FinMMR: Benchmarking Financial Numerical Reasoning More Multimodal, Comprehensive and Challenging

Zichen Tang Haihong E* Jiacheng Liu Zhongjun Yang Rongjin Li Zihua Rong Haoyang He Zhuodi Hao Xinyang Hu Kun Ji Ziyan Ma Mengyuan Ji Jun Zhang Chenghao Ma Qianhe Zheng Yang Liu Yiling Huang Xinyi Hu Qing Huang Zijian Xie Shiyao Peng

Beijing University of Posts and Telecommunications

bupt-reasoning-lab.github.io/FinMMR

BUPT-Reasoning-Lab/FinMMR BUPT-Reasoning-Lab/FinMMR

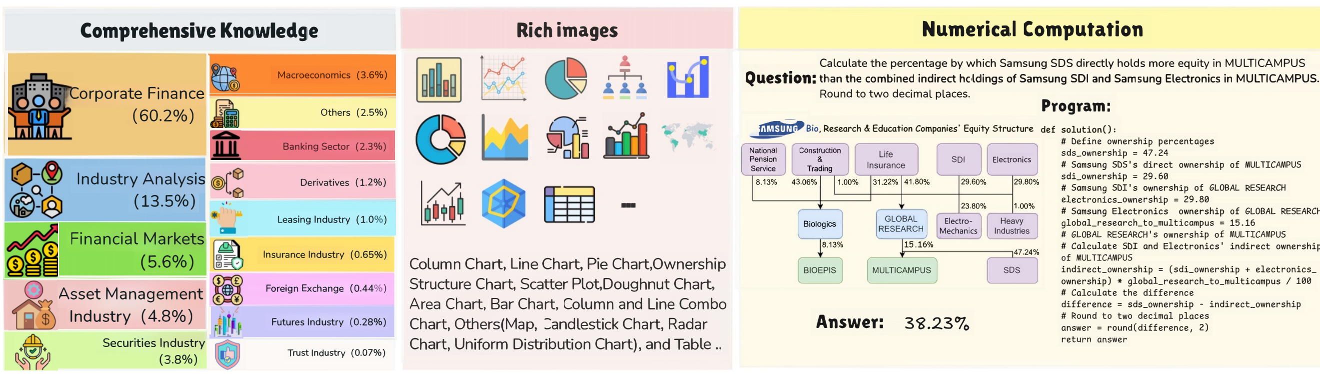


Figure 1. Overview of the FinMMR dataset. FinMMR presents three challenges: (1) **visual perception**: 8.7K financial images of 14 categories; (2) **knowledge reasoning**: 4.3K financial questions of 14 subdomains; (3) **numerical computation**: multi-step precise calculation.

Abstract

*We present **FinMMR**, a novel bilingual multimodal benchmark tailored to evaluate the reasoning capabilities of multimodal large language models (MLLMs) in financial numerical reasoning tasks. Compared to existing benchmarks, our work introduces three significant advancements. (1) **Multimodality**: We meticulously transform existing financial reasoning datasets, and construct novel questions from the latest Chinese financial research reports. The dataset comprises 4.3K questions and 8.7K images spanning 14 categories, including tables, bar charts, and ownership structure charts. (2) **Comprehensiveness**: FinMMR encompasses 14 financial subdomains, including corporate finance, banking, and industry analysis, significantly exceeding existing benchmarks in financial domain knowledge breadth. (3) **Challenge**: Models are required to perform multi-step precise numerical reasoning by integrating financial knowledge with the understanding of complex financial images and text. The best-performing MLLM achieves only 51.4% accuracy on* Hard *problems. We believe that FinMMR will drive advancements in enhancing the reasoning capabilities of MLLMs in real-world scenarios.*

1. Introduction

Recently, large reasoning models (LRMs) [21, 43, 44, 52, 54, 60], show powerful reasoning capabilties over multi-step reasoning tasks, with train-time scaling and test-time scaling [26, 41]. These reasoning models are proficient in code [7, 24], math [30, 36], and science [57]. Multimodal large language models (MLLMs) [2, 18, 42] also exhibit greater performance on multimodal reasoning [34, 63].

Despite significant advancements, there remains a notable gap in understanding the practical applicability of MLLMs in numerical reasoning within real-world scenarios, particularly in high-stakes fields such as finance and healthcare. As depicted in Fig. 1, financial analysts in their daily work are required to read visually enriched financial documents, extract key financial indicators from tables, images, and contextual texts, and perform precise multi-step

*Corresponding author.

Figure 2. Sampled FinMMR examples with two language (*i.e*. English and Chinese), rich images and different knowledge. The questions and images need expert-level visual perception, knowledge reasoning and numerical computation.

numerical calculations, supporting professional decision-making. Similarly, to achieve expert artificial general intelligence (AGI) [4, 16, 37, 38, 63], MLLMs are expected to comprehend complex domain-specific images akin to human experts, and apply domain knowledge to perform accurate numerical reasoning. This raises the question: **Can current MLLMs effectively integrate visual and textual information to perform deep, domain-specific complex reasoning, similar to the significant progress made by LRMs in text-based reasoning?**

Specifically, we choose the financial domain to evaluate the complex reasoning capabilities of MLLMs, where precision and transparent reasoning are paramount [27]. Existing numerical reasoning benchmarks for finance are limited in their text-based reasoning, coverage of specific financial knowledge, and complexity of reasoning [10, 12, 27, 65, 67]. FAMMA [61] is mainly modelled after textbook and CFA exam questions, MathVista [34] does not involve the application of financial knowledge, MMMU [63] and MMMU-Pro [64] are all multiple-choice questions, still showing a significant gap from the real-world scenario. The lack of high-quality, knowledge-intensive multimodal financial numerical reasoning datasets makes it challenging to objectively evaluate the actual reasoning capabilities of MLLMs and analyze their shortcomings.

Therefore, we propose **FinMMR**, a bilingual multimodal numerical reasoning benchmark designed to evaluate the reasoning capabilities of MLLMs in the finance domain. The dataset comprises 4.3K problems, covering 14 financial subdomains (*e.g*. corporate finance and industry analysis), with 8.7K images derived from 14 categories (*e.g*. tables and ownership structure charts). Each problem includes rich images, an unambiguous question, a Python-formatted solution, and a precise answer.

For multimodality, without representing financial tables as structured text, FinMMR represent all tables, charts, and diagrams as images. **For comprehensiveness**, FinMMR covers 14 financial subdomains and two languages (*i.e*. English and Chinese), demanding domain knowledge such as option pricing and portfolio management. **For challenge**, FinMMR focus on multi-step numerical reasoning, requiring models to provide exact numerical answers under strict evaluation criteria (emphasizing units, percentages, and decimal places). Furthermore, we **mix each Chinese questions with two distractor images** that are contextually adjacent to the ground images, approaching real-world multimodal reasoning scenarios.

We evaluate 12 current state-of-the-art MLLMs [17–19, 21, 42–44, 54], utilizing Chain-of-Thought (CoT) [58] and Program-of-Thought (PoT) [9]. The experimental results on FinMMR reveals three key findings:

- **MLLMs Face Significant Challenges in Domain-Specific Multimodal Numerical Reasoning:** All evaluated models underperform on FinMMR with CoT or PoT. On the *Hard* set, the best-performing model, Claude 3.7 Sonnet with 64K extended thinking, achieves only 51.4% accuracy, while OpenAI-o1 achieves merely 44.7%. Through error analysis, we identify that visual perception, knowledge reasoning, and numerical computation collectively pose challenges to MLLMs. Current MLLMs still struggle with complex multimodal reasoning tasks in specialized domains, compared to text-based reasoning.
- **Better Synergy Between Visual Perception and Complex Reasoning is Needed:** Distracting images lead to a greater than 10% drop in accuracy for Qwen2.5-VL-72B compared to ground images alone, indicating that irrelevant visual information severely impacts multimodal reasoning. By decoupling visual filtering and reasoning, Qwen2.5-VL-72B improved from 64.73% to 71.5%. Combining MLLMs with LRMs, by efficiently parsing visual information into structured text and leveraging the LRM's text-based reasoning capabilities, can also yield better performance. The combination of GPT-4o and DeepSeek-R1 achieves 86.72% accuracy on 1,160 tabular questions, outperforming Claude 3.7 Sonnet (83.53%).
- **Refined Structured Domain Knowledge Enhances MLLMs' Complex Reasoning:** Leading MLLMs lack sufficient experience in applying rich domain knowledge when solving complex reasoning tasks. By annotating structured financial functions and leveraging the model's ability to generate retrieval questions and make judgments, knowledge augmentation can significantly improve MLLMs' performance. MLLMs achieve improvements ranging from 2.76% to 4.31%, weaker models can approach state-of-the-art (SoTA) performance, while SoTA model can also achieve further gains.

These findings highlight the bottlenecks of MLLMs in complex multimodal reasoning tasks in expert domains closer to real-world scenarios. They emphasize the need for continuous improvements in three key areas: more intricate visual perception, more specialized knowledge reasoning, and more accurate numerical computation. Alternatively, leveraging tools or model combinations can help achieve a balance between performance and cost, enabling MLLMs to perform expert-level reasoning tasks like human experts.

2. Related Work

2.1. LRM and MLLM

By integrating train-time scaling and test-time scaling [26, 41], LRMs have demonstrated remarkable reasoning capabilities [60]. However, most LRMs are limited to handling text-based problems. The growing demand for solving real-world tasks has spurred the development of multimodal large reasoning models [2, 18, 53] and benchmarks designed to evaluate the perception and reasoning abilities of MLLMs [6, 14, 22, 25, 28, 31, 33, 62–64]. For instance, MME-CoT [25] evaluates models' space-time understanding, while EMMA [22] focuses STEM subjects. Following this trend, domain-specific benchmarks which require deep domain expertise have emerged, such as MathVista [34] for mathematics and GMAI-MMBench [8] for medicine. Yet, financial reasoning remains an unexplored area in the current landscape of MLLM benchmarks.

2.2. Financial Benchmarks

The financial domain presents a distinct and more formidable set of challenges for model evaluation, which arise from its inherent complexity, information density, and dependence on expertise. The majority of existing text-only financial numerical reasoning benchmarks [11, 12, 27, 65, 66, 68] are constrained by limitations such as suboptimal annotation quality, narrow domain knowledge coverage, and overly simplistic reasoning tasks. Although Fi-

Property	Value
# Total Questions	4,300
# Difficulties (Hard/Medium/Easy)	1,300/1,500/1,500
# Validation / Test	3,400/900
# Chinese / English	2,150/2,150
# Operators (Hard/Medium/Easy)	**5.34**/2.97/1.75
# Lines of Code (Hard/Medium/Easy)	**7.34**/5.06/4.14
# Parentheses (Hard/Medium/Easy)	**4.25**/3.11/0.88
# Difficulty (Hard/Medium/Easy)	**3.79**/2.96/1.93

Table 1. Key statistics of FinMMR (Avg values of three subsets).

nanceReasoning [51] offers complex tasks with rich knowledge and high-quality annotations, its text-only modality limits cross-modal reasoning evaluation.

Recent multimodal financial benchmarks have sought to bridge this gap but still possess limitations. FAMMA [61] being sourced from textbooks and examinations does not mirror the real-world tasks. FinMME [35] uses a multiple-choice format, which may overestimate model reasoning due to guesswork. MME-Finance[15] is constrained by coarse annotations and an isolated assessment of domain knowledge, limiting holistic evaluation of real-world financial reasoning.

3. The FinMMR Benchmark

3.1. Overview of FinMMR

We introduce FinMMR, a bilingual (English and Chinese) multimodal benchmark for evaluating the reasoning capabilities of MLLMs in financial numerical reasoning tasks. FinMMR consists of 4,300 questions covering 14 financial subdomains including corporate finance, industry analysis, financial markets, and asset management. The key statistics are summarized in Tab. 1, and the composition of subdomains and images is illustrated in Fig. 1. As illustrated in Fig. 2, FinMMR introduces three key challenges:

- **Rich Images Challenge Visual Perception:** FinMMR comprises 8.7K images from 14 categories, including bar charts, line charts, ownership structure chart, etc. MLLMs must identify relevant images among distractors and extract critical information from the correct images.
- **Comprehensive Domains Challenge Knowledge Reasoning:** MLLMs need to flexibly apply diverse domain-specific financial knowledge from 14 sub-domains to solve multi-step reasoning tasks.
- **Complex Formulas Challenge Numerical Computation:** All questions require precise numerical answers, eliminating the potential bias from lucky guesses that could occur in multiple-choice formats.

3.2. Data Curation Process

We first curated a subset of questions from public text-based financial reasoning benchmarks and systematically transformed them into multimodal problems using a unified standard. Subsequently, we constructed a novel multimodal Chinese Research Report Question Answering (CR-RQA) dataset from scratch, merging two data sources into FinMMR. Each question is accompanied by an executable Python solution, yields a numerical answer and delineates a clear reasoning pathway.

Update to Public Datasets We re-annotate 124 and 163 financial questions from MMMU [63] and MMMU-Pro [64], respectively. Following rigorous verification, these questions were directly incorporated into our dataset. Furthermore, we extracted 77, 288, and 795 questions from FinanceMath [65], CodeTAT-QA [27], and CodeFinQA [27], respectively. From DocMath-Eval [66], we further obtained 703 questions from its four subsets. For each question, we rendered any tabular data as images while removing the corresponding table information from the text, ensuring that MLLMs cannot rely solely on textual content.

Building a Novel Dataset from Scratch We collect 90 research reports, all of which are obtained through authorized access and cover diverse topics such as industry research, macroeconomic analysis, and strategy research. We use 360LayoutAnalysis [39] to extract informative images and discard those lacking explicit numerical data, reducing ambiguity. For each retained image, we prompt Qwen-VL-Max [47] to formulate questions requiring multiple reasoning steps or complex calculations. Each question is accompanied by an executable Python solution and a definitive numerical answer.

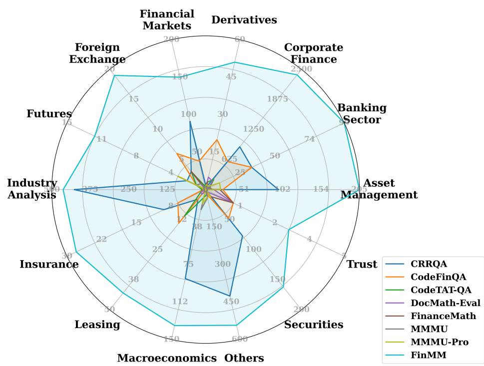
Futhermore, we introduce distractor images—sourced from the same reports adjacent to ground images—to challenge MLLMs to extract relevant numerical information from structured, densely packed visuals.

Data Quality Assurance This process ensured that every question was clearly written, featured a detailed reasoning solution, and included an accurate final answer. The annotators comprised 16 master's students in finance and two experts holding CFA certifications. With the aid of LLMs, this meticulous verification process spanned three months, culminating in a dataset that meets high standards of clarity and correctness.

Dataset Splitting and Release To classify the problems by difficulty, we employ a heuristic algorithm that takes into account the number of operators (o), code lines (l) and parentheses pairs (p) in the Python solution. Specifically, the difficulty rating rc of a problem is defined as:

$$rc = \ln(\max(o, 1)) + \ln(\max(l + p, 1)) \tag{1}$$

FinMMR is classified as *Easy* (1,300), *Medium* (1,500) and *Hard* (1,500) based on this formula, with each level ran-



Dataset	**Size (Fin)**	**Domain Coverage**	**Modalities**	**Question Type**
MMMU [63]	11,550 (1,603)	10	T+I	MC
MMMU-Pro [64]	1,730 (286)	10	T+I, P.I.	MC
FinanceMath [65]	1,200 (1,200)	10	T	NUM
CodeTAT-QA [27]	3,144 (3,144)	6	T	NUM
CodeFinQA [27]	5,463 (5,463)	13	T	NUM
DocMath-Eval [66]	4,000 (4,000)	12	T	NUM
FAMMA [61]	1,758 (1,758)	8	T+I	MC, NUM
CRRQA (ours)	2,150 (2,150)	13	T+I, P.I.	NUM
FinMMR (ours)	4,300 (4,300)	14	T+I, P.I.	NUM

Figure 3. The comparison between finance-related datasets. These datasets vary in size, domain coverage, modalities, and question type, with some focusing on text-only data while others include images. Each axis has scale labels with varying ranges to measure the number of questions from each dataset across different subdomains. In the Modalities, T means text input, I means Images input, P.I. means pure images input. In the Question Type, NUM means numerical answer, MC means multi-choice answer.

domly split into *test* and *validation* sets. All questions are publicly available, while only the 300 validation answers per level are released, while test answers remain private to prevent data leakage [13, 48, 50]. We maintain an online evaluation platform that enables researchers to assess their models.

3.3. Comparisons with Existing Benchmarks

As illustrated in Fig. 3, previous work has studied multi-discipline multimodal reasoning [63, 64], general mathmatical reasoning [34] or text-based financial QA [27, 65, 66]. FinMMR focus on multimodal financial numerical reasoning, curating 4,300 questions requiring a deep understanding of domain-specific images (*e.g.* earnings reports, stock charts). To mimic this scenario, we deliberately incorporate 3,938 distractors into 2,150 questions to rigorously evaluate MLLMs' visual perception capability. Compared to existing financial benchmarks, they suffer from narrow domain coverage[11, 67]. FinMMR encompasses 14 financial subdomains and 14 image categories, comprehensively evaluating MLLMs' domain-speific reasoning capabilities.

4. Evaluation System

To facilitate the evaluation of complex reasoning on FinMMR, we established a dedicated evaluation system. All MLLMs evaluated were accessed through official APIs.

4.1. Multimodal Large Language Models

We systematically evaluate the multimodal reasoning capabilities of twelve recent MLLMs under the zero-shot setting. The evaluated MLLMs are: OpenAI-o1 [42], GPT-4o [40], Claude 3.7 Sonnet (including thinking mode) [2], Gemini 2.0 Flash Thinking [18], Gemini 2.0 Pro [19], Gemini 2.0 Flash [17], InternVL2.5-78B [45], Grok-2 [59], Pixtral Large [1], Qwen2.5-VL-72B [3], Qwen-QVQ-72B-Preview [53], and Qwen-Omni-Turbo [55].

4.2. Prompting Methods

Following Zhao et al. [65], our evaluation adopts Chain-of-Thought (CoT) [58] and Program-of-Thought (PoT) [9] two prompting methods. Due to budget constraints, we report OpenAI-o1 performance with PoT prompts on *Hard* set only. All other models are evaluated on the entire benchmark with CoT prompts and PoT prompts. Detailed prompts can be found in the Appendix.

4.3. Answer Extraction and Evaluation

Following Zhao et al. [65], we extract answers based on the prompting technique. For CoT outputs, we employ GPT-4o-mini to extract numerical answers. For PoT, we run the generated program for numerical results. Finally, we conduct a strict accuracy evaluation, comparing numerical results with ground truth and deeming the results accurate within a stringent error tolerance of 0.2%.

5. Experiments

We answer the following research questions (RQs): **RQ1**: Are MLLMs multimodal reasoners with extended thinking and PoT prompts? **RQ2**: What are the primary challenges facing MLLMs? **RQ3**: How can the visual perception difficulties of MLLMs be mitigated? **RQ4**: How can the knowledge reasoning capabilities of MLLMs be enhanced? **RQ5**: How can the numerical computation abilities of MLLMs be compensated for?

5.1. Main Results (RQ1)

The performance of the MLLMs evaluated using two initiation methods on the FinMMR is shown in Tab. 2.

Model	Size	Extended thinking	Hard			Medium		Easy		Avg.		Token (M)	
			IO	CoT	PoT	CoT	PoT	CoT	PoT	CoT	PoT	CoT	PoT
Proprietary MLLMs													
Claude 3.7 Sonnet		✔ (64K)	**53.00**	**51.00**	**51.40**	62.50	62.17	**78.50**	**78.50**	**64.00**	**64.02**	**8.51**	**11.25**
Claude 3.7 Sonnet		✘	49.80	50.80	48.50	62.25	58.83	77.00	76.92	63.35	61.42	0.99	0.89
GPT-4o		✘	–	45.40	47.80	63.33	59.92	78.00	76.00	62.24	61.24	0.85	0.28
Gemini 2.0 Flash Thinking		✔	–	46.00	46.00	60.75	56.58	77.17	74.17	61.31	58.92	1.30	0.48
Gemini 2.0 Pro		✘	–	46.50	47.30	60.58	57.92	75.50	75.67	60.86	60.30	0.85	0.45
Gemini 2.0 Flash		✘	–	44.40	45.90	57.83	53.42	74.92	73.75	59.05	57.69	0.79	0.43
OpenAI-o1		✔	48.00	–	44.70	–	–	–	–	–	–	–	0.21
Qwen-Omni-Turbo		✘	–	17.50	27.30	35.83	48.00	57.50	61.67	36.94	45.66	0.90	0.42
Open Source MLLMs													
Llama 4 Maverick	17B	✘	–	48.70	47.80	63.30	59.20	77.80	77.80	63.27	61.60	–	–
Qwen-QVQ-72B-Preview	72B	✔	43.30	40.30	6.20	55.67	9.67	75.42	12.42	57.13	9.43	5.43	5.70
Qwen2.5-VL-72B	72B	✘	–	43.30	46.20	**63.42**	**64.17**	77.42	75.83	61.38	62.07	1.05	0.44
Gemma 3	27B	✘	–	23.40	22.30	45.20	36.40	69.10	61.60	45.90	40.10	–	–
InternVL2.5-78B	78B	✘	–	37.40	44.00	60.50	61.17	70.92	70.58	56.27	58.58	–	–
Grok-2		✘	–	27.80	25.50	41.50	35.83	73.08	72.83	47.46	44.72	1.13	0.60
Pixtral Large	124B	✘	–	19.70	25.50	41.50	35.83	73.08	72.83	47.46	44.72	1.15	0.75
Mistral 3.1	24B	✘	–	19.70	15.20	38.40	29.80	67.70	49.40	41.93	31.47	–	–

Table 2. Results of different models using IO, CoT and PoT prompting methods on the *test* set of FinMMR. We use average Accuracy using CoT prompting as the ranking indicator of model performance. The results underscore the superior performance of reasoning-enhanced MLLM (*i.e.* Claude 3.7 Sonnet with 64K extended thinking) with PoT in complex multimodal numerical reasoning task.

Challenges of MLLMs in Domain-Specific Complex Numerical Reasoning As the difficulty increases, the average accuracy shows a continuous and significant decline. In the CoT setting, the average accuracy rates on the *Easy*, *Medium*, and *Hard* sets are 73.79%, 53.33%, and 39.18%, respectively. The currently best-performing model (*i.e.* Claude 3.7 Sonnet with 64K extended thinking) consistently demonstrates superior performance across all difficulty sets using the CoT prompting method. However, **However, its accuracy on the *Hard* set remains below the 60% passing threshold under both prompting methods.** On the overall *test* set, Claude 3.7 Sonnet achieves only 64% accuracy. These results highlight the challenging nature and rigorous standards of FinMMR, effectively assessing the limits of MLLMs' reasoning capabilities and the disparities among models compared to previous multimodal reasoning datasets.

Does extended thinking help? Reasoning models show consistent improvements, compared with non-reasoning MLLMs. Claude 3.7 Sonnet with 64K extended thinking achieves a 2.9% improvement compared to the model without extended thinking (*i.e.* 51.40% vs. 50.80%) on the *Hard* set, using PoT prompts. This enhancement comes at the cost of using nearly 15 times more tokens for intricate thinking (*i.e.* 448K vs. 30K). This trend also persists in the Gemini 2.0 Flash series.

We observe that Qwen-QVQ-72B-Preview lose basic code generation capabilities due to the reinforcement learning of text-based long thinking, which is likely attributed to biases in training strategies and training data. On the *Hard* set, this model achieves a code execution success rate of only 10.9%, resulting in an accuracy of merely 6.2% in the PoT setting, significantly lower than the 40.3% accuracy achieved with CoT. This finding highlights the importance of maintaining foundational capabilities, such as programming, while enhancing the reasoning abilities of LLMs, to avoid rendering them ineffective in performing other basic tasks.

Does PoT help? Experimental results strongly validate the superiority of PoT prompting over CoT in numerical reasoning tasks, especially on the *Hard* set. After removing the highest and lowest outliers, PoT achieves a mean accuracy of 40.75% versus 40.16% for CoT, representing an improvement of 0.59%. Furthermore, PoT encourages MLLMs to leverage structured code generation to reduce token consumption during reasoning. Under similar or lower token usage, PoT achieves similar or greater accuracy. For instance, GPT-4o achieves a 2.4% improvement in accuracy over CoT while consuming significantly fewer tokens under the PoT setting. Similarly, the Qwen2.5-VL-72B demonstrates the most pronounced efficiency gains: PoT improves accuracy to 64.17% from 63.42% while reducing token consumption by 59% (153K vs. 373K) on the *Medium* set. **When addressing complex numerical reasoning problems, PoT avoids precise numerical calculations by utilizing external tools (*i.e.* Python interpreter) and reduces the need for repetitive text-based reasoning, which is beneficial for most MLLMs.**

However, we also observe that for certain reasoning-enhanced models (*e.g.* Claude 3.7 Sonnet with 64K extended thinking and OpenAI-o1), due to the enforced requirement for long thought, they still engage in extensive

Dataset	Test			Validation		
	Ground Images (%)	Distractor Images (%)	Degradation	Ground Images (%)	Distractor Images (%)	Degradation
Hard	57.18	47.23	↓ 9.95	56.74	45.58	↓ 11.16
Medium	61.36	73.01	↓ 11.65	64.73	77.08	↓ 12.35
Easy	53.64	61.59	↓ 7.95	51.52	60.61	↓ 9.09
The improvement achieved by the filtering-reasoning pipeline on the *medium validation* set: 64.73 → 71.56 **↑ 7.58**						

Table 3. Degradation of Qwen2.5-VL-72B on all subsets due to distractor images and improvement achieved by the filtering-reasoning pipeline on the *medium validation* set under PoT setting.

text-based reasoning before generating code even on the PoT setting, resulting in exceptionally high token consumption (448K and 212K), which is more than 10 times that of other MLLMs. To further investigate this, we added an IO baseline without any external prompts for reasoning models on the hard test set. The IO group achieved the highest accuracy, which we attribute to the comprehensive built-in system prompts embedded in the tested closed-source models. This highlights the need for future research to balance reasoning performance with the control of inefficient and redundant token generation, aiming to achieve a good trade-off between performance and cost, as well as to investigate whether PoT prompting can yield significant performance gains on open-source reasoning models.

5.2. Error Analysis (RQ2)

To better analyze the capabilities and limitations of MLLMs on FinMMR, we conduct a detailed error analysis for the Claude 3.7 Sonnet with 64K extended thinking in the PoT setting. Error analysis is based on 100 sampled failure cases, which we categorize into three main error types, some of which involve compound errors. More details of error cases are provided in the Appendix.

- **Visual Perception Error** (30/100): The model incorrectly perceives, identifies, or interprets visual information from images, or mistakenly recognizes incorrect data, subsequently causing errors in calculations, broken reasoning chains, or incorrect conclusions.
- **Knowledge Reasoning Error** (38/100): Due to insufficient domain-specific knowledge, the model exhibits logical confusion or conceptual misunderstandings during reasoning, leading to incorrect answers.
- **Numerical Computation Error** (32/100): In problems involving mathematical operations and numerical reasoning, the model produces significant deviations from the correct answers due to errors in the calculation steps, precision control, or numerical hallucination.

5.3. Visual Filtering for Reasoning (RQ3)

As shown in Tab. 3, when processing multi-image inputs containing distractor images, Qwen2.5-VL-72B demonstrates significantly lower accuracy across all difficulty levels compared to ground images scenarios. This finding aligns with conclusions from previous work [32, 49], indicating that irrelevant visual information substantially interferes with MLLMs' reasoning capabilities. In particular, the *Medium* set exhibits the most pronounced performance drop (77.78% ground images vs. 64.73% distractor images), attributed to two key characteristics: (1) moderate complexity making visual perception quality the performance bottleneck; (2) semantic relevance between distractors and questions increasing visual filtering difficulty. To address distractor image interference, we propose a two-stage multimodal reasoning pipeline:

- **Visual Filtering**: We first instruct MLLM to analyze the set of images $\mathcal{I}$ and the question q, assessing the relevance of the image (relevant / irrelevant). Irrelevant images are excluded from subsequent reasoning.
- **Enhanced Reasoning**: Then, the filtered subset $\mathcal{I}'$ and the question q are input into the MLLM for the final reasoning. The system automatically reverts to the original inputs $\mathcal{I}$ if all images are mistakenly filtered.

Does the Pipeline help? As illustrated in Tab. 3, we evaluate Qwen2.5-VL-72B on the 207 English questions with distractor images of the *Medium* validation subset. Our method achieves an overall accuracy of 71.56%, representing a 6.83 percentage point improvement over direct reasoning. This result is only 6.22% away from the ideal accuracy of the ground images scenarios (77.78%). Detailed analysis reveals 73.4% (152/207) ground image recognition accuracy during filtering. When correctly identified, the accuracy of these problems increases to 81.58% (124/152). **This finding underscores the necessity to enhance the ability of MLLMs to filter out irrelevant image information, thereby strengthening their robustness in reasoning within more complex real-world scenarios.**

5.4. Knowledge Augmentation (RQ4)

To enhance the understanding and application capabilities of financial knowledge of MLLMs, we explore a method of enhancing refined knowledge to improve the performance of MLLM in domain-specific complex reasoning tasks.

- **Function Library Construction**: We annotate a comprehensive financial function library containing 3,133

Setting	PoT	RAG with PoT
Gemini 2.0 Flash Thinking	78.71	83.02 (+4.31)
GPT-4o	80.60	83.62 (+3.02)
Claude 3.7 Sonnet (*wo.*)	81.21	85.43 (+4.22)
Claude 3.7 Sonnet (64K)	83.53	86.29 (+2.76)

Table 4. Improvements of different MLLMs with knowledge augmentation on the 1,160 problems of FinMMR under PoT setting.

Python functions from financial encyclopedia. Each function includes precise functional descriptions, parameter explanations, and step-by-step implementation code.

- **MLLM-Instructed Knowledge Retrieval**: In financial problems with hybrid contexts, using short questions or full contexts for retrieval often fails to retrieve directly relevant knowledge [5, 46]. We observed that powerful MLLMs (*e.g*. GPT-4o) can effectively summarize rich semantic information from contexts. Therefore, we first prompt the MLLM to generate precise retrieval queries based on the context [29, 56]. Then we use Contriever [23] to retrieve relevant financial Python functions, based on the semantic similarity between the refined queries and function descriptions.
- **MLLM as Retrieval Judge**: Recent studies have shown that models are capable of judging the relevance of candidates retrieved for the question [20]. In this setting, we first retrieved the Top-30 financial functions and then prompted the MLLM to select the Top-3 functions most useful to answer the question, if any.

Does Knowledge Augmentation help? As shown in Tab. 4, all evaluated MLLMs enhanced with financial function knowledge achieved significant performance improvements, ranging from 2.76% to 4.31%. Leveraging the improved retrieval efficiency enabled by *MLLM-Instructed Knowledge Retrieval* and *MLLM as Retrieval Judge*, the knowledge augmentation approach achieves greater performance, boosting the accuracy of Claude 3.7 Sonnet with 64K thinking to 86.29%. Notably, Gemini 2.0 Flash Thinking, which has relatively weaker reasoning capabilities, also improved from 78.71% to 83.02%, approaching the performance of Claude 3.7 Sonnet (83.53%) without knowledge augmentation. **The results further illustrate that refined domain-specific reasoning knowledge can significantly enhance the performance of MLLMs in complex reasoning tasks within expert domains.**

5.5. Visual Parser with Reasoner (RQ5)

In complex multimodal numerical reasoning tasks, single models often struggle to simultaneously achieve visual perception, knowledge reasoning, and numerical computation. To investigate the potential of model collaboration, we instruct the MLLM to act as the **Visual Parser**, responsible for carefully converting images into structured textual data. Then, an LRM acts as the **Reasoner**, performing multi-step numerical reasoning based on the textual context.

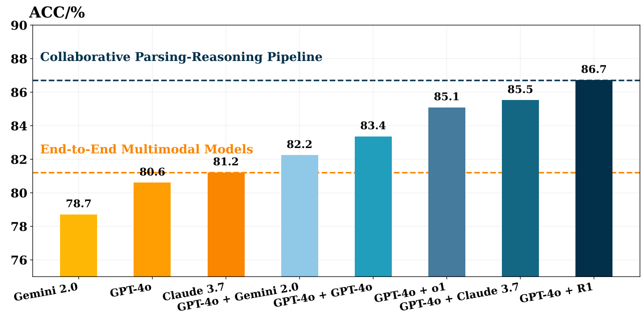


Figure 4. Results of model combinations and individual models.

Specifically, we filter 1,160 table question-answering problems from FinMMR and employ **GPT-4o** as Visual Parser, instructing it to separate table headers or cells with vertical bars (|) and rows with newlines. For Reasoners, in addition to **GPT-4o**, we evaluate several LRMs (*i.e*. **Claude 3.7 Sonnet**, **Gemini 2.0 Flash Thinking**, **DeepSeek-R1**, and **OpenAI-o1**).

Does Model Collaboration help? As shown in Table Fig. 4, the structured data from GPT-4o's visual parsing significantly enhances downstream reasoning. The individual model (*i.e*. GPT-4o with PoT) achieves an accuracy of 80.6%, while the combination of models improves the accuracy to 86.72% (*i.e*. DeepSeek-R1 as Reasoner with PoT). Performance variance emerges across reasoning models using identical visual inputs. Claude 3.7 Sonnet reaches 85.52%, outperforming Gemini 2.0 Flash Thinking (82.24%), confirming the decisive impact of text-based reasoning capabilities. **This evidences that model collaboration effectively compensates for individual model limitations through complementary strengths.**

6. Conclusion

We introduce FinMMR, a multimodal, comprehensive, and challenging benchmark for evaluating the financial numerical reasoning capabilities of MLLMs. FinMMR challenges MLLMs' intricate visual perception, specialized knowledge reasoning, and accurate numerical computation through its rich images, comprehensive domains, and complex formulas embedded in each multimodal financial question. The evaluation results reveal that 12 state-of-the-art MLLMs still struggle with complex multimodal reasoning tasks in specialized domains. FinMMR highlights the bottlenecks of MLLMs and the need for continuous improvements, including reasoning-enhanced training, tool use, refined structured knowledge augmentation and model combinations, allowing models to perform expert-level reasoning tasks closer to the real-world scenarios like human experts.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant Nos. 62176026, 62473271), the Beijing Natural Science Foundation (Grant No. QY24214), and the BUPT Innovation and Entrepreneurship Support Program (Grant Nos. 2025-YC-A033, 2025-YC-A042). This work is also supported by the Engineering Research Center of Information Networks, Ministry of Education, China. We would also like to thank the anonymous reviewers and area chairs for constructive discussions and feedback.

References

[1] Mistral AI. Pixtral large, 2024. 5

[2] Anthropic. Claude 3.7 sonnet and claude code, 2025. 1, 3, 5

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 5

[4] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023. 2

[5] Liang Chen, Yang Deng, Yatao Bian, Zeyu Qin, Bingzhe Wu, Tat-Seng Chua, and Kam-Fai Wong. Beyond factuality: A comprehensive evaluation of large language models as knowledge generators. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. 8

[6] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models?, 2024. 3

[7] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code, 2021. 1

[8] Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, Benyou Wang, Shaoting Zhang, Bin Fu, Jianfei Cai, Bohan Zhuang, Eric J Seibel, Junjun He, and Yu Qiao. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai, 2024. 3

[9] Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*, 2023. 3, 5

[10] Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 2

[11] Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan R Routledge, et al. Finqa: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, 2021. 3, 5

[12] Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 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, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. 2, 3

[13] Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. Investigating data contamination in modern benchmarks for large language models, 2024. 5

[14] 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. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024. 3

[15] Ziliang Gan, Yu Lu, Dong Zhang, Haohan Li, Che Liu, Jian Liu, Ji Liu, Haipang Wu, Chaoyou Fu, Zenglin Xu, Rongjunchen Zhang, and Yong Dai. Mme-finance: A multimodal finance benchmark for expert-level understanding and reasoning, 2024. 4

[16] Yingqiang Ge, Wenyue Hua, Kai Mei, Juntao Tan, Shuyuan Xu, Zelong Li, Yongfeng Zhang, et al. Openagi: When llm meets domain experts. *Advances in Neural Information Processing Systems*, 36:5539–5568, 2023. 2

[17] Gemini. Gemini 2.0: Flash, flash-lite and pro, 2025. 3, 5

[18] Gemini. Gemini 2.0 flash thinking, 2025. 1, 3, 5

[19] Gemini. Gemini 2.0, 2025. 3, 5

[20] Jian Guan, Wei Wu, zujie wen, Peng Xu, Hongning Wang, and Minlie Huang. AMOR: A recipe for building adaptable modular knowledge agents through process feedback. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 8

[21] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. 1, 3

[22] Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark, 2025. 3

[23] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*, 2022. 8

[24] Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama,

Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code, 2024. 1

[25] Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, Bo Zhang, Chaoyou Fu, Peng Gao, and Hongsheng Li. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency, 2025. 3

[26] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. 1, 3

[27] Michael Krumdick, Rik Koncel-Kedziorski, Viet Dac Lai, Varshini Reddy, Charles Lovering, and Chris Tanner. BizBench: A quantitative reasoning benchmark for business and finance. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8309–8332, Bangkok, Thailand, 2024. Association for Computational Linguistics. 2, 3, 4, 5

[28] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension, 2023. 3

[29] Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. Search-o1: Agentic search-enhanced large reasoning models, 2025. 8

[30] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023. 1

[31] Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning, 2024. 3

[32] Haowei Liu, Xi Zhang, Haiyang Xu, Yaya Shi, Chaoya Jiang, Ming Yan, Ji Zhang, Fei Huang, Chunfeng Yuan, Bing Li, and Weiming Hu. MIBench: Evaluating multimodal large language models over multiple images. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22417–22428, Miami, Florida, USA, 2024. Association for Computational Linguistics. 7

[33] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024. 3

[34] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2, 3, 5

[35] Junyu Luo, Zhizhuo Kou, Liming Yang, Xiao Luo, Jinsheng Huang, Zhiping Xiao, Jingshu Peng, Chengzhong Liu, Jiaming Ji, Xuanzhe Liu, Sirui Han, Ming Zhang, and Yike Guo. Finmme: Benchmark dataset for financial multi-modal reasoning evaluation, 2025. 4

[36] Yujun Mao, Yoon Kim, and Yilun Zhou. CHAMP: A competition-level dataset for fine-grained analyses of LLMs' mathematical reasoning capabilities. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13256–13274, Bangkok, Thailand, 2024. Association for Computational Linguistics. 1

[37] Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: A benchmark for general ai assistants. In *12th International Conference on Learning Representations, ICLR 2024*, 2024. 2

[38] Meredith Ringel Morris, Jascha Sohl-Dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. Position: levels of agi for operationalizing progress on the path to agi. In *Proceedings of the 41st International Conference on Machine Learning*, pages 36308–36321, 2024. 2

[39] 360 AILAB NLP. 360layoutanalysis: A layout analysis toolkit for document understanding. `https://github.com/360AILAB-NLP/360LayoutAnalysis`, 2025. 4

[40] OpenAI. Hello gpt-4o, 2024. 5

[41] OpenAI. Learning to reason with llms, 2024. 1, 3

[42] OpenAI. Openai o1 system card, 2024. 1, 3, 5

[43] OpenAI. Openai o1-mini, 2024. 1

[44] OpenAI. Openai o3-mini system card, 2025. 1, 3

[45] OpenGVLab. Internvl2.5: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2024. 5

[46] Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. Check your facts and try again: Improving large language models with external knowledge and automated feedback, 2023. 8

[47] QwenLM. Qwen-vl: A vision-language model by qwenlm. `https://github.com/QwenLM/Qwen-VL`, 2024. 4

[48] Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore, 2023. Association for Computational Linguistics. 5

[49] Aditya Sharma, Michael Saxon, and William Yang Wang. Losing visual needles in image haystacks: Vision language models are easily distracted in short and long contexts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5429–5451, Miami, Florida, USA, 2024. Association for Computational Linguistics. 7

[50] Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*, 2024. 5

[51] Zichen Tang, Haihong E, Ziyan Ma, Haoyang He, Jiacheng Liu, Zhongjun Yang, Zihua Rong, Rongjin Li, Kun Ji, Qing Huang, Xinyang Hu, Yang Liu, and Qianhe Zheng. Financereasoning: Benchmarking financial numerical reasoning more credible, comprehensive and challenging, 2025. 4

[52] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms, 2025. 1

[53] Qwen Team. Qvq: To see the world with wisdom, 2024. 3, 5

[54] Qwen Team. Qwq: Reflect deeply on the boundaries of the unknown, 2024. 1, 3

[55] Qwen Team. qwen-omni, 2025. 5

[56] Prakhar Verma, Sukruta Prakash Midigeshi, Gaurav Sinha, Arno Solin, Nagarajan Natarajan, and Amit Sharma. Plan*rag: Efficient test-time planning for retrieval augmented generation, 2025. 8

[57] Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. In *Forty-first International Conference on Machine Learning*, 2024. 1

[58] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022. 3, 5

[59] xAI. Models, 2024. 5

[60] Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, et al. Towards large reasoning models: A survey of reinforced reasoning with large language models, 2025. 1, 3

[61] Siqiao Xue, Tingting Chen, Fan Zhou, Qingyang Dai, Zhixuan Chu, and Hongyuan Mei. Famma: A benchmark for financial domain multilingual multimodal question answering, 2024. 2, 4, 5

[62] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities, 2024. 3

[63] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 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, 2024. 1, 2, 3, 4, 5

[64] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark, 2024. 3, 4, 5

[65] Yilun Zhao, Hongjun Liu, Yitao Long, Rui Zhang, Chen Zhao, and Arman Cohan. Financemath: Knowledge-intensive math reasoning in finance domains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12841–12858, Bangkok, Thailand, 2024. Association for Computational Linguistics. 2, 3, 4, 5

[66] Yilun Zhao, Yitao Long, Hongjun Liu, Ryo Kamoi, Linyong Nan, Lyuhao Chen, Yixin Liu, Xiangru Tang, Rui Zhang, and Arman Cohan. DocMath-eval: Evaluating math reasoning capabilities of LLMs in understanding long and specialized documents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16103–16120, Bangkok, Thailand, 2024. Association for Computational Linguistics. 3, 4, 5

[67] Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3277–3287, Online, 2021. Association for Computational Linguistics. 2, 5

[68] Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance, 2021. 3