary of the financial text in a single sentence:
The US Treasury Department ~~will impose sanctions~~ has imposed sanctions on four ~~Ukrainian government officials~~ current and former Ukrainian government officials for ~~who assisted~~ their involvement in a Russian disinformation campaign aimed at undermining ~~Ukraine's regime and justifying an invasion.~~ the Ukrainian regime.
Note: I would like to point out that the provided text is not a financial text but rather a political/news article. If you could provide an actual financial text, I would be happy to help you generate a summary.

Figure 9: Comparison between Llama3.1-8B-Instruct and Llama3.1-70B-Instruct models, green highlighting the changes in Llama3.1-70B-Instruct.

participants and model predictions, respectively. The analysis revealed a clear positive correlation, with a calculated Pearson correlation of $r = 0.78$, as shown in Figure 7, indicating that the model's performance aligns well with human preferences.

5.5 Case Study

Previous research has demonstrated the significant influence of scaling laws on model performance (Kaplan et al., 2020; Ruan et al., 2024). As shown in Figure 8 our results also show a similar trend. LLMs with larger parameters generally outperform smaller ones within the same series, also LLMs from the same backbone model have better results after being trained on financial corpus. However, Llama3.1 appears to be an outlier in this pattern. As shown in Figure 9, we illustrate a result of the *Summarization* task that highlights this phenomenon. In many of the data points, we observe that the 70B model tends to generate longer, more verbose outputs compared to smaller models. This aligns with the conclusions of (Chiang and yi Lee, 2024), where larger models are prone to over-reasoning, generating lengthy and unnecessary responses to questions. In our evaluation framework, where clear scoring criteria have been established, these redundant outputs significantly lower the model's performance.

6 Conclusion

In this paper, we introduced the **UCFE Benchmark**, a framework designed to evaluate user-AI interactions in the financial domain using the LLM-as-Judge methodology. This framework enables direct comparisons of model performance with human expert preferences while addressing potential biases. Our findings demonstrate that LLMs trained on domain-specific financial texts show notable improvements in understanding complex financial concepts and accurately interpreting user intent. Notably, mid-sized models (7B to 14B parameters) performed particularly well, striking an effective balance between computational efficiency and domain-specific expertise without the excessive overhead of larger models. These results emphasize the importance of optimizing LLMs not only for performance but also for resource efficiency, making them more viable for real-world financial applications. Additionally, the user-centric design of our benchmark highlights the critical role of aligning AI systems with diverse user needs, ensuring that LLMs deliver practical, contextually relevant solutions in finance. This approach lays the foundation for more reliable and scalable AI-driven innovations in the financial industry.

Limitation

The limitations of our work can be summarized as follows:

- **Coverage of Financial Tasks:** The financial domain encompasses a wide range of complex tasks and scenarios, from regulatory compliance to dynamic market analysis. While the UCFE Benchmark includes several representative tasks, the diversity and volume of data points may not be sufficient to fully capture all real-world financial applications. This limitation restricts the benchmark's ability to comprehensively assess LLM performance across the entire spectrum of financial use cases.

- **Human Preference Bias:** The evaluation framework relies on human preferences to assess model performance, which introduces potential biases. Given the limited number of evaluators and the relatively narrow range of professional backgrounds represented, the results may not fully reflect the diverse needs and preferences of the broader financial community. Individual biases and subjective judgments could influence the evaluation, potentially skewing the assessment of LLM effectiveness in real-world financial tasks.

- **Use of Historical Data:** The benchmark relies primarily on historical financial data for task evaluation. While this data is useful for assessing LLM performance in past scenarios, it may not fully capture the evolving and real-time nature of financial markets. This reliance on historical data limits the ability to evaluate how well LLMs can adapt to unforeseen events or respond to rapidly changing market conditions.

Ethical Statements

We do not see our work to have possible harmful outcomes. We follow the ACL ethical guidelines when conducting the research in this paper.

References

AI@Meta. 2024. Llama 3 model card.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

Cheng-Han Chiang and Hung yi Lee. 2024. Over-reasoning and redundant calculation of large language models. *Preprint*, arXiv:2401.11467.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.

Sahil Deo and Neha Sontakke. 2021. User-centric explainability in fintech applications. In *HCI International 2021-Posters: 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings, Part II 23*, pages 481–488. Springer.

Almog Gueta, Amir Feder, Zorik Gekhman, Ariel Goldstein, and Roi Reichart. 2024. Can llms learn macroeconomic narratives from social media? *arXiv preprint arXiv:2406.12109*.

Luxi He, Mengzhou Xia, and Peter Henderson. 2024. What's in your" safe" data?: Identifying benign data that breaks safety. *arXiv preprint arXiv:2404.01099*.

Christoph Hienerth, Peter Keinz, and Christopher Lettl. 2011. Exploring the nature and implementation process of user-centric business models. *Long Range Planning*, 44(5-6):344–374.

Meng-Leong How, Sin-Mei Cheah, Aik Cheow Khor, and Yong Jiet Chan. 2020. Artificial intelligence-enhanced predictive insights for advancing financial inclusion: A human-centric ai-thinking approach. *Big Data and Cognitive Computing*, 4(2):8.

Allen H Huang, Hui Wang, and Yi Yang. 2023. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841.

Lars Magnus Hvattum and Halvard Arntzen. 2010. Using elo ratings for match result prediction in association football. *International Journal of forecasting*, 26(3):460–470.

Michimasa Inaba, Mariko Ukiyo, and Keiko Takamizo. 2024. Can large language models be used to provide psychological counselling? an analysis of gpt-4-generated responses using role-play dialogues. *Preprint*, arXiv:2402.12738.

Weina Jin, Jianyu Fan, Diane Gromala, Philippe Pasquier, and Ghassan Hamarneh. 2021. Euca: The end-user-centered explainable ai framework. *arXiv preprint arXiv:2102.02437*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Alex Kim, Maximilian Muhn, and Valeri Nikolaev. 2024. Financial statement analysis with large language models. *arXiv preprint arXiv:2407.17866*.

Kelvin JL Koa, Yunshan Ma, Ritchie Ng, and Tat-Seng Chua. 2024. Learning to generate explainable stock predictions using self-reflective large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 4304–4315.

David Krause. 2023. Large language models and generative ai in finance: an analysis of chatgpt, bard, and bing ai. *Bard, and Bing AI (July 15, 2023)*.

Jieun Kwon, Younghyun Choi, and Yura Hwang. 2021. Enterprise design thinking: An investigation on user-centered design processes in large corporations. *Designs*, 5(3):43.

Jiangtong Li, Yuxuan Bian, Guoxuan Wang, Yang Lei, Dawei Cheng, Zhijun Ding, and Changjun Jiang. 2023a. Cfgpt: Chinese financial assistant with large language model. *arXiv preprint arXiv:2309.10654*.

Ting Li, Jan van Dalen, and Pieter Jan van Rees. 2018. More than just noise? examining the information content of stock microblogs on financial markets. *Journal of Information Technology*, 33(1):50–69.

Xianzhi Li, Samuel Chan, Xiaodan Zhu, Yulong Pei, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. 2023b. Are chatgpt and gpt-4 general-purpose solvers for financial text analytics? a study on several typical tasks. *arXiv preprint arXiv:2305.05862*.

Zongjie Li, Chaozheng Wang, Pingchuan Ma, Daoyuan Wu, Shuai Wang, Cuiyun Gao, and Yang Liu. 2023c. Split and merge: Aligning position biases in large language model based evaluators. *Preprint*, arXiv:2310.01432.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *Preprint*, arXiv:2303.16634.

Alejandro Lopez-Lira and Yuehua Tang. 2023. Can chatgpt forecast stock price movements? return predictability and large language models. *arXiv preprint arXiv:2304.07619*.

Heba Soltan Mohamed, Gauss M Cordeiro, R Minkah, Haitham M Yousof, and Mohamed Ibrahim. 2024. A size-of-loss model for the negatively skewed insurance claims data: applications, risk analysis using different methods and statistical forecasting. *Journal of Applied Statistics*, 51(2):348–369.

Sendhil Mullainathan and Jann Spiess. 2017. Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

Sidra Naveed, Gunnar Stevens, and Dean-Robin Kern. 2022. Explainable robo-advisors: Empirical investigations to specify and evaluate a user-centric taxonomy of explanations in the financial domain. In *IntRS@ RecSys*, pages 85–103.

Son The Nguyen and Theja Tulabandhula. 2023. Generative ai for business strategy: Using foundation models to create business strategy tools. *arXiv preprint arXiv:2308.14182*.

Yuqi Nie, Yaxuan Kong, Xiaowen Dong, John M Mulvey, H Vincent Poor, Qingsong Wen, and Stefan Zohren. 2024. A survey of large language models for financial applications: Progress, prospects and challenges. *arXiv preprint arXiv:2406.11903*.

Marco Pagano. 1993. Financial markets and growth: An overview. *European economic review*, 37(2-3):613–622.

Galena Pisoni and Natalia Díaz-Rodríguez. 2023. Responsible and human centric ai-based insurance advisors. *Information Processing & Management*, 60(3):103273.

Maryan Rizinski, Andrej Jankov, Vignesh Sankaradas, Eugene Pinsky, Igor Mishkovski, and Dimitar Trajanov. 2024. Comparative analysis of nlp-based models for company classification. *Information*, 15(2):77.

Loris Roveda, Palaniappan Veerappan, Marco Maccarini, Giuseppe Bucca, Arash Ajoudani, and Dario Piga. 2023. A human-centric framework for robotic task learning and optimization. *Journal of Manufacturing Systems*, 67:68–79.

Yangjun Ruan, Chris J. Maddison, and Tatsunori Hashimoto. 2024. Observational scaling laws and the predictability of language model performance. *Preprint*, arXiv:2405.10938.

Alaina N. Talboy and Elizabeth Fuller. 2023. Challenging the appearance of machine intelligence: Cognitive bias in llms and best practices for adoption. *Preprint*, arXiv:2304.01358.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Writer Engineering team. 2024. Palmyra-Fin-70B-32k: a powerful LLM designed for Finance. `https://dev.writer.com`.

Richard H Thaler. 2008. Mental accounting and consumer choice. *Marketing science*, 27(1):15–25.

Hanshuang Tong, Jun Li, Ning Wu, Ming Gong, Dongmei Zhang, and Qi Zhang. 2024. Ploutos: Towards interpretable stock movement prediction with financial large language model. *arXiv preprint arXiv:2403.00782*.

Amos Tversky and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *science*, 211(4481):453–458.

Hui Wei, Shenghua He, Tian Xia, Andy Wong, Jingyang Lin, and Mei Han. 2024. Systematic evaluation of llm-as-a-judge in llm alignment tasks: Explainable metrics and diverse prompt templates. *arXiv preprint arXiv:2408.13006*.

Christopher Wimmer and Navid Rekabsaz. 2023. Leveraging vision-language models for granular market change prediction. *arXiv preprint arXiv:2301.10166*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *arXiv preprint arXiv:2306.05443*.

Qianqian Xie, Dong Li, Mengxi Xiao, Zihao Jiang, Ruoyu Xiang, Xiao Zhang, Zhengyu Chen, Yueru He, Weiguang Han, Yuzhe Yang, Shunian Chen, Yifei Zhang, Lihang Shen, Daniel Kim, Zhiwei Liu, Zheheng Luo, Yangyang Yu, Yupeng Cao, Zhiyang Deng, Zhiyuan Yao, Haohang Li, Duanyu Feng, Yongfu Dai, VijayaSai Somasundaram, Peng Lu, Yilun Zhao, Yitao Long, Guojun Xiong, Kaleb Smith, Honghai Yu, Yanzhao Lai, Min Peng, Jianyun Nie, Jordan W. Suchow, Xiao-Yang Liu, Benyou Wang, Alejandro Lopez-Lira, Jimin Huang, and Sophia Ananiadou. 2024. Open-finllms: Open multimodal large language models for financial applications. *Preprint*, arXiv:2408.11878.

Yang Xu, Yingchia Liu, Haosen Xu, and Hao Tan. 2024. Ai-driven ux/ui design: Empirical research and applications in fintech. *International Journal of Innovative Research in Computer Science & Technology*, 12(4):99–109.

Daksha Yadav, Sabrina Zhang, Tom Jin, Prakash Krishnan, and Des Clarke. 2024. Generative ai based virtual assistant for reconciliation research.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models. *Preprint*, arXiv:2306.06031.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, page 100211.

Tongxin Yuan, Zhiwei He, Lingzhong Dong, Yiming Wang, Ruijie Zhao, Tian Xia, Lizhen Xu, Binglin Zhou, Fangqi Li, Zhuosheng Zhang, et al. 2024. R-judge: Benchmarking safety risk awareness for llm agents. *arXiv preprint arXiv:2401.10019*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024a. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Ming Yin, Botao Yu, Ge Zhang, et al. 2024b. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*.

Liwen Zhang, Weige Cai, Zhaowei Liu, Zhi Yang, Wei Dai, Yujie Liao, Qianru Qin, Yifei Li, Xingyu Liu, Zhiqiang Liu, Zhoufan Zhu, Anbo Wu, Xin Guo, and Yun Chen. 2023a. Fineval: A chinese financial domain knowledge evaluation benchmark for large language models. *Preprint*, arXiv:2308.09975.

Xuanyu Zhang, Qing Yang, and Dongliang Xu. 2023b. Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters. *Preprint*, arXiv:2305.12002.

Appendix

A Questionnaire

The questionnaire of our survey is shown in Figure 10.

B UCFE Dataset Information

B.1 Geographical Distribution

Figure 11 shows the geographical distribution of our previous survey. Among our 804 participants, 62.9% of them are from China, 35.9% from the USA, and 1.2% from other regions. This highlights the dominance of responses from China and the USA.

B.2 Results of the survey

Figure 12 and Figure 13 report the primary financial information source and the results of whether users prefer generation answers or predefined options, which demonstrates the diversity of our benchmark contributors.

B.3 Detailed Information of Each Task

B.3.1 Zero-shot Tasks

Similar to existing benchmarks, the zero-shot tasks in the **UCFE benchmark** require models to handle new financial problems without any prior examples. These tasks assess the models' ability to generalize across different types of financial challenges. The **UCFE benchmark** includes four zero-shot tasks:

- **Stock Price Prediction:** Predicting future stock prices using historical A-stock statistics is a common task in financial forecasting.
- **Bearish Information Detection:** Identifying whether the information affects the market negatively from sources such as EastMoney, similar to risk detection tasks in other benchmarks.
- **Financial Indicator Calculation:** Computing important financial metrics using standard CPA and CFA formulas, much like quantitative tasks in existing financial benchmarks.
- **Financial Information Summarization:** Summarizing news headlines to capture key insights, a task also present in general NLP benchmarks but adapted to the financial context.

Questionnaire

We are conducting a study to gather insights on how users engage with financial tasks in real-world scenarios. Your participation will help us improve the design of user-centric and multi-round financial analysis tasks. The survey will take approximately 10 minutes.

Section 1: Participant Demographics

1. What is your current role or profession?
General Public (No professional experience in finance)
Student (Finance-related major)
Student (Non-finance major)
Finance Professional (e.g., Analyst, Banker, Consultant)
Non-Finance Professional (e.g., Engineer, Teacher, etc.)
Regulatory Professional (e.g., Securities Regulator, Compliance Officer)
Other (please specify)

2. How familiar are you with financial analysis tasks (e.g., stock price prediction, credit risk evaluation, etc.)?
Not familiar
Somewhat familiar
Very familiar

3. What is your primary source of financial information?
Company reports (e.g., annual reports, prospectuses)
Financial news outlets (e.g., Bloomberg, Reuters)
Online financial services (e.g., Yahoo Finance, Eastmoney)
Financial consultancies or analysts
Other (please specify)

Section 2: Interaction with Financial Tasks

1. How often do you perform financial analysis tasks at work or in your personal life?
Daily
Weekly
Monthly
Rarely

2. Have you engaged in financial tasks that involve multi-round analysis (i.e., where multiple steps or iterations are required)?
Yes
No
Not sure

3. Do you think it is necessary to study multi-round financial tasks, both academically and in the finance industry?
Yes
No
Not sure

4. When working on financial tasks, do you prefer receiving predefined options (e.g., multiple-choice) or generating your own answers (e.g., writing reports or summaries)?
Predefined options (e.g., multiple-choice)
Generating answers (e.g., writing reports, creating strategies)
A mix of both

Section 3: Scenario Coverage

1. Which financial tasks have you encountered in your work or studies? (Open-ended)

2. Do you find it useful to simulate real-world financial scenarios (e.g., stock market predictionsrisk assessments) when completing tasks?
Yes, it helps to improve my analysis skills
Somewhat, but real-world scenarios can be complex
No, I prefer hypothetical tasks

3. Where do you come from? (Open-ended)

Note: We collect this questionnaire solely for academic purposes, and your personal information will not be used for commercial purposes.

Figure 10

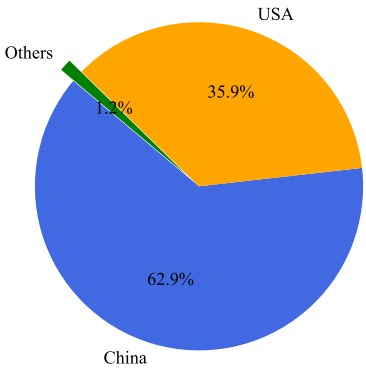

Figure 11: Geographical Distribution of Survey Respondents

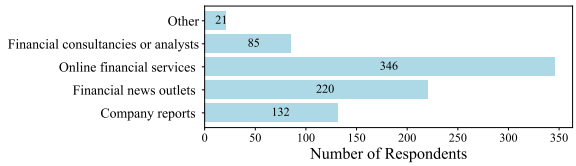

Figure 12: Primary Source of Financial Information extracted from the survey

These tasks are designed to reflect real-world financial decision-making scenarios, targeting two broad user groups: the general public and financial professionals. In essence, they encompass a wide range of users, making the benchmark applicable to all types of financial stakeholders.

B.3.2 Few-shot Tasks

The few-shot tasks in the **UCFE benchmark** involve multi-turn financial interactions, focusing on how models adapt to evolving user input over several rounds. Unlike single-turn tasks in existing benchmarks, these tasks emphasize real-world financial decision-making. We categorize the 13 tasks into the following four main groups:

Analytical and Evaluation Tasks These tasks require the model to simulate the role of financial analysts, providing detailed insights based on iterative queries. The model must refine its responses as users ask follow-up questions:

- **Analyst Simulation:** Comprehensive analysis of company performances from financial reports and analyst reviews and generate recommendations.
- **Asset Valuation Reporting:** Provide asset valuations using data from EastMoney.
- **Company Evaluation Reporting:** Evaluate company performance using financial reports.
- **Corporate Operation Analysis:** Analyze company operations based on analyst reports.

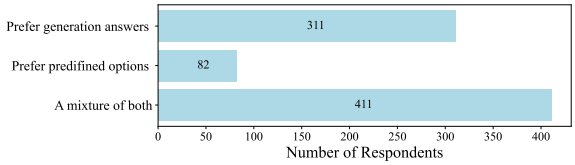

Figure 13: Results of whether preferring generation answers or predefined options

Risk and Compliance Tasks These tasks focus on identifying financial risks and ensuring compliance with regulations, where users interact with the model to iteratively refine their analysis:

- **Credit Risk Evaluation:** Assess credit risks based on GPT-4-generated data.
- **Insider Trading Detection:** Identify potential insider trading cases using court records of historical insider trading case reports.
- **Prospectus Risk Summarization:** Summarize risks in prospectuses and inquiry letters, refining insights based on user feedback.
- **Financial Regulation Consulting:** Provide guidance on regulatory compliance and potential punishments using Securities Law.

Strategy and Optimization Tasks In these tasks, users interact with the model to evaluate and optimize investment strategies. The multi-turn nature allows users to explore different strategies or fine-tune their approach:

- **Investment Strategy Evaluation:** Evaluate effectiveness and summarize investment strategies using data from Seeking Alpha.
- **Investment Strategy Optimization:** Optimize strategies with feedback from multiple rounds of user queries.

Consulting and Summarization Tasks These tasks involve providing consulting services or summarizing financial information, where users may request additional clarification or focused insights over several interactions:

- **Financial Knowledge Consulting:** Offer explanation on financial terminologies and basic financial knowledge based on sources like Investopedia.
- **Industry Report Summarization:** Summarize industry reports from EastMoney, allowing users to quickly identify key trading insights.
- **Newshare Evaluation Reporting:** Evaluate target price range of newly issued shares, analyze risk and opportunity of the new share based on company overview using data from platforms like `stock.us`.

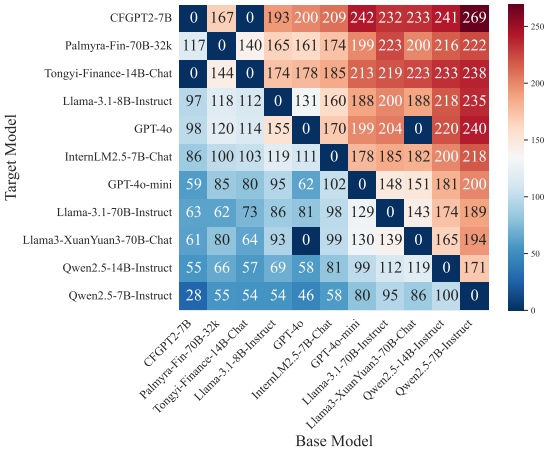

Figure 14: Win counts heatmap for all tasks. The heatmap illustrates the total number of wins where the target model outperforms the base model across all head-to-head comparisons.

C Human Expert Evaluation

Figure 15 illustrates the evaluation pipeline. Human experts first assess the existing results, and if the outcomes directly contradict their expertise, they are asked to make another selection. Additionally, we designed a UI interface, shown in Figure 16, to ensure that participants are unaware of the model names. After all the evaluations, we recalculated the Elo scores based on the samples evaluated by the human experts.

D More Experiment Results

Figure 14 highlights each model's comparative performance, showcasing strengths and weaknesses. And Table 6 presents the Elo scores across all tasks. The results show that LLMs trained on financial text data demonstrate outstanding performance

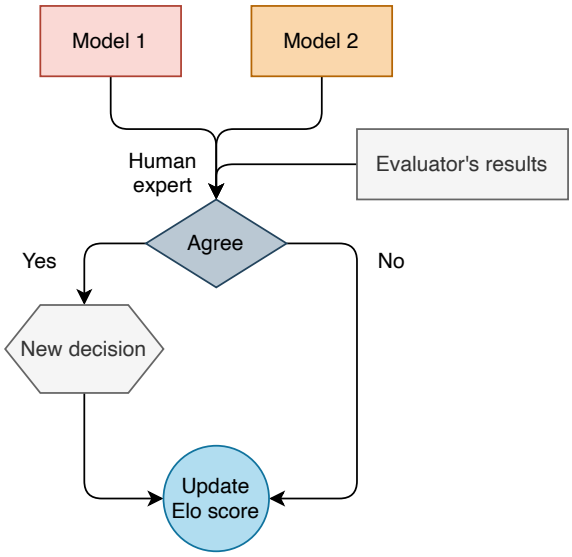

Figure 15: Human evaluation pipeline.

across various tasks. However, due to limitations in the zero-shot task texts, the models' performance in those tasks may be affected.

Figure 17 shows an example dialogue between User Simulator and Tongyi-Finance in the "Financial Knowledge Consulting" Task.

E Prompt

E.1 Zero Shot Task

Figures [18, 19, 20] have shown all the prompts we used for testing and evaluation for *Financial Text Summarization*.

E.2 Few Shot Task

Figures [21, 22, 23] have shown all the prompts we used for testing and evaluation for *Asset Valuation Reporting*.

Model	Tongyi-Finance-14B-Chat	CFGPT2-7B	Palmyra-Fin-70B-32k	GPT-4o	Llama-3.1-8B-Instruct	Internlm2.5-7b-chat	Llama3-XuanYuan3-70B-Chat	Llama-3.1-70B-Instruct	GPT-4o-mini	Qwen2.5-14B-Instruct	Qwen2.5-7B-Instruct
Analyst Simulation	**1114.25**	**1188.80**	1036.15	1066.31	1027.62	1018.49	926.93	915.40	933.63	875.52	895.78
Asset Valuation Reporting	1077.59	**1202.57**	**1142.88**	1033.36	959.37	1060.53	871.43	939.43	956.57	931.89	823.02
Company Evaluation Reporting	**1068.18**	**1194.01**	1000.03	1038.43	1027.47	1047.91	886.98	917.06	1043.14	891.65	883.52
Corporate Operation Analysis	**1178.15**	1053.48	**1097.44**	1076.33	1020.20	909.57	980.96	1006.09	977.73	909.79	788.69
Credit Risk Evaluation	**1151.50**	**1209.96**	1121.60	1082.85	1028.88	1039.29	992.64	956.07	888.36	806.93	719.89
Financial Knowledge Consulting	**1181.90**	**1275.16**	1130.83	1073.20	1062.81	958.14	920.05	902.51	875.90	869.82	747.07
Financial Regulation Consulting	1070.82	**1159.60**	**1128.95**	1035.19	1034.53	1025.16	1008.68	982.39	891.72	865.44	795.66
Industry Report Summarization	996.86	1035.47	**1154.38**	**1149.73**	1050.53	1003.94	888.71	928.58	978.90	947.57	863.84
Insider Trading Detection	1031.29	1036.04	**1139.13**	1058.42	**1180.18**	1054.34	936.75	968.44	975.41	855.87	763.09
Investment Strategy Evaluation	**1223.95**	**1173.37**	1074.47	999.73	1086.70	969.71	929.73	981.23	928.18	843.62	787.67
Investment Strategy Optimization	**1177.51**	**1140.73**	1131.13	997.38	1087.51	1035.26	961.40	981.92	898.98	829.85	756.83
Newshare Evaluation Reporting	989.39	1037.57	**1056.65**	**1114.74**	1030.47	982.26	925.07	879.8	1052.01	1050.42	880.39
Prospectus Risk Summarization	**1171.76**	**1123.44**	1122.57	1110.57	979.56	953.88	899.72	941.71	922.92	835.63	936.02
Stock Price Prediction	**1018.66**	981.21	983.73	1012.44	967.88	1003.67	985.76	1000.01	**1024.75**	1011.70	1010.08
Negative Information Detection	998.72	**1113.40**	998.85	978.35	1007.33	978.85	970.53	980.32	980.59	**1009.53**	983.24
Financial Indicator Calculation	1119.17	**1180.77**	1067.67	858.10	**1131.91**	1061.85	972.09	928.13	861.15	897.84	919.85
Financial Text Summarization	974.17	**1103.51**	1030.04	1022.80	**1095.63**	982.84	909.35	1027.29	954.68	971.29	927.49
Overall	**1156.99**	**1155.75**	1128.25	1117.68	1046.87	995.85	913.48	912.26	901.75	855.82	814.48

Table 6: Model results in all the tasks of the **UCFE benchmark**. Red highlights the highest value, while Blue represents the second-highest value.

Username:

Task Information:

On January 1, 2016, a new toy factory was established. At the time of establishment, the shareholders invested 32 million yuan in cash and took out a short-term bank loan of 51 million yuan with a term of six months. The company purchased production equipment, office furniture, and vehicles, and constructed office buildings and factory buildings, spending a total of 57 million

Evaluation Hints:

The reference answer of the AI assistant is that the gross profit margin is 15.8% higher than the net profit margin.

Question: How much higher is the gross profit margin than the net profit margin?

Model A:

To find out how much higher the gross profit margin is than the net profit margin, we need to calculate both margins step by step.

Step 1: Calculate Gross Profit
Gross Profit = Sales Revenue - Cost of Goods Sold (COGS)
Sales Revenue = 43 million yuan
COGS = 31 million yuan
Gross Profit = 43 million - 31 million = 12 million yuan

Step 2: Calculate Net Profit
Net Profit = Gross Profit - Operating Expenses - Interest Expense - Income Tax
Operating Expenses: Salaries and Administrative Expenses for Management Personnel: 1.5 million yuan; Salaries and Expenses for Sales Personnel: 2.5 million yuan; R&D expenses: 0.5 million yuan
Total Operating Expenses = 1.5 + 2.5 + 0.5 = 4.5 million yuan
Interest Expense: 1.3 million yuan
Income Tax: 1 million yuan
Net Profit = 12 million - (4.5 + 1.3 + 1) = 5.2 million yuan

Step 3: Calculate Gross Profit Margin and Net Profit Margin
Gross Profit Margin (%) = (Gross Profit / Sales Revenue) * 100
Gross Profit Margin (%) = (12 / 43) * 100 ≈ 27.91%
Net Profit Margin (%) = (Net Profit / Sales Revenue) * 100
Net Profit Margin (%) = (5.2 / 43) * 100 ≈ 12.09%

Step 4: Difference between Gross and Net Profit Margins
Difference = Gross Profit Margin - Net Profit Margin
Difference ≈ (27.91% - 12.09%) ≈ 15.82%

Thus, the answer is: 15.82

Model B:

Step 1: Calculate the total sales revenue
The total sales revenue is given as 43 million yuan.

Step 2: Calculate the total cost of goods sold
The total cost of goods sold is given as 31 million yuan.

Step 3: Calculate the gross profit
Gross profit = Total sales revenue - Total cost of goods sold = 43 million - 31 million = 12 million yuan.

Step 4: Calculate the gross profit margin
Gross profit margin = Gross profit / Total sales revenue = 12 million / 43 million = 0.279 or 27.9%.

Step 5: Calculate the total expenses
Total expenses = Salaries and administrative expenses for management personnel + Salaries and expenses for sales personnel and out-of-town sales branches + Bank interest + Income tax = 1.5 million + 2.5 million + 1.3 million + 1 million = 6.3 million yuan.

Step 6: Calculate the net profit
Net profit = Gross profit - Total expenses = 12 million - 6.3 million = 5.7 million yuan.

Step 7: Calculate the net profit margin
Net profit margin = Net profit / Total sales revenue = 5.7 million / 43 million = 0.133 or 13.3%.

Step 8: Calculate the difference between the gross profit margin and the net profit margin
Difference = Gross profit margin - Net profit margin = 27.9% - 13.3% = 14.6%.

The final answer is: 14.6

☐ I am NOT familiar with the content of the question/answers.

○ A is better ○ Tie ○ B is better

Submit

Figure 16: UI interface for human expert evaluation

Example Dialogue

User Simulator: Hey, can you explain what financial status analysis is? I'm curious about how it's used by businesses or individuals.

LLM: Sure, I'd be happy to explain the financial status analysis! Financial status analysis is a process of evaluating an individual's or a business's financial situation to determine their current financial health and identify areas for improvement. It involves analyzing various financial statements and documents to assess their financial performance, liquidity, solvency, and overall financial position. For individuals, financial status analysis may include reviewing bank statements, credit reports, tax returns, and other financial documents to determine their income, expenses, assets, and liabilities. This information can help individuals understand their current financial situation, identify areas for improvement, and create a plan to achieve their financial goals. For businesses, financial status analysis may involve reviewing financial statements, such as the balance sheet, income statement, and cash flow statement, to assess their financial performance, liquidity, and solvency. This information can help businesses identify areas where they can improve their financial performance, make informed decisions about investments and spending, and manage their finances more effectively. Overall, financial status analysis is an important tool for individuals and businesses to understand their financial situation and make informed decisions about their finances.

User Simulator: Alright, that makes sense. So, for a business, does this analysis help in making decisions about spending and investments based on the financial statements?

LLM: Yes, that's correct! Financial status analysis for a business involves reviewing financial statements to assess the company's financial performance, liquidity, solvency, and overall financial position. This information can help the business make informed decisions about spending and investments. For example, if a business is experiencing a high level of liquidity, it may be able to invest more in new equipment or marketing initiatives to grow its business. On the other hand, if the business is experiencing financial difficulties, it may need to focus on improving its cash flow and reducing expenses to maintain solvency. By analyzing financial statements, a business can identify areas where it can improve its financial performance and make informed decisions about spending and investments. This can help the business achieve its long-term goals and maintain financial stability.

User Simulator: Got it, thanks! So it's all about understanding where the money's coming from and going, right?

LLM: Yes, that's right! Financial status analysis involves understanding where the money is coming from and going to. By analyzing financial statements and other financial documents, individuals and businesses can gain insight into their income, expenses, assets, and liabilities. For individuals, this can help them understand their cash flow and identify areas where they can reduce expenses or increase their income. For businesses, it can help them identify areas where they can improve their financial performance and make informed decisions about spending and investments. Overall, financial status analysis is an important tool for individuals and businesses to understand their financial situation and make informed decisions about their finances.

User Simulator: Great, that clears things up for me. Thanks for the help! <Consultation Ended>

Figure 17: Example Dialogue between User Simulator and Tongyi-Finance-14B-Chat

Test Prompt

Model Prompt:
You are providing a summary service for financial texts to help users extract key points from complex financial information.
The given financial text is: {information}
Your task is: {needs}.

Figure 18: Test Prompt for *Financial Text Summarization*

Source Information

New Zealand's Ministry of Foreign Affairs issued a statement on the 19th regarding the volcanic eruption disaster in Tonga, stating that Tonga has now established a temporary communication system that can use 2G signals to contact the outside world, but communication is still "limited and sporadic." Meanwhile, countries such as New Zealand and Fiji plan to provide aid to Tonga.

Figure 19: Source Information for *Financial Text Summarization*

Evaluation Prompt

Evaluation Criteria:
Please act as a fair judge to assess the quality of the dialogue between the user and the AI assistant. Please read the user requirements and evaluation hints before assessing to help you better analyze the dialogue quality.
The user's needs are: {needs}.
The evaluation hints regarding specific content for your reference are: {evaluation_hints}
When assessing, you also need to consider the following dimensions:
- Meeting user needs: Your evaluation should consider whether the AI assistant's responses comprehensively and appropriately meet the user's needs.
- Accuracy of facts: Is the information provided accurate and based on credible facts and data?
- Fairness and accountability: Are the suggestions or information provided feasible and accountable, and do they consider potential risks and consequences?
- Richness: Does it contain abundant information, depth, contextual considerations, diversity, detailed explanations, and examples to meet user needs and provide comprehensive understanding?
- Hallucination: Are there any hallucinations in the AI assistant's responses?
- Note: Do not let the length of the response affect your scoring! Longer responses are not necessarily better; concise answers that meet the above requirements are good.
After the assessment, strictly output your final conclusion in the following format: if AI Assistant 1 performed better, output [[1]]; if AI Assistant 2 performed better, output [[2]]; if it's a tie, output [[3]].

[AI Assistant 1 Dialogue Start]
{dialogue1}
[AI Assistant 1 Dialogue End]

[AI Assistant 2 Dialogue Start]
{dialogue2}
[AI Assistant 2 Dialogue End]

Evaluation Points:
AI assistant's suggested reference answer is: "The New Zealand Ministry of Foreign Affairs stated on the 19th that Tonga has established a temporary communication system, but communication is still limited. New Zealand and countries like Fiji plan to provide aid to Tonga." This answer is an ideal response example.
The AI assistant's response must include the following key content (the expression can vary): "Tonga has established a temporary communication system," and the absence of this content will result in the answer being directly judged as incorrect.
Ideally, the AI assistant's response should include the following points: "communication is limited, New Zealand and countries like Fiji plan to provide aid to Tonga," to ensure the comprehensiveness of the answer.

Figure 20: Evaluation Prompt for *Financial Text Summarization*

Test Prompt

Role Prompt:
You are role-playing as a writer.
You are conversing with an AI assistant, hoping it can help generate an asset evaluation report.
The purpose, object, and scope of the evaluation are: information. Your needs are: needs.
Ensure to converse with the AI assistant in the tone of a writer, avoid unnecessary chatter, and try to be as realistic as possible.
If you feel the AI assistant's response meets your needs, you can output the corresponding characters as instructed by the prompt. If not, raise your concerns based on the AI assistant's response.
Note: What you need to do is simulate a user asking the AI assistant questions based on the provided information and needs (if any) rather than answering or solving problems.
You do not need to perform any calculations, analysis, or generate report content. If the AI assistant asks questions or needs additional information, please answer truthfully.
Please start your conversation.

User Intention:
Generate the purpose, object, and scope sections in the asset evaluation report.

Model Prompt:
You are providing document services to a writer. During the service provision process, you can ask the other party for more information. The template must be in the form of an asset appraisal report.

Figure 21: Test Prompt for *Asset Valuation Reporting*

Source Information

The evaluation object is Keda Guochuang Xinneng Technology Co., Ltd., and the reason for evaluating the assets is that Keda Guochuang Software Co., Ltd. is issuing shares to purchase 100% equity of Keda Guochuang Xinneng Technology Co., Ltd. Keda Guochuang Software Co., Ltd. has signed the "Share Issuance and Asset Purchase Agreement" with Hefei Guibo Equity Investment Partnership (Limited Partnership) and Hefei Zixu Investment Partnership (Limited Partnership). The partners of Hefei Guibo Equity Investment Partnership (Limited Partnership) and Hefei Zixu Investment Partnership (Limited Partnership) include Sun Lu, Shi Xingling, Xu Genyi, Chen Xuexiang, Zhang Qiyun, and Dong Xianquan. As of December 31, 2020, the book value of the company's total assets was 461.3236 million yuan, the book value of total liabilities was 161.9956 million yuan, and the book value of net assets was 299.3280 million yuan.

Figure 22: Source Information for *Asset Valuation Reporting*

Evaluation Prompt

Evaluation Criteria:
Please act as a fair judge to assess the quality of the dialogue between the user and the AI assistant. Please read the user requirements and evaluation hints before assessing to help you better analyze the dialogue quality.
The user's needs are: {needs}.
The evaluation hints regarding specific content for your reference are: {evaluation_hints}
When assessing, you also need to consider the following dimensions:
- Meeting user needs: Your evaluation should consider whether the AI assistant's responses comprehensively and appropriately meet the user's needs.
- Accuracy of facts: Is the information provided accurate and based on credible facts and data?
- Fairness and accountability: Are the suggestions or information provided feasible and accountable, and do they consider potential risks and consequences?
- Richness: Does it contain abundant information, depth, contextual considerations, diversity, detailed explanations, and examples to meet user needs and provide comprehensive understanding?
- Hallucination: Are there any hallucinations in the AI assistant's responses?
- Note: Do not let the length of the response affect your scoring! Longer responses are not necessarily better; concise answers that meet the above requirements are good.
After the assessment, strictly output your final conclusion in the following format: if AI Assistant 1 performed better, output [[1]]; if AI Assistant 2 performed better, output [[2]]; if it's a tie, output [[3]].

[AI Assistant 1 Dialogue Start]
{dialogue1}
[AI Assistant 1 Dialogue End]

[AI Assistant 2 Dialogue Start]
{dialogue2}
[AI Assistant 2 Dialogue End]

Evaluation Points:
AI assistant's response must include the following key content (expression can vary): 1. The assessment purpose must include the agreement signing time; 2. The partner's name must be included in the assessment purpose section; 3. The assessment purpose must be a single section; 4. The assessment object and scope must be the second section, and content from different sections should not be confused, missing these contents will result in the answer being directly judged as incorrect.
Encourage the AI assistant to mention the following content in the response: 1. Add background information to enhance the completeness of the report. 2. Provide a detailed explanation of the assessment purpose, including an explanation of the importance or necessity of the assessment purpose, to improve the persuasiveness of the report. 3. Any key factors within the assessment scope that may affect asset value, such as market conditions, industry trends, etc., to provide a more comprehensive assessment perspective, this will help improve the quality of the answer.
The AI assistant's response should avoid including the following content: 1. Avoid including detailed company history or unrelated business introductions that are not related to the assessment purpose, object, and scope in the report. 2. Vague or uncertain language. 3. Ensure that the assessment purpose, object, and scope are each independent, do not mix the content together. Mentioning these contents will result in the answer being judged as inappropriate.

Figure 23: Evaluation Prompt for *Asset Valuation Reporting*