

CFBenchmark-MM: Chinese Financial Assistant Benchmark for Multimodal Large Language Model

Jiangtong Li[†], Yiyun Zhu[†], Dawei Cheng^{*}, Zhijun Ding, Changjun Jiang

Abstract: Multimodal Large Language Models (MLLMs) have rapidly evolved with the growth of Large Language Models (LLMs) and are now applied in various fields. In finance, the integration of diverse modalities such as text, charts, and tables is crucial for accurate and efficient decision-making. Therefore, an effective evaluation system that incorporates these data types is essential for advancing financial application. In this paper, we introduce CFBenchmark-MM, a Chinese multimodal financial benchmark with over 9,000 image-question pairs featuring tables, histogram charts, line charts, pie charts, and structural diagrams. Additionally, we develop a staged evaluation system to assess MLLMs in handling multimodal information by providing different visual content step by step. Despite MLLMs having inherent financial knowledge, experimental results still show limited efficiency and robustness in handling multimodal financial context. Further analysis on incorrect responses reveals the misinterpretation of visual content and the misunderstanding of financial concepts are the primary issues. Our research validates the significant, yet underexploited, potential of MLLMs in financial analysis, highlighting the need for further development and domain-specific optimization to encourage the enhanced use in financial domain.

Key words: Multimodal Large Language Model; Financial Benchmark; Visual Evaluation

1 Introduction

In recent years, MLLMs such as GPT-4V [1], Gemini [2], and Claude [3] have grown significantly, excelling at integrating text and visuals for processing and generating information. Similarly, open-source MLLMs like LLaVA [4], CogVLM [5], and InternLM-XComposer [6] advance by leveraging the foundations of LLMs to enhance visual content integration across domains.

In financial domain, analyzing charts such as

stock trends, financial statements, and economic data visualizations becomes crucial for decision-making, which relies on both structured and unstructured data. Therefore, MLLMs need to manage complex financial context that combine textual data with visual charts, requiring high levels of analysis and comprehension. However, a comprehensive benchmark to evaluate MLLMs in handling complex multimodal financial context is still absence.

Early financial benchmarks [7, 8] focus on specific tasks like sentiment analysis of news [9, 10] or statistics reasoning [7, 11], evaluating single model performance within narrow domains. As benchmarks evolve [12–14], they begin assessing the generalization of models across diverse professional tasks [15, 16]. Recent developments in benchmarks [17] include non-text data like charts, enhancing realism and accuracy in multimodal scenarios. However, most financial benchmarks [7, 17] still either concentrate solely on text or are limited to specific multimodal tasks such as

Jiangtong Li, Yiyun Zhu, Dawei Cheng, Zhijun Ding, and Changjun Jiang are with the School of Computer Science and Technology, Tongji University, Shanghai 201804, China. E-mail: jiangtongli@tongji.edu.cn; 2533994@tongji.edu.cn; dcheng@tongji.edu.cn; dingzj@tongji.edu.cn; cjjiang@tongji.edu.cn.

[†] These authors contributed equally to this work.

^{*} To whom correspondence should be addressed.

Manuscript received: 2025-09-29; revised 2026-01-11; accepted: 2026-04-07

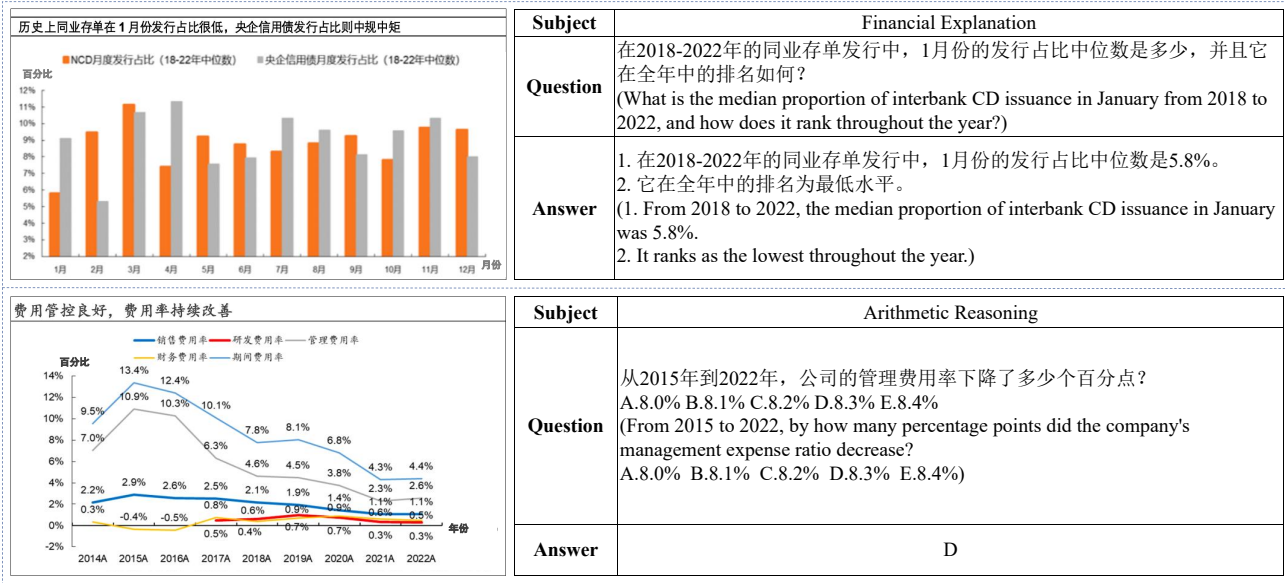


Fig. 1 Two examples from CFBenchmark-MM.

stock tracking. This underscores the need for a more comprehensive multimodal financial benchmark that not only evaluates financial knowledge and analytical skills but also measures effectiveness in utilizing and integrating multimodal information from various visual formats like pie charts, line charts, tables, and structural diagrams. In this paper, we introduce CFBenchmark-MM, which bridges gaps in existing benchmarks, providing a detailed and realistic evaluation for MLLMs in complex financial context with diverse data types. CFBenchmark [15] serves as a fundamental framework for evaluating LLMs within the text-based financial domain. As a progressive extension of the CFBenchmark series, CFBenchmark-MM expands this rigorous evaluation scope into the multi-modal landscape. Specifically, CFBenchmark-MM includes five core tasks, *i.e.*, arithmetic reasoning, statistical reasoning, structural reasoning, financial explanation, and financial knowledge, with 9,356 image-question pairs. To construct CFBenchmark-MM, we develop a three-step annotation pipeline: 1) we collect charts and their analyses from financial research reports; 2) we exploit GPT-4 to generate questions and answers based on the text analyses; 3) we manually verify whether the questions can be answered using the charts. In Figure 1, we present two cases from CFBenchmark-MM.

To assess the capability and efficiency of MLLMs in handling multimodal financial information, we develop a staged evaluation system including multiple settings. Initially, MLLMs answer questions without visual

content to evaluate their basic memory. Progressively, they respond based on a) textual descriptions of charts, b) charts, and c) both charts and their textual descriptions. This staged approach helps measure how MLLMs is enhanced with multimodal financial information in answering questions. As for the evaluation metrics, we employ a multiple-choice format, measuring accuracy, for objective questions in arithmetic, statistical, and structural reasoning. And we compare the generated answers to reference answers, using GPT-4 as a scorer, for subjective financial explanation and knowledge questions. Scores reflect the proportion of key information matched.

Our experiments reveal that most MLLMs have a fundamental memory of financial knowledge, allowing them to answer questions such as "What is the increase rate of Chinese GDP over 2022?" or "Which stock grew the second largest in 2021?" Including multimodal information improves response accuracy, yet overall performance remains moderate. Even the advanced GPT-4V only achieves a 52% accuracy and 38% scores on objective and subjective questions, highlighting the need for further enhancements in financial MLLMs. Additionally, our analysis of incorrect responses pinpoints that misinterpretation of visual content and misunderstanding of financial concepts are the primary issues. These findings emphasize the importance of MLLMs to analyze multimodal content in financial contexts. Our contributions can be summarized as:

Table 1 Comparison of CFBenchmark-MM with existing multimodal benchmarks. The ✓ and × denote the presence and absence of the corresponding feature, respectively.

Benchmarks	Question Type		Financial Competence				Evaluation Method			QA Number	Evaluation Model Number
	Objective	Subjective	Chart Interpretation	Complex Calculation	Structural Reasoning	Structural Reasoning	QA Accuracy	LLM Scoring	Staged Evaluation		
MMMU	✓	×	✓	×	×	×	✓	×	×	11,500	16
MMBench	✓	×	✓	×	×	×	✓	×	×	3,217	21
MME	✓	×	×	×	×	×	✓	×	×	2,374	30
FinVQA	×	✓	✓	✓	×	×	✓	×	×	1,025	9
MME-Finance	×	✓	✓	✓	✓	✓	✓	✓	×	2,274	19
CFBenchmark-MM	✓	✓	✓	✓	✓	✓	✓	✓	✓	9,356	13

- We introduce CFBenchmark-MM, a comprehensive benchmark with over 9,000 image-question pairs, focusing on evaluate MLLMs in real multimodal financial context.
- We establish an elaborated annotation process using real multimodal financial data, which includes automated tagging and manual verification to ensure dataset quality.
- We develop a staged evaluation system with multiple settings to assess the financial abilities of MLLMs and their use of multimodal information step by step, providing insights into existing challenges.

In Table 1, we present a comprehensive comparison of CFBenchmark-MM with representative multimodal benchmarks. This comparison serves to highlight our benchmark’s unique contributions, specifically its larger scale, specialized focus on complex financial reasoning tasks, and the novel staged evaluation method that ensures a robust assessment of multimodal capabilities in the financial domain.

2 Related Work

2.1 Large Language Models in Finance

Large Language Models (LLMs) [18–24], such as the GPT series by OpenAI [1, 25–27], feature billions of parameters and extensive training on text data. Notable among them, BloombergGPT [28] stands out as a FinLLM with 50 billion parameters. Successor models like FinGPT [29] and PIXIU [30] improve performance through supervised fine-tuning, excelling in zero-shot or few-shot settings. In the Chinese financial domain, XuanYuan [31] specializes in question-answering with domain-specific fine-tuning, while DISC-FinLLM [8] enhances reasoning in finance using multi-expert fine-tuning with LoRA [32]. Despite these advancements, current FinLLMs show a limited focus on multimodal data. Beyond finance, recent

studies like M2WLLM [33] and EV-STLLM [34] have also explored leveraging LLMs for data-driven forecasting in energy-related scenarios. Although these works demonstrate the general potential of LLMs in time-series modeling, their problem formulation differs from our focus on multimodal financial understanding and evaluation. Thus, this paper focuses on multimodal evaluation in finance, aiming to guide the development of MLLMs in this domain.

2.2 Multimodal Large Language Models

MLLMs [35], evolving from LLMs, focus on integrating information from various modalities with large language models, which can be divided into pre-training for visual-language alignment and fine-tuning for visual instructions. For visual-language pre-training methods, models like CLIP [36] use dual encoders to evaluate matches between images and text, while others like ALBEF [37] and VLMO [38] enhance interaction between these modalities. BLIP [39] and its successor, BLIP-2 [40], improve modality integration with query transformers and noise reduction, and CogVLM [5] introduces visual expert modules for deeper alignment. For visual-instruction fine-tuning methods, models such as InstructBLIP [41] and LLaVA [4] enhance instruction awareness and align image features with linguistic embeddings, respectively. LaVIT [42] and MiniGPT-4 [43] process visual and language data by converting visuals into discrete tokens and adjusting pre-trained models for multimodal instructions. The development of models like GPT-4 [1] showcases advanced capabilities in handling visual tasks and improving multimodal interaction across various benchmarks.

Despite the progress made by MLLMs, research on multimodal charts in financial domain remains sparse. Therefore, this work focuses on multimodal evaluation in real financial contexts, aiming to enhance the development of MLLMs specifically tailored for the financial domain.

2.3 Benchmarks

Research on evaluation benchmarks is essential for developing specific domains and provides a crucial framework for assessing the performance of large models. In this section, we discuss the evolution of benchmarks in financial domains for MLLMs, highlighting their significance and impact. In financial domain, benchmarks like FLUE [7], CFLEB [12], and DISC-FinLLM [8] evaluate NLP capabilities, while FinEval [13], investorbench [44] and FinQA [11] focus on financial analysis and knowledge. Other benchmarks such as FinBen [14], SuperCLUE-Fin [45], and FinanceIQ [31] merge NLP and financial skills assessment across various applications, significantly advancing LLMs in the financial domain. However, their limited focus on multimodal domains does not fully meet the practical requirements of financial applications. In general domains, several authoritative benchmarks [46–51] exist for multimodal evaluation. MME [52] is a widely used benchmark for evaluating MLLMs, assessing 14 sub-tasks of perceptual and cognitive abilities. ShareGPT4V [53] includes 1.2M image-text pairs, from diverse aspects such as world knowledge and artistic evaluations. Other multimodal evaluation like MME-industry [54] focus on specific domain. These benchmarks provide comprehensive evaluations and highlight improvement of MLLMs in general domain.

However, few studies, like FinVQA [17] and MME-Finance [55] focus on multimodal evaluation in finance. Specifically, FinVQA assesses how models handle different financial modalities but is limited by its small dataset size and narrow scope. MME-Finance proposes a bilingual financial multimodal benchmark, but also has the limitation of insufficient amount. This underscores the need to develop comprehensive multimodal benchmarks to thoroughly evaluate MLLMs in managing real financial context.

3 Methodology

In order to construct a multimodal evaluation benchmark that aligns with the practical requirements of financial application, we engage with professionals from the different financial domain, including consultants, bankers, investment bankers, and securities experts. Through extensive discussions with these industry specialists, we design and analyze our CFBenchmark-MM from two key dimensions: the

types of questions and the types of multimodal data.

In terms of the question type, we classify the questions in our benchmark into five categories, which includes three categories involving objective evaluations, *i.e.*, **Arithmetic Reasoning**, **Statistical Reasoning**, and **Structural Reasoning**, and two categories involving subjective evaluations, *i.e.*, **Financial Explanation** and **Financial Knowledge**. Regarding the multimodal data, we further category them into various types such as tables, histogram charts, line charts, pie charts, composite charts, and structural diagrams (such as equity structure charts and development path charts). This classification is carefully crafted to ensure that our benchmark comprehensively covers the diverse needs and challenges encountered in multimodal financial applications.

Here is a precise definition of the five question types:

- (1) **Arithmetic Reasoning:** This category evaluates the MLLMs in basic arithmetic operations such as addition, subtraction, multiplication, and division, as well as the understanding of numerical properties and relationships within financial charts.
- (2) **Statistical Reasoning:** This category involves to interpret and analyze statistical data in financial domain. It focuses on statistical measures such as mean, median, mode, variance, standard deviation, and correlation coefficients in financial charts.
- (3) **Structural Reasoning:** This category pertains to interpret and analyze structural diagrams and flowcharts in financial domain. This includes assessing diagrams that illustrate complex relationships such as shareholding structures or company development pathways.
- (4) **Financial Explanation:** This category entails to interpret and analyze the financial trends, fluctuations, and anomalies. It requires synthesizing multimodal information to provide comprehensive explanations of financial phenomena and their underlying causes.
- (5) **Financial Knowledge:** This category concerns the application and extension of financial knowledge through the analysis of multimodal data. It challenges the MLLMs to integrate and extrapolate given information in understanding financial concepts.

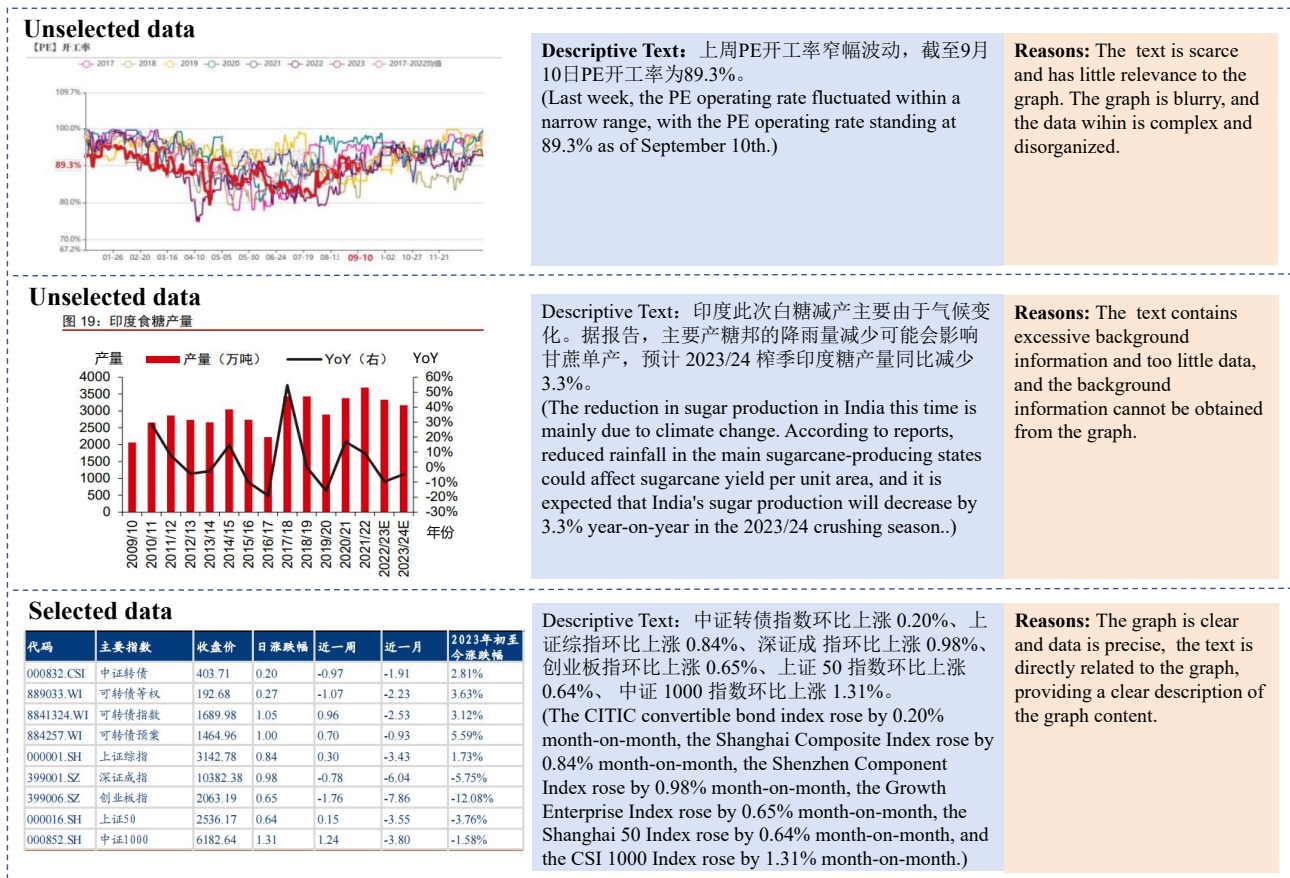


Fig. 2 The criteria we use when selecting data from financial research reports. We choose financial data that are clear, precise, and contain analyzable content.

3.1 Benchmark Construction

3.1.1 Data Collection

To construct a comprehensive benchmark, we first collect multimodal charts from publicly available financial research reports, which reflect actual usage patterns within the financial industry, providing an authentic basis for our evaluations. Besides, during benchmark construction from on these charts, we have two options, 1) employ annotators to propose questions and answers directly from the chart content; 2) extract the analyses of charts from the reports and automatically generate corresponding questions and answers. However, the former method requires substantial manpower and could introduce annotator biases. To align with the actual analysis methods used in financial application, we opt to collect the textual analyses accompanying these multimodal charts. These textual analyses, while not encompassing all information presented in the charts, represent the most critical and significant insights according to the

use cases of these reports.

To ensure the quality of the collected charts and their corresponding textual analyses, we establish specific screening criteria, 1) information in the charts is clear and visible without any obstructions, blurriness, or disorder; 2) the textual analysis has to correspond closely with the information in the charts and be relatively extensive (exceeding 50 Chinese characters). In Figure 2, we present several selected and unselected data to further explain our criteria. Based on these principles, we organize each collection group with two undergraduate students. Only when both individuals agree on the clarity and relevance of a chart and its analysis, the collect charts-analysis pair is included. Through this meticulous approach, we compile a total of 2500 pairs of charts-analysis.

3.1.2 Data Annotation

To ensure the quality of our benchmark, we utilize the advanced LLM, GPT-4, to generate questions and answers from the textual analysis of collected chart-analysis pairs. Based on the characteristics of

the charts, the association between chart types and question categories are defined as 1) line charts and tables are used to generate **Arithmetic Reasoning** and **Financial Knowledge** question, 2) histogram charts and other types like pie charts and composite charts are used to generate **Statistical Reasoning** and **Financial Explanation** questions, 3) structural diagrams are used to generate **Structural Reasoning** and **Financial Explanation** questions. While proposing the questions and answers, we design five specific instructions for GPT-4, which require the model to first assess if the provided analysis is suitable for question generation. If deemed appropriate, the model then proceed to craft questions aligned with the requirements of each of the five question categories. More details about these instructions are available in appendix. For each analysis that qualified, the model is instructed to generate two questions per question category, ensuring a comprehensive assessment.

For subjective questions (financial knowledge and explanation), the answers are formatted in bullet points to facilitate clarity, while for objective questions (arithmetic, statistical, and structural reasoning), the answers are structured as multiple-choice. Through this methodical approach, we successfully identified 2,421 chart-analysis pairs suitable for this exercise, ultimately generating a robust dataset of 9,684 image-question pairs. This annotation procedure ensures the quality of question-answer pairs and their relation with financial charts and provides solid support for evaluating MLLMs in multimodal financial context.

3.1.3 Manual Examination

After completing data annotation, we conduct manual verification through the following steps, 1) we verify whether the proposed questions can be answered using only the provided multimodal information; 2) for objective questions, checking the correctness of the option format and ensuring that the correct answer is included among the options; 3) for subjective questions, confirming the accuracy of answers formatted in bullet points; 4) we manually label questions that are unrelated to the charts or where the correct information cannot be obtained from the charts. All manual checks are conducted by PhD students with financial expertise, with high inter-annotator agreement (Cohen’s $\kappa = 0.810$). Detailed instructions for examination can be found in Section S.1, S.2 and S.3, which are in the

Table 2 The statistics about CFBenchmark-MM in terms of question category.

Category	Chart Number	Question Number	Avg. Token (Q)	Avg. Token (A)
Arithmetic Reasoning	1020	2040	79.7	1.1
Statistical Reasoning	1151	2302	69.9	1.4
Structural Reasoning	168	336	82	1.2
Financial Explanation	1319	2638	33.9	110.8
Financial Knowledge	1020	2040	28.4	275.3

Electronic Supplementary Material (ESM) of the online version of this article. After the manual examination, we finally obtain 2,339 valid charts along with 9,356 questions to construct CFBenchmark-MM.

3.1.4 Statistics

In CFBenchmark-MM, the most common type, histogram charts, totals 677 charts, accounting for 29%, and the least common, structural diagram, includes only 168 charts, making up 7%. This variation reflects the diversity in financial research reports. This distribution is intentionally designed to balance chart diversity and reasoning coverage. While certain chart types appear more frequently due to their natural prevalence in financial reports, we do not enforce rigid one-to-one mappings between chart types and reasoning categories. In Table 2, we present the statistics for different categories of questions. Financial explanation questions are most frequent due to the higher availability of corresponding charts, while structural reasoning questions are least frequent, reflecting the scarcity of structural diagrams. Additionally, subject questions often feature multiple correct answers, resulting in answers longer than one word. Our CFBenchmark-MM, with its extensive range and high volume, is the most comprehensive to date and plays a significant role in multimodal financial context.

3.2 Benchmark Evaluation

Most existing MLLMs have been trained on financial text data and may inherently possess some capabilities to answer benchmark-related questions. To fully evaluate the financial and multimodal processing capabilities, we develop a staged evaluation system with multiple settings for CFBenchmark-MM. In Figure 3, we illustrate the procedure of our staged evaluation system, as detailed below:

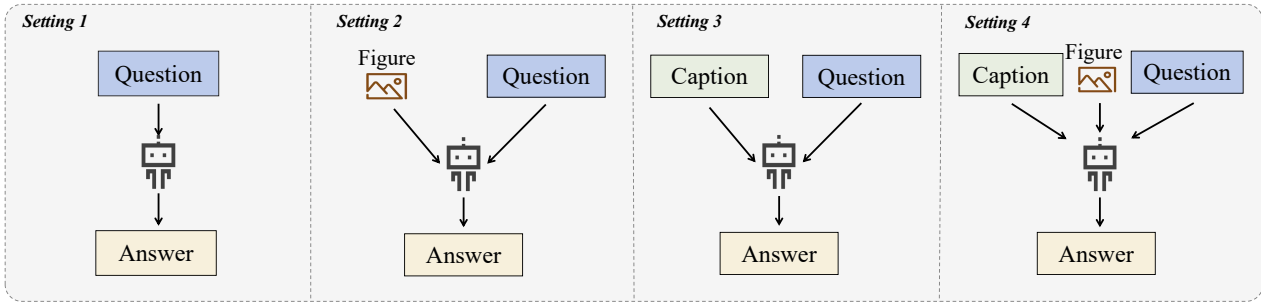


Fig. 3 The illustration of benchmark evaluation system.

- (1) **Question (setting-Q):** This baseline evaluation assesses models MLLMs on questions without additional multimodal information. It reflects the inherent knowledge of MLLMs.
- (2) **Question + Image (setting-Q+I):** In this setting, we introduce original charts into the evaluation to examine how models integrate visual content with textual queries, testing their multimodal processing capabilities.
- (3) **Question + Caption (setting-Q+C):** In this setting, we enhance the assessment by adding textual descriptions generated by advanced MLLM, GPT-4V, from the charts, evaluating how well models can utilize synthesized textual representations alongside the queries.
- (4) **Question + Image + Caption (setting-Q+I+C):** This setting combines both the original visual content and its textual description into assessment, measuring the ability of MLLMs to synergize and scale their understanding across complex multimodal inputs.

In the Q+C and Q+I+C settings, GPT-4V-generated captions provide high-level semantic context like chart topics and overall trends rather than exact values, such information constitutes essential background knowledge for multimodal reasoning, rather than explicit cues for direct answer extraction. Through this staged evaluation, we not only gauge the foundational knowledge of MLLMs in financial domain but also critically assess their proficiency in handling both direct multimodal inputs and derived textual representations. This evaluation strategy provides valuable insights into model performance, aiding in the effective selection of models for financial applications.

For evaluation metric, objective questions are assessed through a multiple-choice format, with

accuracy determined by correct options. Subjective questions, like financial explanation and knowledge, are evaluated by matching generated answers to reference answers using GPT-4 as a scorer. To validate the reliability of this automated scoring, we conduct a calibration study where two finance-trained human raters independently score a stratified random subset of 1,000 subjective samples under the same rubric, and the resulting correlations with GPT-4 scores are $r = 0.81$ and $r = 0.87$. Compare to pure human evaluation, which is prone to subjectivity and non-reproducibility at scale, GPT-4 with explicit rubrics provides standardized and reproducible assessments, ensuring fair model comparison. The LLM-as-a-Judge paradigm is well-established: MMBench [51] demonstrates that GPT-4 achieves 91.5% agreement with human judgments. G-Eval [56] shows that GPT-4 correlates better with human experts than metrics like BLEU. Scoring reflects the proportion of key points accurately captured. For instance, if three out of five key points in a reference answer are matched, the score is 0.6.

4 Experiments

4.1 Experimental Setup

We evaluate the zero-shot performance of 13 cutting-edged MLLMs: (1) LLaVA-NeXT-mistral-7B [4]; (2) LLaVA-NeXT-vicuna-7B [4]; (3) LLaVA-NeXT-vicuna-13B [4]; (4) LLaVA-NeXT-34B [4]; (5) CogVLM2 [5]; (6) DeepSeek-VL-7B [57]; (7) InternLM-XComposer2-4KHD-7B [6]; (8) InternLM-XComposer2-VL-7B [58]; (9) mPLUG-Owl [46]; (10) GPT-4V [1]; (11) Llama-3.2-11B-Vision [59]; (12) Qwen2.5-VL-7B-Instruct [60]; (13) Baichuan-Omni-1d5 [61]. For open-source MLLMs, we directly utilize their released model weights on single Nvidia A100 GPU during the evaluation. For GPT-4V, we utilize the API for evaluation. Moreover, some popular MLLMs,

Table 3 Answer accuracy comparison of 13 cutting-edge MLLMs under Q (question) and Q+I (question and image) settings.

Model	Q					Q+I				
	Statistical	Arithmetic	Structural	Explanation	Knowledge	Statistical	Arithmetic	Structural	Explanation	Knowledge
CogVLM2	30.1	29.2	34.7	7.6	24.5	<u>46.6</u>	40.6	<u>55.4</u>	34.2	38.6
DeepSeek-VL-7B	30.9	29.1	39.2	5.4	30.9	35.4	31.8	44.9	12.6	39.3
GPT-4V	38.0	26.9	41.9	8.2	<u>31.8</u>	52.0	46.0	58.1	24.8	52.4
InternLM-XComposer2-4KHD-7B	27.4	24.4	35.9	9.8	29.9	44.6	38.9	52.1	<u>27.7</u>	35.5
InternLM-XComposer2-VL-7B	30.4	29.1	39.5	12.7	33.2	38.6	33.0	46.7	21.7	41.6
LLaVA-NeXT-34B	<u>36.2</u>	34.0	<u>40.8</u>	6.3	29.0	42.8	39.2	50.3	15.8	44.2
LLaVA-NeXT-mistral-7B	16.6	18.2	17.4	2.2	15.4	22.4	24.4	26.6	5.6	18.1
LLaVA-NeXT- <i>vicuna</i> -13B	22.6	19.1	27.2	4.0	21.3	29.6	26.7	29.9	12.3	26.3
LLaVA-NeXT- <i>vicuna</i> -7B	29.1	28.4	30.5	6.3	10.4	31.8	32.3	30.8	11.3	15.2
mPLUG-0w1	24.8	22.2	29.0	7.0	25.9	28.0	25.2	31.4	12.1	30.9
Qwen2.5-VL-7B-Instruct	29.3	27.5	30.2	10.2	29.4	43.1	<u>40.7</u>	49.2	22.1	<u>49.5</u>
Llama-3.2-11B	25.3	<u>29.3</u>	27.7	8.1	28.4	45.1	42.3	52.9	18.1	46.5
Baichuan-Omni-1d5	27.9	25.7	28.8	<u>12.6</u>	25.9	42.5	38.6	46.7	21.0	46.6
Average	28.4	26.4	32.5	7.7	25.8	38.7	35.4	44.2	18.4	37.3

like, BLIP [39, 40], and InstructBLIP [41], cannot be evaluated in our benchmark due to the language support. The instruction for evaluation is shown below:

“Here is a question about {subject}, featuring a chart from financial domain and its corresponding textual description. Based on the chart and text provided, please answer the question. The chart and text may be absent; if so, ignore their content. The question may be multiple-choice or a short-answer type. For multiple-choice questions, select all correct options; for short-answer questions, respond to each point in the question. The chart could be a table, structural diagram, pie chart, line chart, or bar chart. When answering, carefully examine each component of the chart, including but not limited to: the title, labels on the X and Y axes, the legend, the colors and shapes of data series, and the data points and values displayed. If it is a table, closely review the title and content. If it is pie charts, specify the percentage and name of each sector. If values are not marked on chart, determine the size of the values based on the X, Y axes and answer based on these estimated values. The question is {question}. The chart is {<IMAGE>}. The chart description is {caption}.”

4.2 Main Results

In Table 3, we display the performance of 13 MLLMs under setting-Q and setting-Q+I, showcasing their capabilities within multimodal financial contexts. Despite only having questions as input, most MLLMs achieve over 20% accuracy (accuracy for random choice) in objective questions and score above 10% in subjective assessments, demonstrating their foundational memory in financial knowledge. With the

addition of original charts (setting-Q+I), all MLLMs exhibit notable improvements, indicating the value of integrating visual data for enhanced comprehension. Among them, GPT-4V stands out, achieving the largest improvement upon integrating charts, with over 15% accuracy and 20% score increases in objective and subjective questions, respectively. This underscores its superior multimodal processing capabilities. In contrast, LLaVA-NeXT-mistral-7B shows the weakest overall performance, with minimal improvement upon integrating charts, likely due to its limited Chinese language processing capabilities. However, even advanced MLLMs like GPT-4V achieve only 52% accuracy and 38% score in objective and subjective questions, falling short of real-world application requirements. Moreover, under both evaluation settings, most models perform best on Structural Reasoning questions, while their performance is weakest on tasks related to Financial Explanation. Although the Structural Reasoning subset is relatively small, the consistent performance gains across multiple MLLMs indicate that our main conclusion on the benefit of multimodal input remains robust. These findings highlight a substantial opportunity for further enhancements in multimodal financial applications, emphasizing the need for exploiting multimodal data in financial analysis and bridging the gap between capabilities and practical utility.

In Table 4, we present the performance of three representative MLLMs, *i.e.*, GPT-4V, LLaVA-NeXT-34B, and InternLM-XComposer2-4KHD-7B, across different evaluation settings in multimodal financial

Table 4 Evaluation results (answer accuracy) on GPT-4V, LLaVA-NeXT-34B, and InternLM-XComposer2-4KHD-7B (InternLM-X2-4K-7B for short) in different evaluation settings. Q, C, and I indicate question, caption, and image, respectively. The best performance in different tasks are highlight in bold.

Question	GPT-4V				LLaVA-NeXT-34B				InternLM-X2-4K-7B			
	Q	Q+C	Q+I	Q+C+I	Q	Q+C	Q+I	Q+C+I	Q	Q+C	Q+I	Q+C+I
Arithmetic	26.9	35.4	46.0	38.1	34.0	38.0	39.2	42.1	24.4	32.8	38.8	40.1
	-	+8.5	+19.1	+11.2	-	+4.0	+5.2	+8.1	-	+8.4	+14.4	+15.7
Statistical	38.0	41.5	52.1	45.2	36.2	40.2	42.8	45.2	27.4	36.6	44.6	45.9
	-	+3.5	+14.1	+7.2	-	+4.0	+6.6	+9.0	-	+9.2	+14.2	+15.5
Structural	41.9	50.1	58.1	54.6	40.8	47.1	50.3	55.7	35.9	47.3	52.1	54.2
	-	+8.2	+16.2	+12.7	-	+6.3	+9.5	+15.7	-	+11.4	+16.2	+18.1
Explanation	8.2	11.5	24.8	23.8	6.3	13.6	15.8	17.1	9.8	12.9	27.7	31.1
	-	+3.3	+16.6	+15.6	-	+7.3	+9.5	+10.8	-	+3.1	+17.9	+21.3
Knowledge	31.8	38.1	52.4	49.9	29.0	34.6	44.2	48.3	29.9	33.5	35.5	38.9
	-	+6.3	+20.6	+18.1	-	+5.6	+15.2	+19.3	-	+3.6	+5.6	+9.0
Average	29.4	35.3	46.7	42.3	29.3	34.7	38.5	41.7	25.5	32.6	39.7	42.0
	-	+5.9	+17.3	+12.9	-	+5.4	+9.2	+12.4	-	+7.1	+14.2	+16.5

Table 5 Evaluation results (answer accuracy) on GPT-4V, LLaVA-NeXT-34B, and InternLM-XComposer2-4KHD-7B (InternLM-X2-4K-7B for short) under Q+I setting in term of different instruction format.

Question	GPT-4V			LLaVA-NeXT-34B			InternLM-X2-4K-7B		
	Original	Simple	ICL	Original	Simple	ICL	Original	Simple	ICL
Arithmetic	46.0	46.2	49.1	39.2	38.7	41.9	38.8	38.2	40.3
Statistical	52.1	51.7	54.8	42.8	42.5	44.3	44.6	43.9	46.1
Structural	58.1	57.6	60.9	50.3	49.2	53.1	52.1	51.8	53.2
Explanation	24.8	24.4	28.3	15.8	14.8	18.4	27.7	27.3	29.2
Knowledge	52.4	51.1	57.2	44.2	44.6	50.2	35.5	34.9	42.1
Average	46.7	46.2	50.1	38.5	38.0	41.6	39.7	39.2	42.0

contexts. Despite the highest performance of GPT-4V at setting-Q, the inclusion of chart information in setting-Q+I significantly improves performance, illustrating its strong capability to integrate visual content. Conversely, InternLM-XComposer2-4KHD-7B also demonstrates significant improvements, attributed to its higher resolution for visual processing. In contrast, LLaVA-NeXT-34B shows minimal gains, suggesting the need to enhance its CLIP-based visual alignment for charts. Additionally, all MLLMs improve when captions is added to questions (from setting-Q to setting-Q+C), but starting from setting-Q+I, only LLaVA-NeXT-34B and InternLM-XComposer2-4KHD-7B could further benefit from the addition of captions (setting-Q+I+C), while GPT-4V experiences a decrease. This indicates that captions generated by GPT-4V can provide additional useful information to other MLLMs. However, since the captions only contain part of the information from the charts, they naturally struggle to offer more information to GPT-4V

itself.

Overall, MLLMs still show limited capabilities in integrating visual content in financial applications. Supplementing effective textual information can also significantly aid MLLMs in achieving better performance. These findings highlight the importance of fine-tuning multimodal architectures and optimizing caption integration methods to bridge the gap between visual and textual.

4.3 Ablation Study

4.3.1 The Effect of Instruction

To further explore the MLLMs to different instructions, we design two extra types of instructions and display the results in Table 5 (Detailed instructions can be found in Section S.4 in the ESM). The first type, **Simple**, simplifies existing instructions by directly requiring models to answer questions based on provided charts, removing any additional guidance or contextual cues to test the models' ability to perform with minimal

Table 6 Evaluation results (answer accuracy) on GPT-4V, LLaVA-NeXT-34B, and InternLM-XComposer2-4KHD-7B (InternLM-X2-4K-7B for short) under Q+I setting in term of different instruction format.

Question	GPT-4V		LLaVA-NeXT-34B		InternLM-X2-4K-7B	
	zh	en	zh	en	zh	en
Arithmetic	46.0	45.5	39.2	40.6	38.8	39.3
Statistical	52.1	53.8	42.8	45.9	44.6	42.8
Structural	58.1	57.2	50.3	53.8	52.1	49.8
Explanation	24.8	25.2	15.8	16.1	27.7	26.5
Knowledge	52.4	50.7	44.2	45.5	35.5	36.7
Average	46.7	46.5	38.5	40.4	39.7	39.0

input. The second type, **ICL**, adds three sets of question-reason-answer combinations as prompts to study the in-context learning ability of MLLMs, enabling models to leverage contextual examples for better reasoning. We only include question-reason-answer pairs in few-shot instructions, as some MLLMs do not support multi-image input, which could otherwise complicate the evaluation process. This table reveals that simplified instructions lead to a slight decrease in model performance, about 1%. This indicates that without additional information or example templates, the benefits of detailed instructions are limited. Conversely, the ICL results demonstrate that models achieve 3%-4% improvement from the original performance. Although image is not included in instructions, the improvement still suggests potential for enhancing instruction design. It also highlights opportunities for improvement through more innovative instruction design and comprehensive model training.

4.3.2 The Effect of Language

To further verify the impact of language on model performance, we translate all instructions, questions, and answers into English using GPT-4 and conduct experiments based on the translation. The experimental results are shown in Table 6. From this table, we can find that LLaVA-NeXT-34B achieves an improvement of 1%-3% with English input, likely due to its stronger English proficiency compared to Chinese. In contrast, GPT-4V and InternLM-XComposer2-4KHD do not show significant changes, indicating their robustness in different languages. This suggests that language itself does not significantly affect the difficulty of answering questions, as the core challenge lies in the complexity of the financial and multimodal tasks rather than linguistic barriers. However, the results also highlight that LLaVA-NeXT-34B performs poorly in Chinese, which significantly affects its performance

in CFBenchmark-MM, underscoring the importance of language-specific optimization for models targeting non-English domains. These findings emphasize the need for future research to address language-specific limitations and enhance multilingual capabilities in existing MLLMs, particularly for applications in diverse linguistic and cultural contexts. Overall, our findings reinforce that while language translation offers benefits to some MLLMs, the core difficulty lies in the complexity of the multimodal financial tasks.

4.4 Case Study

To further explore issues with MLLMs in multimodal financial contexts, we present two typical cases of incorrect responses from GPT-4V in Figure 4 and analyze the reasons to these errors. In Figure 4(a), we showcase an example from statistical reasoning, where the response from GPT-4V confounds the traffic data from different periods and airports, failing to extract the correct visual information to answer the question. Furthermore, in the response, the calculation of GPT-4V is also incorrect, indicating that its mathematical abilities need further improvement, particularly in handling complex numerical data. In Figure 4 (b), we present an example from financial explanation, where the response from GPT-4V reveals the failure in distinguishing “compound growth rate” and “average growth rate” from the question and chart, highlighting an incomplete understanding of GPT-4V in financial concepts. Moreover, when we query the relationship between these two concepts in textual form with GPT-4V, it correctly identifies their relationship, which shows that aligning visual-textual concepts poses more challenges than textual-textual concepts. This discrepancy underscores the need for improved multimodal alignment in financial contexts, particularly in integrating visual data with

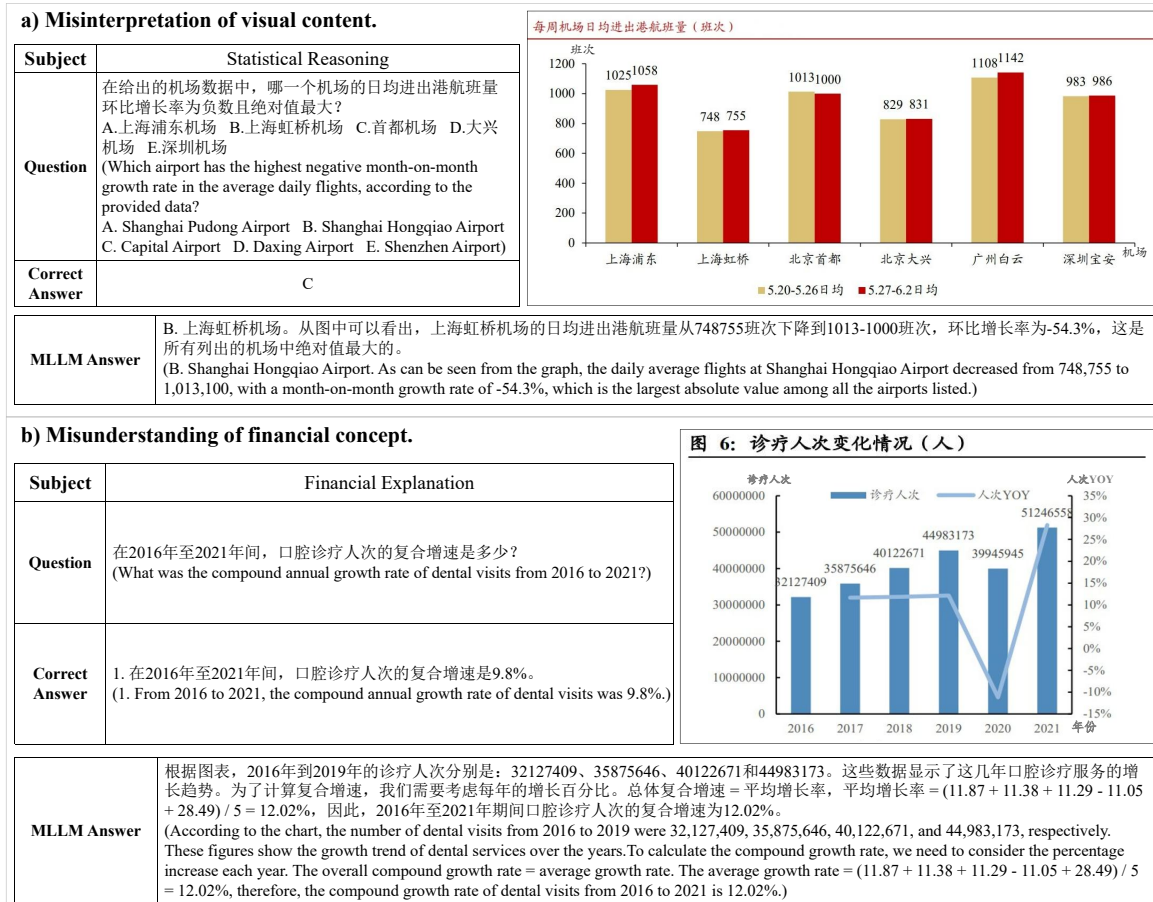


Fig. 4 Failure cases in CFBenchmark-MM by GPT-4V with reasons and answers.

textual analysis to enhance the accuracy and reliability of MLLMs in real-world financial applications. Also, this explains the lower accuracy observed in the Financial Explanation category, which requires models to go beyond surface-level data extraction and perform concept-aware multimodal reasoning. Specifically, models must align visual patterns in charts with abstract financial concepts and generate coherent explanatory narratives. Such reasoning often involves implicit assumptions and causal semantics that are not explicitly encoded in numerical trends, making this task particularly challenging for MLLMs.

We also provide a quantitative distribution of major error types for CogVLM2 as an example. As shown in Figure 5, visual extraction errors account for the largest proportion (58.32%), which indicates that the primary bottleneck of current MLLMs lies in accurately identifying and extracting key information from financial charts, such as values, labels, and structural relations. In comparison, numerical calculation errors and financial concept misunderstanding suggest that

reasoning and concept grounding remain challenging even after partial visual information is captured. Incomplete answers show that subjective tasks often fail because models cannot cover all required points. These findings suggest that while MLLMs have made significant strides, there is still considerable room for improvement in their ability to process and interpret complex multimodal financial data. Additionally, our analysis reveals that the errors may stem from limitations in training data diversity and challenges in visual processing, which indicate a critical need for advanced fine-tuning strategies. These shortcomings illustrate that even state-of-the-art MLLMs require further enhancement to handle the intricacies of financial data analysis and multimodal integration.

4.5 Computational Cost and Scalability Analysis

We estimated the theoretical inference cost based on standard scaling laws for transformer models in Table 7. The total cost per inference (C_{total}) is decomposed into the vision encoder cost (C_{vis}), the LLM prefill cost

Table 7 Theoretical computational cost analysis of evaluated MLLMs. Comparisons are conducted under a standardized financial scenario: processing one high-resolution image (1024×1024) with a 500-token prompt and generating a 200-token response.

Model	Model Parameters (B)		Vision Tokens (N_v)	Estimated Inference Cost (TFLOPs)		
	Vision (P_v)	LLM (P_{llm})		Vision Enc.	LLM Prefill	Total
mPLUG-Owl	0.3	7	576	0.08	15.1	17.9
DeepSeek-VL-7B	0.4	7	576	0.14	15.1	18.0
InternLM-XComposer2-VL-7B	0.3	7	1,024	0.08	21.3	24.2
LLaVA-NeXT-Mistral-7B	0.3	7	$\sim 2,880$	0.32	47.3	50.4
LLaVA-NeXT-Vicuna-7B	0.3	7	$\sim 2,880$	0.32	47.3	50.4
LLaVA-NeXT-Vicuna-13B	0.3	13	$\sim 2,880$	0.32	87.9	93.4
LLaVA-NeXT-34B	0.3	34	$\sim 2,880$	0.32	229.8	243.7
InternLM-XComposer2-4KHD	0.3	7	$\sim 8,737$	2.00	129.3	134.1
CogVLM2	4.4	8	1,225	1.90	27.6	32.7
Llama-3.2-11B-Vision	1.0	11	$\sim 1,610$	0.60	46.4	51.4
Qwen2.5-VL-7B-Instruct	0.6	7	$\sim 1,300$	0.50	25.2	28.5
Baichuan-Omni-1.5	0.4	7	$\sim 1,500$	0.40	28.0	31.2

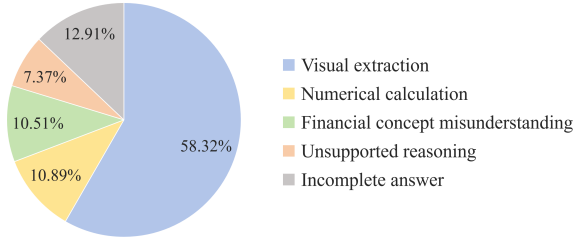


Fig. 5 Quantitative distribution of major error types for CogVLM2 on CFBenchmark-MM.

(C_{prefill}), and the generation cost (C_{gen}). The calculation formula is defined as follows:

$$C_{\text{total}} \approx C_{\text{vis}} + C_{\text{prefill}} + C_{\text{gen}}$$

$$C_{\text{total}} \approx C_{\text{vis}} + 2 \cdot P_{\text{llm}} \cdot (N_{\text{vis}} + N_{\text{in}}) + 2 \cdot P_{\text{llm}} \cdot N_{\text{out}}$$

Where: P_{llm} denotes the parameter size of the LLM backbone (e.g., 34×10^9 for 34B). N_{vis} is the number of visual tokens representing the image (e.g., 2,880 tokens for LLaVA-NeXT at 1024×1024 resolution). N_{in} and N_{out} represent the input prompt length (fixed at 500) and generated response length (fixed at 200), respectively. C_{vis} is the forward pass FLOPs of the vision encoder (e.g., ~ 0.08 TFLOPs for CLIP-L). For instance, the calculation for LLaVA-NeXT-34B is derived as:

$$C_{\text{total}} \approx 0.32 + 2 \cdot 34 \cdot (2880 + 500) \cdot 10^{-3} + 2 \cdot 34 \cdot 200 \cdot 10^{-3} \\ \approx 243.7 \text{ TFLOPs}$$

This quantitative breakdown confirms that the high computational cost of LLaVA-NeXT-34B is driven by the interaction between its large parameter count and dense visual token representation.

Our analysis reveals three distinct computational bottlenecks affecting scalability.

- **Parameter-Dominant Cost:** As shown in the table, LLaVA-NeXT-34B exhibits the highest inference cost (~ 243.7 TFLOPs). This is driven primarily by its massive language backbone (34B parameters). While it offers superior reasoning capabilities, the high computational overhead necessitates significant GPU memory bandwidth, making it less suitable for low-latency, real-time applications compared to 7B-parameter models.
- **Token-Dominant Cost:** InternLM-XComposer2-4KHD presents a unique case. Despite having a smaller 7B backbone, its dynamic resolution strategy generates over 8,700 visual tokens to preserve fine-grained chart details. This results in a prefill computation cost (~ 129.3 TFLOPs) that rivals or exceeds that of much larger models (e.g., 13B scale). This highlights a critical trade-off: capturing high-resolution visual details in financial documents comes at the cost of significantly increased prefill latency.
- **Encoder-Dominant Cost:** CogVLM2 employs a massive 4.4B-parameter vision encoder (EVA2-CLIP-E). Unlike other models that use lightweight encoders (e.g., CLIP-L, ~ 0.3 B), CogVLM2 incurs a computational cost during the visual encoding stage that is approximately $6 \times$ to $20 \times$ higher than its peers. This encoder-heavy design may create throughput bottlenecks in scenarios requiring high-concurrency image processing.

5 Conclusion

This paper has explored the development and implication of Multimodal Large Language Models (MLLMs) in multimodal finance application, highlighting the crucial role of integrating texts, charts and figures to improve decision-making accuracy and efficiency. Therefore, we introduce CFBenchmark-MM, a comprehensive financial benchmark with over 9,000 image-question pairs, including various data representations such as tables, line charts, pie charts, and structural diagrams. Our evaluation settings have assessed the capabilities of MLLMs in handling multimodal information, from textual data to visual content integration, and further reveal their strengths and limitations across different financial reasoning tasks. These findings indicate that although MLLMs are capable of possessing inherent financial knowledge, their efficiency in processing multimodal information is still limited, highlighting the need for further improvements in the multimodal context. Our research underscores the significant yet underexploited potential of MLLMs in financial analysis and advocates for their enhanced application in the financial domain. The insights call for an ongoing refinement of the MLLMs to better utilize multimodal information, with the aim of achieving more robust and accurate financial decision making.

6 Limitation and Future Work

In this paper, we investigate the evaluation of Multimodal Large Language Models (MLLMs) in financial application and propose a benchmark and an evaluation framework, CFBenchmark-MM. The benchmark segregates into five question categories, analyzing 12 open-source MLLMs and 1 close-source MLLM, GPT-4V. The benchmark is scheduled for release following the publication of this paper. However, our evaluations predominantly concentrate on open-source MLLMs, with a noticeable absence of testing on certain proprietary models such as Claude and Gemini. Additionally, while our five categories of questions encompass a broad range of real-world financial applications, they do not extend into specialized sub-disciplines within finance, such as investments and banking. In constructing our benchmark, we effectively utilize multimodal data from financial research reports to generate questions

and answers, primarily focusing on Chinese-language reports. Although an English version is provided, the exploration of multilingual benchmark remains constrained.

In the future, the research directions can be discussed from the following two perspective:

- **Benchmark in finance:** 1) Enhancing the task classification by breaking down the current five question types into more detailed financial sub-domains; 2) Extending the linguistic scope of benchmark system by including additional languages such as French, Japanese, and German, moving beyond the current bilingual (Chinese-English) framework; 3) Incorporating considerations of security and robustness in the design of future benchmarks to facilitate safer and more reliable applications of MLLMs in financial domain.
- **MLLMs in finance:** 1) Expanding the application of MLLMs within the financial domain by leveraging more domain-specific multimodal data to optimize their performance; 2) Developing specialized visual alignment modules as current reliance on pre-trained models like CLIP and ViT offers limited functionality in accurately representing financial charts; 3) Investigating the potential of multi-agent systems in multimodal financial settings to understand how collaborative interactions among different agents can enhance model effectiveness and decision-making.

These proposed enhancements aim to deepen the integration, utility, and adaptability of MLLMs in financial domain, thereby augmenting their applicability in sophisticated financial analyses.

Acknowledgment

This work was supported by the National Science Foundation of China (62522213, 62402341) and the Fundamental Research Funds for the Central Universities.

Electronic Supplementary Material

We present the instructions and prompts employed during the benchmark construction and evaluation in Supplementary. For the benchmark construction, we provide human annotators with textual instructions

to ensure the quality of the benchmark during the multi-modal data collection (Section S.1) and manual examination (Section S.3) stages. Additionally, corresponding to five question type, we design five different annotation prompts (Section S.2) for GPT-4V to convert financial analyses into corresponding questions and answers. For the evaluation of the benchmark, we devise three types of prompts (Section S.4) for MLLMs to answer the questions, namely, original, simple, and ICL. These prompts are structured to assess the ability of MLLMs to process and interpret the information under varying levels of complexity and context-specific guidance.

References

- [1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, “GPT-4 technical report,” *arXiv preprint arXiv:2303.08774*, 2023.
- [2] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, *et al.*, “Gemini: a family of highly capable multimodal models,” *arXiv preprint arXiv:2312.11805*, 2023.
- [3] “Introducing the next generation of claude,” 2024.
- [4] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” *Advances in neural information processing systems*, vol. 36, 2023.
- [5] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, X. Song, *et al.*, “Cogvlm: Visual expert for pretrained language models,” *arXiv preprint arXiv:2311.03079*, 2023.
- [6] X. Dong, P. Zhang, Y. Zang, Y. Cao, B. Wang, L. Ouyang, S. Zhang, H. Duan, W. Zhang, Y. Li, *et al.*, “InternLM-XComposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd,” *arXiv preprint arXiv:2404.06512*, 2024.
- [7] R. S. Shah, K. Chawla, D. Eidnani, A. Shah, W. Du, S. Chava, N. Raman, C. Smiley, J. Chen, and D. Yang, “When flue meets flang: Benchmarks and large pre-trained language model for financial domain,” *arXiv preprint arXiv:2211.00083*, 2022.
- [8] W. Chen, Q. Wang, Z. Long, X. Zhang, Z. Lu, B. Li, S. Wang, J. Xu, X. Bai, X. Huang, *et al.*, “Disc-finllm: A chinese financial large language model based on multiple experts fine-tuning,” *arXiv:2310.15205*, 2023.
- [9] K. Mishev, A. Gjorgjevikj, I. Vodenska, L. T. Chitkushev, and D. Trajanov, “Evaluation of sentiment analysis in finance: from lexicons to transformers,” *IEEE access*, vol. 8, pp. 131662–131682, 2020.
- [10] A. Sinha, S. Kedas, R. Kumar, and P. Malo, “Sentfin 1.0: Entity-aware sentiment analysis for financial news,” *Journal of the Association for Information Science and Technology*, vol. 73, no. 9, pp. 1314–1335, 2022.
- [11] Z. Chen, W. Chen, C. Smiley, S. Shah, I. Borova, D. Langdon, R. Moussa, M. Beane, T.-H. Huang, B. Routledge, *et al.*, “Finqa: A dataset of numerical reasoning over financial data,” *arXiv preprint arXiv:2109.00122*, 2021.
- [12] D. Lu, H. Wu, J. Liang, Y. Xu, Q. He, Y. Geng, M. Han, Y. Xin, and Y. Xiao, “Bbt-fin: Comprehensive construction of chinese financial domain pre-trained language model, corpus and benchmark,” *arXiv preprint arXiv:2302.09432*, 2023.
- [13] L. Zhang, W. Cai, Z. Liu, Z. Yang, W. Dai, Y. Liao, Q. Qin, Y. Li, X. Liu, Z. Liu, *et al.*, “Fineval: A chinese financial domain knowledge evaluation benchmark for large language models,” *arXiv preprint arXiv:2308.09975*, 2023.
- [14] Q. Xie, W. Han, Z. Chen, R. Xiang, X. Zhang, Y. He, M. Xiao, D. Li, Y. Dai, D. Feng, *et al.*, “The finben: An holistic financial benchmark for large language models,” *arXiv preprint arXiv:2402.12659*, 2024.
- [15] Y. Lei, J. Li, D. Cheng, Z. Ding, and C. Jiang, “Cfbenchmark: Chinese financial assistant benchmark for large language model,” *arXiv preprint arXiv:2311.05812*, 2023.
- [16] J. Ding, C. Ding, W. Pang, B. Xiao, Z. Liu, P. Chen, J. Chen, T. Yuan, J. Guan, Y. Jiang, *et al.*, “Cnfinbench: A benchmark for safety and compliance of large language models in finance,” *arXiv preprint arXiv:2512.09506*, 2025.

- [17] G. Bhatia, E. M. B. Nagoudi, H. Cavusoglu, and M. Abdul-Mageed, “Fintral: A family of gpt-4 level multimodal financial large language models,” *arXiv preprint arXiv:2402.10986*, 2024.
- [18] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.
- [19] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, “Llama: Open and efficient foundation language models,” *arXiv preprint arXiv:2302.13971*, 2023.
- [20] Z. Du, Y. Qian, X. Liu, M. Ding, J. Qiu, Z. Yang, and J. Tang, “Glm: General language model pretraining with autoregressive blank infilling,” *arXiv preprint arXiv:2103.10360*, 2021.
- [21] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, “Qlora: Efficient finetuning of quantized llms,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [22] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, “Xlnet: Generalized autoregressive pretraining for language understanding,” *Advances in neural information processing systems*, vol. 32, 2019.
- [23] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, *et al.*, “A survey on evaluation of large language models,” *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, pp. 1–45, 2024.
- [24] J. Li, Y. Lei, Y. Bian, D. Cheng, Z. Ding, and C. Jiang, “Ra-cfgpt: Chinese financial assistant with retrieval-augmented large language model,” *Frontiers of Computer Science*, vol. 18, no. 5, p. 185350, 2024.
- [25] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, “Improving language understanding by generative pre-training,” 2018.
- [26] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
- [27] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [28] S. Wu, O. Irsoy, S. Lu, V. Dabrovolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann, “Bloomberggpt: A large language model for finance,” *arXiv:2303.17564*, 2023.
- [29] H. Yang, X.-Y. Liu, and C. D. Wang, “Fingpt: Open-source financial large language models,” *FinLLM at IJCAI*, 2023.
- [30] Q. Xie, W. Han, X. Zhang, Y. Lai, M. Peng, A. Lopez-Lira, and J. Huang, “Pixiu: A comprehensive benchmark, instruction dataset and large language model for finance,” in *Advances in Neural Information Processing Systems*, 2024.
- [31] X. Zhang and Q. Yang, “Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters,” in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 4435–4439, 2023.
- [32] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “Lora: Low-rank adaptation of large language models,” in *ICLR*, 2022.
- [33] H. Fan, M. Li, Z. Zhang, L. Cheng, Y. Ye, W. Liu, and D. Liu, “M2wllm: Multi-modal multi-task ultra-short-term wind power prediction algorithm based on large language model,” *Information Fusion*, p. 103541, 2025.
- [34] H. Fan, Y. Chai, C. Liu, W. Liu, Z. Zhang, W. Run, and D. Liu, “Ev-stllm: Electric vehicle charging forecasting based on spatio-temporal large language models with multi-frequency and multi-scale information fusion,” *arXiv preprint arXiv:2507.09527*, 2025.
- [35] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, “A survey on multimodal large language models,” *arXiv preprint arXiv:2306.13549*, 2023.

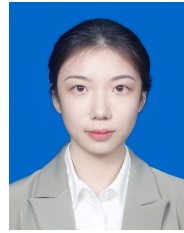
- [36] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021.
- [37] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, “Align before fuse: Vision and language representation learning with momentum distillation,” *Advances in neural information processing systems*, vol. 34, pp. 9694–9705, 2021.
- [38] H. Bao, W. Wang, L. Dong, Q. Liu, O. K. Mohammed, K. Aggarwal, S. Som, S. Piao, and F. Wei, “Vlmo: Unified vision-language pre-training with mixture-of-modality-experts,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 32897–32912, 2022.
- [39] J. Li, D. Li, C. Xiong, and S. Hoi, “Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” in *International conference on machine learning*, pp. 12888–12900, PMLR, 2022.
- [40] J. Li, D. Li, S. Savarese, and S. Hoi, “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” in *International conference on machine learning*, pp. 19730–19742, PMLR, 2023.
- [41] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, and S. Hoi, “Instructblip: Towards general-purpose vision-language models with instruction tuning,” *Advances in Neural Information Processing Systems*, vol. 36, 2023.
- [42] Y. Jin, K. Xu, L. Chen, C. Liao, J. Tan, B. Chen, C. Lei, A. Liu, C. Song, X. Lei, *et al.*, “Unified language-vision pretraining with dynamic discrete visual tokenization,” *arXiv preprint arXiv:2309.04669*, 2023.
- [43] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, “Minigt-4: Enhancing vision-language understanding with advanced large language models,” *arXiv preprint arXiv:2304.10592*, 2023.
- [44] H. Li, Y. Cao, Y. Yu, S. R. Javaji, Z. Deng, Y. He, Y. Jiang, Z. Zhu, K. Subbalakshmi, G. Xiong, J. Huang, L. Qian, X. Peng, Q. Xie, and J. W. Suchow, “Investorbench: A benchmark for financial decision-making tasks with llm-based agent,” 2024.
- [45] L. Xu, L. Zhu, Y. Wu, and H. Xue, “Supercluefin: Graded fine-grained analysis of chinese llms on diverse financial tasks and applications,” *arXiv preprint arXiv:2404.19063*, 2024.
- [46] Q. Ye, H. Xu, J. Ye, M. Yan, H. Liu, Q. Qian, J. Zhang, F. Huang, and J. Zhou, “mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration,” *arXiv preprint arXiv:2311.04257*, 2023.
- [47] W. Yu, Z. Yang, L. Li, J. Wang, K. Lin, Z. Liu, X. Wang, and L. Wang, “Mm-vet: Evaluating large multimodal models for integrated capabilities,” *arXiv preprint arXiv:2308.02490*, 2023.
- [48] P. P. Liang, Y. Lyu, X. Fan, Z. Wu, Y. Cheng, J. Wu, L. Chen, P. Wu, M. A. Lee, Y. Zhu, *et al.*, “Multibench: Multiscale benchmarks for multimodal representation learning,” *arXiv preprint arXiv:2107.07502*, 2021.
- [49] L. Li, J. Lei, Z. Gan, L. Yu, Y.-C. Chen, R. Pillai, Y. Cheng, L. Zhou, X. E. Wang, W. Y. Wang, *et al.*, “Value: A multi-task benchmark for video-and-language understanding evaluation,” *arXiv preprint arXiv:2106.04632*, 2021.
- [50] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, C. Wei, B. Yu, R. Yuan, R. Sun, M. Yin, B. Zheng, Z. Yang, Y. Liu, W. Huang, H. Sun, Y. Su, and W. Chen, “Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi,” 2024.
- [51] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, K. Chen, and D. Lin, “Mmbench: Is your multimodal model an all-around player?,” 2024.
- [52] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, J. Yang, X. Zheng, K. Li, X. Sun, *et al.*, “Mme: A comprehensive evaluation benchmark for multimodal large language models,” in *The Thirty-ninth Annual Conference on Neural*

Information Processing Systems Datasets and Benchmarks Track, 2025.

- [53] L. Chen, J. Li, X. Dong, P. Zhang, C. He, J. Wang, F. Zhao, and D. Lin, “Sharegpt4v: Improving large multi-modal models with better captions,” *arXiv preprint arXiv:2311.12793*, 2023.
- [54] D. Yi, G. Zhu, C. Ding, Z. Li, D. Yi, and J. Wang, “Mme-industry: A cross-industry multimodal evaluation benchmark,” 2025.
- [55] Z. Gan, Y. Lu, D. Zhang, H. Li, C. Liu, J. Liu, J. Liu, H. Wu, C. Fu, Z. Xu, R. Zhang, and Y. Dai, “Mme-finance: A multimodal finance benchmark for expert-level understanding and reasoning,” 2024.
- [56] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, “G-eval: Nlg evaluation using gpt-4 with better human alignment,” *arXiv preprint arXiv:2303.16634*, 2023.
- [57] H. Lu, W. Liu, B. Zhang, B. Wang, K. Dong, B. Liu, J. Sun, T. Ren, Z. Li, Y. Sun, *et al.*, “DeepSeek-VL: towards real-world vision-language understanding,” *arXiv preprint arXiv:2403.05525*, 2024.
- [58] X. Dong, P. Zhang, Y. Zang, Y. Cao, B. Wang, L. Ouyang, X. Wei, S. Zhang, H. Duan, M. Cao, *et al.*, “InternLM-XComposer2: Mastering free-form text-image composition and comprehension in vision-language large model,” *arXiv preprint arXiv:2401.16420*, 2024.
- [59] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, and *et al.*, “The llama 3 herd of models,” 2024.
- [60] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, and *et al.*, “Qwen2.5-vl technical report,” 2025.
- [61] Y. Li, J. Liu, T. Zhang, T. Zhang, S. Chen, T. Li, Z. Li, L. Liu, and *et al.*, “Baichuan-omni-1.5 technical report,” 2025.



Jiangtong Li received the BEng degree and PhD degree in computer science from Shanghai Jiao Tong University in 2019, and 2023. He is currently a postdoctoral associate in Computer Science and Technology Department in Tongji University, Shanghai, China. His current research interests cover multi-modal modeling, large language model, graph learning and big data in finance.



Yiyun Zhu Yiyun Zhu received the BEng degree from Tongji University, Shanghai, China. She is currently pursuing the master’s degree in computer science and technology with the School of Computer Science and Technology, Tongji University. Her research interests include large language model, financial technology, and big data.



Dawei Cheng received the PhD Degree in computer science from Shanghai Jiao Tong University, Shanghai, China, in 2018. He was a postdoctoral associate at MoE key lab of artificial intelligence, department of computer science, Shanghai Jiao Tong University from 2018 to 2020. He is an associate professor with the department of computer science and technology, Tongji University, Shanghai, China. His research fields include data mining, graph learning and big data in finance.



Zhijun Ding (Senior Member, IEEE) received the PhD degree from Tongji University, Shanghai, China, in 2007. Currently, he is a professor with the Department of Computer Science and Technology, Tongji University, Shanghai, China. His research interests include formal methods, petri nets, services computing, and workflow. He has published more than 100 papers in domestic and international academic journals and conference proceedings.



Changjun Jiang received the PhD degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 1995. He is currently a Professor with the Department of Computer Science and Technology, Tongji University, Shanghai, China. His current research interests include concurrency theory, formal verification of software, service-oriented computing, big data in finance, intelligent systems, financial risk management and big data computing.