FinMR: A Comprehensive Benchmark for Multimodal Financial Reasoning with Insights from Error Feedback Learning

Anonymous ACL submission

Abstract

We introduce **FinMR**, a novel multimodal benchmark designed to evaluate the reasoning capabilities of multimodal LLMs in financial problem-solving. **FinMR** features 3,200 college-level question-answer pairs, 1,049 focused on financial math and 2,151 on financial expertise, integrating textual and visual content, such as stock price trends. To enhance financial reasoning, we employ Error Feedback Learning (**EFL**), which incorporates negative examples and feedback for iterative improvement. Through evaluations of open-source and closed-source models, we demonstrate that MLLMs outperform LLMs and that **EFL** is more effective than CoT prompting. Our error analysis highlights key challenges in image recognition, question understanding, and formula application, providing insights for future research. **FinMR** establishes a robust foundation for advancing financial reasoning capabilities and developing more effective multimodal reasoning techniques. Our code and data can be found at https://anonymous/FinMR/Code&Data. The leaderboard can be found at https://anonymous/FinMR/Leaderboard.

.

1 Introduction

Financial analysis involves leveraging expert knowledge to derive insights from diverse financial data and reach logical decisions. Effective financial analysis can yield substantial monetary gains or prevent billion-dollar losses (Jerven, 2013; MacKenzie, 2008; Chen et al., 2021). However, analyzing financial data is inherently complex, requiring integration of structured data (e.g., tables), semi-structured data (e.g., filings), and unstructured data (e.g., reports). It demands advanced mathematical rigor, including multi-step calculations, statistical analysis, and domain-specific formulas, along with a deep understanding of concepts such as portfolio optimization and risk modeling.

Recent advancements in Large Language Models (LLMs), such as GPT-o1 and DeepSeek-R1, have eased reasoning over financial text (OpenAI, 2024a; DeepSeek-AI et al., 2025). However, many financial tasks also involve visual data like stock trends, tables, and charts, requiring multimodal integration. Multimodal LLMs (MLLMs), capable of processing both text and visuals, have shown promising results on benchmarks like MMMU (Yue et al., 2024) and MathVista (Lu et al., 2024) in some technical reports (OpenAI, 2024a; GeminiTeam et al., 2025; MetaAI, 2025; Anthropic, 2025; Li et al., 2024a). However, existing multimodal benchmarks (see Table 1) mainly assess general reasoning in open domains. The only financial-specific benchmark, FAMMA (Xue et al., 2024), covers just 1,758 examples across 8 topics, lacking coverage of key areas like risk management and valuation. Thus, the financial reasoning capabilities of MLLMs remain largely unexplored. This paper seeks to address the critical question: ***What are the multimodal reasoning capabilities and limitations of MLLMs in the financial domain?***

To answer this, we first introduce **FinMR**, a benchmark specifically designed for evaluating MLLMs on financial reasoning. **FinMR** spans 15 financial topics with 3,200 college-level QA pairs combining text and visual content. The dataset includes diverse visuals such as table images, stock trends, and statistical distributions (see Figure 1(a)). It contains 1,049 financial math questions requiring advanced mathematical skills and 2,151 financial expertise questions needing deep domain knowledge. Each example includes manually annotated explanations to support reasoning evaluation and error analysis. The dataset is split into a 80% development set (2,560 examples) and a 20% test set

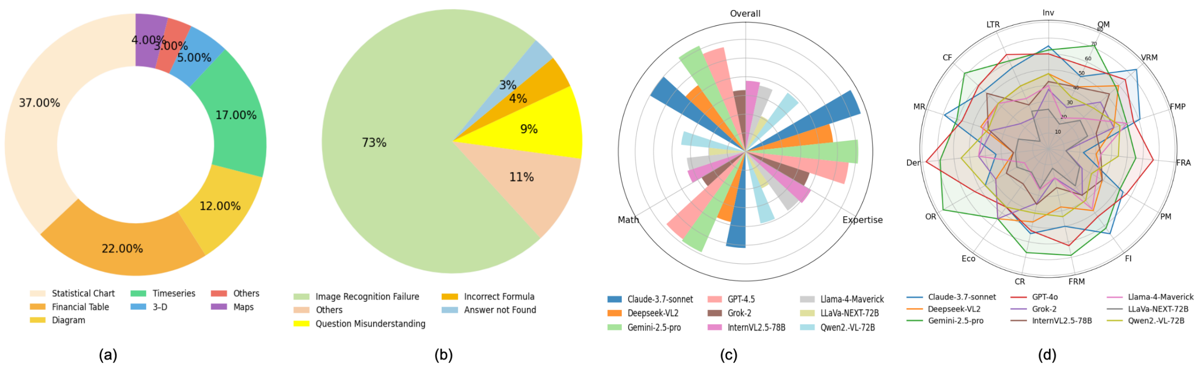


Figure 1: **FinMR** features a wide variety of visual data types, as summarized in panel (a), with representative examples provided in Appendix A. Evaluation of financial reasoning abilities of MLLMs covers mathematical and expertise-based tasks (see panel (b)), and performance varies across 15 financial domain topics (see panel (c), the abbreviation list of topics provided in Table 2). Panel (d) highlights the major categories of errors, including image recognition failures, incorrect formula applications, and misunderstandings of questions.

(640 examples).

Second, we comprehensively evaluate state-of-the-art LLMs and MLLMs on **FinMR**, identifying performance gaps across models like GPT-o1 (OpenAI, 2024b), DeepSeek-R1 (DeepSeek-AI et al., 2025), Gemini-2.5-Pro (GeminiTeam et al., 2025), and Claude-3.7-Sonnet (Anthropic, 2025). We apply two methods: Chain-of-Thought (CoT) prompting (Wei et al., 2022b; Li et al., 2024b) and our proposed *Error Feedback Learning (EFL)*. Inspired by prior studies (Wang et al., 2024b), we construct an *error database* using development data, pairing negative examples with AI-driven feedback. EFL retrieves similar mistakes and feedback to guide model reasoning, improving performance without additional training.

The main contributions are summarized as follows:

- **FinMR**: the first comprehensive multimodal benchmark covering 15 financial topics, comprising 3,200 QA pairs with manually annotated explanations, each integrating text and diverse image types, enabling evaluation of MLLMs' financial reasoning and intermediate reasoning errors.
- **EFL**: an error feedback learning strategy stimulating models' learning ability on **FinMR** by utilizing detailed explanations for error correction. The constructed error database enables retrieval-based self-correction, evidencing the effectiveness of error feedback and the value of expert annotations.
- **Comprehensive experiments**: evaluations across LLMs and MLLMs demonstrate that MLLMs consistently outperform LLMs on FinMR. Gemini-2.5-Pro emerges as the top-performing MLLM on **FinMR**, while Claude-3.7-sonnet gain notable improvement in mathematical reasoning when direct image inputs are utilized.

2 Related Works

2.1 LLMs and MLLMs for Financial Reasoning

Reasoning-specific LLMs, such as GPT-o1 (OpenAI, 2024b), have shown considerable competence in text-only reasoning tasks across various benchmarks. Other notable works, including Deepseek-R1 (DeepSeek-AI et al., 2025), have further advanced the reasoning capabilities of LLMs. Building on these advancements, researchers have extended textual LLMs to incorporate diverse modalities, such as images. Notable examples of this effort include GPT-4o (OpenAI, 2024a), Genimi-2.5-Pro (GeminiTeam et al., 2025), Qwen2.5-VL (Bai et al., 2025), InternLMXComposer-VL (Zhang et al., 2024a) and SPHINX (Lin et al., 2023). These models have demonstrated remarkable visual performance on multimodal reasoning benchmarks, including MMMU (Yue et al., 2024), MathVista (Lu et al., 2024), and Math-V (Wang et al., 2024a). Interestingly, several MLLMs have achieved performance levels that surpass GPT-4o on these benchmarks. We evaluate various state-of-the-art MLLMs' reasoning performance on the **FinMR** benchmark and analyze their limitations.

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., 2021)	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

2.2 Methods for Stimulating Inherent Reasoning Capability

Chain of Thought (CoT). CoT(Wei et al., 2022b) 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. (2024c), 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 employ 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). Particularly, Chia et al. (2023), Sun et al. (2024a), and Zhang et al. (2024b) proposed some fixed mistake examples as demonstration for LLM to improve reasoning capability. 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., 2024c). 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 AI feedback in enhancing model performance on multimodal mathematical reasoning tasks (Yan et al., 2024; Wang et al., 2024b; Lu et al., 2023; Sun et al., 2024b). Building on this foundation, our work employ a EFL strategy to retrieve similar error feedback from an error database. As fixed mistake examples can not solve diverse errors, this approach allows models to iteratively refine their reasoning capabilities by analyzing different 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. This benchmark 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.

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

Table 2: Financial Topic Distribution of FinMR.

Topics & 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	-

3.2 Data Curation Process

Data Collection and Compliance. The data collection process for **FinMR** consisted of two stages. First, we compiled financial exam papers from college-level courses and professional certification programs, focusing on business schools officially collaborating with Chartered Financial Analyst (CFA) and Financial Risk Management (FRM) programs. As these institutions' curricula are often integrated into university courses, their final exams serve as valuable sources for expert-level reasoning tasks. We extracted QA pairs from past final exams, and for PDF-formatted papers, used the Mathpix API (Wang et al., 2024a) to retrieve text, formulas, and images. To maintain consistent formatting, only images from questions were included, excluding those from options and explanations. In the second stage, two Finance PhD students—both CFA and FRM holders—manually verified the explanations and filtered out QA pairs lacking correct answers or high-quality reasoning. This rigorous verification ensured that the dataset comprised QA pairs with expert-validated explanations.

Data Quality Assurance. We employed a three-stage data curation process involving six annotators (four master's students and two PhD students from Computer Science and Finance). First, question texts and images were separately extracted and manually checked for alignment due to occasional inaccuracies from Mathpix, such as misaligned or out-of-order visuals. Images with suboptimal clarity were enhanced using Image Generator Pro. QA pairs missing explanations or written in non-English were removed, resulting in 4,470 refined QA pairs, comprising 30% math-focused and 70% financial expertise tasks. Second, two PhD students reviewed the validity, completeness, and clarity of explanations. Incomplete or overly concise explanations were expanded with detailed reasoning, and grammatical and stylistic issues were corrected, yielding a final dataset of 3,200 high-quality QA pairs. Six examples are provided in Appendix A. Finally, each question was labeled with metadata including ID, source topic, and question type. To facilitate evaluation, the dataset was split by topic: 80% allocated to the development set (2,560 samples) and 20% to the test set (640 samples) (Zhao et al., 2024a). Throughout the process, *we strictly adhered to copyright and licensing regulations*, excluding any data from sources that explicitly prohibit copying or commercial use.

3.3 Comparisons with Existing Benchmarks

We compare **FinMR** with nine reasoning benchmarks: MathVista (Lu et al., 2024), MMMU (Yue et al., 2024), MATH-V (Wang et al., 2024a), FinQA (Chen et al., 2021), 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 detailed comparisons.

Comparison with Multimodal Benchmarks. MathVista (Lu et al., 2024) includes 6,141 examples across seven reasoning types but mainly targets elementary to high school levels. MMMU (Yue et al., 2024) expands to 11,500 college-level examples across 30 subjects, emphasizing multidisciplinary multimodal reasoning. MATH-V (Wang et al., 2024a) offers 2,252 expert-level questions from competition datasets, covering 16 subjects.

Unlike these benchmarks, **FinMR** focuses specifically on financial reasoning, combining domain expertise with mathematical reasoning. It also provides detailed, manually verified explanations averaging 61.43 words. As shown in Table 3, **FinMR** questions are longer (33.97 words average, 327.74 words of context) than MathVista (15.6), MMMU (59.33), and MATH-V (42.3), enabling more advanced and realistic financial reasoning tasks.

Comparison with Financial QA Benchmarks. Existing financial QA datasets mainly target LLMs and cover limited subdomains, often without multimodal inputs. FAMMA (Xue et al., 2024) is an exception but includes only 1,758 QA pairs and omits key areas like risk management. In contrast, **FinMR** spans 15 topics from CFA and FRM exams, offering broader financial coverage for evaluating MLLMs.

Benchmarks such as FinQA (Chen et al., 2021), TAT-QA (Zhu et al., 2021), and MultiHiertt (Zhao et al., 2022) focus on numerical reasoning over real-world financial tables but emphasize simpler calculations. In contrast, **FinMR** includes advanced mathematical reasoning tasks involving calculus and statistics, and covers seven image types including complex financial tables (see Figure 1(a)). In terms of reasoning format, datasets like DocMath-Eval (Zhao et al., 2024b) and FinanceMath (Zhao et al., 2024a) use Python-based solutions, which, while precise, often lack interpretability. **FinMR** addresses this by providing manually annotated textual explanations for all 3,200 QA pairs, offering clearer insights into models' reasoning processes.

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, integral to the Error Feedback Learning (EFL) method. This database enables systematic error analysis and iterative refinement of reasoning capabilities. As shown in Figure 2, the construction involves three stages: *data input*, *feedback generation*, and *storage*.

In the data input stage, we feed context, questions, images, and options from the **FinMR** *development dataset* into evaluated models. For LLMs that cannot process images directly, we use GPT-4o to convert images into textual descriptions; MLLMs directly process the visual data alongside text, following methods like Wang et al. (2024a). In feedback generation, models produce step-by-step reasoning and answers, which are compared against ground truth. For incorrect answers, a feedback prompt (*1. Refine correct reasoning, 2. Compare correct and incorrect reasoning, 3. Summarize hints for future questions*) guides the model in refining its reasoning, supported by manually annotated explanations. The feedback captures reasoning flaws and offers actionable improvements. In the storage stage, we save input examples, refined feedback, and metadata (e.g., question ID, model information) in an external database.

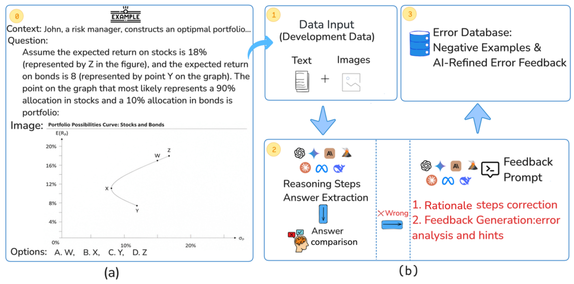


Figure 2: Panel (a) provides an typical example of FinMR. Panel (b) exhibits three stages of error database construction. In the second stage, Large models leverage the annotated explanations to generate correct reasoning steps and hits.

4.2 Evaluation Process

The evaluation process consists of four stages: *test data input, reasoning, output, and evaluation*, as displayed in Figure 3. 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 a simple instruction, such as *"Let's think step by step"* followed by the user input. In EFL, the model retrieves the most similar negative example 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) and Wang et al. (2024a).

In the output stage, we employ 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.

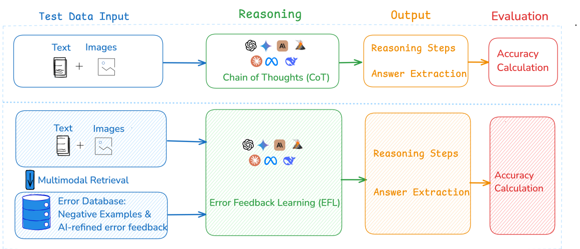


Figure 3: Four stages of the evaluation process. For LLMs with no visual ability, we leverage GPT-4o to generate image captions to support reasoning tasks. The process employs two methods, CoT and EFL. The latter retrieves the most similar (i.e., top-1 semantic similarity) negative examples and error feedback for learning.

[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

5 Results and Analysis

5.1 LLM, MLLM, and Experiment Setup

We evaluate following reasoning-focused LLMs on **FinMR**:

- Closed-source: GPT-o1 (OpenAI, 2024b), Grok-3 (Xai, 2025).
- Open-source: Deepseek-R1 (DeepSeek-AI et al., 2025).

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

- Closed-source: GPT-4o (OpenAI, 2024a), Gemini-2.5-Pro (GeminiTeam et al., 2025), Claude-3.7-Sonnet *extended thinking* (Anthropic, 2025), Grok-2 (Xai, 2024);
- Open-source: Deepseek-VL-2 (Wu et al., 2024), Llama-4-Maverick (MetaAI, 2025), InternVL3-78B (Zhu et al., 2025), Qwen-2.5-VL-72B (Bai et al., 2025), LLaVa-NEXT-72B (Li et al., 2024a).

All experiments on open-source models were conducted using 1 A100 GPU, while experiments on closed-source models were performed using 3 4090 GPUs. Additionally, we used LangSmith to trace all the experiments and set the temperature of large models to 0.7. For the human evaluation, we invited one CFA candidate and one FRM candidate. To simulate the behavior of large models following the EFL strategy, they were permitted to consult their personal error logs accumulated during their exam preparation. The handwritten copies of two human tests are provided in Appendix B.

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). Most notably, all models employing the two methods exhibit substantially lower performance than human experts. Additionally, we summarize and analyze other key findings as follows:

Disparity between Open-source and Closed-source Models: The results on **FinMR** (Table 4, marked in blue) reveal that closed-source MLLMs generally outperform open-source counterparts. For example, Gemini-2.5-Pro achieved 61.93%

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

Models	Open Source?	Method	Overall	Expertise	Math	Inv	QM	VRM	FMP	FRA	PM	FI	FRM	CR	Eco	OR	Der	MR	CF	LTR
Textual Modality: Text + Image Caption																				
Deepseek-R1	✓	CoT	44.84	48.51	42.20	47.50	50.00	46.15	52.17	44.44	41.77	40.17	50.00	46.23	36.36	23.08	11.11	49.17	42.86	43.83
Deepseek-R1	✓	EFL	50.78	55.97	47.04	50.00	78.57	51.92	73.91	55.56	45.57	42.17	62.50	50.00	63.64	38.46	44.44	53.33	50.00	47.48
GPT-o1	✗	CoT	45.78	54.84	33.21	55.00	28.57	53.85	39.13	33.33	34.97	50.60	18.75	45.28	16.18	33.46	60.67	51.67	40.48	30.43
GPT-o1	✗	EFL	47.97	57.30	39.18	60.00	35.71	57.69	52.17	55.56	37.18	54.22	31.25	47.17	18.18	38.46	66.67	55.00	45.24	47.83
Grok-3	✗	CoT	21.88	22.58	20.90	30.00	21.57	19.23	13.04	33.33	11.39	31.33	25.00	18.87	18.18	7.69	44.44	20.83	22.57	15.39
Grok-3	✗	EFL	28.91	29.03	28.73	32.50	28.43	34.62	26.09	44.44	17.72	37.35	31.25	25.47	27.27	46.15	55.56	29.17	28.57	17.39
Multimodality: Text + Image																				
Claude-3.7-sonnet	✗	CoT	53.91	61.83	42.91	52.50	42.86	65.38	52.17	20.22	48.10	60.24	37.50	52.83	18.18	30.77	20.02	65.00	47.62	52.52
Claude-3.7-sonnet	✗	EFL	59.84	65.86	51.49	65.00	50.00	75.00	60.87	22.22	54.43	66.27	50.00	54.72	45.45	46.15	33.33	69.17	54.76	56.17
Deepseek-VL2	✓	CoT	37.81	42.47	31.34	45.00	35.71	46.15	17.39	25.22	34.18	43.37	18.75	38.68	18.18	15.38	22.62	38.33	30.95	30.43
Deepseek-VL2	✓	EFL	43.91	47.85	38.43	47.50	42.86	59.62	43.48	30.33	39.24	48.19	37.50	47.17	50.55	38.46	24.22	45.00	42.86	43.48
Gemini-2.5-pro	✗	CoT	54.76	55.11	54.28	47.50	42.86	51.92	52.17	33.33	46.84	51.19	59.60	59.43	45.45	38.46	57.00	55.00	64.29	47.83
Gemini-2.5-pro	✗	EFL	61.93	61.29	58.83	62.50	71.43	57.69	52.17	55.56	51.90	61.90	68.75	66.98	54.55	76.92	68.89	60.83	71.43	60.87
GPT-4o	✗	CoT	50.00	50.68	47.25	60.00	42.86	44.23	52.17	33.33	53.16	46.99	60.75	50.94	45.45	23.08	55.56	49.17	50.00	62.22
GPT-4o	✗	EFL	57.03	56.37	57.89	60.00	57.14	65.38	56.52	66.67	56.96	53.01	62.50	52.83	45.45	53.85	77.78	57.50	59.52	65.22
Grok-2	✗	CoT	27.81	33.33	20.15	30.00	21.43	40.38	21.74	10.11	21.05	30.12	6.25	27.36	27.27	15.38	24.44	33.50	11.90	4.35
Grok-2	✗	EFL	32.87	36.63	27.61	37.50	28.57	44.23	43.48	11.11	24.52	36.47	18.75	34.91	53.55	38.46	44.44	37.67	23.81	23.78
InternVL3-78B	✓	CoT	31.87	33.06	30.22	32.50	14.29	28.85	20.43	16.22	27.85	33.73	18.75	34.91	18.18	30.77	22.22	36.33	40.38	20.74
InternVL3-78B	✓	EFL	37.81	41.13	33.21	42.50	42.86	51.92	31.74	33.33	39.24	36.14	25.00	35.85	27.27	33.62	28.22	40.00	52.48	30.43
Llama-4-Maverick	✓	CoT	27.34	30.11	23.51	27.50	21.43	23.08	39.13	0.00	25.32	36.14	14.75	23.58	10.09	15.69	11.11	32.50	35.71	21.74
Llama-4-Maverick	✓	EFL	36.09	38.98	32.09	40.00	21.43	28.85	52.17	33.33	35.44	46.99	18.75	26.42	19.09	17.49	44.44	43.33	42.86	34.78
LLaVa-NEXT-72B	✓	CoT	16.72	18.55	14.18	20.00	7.14	19.23	13.09	0.00	18.99	21.69	10.50	16.04	9.09	7.69	10.01	18.13	4.76	4.35
LLaVa-NEXT-72B	✓	EFL	20.78	21.51	19.78	25.00	17.15	26.92	26.04	11.11	22.78	28.92	12.50	25.47	18.18	23.08	21.00	20.00	7.14	22.09
Qwen2.5-VL-72B	✓	CoT	34.84	38.71	28.73	42.50	28.57	33.46	43.83	22.22	30.65	30.96	18.75	33.02	27.27	7.69	22.22	39.17	28.57	26.09
Qwen2.5-VL-72B	✓	EFL	38.91	39.25	35.18	47.50	35.71	38.46	47.48	44.44	31.38	40.12	43.75	41.51	45.45	46.15	55.56	40.00	42.86	43.48
Human Evaluation	-	-	88.50	91.00	86.00	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

overall accuracy and 58.83% math accuracy, while open-source models like LLaVa-NEXT-72B scored below 25%. Open-source models such as LLaVa-NEXT-72B, and LLaMa-4-Maverick, underperformed, highlighting challenges faced by open models in complex multimodal reasoning.

Effectiveness of Error Feedback Learning: Comparisons between CoT prompting and EFL demonstrate the effectiveness of leveraging error feedback to boost performance. Across all models, accuracy improved notably with EFL. For instance, Grok-3 improved from 21.88% to 28.91% after incorporating negative examples and feedback. Similarly, Gemini-2.5-Pro and InternVL3-78B achieved nearly 6% gains. These consistent improvements across diverse models confirm EFL as a robust and generalizable method for enhancing financial reasoning.

Impact of Image Captions and Direct Image Inputs: For LLMs lacking visual capabilities, GPT-4o-generated captions supplemented image understanding. Most textual models achieved over 40% accuracy, except Grok-3, which lagged behind. Among textual models, DeepSeek-R1 (EFL) stood out, reaching 50.78% and outperforming GPT-o1. In contrast, MLLMs processing direct images achieved significantly higher accuracies, with Gemini-2.5-Pro (EFL) leading at 61.93%, slightly ahead of Claude-3.7-Sonnet (EFL). These results underscore the advantage of integrated text-visual reasoning for complex tasks.

Challenge of Financial Math Reasoning: As shown in Table 4, a significant gap exists between mathematical and expertise reasoning. GPT-o1 (CoT) and Claude-3.7-Sonnet (CoT) scored 33.21% and 42.91% in math, compared to 54.84% and 61.83% in expertise tasks. Mathematical reasoning demands higher logical rigor, multi-step calculations, and precise chains, leading to lower performance. Figure 1(b) further shows that large closed-source MLLMs achieve more balanced performance across tasks when using multimodal inputs. Strengthening text-image interaction is crucial for further improving multimodal mathematical reasoning.

Comparison among Different Topics: Table 4 and Figure 1(c) reveal varying model performance across 15 financial topics. Closed-source models consistently outperform across all topics. Multimodal inputs particularly enhance performance in mathematically intensive areas like Quantitative Methods (QM), Derivatives (Der), Credit Risk (CR), and Liquidity and Treasury Risk (LTR), which require precise reasoning and domain expertise. In contrast, gains are less notable in loosely structured topics like Operational Risk and Economics, where contextual understanding is more critical. These findings highlight the need for further strengthening multimodal capabilities in complex reasoning tasks that combine mathematics and domain knowledge.

5.3 Error Type Analysis

To better understand the limitations of the tested models, we conducted an error analysis, categorizing failures into five types: ***image recognition failure, question misunderstanding, incorrect formula application, answer not found***, and ***others***. Figure 1(d) summarizes the prevalence of each error type, and Table 5 provides representative examples. More detailed examples are from Table 6 to Table 10 in Appendix C.

Among them, **image recognition failures** are the most significant, accounting for 73% of total errors. As shown in Table 5, many reasoning steps indicate that the provided images lack sufficient direct information. Moreover, many images require domain-specific expertise to extract implicit information, which current models struggle with, highlighting the need for more sophisticated visual understanding, especially for specialized financial visuals such as charts and tables.

Another major issue is **question misunderstanding**, particularly for specialized financial domain questions at the college level. Models often misinterpret the intent or nuances, leading to incorrect reasoning. This emphasizes the importance of deeper contextual understanding for domain-specific tasks.

Even when questions and images are correctly interpreted, models frequently fail in **applying the correct formulas**, especially in harder questions that require integrating knowledge across topics or applying cross-domain financial formulas. These errors suggest current models lack the logical rigor and multi-step calculation ability needed for complex mathematical reasoning.

The ***answer not found*** error is another recurring problem, especially for models like LLaMa-4-Maverick and LLaVa-NEXT, which often fail to produce a final answer. Our study found that unrestricted output tokens led to a **repetition problem**, resulting in inconsistent, overly lengthy, and incomplete reasoning outputs, as shown in the first example of Table 5. This issue is particularly harmful for tasks requiring long reasoning chains, highlighting the need to manage token limits.

In summary, these errors stem from both technical limitations, such as image recognition failures and output repetition, and a lack of financial domain expertise. Although our analysis offers a foundational understanding of error types, space constraints prevent a full systematic analysis. We believe such an in-depth analysis would provide valuable insights into the reasoning capabilities and shortcomings of large models and should be pursued in future work.

Table 5: Model Reasoning Error Analysis

Error Type	**Model**	**Model Reasoning Steps**	**Human Check / Explanation**
Answer Not Found: Repetition Problem	LLaMa-4-Maverick	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-70B; Qwen2.5-VL-70B	The distance to default: $\text{DD} = \frac{\ln\left(\frac{V}{D}\right)+\left(r-\frac{\sigma^2}{2}\right)T}{\sigma\sqrt{T}}$	The distance to default : $\text{DD} = \frac{\text{Asset Value} - \text{DP}}{\text{AssetVolatility}}$
Question Misunderstanding	LLaVa-NEXT-70B	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-70B; Qwen2.5-VL-70B	... 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.

6 Conclusion and Future Work

This paper introduced **FinMR**, a benchmark for evaluating the financial reasoning capabilities of multimodal models. Evaluations across open- and closed-source LLMs and MLLMs revealed key insights and challenges.

Our findings highlight three main conclusions: **(1)** MLLMs significantly outperform LLMs by integrating textual and visual information, though image recognition remains a major bottleneck. **(2)** Error Feedback Learning (EFL) consistently outperforms Chain of Thought (CoT) prompting, validating the benefit of leveraging negative examples with feedback. **(3)** Financial math reasoning tasks remain more challenging, with about 5% lower accuracy than expertise reasoning, due to errors like incorrect formula application and question misunderstanding.

Future work may focus on enhancing visual reasoning, developing efficient training-free methods, and systematically addressing identified challenges. A deeper analysis of reasoning errors could further reveal the limitations of both LLMs and MLLMs, especially for complex financial tasks.

7 Limitations

Our work has several limitations that should be addressed in future research. First, all questions in **FinMR** are derived from past exam papers at universities where English is the primary language of instruction. This restricts the dataset's applicability to multilingual financial reasoning tasks and limits its ability to evaluate models' performance in other languages. Second, the scope of **FinMR** is narrowly focused on the financial domain, excluding broader business-related topics, like marketing. While this ensures domain specificity, it may reduce the dataset's generalization to other fields where financial reasoning intersects with broader business concepts. Finally, we deliberately excluded questions with multiple answers to simplify the reasoning evaluation. However, such questions are common in real-world financial examinations and collegiate-level assessments, and their inclusion would better reflect the complexities and uncertainties encountered in practical scenarios.

References

Anthropic. 2025. Introducing computer use, a new claude 3.7 sonnet, and claude 3.7 haiku \ anthropic.

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, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024. M^{3}CoT: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. arXiv.

Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Hao Zhang, and Chuang Gan. 2023. See, think, confirm: Interactive prompting between vision and language models for knowledge-based visual reasoning. *Preprint*, arXiv: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. 2021. FinQA: A dataset of numerical reasoning over financial data. pages 3697–3711. Association for Computational Linguistics.

Yew Ken Chia, Guizhen Chen, Luu Anh Tuan, Soujanya Poria, and Lidong Bing. 2023. Contrastive chain-of-thought prompting. *Preprint*, arXiv:2311.09277.

DeepSeek-AI, Daya Guo, and Dejian Yang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

GeminiTeam, Natalie Clay, Tomas Kocisky, Bartek Perz, Dian Yu, and Howard. 2025. Gemini 2.5: Our most intelligent ai model.

Morten Jerven. 2013. 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.

Kaichen Li, Hao Zhang, Dong Zhang, Renrui Guo, Feng Zhang, Yuanhan Li, Ziwei Zhang, Chunyuan Liu, and Bo Li. 2024a. LLaVA-NeXT: Stronger LLMs supercharge multimodal capabilities in the wild.

Suyan Li, Fuxiang Huang, and Lei Zhang. 2024b. A survey of multimodal composite editing and retrieval. *arXiv preprint*.

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, Jiaming Han, Siyuan Huang, Yichi Zhang, Xuming He, Hongsheng Li, and Yu Qiao. 2023. SPHINX: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint*.

Bingshuai Liu, Chenyang Lyu, Zijun Min, Zhanyu Wang, Jinsong Su, and Longyue Wang. 2024. Retrieval-augmented multi-modal chain-of-thoughts reasoning for large language models. *arXiv preprint*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint*.

Pan Lu, Liang Qiu, Wenhao Yu, Sean Welleck, and Kai-Wei Chang. 2023. A survey of deep learning for mathematical reasoning. arXiv.

Donald MacKenzie. 2008. *An engine, not a camera: How financial models shape markets*, volume 48. Mit Press.

MetaAI. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation.

OpenAI. 2024a. Introducing GPT-4o and more tools to ChatGPT free users | OpenAI.

OpenAI. 2024b. Introduction OpenAI o1.

Hao Sun, Yong Jiang, Bo Wang, Yingyan Hou, Yan Zhang, Pengjun Xie, and Fei Huang. 2024a. Retrieved in-context principles from previous mistakes. *Preprint*, arXiv:2407.05682.

Shilin Sun, Wenbin An, Feng Tian, Fang Nan, Qidong Liu, Jun Liu, Nazaraf Shah, and Ping Chen. 2024b. A review of multimodal explainable artificial intelligence: Past, present and future. arXiv.

Cheng Tan, Jingxuan Wei, Linzhuang Sun, Zhangyang Gao, Siyuan Li, Bihui Yu, Ruifeng Guo, and Stan Z. Li. 2024. Retrieval meets reasoning: Even high-school textbook knowledge benefits multimodal reasoning. *arXiv preprint*.

Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. arXiv.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169.

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

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. 2022a. Emergent abilities of large language models. *Preprint*, arXiv:2206.07682.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H Chi, Quoc V Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, and 8 others. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *Preprint*, arXiv:2412.10302.

Xai. 2024. Grok-2 beta release | xai.

Xai. 2025. Grok 3 beta — the age of reasoning agents | xai.

Siqiao Xue, Tingting Chen, Fan Zhou, Qingyang Dai, Zhixuan Chu, and Hongyuan Mei. 2024. Famma: A benchmark for financial domain multilingual multimodal question answering. *Preprint*, arXiv:2410.04526.

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

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2024. VisRAG: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint*.

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, and 3 others. 2024. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. *arXiv preprint*.

Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, Songyang Zhang, Wenwei Zhang, Yining Li, Yang Gao, Peng Sun, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, and 8 others. 2024a. InternLM-XComposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint*.

Tianjun Zhang, Aman Madaan, Luyu Gao, Steven Zheng, Swaroop Mishra, Yiming Yang, Niket Tandon, and Uri Alon. 2024b. In-context principle learning from mistakes. *Preprint*, arXiv:2402.05403.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2024c. Multimodal chain-of-thought reasoning in language models. *arXiv preprint*.

Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Xuan Long Do, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, and Shafiq Joty. 2023a. Retrieving multimodal information for augmented generation: A survey. *arXiv preprint*.

Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. 2022. MultiHiertt: Numerical reasoning over multi hierarchical tabular and textual data. *arXiv preprint*.

Yilun Zhao, Hongjun Liu, Yitao Long, Rui Zhang, Chen Zhao, and Arman Cohan. 2023b. KnowledgeMath: Knowledge-intensive math word problem solving in finance domains. *arXiv preprint*.

Yilun Zhao, Hongjun Liu, Yitao Long, Rui Zhang, Chen Zhao, and Arman Cohan. 2024a. FinanceMath: Knowledge-intensive math reasoning in finance domains. *arXiv preprint*.

Yilun Zhao, Yitao Long, Hongjun Liu, Ryo Kamoi, Linyong Nan, Lyuhao Chen, Yixin Liu, Xiangru Tang, Rui Zhang, and Arman Cohan. 2024b. DocMath-eval: Evaluating math reasoning capabilities of LLMs in understanding long and specialized documents. *arXiv preprint*.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. *arXiv preprint*.

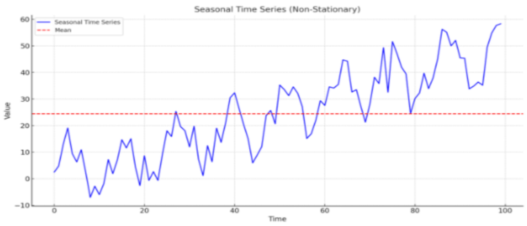
Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Financial Questions in FinMR

Question:

```
In regard to modeling and forecasting seasonality, each of the following
is true EXCEPT which is not accurate?
```

Image:



Options:

A: A seasonal time series is, by definition, covariance stationary,
B: Trading-day variation is a type of seasonality that refers to the fact that different months contain different numbers of trading days,
C: A key technique for modeling seasonality is regression on seasonal dummy variables; dummy variables assume a value of zero or one.
D: Log transformation is useful in both trend models and seasonal models, but for different reasons; in a seasonal model,log transformation can stabilize seasonal patterns whose variance is growing over time.

Question:

```
A portfolio manager wants to invest a small amount of new money that has
recently come into a fund. The fund is benchmarked to an index and, rather
than adding a new holding, the manager is considering increasing the
holdings of one of the four assets described in the following table. The
portfolio manager wants to select the asset that has the lowest marginal
VaR as long as its Treynor ratio is at least 0.1. Assuming the risk free
rate is 2%, which asset should the portfolio manager select?
```

Image:

Asset	Portfolio Weight	Expected Return	Beta to the Index	Beta to the Portfolio
A	1.2%	12%	1.2	0.90
B	0.8%	10%	0.7	0.90
C	0.75	10%	0.6	0.85
D	0.35	8%	0.3	1.10

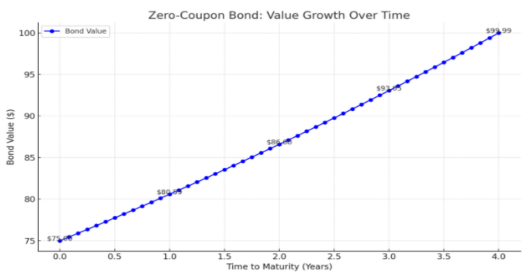
Options:

A. Asset A
B. Asset B
C. Asset C
D. Asset D

Question:

```
The annual yield-to-maturity, stated for with a periodicity of 12, for a
4-year, zero-coupon bond priced at 75 per 100 of par value is closest to:
```

Image:



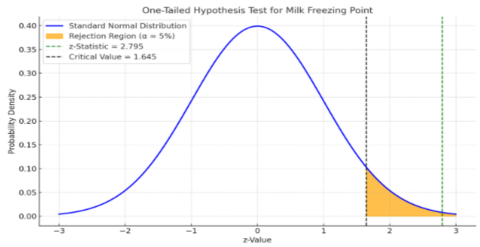
Options:

A. 6.25%
B. 7.21%
C. 7.46%

Question:

```
One company suspects the producers of adulterating milk are not doing
their jobs. By measuring the freezing point of milk, one can detect
whether the milk has been mixed with water. The freezing temperature of
natural milk approximately follows a normal distribution, with a mean
value of -0.545℃ and a standard deviation of 0.008℃. The addition of
water to milk can raise the freezing temperature close to the freezing
temperature of thewater. The company conducts a hypothesis test with a
null hypothesis that the mean value of the freezing temperature of natural
milk is no greater than -0.545℃. At the 5% significance level, the
company measures the freezing temperature of five batches of milk
submitted by the producers and calculates the sample mean of -0.535℃. Can
the producers be considered to deceit the company?
```

Image:



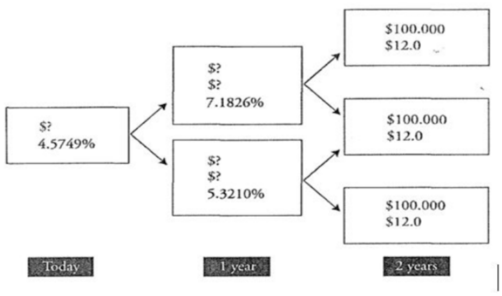
Options:

A: Yes, because the test statistic is larger than 1.96.
B: Yes, because the p-value is smaller than the size of the test.
C: No, because -0.545℃ is not falling into the confidence interval.
D': No, because the test statistic is smaller than 1.96.

Question:

```
The value today of an option-free, 12% annual coupon bond with two years
remaining until maturity is closest to:
```

Image:



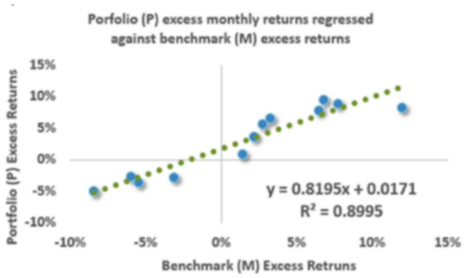
Options:

A: 109.927.
B: 111.485.
C: 112.282.
D: 113.394.

Question:

```
Over the previous twelve (12) months, Analyst Robert regressed Portfolio
(P) excess returns against the Benchmark (M) excess returns:
```

Image:

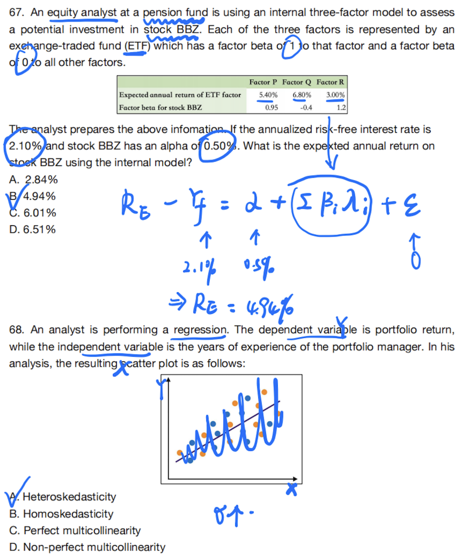


Options:

A. 0.655
B. 0.728
C. 0.833
D. 0.950

B Human Evaluation

We invited two financial experts to participate in the human evaluation. Test Paper 1, shown in Figure 5, contains 100 questions, while Test Paper 2, shown in Figure 6, contains another 100 questions. The handwritten responses illustrate that human reasoning typically involves identifying and circling critical information in the question to aid comprehension. Additionally, experts frequently make brief annotations or calculations in the margins to support quick reasoning and ensure the correctness of the final answer. This behavior highlights the structured and interpretable nature of human problem solving in financial contexts.



67. An equity analyst at a pension fund is using an internal three-factor model to assess a potential investment in stock BBZ. Each of the three factors is represented by an exchange-traded fund (ETF) which has a factor beta of 1 to that factor and a factor beta of 0 to all other factors.

	Factor P	Factor Q	Factor R
Expected annual return of ETF factor	5.40%	6.80%	3.00%
Factor beta for stock BBZ	0.95	-0.4	1.2

The analyst prepares the above infomation. If the annualized risk-free interest rate is 2.10% and stock BBZ has an alpha of 0.50%. What is the expexted annual return on stock BBZ using the internal model?

A. 2.84%
B. 4.94%
C. 6.01%
D. 6.51%

$R_E - r_f = \alpha + \boxed{\Sigma \beta_i \lambda_i} + \varepsilon$

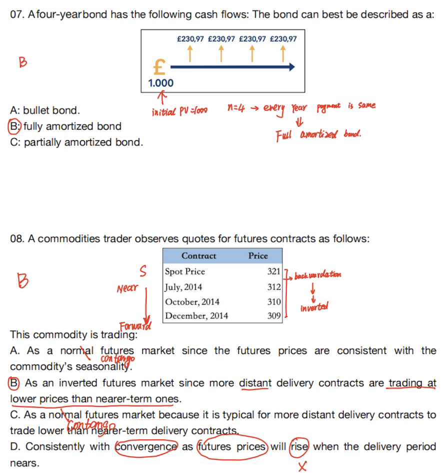
2.1% 0.5% 0

$\Rightarrow R_E = 4.94\%$

68. An analyst is performing a regression. The dependent variable is portfolio return, while the independent variable is the years of experience of the portfolio manager. In his analysis, the resulting scatter plot is as follows:

A. Heteroskedasticity
B. Homoskedasticity
C. Perfect multicollinearity
D. Non-perfect multicollinearity

Figure 5: Handwritten response from Expert A for Test Paper 1



07. Afour-yearbond has the following cash flows: The bond can best be described as a:

B

A: bullet bond.
B: fully amortized bond
C: partially amortized bond.

initial PV =1000 n=4 → every Year payment is same ⇓ Full amortized bond.

08. A commodities trader observes quotes for futures contracts as follows:

Contract	Price
Spot Price	321
July, 2014	312
October, 2014	310
December, 2014	309

B

S Near Forward

backwardation ↓ inverted

This commodity is trading:

A. As a normal futures market since the futures prices are consistent with the commodity's seasonality. contango
B. As an inverted futures market since more distant delivery contracts are trading at lower prices than nearer-term ones.
C. As a normal futures market because it is typical for more distant delivery contracts to trade lower than nearer-term delivery contracts. Contango
D. Consistently with convergence as futures prices will rise when the delivery period nears. X

Figure 6: Handwritten response from Expert B for Test Paper 2

C Error Examples

We provide different error examples from MLLMs, better exhibiting their inner problems in Table 6, 7, 8, 9, 10.

Model	LLaMa-4-Maverick
Error Type	Answer not found: repetition problem
Question	A bond offers an annual coupon rate of 5%, with interest paid semiannually. The bond matures in seven years. At a market discount rate of 3%, the price of this bond per 100 of par value is closest to:
Option	{'A': '106.60.', 'B': '112.54.', 'C': '143.90.'}
Model Reasoning Step	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: $PV = \frac{\$100}{(1+0.03)^7} \approx \$64.91$
Human Check	The reasoning step repetition resulted in no final answer.

Table 6: Error of Repetition Problem

<table>
<tr><td>Model</td><td>LLaVa-NEXT-72B; Qwen2.5-VL-72B</td></tr>
<tr><td>Error Type</td><td>Wrong financial math formula</td></tr>
<tr><td>Context</td><td>Large bank uses the KMV model to measure credit risk exposure. The bank has exposure to Company X. Company X’s firm value, expected firm value, beta, and outstanding debt are shown below.
• Current firm value: 4,000
• Expected firm value: 5,000
• Std. dev: 500
• Company X Beta: 1.4
• Debt 1 Short-term: 1,300
• Debt 2 Long-term: 1,800</td></tr>
<tr><td>Question</td><td>What is the distance to default and the default point for Company X?</td></tr>
<tr><td>Image</td><td><table><tr><td></td><td>Default Point</td><td>Distance to Default</td></tr><tr><td>a.</td><td>2,200</td><td>3.8</td></tr><tr><td>b.</td><td>3,100</td><td>5.6</td></tr><tr><td>c.</td><td>2,200</td><td>5.6</td></tr><tr><td>d.</td><td>3,100</td><td>3.8</td></tr></table></td></tr>
<tr><td>Option</td><td>{’A’: "a", ’B’: "b", ’C’: "c", ’D’: "d"}</td></tr>
<tr><td>Model Reasoning Step</td><td>To solve this problem, we use the KMV model for predicting credit risk.

Step 1: Calculate Default Point
Default Point = Short-term Debt + 0.5 × Long-term Debt
= 1,300 + 0.5 × 1,800 = 2,200
Step 2: Wrong Distance to Default formula used:
$\text{DD} = \frac{\ln\left(\frac{V}{D}\right)+\left(r-\frac{\sigma^2}{2}\right)T}{\sigma\sqrt{T}}$
Where: $V = 4{,}000$, $D = 2{,}200$, $\sigma = 500$, $r = 0.05$, $T = 1$
Wrong term: $0.05 - \frac{500^2}{2} = -124{,}999.95$
DD result = $\frac{0.600-124{,}999.95}{500} = -249.9987$ (illogical)</td></tr>
<tr><td>Human Explanation</td><td>Correct formula:

Default Point = 1,300 + 0.5 × 1,800 = 2,200
Distance to default:
$\text{DD} = \frac{(5{,}000-2{,}200)}{500} = 5.6$</td></tr>
</table>

Table 7: Error of Wrong Math Formula

<table>
<tr><td>Model</td><td>LLaVa-NEXT-72B</td></tr>
<tr><td>Error Type</td><td>Question Misunderstanding</td></tr>
<tr><td>Question</td><td>The exhibit below presents a summary of bilateral mark-to-market (MtM) trades for three counterparties.
If netting agreements exist between all pairs of counterparties shown, what is the correct order of net exposure per counterparty, from highest to lowest?</td></tr>
<tr><td>Image</td><td><table><tr><td colspan="4">MtM Trades for Four Counterparties (USD Million)</td></tr><tr><td></td><td></td><td colspan="2">Opposing Counterparty</td></tr><tr><td></td><td></td><td>B</td><td>C</td></tr><tr><td>Counterparty A</td><td>Trades with positive MtM</td><td>10</td><td>10</td></tr><tr><td></td><td>Trades with negative MtM</td><td>-10</td><td>0</td></tr><tr><td></td><td></td><td>A</td><td>C</td></tr><tr><td>Counterparty B</td><td>Trades with positive MtM</td><td>10</td><td>0</td></tr><tr><td></td><td>Trades with negative MtM</td><td>-10</td><td>-5</td></tr><tr><td></td><td></td><td>A</td><td>B</td></tr><tr><td>Counterparty C</td><td>Trades with positive MtM</td><td>0</td><td>5</td></tr><tr><td></td><td>Trades with negative MtM</td><td>-10</td><td>0</td></tr></table></td></tr>
<tr><td>Option</td><td>{’A’:"A-B-C", ’B’:"A-C-B", ’C’:"C-A-B", ’D’:"C-B-A"}</td></tr>
<tr><td>Model Reasoning Step</td><td>This appears to be a task related to logic puzzles, ✗ specifically an example of a "river crossing" problem where you need to determine which item(s) can be carried across a river without getting them wet or damaged.
The options are labeled as ’A’, ’B’, ’C’, and ’D’.
To solve this type of problem, we typically follow these steps: 1. List all the items that can be directly carried across the river without any issues.
2. Identify which items can be combined to make something that can cross the river.
3. If necessary, rearrange the items or break them down into smaller components to find a solution.
Let’s apply these steps to this problem:
1. Directly carryable items are ’A’,’C’, and ’D’.
2.Combining ’B’ with ’C’ can make ’BC’ , which is not directly carryable across the river.
However, we can break down ’B’ into smaller components if possible.
3.To break down ’B’, we need to find an item that can be combined with ’B’ to form something that can cross the river. In this case, ’A’ can be combined with ’B’ to create ’AB’.
Therefore, ’AB’ is not directly carryable across the river either.
Without further information or clues about the items, it seems like we cannot directly carry any of the options across the river without getting them wet or damaged.
However, if there are additional constraints or rules that are not provided in the image, they might be necessary to solve this puzzle.</td></tr>
<tr><td>Human Explanation</td><td>One must properly net the positive and negative tradesper counterparty for all three counterparties shown.
The properly netted amounts are: For counterparty A: exposure to B = USD 0, exposure to C=USD 10 for a sum of USD 10;
For counterparty B exposure to A=USD 0, exposure to C=USD 0 for a sum of USD 0;
For counterparty C: exposure to A=USD 0, exposure to B=USD 5 for a sum of USD 5;
Therefore, the correct sequence is A-C-B.</td></tr>
</table>

Table 8: Error of Question Misunderstading

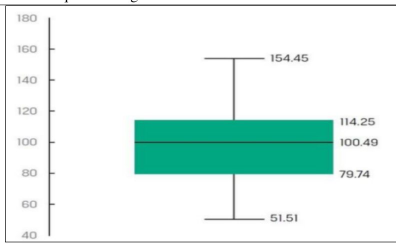
<table>
<tr><td>Model</td><td>LLaVa-NEXT-72B; Qwen2.5-VL-72B</td></tr>
<tr><td>Error Type</td><td>Image Recognition Problem</td></tr>
<tr><td>Question</td><td>The interquartile range is closest to:</td></tr>
<tr><td>Image</td><td>
</td></tr>
<tr><td>Option</td><td>{"A": 13.76, "B": 25.74, "C": 34.51}</td></tr>
<tr><td>Model Reasoning Step</td><td>To determine the interquartile range (IQR), we need to know the first quartile (Q1) and the third quartile (Q3). The IQR is given by:
IQR = Q3 − Q1
ding55 However, the problem does not provide the values of Q1 and Q3. Without these specific values, it’s impossible to directly calculate the IQR.
Given the options:
- A: 13.76
- B: 25.74
- C: 34.51
We need more context or data to determine which of these values is the correct IQR. Since the necessary data (Q1 and Q3) is not provided, we cannot definitively choose an option based on the information given.
Since the problem lacks the required data, we cannot make a selection.
However, if we hypothetically had the values, we would follow the steps:
1. Identify Q1 and Q3.
2. Calculate IQR using IQR = Q3 − Q1.
3. Compare the calculated IQR with the given options.
Without the actual data, we cannot complete the calculation. Therefore, the correct option cannot be determined from the given information.
Therefore, the correct option is Cannot be determined.</td></tr>
<tr><td>Human Check</td><td>This image is inputted.</td></tr>
</table>

Table 9: Error 1 of Image Recognition

<table>
<tr><td>Model</td><td>GPT-4o</td></tr>
<tr><td>Error Type</td><td>Image Recognition Problem</td></tr>
<tr><td>Question</td><td>Based on Exhibit 1, the five-year spot rate is closest to:</td></tr>
<tr><td>Image</td><td>
</td></tr>
<tr><td>Option</td><td>{"A": 4.40%, "B": 4.45%, "C": 4.50%}</td></tr>
<tr><td>Model Reasoning Step</td><td>To find the five-year spot rate from Exhibit 1, we need to look at the graph labeled "Par Rate and Spot Rate by Maturity."
1. Identify the Five-Year Spot Rate:
- The graph shows the spot rate as a green dotted line.
- For the five-year maturity, the spot rate is 4.0%.
2. Compare with Options:
- The given options are:
- A: 4.40%
- B: 4.45%
- C: 4.50%
- The closest option to 4.0% is not directly listed, indicating a potential misalignment in options or a need for assumptions.
3. Conclusion:
- Given the options provided and the graph, there’s a discrepancy. However, based on the graph alone, the spot rate should be understood as 4.0%.
No option directly matches from the given choices, but based on the graph, the spot rate aligns more closely with the understanding of 4.0% from the graph.
Therefore, the correct option should reflect the closest understanding, even if not listed. Since 4.0% is the spot rate, none of the options directly align.
Therefore, the correct option is [A] based on the graph.</td></tr>
<tr><td>Human Check</td><td>The five-year spot rate is determined by using forward substitution and the known values of the one-year, two-year, three-year, and four-year spot rates as follows:
$1 = \frac{0.0437}{(1.025)} + \frac{0.0437}{(1.030)^2} + \frac{0.0437}{(1.035)^3} + \frac{0.0437}{(1.040)^4} + \frac{1+0.0437}{[1+r(5)]^5}$
$r(5) = 5\sqrt{\frac{1.0437}{0.8394}} - 1 = 4.453\%$
We need to use the information from the image to do calculation instead of capturing the image description.</td></tr>
</table>

Table 10: Error 2 of Image Recognition