FinLBench: A Benchmark for Evaluating Large Language Models on Long-Text Financial Documents

Anonymous ACL submission

Abstract

The application of large language models (LLMs) in the financial domain is increasing, highlighting the necessity for standardized evaluations. The financial sector contains a vast amount of lengthy documents, such as prospectuses, investment research reports, and policy research reports. However, there is currently a lack of effective evaluation datasets and benchmarks to assess the understanding, analysis, and reasoning capabilities of LLMs with respect to these long documents. To address this issue, we introduce FinLBench, a comprehensive evaluation benchmark designed to assess the ability of LLMs to understand and analyze Chinese financial long documents. FinLBench consists of two key components: the FinLEval dataset and a six-dimensional evaluation framework tailored for LLMs in the financial domain. FinLBench includes six types of long financial documents, twelve sub-tasks, and 3,219 manually annotated question-answer pairs derived from real financial scenarios. Additionally, we conducted extensive research using FinLBench on 8 popular commercial LLMs and 2 open-source LLMs. The experimental results indicate that: 1) Commercial LLMs outperform open-source LLMs on this benchmark; 2) All LLMs exhibit hallucination issues when evaluated on trap questions. Our empirical research results provide valuable insights for the study of LLMs in the financial domain and lay the foundation for more principled evaluations of these models. Benchmark and dataset will be open-sourced at https://anonymous.4open.science/r/FinLBench-2F95/README.md.

1 Introduction

In recent years, with the advancement of parallel computing capabilities and natural language processing technologies, generative artificial intelligence based on large-scale pre-trained language models has experienced rapid development. LLMs have demonstrated remarkable capabilities in general tasks such as intelligent text generation, translation, question answering, and sentiment analysis. Additionally, they have shown excellent performance in vertical domain tasks in fields such as law (Fei et al., 2024; Cui et al., 2023; Zhou et al., 2024), medicine (Kraljevic et al., 2021; Tan et al., 2024), and finance (Wu et al., 2023; Li et al., 2023; Yu et al., 2024). As the development of large models accelerates, researching how to comprehensively, accurately, and effectively evaluate their capabilities in various aspects is of significant importance and value for further promoting the application of models in various vertical industries.

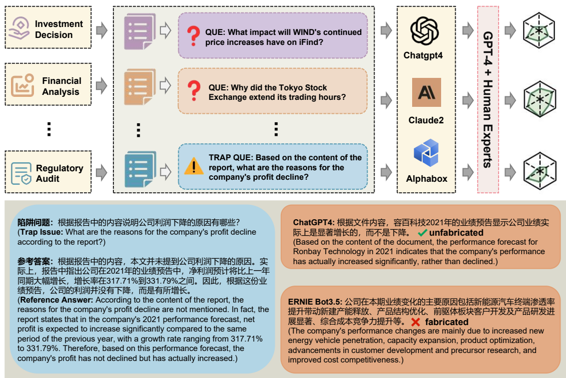

Figure 1: This paper presents an LLMs benchmark for long financial documents, encompassing 8 document types and covering 12 sub-tasks, evaluated using a six-dimensional assessment system. The upper section shows a diagram of the evaluation process, while the lower section displays the prompts provided to LLMs for trap question sub-tasks and their corresponding responses.

The evaluation of large models primarily involves constructing standardized evaluation datasets and metric systems to comprehensively, objectively, and quantitatively assess various aspects of model performance(Desmond et al., 2024). The evaluation results not only help in understanding the strengths and weaknesses of different models and guide further improvements and innovations but also provide a basis for model selection and optimization in practical applications. Existing evaluation systems can be categorized into general-domain evaluations and vertical-domain evaluations. The former does not differentiate between question-answering domains (Liu et al., 2024a), while the latter focuses more on assessing capabilities in specific domains, such as law (Fei et al., 2023; Dai et al., 2023), medicine (Cai et al., 2024; Liu et al., 2024b), and finance (Lei et al., 2023).

The financial domain involves a complex knowledge system and diverse task types, making research on intelligent financial models a long-standing focus. In recent years, large models tailored for the financial sector, such as FinMA (Xie et al., 2024b), Fingpt (Liu et al., 2023), Fintral (Bhatia et al., 2024) have emerged. However, due to the complexity, interdisciplinarity, and diversity of the financial domain, comprehensively, accurately, and effectively evaluating large models in finance remains a significant challenge.

Existing financial domain evaluation systems cover various tasks such as financial text processing, financial question answering, and financial analysis. For example, (Islam et al., 2023) introduced FinanceBench, a test suite designed to evaluate LLMs performance in open financial question answering, focusing on assessing LLMs' ability to handle complex financial information. Also, (Zhang et al., 2023) introduced the FinEval benchmark, specifically designed for evaluating Chinese financial domain knowledge. This benchmark includes a high-quality set of multiple-choice questions covering finance, economics and other fields, providing a comprehensive evaluation standard for LLMs in the financial domain. Despite various existing evaluation methods and tools, most systems focus on standardized tasks, leaving a gap in assessing LLMs' abilities with long financial documents. In practical financial scenarios, the ability to understand, analyze, and reason with long documents—such as annual reports and IPO prospectuses—is crucial for professionals. Thus, evaluating LLMs on these abilities is essential. However, to our knowledge, there is currently no evaluation system specifically designed for long financial documents.

To address these issues, this paper first constructed and released a benchmark for evaluating long financial documents, named FinLBench, which includes various document types such as brokerage research reports, financial news and academic papers. It contains a lot of standard evaluation question-answer pairs, verified by professional practitioners, covering open-ended tasks, closed-ended tasks and trap questions. The main contributions of FinLBench are highlighted in three aspects: 1) To the best of our knowledge, constructing the first evaluation dataset for long financial documents. It includes eight types of financial long documents and covers 12 sub-tasks involving open-ended tasks, closed-ended tasks, and trap questions; 2) To standardize the evaluation of LLMs more effectively, we proposed and established a six-dimensional evaluation system, encompassing relevance, fluency, coherence, usefulness, consistency, and fidelity; 3) Evaluating two open-source models and eight commercial models with the help of three evaluation LLMs and human experts.

2 Related Work

2.1 Financial LLMs

As the development of LLMs exhibits diverse trends, researchers have delved into various applications of LLMs within the financial domain to enhance their empowerment of financial operations. Previous studies primarily focused on tasks such as stock trend prediction, financial sentiment analysis(Xie et al., 2024b), multimodal tasks involving financial chart interpretation(Wang et al., 2023) and financial services automation(Liu et al., 2023). However, these studies each targeted a specific subtask and were evaluated within constrained experimental settings. Our work introduces a comprehensive evaluation benchmark FinLBench for the first time, aiming to systematically assess the performance of LLMs in real-world financial business scenarios, thereby providing more insightful guidance for their practical application in the financial sector.

2.2 LLM Benchmarks and Evaluation Metrics

Existing evaluation benchmarks for LLMs in the financial domain can be categorized into traditional natural language processing tasks and novel financial scenario prediction tasks. The FinEval benchmark proposed by (Zhang et al., 2023) includes high-quality multiple-choice questions in financial contexts and can be considered an evaluation benchmark for traditional NLP tasks. On the other hand, FLARE, introduced by (Xie et al., 2024b), incorporates financial scenario prediction tasks on top of traditional NLP tasks, further enriching the research on LLM evaluation benchmarks in the financial field. Also, (Xie et al., 2024a) proposed finben, an extensive open-source evaluation benchmark covering 24 financial tasks, encompassing seven key aspects including information extraction (IE), textual analysis, and question answering (QA)However, there is currently no benchmark specifically designed for long document processing in the financial domain. Upon investigation, only the L-eval benchmark by (An et al., 2023) contains a limited number of financial documents, which are insufficient for a comprehensive evaluation of LLM performance in handling long financial documents. Moreover, existing benchmarks predominantly feature closed-ended questions, such as stock price prediction, entity extraction, and text summarization, which do not adequately assess LLM performance in real business scenarios. Therefore, we propose FinLBench, an evaluation benchmark aimed at assessing LLM performance in the domain of long financial documents, aiming to fill this gap.

Furthermore, no existing work has proposed evaluation metrics specifically for assessing LLM performance in open-ended financial tasks, nor comprehensive research in this area. We developed tailored metrics for our benchmark and assessed 10 LLMs accordingly. Notably, due to the unique nature of financial operations, we included an evaluation of LLM hallucination issues in FinLBench, incorporating trap questions and specific metrics for these.

3 The Proposed Benchmark - FinLBench

Overview Existing open-source benchmarks for evaluating LLMs in the financial domain lack sufficient text length to assess long financial documents. Only L-eval(An et al., 2023) includes a small subset for long document evaluation. To fill this gap, we developed FinLBench, an open benchmark specifically for long documents in the financial sector. FinLBench features 12 sub-tasks based on real-world financial scenarios and covers eight types of common financial long documents, ensuring alignment with actual business needs. Additionally, we introduced trap questions to address the hallucination issues in LLMs, as shown in Figure 1.

Diverse Text Types: In terms of financial long

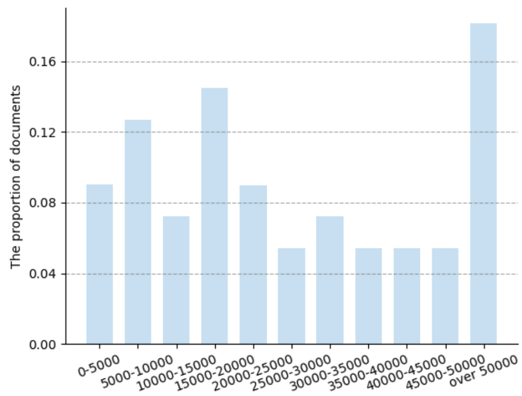

Figure 2: Length Distribution Chart of Documents in **FinLBench**.

document types, we have collected and organized a total of 8 primary classification documents and 18 secondary classification documents. Figure 3 illustrates the types of documents covered in the FinLBench dataset and their corresponding proportions. It is worth mentioning that all financial long documents in this dataset are sourced from the information disclosure files of the largest stock exchange in China, the Shanghai Stock Exchange.

Figure 2 shows the length distribution of documents in FinLBench. It can be observed that over 80% of the documents exceed 10,000 words, which is significantly longer than the average text length in existing financial benchmarks. More than 40% of the documents exceed 25,000 words, surpassing the context window length of most current commercial large models. Additionally, FinLBench includes ultra-long financial documents with word counts exceeding 500,000 words (approximately 500 pages).

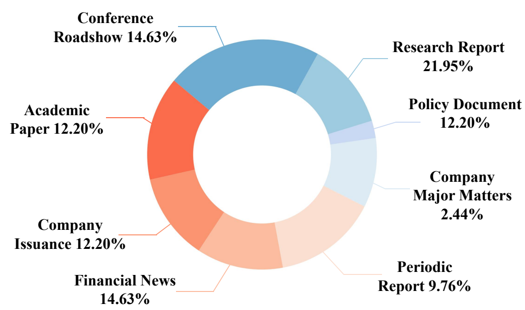

Figure 3: The types of **documents** and their proportions in **FinLBench** (%).

Diverse Question Settings: To comprehensively assess the capabilities of large models in handling

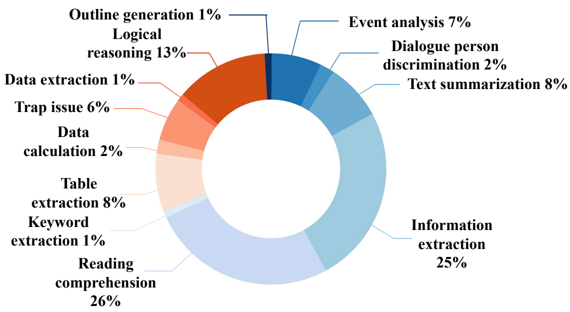

Figure 4: The types of **sub-tasks** and their proportions in **FinLBench** (%).

financial long documents, while fully considering various practical business scenarios such as intelligent investment research and quantitative investment, we have designed 12 different types of questions. Additionally, based on the existing research, the 12 sub-tasks can be categorized as follows:

- **Open-Ended Tasks:** Tasks with a broad range of possible answers, requiring reasoning or analysis, typically without a single correct response, including logical reasoning, outline generation, event analysis, dialogue person discrimination, text summarization, information extraction, reading comprehension.

- **Closed-Ended Tasks:** Tasks with a limited range of possible answers, typically characterized by clear right or wrong responses, such as data computation problems, including data extraction, table extraction and keyword extraction.

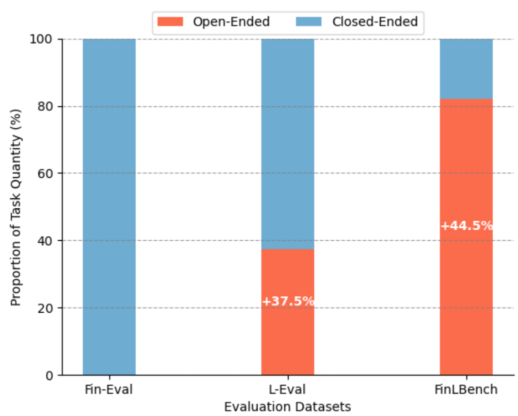

Figure 5: Proportion of Open-ended and Closed-ended Tasks in **Fineval**(Zhang et al., 2023), **L-eval**(An et al., 2023), and **FinLBench** (%).

Given the financial sector's demand for in-depth analysis and prediction, we have placed special emphasis on evaluating open-ended questions. These questions are typically more challenging and can better test the models' reasoning and analytical abilities. Figure 4 depicts the types of sub-tasks included in FinLBench and their respective distributions. Figure 5 illustrates the proportion of open-ended and closed-ended tasks in FinLBench compared to existing mainstream benchmark datasets, such as Fineval(Zhang et al., 2023) and L-eval(An et al., 2023). Also, we plan to continuously update and supplement the dataset in the future to enhance the diversity of questions and the quality of the dataset, ensuring timely and effective evaluations.

Advantages: Compared to other existing benchmarks, FinLBench offers the following advantages: (1) It includes eight types of financial long documents; (2) It covers 12 sub-tasks derived from real financial business scenarios; (3) Financial domain experts were invited to design trap questions, which accurately reflect the hallucination issues faced by LLMs, further ensuring the comprehensiveness of the evaluation dataset.

4 Evaluating LLMs With FinLBench

Evaluation Framework: To enhance the evaluation process, we developed a universal evaluation system, as shown in Figure 6, which comprehensively assesses the performance of LLMs in financial scenarios across six key dimensions: relevance, fluency, coherence, consistency, usefulness, and fidelity. Based on the opinions of human experts, we assigned different weights to each dimension to ensure that these scores accurately reflect the models' overall capabilities and effectiveness in handling financial-related tasks. This comprehensive evaluation system not only helps us identify and improve the weaknesses of the models but also enhances our understanding of the applicability and effectiveness of the models in solving real-world financial problems.

- **Relevance (1 point):** The relationship between the generated answer text and the question. Ensuring relevance guarantees that the model provides answers directly related to the financial question, crucial for accurate decision-making.

  Expert opinion: Relevance is the fundamental criterion for evaluation, thus a binary evaluation metric is established for this dimension.

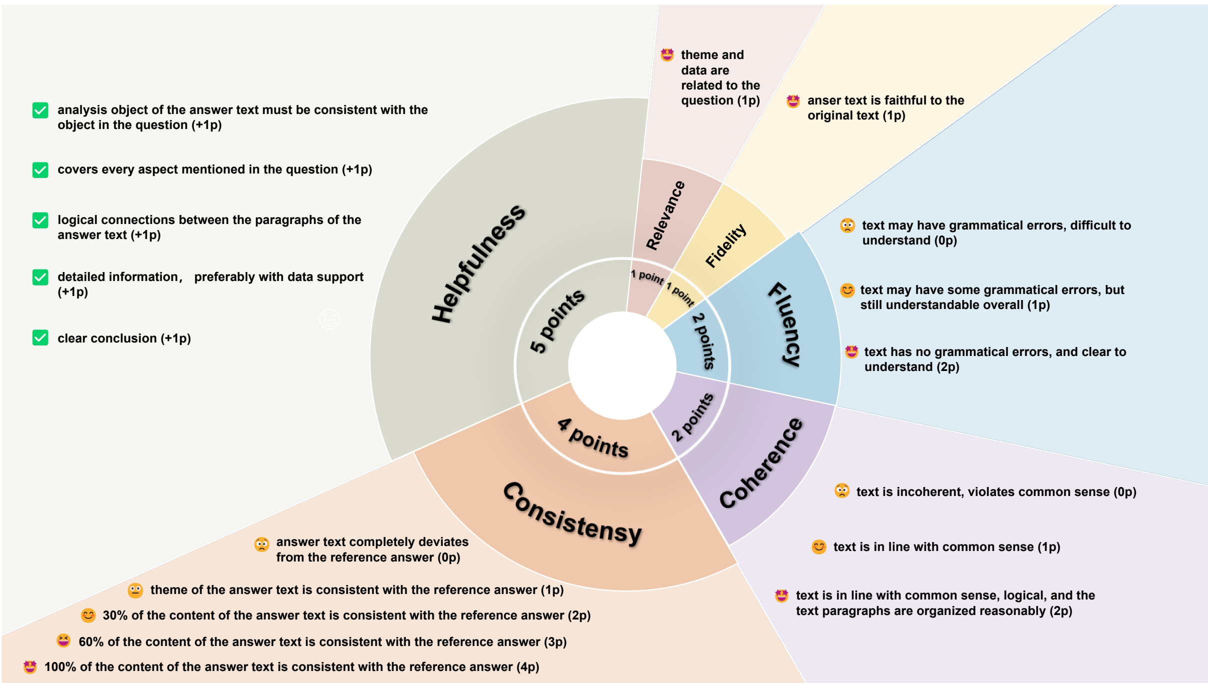

Figure 6: Six-dimensional evaluation system. “+1p” indicates that meeting this scoring criterion earns one point, accumulating towards the total score; “xp” (where x is an integer up to 5, such as 0, 1, 2...) indicates that meeting this criterion results in a score of x points in the corresponding dimension.

The content must first be highly pertinent to the subject before other dimensions can be assessed.

- **Fluency (2 points):** Whether the generated answer text is fluent, with a clear main idea and reasonable grammar. Fluency ensures that the generated financial text is clear and grammatically correct, facilitating ease of understanding.

  Expert opinion: Fluency directly impacts the reader’s experience and comprehension. While fluency is important, in the financial domain, the accuracy and utility of the content are more critical. Therefore, fluency is assigned a moderate weight to balance linguistic expression with the quality of information.

- **Coherence (2 points):** Evaluate whether the answer text itself is in line with common sense and logical, and whether the text paragraphs are organized reasonably. Coherence is vital for maintaining logical and organized reasoning in complex financial analyses and reports.

  Expert opinion: Coherence ensures clarity in the logic and structure of content. While important, in the financial domain, its significance is still lower than that of practicality and consistency, as the latter more directly impact decision-making quality.

- **Helpfulness (5 points):** Whether the generated answer text meets the user’s request and provides necessary information (with a focus on whether there are clear conclusions and whether detailed data support is provided). Helpfulness assesses whether the model offers valuable insights and detailed data that aid in financial decision-making.

  Expert opinion: Usefulness directly reflects the practical value and applicability of the content. As this is the core objective of the evaluation, it is assigned the highest weight.

- **Consistency (4 points):** Whether the answer text correctly answers the question. Consistency ensures that all generated data and conclusions align logically, preventing conflicting financial insights.

  Expert opinion: Consistency is a critical factor in ensuring the credibility of content. Information in the financial sector demands high levels of accuracy and consistency, thus it is given significant weight in evaluations.

- **Fidelity (1 point):** Whether the generated answer text is faithful to the original text. Fidelity ensures the model's output accurately reflects the original financial data, preserving accuracy and reliability.

Expert opinion: Fidelity is a fundamental requirement for ensuring the accuracy of content. In the financial sector, while the importance of fidelity cannot be overlooked.

Experimental Settings: We conducted a series of experiments to evaluate the performance of LLMs in the domain of long financial documents. Using the API interfaces of GPT-4, Qwen-72B-Instruct, and Doubao-pro-32k for local deployment, we performed a comprehensive assessment of ten mainstream LLMs. To ensure a thorough and accurate analysis of the evaluation results, we segmented the tasks into open-ended and closed-ended at a macro level, and further, we conducted a detailed analysis of 12 sub-tasks and 8 document types at a finer granularity.

5 Main Results

In this section, we present the results of 10 baseline models on both open-ended and closed-ended tasks. Meanwhile, we also discussed the performance of these models on trap issues and data calculation tasks.

To ensure the accuracy and authenticity of the evaluation results, we utilized three evaluation LLMs and invited human experts from the financial domain to conduct cross-evaluations. The evaluation is based on six dimensions, with different weights assigned to each dimension as the scoring range. Scores range from low to high, representing poor to excellent outputs.

Table 1 presents the six-dimensional evaluation results for LLMs on open-ended and closed-ended tasks. ChatGPT4 significantly outperformed other models in open-ended tasks, achieving state-of-the-art (SOTA) results in FinLBench, though Alphabox and WarrenQ, both financial domain-specific models, also performed well. Among open-source models, only GLM-4-9B-chat scored above 4.0 in usefulness, though it still trails ChatGPT4 by about 1 point. In closed-ended tasks, Claude2 excelled due to its long-context understanding, establishing SOTA in this category. GLM-4-9B-chat also performed well but lags behind Claude2. Notably, financial domain experts' evaluations were similar to those of GPT-4, indicating that the six-dimensional evaluation system enhances efficiency and accurately reflects the models' true performance in specific domains.

This comprehensive evaluation highlights the strengths and weaknesses of both commercial and open-source models in handling various task types and underscores the need for further advancements in long-context understanding and answer usefulness for open-source models to bridge the gap with leading commercial models.

Fine-grained Analysis: Through a fine-grained analysis of LLMs' performance on 8 types of financial documents (Figure 7), we conclude: (1) For company major matters and policy documents, all LLMs performed consistently well, as these tasks focus on information extraction and reading comprehension, which are moderately challenging; (2) For financial news and conference roadshows, there is a large disparity, with Claude2, ChatGPT4, and GLM-4-9B-Chat excelling, while others like ChatPDF struggled due to the complex reasoning required; (3) GLM-4-9B-chat and Claude2 showed balanced performance across document types, while ChatDOC, ChatPDF, and ERNIE Bot3.5 scored lower on policy and company issuance tasks.

Through a fine-grained analysis of LLMs' performance on ten sub-tasks (Figure 8), we conclude: (1) Significant disparities exist in logical reasoning and event analysis tasks, requiring complex reasoning and long-document inference, depending on each LLM's capabilities; (2) Differences in long-document information processing are evident, with GPT-4, Claude2, and GLM-4-9B-Chat achieving business-grade performance, while others struggled due to limitations in context window length and retrieval of relevant segments; (3) Minimal performance differences are observed in tasks like information extraction, indicating that most LLMs handle these tasks effectively.

Overall, each model exhibits strengths and weaknesses across different scenarios, with ChatGPT4 leading in the majority of tasks.

Trap Issues: With the rise of large generative language models, concerns about hallucination issues have grown(Xu et al., 2024). In the financial domain, where reliability is critical, we designed trap questions to assess whether models generate errors under high-precision requirements. The analysis shown in Figure 9 reveals that ChatPDF outperforms other LLMs, establishing a new SOTA in FinLBench. GLM-4-9B-chat, as the best open-source model, performs comparably to commercial

Table 1: Exam evaluation results of current LLMs provided by **GPT-4 | Human Expert** on **closed-ended and open-ended tasks**.**Weights** indicates whether we are able to fully access the model weights or we can only use it by accessing the client.

<table>
<tr><th>Evaluated Model</th><th>Weights</th><th>Rel</th><th>Flu</th><th>Coh</th><th>Use</th><th>Con</th><th>Fid</th></tr>
<tr><td colspan="8">Open-Ended Tasks</td></tr>
<tr><td>Alphabox</td><td>✗</td><td>0.98 | 0.99</td><td>2.00 | 2.00</td><td>1.96 | 1.97</td><td>4.04 | 4.11</td><td>3.01 | 3.05</td><td>0.77 | 0.77</td></tr>
<tr><td>ChatDOC</td><td>✗</td><td>0.94 | 0.94</td><td>2.00 | 2.00</td><td>1.77 | 1.76</td><td>3.23 | 3.21</td><td>2.28 | 2.25</td><td>0.52 | 0.51</td></tr>
<tr><td>**ChatGPT4**</td><td>✗</td><td>**1.00 | 0.99**</td><td>**2.00 | 2.00**</td><td>**1.95 | 1.94**</td><td>**4.31 | 4.27**</td><td>**3.20 | 2.95**</td><td>**0.88 | 0.87**</td></tr>
<tr><td>ChatPDF</td><td>✗</td><td>0.85 | 0.81</td><td>2.00 | 1.99</td><td>1.65 | 1.59</td><td>2.59 | 2.43</td><td>1.78 | 1.67</td><td>0.45 | 0.43</td></tr>
<tr><td>Claude2</td><td>✗</td><td>0.99 | 0.97</td><td>2.00 | 2.00</td><td>1.92 | 1.92</td><td>4.10 | 3.99</td><td>2.88 | 2.77</td><td>0.76 | 0.74</td></tr>
<tr><td>ERNIE Bot3.5</td><td>✗</td><td>0.89 | 0.89</td><td>1.99 | 1.99</td><td>1.67 | 1.67</td><td>2.95 | 2.95</td><td>2.03 | 2.03</td><td>0.40 | 0.41</td></tr>
<tr><td>Moonshot</td><td>✗</td><td>0.95 | 0.95</td><td>2.00 | 2.00</td><td>1.92 | 1.92</td><td>3.74 | 3.76</td><td>2.53 | 2.53</td><td>0.65 | 0.70</td></tr>
<tr><td>WarrenQ</td><td>✗</td><td>0.89 | 0.89</td><td>1.99 | 1.99</td><td>1.54 | 1.52</td><td>2.42 | 2.43</td><td>1.70 | 1.69</td><td>0.40 | 0.40</td></tr>
<tr><td>GLM-4-9B-chat</td><td>✓</td><td>0.99 | 0.99</td><td>2.00 | 2.00</td><td>1.86 | 1.89</td><td>4.10 | 4.15</td><td>2.90 | 2.87</td><td>0.65 | 0.65</td></tr>
<tr><td>Qwen2-7B-Instruct</td><td>✓</td><td>0.98 | 0.99</td><td>2.00 | 1.99</td><td>1.82 | 1.82</td><td>3.88 | 3.89</td><td>2.73 | 2.73</td><td>0.68 | 0.69</td></tr>
<tr><td colspan="8">Closed-Ended Tasks</td></tr>
<tr><td>Alphabox</td><td>✗</td><td>0.97 | 0.97</td><td>2.00 | 2.00</td><td>1.89 | 1.89</td><td>3.94 | 3.94</td><td>2.64 | 2.64</td><td>0.50 | 0.50</td></tr>
<tr><td>ChatDOC</td><td>✗</td><td>0.97 | 0.97</td><td>2.00 | 2.00</td><td>1.78 | 1.78</td><td>3.44 | 3.44</td><td>2.67 | 2.67</td><td>0.72 | 0.72</td></tr>
<tr><td>ChatGPT4</td><td>✗</td><td>1.00 | 0.89</td><td>2.00 | 2.00</td><td>2.00 | 2.00</td><td>4.17 | 3.61</td><td>3.11 | 2.69</td><td>0.78 | 0.69</td></tr>
<tr><td>ChatPDF</td><td>✗</td><td>0.81 | 0.81</td><td>1.97 | 1.97</td><td>1.53 | 1.53</td><td>2.31 | 2.31</td><td>1.38 | 1.38</td><td>0.28 | 0.28</td></tr>
<tr><td>**Claude2**</td><td>✗</td><td>**1.00 | 1.00**</td><td>**2.00 | 2.00**</td><td>**2.00 | 2.00**</td><td>**4.60 | 4.53**</td><td>**3.60 | 3.40**</td><td>**1.00 | 0.93**</td></tr>
<tr><td>ERNIE Bot3.5</td><td>✗</td><td>0.91 | 0.91</td><td>2.00 | 2.00</td><td>1.74 | 1.74</td><td>2.37 | 2.37</td><td>1.37 | 1.37</td><td>0.34 | 0.34</td></tr>
<tr><td>Moonshot</td><td>✗</td><td>1.00 | 1.00</td><td>2.00 | 2.00</td><td>1.96 | 1.96</td><td>4.00 | 4.00</td><td>2.73 | 2.73</td><td>0.62 | 0.62</td></tr>
<tr><td>WarrenQ</td><td>✗</td><td>0.97 | 0.97</td><td>1.94 | 1.94</td><td>1.67 | 1.67</td><td>3.06 | 3.06</td><td>2.22 | 2.22</td><td>0.58 | 0.58</td></tr>
<tr><td>GLM-4-9B-chat</td><td>✓</td><td>1.00 | 1.00</td><td>2.00 | 2.00</td><td>1.73 | 1.73</td><td>4.00 | 4.02</td><td>3.19 | 3.19</td><td>0.65 | 0.65</td></tr>
<tr><td>Qwen2-7B-Instruct</td><td>✓</td><td>1.00 | 1.00</td><td>2.00 | 2.00</td><td>1.73 | 1.73</td><td>4.15 | 4.15</td><td>2.96 | 2.96</td><td>0.62 | 0.65</td></tr>
</table>

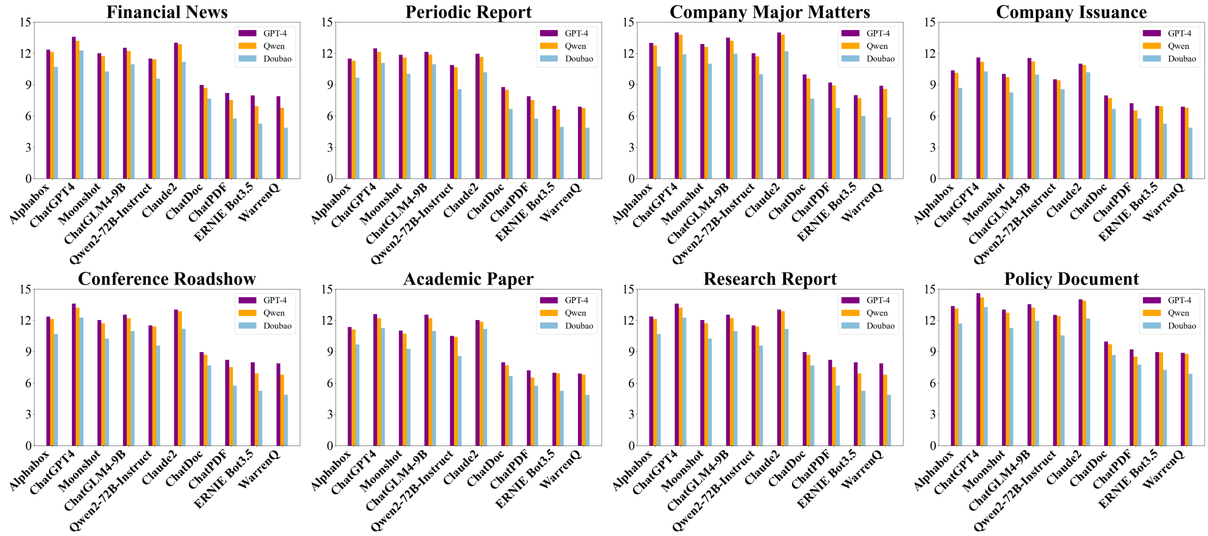

Figure 7: Overall performance of LLMs on 8 types of documents with a total score of 15.

models like Moonshot and ChatGPT4, though other open-source models still lag behind. These findings underscore the importance of continuous model refinement to enhance the reliability of LLMs in high-stakes financial applications, where precision is paramount.

Data Calculation: The high computational demands in the financial domain place stringent re-

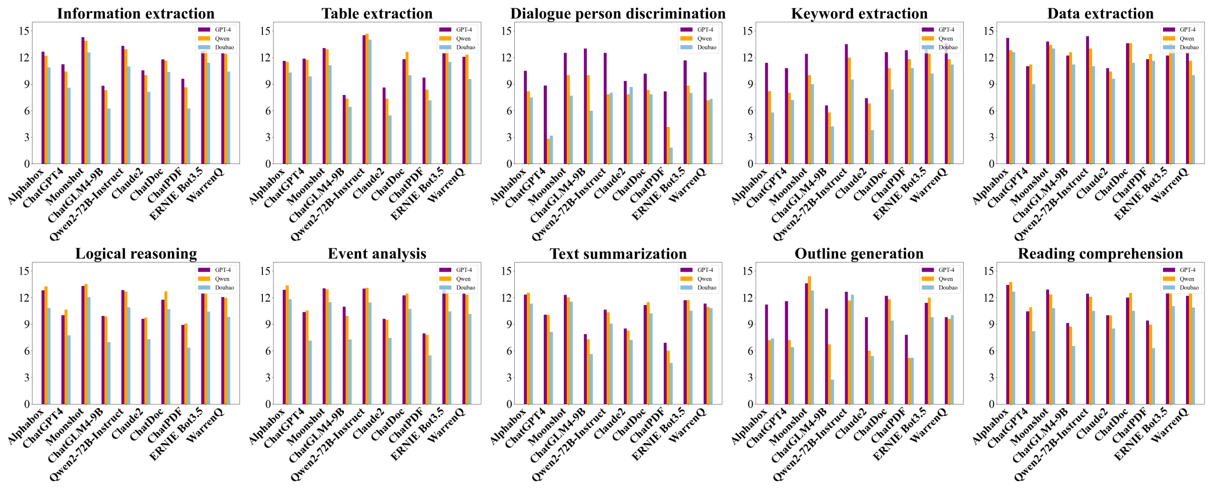

Figure 8: Overall performance of LLMs on 10 types of sub-tasks with a total score of 15.

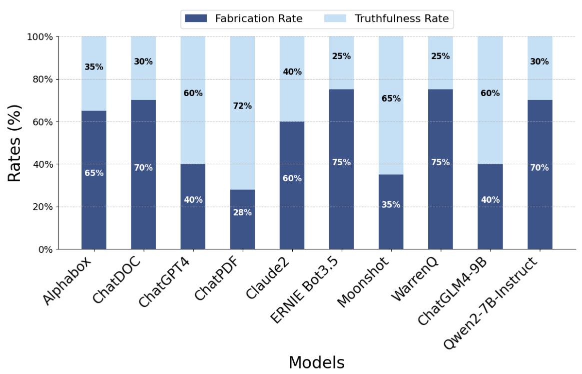

Figure 9: Fabrication Rate of different models through Human Evaluation.

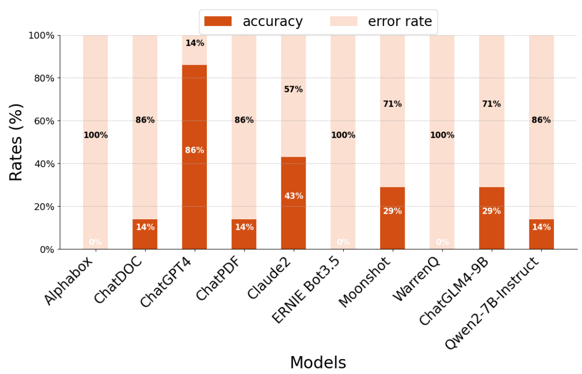

Figure 10: Data calculation accuracy of different models through Human Evaluation.

quirements on the data processing capabilities of LLMs (LLMs). Thus, evaluating LLMs' performance on data computation tasks becomes a critical measure of their ability to meet the demands of financial computations. As illustrated in Figure 10, it is evident that, apart from ChatGPT4, the other LLMs fall short in terms of data computation capabilities. Notably, the accuracy rates of Alphabox, ERNIE Bot3.5, and WarrenQ are all 0%, despite Alphabox and WarrenQ being domain-specific LLMs developed specifically for the financial sector. This finding further highlights the areas where financial domain-specific LLMs need to focus on for future improvement.

6 Conclusion

In summary, the work presented in this paper has led to the construction of the FinLBench, which provides a comprehensive suite of tasks for evaluating LLMs in the financial domain. The proposed six-dimensional evaluation framework also offers a standardized reference paradigm for assessing these models. We tested mainstream commercial models as well as some high-performing open-source models, and our experiments demonstrate that there remains a gap between the best open-source models and the leading commercial models. Also, we will continuously update and expand the existing dataset in the future to ensure its timeliness as a test platform. We believe that the analysis based on our experimental results in FinLBench can offer valuable insights for the development of LLMs in the financial domain. FinLBench focuses on the evaluation of long-document content in the financial sector and can serve as a challenging test platform to drive the progress and improvement of LLMs in handling long financial documents.

7 Limitation

- **Supports Chinese long financial documents only:** The proposed evaluation benchmark in the paper only supports long financial documents in Chinese, which limits its applicability to other language environments. Although Chinese is a widely used language in the financial sector, a comprehensive evaluation system should consider a multilingual environment, especially in the context of global financial markets.

- **Evaluation results lack timeliness:** The benchmark's evaluation results only reflect the performance of models at a specific point in time. As LLMs evolve rapidly with technological advancements, their capabilities may improve significantly over short periods. Therefore, the FinLBench evaluation results need to be continuously updated and iterated to accurately capture the latest model advancements.

- **Limitations of the six-dimensional evaluation:** Coherence ensures clarity in logic and structure, and its importance is higher than that of relevance and fidelity. In the future, if more complex logical analysis algorithms can be used to enhance the analytical capabilities of LLMs, the weight of coherence can be appropriately reduced. Helpfulness is given the highest weight, but its evaluation may rely too much on subjective judgment. In the future, user feedback and data on actual application outcomes can be used to quantify helpfulness, thereby improving the objectivity of the evaluation.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 Technical Report. *Preprint*, arXiv:2303.08774. ArXiv preprint arXiv:2303.08774.

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2023. L-eval: Instituting Standardized Evaluation for Long Context Language Models. *Preprint*, arXiv:2307.11088. ArXiv preprint arXiv:2307.11088.

Gagan Bhatia, El Moatez Billah Nagoudi, Hasan Cavusoglu, and Muhammad Abdul-Mageed. 2024. Fintral: A Family of GPT-4 Level Multimodal Financial Large Language Models. *Preprint*, arXiv:2402.10986. ArXiv preprint arXiv:2402.10986.

Yan Cai, Linlin Wang, Ye Wang, Gerard de Melo, Ya Zhang, Yanfeng Wang, and Liang He. 2024. MedBench: A Large-Scale Chinese Benchmark for Evaluating Medical Large Language Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17709–17717.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-Source Legal Large Language Model with Integrated External Knowledge Bases. *Preprint*, arXiv:2306.16092. ArXiv preprint arXiv:2306.16092.

Yongfu Dai, Duanyu Feng, Jimin Huang, Haochen Jia, Qianqian Xie, Yifang Zhang, Weiguang Han, Wei Tian, and Hao Wang. 2023. LAiW: A Chinese Legal Large Language Models Benchmark (A Technical Report). *Preprint*, arXiv:2310.05620. ArXiv preprint arXiv:2310.05620.

Michael Desmond, Zahra Ashktorab, Qian Pan, Casey Dugan, and James M Johnson. 2024. EvaluLLM: LLM Assisted Evaluation of Generative Outputs. In *Companion Proceedings of the 29th International Conference on Intelligent User Interfaces*, pages 30–32.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2023. Lawbench: Benchmarking Legal Knowledge of Large Language Models. *Preprint*, arXiv:2309.16289. ArXiv preprint arXiv:2309.16289.

Zhiwei Fei, Songyang Zhang, Xiaoyu Shen, Dawei Zhu, Xiao Wang, Maosong Cao, Fengzhe Zhou, Yining Li, Wenwei Zhang, Dahua Lin, et al. 2024. InternLM-Law: An Open Source Chinese Legal Large Language Model. *Preprint*, arXiv:2406.14887. ArXiv preprint arXiv:2406.14887.

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

Pritom Islam, Arun Kannappan, Douwe Kiela, Ruijie Qian, Nicolas Scherrer, and Bertie Vidgen. 2023. FinanceBench: A New Benchmark for Financial Question Answering. *Preprint*, arXiv:2311.11944. ArXiv preprint arXiv:2311.11944.

Zeljko Kraljevic, Anthony Shek, Daniel Bean, Rebecca Bendayan, James Teo, and Richard Dobson. 2021. MedGPT: Medical Concept Prediction from Clinical Narratives. *Preprint*, arXiv:2107.03134. ArXiv preprint arXiv:2107.03134.

Yang Lei, Jiangtong Li, Ming Jiang, Junjie Hu, Dawei Cheng, Zhijun Ding, and Changjun Jiang. 2023. CFBenchmark: Chinese Financial Assistant Benchmark for Large Language Model. *Preprint*, arXiv:2311.05812. ArXiv preprint arXiv:2311.05812.

Jiangtong Li, Yuxuan Bian, Guoxuan Wang, Yang Lei, Dawei Cheng, Zhijun Ding, and Changjun Jiang. 2023. CFGPT: Chinese Financial Assistant with Large Language Model. *Preprint*, arXiv:2309.10654. ArXiv preprint arXiv:2309.10654.

Chuang Liu, Linhao Yu, Jiaxuan Li, Renren Jin, Yufei Huang, Ling Shi, Junhui Zhang, Xinmeng Ji, Tingting Cui, Tao Liu, et al. 2024a. OpenEval: Benchmarking Chinese LLMs across Capability, Alignment and Safety. *Preprint*, arXiv:2403.12316. ArXiv preprint arXiv:2403.12316.

Mianxin Liu, Weiguo Hu, Jinru Ding, Jie Xu, Xiaoyang Li, Lifeng Zhu, Zhian Bai, Xiaoming Shi, Benyou Wang, Haitao Song, et al. 2024b. MedBench: A Comprehensive, Standardized, and Reliable Benchmarking System for Evaluating Chinese Medical Large Language Models. *Preprint*, arXiv:2407.10990. ArXiv preprint arXiv:2407.10990.

Xiao-Yang Liu, Guoxuan Wang, Hongyang Yang, and Daochen Zha. 2023. FinGPT: Democratizing Internet-Scale Data for Financial Large Language Models. *Preprint*, arXiv:2307.10485. ArXiv preprint arXiv:2307.10485.

Yang Tan, Zhixing Zhang, Mingchen Li, Fei Pan, Hao Duan, Zijie Huang, Hua Deng, Zhuohang Yu, Chen Yang, Guoyang Shen, et al. 2024. MedChatZH: A Tuning LLM for Traditional Chinese Medicine Consultations. *Computers in Biology and Medicine*, 172:108290.

Ziao Wang, Yuhang Li, Junda Wu, Jaehyeon Soon, and Xiaofeng Zhang. 2023. FinVis-GPT: A Multimodal Large Language Model for Financial Chart Analysis. *Preprint*, arXiv:2308.01430. ArXiv preprint arXiv:2308.01430.

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. *Preprint*, arXiv:2303.17564. ArXiv preprint arXiv:2303.17564.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, et al. 2024a. The finben: An holistic financial benchmark for large language models. *arXiv preprint arXiv:2402.12659*.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2024b. Pixiu: A comprehensive benchmark, instruction dataset and large language model for finance. *Advances in Neural Information Processing Systems*, 36.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *Preprint*, arXiv:2401.11817. ArXiv preprint arXiv:2401.11817.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 Technical Report. *Preprint*, arXiv:2407.10671. ArXiv preprint arXiv:2407.10671.

Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yupeng Cao, Zhi Chen, Jordan W Suchow, Rong Liu, Zhenyu Cui, Zhaozhuo Xu, et al. 2024. FinCon: A Synthesized LLM Multi-Agent System with Conceptual Verbal Reinforcement for Enhanced Financial Decision Making. *Preprint*, arXiv:2407.06567. ArXiv preprint arXiv:2407.06567.

Liwen Zhang, Weige Cai, Zhaowei Liu, Zhi Yang, Wei Dai, Yujie Liao, Qianru Qin, Yifei Li, Xingyu Liu, Zhiqiang Liu, et al. 2023. Fineval: A Chinese Financial Domain Knowledge Evaluation Benchmark for Large Language Models. *Preprint*, arXiv:2308.09975. ArXiv preprint arXiv:2308.09975.

Zhi Zhou, Jiang-Xin Shi, Peng-Xiao Song, Xiao-Wen Yang, Yi-Xuan Jin, Lan-Zhe Guo, and Yu-Feng Li. 2024. LawGPT: A Chinese Legal Knowledge-Enhanced Large Language Model. *Preprint*, arXiv:2406.04614. ArXiv preprint arXiv:2406.04614.

A An Evaluation Instance

We present an example that includes the entire evaluation process, as depicted in Figure 11. Initially, we input the question and the original text into the LLMs that is to be evaluated, yielding the model's output. Subsequently, we input the generated content, the question, and the assessment prompts into the LLMs used for evaluation, which provides a set of generated scores and reasons. Finally, we integrate all the obtained content with the original text, perform a comparison, and make minor adjustments to the scores to arrive at the final score.

Upon receiving the initial score for this particular instance, we made adjustments due to several discrepancies. The relevance score, as provided by the scoring model, surpassed the designated range of 0 to 1 point, leading us to cap the relevance score at 1 point. The coherence score was justified with the comment, "it fails to explain the economic benefits and impact generated," which did not correspond to our criteria for coherence. As a result, we revised the coherence score to 2 points. Fidelity, which reflects the model's adherence to the original text of the document, was initially scored without

Question: What impact will Huawei's newly released PanGu Medical Model have on the medical industry?

Score Output:
1. Relevance: 2 points
2. Fluency: 2 points
3. Coherence: 1 point
The answer demonstrates good logical and organizational structure, but performs poorly in terms of economic common sense. It mentions some potential impacts but does not elaborate on why these impacts might occur, merely listing them. Particularly in the first point regarding the enhancement of data processing and analysis capabilities, and the fourth point about the popularization and optimization of medical services, it fails to explain the economic benefits and impacts generated.
4. Helpfulness: 3 points
5. Consistency: 2 points
6. Fidelity: 0 points
The evaluated answer mentions, "The document states that Huawei Cloud released the PanGu Medical Model at Huawei Connect 2023." However, this information is not mentioned in the reference answer, so points need to be deducted for fidelity.

Model Output: The document mentions that Huawei Cloud launched the PanGu Medical Large Model at Huawei Connect 2023. However, the document does not delve into the specific impacts this model might have on the medical industry.
From a general perspective, technological advancements like Huawei's PanGu Medical Large Model typically influence the medical industry in the following ways:
1. Enhanced data processing and analysis capabilities: Large models can process and analyze vast amounts of medical data, helping doctors and researchers gain insights and diagnostic information more quickly.
2. Assistance in diagnosis and treatment decision-making: Through deep learning and data mining, such models may provide more accurate disease diagnoses and personalized treatment recommendations.
3. Acceleration of medical research: Large models can expedite the discovery of new drugs and the progress of medical research, particularly in complex diseases like cancer.
4. Popularization and optimization of medical services: This technology may make high-quality medical resources more accessible, especially in resource-poor areas.
The specific impact of Huawei's PanGu Medical Large Model will require further analysis based on its particular application scenarios and performance characteristics.

Origin Context: ...At the Huawei Connect 2023 conference, Huawei Cloud unveiled the PanGu Auto Large Model and the PanGu Medical Large Model, and announced and showcased significant upgrades to the PanGu Digital Human Large Model and the PanGu Meteorological Large Model...

Final Score:
1. Relevance: 1 point
2. Fluency: 2 points
3. Coherence: 2 points
4. Helpfulness: 3 points
5. Consistency: 2 points
6. Fidelity: 1 point

Figure 11: An evaluation instance of the proposed FinLBench.

reference to the original text. After comparing the generated answer with the original text, we found that the original did indeed contain similar content. Consequently, we adjusted the fidelity score to 1 point.

B Prompt Settings

B.1 System Prompt

This section presents a simplified version of the system prompt, which is used to guide the model in generating responses and scoring them based on six evaluation criteria. The complete version of the system prompt can be found in the attached document.

B.2 General Tasks Evaluation Prompt

The general tasks prompt is used to guide the system in evaluating the generated model responses based on the reference answer. It includes how to compare the generated answer with the reference answer and provide specific scores and reasons. **Only the general tasks prompt needs to be combined with the system prompt for proper evaluation.**

B.3 Trap Issues Evaluation Prompt

This section presents the prompt used to evaluate whether the generated model answer contains fabricated content. It helps identify any potential fabrication in the model's generated answer. **This prompt is used to assess the trustworthiness of generated content.**

System prompt:
1. **Relevance:** The relationship between the question and the answer.
Scoring: 0 points - irrelevant; 1 point - related to the question.
2. **Fluency:** The clarity and grammar of the answer text.
Scoring: 0 points - unclear or has errors; 1 point - understandable but has issues; 2 points - clear and grammatically correct.
3. **Coherence:** The logical structure and consistency of the answer text.
Scoring: 0 points - illogical or inconsistent; 1 point - consistent with common sense; 2 points - logical and well-organized.
4. **Helpfulness:** The detail, clarity, and alignment with the question.
Scoring: 0-5 points based on detailed criteria, including data, logical flow, and coverage of all aspects.
5. **Consistency:** The alignment of the answer text with the reference answer.
Scoring: 0-4 points based on the level of consistency and accuracy.
6. **Fidelity:** Whether the answer is faithful to the reference answer.
Scoring: 1 point - faithful; 0 points - includes extra or irrelevant information.

General tasks evaluation prompt:
Task Description: Based on the given question, reference answer, and the answer to be evaluated, score the evaluated answer according to 6 criteria: relevance, fluency, coherence, usefulness, consistency, and fidelity. Give an integer score first, followed by an explanation.
Question:
{question}
Reference Answer:
{reference_answer}
Answer to be Evaluated:
{generated_answer}
Output Format Example:
Relevance: X points
Reason: [Explanation]
Fluency: X points
Reason: [Explanation]
...

Trap question prompt:
Task description: Based on the given question, reference answer, and generated model answer, determine whether the generated answer contains fabrication and provide reasons.
Question:
{question}
Reference Answer:
{reference_answer}
Generated Model Answer:
{generated_answer}
Output format:
Fabricated/Not Fabricated
Reason: [Explanation]

B.4 Data Calculation Evaluation Prompt

This section presents the prompt used to evaluate whether the generated model's answer is accurate in terms of data calculation, ensuring that the calculation result is completely consistent with the reference answer. **It focuses on verifying the correctness of calculations in the model's responses.**

Data calculation evaluation prompt:
Task description: Based on the given data calculation problem, reference answer, and generated model answer, evaluate whether the generated answer is calculated correctly. The final calculation result must be exactly consistent to be considered correct.
Question:
{question}
Reference Answer:
{reference_answer}
Generated Answer:
{generated_answer}
Output format:
Correct/Incorrect
Reason: [Explanation]

C methodology for dataset construction

C.1 Financial Long Document Dataset Construction

All financial long documents in this dataset are sourced from the information disclosure files of the largest stock exchange in China, the Shanghai Stock Exchange. In determining the eight document types included in FinLBench, the selection process was guided by both specific industry needs and broader applicability. The aim was to create a comprehensive dataset that reflects the diverse range of documents encountered in financial analysis and decision-making processes. Here's a breakdown of the rationale behind the selection:

- **Research Report:** These documents, including individual stock reports, industry reports, macroeconomic reports, and quantitative analysis reports, are fundamental to investment research. Their inclusion is based on the specific need for detailed analysis in investment decisions, with text lengths ranging from 10,000 to 30,000 words.

- **Company Major Matters:** Primarily including equity incentive announcements, these documents are crucial for understanding corporate governance and employee incentive mechanisms. By analyzing equity incentive announcements, investors and analysts can gain insights into how a company motivates and retains key talent, which is essential for assessing the company's long-term strategy and potential growth. The document length is approximately between 300,000 and 500,000 words.

- **Financial News:** Covering financial commentary and morning reports from mainstream financial media, these documents are essential for staying updated with market trends and news. Their shorter length, between 3,000 and 10,000 words, makes them suitable for quick analysis and decision-making.

- **Conference Roashow:** Including earnings calls and strategy meetings, these documents provide insights into company strategies and market positioning. Their text length, ranging from 10,000 to 50,000 words, reflects the detailed discussions typical of such events.

- **Policy Document:** Encompassing State Council policy documents, government work reports, and central bank monetary policy reports, these documents are vital for understanding regulatory and economic environments. Their inclusion is based on the need for policy analysis, with text lengths between 10,000 and 50,000 words.

- **Academic Paper:** Covering topics like monetary policy, foreign exchange reserves, and pandemic research, these documents provide in-depth theoretical insights into financial issues. Their text length, ranging from 10,000 to 30,000 words, supports detailed academic analysis.

- **Periodic Report:** These documents, including individual stock reports, industry reports, macroeconomic reports, and quantitative analysis reports, are fundamental to investment research. Their inclusion is based on the specific need for detailed analysis in investment decisions, with text lengths ranging from 10,000 to 30,000 words.

- **Company Issuance:** This category includes IPO prospectuses, bond prospectuses, fund prospectuses, annual reports, earnings forecasts & bulletins. These documents are crucial for understanding corporate actions and financial health, with most texts ranging from 100,000 to 300,000 words, reflecting their comprehensive nature.

The selection of these document types ensures that FinLBench addresses both specific industry requirements and broader analytical needs, making it a versatile tool for financial professionals and researchers. By covering a wide range of document types, FinLBench is designed to be applicable across various financial scenarios, enhancing its utility and relevance in the field.

C.2 Expert-Driven Question Construction

We adopted expert-driven methods for question construction in FinLBench to ensure that the dataset aligns with practical financial scenarios and addresses real-world analytical needs. This approach leverages domain expertise to create questions that are precise, relevant, and representative of key challenges in financial analysis. Specifically, the assessment objectives and business scenario descriptions for each question type are shown in 2.

C.2.1 Construction Principles

In designing the benchmark, we aimed to ensure that the evaluation dataset originates from frontline financial business scenarios and serves as a robust tool across various use cases. To achieve this, we adhered to the following principles:

- **Scenario Alignment:** Questions are crafted to reflect real challenges and inquiries faced by financial analysts in their day-to-day work.

- **Domain Expertise:** Each question is meticulously designed by financial experts, ensuring accuracy, contextual relevance, and analytical depth.

- **Breadth and Relevance:** The questions are structured to address a wide range of financial scenarios, representing typical needs and diverse problem types within the industry.

C.2.2 Collaborative Development Process

We developed the questions through close collaboration with financial experts from leading securities firms. By engaging with professionals from both business and IT departments, we ensured that the constructed questions resonate with real-world financial practices and challenges. This process ensures that the questions are not only theoretically sound but also practically applicable, forming a strong foundation for evaluating performance in financial long-document analysis.

C.3 Constructing Questions Using LLMs

We have also implemented the method of constructing questions in FinLBench using large language

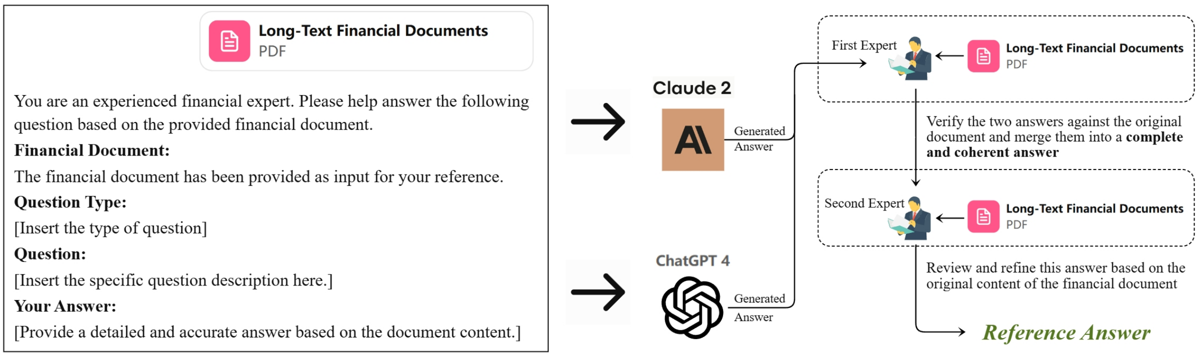

Figure 12: “2+1+1” Reference Answer Compilation Workflow.

models(LLMs). By using the prompt as a template to generate questions using an LLM other than the tested model, we found that LLMs not only have the ability to generate answers based on documents and questions, but also have good performance in constructing questions that meet the requirements. The prompt template is shown in Figure 13. The financial scenario is a description that informs the LLM of the scenarios in which these questions are raised, including the content or focus that the analyst wants to understand in the document. The question type refers to the type and brief introduction of questions that the LLM is designed for.

To ensure the high quality of the questions posed by the LLM and their inclusion in the dataset, we have had financial experts review and filter the questions raised by the LLM, thereby identifying those that are truly valuable.

You are an experienced financial document analyst. Please help me design some useful questions based on the input files and requirements to analyze this financial document.

Financial scenario:
[Description of financial scenario]

Question type:
[Question type: Description of question type]

The questions you designed:

Figure 13: Input template for questions constructed using LLMs.

We have demonstrated an example of constructing event analysis questions using an LLM, as illustrated in Figure 14. The first question given by the LLM is valuable in the given scenario and is adopted, while the second question is rejected due to its low relevance and usefulness in the given scenario.

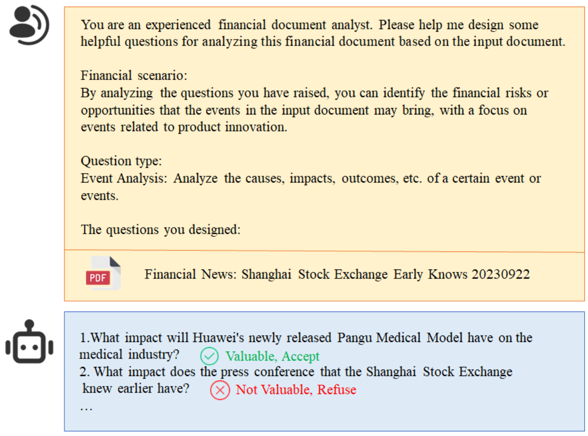

Figure 14: An example of using LLMs to construct questions: We prompt the LLM to generate questions by inputting a prompt.

C.4 Hybrid Workflow for Constructing Reference Answers

In constructing the reference answers for the FinLBench evaluation dataset, we employed a "2+1+1" workflow (illustrated in Figure 12) to ensure high-quality and reliable answers. For closed-ended questions, where clear answers exist within the document, we directly used the relevant content from the original text as the reference answer. For open-ended questions, which constitute 80% of the dataset and often involve financial documents spanning tens of thousands of words, we adopted a hybrid human-AI approach to balance efficiency and quality.

Specifically, we began by leveraging two state-of-the-art large language models, ChatGPT 4 and Claude 2, to independently generate two initial answers based on the question and the original financial document. Next, a human expert reviewed these AI-generated answers, validating them against the document content and integrating the best aspects of both into a cohesive and comprehensive response. Finally, a second human expert conducted a thorough review and optimization of the integrated answer, ensuring accuracy, clarity, and alignment with the original document. This finalized answer was used as the reference answer for the evaluation dataset.

The adoption of this workflow stems from the challenges inherent in creating a predominantly subjective question set with significant document length. While direct human authorship would be prohibitively time-consuming, we observed that advanced AI models, with appropriate human oversight, are capable of producing answers that approach human-level quality. This human-AI collaboration not only ensures efficiency but also maintains the high standard required for the FinLBench dataset.

D Model Introduction

Commercial Models

1. Alphabox [1] is a tool that can automatically generate document summaries based on user-uploaded documents.

2. ChatDOC[2] is a tool with high accuracy in recognizing tables, text, and formulas, capable of interacting with multiple documents simultaneously.

3. ChatGPT4[3] is a tool that supports multimodal document input, including images and text, capable of processing texts exceeding 25,000 words(Achiam et al., 2023).

4. ChatPDF[4] is a tool specifically designed for PDF document search.

5. Claude2[5] is a large language model with a context window of up to 100k tokens.

6. Moonshot[6] is a tool that supports long input, capable of interpreting documents exceeding 200,000 words.

7. WarrenQ[7] is a tool that leverages a large model augmented with search capabilities and the Juyuan Financial Database.

8. ERNIE Bot3.5[8] is a large language model featuring knowledge enhancement, retrieval enhancement, and dialogue enhancement technologies.

Open-source Models

1. GLM-4-9B-chat supports long-text reasoning with a maximum context of 128K tokens (GLM et al., 2024).

2. Qwen2-7B-Instruct has 7.62 billion parameters and supports a context length of up to 131,072 tokens (Yang et al., 2024).

[1] Alphabox - https://www.alphabox.top/
[2] ChatDOC - https://www.chatdoc.com/
[3] ChatGPT4 - https://chatgpt.com/?model=gpt-4
[4] ChatPDF - https://www.chatpdf.com/
[5] Claude2 - https://claude.ai/
[6] Moonshot - https://kimi.moonshot.cn/
[7] WarrenQ - https://www.warrenq.cn/
[8] ERNIE Bot3.5 - https://yiyan.baidu.com/

Table 2: Financial Task Description and Ability Assessment

Question-type	Ability assessment	Financial task description
Table extraction	Parse PDF tables and Understand two-dimensional table data	Accurately identify and refine investment research data. Financial summary data such as regular reports generally exist in tables including profit forecast data in research reports, etc.
Dialogue person discrimination	Understanding in multiplayer dialogue	Meeting roadshows usually involve multiple participants requiring accurate identification of multiple roles in order to accurately summarize the content of each role
Keyword extraction	Accurate retrieval and Entity extraction	Summarize the key words of the main theme in financial documents, including keywords for conference roadshows
Data extraction	Accurate retrieval and Entity extraction	Extract specific numerical information from financial documents such as extracting revenue, gross profit and other indicator data from company performance meeting minutes and extracting business indicators such as production and sales from company reports
Data calculation	Mathematical computing	Based on the known data in the document further perform some data calculations such as calculating derivative indicators
Trap issue	The severity of hallucinations	Hallucination refers to the fact that a large model generates factual information in the answer that is not included in financial documents
Logical reasoning	Logical reasoning especially in financial problems	Assist in future performance analysis based on company fundamentals or assist in industry trend analysis based on industry fundamentals
Outline generation	Text summarization	Serving writing scenarios such as generating brief outlines for government reports and writing comments
Event analysis	Understanding financial concepts and Logical reasoning in financial reasoning	Further consider the events in the information and obtain more inferred information. Common scenarios include quantitative investment policy analysis, ESG factor mining, etc.
Text summarization	Text summarization	Summarize the abstracts in financial investment research documents such as generating document abstracts for conference roadshows
Information extraction	Retrieval and Information extraction	Extract specific information such as identifying company entities, event queries, etc.
Reading comprehension	Comprehensive ability in understanding long texts	Generate relevant data, analytical logic and conclusions for a certain problem, including analysis of changes in the company's gross profit margin