FinMathVL: Evaluating the Financial Mathematical Reasoning of Large Vision-Language Models

Feng Tao Zhaozhao Ma Lirong Gao
Chenglin Li Yin Zhang Yicheng Li

School of Computer Science and Technology, Zhejiang University

Working Paper

Abstract

Mathematical reasoning ability is crucial for large vision language models (LVLMs) as it enables LVLMs to understand numerical information and perform reasoning. Although multiple mathematical benchmarks exist for LVLMs, they often lack integration with specific scenarios, making it challenging to evaluate model performance in real-world applications. Incorporating mathematical calculations into specific scenarios not only evaluates LVLMs' capability in mathematical reasoning but also examines their understanding of domain specific knowledge. To evaluate the mathematical reasoning abilities of LVLMs in the financial domain, we propose FinMathVL, a multimodal mathematical reasoning benchmark that covers various financial topics. We select data from current mainstream financial examinations and classify it based on factors such as the mathematical skills involved, reasoning difficulty, and the complexity of financial concepts. Additionally, to ensure data professionalism and accuracy, financial domain experts further refine and annotate it. Experimental results show that a few closed-source models achieve up to 80% accuracy on this benchmark, while most open-source models perform poorly, with accuracy generally below 40%, except for Qwen2-VL-72B. Further analysis reveals that as the difficulty of mathematical skills increases, the number of reasoning steps grows, and the complexity of financial concepts rises, the overall performance of the models tends to decline. In addition, the experiments also show that introducing knowledge enhancement mechanisms effectively improves the reasoning ability and accuracy of LVLMs in financial mathematical reasoning tasks.

Overall Framework

The overall construction pipeline of FinMathVL is summarized in Figure 1.

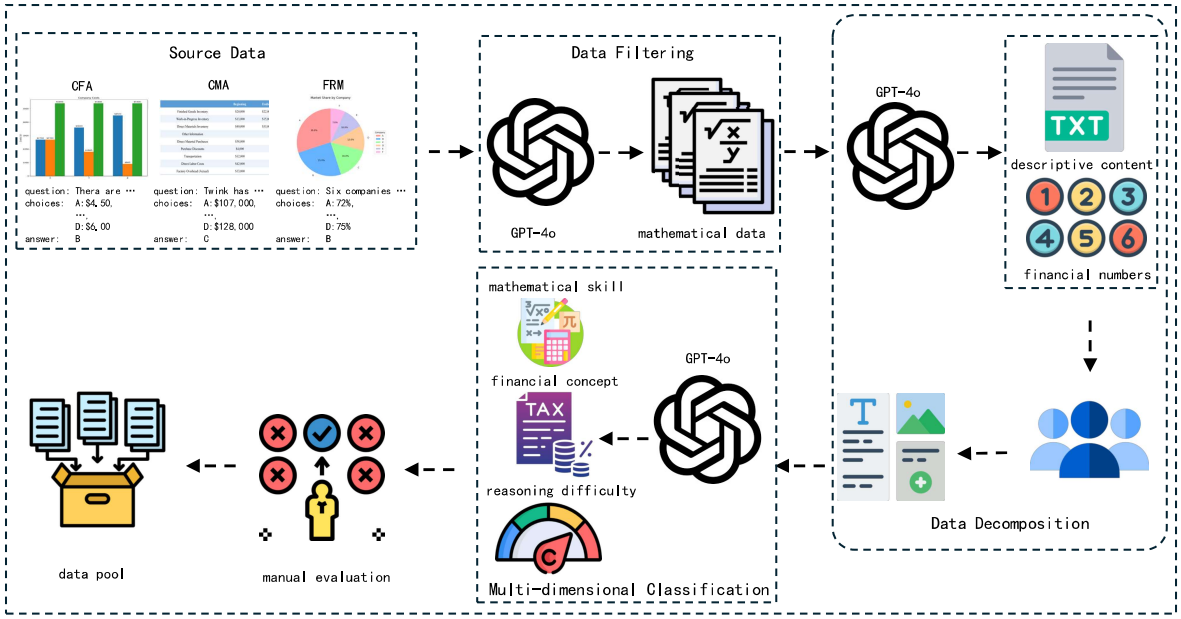

Figure 1: Overall Framework of FinMathVL.

Starting from authoritative financial certification exams (such as CFA, FRM), we first collect source questions and associated materials, and then perform automated data filtering to identify items that genuinely require mathematical reasoning in realistic financial contexts. Next, we decompose each retained question into structured textual and visual components, followed by a multi-dimensional classification according to mathematical skill requirements, financial concept complexity, and reasoning difficulty. Finally, all converted items undergo a rigorous round of manual review to verify correctness, clarity, and label quality before being admitted into the final data pool. A representative data instance in FinMathVL is provided in Figure 2.

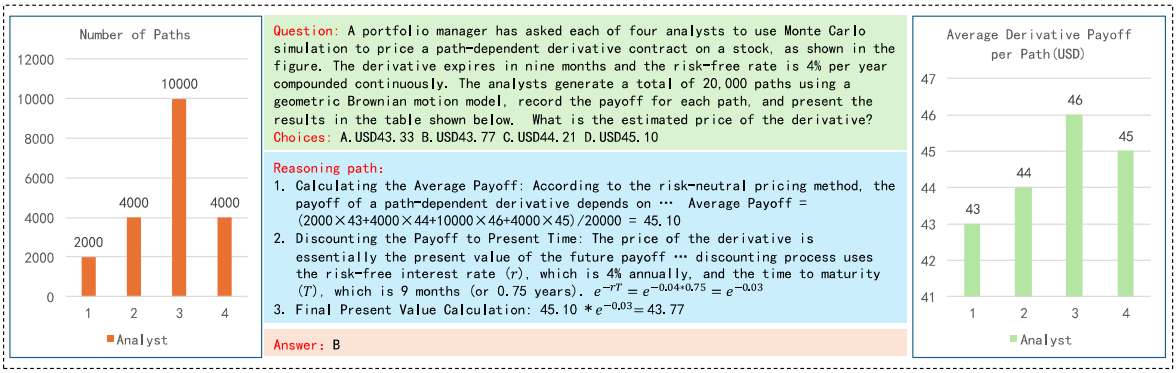

Figure 2: An exmaple data instance in FinMathVL. Each data sample in FinMathVL comprises five parts: the question, choices, images, reasoning path, and answer. LVLMs need to have a precise understanding of financial concepts and accurately extract relevant numerical information from images to construct a logical reasoning chain and derive the final answer.

Selected Experimental Results

This paper proposes FinMathVL, a benchmark for evaluating the financial mathematical reasoning ability of LVLMs. We classify the benchmark along multiple dimensions, including mathematical skills, financial concept complexity, and reasoning difficulty, aiming to provide a comprehensive and systematic evaluation framework for LVLMs. Experimental results show that current models perform poorly on advanced mathematical skills, complex financial concept understanding, and high-difficulty reasoning tasks, indicating that there remains substantial room for improvement in these areas. On the other hand, knowledge enhancement (Table 1) effectively improves the models' understanding of financial concepts and strengthens their financial mathematical reasoning ability.

	Closed-source LVLMs						
	GPT-4o	GPT-4o-mini	Claude-3-Opus	Claude-3.5-Sonnet	Phi3-Vision-4B	Phi3.5-Vision-4B	Llama3.2-11B
Base	79.28%	64.85%	60.46%	78.24%	33.26%	36.19%	32.63%
$\text{Base}_{\text{know}}$	81.23%	69.32%	62.23%	77.12%	40.23%	45.13%	35.12%
	Open-source LVLMs						
	LLaVA-1.6-7B	LLaVA-1.6-13B	InternVL-2.5-4B	InternVL-2.5-8B	Qwen2-VL-3B	Qwen2-VL-7B	Qwen2-VL-72B
Base	30.54%	30.54%	41.84%	37.65%	33.03%	34.93%	60.66%
$\text{Base}_{\text{know}}$	35.23%	37.87%	44.56%	39.23%	38.45%	39.13%	66.78%

Table 1: Performance comparison between baseline and knowledge augmentation on different LVLMs.