FinMME: Benchmark Dataset for Financial Multi-Modal Reasoning Evaluation

Junyu Luo[1,2*], Zhizhuo Kou[3*], Liming Yang[2*], Xiao Luo[4], Jinsheng Huang[1,2], Zhiping Xiao[5†], Jingshu Peng[3], Chengzhong Liu[3], Jiaming Ji[3], Xuanzhe Liu[2], Sirui Han[3†], Ming Zhang[1,2†], Yike Guo[3],

[1] State Key Laboratory for Multimedia Information Processing, PKU-Anker LLM Lab
[2] School of Computer Science, Peking University [3] HKUST
[4] University of California, Los Angeles [5] University of Washington

Dataset: `https://huggingface.co/datasets/luojunyu/FinMME`

Abstract

Multimodal Large Language Models (MLLMs) have experienced rapid development in recent years. However, in the financial domain, there is a notable lack of effective and specialized multimodal evaluation datasets. To advance the development of MLLMs in the finance domain, we introduce FinMME, encompassing more than 11,000 high-quality financial research samples across 18 financial domains and 6 asset classes, featuring 10 major chart types and 21 subtypes. We ensure data quality through 20 annotators and carefully designed validation mechanisms. Additionally, we develop FinScore, an evaluation system incorporating hallucination penalties and multi-dimensional capability assessment to provide an unbiased evaluation. Extensive experimental results demonstrate that even state-of-the-art models like GPT-4o exhibit unsatisfactory performance on FinMME, highlighting its challenging nature. The benchmark exhibits high robustness with prediction variations under different prompts remaining below 1%, demonstrating superior reliability compared to existing datasets. Our dataset and evaluation protocol are available at `https://github.com/luo-junyu/FinMME`.

1 Introduction

Multimodal Large Language Models (MLLMs) have demonstrated remarkable progress in data comprehension and understanding (Fu et al., 2024c), with their capabilities being evaluated through various benchmarks such as MME (Fu et al., 2024a), SEED (Li et al., 2024a), MMC (Liu et al., 2023), MMMU (Yue et al., 2024). The establishment of effective datasets and benchmarks has been instrumental in guiding model optimization and comparative analysis, significantly accelerating the development of multimodal large models.

The financial domain(Chen et al., 2022; Li et al., 2023b), characterized by its knowledge-intensive nature and rich multimodal data, presents an ideal application space for MLLMs, particularly in areas such as research report analysis (Zhao et al., 2024), risk forecasting (Sawhney et al., 2020), and market analysis (Liu et al., 2024a). However, the financial sector poses unique challenges due to its inherent complexity, higher data and knowledge density, and extensive domain expertise requirements, necessitating specialized domain-specific evaluation frameworks. Despite this need, there is currently a notable absence of comprehensive, high-quality multimodal datasets specifically designed for evaluating and optimizing MLLMs in the financial domain.

It is non-trivial to design a comprehensive, high-quality financial multimodal dataset, which presents several fundamental challenges:

- **Data Volume:** Limited volume could lead to high variance in results and limited stability.
- **Data Quality:** MLLM annotated datasets may introduce hallucination-based errors. Moreover, high-knowledge-density financial multimodal datasets remain notably underexplored.
- **Domain-specificity and Difficulty:** While MLLMs achieve 80-90% accuracy on general benchmarks (Masry et al., 2022; Li et al., 2023a; Liu et al., 2024b), financial tasks require both higher accuracy and domain expertise, demanding more rigorous evaluation scenarios.

To address these challenges, we introduce FinMME, a comprehensive and high-quality financial multimodal dataset with the following key features: ❶ **Comprehensive Financial Knowledge Coverage**: FinMME incorporates more than 11,000 rigorously selected financial samples spanning 18 core domains and 6 asset classes. Each sample contains financial charts (10 major types with 21

*Equal contribution.
†Corresponding author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 29465–29489

subtypes), professional research descriptions, hierarchical metadata, and QA annotations, reflecting real-world financial analysis workflows. ❷ **High Data Quality**: We employed 20 annotators and implemented carefully designed validation mechanisms, maintaining annotation error rates below 1% for critical questions. ❸ **Innovative Quality Control**: We leverage MLLMs' external consistency to enhance annotation quality and efficiency, with expert review for cases where multiple models and human annotators disagree. ❹ **Novel Evaluation Metrics**: We introduce a hierarchical evaluation framework encompassing comprehensive perception, fine-grained analysis, and cognitive reasoning. Additionally, we designed FinScore, which provides unbiased evaluation across multiple financial domains while incorporating hallucination penalties to address the financial sector's low tolerance for inaccuracies. ❺ **Challenge and Effectiveness**: Extensive experiments on FINMME demonstrate that even leading MLLMs (GPT-4o, Germini Flash and Claude 3.5 Sonnet) achieve just over 50% performance, highlighting the significant challenges and necessity for multimodal research in the financial domain. We tested 6 proprietary models and 11 open-source models, the prediction standard deviation across different prompts remains below 1%, confirming FINMME's robustness.

In summary, FINMME establishes a new benchmark for financial MLLMs through its comprehensive data coverage, rigorous annotation process, and hierarchical evaluation framework, advancing multimodal capabilities in specialized financial applications.

2 Related Work

2.1 Multi-modal Large Language Models

Recent advances in Multi-modal Large Language Models (MLLMs) have demonstrated remarkable capabilities in unified visual-linguistic understanding as agentic AI (Luo et al., 2025), with open-source models like QwenVL (Bai et al., 2023), Vita (Fu et al., 2024b), VILA (Lin et al., 2024), CogVLM (Wang et al., 2023), and LLaVA (Li et al., 2024b), alongside proprietary models including GPT-4o[1], Claude 3.5 Sonnet[2] and Gemini (Team et al., 2023) showing strong performance in general domain tasks. However, despite their sophisticated encoder-decoder architectures for cross-modal understanding, our evaluation reveals that these MLLMs significantly underperform in knowledge-intensive financial tasks, highlighting the need for specialized datasets such as FINMME to advance financial MLLMs.

Dataset	**Dataset Volume**	**Human Anno.**	**Specific Domain**	**GPT-4o Performance**
MMStar	1500	✗	✗	62
MM-Vet	218	✗	✗	72
MME	2374	✓	✗	–
MMBench	3217	✓	✗	83
MMC	2126	✓	✗	76
MMMU (Full)	11550	✓	✗	63
MMMU (Finance)	390	✓	Finance	-
MME-Finance	1171	✗	Finance	63
FINMME (Ours)	11099	✓	Finance	47

Table 1: Comparison with existing benchmarks. FINMME provides a comprehensive and high-quality dataset for the financial multimodal domain.

2.2 Multi-Modal Evaluation Datasets

Recent advancements in MLLMs have demonstrated exceptional capabilities across a wide array of complex tasks, including MMStar (Chen et al., 2024b), MM-Vet (Yu et al., 2023), MME (Fu et al., 2024a), MMBench (Liu et al., 2024b), MMC (Liu et al., 2023), MMMU (Yue et al., 2024), and others (Li et al., 2023a, 2024a; Huang et al., 2024). Comprehensive benchmarks are essential not only to gauge progress in general multimodal reasoning but also to pinpoint areas that require further refinement. However, domain-specific evaluation remains limited, particularly in the finance domain, where the high knowledge density and inherent complexity of financial data demand specialized evaluation frameworks. More related background can be found in Appendix A.

Differences from Existing Datasets As shown in Table 1, existing multimodal benchmarks are constrained by data scale, annotation quality, domain coverage and task complexity[3]. While the concurrent work MME-Finance (Gan et al., 2024) also targets financial multimodal evaluation, it faces limitations in data volume and annotation quality. In contrast, FINMME offers a comprehensive, high-quality large-scale dataset specifically designed for financial multimodal tasks. We provide a detailed comparison with existing financial domain datasets in Appendix E to highlight the advantages.

[1]https://openai.com/index/hello-gpt-4o/

[2]https://www.anthropic.com/news/claude-3-5-sonnet

[3]The performance is from official reports or quoted (Fu et al., 2025). MMMU (Finance) is the domain-specific subset.

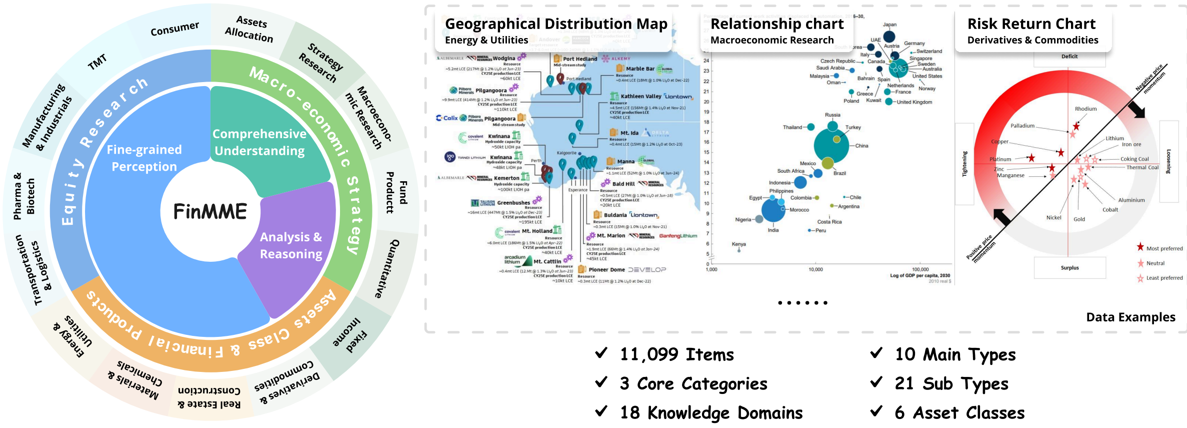


Figure 1: The Comprehensive Taxonomy, Data Examples and Statistical Characteristics of FINMME. The circular taxonomy diagram shows three core cognitive levels, knowledge categories and domains.

3 FINMME Dataset: High-Quality Financial Multi-Modal Dataset

FINMME comprises more than 11,000 high-quality financial multi-modal samples, with each sample consisting of multi-modal metadata and question information. All data undergoes a rigorous quality control process to ensure reliability. Our dataset design was informed by discussions with six financial domain experts (detailed consultation records in Appendix B). This section provides a comprehensive introduction to the FinMME dataset, including detailed data classification and statistics (Section 3.1), question-answer design (Section 3.3), data sources (Section 3.4), annotation process (Section 3.5), and quality control protocols (Section 3.6).

Statistic	Number
Dataset Overview	
Total Samples	11,099
Cognitive Level Distribution	
Comprehensive Understanding	2,333
Fine-grained Perception	6,466
Analysis and Reasoning	2,300
Core Knowledge Domain	
Equity Research	7,601
Macroeconomic Research	1,485
Assets Class and Financial Products	2,013
Unique Images	4,458
Average Question Length	24.1
Average Caption Length	10.8

Table 2: Statistical characteristics of the FINMME dataset, including question types, cognitive levels, and knowledge domains.

3.1 Statistical Characteristics

The multi-modal metadata encompasses financial images, image captions, professional research report descriptions, and fine-grained data labels (*i.e.*, target markets, asset classes, and detailed data class labels). The question information includes problem statements, multiple choice options, standard answers (with unit and error tolerance ranges for calculation questions), and question type labels. Dataset statistics are shown in Table 2.

3.2 Fine-grained Data Labels

Knowledge Domain. FINMME aims to provide comprehensive coverage of financial knowledge domains, encompassing 18 core financial domains: TMT (Technology, Media & Telecom), Consumer, Pharmaceuticals & Biotechnology, Financials, Real Estate & Construction, Industrials & Manufacturing, Energy & Utilities, Materials & Chemicals, Military & Defense, Transportation & Logistics, Macroeconomic Research, Strategy Research, Broad Asset Allocation, Equity Research, Fixed Income, Fixed Income Quantitative, Derivatives & Commodities, and Fund Products. The taxonomy is in Figure 1. This extensive coverage effectively reflects the modern financial knowledge system.

Data Class. FinMME incorporates diverse data classes, categorized into 10 main classes and 21 subclasses. The main classes comprise Time Series, Distribution Charts, Proportional Charts, Relationship Charts, Financial Reports, Risk Analysis, Market Structure, Geographical Charts, Process Flow, *etc.*. To facilitate future research, we have meticulously annotated each image with both main class and subclass categories, with details provided

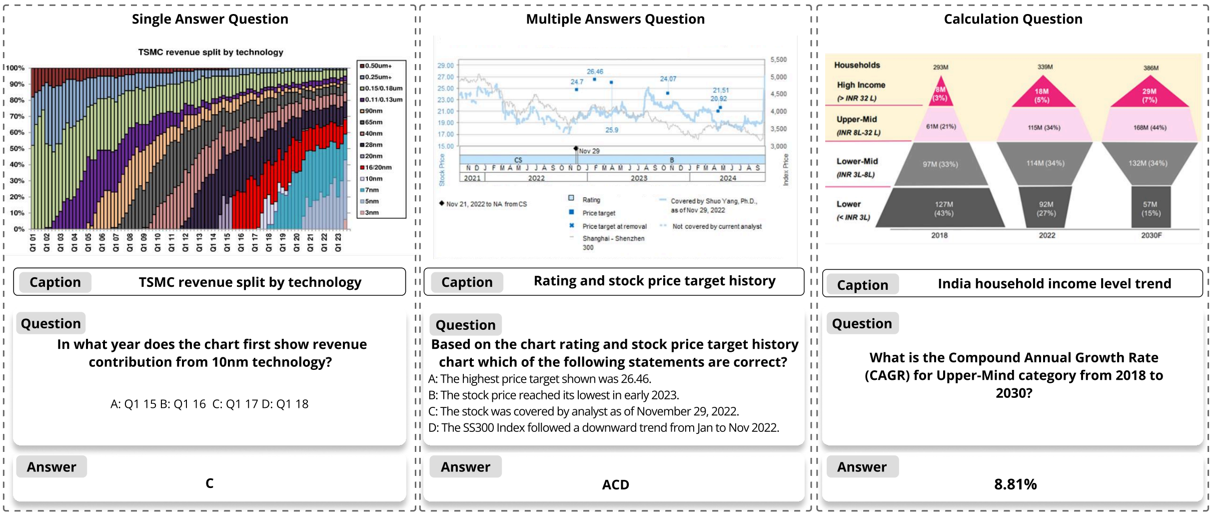


Figure 2: Representative examples of different question types in FINMME dataset.

in Appendix C.

Asset Class. We effectively differentiate the multimodal data according to 6 asset classes to support cross-asset analysis. The dataset covers Equity, Foreign Exchange, Rates, Commodity, Credits, and Cross-Asset. These asset class labels enable targeted model evaluation across different market segments and facilitate the assessment of specialized knowledge in distinct financial instruments.

3.3 Question-Answer Design

We establish a hierarchical evaluation framework to comprehensively assess MLLMs' capabilities in the financial domain. This framework encompasses three fundamental dimensions:

Comprehensive Perception. This dimension evaluates models' ability to perform temporal sequence recognition, horizontal comparisons, holistic discrimination, and multi-chart analysis. The assessment is primarily conducted through multiple-choice questions (single answer and multiple answers), focusing on models' capacity to comprehend and interpret complex financial visualizations and their interrelationships.

Fine-grained Perception. This aspect examines numerical extraction and local variation analysis capabilities. The evaluation utilizes multiple-choice questions (single answer and multiple answers) to assess models' precision in identifying and analyzing specific data points and localized patterns within financial contexts.

Cognition and Reasoning. This dimension encompasses data inference, cross-modal understanding, trend prediction, causal analysis, scenario-based decision support, and hypothesis analysis. The assessment combines computational problems and multiple-choice questions to evaluate models' advanced reasoning capabilities in financial scenarios, including their ability to synthesize information across modalities and make informed predictions.

3.4 Data Sources

Adhering to compliance principles, we collected over $7,000$ professional research reports and web page screenshots through a hybrid approach combining manual curation and automated crawling, from which we extracted high-quality financial images and associated text. Throughout the collection process, we prioritized copyright compliance and selected materials authorized for public dissemination. All data underwent a rigorous three-stage cleaning process: automated deduplication, format standardization, and manual review, ensuring the authoritativeness and legality of data sources.

3.5 Annotation Process

Annotation Team. We recruited a team of 20 annotators, consisting of 12 Junior annotators and 8 Experts. Junior annotators with basic finance knowledge were responsible for question review, reformulation, and independent problem-solving. The Expert group included (i) 4 people from academia specializing in STEM and finance, holding at least a master's degree, and (ii) 4 finance industry professionals. These experts were tasked with dataset question selection, quality assessment, and answer verification.

Time Investment. The annotation and review process required approximately 800 cumulative hours

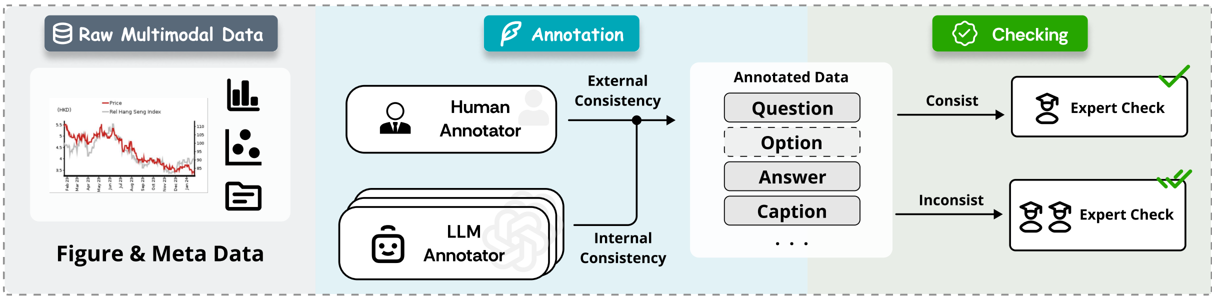


Figure 3: The annotation pipeline of FINMME. The process consists of three main stages: (1) Raw Multimodal Data collection, (2) Annotation through parallel human and LLM annotators to ensure external and internal consistency, and (3) Quality Control checking where expert reviewers validate consistent annotations and resolve inconsistencies.

of work from the 20-member team, with time estimates aggregated from individual contributions.

3.6 Quality Control Protocol

We designed an innovative quality control methodology, as illustrated in Figure 3. While ensuring dataset quality, we use LLMs to achieve a more efficient dataset construction process through a three-stage pipeline. First, we collect and prepare the raw multimodal data. Second, in the annotation stage, we employ a parallel annotation strategy where both human annotators and multiple LLM annotators independently process the data. This dual-track approach helps establish both external consistency (through human annotations) and internal consistency (through multiple LLM predictions). The annotated data includes questions, options, answers, captions, and other relevant metadata. Finally, in the quality control stage, we implement a consistency-based review process: when human and LLM annotations align, a single expert performs a validation check; when discrepancies occur, multiple experts conduct a thorough review to determine the final ground truth. This systematic approach ensures high-quality annotations while optimizing the efficiency of expert involvement.

3.7 Summary

FINMME distinguishes itself from existing datasets through three key characteristics: superior quality, comprehensive coverage, and fine-grained label annotations. The dataset features high-quality multi-modal data spanning diverse financial knowledge domains, accompanied by meticulously annotated classifications and question-answer pairs. These distinctive attributes enable effective evaluation of MLLMs' performance in complex financial scenarios. The combination of the above positions FINMME as a robust benchmark for assessing multi-modal language models' capabilities in professional financial applications.

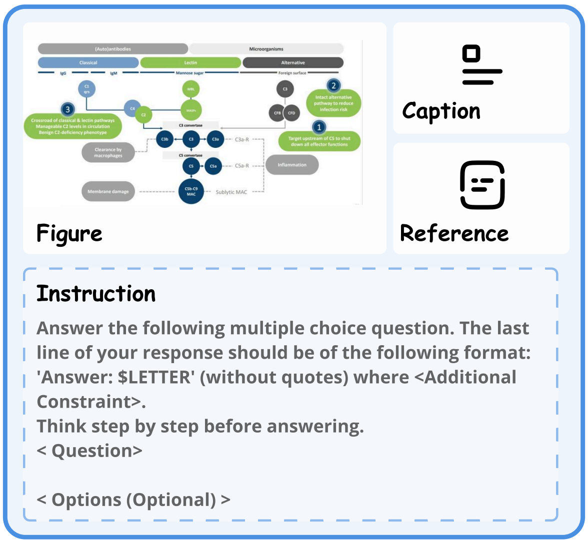


Figure 4: Illustration of the evaluation prompt template.

4 FINMME Benchmark: Comprehensive Financial Multi-Modal Evaluation

To ensure comprehensive evaluation, we employ a combination of multiple-choice questions (MCQs) and computational problems. The MCQs include both single-answer and multiple-answer formats, with an increased emphasis on multiple-answer questions compared to existing datasets. This design choice aims to better challenge models and reduce hallucination tendencies, as multiple-answer questions require more precise understanding and exhibit lower tolerance for incorrect selections.

4.1 Hallucination Penalty

For multiple-answer questions, we introduce a scoring mechanism that effectively balances reward for correct answers with penalties for over-selection. The raw score for a single multiple-choice question is calculated as:

$$S_q = \max(0, \frac{c}{n} - \frac{i}{s}), \quad (1)$$

where S_q represents the raw score for a single multiple-choice question, c is the number of correct selections, n is the total number of options, i is the number of incorrect selections, and s is the total selections made by the model. This formulation penalizes hallucination by reducing scores proportionally to incorrect selection ratios while normalizing based on the total options available.

4.2 Knowledge-unbiased Evaluation

Financial knowledge domains inherently vary in complexity and difficulty. For instance, quantitative analysis in derivatives typically presents greater challenges than basic equity research. To address these variations and ensure fair evaluation, we implement domain-normalized scoring:

$$F = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N_k}\sum_{i=1}^{N_k} S_{k,i}\,, \quad (2)$$

where $S_{k,i}$ represents the score of the i-th question in domain k, N_k is the total number of questions in domain k, and K is the total number of domains. This formulation first calculates the average performance within each domain, then takes the mean across all domains, ensuring each knowledge domain contributes equally to the final score regardless of its number of questions.

4.3 FinScore

Financial applications demand both high accuracy and low hallucination due to the critical nature of investment decisions. To address this dual requirement, we introduce FinScore ($\mathcal{F}$) that combines domain-normalized performance with hallucination penalties, reflecting a model's practical value in financial contexts.

We first define the hallucination penalty rate P_H, which represents the average ratio of incorrect selections across the dataset:

$$P_H = \text{mean}\left(\frac{i}{s}\right), \quad (3)$$

where the mean is calculated across all questions in the dataset. The final FinScore combines the domain-normalized score with the hallucination penalty:

$$\mathcal{F} = F \cdot (1 - P_H)\,, \quad (4)$$

where F is the domain-normalized average score across all questions and P_H is the hallucination penalty rate. This multiplicative combination ensures that models are evaluated on both accuracy and reliability, with a strong emphasis on penalizing hallucination. In financial applications where incorrect predictions can lead to significant risks, models that hallucinate receive substantially lower scores regardless of their knowledge accuracy, reflecting the critical importance of reliable analysis.

5 Experiment

5.1 Competing MLLMs

To comprehensively evaluate the performance of current multimodal large language models in the financial domain, we conducted experiments across a diverse range of model architectures and parameter scales. Our evaluation encompasses both proprietary and open-source models. The proprietary models include GPT4o[4], GPT4o-mini, Gemini Flash 2.0 (Team et al., 2023), Claude 3.5 Sonnet[5], Claude 3.5 Haiku[6] and Doubao-1.5V Pro[7]. For open-source alternatives, we selected Qwen2.5 VL 72B (Yang et al., 2024), InternVL 25-8B[8],MiniCPM-O26 (Hu et al., 2024), DeepSeekVL-2 (Wu et al., 2024), Qwen-2-VL-72B (Wang et al., 2024), Qwen-2-VL-7B (Wang et al., 2024), DeepseekVL-2 Small (Wu et al., 2024), Phi-3 128K (Abdin et al., 2024), Phi-3.5 V (Abdin et al., 2024) and DeepSeekVL-2 Tiny (Wu et al., 2024).

5.2 Evaluation Methods

Our experimental evaluation was conducted separately for proprietary and open-source models. Proprietary models and larger open-source models were evaluated through commercial API calls, while smaller open-source models were deployed locally. All local experiments were performed on a single NVIDIA H100-level GPU. We utilized vLLM for efficient local deployment and inference.

5.3 Main Results and Key Insights

Proprietary Models' Performance. Proprietary models demonstrate superior performance, with

[4]https://openai.com/index/hello-gpt-4o/
[5]https://www.anthropic.com/news/claude-3-5-sonnet
[6]https://www.anthropic.com/claude/haiku
[7]https://team.doubao.com/zh/special/doubao_1_5_pro
[8]https://internvl.opengvlab.com/

Method	Compre.	FG	Reason.	Single.	Multi.	Cal.	Avg.	FinScore
Proprietary Models								
Gemini Flash 2.0	**49.89**	**59.07**	**48.71**	**63.73**	**54.11**	35.59	**51.85**	**20.10**
Claude 3.5 Sonnet	45.99	55.28	43.35	59.61	47.59	**37.35**	48.20	15.61
GPT-4o	44.33	53.49	42.24	58.49	45.74	35.06	46.56	15.34
DouBao-1.5V Pro	44.42	54.33	43.48	58.55	47.36	35.43	47.26	15.03
GPT-4o Mini	41.91	48.47	42.88	52.38	45.42	31.27	43.72	11.70
Claude 3.5 Haiku	29.09	36.21	28.22	41.75	34.98	6.71	29.49	6.41
Open-source Models								
Qwen2.5-VL 72B	**49.64**	**60.25**	**49.44**	**65.06**	**54.26**	**36.60**	**52.54**	**20.87**
Qwen2-VL 72B	37.11	51.68	33.92	58.05	36.81	32.77	41.72	11.50
InternVL 2.5-8B	37.96	51.83	35.33	59.43	38.60	28.24	41.90	10.42
MiniCPM-O 2.6	37.71	53.17	35.98	60.21	39.05	30.31	42.74	9.77
DeepSeekVL-2	32.91	51.46	29.63	60.41	35.73	18.33	38.08	8.28
Qwen2-VL 7B	34.14	48.17	31.73	54.88	35.07	26.32	41.80	6.91
Qwen2.5-VL 3B	32.53	52.55	30.70	61.29	31.98	30.15	39.87	6.95
DeepSeekVL-2 Small	34.14	51.00	31.55	59.73	34.81	26.85	38.18	6.11
Phi-3 V	27.52	45.59	26.97	54.35	26.57	25.73	34.45	3.87
Phi-3.5 V	25.73	43.37	26.46	51.84	24.24	27.12	33.13	2.85
DeepSeekVL-2 Tiny	23.06	31.48	21.14	37.14	25.97	7.88	24.45	2.05

Table 3: **Performance Comparison** across different evaluation dimensions.

Gemini Flash 2.0 leading at average score and FinScore. The performance gap between proprietary and open-source models is most pronounced in multi-turn reasoning tasks.

Open-source Models' Performance. Qwen2.5-VL 72B achieves competitive performance comparable to proprietary models, particularly excelling in fine-grained perception and single-turn tasks.

Task-Specific Performance. All models perform better in single-turn tasks compared to multi-turn reasoning, with an average performance gap of 20-25%. Calculation questions remain the most challenging dimension, with even top models achieving below 40% accuracy.

Financial Domain Adaptation. FinScore reveals significant gaps in financial domain expertise, with most open-source models scoring below 12, indicating room for improvement in financial knowledge and hallucination control.

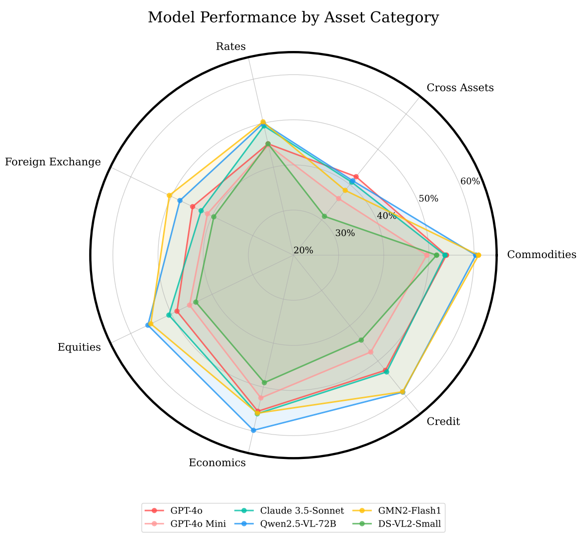


Figure 5: The radar chart of the asset class distribution of the dataset.

5.4 Domain-specific Performance Analysis

Through performance evaluation across 16 different industry domains, we observe significant variations in model capabilities. In traditional industrial sectors such as pharmaceuticals, energy, and metals, models generally demonstrate strong performance (Gemini Flash 2.0), and energy and metals sectors consistently maintain scores above 50. However, economics and fixed income sectors present significant challenges, with even top mod-

Method	Energy	Estate	Constr.	Metals	Chem.	Econo.	Asset	Fixed	Equity	Industrials	TMT	Trans.	General	Cons.	Pharma	Others
Method	Energy	Estate	Deriva.	Meteri.	Macroe.	Assets	Strate.	Fixed.	Equity	Indust.	TMT	Trans.	Financial	Consum.	Pharma	Others
Proprietary Models																
Gemini Flash 2.0	**55.57**	**56.39**	**52.63**	**60.48**	**54.54**	40.57	**57.26**	42.07	**52.24**	**52.32**	**60.75**	**54.75**	**53.64**	**65.50**	**55.43**	**63.33**
Claude 3.5 Sonnet	49.83	50.38	46.20	57.19	53.35	41.51	54.03	42.07	47.86	47.54	56.65	52.33	49.21	59.91	48.91	53.33
GPT-4o	50.00	52.63	47.95	57.19	52.57	41.51	54.03	42.07	46.03	44.02	51.93	48.88	49.04	57.81	45.59	60.00
DouBao-1.5V Pro	48.61	50.38	46.78	56.29	54.22	**44.34**	47.58	**44.14**	48.13	46.55	50.43	48.70	49.33	58.97	46.02	53.33
GPT-4o Mini	45.82	48.12	45.61	49.40	49.17	33.96	45.97	37.24	44.31	45.71	43.85	47.84	45.14	51.63	44.28	53.33
Claude 3.5 Haiku	29.79	32.33	33.92	35.33	40.49	32.08	40.32	32.41	31.96	30.94	34.29	39.21	29.30	33.92	28.22	36.67
Open-source Models																
Qwen-VL-2.5 72B	**56.79**	**57.14**	**51.46**	**60.78**	**58.17**	**44.34**	**54.84**	**38.62**	**53.05**	**53.87**	**58.01**	**55.27**	**54.57**	**64.80**	**57.16**	**66.67**
Qwen-VL-2 72B	43.55	41.35	44.44	56.89	49.88	30.19	39.52	37.24	42.32	43.18	45.96	47.15	42.40	53.73	42.69	30.00
InternVL-2.5 8B	44.77	42.86	46.20	50.30	47.20	37.74	51.61	37.24	43.60	41.77	46.83	45.08	44.85	51.52	47.03	56.67
MiniCPM-O 2.6	47.21	49.62	43.27	50.00	47.43	35.85	49.19	**38.62**	44.18	42.05	47.33	45.94	47.00	55.36	44.86	43.33
DeepSeekVL-2	39.72	47.37	42.69	47.01	43.80	28.30	50.00	37.93	41.92	42.48	45.84	42.66	41.29	49.53	41.82	53.33
Qwen-VL-2 7B	45.99	42.11	41.52	44.01	42.38	30.19	41.13	34.48	38.91	42.05	45.09	40.93	41.35	48.72	41.24	40.00
Qwen-VL-2.5 3B	40.94	48.87	39.77	50.90	45.46	29.25	44.35	37.24	41.82	41.49	47.20	40.93	42.92	51.98	44.57	43.33
DeepSeekVL-2 Small	45.99	48.12	45.03	51.20	45.15	35.85	51.61	33.79	42.73	44.59	47.20	40.59	42.17	51.40	40.52	40.00
Phi-3 V	36.76	39.10	40.94	43.11	42.07	27.36	43.55	31.03	35.67	36.85	38.14	37.31	36.92	42.66	37.05	46.67
Phi-3.5 V	34.15	41.35	34.50	40.42	40.17	32.08	38.71	33.79	33.72	33.47	37.02	37.82	35.06	42.07	34.01	46.67
DeepSeekVL-2 Tiny	31.53	30.08	26.90	29.64	28.89	19.81	36.29	27.59	27.57	26.44	27.20	27.29	24.64	29.95	25.47	43.33

Table 4: **Domain-specific Performance Comparison** across different sectors and industries.

els scoring below 45 points, indicating persistent difficulties in complex financial reasoning tasks. Notably, while smaller models consistently underperform across all domains, open-source models such as Qwen-VL-2.5 72B demonstrate competitive performance against proprietary models in specific domains, particularly in energy and metals. These findings not only reveal the current importance of model scale for domain expertise but also suggest promising developments in open-source models' ability to handle specialized tasks.

5.5 Asset Class Analysis

Analysis of the asset class distribution radar chart reveals notable performance variations across financial asset types. Models demonstrate strongest performance in the Commodities sector, followed by moderate performance in Credit and Rates categories. However, models show relatively weaker performance in the Foreign Exchange and Economics domains.

Notably, GPT4o and Claude 3.5-Sonnet exhibit robust overall capabilities across most asset classes. In contrast, smaller-scale models show acceptable performance only in specific categories like Commodities, while demonstrating lower overall effectiveness. These findings highlight the persistent disparities in multimodal large language models' comprehension capabilities within the financial domain, particularly in more complex areas like Foreign Exchange and Economics, indicating substantial room for improvement.

Method	Single.	Multi.	Cal.	Avg.
GPT-4o	58.49±0.93	45.74±0.77	35.06±0.58	46.56±0.64
Qwen2-VL 7B	58.05±0.85	36.81±0.86	32.77±0.62	41.72±0.57

Table 5: **Model Performance with Standard Deviations** with 5 runs.

5.6 Stability Analysis

To assess the robustness and reliability of our evaluation framework, we conducted multiple rounds of testing and analyzed the standard deviations of model performance across different dimensions. As shown in Table 5, both GPT-4o and Qwen2-VL 7B demonstrate remarkable stability in their performance. The standard deviations across all evaluation dimensions remain consistently below 1%, with GPT-4o showing variations between 0.58% and 0.93%, and Qwen2-VL 7B exhibiting fluctuations between 0.57% and 0.86%. These low variance levels indicate the high reliability and reproducibility of our evaluation framework, while also confirming the consistency of model behaviors across multiple test runs. The consistently low standard deviations across different model scales further validate the robustness of our evaluation methodology and the quality of our dataset.

6 Conclusion

This paper introduces FinMME, a comprehensive multimodal evaluation framework for the financial domain, comprising high-quality samples across 18 core financial domains. Our experiments demonstrate that leading MLLMs achieve unsatisfactory performance on FINMME, highlighting significant room for improvement in financial applications. The proposed FinScore metric, incorporating hallucination penalties and domain-normalized scoring, provides a robust evaluation framework for financial tasks, while maintaining prediction stability with low standard deviations across different prompts. Future work will focus on expanding dataset coverage, enhancing evaluation metrics, and promoting FINMME's application in real-world financial analysis scenarios.

Limitations

Despite FINMME's carefully curated nature and substantial sample size, we acknowledge several limitations. Our evaluation methodology relies primarily on multiple-choice questions and calculations, which enables objective assessment but may not fully capture the complexity of real-world financial analysis tasks. Complex financial concepts posed interpretation difficulties even for knowledgeable annotators, potentially introducing subtle biases despite our quality control protocols. While FINMME covers diverse financial domains, it may not capture all scenarios encountered in financial work due to the vast and evolving nature of the industry, and currently lacks integration with audio/video content and real-time data analysis. Finally, although our stability analysis demonstrates robustness with high-quality inputs, these findings may not generalize to noisy or distorted inputs, highlighting that robustness to perturbations represents an important research direction building upon FINMME.

Acknowledgments

This paper is partially supported by grants from the National Key Research and Development Program of China with Grant No. 2023YFC3341203, the National Natural Science Foundation of China (NSFC Grant Number 62276002), HKUST Start-up Fund (R9911), Theme-based Research Scheme grant (No.T45-205/21-N), InnoHK funding for Hong Kong Generative AI Research and Development Center, Hong Kong SAR. The authors are grateful to the anonymous reviewers for their efforts and insightful suggestions to improve this article.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Gagan Bhatia, Hasan Cavusoglu, Muhammad Abdul-Mageed, et al. 2024. Fintral: A family of gpt-4 level multimodal financial large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13064–13087.

Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. 2010. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342.

Jian Chen, Peilin Zhou, Yining Hua, Yingxin Loh, Kehui Chen, Ziyuan Li, Bing Zhu, and Junwei Liang. 2024a. Fintextqa: A dataset for long-form financial question answering. *Preprint*, arXiv:2405.09980.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. 2024b. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6279–6292.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024a. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Shaoqi Dong, Xiong Wang, Di Yin, Long Ma, et al. 2024b. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Yangze Li, Zuwei Long, Heting Gao, Ke Li, et al. 2025. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*.

Chaoyou Fu, Yi-Fan Zhang, Shukang Yin, Bo Li, Xinyu Fang, Sirui Zhao, Haodong Duan, Xing Sun, Ziwei Liu, Liang Wang, et al. 2024c. Mme-survey: A comprehensive survey on evaluation of multimodal llms. *arXiv preprint arXiv:2411.15296*.

Ziliang Gan, Yu Lu, Dong Zhang, Haohan Li, Che Liu, Jian Liu, Ji Liu, Haipang Wu, Chaoyou Fu, Zenglin Xu, et al. 2024. Mme-finance: A multimodal finance benchmark for expert-level understanding and reasoning. *arXiv preprint arXiv:2411.03314*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Shengding Hu, Yuge Tu, Xu Han, et al. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *Preprint*, arXiv:2404.06395.

Jinsheng Huang, Liang Chen, Taian Guo, Fu Zeng, Yusheng Zhao, Bohan Wu, Ye Yuan, Haozhe Zhao, Zhihui Guo, Yichi Zhang, et al. 2024. Mmevalpro: Calibrating multimodal benchmarks towards trustworthy and efficient evaluation. *arXiv preprint arXiv:2407.00468*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. Financebench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*.

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. *arXiv preprint arXiv:2311.05812*.

Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. 2024a. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024b. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023b. Large language models in finance: A survey. In *Proceedings of the fourth ACM international conference on AI in finance*, pages 374–382.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. Vila: On pretraining for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699.

Chenghao Liu, Arunkumar Arulappan, Ranesh Naha, Aniket Mahanti, Joarder Kamruzzaman, and In-Ho Ra. 2024a. Large language models and sentiment analysis in financial markets: A review, datasets and case study. *IEEE Access*.

Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. 2023. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning. *arXiv preprint arXiv:2311.10774*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2024b. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

J. Luo, W. Zhang, Y. Yuan, et al. 2025. Large language model agent: A survey on methodology, applications and challenges. *arXiv preprint arXiv:2503.21460*.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Ramit Sawhney, Puneet Mathur, Ayush Mangal, Piyush Khanna, Rajiv Ratn Shah, and Roger Zimmermann. 2020. Multimodal multi-task financial risk forecasting. In *Proceedings of the 28th ACM international conference on multimedia*, pages 456–465.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, et al. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *Preprint*, arXiv:2412.10302.

Qianqian Xie, Weiguang Han, Zhengyu Chen, et al. 2024. Finben: A holistic financial benchmark for large language models. *Preprint*, arXiv:2402.12659.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Zhengyuan Yang, Yijuan Lu, Jianfeng Wang, Xi Yin, Dinei Florencio, Lijuan Wang, Cha Zhang, Lei Zhang, and Jiebo Luo. 2021. Tap: Text-aware pre-training for text-vqa and text-caption. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8751–8761.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

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

Huaqin Zhao, Zhengliang Liu, Zihao Wu, Yiwei Li, Tianze Yang, Peng Shu, Shaochen Xu, Haixing Dai, Lin Zhao, Gengchen Mai, et al. 2024. Revolutionizing finance with llms: An overview of applications and insights. *arXiv preprint arXiv:2401.11641*.

A More Related Work

While traditional multimodal benchmarks focused on specific tasks like captioning (Chen et al., 2015; Plummer et al., 2015), VQA (Hudson and Manning, 2019; Goyal et al., 2017; Bigham et al., 2010), and specialized capabilities (Sidorov et al., 2020; Yang et al., 2021; Li et al., 2019), financial datasets such as ConvFinQA (Chen et al., 2022), FINANCEBENCH (Islam et al., 2023), FinBen (Xie et al., 2024), CFBenchmark (Lei et al., 2023), FinTextQA (Chen et al., 2024a) and MME-Finance (Gan et al., 2024), FinVQA (Bhatia et al., 2024) either focus solely on language models or provide limited coverage of multimodal financial tasks, highlighting the need for comprehensive financial multimodal evaluation frameworks.

B Expert Consultation Process Record

Our research design and validation process was strengthened through extensive consultation with financial industry experts. Through in-depth interviews, we gained valuable insights into real-world financial analysis workflows and information consumption patterns, which directly informed the design of FinMME. The expert panel included diverse professionals from investment banking, hedge funds, and asset management, with experience ranging from 5 to 10+ years:

A Investment Banking Professional with 5+ years of experience in ECM (Equity Capital Markets) and primary market equity issuance

B Hedge Fund Sector Analyst with 10+ years of experience in industry research, specializing in new energy sectors

C Hedge Fund Industry Researcher with 5+ years of experience

D Investment Banking Professional with 5 years of experience in strategic equity and derivatives

E Hedge Fund Industry Researcher with 10 years of experience

F Asset Management Fund Manager with 5 years of experience

Key findings from our expert consultations highlighted several critical aspects that shaped our dataset design:

Information Hierarchy and Consumption Patterns: Experts consistently emphasized the importance of structured information access, typically beginning with executive summaries and investment views before diving into specific areas of interest. This insight directly influenced our hierarchical annotation structure in FINMME.

Visual Data Interpretation: Financial professionals heavily rely on charts and visualizations for trend analysis and comparative studies. Expert A and E particularly noted that visual representations often provide more intuitive insights than textual information, supporting our focus on diverse chart types and comprehensive visual analysis tasks.

Multi-source Validation: Expert C highlighted the practice of cross-referencing multiple sources and independently verifying data, emphasizing the importance of accuracy in financial analysis. This insight reinforced our rigorous quality control mechanisms and the inclusion of hallucination penalties in our evaluation metrics.

Domain-specific Requirements: Experts B and F emphasized the critical role of industry-specific knowledge and policy understanding, validating our approach to include comprehensive coverage across multiple financial domains and asset classes.

Report Quality Variation: Multiple experts noted significant variations in report quality across different sources, particularly between domestic and international research reports. This observation supported our decision to implement strict quality control measures and expert validation processes.

These expert insights were instrumental in developing FINMME's comprehensive structure, ensuring its relevance to real-world financial analysis needs while maintaining high standards of quality and reliability. The consultation process validated our approach to creating a benchmark that effectively evaluates MLLMs' capabilities in handling complex financial tasks.

C Dataset Details

Our dataset organizes financial charts into 10 main categories, each with specific subcategories to facilitate precise classification and analysis. The main categories include Distribution Charts, Financial Charts, Flow Charts, Geographical Charts, Line Charts, Market Structure Charts, Proportional Charts, Relational Charts, Risk Distribution Charts, and Others. Each main category is further divided into specialized subcategories that capture specific visualization techniques and purposes. For example, Distribution Charts include histograms, box plots, and violin plots, while Financial Charts encompass line charts, K-line charts, and area charts. This hierarchical organization enables systematic evaluation of models' capabilities across different visualization types while maintaining clear categorization of financial data representation methods.

D Additional Results

The experimental results demonstrate significant variations in model performance across different knowledge domains in financial analysis. Qwen25vl72b emerged as the leading performer, achieving exceptional scores particularly in consumer sectors and other specialized categories, suggesting that its architectural design and training approach are particularly well-suited for financial multimodal tasks. This performance advantage persisted across multiple domains, indicating robust and generalizable capabilities.

Notably, model size did not consistently correlate with performance effectiveness. This suggests that architectural choices and training strategies may be more crucial than raw model size for financial analysis tasks.

Domain complexity emerged as a significant factor in model performance patterns. Models generally excelled in sectors requiring straightforward analysis, such as consumer goods and TMT sectors, where performance consistently exceeded 50% across leading models. However, significant challenges were observed in complex domains like broad asset allocation and strategy research, where most models struggled to achieve scores above 45%. This performance gap highlights the increasing difficulty models face when dealing with multi-factor analysis and complex financial reasoning.

These findings carry important implications for the future development of multimodal models in finance. The success of specialized architectures like Qwen25vl72b suggests that domain-specific optimization may be more valuable than pursuing larger model sizes. Future research should focus on improving model performance in complex analytical domains while maintaining the strong performance observed in straightforward tasks. Additionally, the results emphasize the need for balanced capabilities across different financial sectors, particularly in areas requiring sophisticated reasoning and multi-factor analysis.

E Dataset Comparison

As shown in Figure 7, we compare FINMME with two other prominent financial multimodal datasets: MME-Finance and MMMU-Finance. The comparison reveals distinct characteristics and use cases for each dataset:

FINMME stands out with its high data quality and comprehensive coverage, containing more than 11,000 items across 3 core categories and 15 knowledge domains. It features professional-grade labeling with fine-grained annotations across 21 sub-types, providing detailed categorization of financial content. The dataset's comprehensive coverage spans 6 asset classes, establishing a structured hierarchy across multiple financial domains. A distinguishing aspect is its rigorous quality control system, implemented through expert validation processes that ensure the highest standards of financial accuracy and relevance.

MME-Finance offers a different focus with 4,080 items and 38 class labels. This dataset primarily emphasizes technical charts and trading data, making it particularly suited for market analysis applications. However, it employs general-purpose labeling without fine-grained annotations, resulting in less detailed categorization compared to FINMME. While it covers various financial aspects, its domain coverage is more limited, and the overall data quality is lower than FINMME, particularly in terms of annotation depth and expert validation.

MMMU-Finance is the most specialized of the three datasets, containing 390 items with a focused scope. It concentrates on fundamental business metrics such as sales, dividends, and investments, making it particularly relevant for corporate financial analysis. The dataset is structured around two primary question types and image types, with coverage limited to two sub-fields. Like MME-Finance, it employs general-purpose labeling without detailed annotations, which constrains its utility for complex financial analysis tasks.

This comparison highlights FINMME's unique position in providing comprehensive, high-quality financial multimodal data with professional-grade annotations. While MME-Finance offers broader coverage of technical trading data and MMMU-Finance specializes in business metrics, FINMME delivers the depth and quality necessary for advanced financial analysis and model evaluation across multiple domains and asset classes. The combination of extensive coverage, detailed annotations, and rigorous quality control makes FINMME particularly well-suited for developing and evaluating sophisticated financial analysis models.

F Additional Cases

To provide a comprehensive view of model performance across different question types and difficulty levels, we present additional examples from our evaluation in Figures 8-17. These examples showcase various financial analysis scenarios. The cases demonstrate both successful and failed attempts by models, highlighting where current models excel and where they still face challenges. For instance, while models generally perform well on straightforward chart reading tasks, they often struggle with complex numerical calculations or when deep financial domain knowledge is required. These examples also illustrate common error patterns, such as hallucination in numerical responses and misinterpretation of complex financial relationships, providing valuable insights for future model improvements.

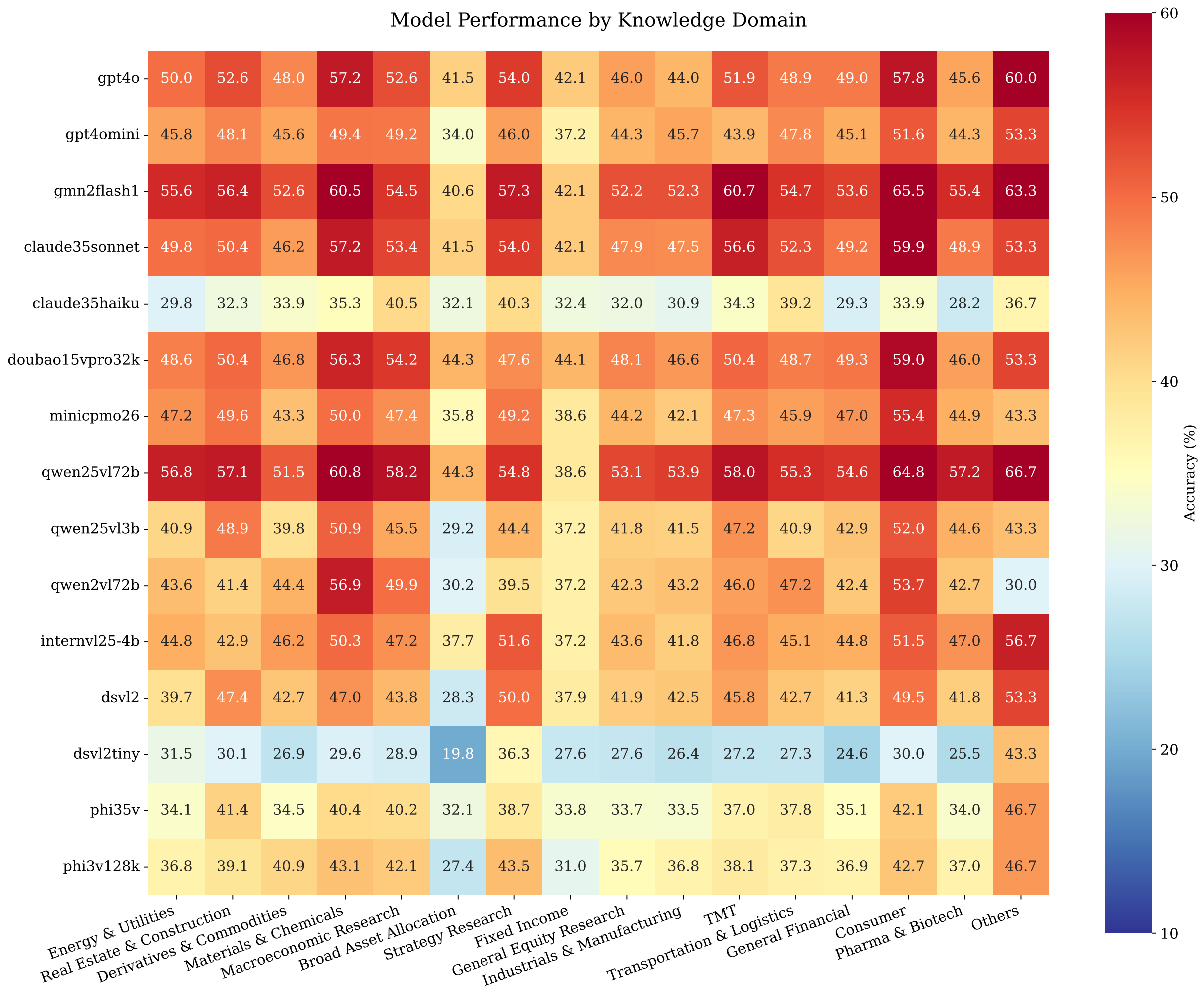


Figure 6: The heatmap of the knowledge domain distribution of the dataset.

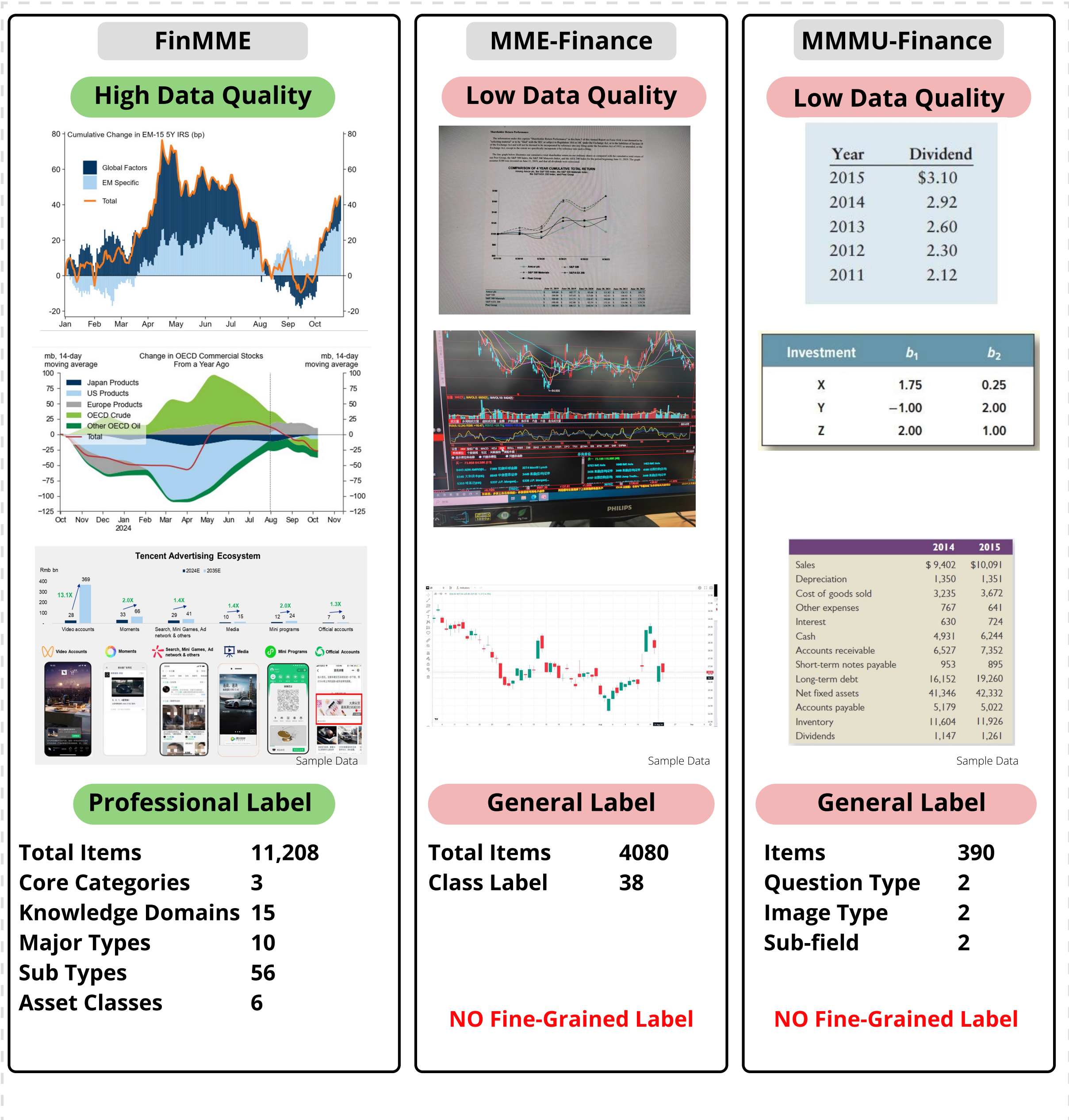
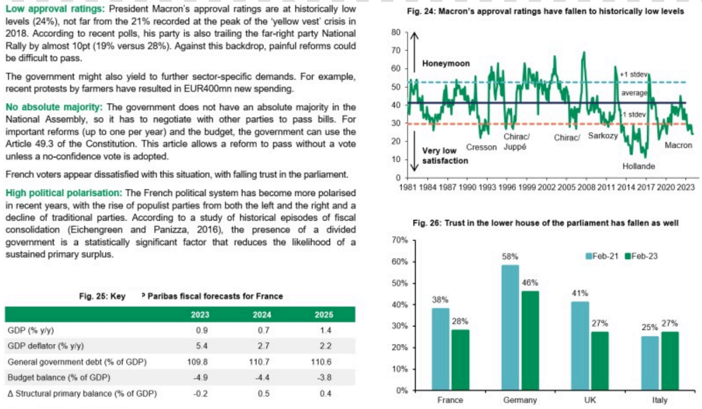


Figure 7: Data Comparison with related works.



Low approval ratings: President Macron's approval ratings are at historically low levels (24%), not far from the 21% recorded at the peak of the 'yellow vest' crisis in 2018. According to recent polls, his party is also trailing the far-right party National Rally by almost 10pt (19% versus 28%). Against this backdrop, painful reforms could be difficult to pass.

The government might also yield to further sector-specific demands. For example, recent protests by farmers have resulted in EUR400mn new spending.

No absolute majority: The government does not have an absolute majority in the National Assembly, so it has to negotiate with other parties to pass bills. For important reforms (up to one per year) and the budget, the government can use the Article 49.3 of the Constitution. This article allows a reform to pass without a vote unless a no-confidence vote is adopted.

French voters appear dissatisfied with this situation, with falling trust in the parliament.

High political polarisation: The French political system has become more polarised in recent years, with the rise of populist parties from both the left and the right and a decline of traditional parties. According to a study of historical episodes of fiscal consolidation (Eichengreen and Panizza, 2016), the presence of a divided government is a statistically significant factor that reduces the likelihood of a sustained primary surplus.

Fig. 25: Key ᴾ Paribas fiscal forecasts for France

	2023	2024	2025
GDP (% y/y)	0.9	0.7	1.4
GDP deflator (% y/y)	5.4	2.7	2.2
General government debt (% of GDP)	109.8	110.7	110.6
Budget balance (% of GDP)	-4.9	-4.4	-3.8
Δ Structural primary balance (% of GDP)	-0.2	0.5	0.4

Fig. 24: Macron's approval ratings have fallen to historically low levels

Fig. 26: Trust in the lower house of the parliament has fallen as well

Question Type

Single Answer Question

Caption

France Political Research

Question

What is the historical low point of Macron's approval ratings?

A: 21%
B: 24%
C: 28%
D: 19%

Ground Truth

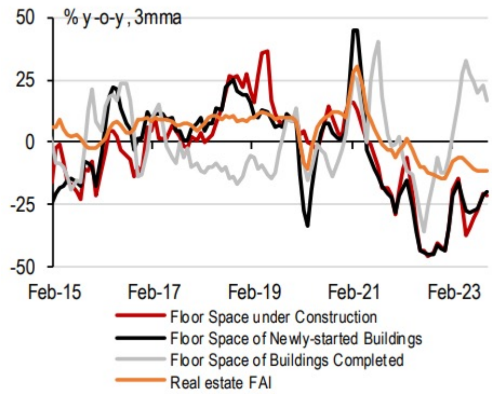
B

GPT4o Output

B

Macron's lowest approval rating was 21%, recorded during the 2018 'yellow vest' crisis. The current 24% rating is close but not the lowest. Graphical data in Fig. 24 confirms this. The correct answer is 21% (Option A).

Figure 8: Additional examples of questions in FinMME.



Question Type **Single Answer Question**

Caption **Trends in China's real estate sector**

Question **In February 2023, which line showed the highest negative percentage change in year-on-year floor space?**

A: Floor Space under Construction
B: Floor Space of Newly-started Buildings
C: Floor Space of Buildings Completed
D: Real Estate FAI

Ground Truth **A**

GPT4o Output **B**

In February 2023, the black line (Floor Space of Newly-started Buildings) shows the steepest decline, reaching the lowest negative percentage change among all categories. Thus, the correct answer is B: Floor Space of Newly-started Buildings.

Figure 9: Additional examples of questions in FinMME.

Question Type

Calculation Question

Caption

Salesforce's Historical Closing Price and Target Price

Question

What is the difference between the highest closing price and the lowest closing price observed in the chart?

Ground Truth

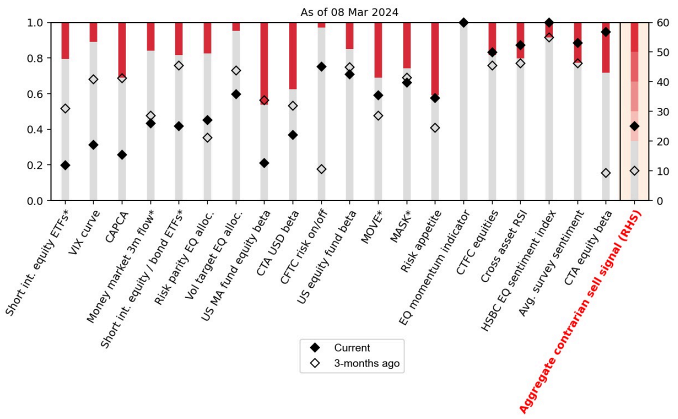
75

GPT4o Output

250

To determine the difference between the highest and lowest closing prices in the chart, we identify the peak and trough of the closing price line (teal color). The highest observed closing price is around USD 350, while the lowest is approximately USD 100. The difference is: 350−100=250350−100=250 Thus, the difference between the highest and lowest closing prices is USD 250.

Figure 10: Additional examples of questions in FinMME.



Question Type

Multiple Answers Question

Caption

Aggregate contrarian sell signal as of 08 Mar 2024

Question

Which inputs are associated with contrarian sell signals as of 08 Mar 2024?

A: Short int. equity ETFs
B: Risk appetite
C: Vol target EQ alloc.
D: Money market 3m flow

Ground Truth

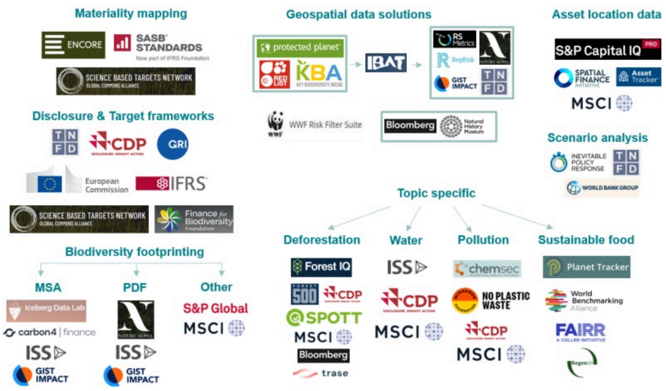
BC

GPT4o Output

ABC

The contrarian sell signals are highlighted in red bars, indicating elevated risk levels. Observing the chart, Short int. equity ETFs, Risk appetite, and Vol target EQ alloc. have strong red bars, suggesting they contribute to the contrarian sell signal. Money market 3m flow does not show a similar impact. Thus, the correct answers are: A: Short int. equity ETFs, B: Risk appetite, C: Vol target EQ alloc.

Figure 11: Additional examples of questions in FINMME.



Question Type | **Multiple Answers Question**

Caption | **The nature data ecosystem**

Question | **Which organizations contribute to the 'Sustainable food' topic?**

A: Planet Tracker
B: Bloomberg
C: FAIRR
D: Regen10

Ground Truth | **ACD**

GPT4o Output | **ACD**

The 'Sustainable food' topic in the image includes contributions from Planet Tracker, FAIRR, World Benchmarking Alliance, and Regen10. Comparing these with the answer choices, Planet Tracker (A), FAIRR (C), and Regen10 (D)are correct, while Bloomberg (B) is not listed under this category. Thus, the correct answers are A, C, and D.

Figure 12: Additional examples of questions in FINMME.

EBIT Margins (ex Fin/RE)	Pre-Earnings Season		Post-Earnings Season	
	FY2023	FY2024	FY2023	FY2024
Semi. & Semi. Eqpt.	22.2%	28.2%	22.4%	28.7%
Tech. Hardware & Eqpt.	3.1%	6.2%	3.0%	6.0%
Media & Ent.	20.3%	21.8%	20.2%	21.9%
Cons Discretionary Dist/Retail	8.9%	9.5%	8.0%	8.9%
Energy	8.0%	7.8%	8.1%	7.8%
Consumer Services	10.8%	12.7%	10.7%	12.2%
Software & Svcs.	17.1%	18.0%	17.0%	17.9%
Automobiles & Comp.	6.3%	6.3%	6.3%	6.4%
Household & Personal Products	13.6%	15.3%	13.5%	15.1%
Materials	6.3%	7.8%	5.7%	7.1%
Telecommunication Svcs.	15.7%	16.6%	15.6%	16.6%
Pharma. Biotec. & Life Sci.	14.8%	15.8%	14.5%	15.8%
Capital Goods	6.7%	7.9%	6.3%	7.6%
Food Beverage & Tobacco	9.5%	10.7%	9.4%	10.5%
Consumer Durables & Apparel	8.6%	9.3%	8.2%	8.8%
Utilities	10.7%	13.3%	11.0%	14.6%
Transportation	12.7%	11.9%	12.6%	13.2%
Cons Staples Dist/Retail	4.2%	4.5%	4.2%	4.5%
Health Care Eqpt. & Svcs.	5.0%	5.0%	5.0%	5.0%
Commercial & Professional Svcs.	19.9%	19.6%	19.9%	19.8%

Question Type

Single Answer Question

Caption

EBIT Margins for Various Sectors Pre and Post Earnings Season

Question

Which sector had the highest EBIT margin for FY2023 post-earnings season?

A: Semi. & Semi. Eqpt.
B: Media & Ent.
C: Software & Svcs.
D: Commercial & Professional Svcs.

Ground Truth

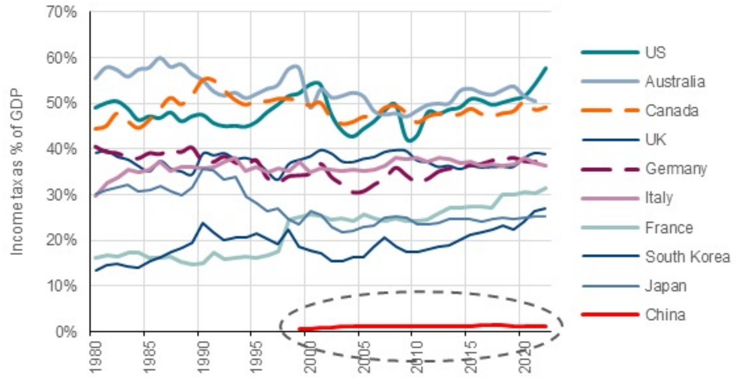
A

GPT4o Output

A

For FY2023 post-earnings season, the Semi. & Semi. Equipment sector had the highest EBIT margin at 22.4%, followed by Media & Entertainment at 20.2%, and Commercial & Professional Services at 19.9%. Since Semi. & Semi. Equipment recorded the highest margin, the correct answer is A: Semi. & Semi. Eqpt.

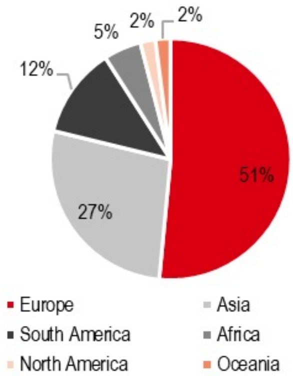
Figure 13: Additional examples of questions in FinMME.



Question Type	**Calculation Question**
Caption	**Personal income tax of GDP for the top 10 economies**
Question	**What is the approximate income tax as a percentage of GDP for the US in 2020?**
Ground Truth	**60**
GPT4o Output	**45**

The US is represented by the teal line in the chart. Observing its position in 2020, it is slightly above the 40% mark but below 50%. Comparing with nearby labeled countries, the US income tax as a percentage of GDP appears to be around 45% in 2020.

Figure 14: Additional examples of questions in FINMME.



Question Type

Single Answer Question

Caption

China's solar module exports by region in 2023.

Question

Which region has the highest share of China's solar module exports in 2023?

A: Europe
B: Asia
C: North America
D: South America

Ground Truth

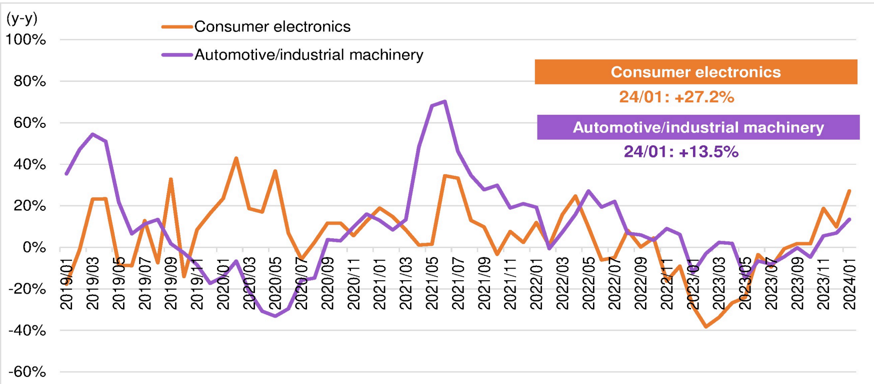
A

GPT4o Output

A

The pie chart shows that Europe has the highest share of China's solar module exports in 2023, accounting for 51%, followed by Asia (27%). Since Europe has the largest portion, the correct answer is: A: Europe.

Figure 15: Additional examples of questions in FinMME.



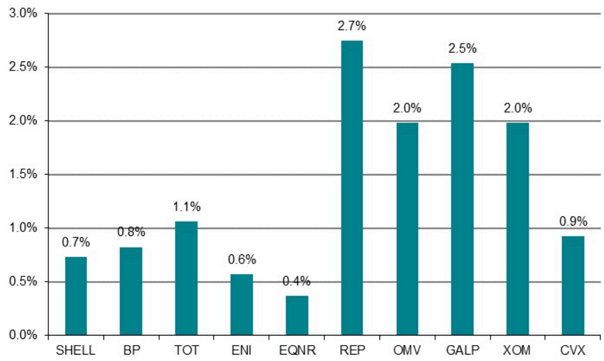
Question Type	**Calculation Question**
Caption	**Exports of MLCC and inductors to the consumerial**
Question	**Calculate the percentage difference between the peaks of both sectors in 2023/11**
Ground Truth	**13.7%**

GPT4o Output

100%

In November 2023, consumer electronics peaked at 20%, and automotive/industrial machinery at 10%. The percentage difference is:
20−1010×100=100%1020−10×100=100%
Thus, the difference is 100%.

Figure 16: Additional examples of questions in FinMME.



Question Type

Calculation Question

Caption

cash flow from operations for major oil companies

Question

What is the average relative exposure of the companies listed in the chart?

Ground Truth

1.37%

GPT4o Output

1.3%

The average relative exposure of the companies listed in the chart is calculated by summing their individual exposures and dividing by the total number of companies. The total exposure sums up to 13.7% across 10 companies, including Shell, BP, TOT, ENI, EQNR, REP, OMV, GALP, XOM, and CVX. Dividing this total by the number of companies gives an average relative exposure of 1.37%.

Figure 17: Additional examples of questions in FINMME.