FinMMDocR: Benchmarking Financial Multimodal Reasoning with Scenario Awareness, Document Understanding, and Multi-Step Computation

Zichen Tang[1], Haihong E[1]*, Rongjin Li[1], Jiacheng Liu[1], Linwei Jia[1], Zhuodi Hao[1], Zhongjun Yang[1], Yuanze Li[1], Haolin Tian[1], Xinyi Hu[1], Peizhi Zhao[1], Yuan Liu[1], Zhengyu Wang[1], Xianghe Wang[1], Yiling Huang[1], Xueyuan Lin[2], Ruofei Bai[1], Zijian Xie[1], Qian Huang[1], Ruining Cao[1], Haocheng Gao[1]

[1]Beijing University of Posts and Telecommunications
[2]Hithink RoyalFlush Information Network Co., Ltd.

Abstract

We introduce **FinMMDocR**, a novel bilingual multimodal benchmark for evaluating multimodal large language models (MLLMs) on real-world financial numerical reasoning. Compared to existing benchmarks, our work delivers three major advancements. (1) **Scenario Awareness**: 57.9% of 1,200 expert-annotated problems incorporate 12 types of implicit financial scenarios (e.g., Portfolio Management), challenging models to perform expert-level reasoning based on assumptions; (2) **Document Understanding**: 837 Chinese/English documents spanning 9 types (e.g., Company Research) average 50.8 pages with rich visual elements, significantly surpassing existing benchmarks in both breadth and depth of financial documents; (3) **Multi-Step Computation**: Problems demand 11-step reasoning on average (5.3 extraction + 5.7 calculation steps), with 65.0% requiring cross-page evidence (2.4 pages average). The best-performing MLLM achieves only 58.0% accuracy, and different retrieval-augmented generation (RAG) methods show significant performance variations on this task. We expect FinMMDocR to drive improvements in MLLMs and reasoning-enhanced methods on complex multimodal reasoning tasks in real-world scenarios.

Project Resources —
https://bupt-reasoning-lab.github.io/FinMMDocR

1 Introduction

Recently, multimodal large language models (MLLMs) (Liu et al. 2023; Bai et al. 2025) have advanced multimodal reasoning, excelling in visual commonsense reasoning (Zellers et al. 2019; Yu et al. 2024) and visual question answering (Goyal et al. 2017; Singh et al. 2019) end-to-end. Large multimodal reasoning models (LMRMs) (OpenAI 2025), enhanced via reinforcement learning, show promise for complex real-world tasks. They demonstrate superior visual understanding and expert-level reasoning capabilities in domain-specific tasks, operating human-like (Li et al. 2025).

Despite LMRMs' success, current domain-specific reasoning benchmarks remain confined to STEM disciplines (Lu et al. 2024; Wang et al. 2024), often using abstract exam-style questions. They inadequately model the real-world tasks that experts routinely handle. As shown in Figure 1, financial analysts must integrate contextual knowledge to formulate necessary assumptions, then process visually dense financial documents to extract key information. This is followed by comprehensive analytical reasoning, often involving precise multi-step computations, to support high-stakes decision-making. Table 1 shows existing financial QA and document QA benchmarks' key limitations compared to such complex multimodal reasoning scenarios:

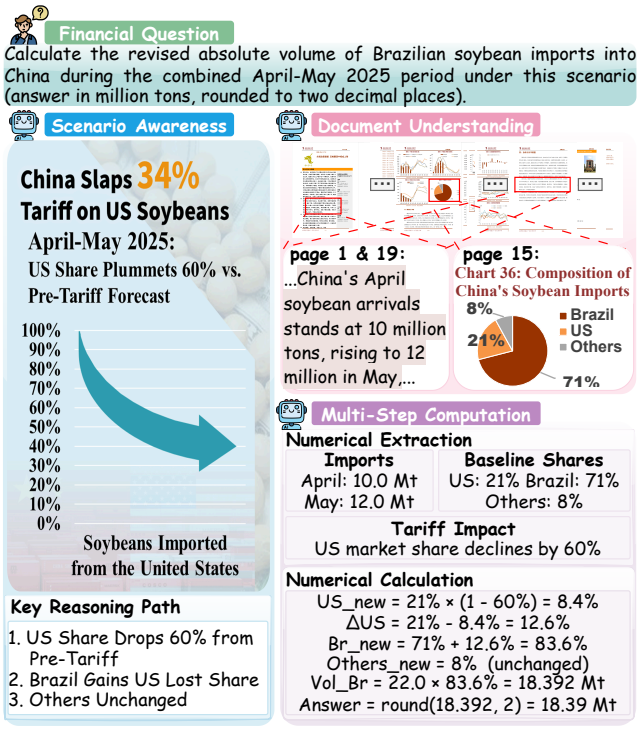


Figure 1: An example of FinMMDocR, integrating a real-world scenario with visually-rich document and multi-step numerical reasoning on China's soybean import shifts between Brazil and the United States during tariff conflicts.

*Corresponding author.

Benchmark	Modalities	Real-World Scenario		Visually-Rich Document			Multi-Step Computation			
		Explicit (%)	Implicit (%)	# Docs	# Pages	# Tokens (k)	Num. Rea. (%)	# Ext.	# Cal.	Cross-Page (%)
Financial QA										
CodeTAT-QA	T	✘	✘	✘	✘	✘	100	2.1	1.0	✘
FinanceMath	T	47.5	39.0	✘	✘	✘	100	3.3	2.5	✘
FinanceReasoning	T	39.1	22.1	✘	✘	✘	100	2.9	2.2	✘
MME-Finance	T+I	✘	✘	✘	✘	✘	15	2.2	1.1	✘
FinMMR	T+I	✘	✘	✘	✘	✘	100	2.6	1.8	✘
DocMath-Eval$_{\text{CompLong}}$	T+TD	15.5	15.1	1,500	61.0	46.5	100	3.0	2.0	52.7
Document QA										
SlideVQA	T+MD	✘	✘	2,619	20.0	2.0	35	$\leq$3	$\leq$3	13.9
MMLongBench-Doc	T+MD	✘	✘	135	47.5	21.2	6	$\leq$3	$\leq$3	33.7
LongDocURL	T+MD	✘	✘	396	85.6	43.6	8	2.6	0.8	52.9
FinMMDocR (ours)	T+MD	**33.7**	**57.9**	**837**	**50.8**	**38.8**	**100**	**5.3**	**5.7**	**65.0**

Table 1: Comparison of FinMMDocR and related benchmarks. **T**: text; **I**: images; **TD**: text document; **MD**: multimodal document; **Explicit**: scenarios with directly given conditions; **Implicit**: scenarios requiring inferred assumptions; **Pages**: pages/doc; **Tokens**: tokens/doc; **Num. Rea.**: numerical reasoning questions; **Ext.**: average extraction steps; **Cal.**: average calculation steps.

- **Absence of Real-World Financial Scenario** *Financial analysts must analyze real-time financial environments to make professional judgments and plausible assumptions.* However, traditional benchmarks (Krumdick et al. 2024; Gan et al. 2025; Tanaka et al. 2023; Ma et al. 2024; Deng et al. 2025) only extract explicitly stated information.
- **Deficiency in Multimodal Document Understanding** *Financial analysts rely on extensive professional documents to extract key information and diverse indicators.* Some benchmarks (Krumdick et al. 2024; Zhao et al. 2024a; Tang et al. 2025b) use text-only inputs, while multimodal ones (Luo et al. 2025; Gan et al. 2025) contain sparse isolated charts or tables. Long-document benchmarks (Ma et al. 2024; Deng et al. 2025) lack diverse financial documents and numerical reasoning tasks.
- **Neglect of Precise Multi-Step Computation** *Financial decision-making, unlike qualitative analysis, requires exact multi-step computations.* In this high-stakes domain (Krumdick et al. 2024), models must deliver numerically exact answers under strict criteria. Prior benchmarks (Zhao et al. 2024a; Krumdick et al. 2024) ignore units, percentages, and decimals or allow 1.0% error margins, diverging from real-world needs.

To fill this gap, we construct FinMMDocR, a more challenging and realistic financial multimodal reasoning benchmark featuring contextual awareness, document understanding, and multi-step computation. FinMMDocR consists of 1,200 numerical reasoning questions (1:1 Chinese-English), equipped with real-world scenarios, visually-rich financial documents, detailed evidence page annotations, golden Python solutions for problem-solving, and exact answers.

- **Scenario Awareness** 57.9% of questions incorporate carefully designed implicit financial scenarios from 12 categories (e.g., Portfolio Management), with an average of 1.9 scenarios per question, significantly surpassing existing benchmarks in density, richness, and complexity.
- **Document Understanding** FinMMDocR contains 837 financial long-documents covering 9 bilingual (Chinese/English) categories (e.g., Financial Engineering, Futures & Options). These documents feature high information density (50.8 pages/doc and 38.8k tokens/doc) and professional visual elements (e.g., candlestick charts).
- **Multi-Step Computation** FinMMDocR averages 11 reasoning steps (5.3 extraction, 5.7 calculation), surpassing other financial reasoning tasks. It enforces strict evaluation (units, percentages, decimals) with 0.2% error tolerance, matching real-world needs. 65.0% of questions require cross-page reasoning (2.4 evidence pages each).

We evaluate 11 proprietary and open-source MLLMs with image inputs using Program-of-Thought (PoT) (Chen et al. 2023), along with 15 LLMs with text inputs using OCR. Beyond end-to-end reasoning, we also evaluate 6 embedding models and 5 agentic retrieval-augmented generation (Agentic RAG) frameworks (Singh et al. 2025). The experimental results reveal three key findings:

- **MLLMs Are Not Qualified Financial Experts for Multimodal Numerical Reasoning.** No model exceeds 60.0% accuracy (OpenAI o4-mini-high: 58.0%), with open-source models particularly struggling, while reasoning-enhanced models show consistent advantages.
- **The More Complex the Task, the Worse Models Perform.** Multimodal models show accuracy degradation in multi-scenario tasks and document understanding failures (78.0% of errors), with extraction errors being the main bottleneck in PoT settings.
- **Vision Is Stronger Than Text, But Complex Agents Underperform Simple RAG.** Vision RAGs surpass text-only methods by utilizing critical document visual cues, yet longer pipelines introduce error propagation that degrades performance, while iterative Agentic RAGs suffer from prohibitive latency without corresponding accuracy improvements for practical deployment.

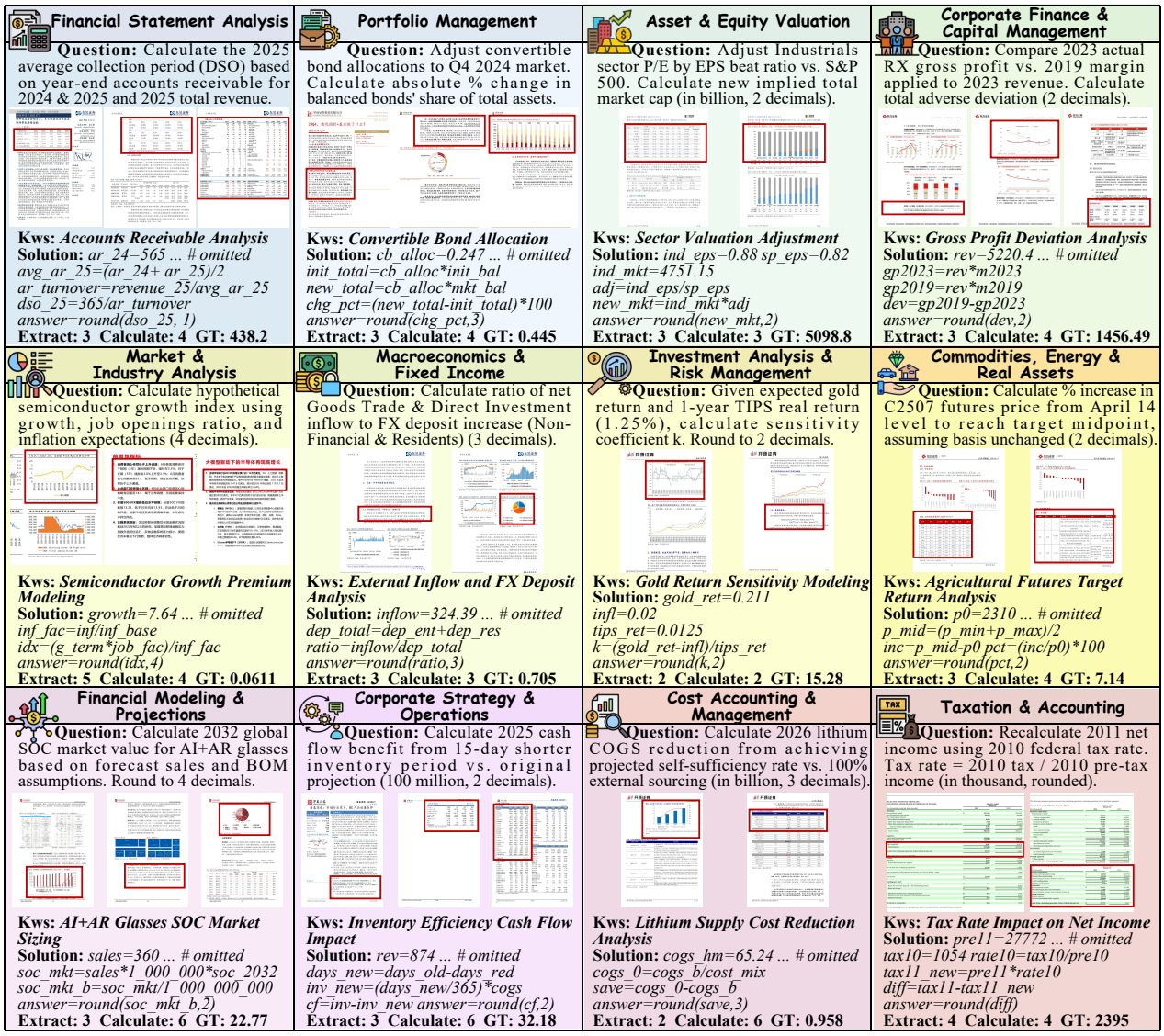


Figure 2: 12 financial scenarios with FinMMDocR examples, covering 9 document categories and cross-page computations. Requires expert *scenario awareness*, *document understanding*, and *multi-step computation*. **Kws**: keywords, **GT**: ground truth.

2 Benchmark Construction

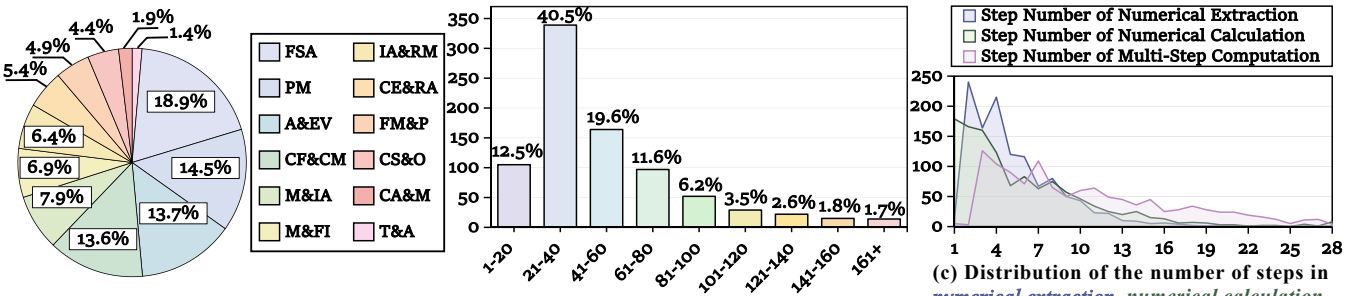
2.1 Overview of FinMMDocR

We introduce FinMMDocR, designed to evaluate the capability of MLLMs to perform complex numerical reasoning when presented with real-world financial scenarios and visually-rich financial documents. Following (Zhao et al. 2024b), each question is accompanied by a Python solution, a standard answer, and page numbers that indicate the locations of relevant visual elements. More examples are shown in Appendix A.

2.2 Data Curation Process

Updates to Public Dataset We selected and re-annotated 600 English questions from the DocMath-Eval$_{\text{CompLong}}$ (Zhao et al. 2024b), comprising all 300 samples from the *testmini* subset and an additional 300 samples chosen from the *test* subset based on diversity and complexity. For the latter, we manually completed previously unreleased solution programs, standard answers, and evidence pages. We retrieved the corresponding documents for all selected examples, rendered each page as an image, and removed original textual inputs to ensure a real multimodal reasoning setting.

Building a Novel Dataset from Scratch We additionally created 600 entirely new Chinese questions. Specifically, we collected 385 Chinese research reports, acquired through authorized channels, covering diverse financial topics (e.g., Company Research, Industry Research). We manually constructed realistic financial scenarios based on document contents (e.g., Financial Modeling & Projections), and further generated knowledge-intensive problems involving complex



(a) Distribution of financial scenarios. (b) Distribution of document lengths. (c) Distribution of the number of steps in *numerical extraction*, *numerical calculation*, and *multi-step computation*.

Note: FSA: Financial Statement Analysis; PM: Portfolio Management; A&EV: Asset & Equity Valuation; CF&CM: Corporate Finance & Capital Management; M&IA: Market & Industry Analysis; M&FI: Macroeconomics & Fixed Income; IA&RM: Investment Analysis & Risk Management; CE&RA: Commodities, Energy & Real Assets; FM&P: Financial Modeling & Projections; CS&O: Corporate Strategy & Operations; CA&M: Cost Accounting & Management; T&A: Taxation & Accounting

Figure 3: Distribution of FinMMDocR: financial scenarios, document lengths, and reasoning steps per question.

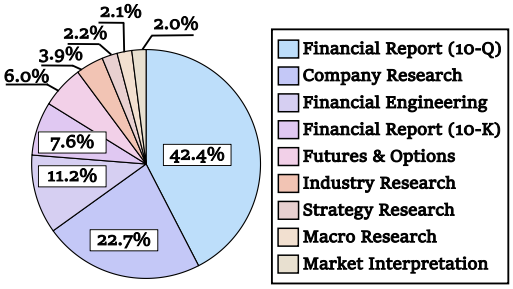


Figure 4: Distribution of FinMMDocR: financial document categories.

Property	Value
# Total Samples	1,200
# Total Document	837
# Financial Scenario (Avg.)	1.9
# Evidence Page (Avg.)	2.4
# Textual Extraction Step (Avg.)	1.0
# Visual Extraction Step (Avg.)	4.3
# Extraction Step (Textual and Visual) (Avg.)	5.3
# Calculation Step (Avg.)	5.7
# Computation Step (Ext. and Cal.) (Avg.)	11.0

Table 2: Basic statistics of FinMMDocR.

numerical reasoning along with corresponding Python solutions, with the assistance of two advanced MLLMs (DeepMind 2025; Anthropic 2025). Documents included in FinMMDocR are exceptionally long, and problems require extracting information dispersed across various sections and modalities (e.g., text, tables, and charts).

Data Quality Assurance Our annotation team comprised 15 master's students majoring in finance and two CFA-certified experts. We implemented a rigorous annotation process to ensure benchmark quality. Specifically, we first fed each sample along with its multimodal document into Gemini 2.5 Pro Preview (DeepMind 2025) and Claude 3.7 Sonnet (Anthropic 2025), the highest-performing MLLMs, to obtain two candidate annotations. Since the model's initial outputs contained numerous logical errors, calculation mistakes, and hallucinations, two annotators cross-reviewed the candidate annotations, selected one for adoption, and subsequently refined it. In cases of disagreement, an additional expert was brought in for arbitration. The selected results underwent further verification and annotation by two annotators. From the initially generated 759 samples, 159 were discarded. Of the remaining 600 samples, 494 underwent modifications: 451 required evidence revision, 80 needed solution adjustment, and 36 had question reformulation. Details are provided in Appendix C.

3 Benchmark Analysis

Table 2 shows FinMMDocR contains 1,200 samples evaluating MLLMs' capabilities across three key dimensions.

Scenario Awareness *FinMMDocR introduces financial reasoning problems with unprecedented scenario density and depth.* 66.2% of problems are scenario-driven across 12 categories (Figure 3(a)). Additionally, all problems feature 1.9 mixed scenarios on average, with 57.9% requiring implicit scenario assumptions rather than given conditions.

Document Understanding *Tasks in FinMMDocR require synthesizing information from multimodal domain-specific documents.* As shown in Figure 3(b) and Figure 4, 837 bilingual (Chinese/English) documents cover 9 categories, averaging 50.8 pages each with 2.4 evidence pages per task, and contain professional charts demanding domain expertise.

Multi-Step Computation *FinMMDocR provides complex financial reasoning tasks requiring cross-page, multimodal, and multi-step reasoning.* As shown in Figure 3(c), each problem requires 11 sequential reasoning steps on average: 5.3 for multimodal numerical extraction (1.0 textual, 4.3 visual) and 5.7 for financial calculation synthesis.

Compared to prior financial QA and document QA benchmarks, FinMMDocR eliminates explicit conditions, limited modalities/types, and excessive focus on information extraction/logical reasoning, better evaluating MLLMs' complex numerical reasoning capabilities in real-world settings.

Model	Size	ACC	Input *Cfg.*	Scenario		*Doc. Len.*		Extract		Compute	
				w/	w/o	≤30	≥31	≤4	≥5	≤4	≥5
MLLM (Image Input)											
Proprietary MLLMs											
OpenAI o4-mini-high		**58.00**	300@F	**55.72**	**62.34**	**57.02**	**58.95**	**63.92**	**51.50**	**63.36**	**52.05**
Doubao-1.5-thinking-pro		38.17	U@F	39.50	35.41	43.99	32.51	40.35	35.93	39.15	37.25
Claude 3.7 Sonnet (Thinking)		37.00	50@1920	35.60	39.40	41.96	32.18	40.66	32.92	39.31	34.40
Doubao-1.5-vision-pro		29.25	U@F	28.81	30.17	32.99	25.62	32.91	25.13	31.92	26.20
Gemini 2.5 Pro Preview		27.42	300@F	27.92	26.43	26.40	28.41	32.91	21.24	31.45	22.82
GPT-4o		17.17	50@1920	12.20	27.18	13.54	20.69	26.42	6.90	25.79	7.49
Grok 2 Vision		2.17	15@1920	2.64	1.25	1.18	3.12	3.16	1.06	3.14	1.07
Open-source MLLMs											
Qwen2.5-VL 72B	72B	12.92	50@F	10.57	17.71	14.04	11.82	18.35	6.90	18.24	6.95
Llama 4 Maverick	400A17B	2.67	300@F	3.65	0.75	1.86	3.45	3.96	1.24	4.09	1.07
Mistral Small 3.1	24B	1.08	15@3840	1.51	0.25	0.51	1.64	1.58	0.53	1.42	0.71
Gemma 3 27B	27B	0.67	15@3840	1.01	0.00	0.17	1.15	0.95	0.35	0.94	0.36
OCR + LLM (Text Input)											
Proprietary LLMs											
Gemini 2.5 Pro Preview		**53.83**	N	**55.22**	**51.12**	**56.01**	**51.72**	**56.80**	**50.62**	**54.09**	**53.65**
Claude 3.7 Sonnet (Thinking)		48.58	N	48.68	48.38	50.42	46.80	51.90	44.96	49.69	47.42
OpenAI o4-mini-high		47.92	200k	50.94	41.90	51.27	44.66	49.53	46.19	47.64	48.31
Doubao-1.5-thinking-pro		42.67	96k	43.52	40.90	44.33	41.05	46.99	37.88	44.65	40.46
Grok 3		41.00	128k	40.13	42.64	41.29	40.72	44.62	36.99	43.87	37.79
Doubao-1.5-vision-pro		32.75	128k	31.70	34.66	30.46	34.98	39.40	25.49	38.36	26.56
GPT-4o		22.17	128k	19.25	28.18	20.14	24.14	28.96	14.69	28.93	14.62
Open-source LLMs											
DeepSeek-R1	671A37B	40.00	64k	41.51	37.16	42.13	37.93	44.46	35.22	42.61	37.25
DeepSeek-V3	671A37B	32.67	128k	30.57	36.66	30.46	34.81	40.03	24.42	39.47	24.96
Llama 4 Maverick	400A17B	29.08	N	27.30	32.42	29.61	28.57	33.23	24.42	32.55	25.13
Qwen3	235A22B	25.08	128k	21.26	32.67	22.00	28.08	34.18	15.04	33.33	15.86
Mistral Small 3.1	24B	15.83	128k	12.45	22.44	14.72	16.91	21.68	9.38	22.33	8.56
Qwen2.5-VL 72B	72B	15.00	128k	12.96	18.95	16.75	13.30	19.62	9.91	19.81	9.63
Llama 3.3 70B	70B	12.17	128k	9.43	17.71	9.14	15.11	18.51	5.13	19.18	4.28
Gemma 3 27B	27B	5.75	128k	5.41	6.48	4.91	6.57	8.39	2.83	8.65	2.50

Table 3: Model performance across input configurations. **Size**: for MoE models, total params and total activated are divided by "A"; **ACC**: accuracy; **Input *Cfg.***: **U@F** = unmerged at full resolution, **X@Y** = merge **X** images (e.g., 300), **Y** = long edge pixels (e.g., 1920), **N** = No cut-off; **Scenario**: **w/** = with contextual scenarios, **w/o** = without; ***Doc. Len.***: document length.

4 Experiments

4.1 Experiments Setting

Models Following (Ma et al. 2024; Deng et al. 2025), we assessed the comprehension capabilities of MLLMs by feeding images directly into models and inputting text extracted by Tesseract OCR engine (Smith 2007). We evaluated 26 different configurations (11 for image input, 15 for text input) on both proprietary and open-source models.

Input Paradigm We designed various configurations to accommodate differences across MLLMs. We tested merging 300, 50, or 15 pages into a single input, alongside an unmerged strategy, while each setting was further tested under three resolution levels (i.e., full resolution, long side 3840/1920 pixels). A fallback strategy that prioritizes preserving page count was applied when models fail to respond in most cases. For text input, we set multiple cut-off lengths to ensure compatibility. Details are provided in Appendix D.

Evaluation Methods We adopt PoT prompts (Chen et al. 2023), which mitigate numerical errors (Zhao et al. 2024a,b), and assess accuracy under a tolerance of 0.2%.

4.2 Main Results

Table 3 presents the results across all models. Our main findings are summarized as follows:

Overall performance across models remains unsatisfactory. None of the models achieved accuracy above the 60% threshold in any of the settings. Within MLLMs, even the SoTA model OpenAI o4-mini-high reached only 58% accuracy. Many models struggled with handling large-scale inputs, both visual and textual. Moreover, open-source models consistently underperformed proprietary models.

Reasoning-enhanced models consistently outperform those without. Across both input settings, reasoning-enhanced models achieved substantially higher accuracy. Among proprietary models, the top three performers were all reasoning-enhanced. Notably, DeepSeek-R1 (Guo et al.

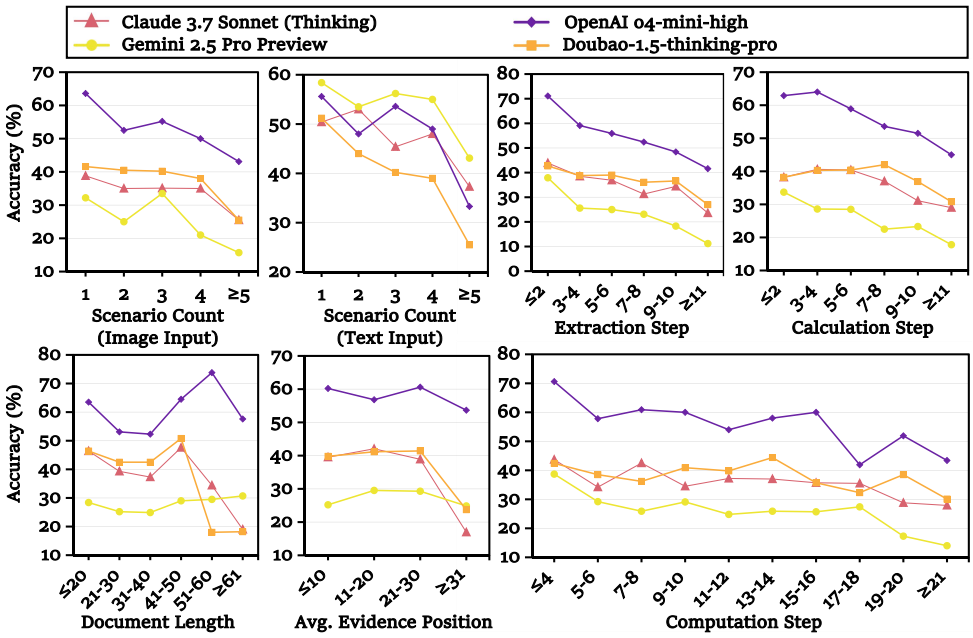


Figure 5: Fine-grained results based on **(top left)** scenario count, **(bottom left)** document length, **(bottom middle)** average evidence position, and **(right)** the number of steps in numerical extraction, numerical calculation, and overall computation.

2025), the only open-source large reasoning model (LRM) in the evaluation, achieved the highest accuracy (40.0%) within its group.

MLLMs face significant bottlenecks in processing long multimodal inputs. While MMLongBenchDoc (Ma et al. 2024) acknowledges the potential information loss introduced by OCR, most MLLMs still perform worse than OCR+LLM models on FinMMDocR, highlighting the bottlenecks MLLMs face when handling image input directly. Specifically, OpenAI o4-mini-high is the only model whose image input performance exceeded its text counterpart, indicating its superior multimodal reasoning capabilities.

Models exhibit substantial disparities in visual understanding. In the OCR+LLM group, the accuracy gap among the top four proprietary models was under 12 points. However, this gap was notably larger in MLLMs (nearly 30 points between OpenAI o4-mini-high and Doubao-1.5-vision-pro). This indicates that visual understanding varies much more significantly across MLLMs, compared to relatively stable language understanding.

4.3 Fine-Grained Analysis

Table 3 and Figure 5 also present the fine-grained results on the further analysis. Detailed results are provided in Appendix E. The key findings are as follows:

Current models struggle with multi-scenario tasks. All exhibit a notable decline in accuracy as the number of scenarios increases. This likely stems from the increased complexity of scenario combinations, requiring more assumptions and associations, thereby better evaluating models' stable reasoning capabilities in complex environments.

Strong document understanding plays a critical role. OpenAI o4-mini-high and Gemini 2.5 Pro Preview maintain stable performance across varying document lengths, likely due to their robust contextual comprehension, while the other two models drop substantially. A similar trend is observed in Figure 5 (bottom middle), where the average index position of evidence positively correlates with document length.

Information extraction, rather than numerical calculation, has a greater impact on model performance in the PoT setting. Accuracy declines progressively with increasing computation steps, following similar patterns to both extraction and calculation performance. Given that calculation typically depends on prior extraction, we hypothesize that this step-dependent accuracy reduction is primarily driven by extraction errors, which aligns with both the PoT's advantage and subsequent error analysis.

4.4 Error Analysis

We randomly sampled 100 failure cases from OpenAI o4-mini-high. Each instance may exhibit multiple error types, which we categorize into four categories. Detailed examples and analysis are provided in Appendix F.

- **Scenario Awareness Error (33/100):** Misinterpretation of task intent, contextual constraints, or key parameters, resulting in flawed reasoning paths.
- **Document Understanding Error (78/100)**: Failure to accurately locate or extract critical information from complex multimodal documents.
- **Knowledge Reasoning Error (44/100)**: Incorrect formula selection or invalid reasoning structures.
- **Numerical Calculation Error (5/100)**: Mistakes in calculation despite correct formulas, often due to precision loss, rounding, or intermediate step errors.

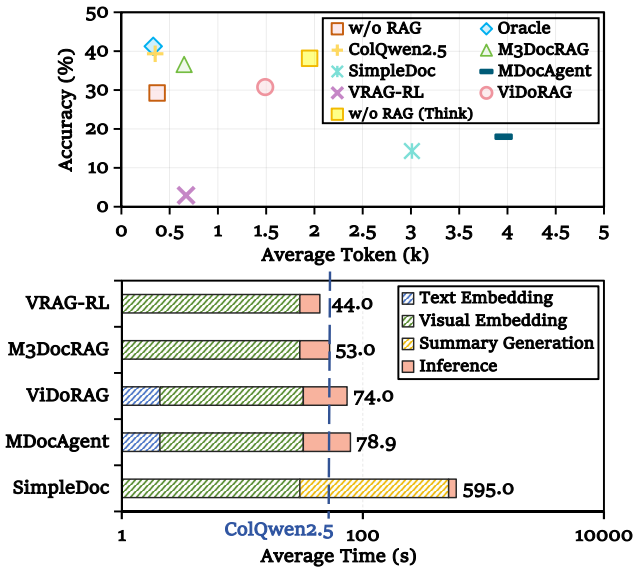


Figure 6: **(Top)** Accuracy and token consumption comparison of RAG methods. **(Bottom)** Runtime composition comparison of Agentic RAGs vs. ColQwen2.5.

4.5 RAG Analysis

We evaluated 6 embedding models (Izacard et al. 2022; Chen et al. 2024; Yu et al. 2025; Faysse et al. 2025) and 5 Agentic RAGs (Cho et al. 2025; Wang et al. 2025a; Han et al. 2025; Jain et al. 2025; Wang et al. 2025b). All Agentic RAGs employed ColQwen2.5 for retrieval and Doubao-1.5-vision-pro for generation. Methods with visual embeddings consistently outperformed text-only approaches, and ColQwen2.5 achieving the best performance. Agentic RAGs underperformed ColQwen2.5, despite consuming more tokens and time, as shown in Figure 6. Detailed analysis is provided in Appendix H. The key findings are as follows:

Agents based solely on semantic retrieval fall short in handling FinMMDocR's complex reasoning demands. SimpleDoc and MDocAgent attempt to enhance semantic representation through multimodal embeddings. However, they often miss the pages containing intermediate variables that are not explicitly stated in the question, resulting in incomplete information retrieval. ViDoRAG partially addresses this issue through an iterative workflow, simulating limited reasoning. Despite lower overall accuracy, it achieves more complete retrieval and reasoning coverage on most of the questions where both models and ColQwen2.5 failed.

Agentic RAGs rely on predefined workflows and fall short of reasoning-enhanced models. ViDoRAG exhibits more numerical errors, like invalid significant figures, likely due to test-based output randomness and context-induced forgetting. Additionally, current frameworks heavily depend on upstream outputs that are rarely questioned or revised downstream, preventing error recovery.

The effectiveness of visually focused strategies remains to be explored. VRAG-RL performed poorly on FinMMDocR, though understandable given the task difficulty. We attribute this to its small base model (7B), and the benefit of scaling up with reinforcement learning remains to be verified.

5 Related Work

Inspired by real-world financial analysis tasks, financial multimodal reasoning demands models to comprehend financial contexts, extract key data from visually dense multimodal financial documents, and perform precise numerical calculations to support multi-step reasoning. However, existing financial QA benchmarks and long-document VQA benchmarks fail to authentically model this task, exhibiting significant gaps. Benchmarks like FinQA (Chen et al. 2021), TAT-QA (Zhu et al. 2021), and ConvFinQA (Chen et al. 2022) only require simple information extraction and arithmetic operations under explicit conditions, while FinanceReasoning (Tang et al. 2025b), FinanceMath (Zhao et al. 2024a), DocMath-Eval (Zhao et al. 2024b), and FinCode (Krumdick et al. 2024) incorporate limited contexts with text-only inputs. FinMMR (Tang et al. 2025a), FinMME (Luo et al. 2025), and MME-Finance (Gan et al. 2025) evaluate models' reasoning capabilities on single or few images. LongDocURL (Deng et al. 2025) and MMLongBench-Doc (Ma et al. 2024) focus on generic multimodal long-document QA, where merely 6% and 8% of tasks involve financial numerical reasoning, further constrained by the scarcity and diversity of domain-specific documents.

MLLMs (ByteDance 2025b; OpenAI 2024; xAI 2024; Bai et al. 2025; AI@Meta 2025; AI 2025; Team et al. 2025) and LMRMs (OpenAI 2025; ByteDance 2025a; Anthropic 2025; DeepMind 2025) offer promising solutions for end-to-end financial multimodal reasoning, leveraging expanded context windows and enhanced reasoning capacities. Concurrently, RAG methods have alleviated models' long-document processing burdens, retrieving relevant pages via semantic similarity between queries and pages. Following text-based RAGs (e.g., BM25, Contriever (Izacard et al. 2022), BGE-M3 (Chen et al. 2024)), vision RAGs like VisRAG (Yu et al. 2025), ColPali (Faysse et al. 2025), and ColQwen2.5 (Faysse et al. 2025) have improved multimodal retrieval performance. Agentic RAG frameworks such as M3DocRAG (Cho et al. 2025), ViDoRAG (Wang et al. 2025a), MDocAgent (Han et al. 2025), SimpleDoc (Jain et al. 2025), and VRAG-RL (Wang et al. 2025b) employ multi-agent collaboration for flexible reasoning.

6 Conclusion

We introduce FinMMDocR, a financial multimodal reasoning benchmark for evaluating MLLMs' professional document understanding and precise multi-step computation in real-world financial scenarios, alongside comprehensive assessments of diverse RAG methods in this complex setting. Extensive experiments reveal significant performance gaps between MLLMs and human experts, with no model exceeding 60% accuracy. While RAG shows promise for information retrieval and reducing visual burdens, fundamental improvements in models' reasoning capabilities and RAG efficiency remain critical future directions. We hope this work establishes foundations for advancing domain-specific multimodal reasoning.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant Nos. 62473271, 62176026), the Beijing Natural Science Foundation (Grant Nos. QY25345, QY25338), the Fundamental Research Funds for the Beijing University of Posts and Telecommunications (Grant No. 2025AI4S03), the BUPT Innovation and Entrepreneurship Support Program (Grant Nos. 2025-YC-A033, 2025-YC-A042), and data support from Hithink RoyalFlush Information Network Co., Ltd. This work is also supported by the Engineering Research Center of Information Networks, Ministry of Education, China. We would also like to thank the anonymous reviewers and area chairs for constructive discussions and feedback.

References

AI, M. 2025. Mistral Small 3.1. https://mistral.ai/news/mistral-small-3-1. Accessed: 2025-03-17.

AI@Meta. 2025. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Accessed: 2025-04-05.

Anthropic. 2025. Claude 3.7 Sonnet and Claude Code. https://www.anthropic.com/news/claude-3-7-sonnet. Accessed: 2025-02-25.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923.

ByteDance. 2025a. Doubao-1.5-thinking-pro Model Card. https://console.volcengine.com/ark/region:ark+cn-beijing/model/detail?Id=doubao-1-5-thinking-pro. Accessed: 2025-04-15.

ByteDance. 2025b. Doubao-1.5-vision-pro Model Card. https://console.volcengine.com/ark/region:ark+cn-beijing/model/detail?Id=doubao-1-5-vision-pro. Accessed: 2025-03-28.

Chen, J.; Xiao, S.; Zhang, P.; Luo, K.; Lian, D.; and Liu, Z. 2024. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 2318–2335. Bangkok, Thailand: Association for Computational Linguistics.

Chen, W.; Ma, X.; Wang, X.; and Cohen, W. W. 2023. Program of Thoughts Prompting: Disentangling Computation from Reasoning for Numerical Reasoning Tasks. *Transactions on Machine Learning Research*.

Chen, Z.; Chen, W.; Smiley, C.; Shah, S.; Borova, I.; Langdon, D.; Moussa, R.; Beane, M.; Huang, T.-H.; Routledge, B.; and Wang, W. Y. 2021. FinQA: A Dataset of Numerical Reasoning over Financial Data. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 3697–3711. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Chen, Z.; Li, S.; Smiley, C.; Ma, Z.; Shah, S.; and Wang, W. Y. 2022. ConvFinQA: Exploring the Chain of Numerical Reasoning in Conversational Finance Question Answering. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 6279–6292. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Cho, J.; Mahata, D.; Irsoy, O.; He, Y.; and Bansal, M. 2025. M3DocVQA: Multi-modal Multi-page Multi-document Understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 6178–6188.

DeepMind, G. 2025. Build rich, interactive web apps with an updated Gemini 2.5 Pro. https://blog.google/products/gemini/gemini-2-5-pro-updates/. Accessed: 2025-05-06.

Deng, C.; Yuan, J.; Bu, P.; Wang, P.; Li, Z.-Z.; Xu, J.; Li, X.-H.; Gao, Y.; Song, J.; Zheng, B.; and Liu, C.-L. 2025. LongDocURL: a Comprehensive Multimodal Long Document Benchmark Integrating Understanding, Reasoning, and Locating. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1135–1159. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.

Faysse, M.; Sibille, H.; Wu, T.; Omrani, B.; Viaud, G.; HUDELOT, C.; and Colombo, P. 2025. ColPali: Efficient Document Retrieval with Vision Language Models. In *The Thirteenth International Conference on Learning Representations*.

Gan, Z.; Zhang, D.; Li, H.; Wu, Y.; Lin, X.; Liu, J.; Wu, H.; Fu, C.; Xu, Z.; Zhang, R.; and Dai, Y. 2025. MME-Finance: A Multimodal Finance Benchmark for Expert-level Understanding and Reasoning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, 12867–12874. New York, NY, USA: Association for Computing Machinery. ISBN 9798400720352.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the v in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Wang, P.; Zhu, Q.; Xu, R.; Zhang, R.; Ma, S.; Bi, X.; et al. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081): 633–638.

Han, S.; Xia, P.; Zhang, R.; Sun, T.; Li, Y.; Zhu, H.; and Yao, H. 2025. MDocAgent: A Multi-Modal Multi-Agent Framework for Document Understanding. arXiv:2503.13964.

Izacard, G.; Caron, M.; Hosseini, L.; Riedel, S.; Bojanowski, P.; Joulin, A.; and Grave, E. 2022. Unsupervised Dense Information Retrieval with Contrastive Learning. *Transactions on Machine Learning Research*.

Jain, C.; Wu, Y.; Zeng, Y.; Liu, J.; Dai, S.; Shao, Z.; Wu, Q.; and Wang, H. 2025. SimpleDoc: Multi-Modal Document Understanding with Dual-Cue Page Retrieval and Iterative Refinement. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 28398–28415. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-332-6.

Krumdick, M.; Koncel-Kedziorski, R.; Lai, V. D.; Reddy, V.; Lovering, C.; and Tanner, C. 2024. BizBench: A Quantitative Reasoning Benchmark for Business and Finance. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8309–8332. Bangkok, Thailand: Association for Computational Linguistics.

Li, Y.; Liu, Z.; Li, Z.; Zhang, X.; Xu, Z.; Chen, X.; Shi, H.; Jiang, S.; Wang, X.; Wang, J.; Huang, S.; Zhao, X.; Jiang, B.; Hong, L.; Wang, L.; Tian, Z.; Huai, B.; Luo, W.; Luo, W.; Zhang, Z.; Hu, B.; and Zhang, M. 2025. Perception, Reason, Think, and Plan: A Survey on Large Multimodal Reasoning Models. arXiv:2505.04921.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 34892–34916. Curran Associates, Inc.

Lu, P.; Bansal, H.; Xia, T.; Liu, J.; Li, C.; Hajishirzi, H.; Cheng, H.; Chang, K.-W.; Galley, M.; and Gao, J. 2024. MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. In *The Twelfth International Conference on Learning Representations*.

Luo, J.; Kou, Z.; Yang, L.; Luo, X.; Huang, J.; Xiao, Z.; Peng, J.; Liu, C.; Ji, J.; Liu, X.; Han, S.; Zhang, M.; and Guo, Y. 2025. FinMME: Benchmark Dataset for Financial Multi-Modal Reasoning Evaluation. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 29465–29489. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.

Ma, Y.; Zang, Y.; Chen, L.; Chen, M.; Jiao, Y.; Li, X.; Lu, X.; Liu, Z.; Ma, Y.; Dong, X.; Zhang, P.; Pan, L.; Jiang, Y.-G.; Wang, J.; Cao, Y.; and Sun, A. 2024. MMLONGBENCH-DOC: Benchmarking Long-context Document Understanding with Visualizations. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 95963–96010. Curran Associates, Inc.

OpenAI. 2024. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/. Accessed: 2024-05-13.

OpenAI. 2025. OpenAI o3 and o4-mini System Card. https://openai.com/index/o3-o4-mini-system-card/. Accessed: 2025-04-16.

Singh, A.; Ehtesham, A.; Kumar, S.; and Khoei, T. T. 2025. Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG. arXiv:2501.09136.

Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; and Rohrbach, M. 2019. Towards VQA Models That Can Read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Smith, R. 2007. An Overview of the Tesseract OCR Engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, 629–633.

Tanaka, R.; Nishida, K.; Nishida, K.; Hasegawa, T.; Saito, I.; and Saito, K. 2023. SlideVQA: A Dataset for Document Visual Question Answering on Multiple Images. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11): 13636–13645.

Tang, Z.; E, H.; Liu, J.; Yang, Z.; Li, R.; Rong, Z.; He, H.; Hao, Z.; Hu, X.; Ji, K.; Ma, Z.; Ji, M.; Zhang, J.; Ma, C.; Zheng, Q.; Liu, Y.; Huang, Y.; Hu, X.; Huang, Q.; Xie, Z.; and Peng, S. 2025a. FinMMR: Make Financial Numerical Reasoning More Multimodal, Comprehensive, and Challenging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3245–3257.

Tang, Z.; E, H.; Ma, Z.; He, H.; Liu, J.; Yang, Z.; Rong, Z.; Li, R.; Ji, K.; Huang, Q.; Hu, X.; Liu, Y.; and Zheng, Q. 2025b. FinanceReasoning: Benchmarking Financial Numerical Reasoning More Credible, Comprehensive and Challenging. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15721–15749. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.

Team, G.; Kamath, A.; Ferret, J.; Pathak, S.; Vieillard, N.; Merhej, R.; Perrin, S.; Matejovicova, T.; Ramé, A.; Rivière, M.; et al. 2025. Gemma 3 Technical Report. arXiv:2503.19786.

Wang, K.; Pan, J.; Shi, W.; Lu, Z.; Ren, H.; Zhou, A.; Zhan, M.; and Li, H. 2024. Measuring Multimodal Mathematical Reasoning with MATH-Vision Dataset. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 95095–95169. Curran Associates, Inc.

Wang, Q.; Ding, R.; Chen, Z.; Wu, W.; Wang, S.; Xie, P.; and Zhao, F. 2025a. ViDoRAG: Visual Document Retrieval-Augmented Generation via Dynamic Iterative Reasoning Agents. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 9124–9145. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-332-6.

Wang, Q.; Ding, R.; Zeng, Y.; Chen, Z.; Chen, L.; Wang, S.; Xie, P.; Huang, F.; and Zhao, F. 2025b. VRAG-RL: Empower Vision-Perception-Based RAG for Visually Rich Information Understanding via Iterative Reasoning with Reinforcement Learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

xAI. 2024. Grok 2 Vision Model Card. https://docs.x.ai/docs/models/grok-2-vision-1212. Accessed: 2024-12-12.

Yu, S.; Tang, C.; Xu, B.; Cui, J.; Ran, J.; Yan, Y.; Liu, Z.; Wang, S.; Han, X.; Liu, Z.; and Sun, M. 2025. VisRAG: Vision-based Retrieval-augmented Generation on Multi-modality Documents. In *The Thirteenth International Conference on Learning Representations*.

Yu, W.; Yang, Z.; Li, L.; Wang, J.; Lin, K.; Liu, Z.; Wang, X.; and Wang, L. 2024. MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities. In *Forty-first International Conference on Machine Learning*.

Zellers, R.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. From Recognition to Cognition: Visual Commonsense Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhao, Y.; Liu, H.; Long, Y.; Zhang, R.; Zhao, C.; and Cohan, A. 2024a. FinanceMATH: Knowledge-Intensive Math Reasoning in Finance Domains. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12841–12858. Bangkok, Thailand: Association for Computational Linguistics.

Zhao, Y.; Long, Y.; Liu, H.; Kamoi, R.; Nan, L.; Chen, L.; Liu, Y.; Tang, X.; Zhang, R.; and Cohan, A. 2024b. DocMath-Eval: Evaluating Math Reasoning Capabilities of LLMs in Understanding Long and Specialized Documents. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 16103–16120. Bangkok, Thailand: Association for Computational Linguistics.

Zhu, F.; Lei, W.; Huang, Y.; Wang, C.; Zhang, S.; Lv, J.; Feng, F.; and Chua, T.-S. 2021. TAT-QA: A Question Answering Benchmark on a Hybrid of Tabular and Textual Content in Finance. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3277–3287. Online: Association for Computational Linguistics.