UNDERSTANDING FINANCIAL REASONING IN AI: A MULTIMODAL BENCHMARK AND ERROR LEARNING APPROACH

Shuangyan Deng, Haizhou Peng, Jiachen Xu, Chouhou Liu,
Ciprian Doru Giurcăneanu, Jiamou Liu
University of Auckland
Auckland, New Zealand
{sden118}@aucklanduni.ac.nz

ABSTRACT

Effective financial reasoning demands not only textual understanding but also the ability to interpret complex visual data such as charts, tables, and trend graphs. This paper introduces a new benchmark designed to evaluate how well AI models—especially large language and multimodal models—reason in finance-specific contexts. Covering 3,200 expert-level question-answer pairs across 15 core financial topics, the benchmark integrates both textual and visual modalities to reflect authentic analytical challenges in finance. To address limitations in current reasoning approaches, we propose an error-aware learning framework that leverages historical model mistakes and feedback to guide inference, without requiring fine-tuning. Our experiments across state-of-the-art models show that multimodal inputs significantly enhance performance and that incorporating error feedback leads to consistent and measurable improvements. The results highlight persistent challenges in visual understanding and mathematical logic, while also demonstrating the promise of self-reflective reasoning in financial AI systems. Our code and data can be found at https://anonymous/FinMR/Code&Data. The leaderboard can be found at https://anonymous/FinMR/Leaderboard.

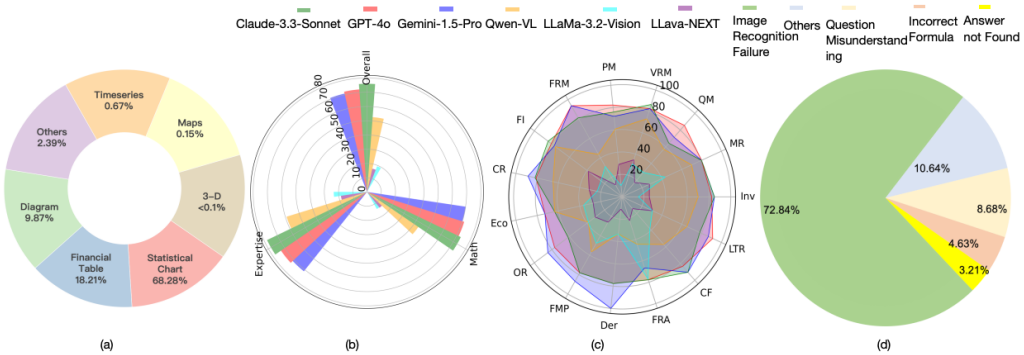

Figure 1: **FinMR** provides diverse visual data, as shown in panel (a). Evaluation of financial reasoning abilities of LLMs and MLLMs covers mathematical and expertise-based tasks (see panel (b)), and performance varies across 15 financial domain topics (see panel (c), the abbreviation list of topics provided in Table 2). The key errors shown in panel (d) categories include image recognition failures, incorrect formula application, question misunderstanding, and answer not found.

1 INTRODUCTION

Financial reasoning is a critical analysis process that involves leveraging expert-level knowledge to derive insights from diverse financial data and logically conclude a decision. Effective financial reasoning can lead to wise decisions that generate substantial monetary benefits or avoid costs

amounting to billions of dollars (Jerven, 2013; MacKenzie, 2008; Chen et al., 2022c). However, analyzing and reasoning from financial data is inherently complex. Unlike general-domain reasoning tasks, financial reasoning requires the integration of information from heterogeneous sources such as structured data (e.g., financial tables), semi-structured data (e.g., regulatory filings), and unstructured data (e.g., economic reports). These tasks demand advanced mathematical rigor, including multi-step calculations, statistical analysis, and domain-specific formulas for valuation, credit risk, and liquidity analysis. Additionally, financial reasoning relies on a deep understanding of complex concepts like portfolio optimization and risk modeling.

Recent advancements in Large Language Models (LLMs), such as GPT-o1 and DeepSeek-R1, have significantly mitigated the challenges associated with reasoning over financial text data (OpenAI, 2024a; DeepSeek-AI et al., 2025). However, many financial problems also involve visual data, such as stock price trends, financial tables, and statistical charts, which require integrating multimodal information for effective reasoning. Multimodal Large Language Models (MLLMs), capable of processing both textual and visual inputs, have demonstrated remarkable capabilities in addressing such complex tasks. These models have been evaluated on prominent multimodal benchmarks such as MMMU (Yue et al., 2024) and Math Vista (Lu et al., 2024), and documented in official technical reports (OpenAI, 2024a; Team et al., 2024; Dubey et al., 2024; Anthropic, 2024; Li et al., 2024a). Despite these advancements, existing multimodal benchmarks (see Table 1) primarily evaluate MLLM's general reasoning capabilities in open domains. The sole exception, FAMMA (Xue et al., 2024), is limited in scope, encompassing only 1758 examples across 8 topics. Key financial areas, such as risk management, valuation, and liquidity analysis, are inadequately represented, leaving the financial reasoning capabilities of MLLMs largely unexplored. To address this gap, we would like to answer the critical question: ***What are the multimodal reasoning capabilities and limitations of MLLMs in the financial domain?***

To answer this question, first, we propose **FinMR**, a comprehensive benchmark specifically designed for evaluating the capabilities of MLLMs in financial reasoning tasks. **FinMR** spans 15 diverse financial topics and includes 3,200 college-level question-answer pairs that combine textual and visual content. These questions are carefully curated to encompass a wide range of visual data, including table images, stock price trends, and statistical distributions, as shown in Figure 1(a). The benchmark is divided into 1,049 financial math questions, each requiring advanced mathematical skills such as calculus and statistics, and 2,151 financial expertise questions, which necessitate a deep understanding of domain-specific financial knowledge. Each example includes manually annotated expert explanations, facilitating both the evaluation of reasoning capabilities and detailed error analysis. To support model development and evaluation, we split the dataset into an 80% *development* set (i.e., 2560 examples) and a 20% *test* set (640 examples).

Second, we conduct a comprehensive evaluation of state-of-the-art LLMs and MLLMs on **FinMR**, identifying significant performance gaps across models. Notably, we test ***LLMs with text input***, including GPT-o1 (OpenAI, 2024b), Gemini-1.5-Pro (GeminiTeam et al., 2024), Claude-3.5-Sonnet (Anthropic, 2024), Llama 3.2(Dubey et al., 2024), Deepseek (DeepSeek-AI et al., 2025), and Qwen (Qwen, 2024), alongside ***MLLMs with text & image input***, such as GPT-4o (OpenAI, 2024a), Gemini-1.5-Pro (GeminiTeam et al., 2024), Claude-3.5-Sonnet (Anthropic, 2024), Llama-3.2-Vision (AI@Meta, 2024), Qwen-VL-Plus (Bai et al., 2023), LLaVa-NEXT (Li et al., 2024a). For evaluation, we adopt two methods, Chain-of-Thought (CoT) prompting (Wei et al., 2022b; Li et al., 2024b) and our proposed *Error Feedback Learning (EFL)* method. Inspired by prior studies that emphasize the value of error feedback in reasoning tasks (Yan et al., 2024b; Wang et al., 2024b; Lu et al., 2023), we construct an *error database* using the development data. This database contains negative examples paired with AI-driven feedback and provides a foundation for the EFL method. EFL leverages our error database to retrieve similar negative examples and associated feedback. This approach improves reasoning performance without the need for additional model training and supports effective retrieval-based reasoning.

Finally, we perform a detailed error analysis to identify key challenges in multimodal financial reasoning and provide guidance for future improvements. Our findings reveal that: (1) Multimodal inputs significantly enhance reasoning capabilities, though image recognition remains a critical bottleneck. Notably, Gemini-1.5-Pro (with text and image) achieves 82.06% accuracy, while Gemini-1.5-Pro (with text and image caption) has only 61.37% accuracy. This verifies that MLLM has higher advantages by direct image input; (2) EFL strategy comprehensively surpasses CoT, and the greatest improvement is 12.44% of Qwen VL, indicating large models benefiting from error feedback;

Table 1: Existing Reasoning Benchmarks versus FinMR

Benchmark	Domain	Modality	Level	Source	Number	Include Math?	Financial Expertise?	Solution Format
MathVista Lu et al. (2024)	Open	Text & Image	Elem. to College	Internet+Expert	6141	Yes	Few	Text
MMMU Yue et al. (2024)	Open	Text & Image	College	Internet, Text-books, Lecture	11500	Yes	Few	Text
MATH-V Wang et al. (2024a)	Math	Text & Image	Elem., High School	Internet	2252	Yes	Few	Text
FinQA Chen et al. (2022c)	Finance	Text Only	College	Expert	8281	Yes	Yes	Math Program
TAT-QA Zhu et al. (2021)	Finance	Text Only	College	Expert	16552	Yes	Yes	Text
MultiHierrt Zhao et al. (2022)	Finance	Text Only	College	Expert	10440	Yes	Yes	Text
DocMath-Eval Zhao et al. (2024b)	Finance	Text Only	College	Internet+Expert	5974	Yes	Yes	Python Program
FinanceMath Zhao et al. (2024a)	Financial Math	Text Only	College	Internet+Expert	1200	Yes	Yes	Python Program
FAMMA Xue et al. (2024)	Finance	Text & Image	College	Textbook	1758	few	Yes	Text
*** FinMR (Ours)**	Finance	Text & Image	College, Profession	Internet+Expert	3700	Yes	Yes	Text

(3) Financial math reasoning remains particularly challenging, with models scoring approximately 10% lower on math-related tasks compared to financial expertise tasks. Figure 1(d) highlights three primary error types in the reasoning process: image recognition failure, question misunderstanding, and incorrect formula application. Among them, image recognition is the most significant bottleneck, underscoring the need for advanced techniques in visual content understanding within the financial domain.

The key contributions of this study are summarized as follows:

- FinMR, the first comprehensive multimodal reasoning benchmark across 15 financial topics. The benchmark includes 3,200 question-answer pairs with explanations annotated manually, and each pair has text and diverse types of images, aiming at evaluating MLLM abilities in knowledge-intensive reasoning and analyzing their error in intermediate reasoning steps.
- EFL, a novel strategy to improve large models' reasoning abilities on **FinMR**. We construct a database of negative examples with AI-driven error feedback, which contributes to self-learning from previous mistakes and next iterate learning, evidencing the effectiveness of error feedback.
- Comprehensive experiments over mainstream LLMs and MLLMs which demonstrate that MLLMs consistently outperform standard LLMs. Our findings highlight the specific performance gap between these models, with Gemini-1.5-Pro emerging as the best-performing MLLM on **FinMR**. Notably, mathematical reasoning performance sees a significant improvement when direct image input is utilized.

2 Related Works

2.1 Reasoning Benchmarks

Early multimodal reasoning benchmarks, such as GeoQA (Chen et al., 2022b) and GeoQA+ (Chen et al., 2022a), are narrow in scope, predominantly addressing plane geometry problems. More recent multimodal math reasoning datasets, such as MMMU (Yue et al., 2024), Math Vista (Lu et al., 2024), and Math-Vision (Wang et al., 2024a), broaden the subjects coverage and difficulty levels. However, these benchmarks focus on general mathematical reasoning and lack domain-specific content. Although FAMMA (Xue et al., 2024) incorporates financial mathematical reasoning examples, it remains limited in scale, with only 1,758 examples across eight topics. Furthermore, FAMMA's scope does not fully capture the complexity of multimodal financial reasoning tasks.

Several benchmarks evaluate textual reasoning in finance, focusing on financial statements and reports. For instance, TAT-QA (Zhu et al., 2021) combines tables and text for numerical reasoning, while FinQA (Chen et al., 2022c) offers 8,281 expert-annotated QA pairs requiring math operations like addition and comparison. These tasks demand significant financial knowledge, making them more complex than typical QA. MultiHiertt (Zhao et al., 2022) and DocMath-Eval (Zhao et al., 2024b) focus on tabular data for financial reasoning, with MultiHiertt incorporating hierarchical tables and unstructured text for complex tasks. FinanceMath (Zhao et al., 2024a) further combines text and tables with expert annotations. However, real-world financial data often includes diverse visuals like price charts, financial statement screenshots, and diagrams, which provide critical insights for reasoning. To address this, we introduce **FinMR**, a comprehensive multi-modal benchmark covering 15 financial topics and integrating seven visual data types, offering a robust foundation for evaluating multimodal reasoning in finance.

2.2 Methods for Stimulating Inherent Reasoning Capability

Chain of Thought (CoT). Reasoning capabilities in large models have traditionally been enhanced through pre-training and fine-tuning methods (Yan et al., 2024a; Liang et al., 2023; Shao et al.,

2024; Liu et al., 2024). While effective, these approaches often involve significant computational and time costs. In contrast, prompt-based methods such as CoT prompting provide a more time-efficient and computationally cost-effective alternative (Wei et al., 2022b; Kojima et al., 2022). CoT enables models to articulate intermediate reasoning steps explicitly, which enhances their ability to process complex queries and arrive at accurate conclusions. This method has been integrated into QA systems, including financial reasoning tasks, to generate detailed reasoning steps before producing an answer (Chen et al., 2021; Zhu et al., 2021; Zhao et al., 2022; 2024a;b). Recent advancements have extended CoT from textual reasoning to multimodal domains, enabling models to process and reason across diverse modalities. Notable contributions in this area include the works of Wang et al. (2024a), Yue et al. (2024), Lu et al. (2024) and Zhang et al. (2024), which leverage CoT to enhance multimodal understanding and decision-making. These approaches enable models to process visual, textual, and other data types, allowing for more complex reasoning processes. We will also adopt CoT prompting to evaluate MLLMs' financial reasoning capabilities on **FinMR**.

Error Feedback. Well-pre-trained LLMs and MLLMs possess an inherent learning capacity, which reduces their hallucination issues by leveraging external materials (Yu et al., 2024; Tan et al., 2024; Zhao et al., 2023a; Liu et al., 2024) and simulating the given examples (Tsimpoukelli et al., 2021; Chen et al., 2023; Wei et al., 2022a). This capacity has been further enhanced through the application of in-context learning, which has been extended to multimodal tasks, including complex reasoning (Liu et al., 2024; Zhao et al., 2023b; Zhang et al., 2024). One promising approach within this paradigm is learning from error feedback, which involves using prior mistakes to improve reasoning performance. Several studies highlight the value of error feedback in enhancing model performance on multimodal mathematical reasoning tasks (Yan et al., 2024b; Wang et al., 2024b; Lu et al., 2023; Sun et al., 2024). Building on this foundation, our work proposes a novel EFL strategy to retrieve similar error feedback from an error database. This approach allows models to iteratively refine their reasoning capabilities by analyzing errors and leveraging corrective feedback.

3 THE FINMR BENCHMARK

3.1 OVERVIEW OF FINMR

We introduce the Financial Multimodal Reasoning (**FinMR**) benchmark, a curated resource designed to evaluate the financial reasoning capabilities of large models across diverse topics and multimodal contexts. **FinMR** encompasses 15 topics in finance, ranging from *Investment* to *Liquidity and Treasury Risk*, as detailed in Table 2. The benchmark includes 3,200 high-quality QA pairs with explanations, split into 2,151 expertise-based QA pairs and 1,049 math QA pairs, as shown in Table 3. All questions in our benchmark are manually collected from financial exam papers at top universities and are available on the website[1], ensuring the dataset represents expert-level financial reasoning tasks. More details are presented in the Appendix A.1

FinMR evaluates three critical skills in MLLMs: (1) visual information understanding, (2) intensive domain-specific knowledge involvement in finance, and (3) reasoning. Unlike traditional benchmarks, **FinMR** presents significant challenges by requiring models to process and integrate diverse, heterogeneous image types, including financial tables, stock price trends, and statistical charts, alongside textual information. This benchmark extends beyond basic visual recognition to demand a sophisticated multimodal approach that combines advanced analytical capabilities with mathematical and financial expertise.

3.2 DATA COLLECTION AND COMPLIANCE.

The data collection process for **FinMR** was conducted in two stages. In the first stage, we compiled a collection of financial exam papers from both college-level courses and professional certification programs. In particular, we focused on exam papers from business schools that officially collaborate with Chartered Financial Analyst (CFA) and Financial Risk Management (FRM), which are internationally recognized institutions that provide exams for financial certificates. The curricula of these institutions are often integrated into university course designs. Consequently, final exam papers from these programs represent a valuable resource for developing expert-level reasoning tasks. We extracted all QA pairs from past final exam papers from these business schools. For exam questions available in PDF format, we utilized the Mathpix API (Wang et al., 2024a) to extract textual

[1] https://www.studocu.com/en-nz

Table 2: Financial Topic Distribution of FinMR.

Topic & Abbreviation	Number	Ratio
Investment (Inv)	371	11.6%
Quantitative Methods (QM)	342	10.7%
Valuation and Risk Models (VRM)	318	9.9%
Financial Markets and Products (FMP)	297	9.3%
Financial Reporting and Analysis (FRA)	264	8.3%
Portfolio Management (PM)	258	8.1%
Fixed Income (FI)	251	7.8%
Credit Risk (CR)	170	5.3%
Foundation of Risk Management (FRM)	169	5.3%
Economics (Eco)	156	4.9%
Operational Risk (OR)	131	4.1%
Derivatives (Der)	126	3.9%
Market Risk (MR)	121	3.8%
Corporate Finance (CF)	119	3.7%
Liquidity and Treasury Risk (LTT)	107	3.3%

Table 3: FinMR Benchmark Statistics.

Statistics	Number	Ratio
Total Questions	3200	100%
* Test	640	20%
* Develop	2560	80%
Total Images	3764	100%
* QA with Single Image/# of Images	2643/2643	70%
* QA with Multiple Images/# of Images	557/1118	30%
Reasoning Type		
* Expertise Reasoning QA	2151	-
* Math Reasoning QA	1049	-
Average Length		
* Context	327.74	-
* Question	33.97	-
* Explanation	63.24	-

content, mathematical formulas, and images. To maintain consistency in dataset formatting, we extracted only images present in the questions while excluding those from options and explanations. In the second stage, we enlisted two PhD students specializing in Finance, both of whom have passed the CFA and FRM exams. These experts manually verified the correctness of the explanations and filtered out QA pairs that lacked correct answers or high-quality explanations. This rigorous verification process ensured that the dataset comprised QS pairs with detailed expert-validated explanations.

3.3 Data Quality Assurance.

We employed a three-stage data curation process with six annotators (four master's and two PhD students from Computer Science and Finance). In the first stage, we categorized questions as expertise-based or math reasoning using GPT-4o and assigned reference topics. Four annotators manually verified the alignment between questions, images, options, answers, and explanations, correcting extraction errors. Image clarity was enhanced using Image Generator Pro, and non-English or incomplete explanations were removed, resulting in 4,470 QA pairs (30% math-focused financial QA, 70% financial expertise QA). In the second stage, two PhD students reviewed the validity, completeness, and clarity of explanations. They expanded reasoning steps and addressed grammatical and stylistic issues to ensure high-quality financial reasoning. After this refinement, the dataset was reduced to 3,200 high-quality QA pairs. In the final stage, each question was labeled with relevant metadata, including question ID, source topics, and question type. The dataset was then split by topic: 80% (2,560 samples) for development and 20% (640 samples) for testing.

4 Evaluation Framework

This section outlines our evaluation framework for assessing the financial reasoning capabilities of large models using **FinMR**. Specifically, we discuss the error feedback database construction, the evaluation process for large models, and the prompting methods used in our experiments.

4.1 Error Database Construction

A key component of our framework is the construction of an error feedback database, which is integral to the Error Feedback Learning (EFL) method. This database enables systematic analysis of model errors and facilitates iterative refinement of reasoning capabilities. The construction process, shown in Figure 2, involves three stages: *data input*, *feedback generation*, and *storage*.

In the data input stage, we input context, questions, images, and options from the **FinMR** *development dataset* into the evaluated models. To accommodate LLMs that cannot directly process images, we use GPT-4o to convert images into textual descriptions, whereas MLLMs directly process the visual data alongside textual content. This approach follows established methodologies, such as those proposed in Wang et al. (2024a). In the feedback generation stage, models generate step-by-step reasoning and derive final answers. We then compare these answers against ground truth. For incorrect responses, we employ a feedback prompt (depicted in Figure 3) to guide the model in refining its reasoning steps. This process is further supported by manually annotated explanations. The

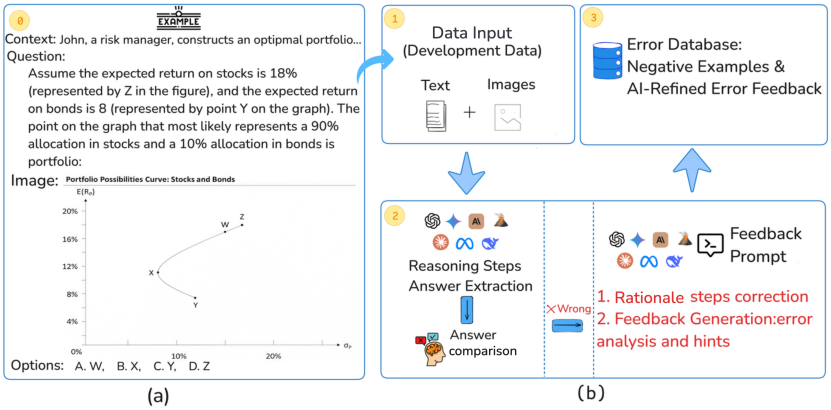

Figure 2: Panel (a) provides a typical example of FinMR. Panel (b) exhibits three stages of error database construction. In the second stage, Large models leverage the annotated explanations to generate correct reasoning steps and hits.

generated feedback includes error analysis which elaborates on the specific reasoning flaws, and actionable guidance for addressing similar problems in the future. Finally, in storage stage, we save all relevant data, including the input examples, refined error feedback, and metadata (e.g., question ID, model information)in an external database.

[System Input]:
You are a financial expert. You will be given questions and options, possibly with context information and images. Also, you will be given wrong reasoning steps and correct reasoning hints. You are supposed to give feedback in Markdown format. The feedback includes:
1. Refine correct reasoning steps with a given explanation.
2. Compare the correct reasoning steps and the model's wrong reasoning steps and highlight the difference.
3. Summarize the hint for future similar questions.

[User Input]:
Context: {context}; Images in Context: {images}; Question: {question}; Images in Question: {images}; Options: {options}; Wrong Answer: {model answer}; Wrong Reasoning Steps: {model reasoning steps}; Correct Answer: {answer}; Explanation: {explanation}.

Figure 3: Feedback Prompt Template

[System Input]:
You are a financial expert and are supposed to answer the given questions with options, context information, and images. Also, You will be given previous learning documents, including questions and options, possibly with context information and images. Please answer the current question. The output reasoning steps are in Markdown format. Finally, you must put the correct option (A, B, C, or D) in **[ ]**.
e.g. Therefore, the correct option is **[B]**.

[User Input]:
Retrieved Example: {example}; Context: {context}; Images in Context: {images}; Question: {question}; Images in Question: {images}; Options: {options}

Let's think step by step to answer the given question.

Figure 4: EFL Prompt Template

4.2 Evaluation Process

The evaluation process consists of four stages: *test data input, reasoning, output, and evaluation*, as displayed in Figure 5. The test data input stage follows the same methodology as the *data input stage* described in the construction of the error feedback database, including how we accommodate both LLMs and MLLMs (see Section 4.1), with the key distinction that the evaluation is performed using the test dataset.

The reasoning stage employs two distinct methods: CoT and EFL. CoT prompting involves guiding models to generate step-by-step reasoning through simple instructions, such as *"Let's think step by step"* followed by the user input. In EFL, the model retrieves the most similar negative example

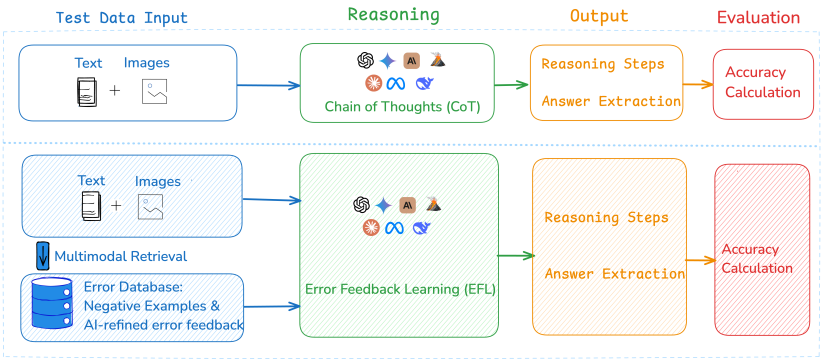

Figure 5: Four stages of the evaluation process. For LLMs with no visual ability, we leverage GPT-4o to generate image captions to support reasoning tasks. The process adopts two methods, CoT and EFL. The latter retrieves the most similar (i.e., top-1 semantic similarity) negative examples and error feedback for learning.

and its error feedback from the previously constructed error database. The EFL prompt, presented in Figure 4, incorporates this feedback into the reasoning process. The goal is to allow the model to learn from prior mistakes and refine their reasoning steps. This iterative retrieval mechanism is a novel contribution. For both reasoning methods, we clarify the format of the reasoning outputs using markdown, following practices outlined in (Zhao et al., 2024b; Wang et al., 2024a).

In the output stage, we adopt the answer extraction pipeline inspired by (Chen et al., 2024; Zhao et al., 2024a). If the final answer is encapsulated in double square brackets (e.g., `[A]`), it is directly identified as the model's response. If no such format is detected, the output is categorized as *"Answer not Found"*, which is regarded as an incorrect response. In the evaluation stage, the extracted answers are compared against the ground truth. The accuracy ratio is computed as the proportion of correct responses to the total number of questions. This is the primary metric for our evaluation.

5 Results

5.1 LLM, MLLM, and Experiment Setup

We evaluate the following LLMs on **FinMR**:

- **Closed-source:** GPT-o1 (OpenAI, 2024b), Gemini 1.5 Pro (GeminiTeam et al., 2024), Claude-3.5-Sonnet (Anthropic, 2024), Deepseek (DeepSeek-AI et al., 2025);
- **Open-source:** Llama 3.2 (Dubey et al., 2024), Qwen (Qwen, 2024).

We also evaluate the following closed-source and open-source MLLMs on **FinMR**:

- **Closed-source:** GPT-4o (OpenAI, 2024a), Gemini-1.5-Pro (GeminiTeam et al., 2024), Claude-3.5-Sonnet (Anthropic, 2024);
- **Open-source:** Llama-3.2-Vision (AI@Meta, 2024), Qwen-VL-Plus (Bai et al., 2023), LLaVa-NEXT (Li et al., 2024a).

All experiments on open-source models were conducted using A100 GPUs, while experiments on closed-source models were performed using 4090 GPUs. Additionally, we used LangSmith to trace all the experiments and set the temperature of large models to 0.7.

5.2 Evaluation Results

We now present and analyze our experiment results in detail. Detailed results for mathematical reasoning and expertise reasoning are presented in Table 4 and Figure 1(b), while performance across different financial topics are shown in Table 4 and Figure 1 (c). We summarize and analyze key findings as follows:

Table 4: Reasoning Performance Comparison of LLMs and MLLMs on **FinMR**. We highlight the best model's performance in green (LLMs) and blue (MLLMs).

	Model	Method	Overall	Math	Expertise	Inv	QM	VRM	FMP	FRA	PM	FI	FRM	CR	Eco	OR	Der	MR	CF	LTR
Textual Modality: Text + Image Caption																				
Closed-source	Claude-3.5-Sonnet	CoT	53.91	42.91	61.83	52.50	30.00	65.38	52.17	22.22	48.10	60.24	37.50	52.83	18.18	30.77	22.22	65.00	47.62	56.52
	Claude-3.5-Sonnet	EFL	64.84	55.97	71.24	65.00	50.00	78.85	65.22	22.22	60.76	71.08	62.50	60.38	45.45	46.15	33.33	75.00	61.90	56.52
	GPT-o1	CoT	46.56	33.96	55.65	55.00	35.71	59.62	56.52	33.33	37.97	53.01	18.75	45.28	18.18	38.46	66.67	65.67	40.48	30.43
	GPT-o1	EFL	60.31	49.25	68.28	67.50	50.00	69.23	65.22	55.56	49.37	68.67	31.25	61.32	36.36	46.15	77.78	65.83	54.76	47.83
	Gemini-1.5-Pro	CoT	47.81	34.70	57.26	47.50	28.57	57.69	60.87	33.33	36.71	53.01	6.25	49.06	54.55	38.46	55.56	61.67	35.71	21.74
	Gemini-1.5-Pro	EFL	61.37	51.49	68.45	60.00	50.00	65.38	69.57	77.78	46.84	70.59	18.75	65.09	72.73	61.54	55.56	74.17	42.86	39.13
	DeepSeek-R1	CoT	61.25	66.42	57.53	65.00	64.29	61.54	78.26	55.56	53.16	59.04	56.25	61.32	54.55	46.15	66.67	72.50	83.33	73.91
	DeepSeek-R1	EFL	71.88	79.10	66.67	80.00	78.57	76.92	91.30	66.67	65.82	63.86	68.75	70.75	72.73	46.15	66.67	72.50	83.33	73.91
Open-source	LLaMa 3.2	CoT	27.34	23.51	30.11	27.50	21.43	23.08	39.13	0.00	25.32	36.14	18.75	23.58	9.09	7.69	11.11	32.50	35.71	21.74
	LLaMa 3.2	EFL	36.09	32.09	38.98	40.00	21.43	28.85	52.17	33.33	35.44	46.99	18.75	26.42	9.09	7.69	44.44	43.33	42.86	34.78
	Qwen	CoT	55.62	52.99	57.53	55.00	35.71	51.92	52.17	33.33	50.63	60.24	31.25	62.26	45.45	30.77	55.56	57.50	71.43	56.52
	Qwen	EFL	67.97	68.28	67.74	75.00	35.71	67.31	69.57	55.56	63.29	66.27	56.25	76.42	72.73	46.15	55.56	65.00	80.95	78.26
Multimodality: Text + Image																				
Closed-source	GPT-4o	CoT	72.19	71.43	73.17	65.00	71.43	71.15	73.91	66.67	73.42	74.70	81.25	71.70	63.64	38.46	77.78	73.33	76.19	78.26
	GPT-4o	EFL	81.72	83.08	81.03	82.50	85.71	82.69	78.26	77.78	82.28	80.72	93.75	81.13	72.73	76.92	77.78	80.00	83.33	91.30
	Gemini-1.5-Pro	CoT	70.83	70.26	71.24	82.50	50.00	69.23	69.57	55.56	64.56	64.29	81.25	74.53	63.64	53.85	88.89	72.50	83.33	69.57
	Gemini-1.5-Pro	EFL	82.06	83.27	81.18	85.00	71.43	82.69	86.96	66.67	72.15	77.38	93.75	88.68	72.73	84.62	100.00	80.00	90.48	86.96
	Claude-3.5-Sonnet	CoT	75.94	73.13	77.96	75.00	57.14	80.77	69.57	77.78	68.35	79.52	75.00	76.42	63.64	53.85	66.67	76.67	88.10	73.91
	Claude-3.5-Sonnet	EFL	80.78	80.22	81.18	85.00	64.29	86.54	78.26	77.78	75.95	84.34	81.25	82.08	63.64	61.54	77.78	80.00	90.48	78.26
Open-source	Qwen-VL	CoT	52.66	43.66	59.14	55.00	35.71	57.69	56.52	22.22	44.30	62.65	25.00	51.89	36.36	23.08	33.33	62.50	47.62	60.87
	Qwen-VL	EFL	65.00	57.09	70.70	70.00	57.14	73.08	56.52	44.44	60.76	75.90	43.75	65.09	63.64	38.46	33.33	70.00	57.14	65.22
	LLaVa-NEXT	CoT	16.72	14.18	18.15	20.00	7.14	19.23	26.09	15.00	18.99	21.69	12.50	16.04	9.09	7.69	11.11	20.00	4.76	4.35
	LLaVa-NEXT	EFL	28.28	26.87	29.30	25.00	17.14	34.62	26.09	22.22	29.11	38.55	12.50	31.13	27.27	30.77	11.11	27.50	14.29	30.43
	LLaMa-3.2-Vision	CoT	19.38	13.81	23.39	5.00	20.30	21.15	43.48	43.00	2.53	18.07	12.50	15.09	27.27	15.38	22.22	34.17	28.57	26.09
	LLaMa-3.2-Vision	EFL	27.19	22.01	30.91	15.00	29.20	30.77	47.83	77.00	10.13	24.10	31.25	23.58	36.36	38.46	33.33	43.33	38.10	30.43

Disparity between Open-source Models and Closed-source Models: The results on **FinMR** reveal insights into the comparative performance of state-of-the-art LLMs and MLLMs, as highlighted in Table 4 (marked in blue and green). Closed-source models consistently outperform open-source counterparts. In particular, the textual LLM DeepSeek-R1 and multimodal Gemini-1.5-Pro gained 71.88% and 82.06% overall accuracy, respectively. In contrast, open-source models such as LLaMa 3.2 and LLaVa-NEXT demonstrate significantly lower overall performances, with accuracies falling below 30%. These results highlight a critical disparity between open-source and closed-source models. Open-source multimodal models like LLaVa_NEXT and LLaMa_3.2_Vision performed below expectations, indicating the challenges faced by open-source approaches in achieving competitive performance on complex multimodal tasks.

Effectiveness of Error Feedback Learning: The comparison between CoT prompting and EFL highlights the effectiveness of leveraging error feedback to enhance model performance. Across all evaluated models, accuracy improved significantly when EFL was applied. For example, GPT_o1, a reasoning-focused model, initially scored 46.56% but saw a significant 13.75% improvement (reaching 60.31%) after incorporating negative examples with constructive feedback. Similarly, multimodal models Gemini-1.5-Pro and Qwen_VL achieved nearly 12% improvements after adopting EFL. This underscores the potential of high-quality error databases in refining reasoning capabilities. Moreover, the consistent performance gains across diverse models demonstrate that EFL is a robust and generalizable approach for improving reasoning in the financial domain.

Impact of Image Captions and Direct Image Inputs: As discussed above, for LLMs that lack inherent visual processing capabilities, images were captioned using GPT-4o to supplement their visual understanding. Textual-modality models trained on diverse datasets delivered moderate performance, with most achieving over 40% accuracy, except for LLaMa 3.2, which struggled to perform on par with its peers. Among textual models, DeepSeek-R1 (EFL) stood out, achieving 71.88% accuracy and outperforming other closed-source LLMs such as GPT-o1. In contrast, MLLMs utilizing direct image input demonstrated significantly higher accuracy. Gemini-1.5-Pro (EFL) excelled as the top-performing multimodal model, achieving 82.06% accuracy, and slightly surpassing Claude-3.5-Sonnet. This highlights the enhanced reasoning capabilities enabled by direct image inputs. By enabling richer context through integrated textual and visual information, multimodal models demonstrate their superiority in addressing intricate, knowledge-intensive tasks.

Challenge of Financial Math Reasoning: Table 4 reveals a significant gap between mathematical and expertise reasoning tasks. Models like GPT-o1, Claude-3.5-Sonnet, and LLaMA-3.2 (EFL) show lower accuracy in mathematical reasoning (e.g., 32.09% for LLaMA-3.2) compared to expertise reasoning (38.98%), due to the former's need for logical rigor and multi-step calculations. In contrast, expertise reasoning relies more on contextual understanding. Closed-source multimodal models like Gemini-1.5-Pro (EFL) achieve over 80% accuracy in mathematical reasoning, outperforming expertise reasoning (83.27% vs. 81.18%). Figure 1(b) confirms that multimodal inputs enhance performance in both tasks, though challenges remain in integrating textual and visual information. Visual results in Figure 1(c) show that multimodal inputs significantly improve performance

in numerically complex topics like Valuation and Risk Models (VRM) and Fixed Income (FI), but less so in less structured topics like Operational Risk and Economics. Although math reasoning tasks are more challenging, visual input makes a contribution to reasoning accuracy. Future work should focus on improving text-image interaction.

5.3 Error Type Analysis

To gain deeper insights into the limitations of the tested models, we conducted an error analysis to identify common challenges encountered during the reasoning process. Errors were categorized into five primary types: ***image recognition failure, question misunderstanding, incorrect formula application, answer not found***, and ***others***. Figure 1(d) summarizes the prevalence of each error type, while Table 5 provides representative examples.

Among these categories, **image recognition failures** emerged as the most significant one, accounting for 72.84% of total errors. As shown in Table 5, many reasoning steps reveal that the provided images lack sufficient direct information for problem-solving. Moreover, a substantial portion of these images require domain-specific expertise to extract implicit information effectively, which current models struggle to achieve. This highlights the need for more sophisticated visual understanding capabilities, particularly in tasks involving specialized financial visuals such as charts and tables.

Another prominent issue is **question misunderstanding**, particularly for questions within specialized financial domains at the college level. Models frequently misinterpret the intent or nuances of these questions, leading to incorrect reasoning steps. This limitation underscores the importance of integrating deeper contextual understanding into models, especially for domain-specific tasks.

Even when questions and images are correctly interpreted, models often fail in **applying the correct formulas**. This issue is especially prevalent in harder-level questions requiring the integration of knowledge across multiple topics or the application of cross-domain financial formulas. These errors suggest that current models lack the ability to handle the logical rigor and multi-step calculations necessary for complex mathematical reasoning tasks.

The ***answer not found*** error type is another recurring issue, particularly for models like LLaMa-3.2-Vision and LLaVa-NEXT, which often fail to produce a final answer during the reasoning process. Our study revealed that for these models, unrestricted output tokens resulted in a **repetition problem**, where the reasoning output becomes repetitive. This issue, as shown in the first example of Table 5, leads to inconsistencies, overly lengthy reasoning steps, and ultimately incomplete answers. This phenomenon is particularly detrimental for tasks requiring long reasoning chains and highlights the importance of managing token limits.

In summary, these errors arise from technical limitations, such as image recognition failures and output repetition, and a lack of financial domain expertise. While our analysis provides a foundational understanding of error types, space limitations prevent us from presenting a full systematic error analysis. We believe that such an in-depth analysis would provide meaningful insights into the reasoning capabilities and shortcomings of large models and should be a focus of future work.

6 Conclusion and Future Work

This paper introduced **FinMR**, a new benchmark tailored to evaluate the financial reasoning capabilities of multimodal models. Through evaluations of open-source and closed-source LLMs and MLLMs, we identified key insights and highlighted critical challenges in this domain. Our findings demonstrate three main conclusions: **(1)** MLLMs significantly outperform LLMs by effectively integrating textual and visual information, underscoring the importance of robust multimodal reasoning frameworks. However, image recognition remains a major bottleneck. **(2)** The Error Feedback Learning (EFL) method consistently outperformed Chain of Thought (CoT) prompting, validating the efficacy of leveraging negative examples with feedback to improve reasoning. **(3)** Financial math reasoning tasks consistently pose greater difficulty, with models achieving approximately 10% lower accuracy compared to expertise reasoning. Key challenges include incorrect formula application and question misunderstandings. Future work should focus on improving models' visual reasoning abilities, exploring more efficient, training-free methods to enhance reasoning, and systematically addressing the challenges identified in this study to enable more robust financial reasoning systems. A more detailed and systematic analysis of errors arising during the reasoning process of large models would provide deeper insights into the specific limitations and failure modes of both LLMs and MLLMs, particularly in complex tasks like financial reasoning.

REFERENCES

AI@Meta. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models, 2024. URL https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/.

Anthropic. Introducing computer use, a new claude 3.5 sonnet, and claude 3.5 haiku \ anthropic, 2024. URL https://www.anthropic.com/news/3-5-models-and-computer-use.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL http://arxiv.org/abs/2308.12966.

Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression, 2022a. URL https://arxiv.org/abs/2212.02746.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P. Xing, and Liang Lin. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning, 2022b. URL https://arxiv.org/abs/2105.14517.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. M^3CoT: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. arXiv, 2024. URL http://arxiv.org/abs/2405.16473.

Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Hao Zhang, and Chuang Gan. See, think, confirm: Interactive prompting between vision and language models for knowledge-based visual reasoning, 2023. URL https://arxiv.org/abs/2301.05226.

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. pp. 3697–3711. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.300. URL https://aclanthology.org/2021.emnlp-main.300.

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. arXiv, 2022c. URL http://arxiv.org/abs/2109.00122.

DeepSeek-AI, Daya Guo, and Dejian Yang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and Kadian. The llama 3 herd of models, 2024. URL http://arxiv.org/abs/2407.21783.

GeminiTeam, Natalie Clay, Tomas Kocisky, Bartek Perz, Dian Yu, and Howard. Gemini: A family of highly capable multimodal models, 2024. URL http://arxiv.org/abs/2312.11805.

Morten Jerven. Poor numbers–how we are misled by african development statistics and what to do about it (uzuazo etemire). *VRÜ Verfassung und Recht in Übersee*, 46(3):336–340, 2013. URL http://www.jstor.org/stable/43239700.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 22199–22213. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf.

Kaichen Li, Hao Zhang, Renrui Zhang, Dong Guo, Feng Li, Yuanhan Zhang, Ziwei Liu, Chun Yuan, and Bo Li. LLaVA-NeXT: Stronger LLMs supercharge multimodal capabilities in the wild, 2024a. URL https://llava-vl.github.io/blog/2024-05-10-llava-next-stronger-llms/.

Suyan Li, Fuxiang Huang, and Lei Zhang. A survey of multimodal composite editing and retrieval, 2024b. URL http://arxiv.org/abs/2409.05405.

Zhenwen Liang, Tianyu Yang, Jipeng Zhang, and Xiangliang Zhang. Unimath. pp. 7126–7133. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.440. URL https://aclanthology.org/2023.emnlp-main.440.

Bingshuai Liu, Chenyang Lyu, Zijun Min, Zhanyu Wang, Jinsong Su, and Longyue Wang. Retrieval-augmented multi-modal chain-of-thoughts reasoning for large language models, 2024. URL http://arxiv.org/abs/2312.01714.

Pan Lu, Liang Qiu, Wenhao Yu, Sean Welleck, and Kai-Wei Chang. A survey of deep learning for mathematical reasoning. arXiv, 2023. doi: 10.48550/arXiv.2212.10535. URL http://arxiv.org/abs/2212.10535.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts, 2024. URL http://arxiv.org/abs/2310.02255.

Donald MacKenzie. *An engine, not a camera: How financial models shape markets*, volume 48. Mit Press, 2008. doi: https://doi.org/10.1353/tech.2007.0154.

OpenAI. Introducing GPT-4o and more tools to ChatGPT free users | OpenAI, 2024a. URL https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/.

OpenAI. Introduction OpenAI o1, 2024b. URL https://openai.com/o1/.

Team Qwen. Qwen2.5-LLM: Extending the boundary of LLMs, 2024. URL http://qwenlm.github.io/blog/qwen2.5-llm/.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models, 2024. URL http://arxiv.org/abs/2402.03300.

Shilin Sun, Wenbin An, Feng Tian, Fang Nan, Qidong Liu, Jun Liu, Nazaraf Shah, and Ping Chen. A review of multimodal explainable artificial intelligence: Past, present and future. arXiv, 2024. doi: 10.48550/arXiv.2412.14056. URL http://arxiv.org/abs/2412.14056.

Cheng Tan, Jingxuan Wei, Linzhuang Sun, Zhangyang Gao, Siyuan Li, Bihui Yu, Ruifeng Guo, and Stan Z. Li. Retrieval meets reasoning: Even high-school textbook knowledge benefits multimodal reasoning, 2024. URL http://arxiv.org/abs/2405.20834.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, and Gulati. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL http://arxiv.org/abs/2403.05530.

Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. arXiv, 2021. URL http://arxiv.org/abs/2106.13884.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with MATH-vision dataset, 2024a. URL http://arxiv.org/abs/2402.14804.

Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. Exploring the reasoning abilities of multimodal large language models (MLLMs): A comprehensive survey on emerging trends in multimodal reasoning, 2024b. URL http://arxiv.org/abs/2401.06805.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022a. URL https://arxiv.org/abs/2206.07682.

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. 2022b. URL http://arxiv.org/abs/2201.11903.

Siqiao Xue, Tingting Chen, Fan Zhou, Qingyang Dai, Zhixuan Chu, and Hongyuan Mei. Famma: A benchmark for financial domain multilingual multimodal question answering, 2024. URL https://arxiv.org/abs/2410.04526.

Xu Yan, Lei Geng, Ziqiang Cao, Juntao Li, Wenjie Li, Sujian Li, Xinjie Zhou, Yang Yang, and Jun Zhang. TabMedBERT: A tabular knowledge enhanced biomedical pretrained language model. IOS Press, 2024a. URL https://ebooks.iospress.nl/doi/10.3233/FAIA240674.

Yibo Yan, Jiamin Su, Jianxiang He, Fangteng Fu, Xu Zheng, Yuanhuiyi Lyu, Kun Wang, Shen Wang, Qingsong Wen, and Xuming Hu. A survey of mathematical reasoning in the era of multimodal large language model: Benchmark, method & challenges, December 2024b. URL http://arxiv.org/abs/2412.11936.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. VisRAG: Vision-based retrieval-augmented generation on multi-modality documents, 2024. URL http://arxiv.org/abs/2410.10594.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI, 2024. URL http://arxiv.org/abs/2311.16502.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models, 2024. URL http://arxiv.org/abs/2302.00923.

Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Xuan Long Do, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, and Shafiq Joty. Retrieving multimodal information for augmented generation: A survey, 2023a. URL http://arxiv.org/abs/2303.10868.

Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. MultiHiertt: Numerical reasoning over multi hierarchical tabular and textual data, 2022. URL http://arxiv.org/abs/2206.01347.

Yilun Zhao, Hongjun Liu, Yitao Long, Rui Zhang, Chen Zhao, and Arman Cohan. KnowledgeMath: Knowledge-intensive math word problem solving in finance domains, 2023b. URL https://arxiv.org/abs/2311.09797.

Yilun Zhao, Hongjun Liu, Yitao Long, Rui Zhang, Chen Zhao, and Arman Cohan. FinanceMath: Knowledge-intensive math reasoning in finance domains, 2024a. URL http://arxiv.org/abs/2311.09797.

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, 2024b. URL http://arxiv.org/abs/2311.09805.

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. URL http://arxiv.org/abs/2105.07624.

A APPENDIX

A.1 COMPARISONS WITH EXISTING BENCHMARKS

We outline the differences between **FinMR** and 9 reasoning benchmarks: MathVista (Lu et al., 2024), MMMU (Yue et al., 2024), MATH-V (Wang et al., 2024a), FinQA (Chen et al., 2022c), TAT-QA (Zhu et al., 2021), MultiHiertt (Zhao et al., 2022), DocMath-Eval (Zhao et al., 2024b), FinanceMath (Zhao et al., 2024a), and FAMMA (Xue et al., 2024). Table 1 provides a detailed comparison.

Comparison with Multimodal Benchmarks. Math Vista (Lu et al., 2024) is a consolidated mathematical reasoning benchmark in visual contexts, containing 6,141 examples categorized into seven reasoning types: *algebraic reasoning, arithmetic reasoning, geometry reasoning, logical reasoning, numeric common sense, scientific reasoning,* and *statistical reasoning*. While valuable, these tasks are primarily designed for elementary and high school levels. MMMU (Yue et al., 2024), with 11500 examples, extends reasoning benchmarks to college-level tasks. It is designed to evaluate the multi-disciplinary multimodal understanding and reasoning capabilities of MLLMs, covering 30 subjects across six disciplines. MATH-V (Wang et al., 2024a) involves 2,252 challenging questions derived from diverse competition datasets, including Math Kangaroo, UK [Grey, Pink, Junior, Senior], ACM, and AIME. These questions emphasize expert-level visual perception and deliberate reasoning across 16 subjects.

Unlike Math Vista, MU, and MATH-V, **FinMR** is designed specifically to focus on financial reasoning. It combines domain-specific expertise with mathematical reasoning. Furthermore, **FinMR** provides detailed, manually verified explanations for all QA pairs, averaging 61.43 words per explanation. As shown in Table 3, our benchmark has an average question length of 33.97 words, alongside extra context (average 327.74 words), which significantly exceeds the averages of Math Vista (15.6 words), MU (59.33 words), and MATH-V (42.3 words). By offering such comprehensive content,**FinMR** supports more advanced reasoning tasks and ensures that models are challenged with realistic financial scenarios.

Comparison with Financial QA Benchmarks. As summarized in Table 1, existing financial QA benchmarks primarily target LLMs and focus on specific subdomains within finance, with limited multimodal content. The sole exception, FAMMA (Xue et al., 2024), includes 1,758 QA pairs across eight topics related to the CFA exam. However, it omits key financial areas such as risk management, constraining its ability to comprehensively evaluate MLLMs. In contrast, **FinMR** incorporates 15 topics derived from CFA and FRM exams, offering broader coverage of essential financial concepts and enabling more thorough assessments of reasoning capabilities in MLLMs.

Benchmarks like FinQA (Chen et al., 2022c), TAT-QA (Zhu et al., 2021), and MultiHiertt (Zhao et al., 2022) primarily focus on numerical reasoning over financial tables extracted from real-world reports, such as earnings statements and financial accounting documents. While these datasets include large numbers of QA pairs: 8,281, 13,215, and 10,440 examples, respectively, they emphasize simpler numerical calculations or specific subtopics like financial reporting without addressing broader or more complex financial reasoning tasks. In contrast, **FinMR** includes mathematical reasoning questions that require advanced knowledge in mathematics and statistics, such as calculus, and spans seven distinct image types, including complex financial tables, as illustrated in Figure 1(a).

Regarding the format of reasoning steps, some benchmarks such as DocMath-Eval (Zhao et al., 2024b) and FinanceMath (Zhao et al., 2024a) utilize Python-based solutions for interpretability. While these code-based explanations aim to provide precision, they often lack intuitive readability, making error analysis more challenging. **FinMR** addresses this limitation by including detailed, manually annotated textual explanations for all 3,200 QA pairs. These explanations offer a richer and more interpretable resource for analyzing reasoning steps, facilitating a deeper understanding of models' strengths and limitations.

A.2 ERROR TYPE ANALYSIS

Table 5: Model Reasoning Error Analysis

Error Type	**Model**	**Model Reasoning Steps**	**Human Check / Explanation**
Answer Not Found: Repetition Problem	LLaMa-3.2-Vision	Now, let's calculate the present value of the face amount at maturity: $PV = \frac{\$100}{(1+0.03)^7} \approx \$64.91$ ... Now, let's calculate the present value of the face amount at maturity:	The reasoning step repetition results in no final answer.
Wrong Financial Math Formula	LLaVa_NEXT; Qwen2_VL	..... The distance to default: DD $= \frac{\ln\left(\frac{V}{D}\right)+\left(r-\frac{\sigma^2}{2}\right)T}{\sigma\sqrt{T}}$	The distance to default : DD $= \frac{\text{Asset value}-\text{Default Point}}{\text{Asset volatility}}$
Question Misunderstanding	LLaVa_NEXT	This appears to be a task related to logic puzzles, specifically an example of a "river crossing" problem where you need to ...	No "river crossing" problem in this case.
Image Recognition Problem	LLaVa_NEXT; Qwen2_VL	... However, the problem does not provide the values of Q1 and Q3. Without these specific values, ...	This image is inputted; the model should recognize the values instead of assuming their absence.
Image Recognition Problem	GPT-4o	.... based on the graph alone, the spot rate should be understood as 4.0%.	$r(5) = 5\sqrt{\frac{1.0437}{0.8394}} - 1 = 4.453\%$ The model needs to extract data from the image for calculation instead of relying solely on textual information.