

---

# FinDocMRE: A Benchmark for Document-Level Financial Multimodal Reasoning Evaluation

---

Jiayong Zhu<sup>1</sup>, Jiangtong Li<sup>1</sup>, Jinru Ding<sup>2</sup>, Dawei Cheng<sup>1</sup>, Jie Xu<sup>2</sup>, Feng Yu<sup>3</sup>

1. School of Computer Science and Technology, Tongji University, Shanghai, China

2. Shanghai Artificial Intelligence Laboratory, Shanghai, China

3. Guotai Haitong Securities Co.,Ltd.

{jiayongz, jiangtongli, dcheng}@tongji.edu.cn

## Abstract

While Large Multimodal Models (LMMs) excel in general visual tasks, their deployment in specialized financial contexts remains insufficient. Existing benchmarks prioritize isolated charts, often overlooking the need to integrate data from text, tables, and images within comprehensive financial documents. To address this limitation, we introduce **FINDOCMRE**, a multi-image document-level benchmark designed for financial multimodal reasoning. We construct the dataset via a semi-automated pipeline that combines Visual-Centric Generation with Expert Verification, thereby minimizing text bias and ensuring high annotation quality. Spanning twelve domains, the benchmark comprises 12,207 samples derived from 2,878 financial reports, designed to evaluate multi-image processing and document-level understanding across five distinct task types. Extensive experiments with eleven representative LMMs reveal that no model surpasses an overall score of 65, highlighting challenges in integrating visual grounding with logical reasoning within complex document environments. Specifically, we observe a significant performance divergence across tasks, where models exhibit proficiency in semantic narrative construction but struggle with numerical estimation and cross-page visual grounding. FINDOCMRE serves as a rigorous benchmark to guide the evolution of financial LMMs towards expert-level document analysis and reasoning.

## 1 Introduction

Recent progress in Large Multimodal Models (LMMs) has significantly improved multimodal understanding and reasoning capabilities [1, 2]. Therefore, applying these models within the financial domain has attracted growing interest. As LMMs evolve from general assistants into specialized financial agents, rigorous evaluation benchmarks become essential. These benchmarks serve as metrics for progress while guiding model development to meet the demands of professional analysis.

Previous research has established various methods for assessing AI models’ financial capabilities. However, existing benchmarks primarily target natural language processing (NLP) tasks [3, 4] or isolated chart comprehension [5]. Consequently, these datasets fail to adequately represent the complexity of real-world **document-level** financial analysis. In contrast to isolated tasks, document-level analysis involves integrating interwoven text, dense data tables, and visual charts, often requiring cross-page reasoning. Despite such evaluation being necessary, constructing a document-level benchmark presents three primary challenges: 1) **Data Availability**: High-quality, open-source financial PDFs providing structural alignments are rare compared to plain text corpora, limiting multimodal evaluation. 2) **Annotation Costs**: Manually designing complex, multi-hop reasoning questions from full documents is highly labor-intensive and demands substantial domain expertise, limiting both dataset scale and diversity. 3) **Generation Quality**: While generating questions via

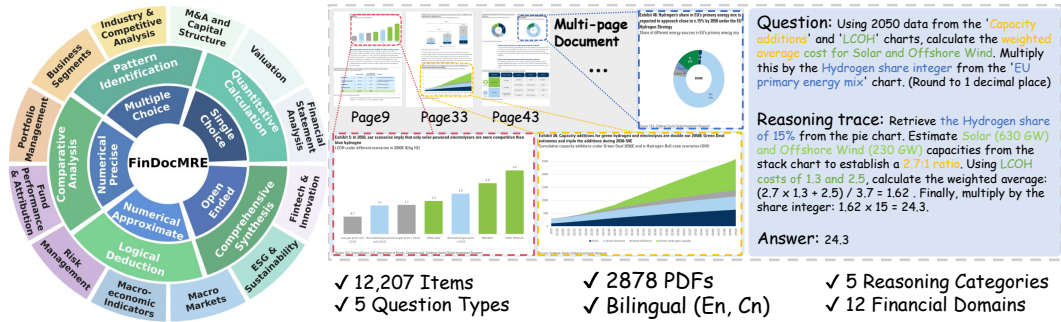


Figure 1: The overall taxonomy and statistics of the FINDOCMRE benchmark. An example illustrates the organization of a multi-image document-level sample.

LMMs is scalable, prior work indicates this method introduces **text bias** [6, 7], yielding coarse-grained queries that lack the numerical precision required for testing expert-level visual reasoning.

To address these challenges, we introduce **FINDOCMRE (Financial Document-level Multimodal Reasoning Evaluation)**, a benchmark designed to assess LMMs within document-level financial reasoning scenarios. We establish a semi-automated construction pipeline to balance scalability and data quality: 1) **Data Preparation**: We collect bilingual financial PDF reports (e.g., annual and research reports) from our collaborators and apply a Joint-Rule Filtering Mechanism (geometric, semantic, and textual indexing) to extract high-value charts and remove noise; 2) **Visual-Centric QA Generation**: We input aggregated chart contexts (excluding surrounding text) into LMMs to ensure the generation of precise, vision-dependent QA pairs and reasoning traces; 3) **Expert Verification**: We employ a human-in-the-loop protocol where senior financial experts verify visual grounding (chart IDs) and correct logical errors, retaining only valid samples for the final dataset. As detailed in Tab. 1, FINDOCMRE stands out as a large-scale financial benchmark supporting **multi-image document-level** inputs, featuring the largest data scale and diverse answer formats. As shown in Fig. 1, the benchmark comprises 12,207 samples derived from 2,878 financial reports spanning twelve industries, covering five question formats and five reasoning categories. We conduct comprehensive evaluations of eleven representative LMMs alongside human expert baselines on FINDOCMRE, employing full PDF formats to simulate realistic financial workflows. Our multidimensional analysis reveals a significant divergence between quantitative calculation and qualitative synthesis: **while models excel at semantic narrative construction, they struggle with precise trend estimation**. Notably, performance deteriorates as visual dependency increases, underscoring that multi-image reasoning remains a bottleneck for current leading multimodal architectures.

To investigate these limitations, we perform multidimensional analyses and ablation studies exploring factors like visual context, model scaling, visual complexity, and prompt robustness, among others. Reducing the visual search space leads to substantial performance gains, indicating a deficiency in aggregating fragmented information from extensive document contexts. With no model surpassing an overall score of 65, our results highlight the formidable challenge of integrating visual grounding with rigorous numerical reasoning. In summary, our contributions are as follows:

- We present FINDOCMRE, the first large-scale benchmark addressing multi-image document-level financial multimodal reasoning, filling the gap in current evaluation standards.
- We develop a semi-automated pipeline integrating Visual-Centric Generation with Expert Verification, reducing textual noise to yield 12,207 high-quality QA samples from 2,878 reports across twelve financial domains.
- We conduct a multidimensional evaluation of eleven representative LMMs against real-world financial expert baselines, employing comparative analysis (Full PDF vs. Cropped) to isolate specific deficiencies in visual grounding and quantitative reasoning.

Table 1: Comparison with representative benchmarks. **Vol.:** Volume; **MM:** Multimodal; **Fin.:** Financial; **M-Img:** Multi-image; **Doc.:** Document-level; **MC:** Multiple-choice; **N<sub>pre</sub>:** Numerical precise; **N<sub>app</sub>:** Numerical approximate; **Open:** Open-ended QA.

Dataset	Dataset Features					Answer Format			
	Vol.	MM	Fin.	M-Img	Doc.	MC	N <sub>pre</sub>	N <sub>app</sub>	Open
MMStar	1,500	✓	✗	✗	✗	✗	✗	✗	✗
MathVision	3,040	✓	✗	✗	✗	✗	✓	✗	✗
FinEval	8,351	✗	✓	✗	✗	✗	✓	✗	✓
FinQA	8,281	✗	✓	✗	✗	✗	✓	✗	✗
FinanceBench	10,231	✗	✓	✗	✗	✗	✓	✗	✓
FAMMA	1,935	✓	✓	✓	✗	✗	✓	✗	✓
MME-Finance	2,274	✓	✓	✗	✗	✗	✓	✓	✓
FinMME	11,099	✓	✓	✗	✗	✓	✓	✗	✗
FinMMDocR	1,200	✓	✓	✓	✓	✗	✓	✗	✗
<b>FinDocMRE</b>	<b>12,207</b>	✓	✓	✓	✓	✓	✓	✓	✓

## 2 Related Work

### 2.1 Large Multi-modal Models

Significant advancements in LMMs, ranging from proprietary models like GPT [8], Gemini [1], Claude [9], GLM [10], and SEED [11], to open-source counterparts like LLaVA [12], Qwen-VL [2], and InternVL [13], have greatly enhanced general visual understanding and multimodal reasoning. However, deploying these general-purpose models in the financial domain faces distinct challenges, as professional analysis demands not only visual recognition but also precise numerical reasoning and trend interpretation within noise-sensitive contexts. Therefore, while these foundation models demonstrate broad capabilities, their reliability as financial agents in realistic, document-intensive scenarios remains under-explored, necessitating evaluation paradigms that go beyond isolated tasks.

### 2.2 Multimodal Benchmarks

To assess the evolving capabilities of LMMs, various general-purpose multimodal benchmarks have been introduced, including MME [14], MathVision [15], MathVerse [16], MMMU [17], MathVista [18], ChartQA [19], MM-Star [20], and others [21–23]. However, these frameworks often fall short in specialized financial contexts, typically confining tasks to isolated visual question answering and omitting the complex grounding demands of the domain. In parallel, document-centric datasets like DocVQA [24] heavily emphasize Optical Character Recognition (OCR) and entity extraction, often bypassing the high-level reasoning required to synthesize cross-modal data (text, tables, and charts). These limitations motivate the development of FINDOCMRE, which aims to bridge basic visual perception with expert-level document analysis and reasoning.

### 2.3 Benchmarks in Finance

In the text domain, benchmarks have evolved from focusing on specific NLP and reasoning tasks, exemplified by FLUE [25], CFLEB [26], InvestorBench [27], and FinanceBench [28], to evaluating holistic financial literacy, as seen in suites like FinEval [29], FinBen [30], CFBenchmark [31], SuperCLUE-Fin [32], OpenFinData [33], and CNFinBench [4]. More recently, initiatives like CFBenchmark-MM [34], MME-Finance [35], FinMME [36], FinMR [37], FinReasoning [38], and FinTBS [39] have extended evaluation to the visual modality, targeting chart and table interpretation. Nevertheless, these datasets typically simplify the task by relying on isolated, pre-localized images. The concurrent work FinMMDocR [40] investigates document-level financial reasoning, focusing on numerical computation. While sharing this setting, FINDOCMRE provides a larger-scale evaluation and introduces more diverse task formats, including multiple-choice and open-ended QA, demanding a more comprehensive understanding of complex financial logic and holistic document intelligence.

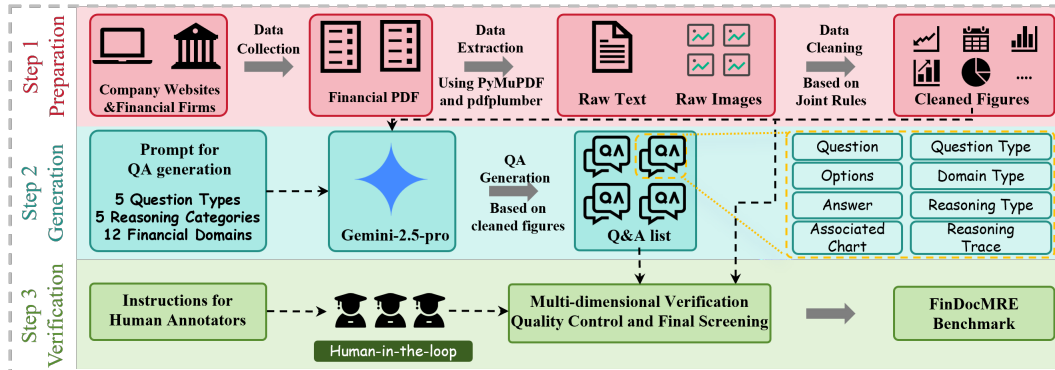


Figure 2: The annotation pipeline of FinDocMRE benchmark .

### 3 FinDocMRE Benchmark

As illustrated in Fig. 2, FinDocMRE is constructed through a semi-automated pipeline: Data Preparation (Sec. 3.1), Visual-Centric Generation (Sec. 3.2), and Expert Verification (Sec. 3.3). This framework combines the scalability of automated generation with strict human-in-the-loop protocols to ensure the accuracy of financial reasoning. Comprehensive statistics are in Sec. 3.4.

#### 3.1 Data Preparation

**Data Acquisition.** To construct a benchmark reflecting real-world scenarios, we collect a corpus of 2,878 financial PDFs (2,469 Chinese, 409 English) from official websites and third-party aggregators. These sources encompass authoritative domestic and international institutions, spanning leading securities firms and investment banks to global economic organizations. We prioritize document types essential for professional analysis, including annual reports, weekly newsletters, and industry research reports, while excluding plain-text files to explicitly target multimodal reasoning.

**Chart-Centric Extraction and Cleaning.** To process these unstructured documents, we implement a Chart-Centric extraction pipeline, using PyMuPDF and pdfplumber to parse embedded images and textual content. To maintain data quality, we apply a Joint-Rule Filtering Mechanism: (1) Geometric Filtering discards layout artifacts based on extreme aspect ratios or low resolutions; (2) Semantic Filtering utilizes image similarity computation to deduplicate visual content, while employing OCR to filter out non-informative elements like corporate logos; and (3) Textual Indexing Verification retains only charts explicitly referenced in the narrative (e.g., “Figure 1”) to ensure analytical relevance. **Further details on data extraction, cleaning and annotation are provided in the Appendix A.**

#### 3.2 Visual-Centric Generation

Preliminary experiments reveal that feeding full PDF pages or images with surrounding text introduces textual bias, compromises accurate chart interpretation. To mitigate this, we employ a visual-centric generation strategy using Gemini-2.5-Pro. As shown in Fig. 2-Step 2, for each document, we combine all extracted charts into an image sequence and prompt (Tab. 7) the model to generate diverse question-answer pairs and reasoning traces, prioritizing cross-chart synthesis where applicable. Excluding surrounding text, we compel the model to derive insights solely from visual data, simulating the workflow of an analyst where visual evidence is interpreted independently of textual summaries.

To ensure a gradient of difficulty levels, we incorporate a multi-dimensional taxonomy directly into the generation prompts. We apply strict JSON schema constraints across five question formats: *single\_choice* and *multiple\_choice* tasks require plausible distractors; *numerical\_precise* and *numerical\_approximate* distinguish exact results from trend estimation; and *open\_ended* queries require synthesizing financial concepts. As illustrated in Fig. 1, this taxonomy spans 12 financial domains (e.g., *Financial Statement Analysis*, *Risk Management*) and five cognitive reasoning categories, progressing from basic *Quantitative Calculation* to advanced *Comprehensive Synthesis*.

Table 2: Statistics of the FinDocMRE dataset.

Category	Count
<i>Questions (QA Pairs)</i>	
Total Questions	12,207
– Chinese	7,105
– English	5,102
<i>Source Documents (PDFs)</i>	
Total PDFs	2,878
– Chinese	2,469
– English	409
<i>Question Types Distribution</i>	
Single Choice	3,439
Multiple Choice	1,423
Numerical Precise	2,405
Numerical Approximate	2,454
Open Ended	2,486

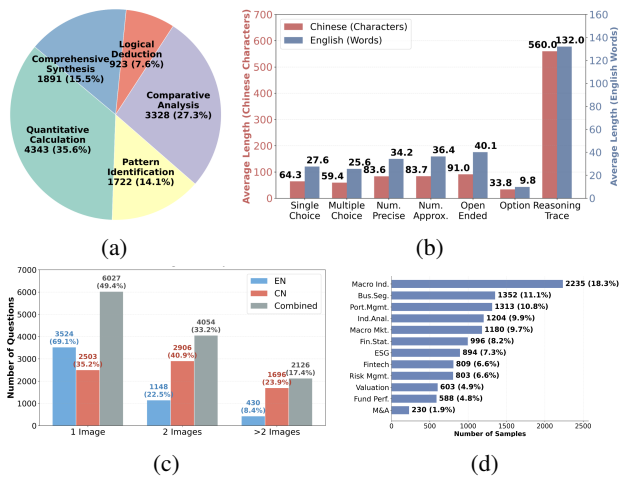


Figure 3: Dataset statistics overview: (a) Reasoning types, (b) Lengths, (c) Image quantity, (d) Domains.

Each generated sample is structured as a metadata object containing the *Question*, *Answer*, *Domain Type*, and *Reasoning Type*. We also require the generation of a detailed *Reasoning Trace*, outlining the step-by-step logic and calculation formulas used to derive the answer. This trace grounds the model inference and assists the subsequent Expert Verification phase, allowing annotators to validate the correctness of complex derivations. Finally, the model records specific `chart_ids` in the *Associated Chart* field, linking each query to its source evidence within the multi-image input.

### 3.3 Expert Verification

To counteract the hallucination risks inherent to visual-centric generation and guarantee expert-level fidelity, we deploy a Human-In-The-Loop (HITL) verification workflow (Fig. 2-Step 3). Each document and its corresponding generated sample are verified by three senior financial researchers, each possessing over three years of professional experience in securities firms. Guided by rigorous annotation protocols in Tab. 8, experts audit each sample by cross-referencing the generated reasoning traces against the source charts. This examination validates the logic of **Question**, the plausibility of **Options**, the accuracy of **Answer**, and the precision of **Associated Chart IDs**.

To ensure the highest data quality, the verification team performs strict filtering rather than manual correction. During the audit, researchers explicitly reject any sample that displays ambiguous phrasing, factual errors, weak option quality, or mismatched chart references. We adopt a unanimous consensus mechanism: if any of the three experts identifies a flaw, the sample is immediately discarded. Therefore, only samples validated by all three researchers without reservation are retained in the final FINDOC MRE benchmark. This rigorous quality control ensures adherence to professional analysis standards, resulting in a final retention rate of approximately 55% for FINDOC MRE.

### 3.4 Dataset Statistics

**Scale and Linguistic Diversity.** FINDOC MRE contains 12,207 annotated samples derived from 2,878 unique financial documents, comprising 7,105 Chinese and 5,102 English QA pairs, carefully balanced to evaluate cross-lingual reasoning capabilities. Rooted in native reports from major global markets, this bilingual composition ensures robust applicability across diverse financial contexts.

**Distribution of Question Types.** In contrast to general multimodal datasets focused on basic visual recognition tasks, FINDOC MRE prioritizes quantitative analysis. As shown in Tab. 2, *Numerical Reasoning* represents a core component: *numerical\_precise* (2,405) and *numerical\_approximate* (2,454) together account for 39.8% of the total samples. Furthermore, *open\_ended* queries (2,486) represent 20.4% of the total, demanding coherent textual synthesis beyond mere option selection.

Table 3: Main evaluation results on FINDOCMRE. We report Accuracy (%) for objective tasks and raw 0-5 scores for open-ended tasks (normalized percentages in parentheses). The Overall Score is the macro-average of the five metrics. The best model results in each column are marked in **bold**.

Model	Single-Ch.	Multi-Ch.	Num. (Pre.)	Num. (App.)	Open-Ended	Overall Score
<i>Proprietary Models</i>						
Gemini-3.1 Pro	<b>79.12</b>	<b>58.05</b>	<b>63.20</b>	<b>37.82</b>	4.12 (82.4)	<b>64.12</b>
GPT-5.4	76.04	56.71	57.17	31.21	<b>4.41</b> (88.2)	61.87
Doubao-1.6 Vision	73.77	57.62	50.31	26.65	4.19 (83.8)	58.43
GPT-5	73.02	51.16	48.07	26.73	4.33 (86.6)	57.12
Gemini-2.5 Pro	72.70	45.33	36.22	20.70	3.90 (78.0)	50.59
Gemini-2.5 Flash	64.55	41.67	25.16	20.95	3.94 (78.8)	46.23
Qwen3-Max	49.96	27.69	3.45	11.86	3.51 (70.2)	32.63
Grok-4.1 Fast	53.97	23.19	3.74	9.90	3.26 (65.2)	31.20
<i>Open-Source Models</i>						
Qwen3-VL-235B	75.25	52.64	52.89	26.28	4.08 (81.6)	57.73
Qwen3-VL-30B	68.94	43.43	41.16	20.70	3.74 (74.8)	49.81
Qwen2.5-VL-72B	59.44	28.95	14.10	14.43	3.19 (63.8)	36.14
<i>Human Performance</i>						
Human Expert	88.37	76.04	85.03	62.59	4.25 (85.0)	79.41

**Reasoning Depth and Visual Dependency.** To assess reasoning difficulty, we examine the length of the expert-verified Chain-of-Thought (CoT) sequences. The average CoT length is 560 characters for Chinese and 132 words for English samples, implying the necessity of multi-step deduction to resolve complex financial queries. Regarding visual grounding, approximately 50.6% of the questions require synthesizing information across multiple charts, necessitating cross-page reasoning within extensive document contexts. Fig. 3 details the distribution across 12 financial domains and 5 reasoning types, illustrating the coverage of FINDOCMRE across different analysis scenarios.

## 4 Experiment

### 4.1 Evaluated LMMs

To assess the performance of state-of-the-art LMMs on FINDOCMRE, we select 9 models ranging from commercial APIs to open-weights architectures. Our evaluation includes nine **proprietary models**: Gemini-3.1-Pro, GPT-5.4, GPT-5 [41], Doubao1.6-Vision [11], Qwen3-Max [42], Grok4.1-Fast [43], Gemini-2.5-Pro, and Gemini-2.5-Flash [1]. We also evaluate three **open-source models**: Qwen2.5-VL-72B [2], Qwen3-VL-30B-A3B, and Qwen3-VL-235B-A22B [42]. All models are tested in a zero-shot setting to assess their inherent financial document reasoning capabilities without external tools. Our preliminary screening excludes smaller architectures (*e.g.*, Qwen2.5-VL-3B/7B) due to limited capabilities, and models like GLM and DeepSeek-VL, where safety filters cause high refusal rates on financial documents. Finally, to establish a human performance baseline, we engage two professional financial analysts from securities firms to complete the benchmark under identical input conditions, utilizing their averaged scores to represent the **Human Expert** results.

### 4.2 Evaluation Metrics

We employ a dual-evaluation strategy aligned with specific task requirements. For deterministic tasks (single-choice, multiple-choice, and numerical reasoning), we report standard **Accuracy**. We apply strict validation protocols: multiple-choice questions require an **Exact Match** to the ground truth option set; numerical estimation allows a **5% relative error tolerance** for visual ambiguity, while precise calculations require exact equality. For open-ended queries, where n-gram metrics (*e.g.*, BLEU [44]) fail to capture semantic validity, we adopt an LLM-as-a-Judge paradigm [45]. To mitigate single-evaluator bias, we employ a panel of three distinct models (Gemini-2.5-Flash, GPT-4o, and Qwen3-Max) to score responses against reference answers from [0,5], assessing correctness, completeness, and logical coherence, taking their average as the final score. **See Appendix B.2 for judge bias and alignment analysis between LMM evaluators and human experts.**

**Overall Performance Metric.** To mitigate sample size imbalances across task categories, we calculate the **Overall Score** using a macro-average strategy. We scale *open\_ended* scores to a 0-100

Table 4: Performance breakdown across five reasoning categories. Model performance is formatted as Accuracy (%) / Score (0-5). Note that the **Quant. Calc.** category consists exclusively of objective questions, so only Accuracy is reported. Best results are marked in **bold**.

Reasoning Type	Quant. Calc.	Pattern Ident.	Comp. Anal.	Logic. Ded.	Comp. Synth.
<i>Proprietary Models</i>					
Gemini-3.1 Pro	<b>51.2</b>	<b>68.5</b> / 3.6	<b>73.8</b> / 3.7	<b>87.3</b> / 4.3	<b>77.3</b> / 4.2
GPT-5.4	47.3	64.4 / <b>3.9</b>	69.5 / <b>3.9</b>	83.5 / <b>4.5</b>	74.8 / <b>4.6</b>
Doubao-1.6 Vision	40.0	60.7 / 3.6	64.7 / 3.7	78.5 / 4.3	68.0 / 4.2
GPT-5	38.6	61.1 / 3.7	60.7 / 3.8	80.3 / 4.5	66.7 / 4.4
Gemini-2.5 Pro	29.9	56.6 / 3.7	58.3 / 3.4	79.2 / 4.3	72.2 / 3.9
Gemini-2.5 Flash	24.2	52.4 / 3.3	51.2 / 3.1	73.3 / 4.3	60.3 / 4.0
Qwen3-Max	8.3	40.3 / 3.1	35.6 / 2.7	64.7 / 3.9	52.0 / 3.5
Grok-4.1 Fast	7.5	39.9 / 2.4	37.1 / 2.1	68.2 / 3.7	48.1 / 3.3
<i>Open-Source Models</i>					
Qwen3-VL-235B	41.1	60.2 / 3.5	64.4 / 3.3	77.2 / 4.4	70.7 / 4.1
Qwen3-VL-30B	32.2	53.4 / 3.4	57.4 / 2.7	72.1 / 4.0	54.7 / 3.8
Qwen2.5-VL-72B	14.3	49.1 / 2.8	42.5 / 2.5	70.0 / 3.7	52.0 / 3.2

range ( $S_{norm} = S_{raw} \times 20$ ) and compute the unweighted mean across all five question types. This method reduces bias from majority classes, providing a balanced assessment of reasoning abilities across dimensions. **Refer to Appendix B for full evaluation setting and prompts.**

### 4.3 Multi-Dimensional Analysis

**Main Results.** Tab. 3 summarizes the performance of the eleven evaluated LMMs against the expert baseline. Results indicate that the proprietary Gemini-3.1 Pro (64.12) and the open-weights Qwen3-VL-235B (57.73) achieve the highest overall scores among the evaluated AI models. Notably, Qwen3-VL-235B rivals commercial models, attaining high accuracy in *single\_choice* (75.25%) and *numerical\_precise* (52.89%). However, with no model surpassing an overall score of 65, a substantial gap of approximately 15 points remains compared to the **Human Expert** baseline (79.41), demonstrating that FINDOCMRE presents a formidable challenge.

A marked performance disparity exists between visual perception and quantitative reasoning. While models perform well on single-choice questions, accuracy declines on *numerical\_approximate* tasks, where the top-performing Gemini-3.1 Pro reaches only 37.82%. Since *numerical\_approximate* tasks require estimating values directly from figures, this suggests that current LMMs lack precision in trend estimation and arithmetic under visual ambiguity. However, in *open\_ended* scenarios, GPT-5.4 leads with a semantic score of 4.41/5.00, even outperforming the human experts (4.25), indicating strong capabilities in synthesizing financial narratives despite calculation limitations.

**Performance by Financial Domain.** Tab. 12 and 13 illustrate performance variations across 12 financial domains. A divergence exists between calculation and synthesis: while Gemini-3.1 Pro leads in quantitative domains such as Valuation (75.5%) and M&A (66.3%), GPT-5.4 outperforms in qualitative reasoning, achieving the highest open-ended scores in 11 categories. This suggests distinct optimization strategies: certain models emphasize visual grounding for extraction, while others focus on semantic synthesis for narrative construction. Furthermore, a performance gap separates standardized "micro" domains (*e.g.*, Risk Management, 68% accuracy) from "macro" domains (*e.g.*, Fund Performance, <60%), implying that LMMs face greater challenges interpreting market trendlines compared to structured formats. Notably, Gemini-3.1 Pro shows strong generalization in Fintech (66.1%), successfully processing the conceptual diagrams.

**Performance by Reasoning Type.** Tab. 4 details model performance across five reasoning categories, revealing distinct performance profiles. While Gemini-3.1 Pro and GPT-5.4 lead in *Quantitative Calculation* (51.2% and 47.3% respectively), surpassing GPT-5 (38.6%), this trend reverses in tasks requiring abstract inference. In *Logical Deduction*, Gemini-3.1 Pro achieves the highest accuracy (87.3%), while GPT-5.4 achieves the highest semantic score (4.5), indicating strong logical reasoning capabilities. In *Comprehensive Synthesis*, results diverge: Gemini-3.1 Pro attains the highest objective accuracy (77.3%), while GPT-5.4 leads in semantic quality (4.6/5.0%), suggesting that while some models excel at verification, GPT-5.4 is effective at constructing coherent financial narratives.

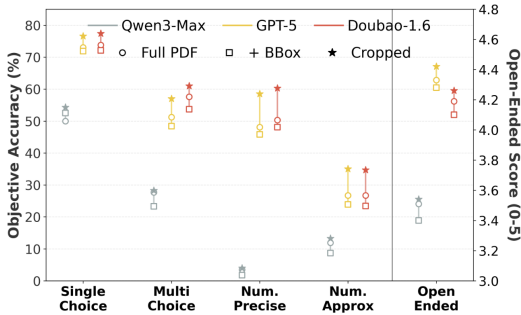


Figure 4: Impact of visual context on objective accuracy. While all models benefit from Cropped Images, Bounding Box annotations degrade the performance of advanced models, due to occlusion or visual noise.

Table 5: Performance stratified by image count per sample. **S**, **M**, and **L** denote 1, 2, and  $\geq 3$  images. Cell format: **Accuracy (%) / Score (0-5)**.

Image Quantity	S (1 img)	M (2 imgs)	L ( $\geq 3$ imgs)
<i>Proprietary Models</i>			
Gemini-3.1 Pro	<b>61.30</b> / 4.09	<b>57.59</b> / 4.10	<b>57.20</b> / 4.16
GPT-5.4	57.32 / <b>4.29</b>	55.40 / <b>4.42</b>	54.75 / <b>4.45</b>
Doubao-1.6 Vision	54.42 / 4.12	53.05 / 4.18	51.89 / 4.23
GPT-5	53.51 / 4.21	50.67 / 4.35	47.50 / 4.37
Gemini-2.5 Pro	46.85 / 4.06	46.21 / 4.02	45.86 / 3.74
Gemini-2.5 Flash	41.59 / 3.80	39.59 / 3.99	36.87 / 3.98
Qwen3-Max	25.21 / 3.49	26.66 / 3.61	24.00 / 3.44
Grok-4.1 Fast	25.76 / 3.19	26.51 / 3.39	24.82 / 3.20
<i>Open-Source Models</i>			
Qwen3-VL-235B	55.09 / 4.14	52.89 / 4.12	51.99 / 4.04
Qwen3-VL-30B	47.72 / 3.72	44.82 / 3.73	41.78 / 3.76
Qwen2.5-VL-72B	32.76 / 3.39	32.91 / 3.29	28.60 / 3.04

**Impact of Model Scale and Evolution.** We use the Qwen series to distinguish the effects of parameter scale from architectural evolution. Within the v3 generation, scaling laws persist; the 235B model (57.73) exceeds the 30B variant (49.81) by approximately 8 points, with the gap widening in calculation-heavy tasks like *Numerical Precise* (52.89% vs 41.16%), suggesting larger capacity supports symbol manipulation. However, inter-generational comparison shows that the Qwen3-VL-30B outperforms the previous Qwen2.5-VL-72B (36.14) by 13.67 points. This gain is evident in fine-grained reasoning, where Qwen3-30B nearly triples the accuracy of Qwen2.5-72B in quantitative calculation (41.16% vs 14.10%), indicating that algorithmic optimizations and multimodal data alignment are more critical for financial reasoning than raw parameter count.

**Impact of Visual Complexity.** We evaluate cross-chart reasoning by categorizing related image samples into Single (S), Dual (M), and Multiple (L,  $\geq 3$ ) groups. As shown in Tab. 5, performance in objective tasks declines as visual dependency increases. For instance, Gemini-3.1 Pro’s accuracy drops from 61.30% (S) to 57.20% (L), indicating difficulty in aggregating information across distributed visual inputs. In contrast, open-ended scores remain stable or improve in complex settings (e.g., GPT-5.4 reaches 4.45 in L). This suggests that while models struggle with precise grounding from multiple charts, they effectively synthesize comprehensive narratives from large visual contexts.

**Performance by Language.** Evaluating across languages (Tabs. 10 and 11 in Appendix C) reveals a *Language Alignment Bias*. Gemini-3.1 Pro leads the English subset (69.56%), while GPT-5.4 excels in Open-Ended tasks (4.58/5.0). Conversely, the Chinese subset exhibits a *Home Field Advantage* for native architectures: Doubao-1.6 Vision remains highly competitive (59.92%), surpassing GPT-5.4 (58.24%) and the open-weights Qwen3-VL-235B (57.65%) due to superior visual grounding in character-dense charts. Remarkably, despite weaker Chinese visual perception, GPT-5.4 retains the highest Open-Ended score across both subsets. This indicates a *decoupling of perception and reasoning*, allowing these models to generate coherent narratives from imperfectly grounded data. **More experiments are in Appendix C.**

#### 4.4 Ablation Studies

**Impact of Visual Context Strategy.** To evaluate the effect of visual context, we conduct an ablation study using three strategies: 1) **Full PDF** (Baseline), using complete document pages; 2) **PDF + BBox**, overlaying red bounding boxes to highlight relevant figures; 3) **Cropped Image**, providing extracted figures at original resolution. As illustrated in Fig. 4, comparing the **Cropped Image** setting with the **Full PDF** baseline reveals a distinct performance gap. Leading models like Doubao-1.6 Vision and GPT-5 achieve their highest accuracy in the Cropped setting (e.g., Doubao-1.6 Vision shows marked gains (over 9%) in Numerical Precise), indicating that locating dispersed visual evidence in multi-page documents remains a major challenge. However, the **PDF + BBox** strategy yields unexpected results. Instead of improving performance, bounding box hints often degrade accuracy for high-capacity models (e.g., GPT-5 scores lower in Multi-Choice with BBox than with

Full PDF). We hypothesize that superimposed frames introduce visual interference or occlusion, disrupting character recognition and feature encoding rather than aiding navigation.

**Robustness to Prompt Variation.** To verify the stability of model rankings against prompt sensitivity, we evaluate performance under prompt variations on a randomly sampled 10% subset of objective questions. We compare the **Standard Prompt** used in our main experiments against a **Chain-of-Thought (CoT) Prompt**, created by appending “Please think step-by-step for the final answer” to the original instructions in order to elicit explicit logical reasoning paths.

Table 6: Stability under prompt variation. Accuracy on a 10% randomly sampled objective subset.

Model	Standard Acc.	CoT Acc.	$\Delta$
<b>Qwen3-VL-235B</b>	54.05%	54.27%	+0.22%
<b>Doubao-1.6</b>	53.71%	53.69%	-0.02%
<b>GPT-5</b>	51.96%	52.11%	+0.15%
<b>Gemini-2.5-Pro</b>	46.54%	46.41%	-0.13%

As detailed in Table 6, performance fluctuations are marginal across all tested models (all absolute  $\Delta < 1\%$ ). More importantly, the relative model rankings remain entirely unchanged. This confirms that the evaluated performance accurately reflects the models’ intrinsic document understanding capabilities, demonstrating robust stability against prompt phrasing variations.

#### 4.5 Case Study

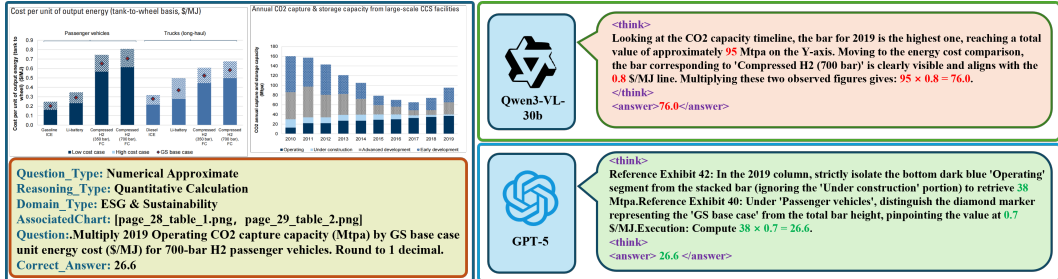


Figure 5: Comparison of GPT-5 and Qwen3-VL-30B on a cross-page calculation task.

Fig. 5 illustrates a cross-page multiplication task testing fine-grained visual grounding, requiring the separation of stacked data series and legend-based markers. Although Qwen3-VL-30b locates the target charts, its reasoning trace indicates susceptibility to **visual saliency bias**. Despite the prompt requesting “Operating” capacity and “GS base case” cost, the model still selects the total bar height (~95) rather than the “Operating” segment (~38) and aligns with the bar’s ceiling (0.8) instead of the diamond marker (0.7). Therefore, disregarding legend constraints results in a calculation error ( $95 \times 0.8 = 76.0$ ) relative to the ground truth (26.6). In contrast, GPT-5 exhibits effective **legend-to-pixel alignment**, filtering visual noise (e.g., the “Under construction” stack) to locate data points defined by semantic constraints. **More case studies are in Appendix D.**

### 5 Conclusion

In this paper, we introduce FINDOCMRE, a benchmark designed to advance multimodal financial reasoning from isolated charts to document-level multimodal contexts. Using a pipeline that combines visual-centric generation with expert verification, we compiled 12,207 samples from 2,878 financial reports to identify limitations in current SOTA models. Extensive experiments across all eleven evaluated LMMs reveal a “Analyst-Calculator Dichotomy”: while models show strong semantic synthesis, they lack the numerical precision and cross-page visual grounding required for professional analysis. Ablation studies identify an “visual grounding bottleneck”, confirming that the primary challenge lies in aggregating fragmented evidence from complex reports. Based on these findings, future research should shift from end-to-end generation to agentic architectures capable of long-context grounding and tool-assisted calculation, fostering trustworthy financial AI.

## 6 Limitations

Despite the scale of FINDOCMRE, we acknowledge two primary limitations. First, our evaluation is restricted to zero-shot settings. Since each sample requires including an entire PDF document and its question and answer, providing few-shot exemplars would produce excessive prompt lengths that exceed most models’ context window constraints. Second, our single-turn QA formulation simplifies complex, iterative professional financial workflows. Our benchmark currently focuses on foundational reasoning units, reserving multi-turn, agentic evaluations for subsequent research.

## References

- [1] Comanici, G., E. Bieber, M. Schaekermann, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.
- [2] Bai, S., K. Chen, X. Liu, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [3] Chen, Z., W. Chen, C. Smiley, et al. Finqa: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711. 2021.
- [4] Ding, J., C. Ding, W. Pang, et al. Cnfinbench: A benchmark for safety and compliance of large language models in finance. *arXiv preprint arXiv:2512.09506*, 2025.
- [5] Xue, S., X. Li, F. Zhou, et al. Famma: A benchmark for financial domain multilingual multimodal question answering. *arXiv preprint arXiv:2410.04526*, 2024.
- [6] Karamcheti, S., S. Nair, A. Balakrishna, et al. Prismatic vlms: Investigating the design space of visually-conditioned language models. In *Forty-first International Conference on Machine Learning*. 2024.
- [7] Huang, W., H. Liu, M. Guo, et al. Visual hallucinations of multi-modal large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9614–9631. 2024.
- [8] Achiam, J., S. Adler, S. Agarwal, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [9] Anthropic. The claude 3 model family: Opus, sonnet, haiku, 2024. Anthropic Model Card.
- [10] Zeng, A., X. Lv, Q. Zheng, et al. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*, 2025.
- [11] Guo, D., F. Wu, F. Zhu, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025.
- [12] Liu, H., C. Li, Q. Wu, et al. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.
- [13] Zhu, J., W. Wang, Z. Chen, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.
- [14] Fu, C., P. Chen, Y. Shen, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.
- [15] Wang, K., J. Pan, W. Shi, et al. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024.
- [16] Zhang, R., D. Jiang, Y. Zhang, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer, 2024.
- [17] Yue, X., Y. Ni, K. Zhang, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567. 2024.
- [18] Lu, P., H. Bansal, T. Xia, et al. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.
- [19] Masry, A., X. L. Do, J. Q. Tan, et al. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279. 2022.

- [20] Chen, L., J. Li, X. Dong, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.
- [21] Liu, Y., H. Duan, Y. Zhang, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024.
- [22] Xu, Z., S. Du, Y. Qi, et al. Chartbench: A benchmark for complex visual reasoning in charts. *arXiv preprint arXiv:2312.15915*, 2023.
- [23] Liu, F., X. Wang, W. Yao, et al. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1287–1310. 2024.
- [24] Mathew, M., D. Karatzas, C. Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209. 2021.
- [25] Shah, R. S., K. Chawla, D. Eidnani, et al. When flue meets flang: Benchmarks and large pre-trained language model for financial domain. In *EMNLP*. 2022.
- [26] Lu, D., H. Wu, J. Liang, et al. Bbt-fin: Comprehensive construction of chinese financial domain pre-trained language model, corpus and benchmark. *arXiv preprint arXiv:2302.09432*, 2023.
- [27] Li, H., Y. Cao, Y. Yu, et al. Investorbench: A benchmark for financial decision-making tasks with llm-based agent. *arXiv preprint arXiv:2412.18174*, 2024.
- [28] Islam, P., A. Kannappan, D. Kiela, et al. Financebench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*, 2023.
- [29] Zhang, L., W. Cai, Z. Liu, et al. Fineval: A chinese financial domain knowledge evaluation benchmark for large language models. *arXiv preprint arXiv:2308.09975*, 2023.
- [30] Xie, Q., W. Han, Z. Chen, et al. Finben: A holistic financial benchmark for large language models. In *NeurIPS*, pages 95716–95743. 2024.
- [31] Lei, Y., J. Li, D. Cheng, et al. Cfbenchmark: Chinese financial assistant benchmark for large language model. *arXiv preprint arXiv:2311.05812*, 2023.
- [32] Xu, L., L. Zhu, Y. Wu, et al. Superclue-fin: Graded fine-grained analysis of chinese llms on diverse financial tasks and applications. *arXiv preprint arXiv:2404.19063*, 2024.
- [33] OpenCompass Project. Openfindata: The open-source financial evaluation dataset for large language models, 2023. Available at <https://github.com/open-compass/OpenFinData>.
- [34] Li, J., Y. Zhu, D. Cheng, et al. Cfbenchmark-mm: Chinese financial assistant benchmark for multimodal large language model. *arXiv preprint arXiv:2506.13055*, 2025.
- [35] Gan, Z., D. Zhang, H. Li, et al. Mmc-finance: A multimodal finance benchmark for expert-level understanding and reasoning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 12867–12874. 2025.
- [36] Luo, J., Z. Kou, L. Yang, et al. Finmme: Benchmark dataset for financial multi-modal reasoning evaluation. *arXiv preprint arXiv:2505.24714*, 2025.
- [37] Deng, S., H. Peng, J. Xu, et al. Finmr: A knowledge-intensive multimodal benchmark for advanced financial reasoning. In *Proceedings of the 6th ACM International Conference on AI in Finance*, pages 168–176. 2025.
- [38] Zhu, Y., Y. Jiang, Z. Xu, et al. From comprehension to reasoning: A hierarchical benchmark for automated financial research reporting. *arXiv preprint arXiv:2603.19254*, 2026.
- [39] Hu, Y., Y. Li, P. Liu, et al. Fintsb: A comprehensive and practical benchmark for financial time series forecasting. *arXiv preprint arXiv:2502.18834*, 2025.
- [40] Tang, Z., E. Haihong, R. Li, et al. Finmmdocr: Benchmarking financial multimodal reasoning with scenario awareness, document understanding, and multi-step computation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, pages 25858–25866. 2026.
- [41] Team, G.-. Gpt-5 is here, 2025.

- [42] Yang, A., A. Li, B. Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.
- [43] Team, G. Grok 4.1 fast and agent tools api, 2025.
- [44] Papineni, K., S. Roukos, T. Ward, et al. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318. 2002.
- [45] Li, D., B. Jiang, L. Huang, et al. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791. 2025.
- [46] Anderson, L. W., D. R. Krathwohl. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition*. Addison Wesley Longman, Inc., 2001.

Sec. A elaborates on the construction of FINDOCMRE, documenting prompts and settings. Sec. B outlines the evaluation protocol, including configuration and scoring prompts. Sec. C presents additional analysis regarding language, visual resolution, and document length. Sec. D examines two additional cases illustrating visual bias and domain-specific capabilities of LMMs.

## A FINDOCMRE Construction

### A.1 Data Preparation

To construct a benchmark reflecting real-world financial analysis, we collect a corpus of 2,878 financial PDFs from official sources. We use a chart-centric extraction framework based on PyMuPDF and pdfplumber to parse both textual and embedded visual content. Extracted text is serialized into JSON format to preserve document structure, while embedded charts are archived by source document. Each image is assigned a filename reflecting its page location and extraction sequence (*e.g.*, `page_1_image_1.png`) as the unique `chart_id`.

The raw extraction output contains noise, including company logos, icons, and fragmented artifacts such as decorative headers or dividers. Furthermore, the initial extraction includes incomplete figures and text blocks incorrectly identified as images. To isolate analytically valuable charts, we apply a Joint-Rule Filtering Mechanism:

1. We apply geometric constraints to eliminate layout artifacts, discarding images with low resolutions or extreme aspect ratios.
2. We integrate image similarity computation with OCR to remove repetitive non-data elements. This hybrid approach enables the deduplication of visual content and the removal of uninformative icons or corporate logos.
3. We perform page-level textual indexing verification. By scanning the text of corresponding pages, we retain only images explicitly referenced by narrative markers (*e.g.*, “Figure x”), ensuring that all “Cleaned Figures” are linked to the document context.

### A.2 Visual-Centric Generation

To reduce textual bias and ground reasoning in visual evidence, we use a visual-centric generation strategy with [GEMINI-2.5-PRO]. Instead of full PDF pages, we input the sequence of “Cleaned Figures” from the data preparation phase, excluding surrounding textual context. Guided by a structured prompt, the model generates a standardized metadata object for each sample, containing eight specific fields: *Question*, *Options* (empty for non-choice tasks), *Answer*, *Question Type*, *Domain Type*, *Reasoning Type*, *Associated Chart* (linking to specific chart IDs), and *Reasoning Trace*. This strict schema provides broad coverage of financial reasoning dimensions. The prompt used for this generation process is presented in Tab. 7.

### A.3 Expert Verification

To reduce hallucination risks associated with visual-centric generation and ensure data quality, we employ human experts to verify each sample. Three senior financial researchers with over three years of professional experience in securities firms verify each document and generated sample. All the experts are Master in Economics or Finance from QS Top-100 universities, who have received compensation exceeding local minimum wage (\$50/hour). Guided by the annotation protocols in Tab. 8, experts audit samples by cross-referencing generated reasoning traces with source charts.

### A.4 Reasoning Type and Financial Industries

To assess the viability of LMMs as professional financial agents, we established a multi-dimensional taxonomy evaluating **Cognitive Depth** (Reasoning Types) and **Knowledge Breadth** (Financial Domains). This framework differentiates elementary visual reasoning from analytical logic.

#### A.4.1 Reasoning Types

Drawing from Bloom’s Taxonomy [46], we classify reasoning skills into five levels, ranging from reasoning to synthesis. This hierarchy enables the identification of specific deficiencies in multimodal models.

- **Quantitative Calculation (Computation):** Fundamental to financial analysis, this category evaluates the capacity for precise multi-step arithmetic (*e.g.*, CAGR, margins) using extracted visual data, verifying computational reliability.
- **Pattern Identification (Perception):** This assesses the recognition of visual trends, volatility, and anomalies (*e.g.*, detecting a “double top” or sudden revenue drop), a capability exclusive to the visual modality.
- **Comparative Analysis (Relation):** Reflecting the benchmarking workflow, this requires aligning and contrasting data points across entities (*e.g.*, “Company A vs. Company B”) or time periods (YoY growth) to test relational reasoning.
- **Logical Deduction (Inference):** This evaluates the application of implicit financial rules (*e.g.*, “If assets increased while liabilities remained flat, equity must have increased”), evaluating adherence to domain logic.
- **Comprehensive Synthesis (Integration):** This requires aggregating fragmented information from heterogeneous charts (*e.g.*, combining macro GDP with micro revenue) to form a unified assessment, similar to investment memo creation.

#### A.4.2 Financial Domains

To guarantee benchmark diversity and prevent overfitting to specific document types (*e.g.*, Annual Reports), we select 12 domains covering the “Micro-Meso-Macro” spectrum of the financial ecosystem.

1. **Micro-Level (Corporate Finance):** Concentrating on firm-specific analysis.
  - *Financial Statement Analysis:* Fundamental accounting evaluation (*e.g.*, Income/Balance Sheets).
  - *Valuation:* Asset pricing methodologies (DCF, P/E ratios).
  - *M&A and Capital Structure:* Corporate transactions and capital composition (*e.g.*, Debt/Equity, Mergers).
  - *Business Segments:* Disaggregated analysis of revenue sources by region or product.
  - *Risk Management:* Assessment of exposure and stability (*e.g.*, VaR, Liquidity).
2. **Meso-Level (Market & Investment):** Analysis of portfolios and industrial sectors.
  - *Industry & Competitive Analysis:* Evaluation of market share and strategic positioning (*e.g.*, SWOT).
  - *Portfolio Management:* Asset allocation and diversification strategies.
  - *Fund Performance:* Attribution analysis (*e.g.*, Alpha, Beta) of investment funds.
  - *Fintech & Innovation:* Emerging trends in digital finance (*e.g.*, Blockchain, AI).
3. **Macro-Level (Global Economy):** Examination of the global economic environment.
  - *Macroeconomic Indicators:* Primary economic drivers (*e.g.*, GDP, Inflation, Rates).
  - *Macro Markets:* Commodity trends (*e.g.*, Gold/Oil) and foreign exchange markets.
  - *ESG & Sustainability:* Non-financial metrics regarding sustainability and governance (*e.g.*, Carbon, Compliance).

This domain diversity enables FINDOCMRE to evaluate LMMs beyond basic OCR capabilities, testing the interpretation of specialized semantics across the financial sector.

---

## System Prompt for Financial Reasoning Generation

---

### # Role

You are a senior financial analyst and an expert in designing challenging logical reasoning problems. Your mission is to create a challenging benchmark capable of effectively distinguishing the multimodal capabilities of current and future State-of-the-Art (SOTA) models.

### # Core Philosophy

- 1. Form Serves Function:** Strictly adhere to question formats: `single_choice`, `multiple_choice`, `numerical_precise` (Calculations must use **explicit values** found directly on charts/tables without ambiguity), `numerical_approximate` (Reasoning requires **visual estimation** *e.g.*, reading axis height where precise labels are absent), and `open_ended` (Pure text answer; strictly forbidden from containing `chart_id` or filenames).
- 2. Evaluate Reasoning, Not Memorization:** The primary objective is evaluating deep analytical, reasoning, and synthesis skills.
- 3. Blind Stem Principle:** The stem is strictly forbidden from mentioning `chart_id` so users emulate real-world blind queries.
- 4. Promote Comprehensive Analysis:** Encourage the design of complex questions that require integrating partial information from multiple distinct charts.
- 5. Information Silo Principle:** All charts must be treated as originating from a fictional, non-public context; do not use external knowledge.
- 6. Abstract Time Principle:** Avoid real dates. Use relative years (*e.g.*, 'Year 1', 'Year 2'). If this conflicts with 'Event Anchoring', the latter takes precedence.
- 7. Quantitative Anchor Principle:** Answers must be uniquely determined by specific information in the charts, avoiding ambiguous estimation scenarios.
- 8. Event Anchoring Principle:** Prioritize using specific events (*e.g.*, "when revenue peaked") to lock time points across multiple charts.
- 9. Context-Free Stem Principle:** The stem must be clear and unambiguous, ensuring solvability whether the input is the full PDF or filtered images.

### # Classification Tags

When generating each question object, you must also add the following two classification tags. The definitions are strict:

#### 1. 5 `reasoning_type` (**Reasoning Type**): [Must choose one]

- **Quantitative Calculation:** The core task is to perform multi-step arithmetic or algebraic operations to arrive at a precise or approximate numerical value.
- **Pattern Identification:** The core task is to identify and locate a specific data point, trend, pattern, or anomaly (*e.g.*, 'peak', 'fastest growth').
- **Comparative Analysis:** The core task is to compare two or more entities or data points to determine their relationship (*e.g.*, 'which is higher', 'are they correlated').
- **Logical Deduction:** The core task is to apply an implicit rule, constraint, or financial identity to deduce a necessary conclusion (*e.g.*, '...must be true?').
- **Comprehensive Synthesis:** The core task is to integrate information from multiple heterogeneous charts to form a high-level, coherent summary, judgment, or evaluation.

#### 2. 12 `domain` (**Domain Tag**): [Must choose one]

- **Financial Statement Analysis:** Focuses on items and their ratios from financial statements like the Income Statement, Balance Sheet, and Cash Flow Statement.
- **Valuation:** Focuses on calculating or comparing a company's value (*e.g.*, DCF, P/E, EV/EBITDA, comparable company analysis, etc.).
- **M&A and Capital Structure:** Focuses on mergers, acquisitions, divestitures, debt, equity, leverage, etc.

- **Industry & Competitive Analysis:** Focuses on market share, industry trends, competitive landscape, SWOT analysis, Porter's Five Forces, etc.
- **Business Segments:** Focuses only on unique metrics or drivers for specific company business segments (*e.g.*, by product, by region), etc.
- **Portfolio Management:** Focuses on asset allocation, portfolio composition, diversification, benchmark comparison, etc.
- **Fund Performance & Attribution:** Focuses on Alpha, Beta, Sharpe Ratio, tracking error, performance attribution, etc.
- **Risk Management:** Focuses on VaR (Value at Risk), credit risk, liquidity risk, operational risk, stress testing, etc.
- **Macroeconomic Indicators:** Focuses on GDP, inflation, interest rates, employment, consumer confidence indices, etc.
- **Macro Markets:** Focuses on the market dynamics of commodities (*e.g.*, oil, natural gas, gold, etc.) and foreign exchange (*e.g.*, currency pairs, exchange rate changes, etc.).
- **ESG & Sustainability:** Focuses on environmental, social, and governance metrics, such as carbon emissions, diversity, governance structure, etc.
- **Fintech & Innovation:** Focuses on digital payments, blockchain, AI in finance, RegTech, etc.

#### # Task Instructions

Generate a JSON list that strictly adheres to the defined schema. **Special Rules:** 1. Return empty [] if review fails. 2. Balance categories.

#### # Step-by-Step Instructions

**Step 1: In-depth Analysis:** Identify chart types, units, and axis meanings. Pay special attention to **real timestamps** and prepare to abstract them (*e.g.*, converting dates to "Year 1"). Understand connections between charts.

**Step 2: Mine Scenarios:** Prioritize "**Event Anchoring**" (using events like "revenue peak" to lock time points across charts), visual geometric features (slopes, intersections), and scenarios requiring active filtering of key info from multiple charts.

**Step 3: Draft Questions:** Select format. **Hide calculation paths** (ask for the final result, not the steps). Ensure **no chart\_id in stems**. Design common-error distractors. Sanitize open\_ended answers (remove filenames). For numerical tasks, ensure multi-step calculation is required.

**Step 3.5: Adversarial Self-Review:** Perform a "**zero-knowledge test**". Ask: "*Based only on the text, can I deduce the answer using common sense without the chart?*" If **YES** (leaks knowledge), discard the question. If **NO**, it passes.

**Step 4: Generate JSON:** Add tags. Construct the `reasoning_trace` to serve as a self-contained proof, explicitly stating which values are estimated vs. precise and explaining the derivation logic step-by-step.

#### # JSON Output Example (1 Shot)

```

1 [
2 {
3   "question_id": "auto_generated_unique_id_1",
4   "source_chart_ids": [ "page_6_image_1.png", "page_5_table_2.png" ],
5   "question_format": "numerical_precise",
6   "reasoning_type": "Quantitative Calculation",
7   "domain": "Financial Statement Analysis",
8   "question_content": {
9     "stem": "Combining the sales trend chart and the cost structure table,
10    in the year when sales revenue reached its historical peak, what was
11    the company's raw material cost as a percentage of total cost? Present
12    the answer as a percentage, rounded to one decimal place.",
13    "options": {}
14  },
15  "answer": "42.5",
16  "reasoning_trace": [
17    "Self-Review: Passed. 'Sales revenue reached its historical peak' is
18    an event requiring visual location. Cost percentage requires multi-step
19    calculation.",
20    "Step 1: Identify the 'Sales Trend Chart' (page_6_image_1.png) and '
21    Cost Structure Table' (page_5_table_2.png).",
22    "Step 2: [Event Anchoring] In Sales Trend Chart, visually inspect the
23    curve to find its highest point, determining the peak occurred in Year
24    5.",
25    "Step 3: [Cross-Chart Query] Using Year 5 as anchor, query Cost
26    Structure Table: Raw Material Cost = $3,400M, Total Cost = $8,000M.",
27    "Step 4: [Precise Calculation] ( $\$3,400M / \$8,000M * 100\% = 42.5\%$ ).",
28  ]
29 }
30 ]

```

Table 7: The full system prompt used for financial reasoning question generation.

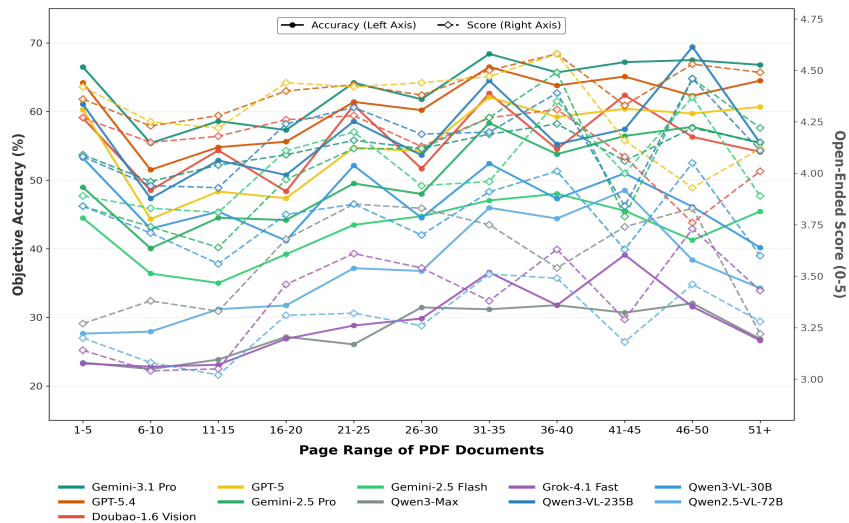


Figure 6: Impact of PDF Length

---

## Expert Verification Guidelines

---

**Role:** You are a senior financial researcher acting as a data auditor. Your goal is to guarantee the benchmark’s "Gold Standard" quality.

**Input Data Package:**

For each verification task, you will be presented with: (1) **Source Context:** The set of extracted Cleaned Charts from the financial report. (2) **Question Metadata:** The generated Stem, Options (for MCQs), and Reference Answer. (3) **Logic Chain:** The step-by-step Reasoning Trace derived by the model. (4) **Grounding:** The Associated Chart IDs indicating visual evidence sources.

---

**Objective:** Perform a binary validation (RETAIN vs. DISCARD). If a sample contains *any* flaw, it must be discarded immediately.

**Phase 1: Visual Grounding Audit (Critical Check)**

- **Chart Alignment:** Verify that the Associated Chart IDs point *exclusively* and *correctly* to the images containing the evidence. If irrelevant charts are cited or key charts are missing → **DISCARD**.
- **Textual Isolation:** Ensure the question is solvable *solely* using the provided visual charts. If the reasoning relies on external knowledge or non-visual PDF text (not present in the input charts) → **DISCARD**.

**Phase 2: Content & Logic Audit**

- **Question Quality:** Is the phrasing clear and unambiguous? If the intent is vague or grammar is poor → **DISCARD**.
- **Factual Correctness:** Is the Reference Answer numerically precise and factually correct based strictly on the chart data? If any calculation error or data hallucination exists → **DISCARD**.
- **Option Quality (for MCQs):** Are the distractors plausible? If options are nonsensical or allow guessing without reasoning → **DISCARD**.
- **Trace Logic:** Does the Reasoning Trace provide a valid, step-by-step derivation? If the logic leaps or cites non-existent values → **DISCARD**.

**Phase 3: Final Decision (Unanimous Consensus Rule)**

- **DISCARD (Action):** Select this if the sample triggers **ANY** violation from Phase 1 or 2.
- **RETAIN (Action):** Select this *only* if the sample is visually grounded, logically flawless, and explicitly verified by your professional judgment.

---

Table 8: Guidelines provided to human experts for verifying the quality of the generated benchmark data.

## B Evaluation Protocol

### B.1 Evaluation Details

In this section, we detail the inference configuration, data input strategy, and scoring protocols.

#### B.1.1 Inference and Input Settings

We evaluate all models in a **Zero-Shot** setting using their official APIs with the inference prompt shown in Tab. 14. The detailed configurations:

- **Visual Input Strategy:** To standardize inputs, we convert all visual content (Full PDF, Bounding Box, or Cropped Charts) into Image Sequences. Considering existing LMMs cannot directly process PDF files, we render visual contexts as high-resolution images.
- **Batching:** We query multiple questions associated with the same document in a single API session. For documents with numerous questions (exceeding 10), we split them into separate batches. For each batch, we re-upload the associated images to maintain consistent visual context.
- **Hyper-parameter:** We use default inference parameters (*e.g.*, `temperature=0`) to reflect baseline capabilities. We fix the random seed at 1234 to support reproducibility. We enable “thinking modes” for models supporting reasoning traces.
- **Error Handling:** We implement a retry mechanism for standard API errors such as connection timeouts. For fatal errors (*e.g.*, input limits or safety filters), we stop retries and record the answer as `null` (treated as incorrect).

#### B.1.2 Scoring Protocol

Given that models output answers directly, we use Regular Expressions for standardization. The scoring protocols are defined as follows:

- **Single Choice: Exact Letter Match.** We compare the selected option with the ground truth.
- **Multiple Choice: Strict Set Match.** The predicted letter set must be identical to the ground truth, without partial credit.
- **Numerical Precise: Exact Equality.** The calculation correspond exactly to the ground truth.
- **Numerical Approximate: Conditional Tolerance.** We apply a two-step validation:
  - (i) If the reference value  $V_{ref} = 0$ , the prediction is correct only if  $V_{pred} = 0$ .
  - (ii) If  $V_{ref} \neq 0$ , the prediction is correct if it falls within a 5% relative error margin:  $|V_{pred} - V_{ref}|/|V_{ref}| \leq 0.05$ .
- **Open Ended: LLM-as-a-Judge.** Three distinct LMMs evaluate the response with reference (prompt in Tab. 15), assigning an **integer score** from [0, 5].

### B.2 Evaluation Bias

To address concerns regarding evaluator subjectivity, we conduct a detailed bias and alignment analysis. Notably, approximately 80% of the tasks in FinDocMRE (*e.g.*, Multiple-Choice, Numerical) rely on deterministic metrics immune to subjectivity. For the remaining 20% open-ended questions, we apply a reference-guided LLM-as-a-Judge approach. To mitigate single-model bias, our protocol ensembles three distinct models (Gemini-2.5-Flash, GPT-4o, and Qwen3-Max) to independently score responses, taking their average as the final metric. To validate this approach and assess its alignment with human experts, we conducted a comparative analysis on a sampled subset of 250 open-ended questions ( $\sim 10\%$  of the open-ended test set). We compared the individual LMM scores against a **Human Expert** baseline, derived from the averaged independent ratings of two professional financial researchers. The evaluation results across five representative target models are presented in Table 9.

As shown in Table 9, while LMM evaluators exhibit slight leniency in absolute scoring (typically +0.2 to +0.6 higher than human experts), their relative rankings maintain alignment with human judgments. Across all evaluated target models, the performance hierarchy remains identical (1st to 5th) with minimal variance among the automated judges. This demonstrates that despite minor scaling differences in absolute scores, the ensembled LMMs reliably discern response quality, confirming that our evaluation framework is robust, objective, and aligned with professional standards.

Table 9: Bias and alignment analysis on a sampled subset (250 questions). Results are **Score / Rank**. Target models evaluated: GPT-5, Doubao-1.6 Vision (DB-1.6), Qwen3-vl-235B (QW-235B), Gemini-2.5-Pro (Gem-Pro), and Qwen3-vl-30B (QW-30B).

Evaluator	GPT-5	DB-1.6	QW-235B	Gem-Pro	QW-30B
<b>Gemini-2.5-Flash</b>	4.32 / 1	4.22 / 2	4.09 / 3	3.91 / 4	3.75 / 5
<b>GPT-4o</b>	4.34 / 1	4.23 / 2	4.10 / 3	3.92 / 4	3.76 / 5
<b>Qwen3-Max</b>	4.36 / 1	4.25 / 2	4.12 / 3	3.95 / 4	3.79 / 5
<b>Human Expert</b>	4.11 / 1	3.96 / 2	3.71 / 3	3.38 / 4	3.19 / 5

Table 10: Evaluation results for the **English subset**. Abbreviations denote question types: Single-Choice (**Single**), Multiple-Choice (**Multi**), Numerical-Precise (**N(p)**), Numerical-Approximate (**N(a)**), and Open-Ended (**Open**). **Overall** score is the weighted average.

Model	Single	Multi	N(p)	N(a)	Open	Overall
<i>Proprietary Models</i>						
Gemini-3.1 Pro	82.32	65.50	69.10	43.90	4.35 (87.0)	69.56
GPT-5.4	79.77	62.23	65.15	37.07	4.58 (91.6)	67.16
GPT-5	76.32	53.93	49.56	27.61	4.37 (87.4)	58.96
Doubao-1.6 Vision	73.77	53.28	45.51	24.88	4.15 (83.0)	56.09
Gemini-2.5 Pro	76.01	48.25	38.30	20.20	4.26 (85.2)	53.59
Gemini-2.5 Flash	68.92	46.94	28.13	21.17	4.03 (80.6)	49.15
Qwen3-Max	55.58	30.79	5.73	14.44	3.76 (75.2)	36.35
Grok-4.1 Fast	62.28	27.73	6.12	11.71	3.65 (73.0)	36.17
<i>Open-Source Models</i>						
Qwen3-VL-235B	76.83	47.60	50.15	27.22	4.29 (85.8)	57.52
Qwen3-VL-30B	70.33	29.69	36.62	19.41	3.75 (75.0)	46.21
Qwen2.5-VL-72B	66.50	36.46	21.42	15.02	3.49 (69.8)	41.84

## C Extra Experimental Results

### C.1 Performance by Language

### C.2 Evaluation Prompts

The full evaluation prompts used in our experiments are provided in Table 14 and Table 15.

### C.3 Impact of Visual Resolution.

We examine model sensitivity to image quality by varying resolution from 60 to 200 DPI. As illustrated in Fig. 7, a performance saturation trend is observed. For advanced models (Doubao-1.6 Vision, GPT-5), accuracy increases between 60 and 110 DPI as details become legible, stabilizing beyond 110 DPI. This supports our selection of 110 DPI for main experiments, balancing visual clarity with token efficiency. In contrast, lower-performing models (*e.g.*, Qwen3-Max) show minimal gains, suggesting limitations in reasoning rather than visual perception.

### C.4 Impact of Document Length.

Increasing source document length typically introduces visual noise, often causing performance degradation known as the “Lost-in-the-Middle” phenomenon. However, as shown in Fig. 6, results indicate a non-monotonic relationship between PDF page count and model performance, where accuracy fluctuates rather than declining consistently. We attribute this behavior to three factors. First, page count correlates with document standardization; longer documents are predominantly official Periodic Reports (*e.g.*, Annual Reports) with professional typesetting and vector charts. In contrast, shorter documents show greater layout variance, increasing parsing difficulty. Second, the Joint-Rule Filtering Mechanism stabilizes performance by reducing noise. By removing decorative elements and text-only pages, the mechanism maintains high information density, preventing the processed visual context from expanding linearly with raw page count. Finally, financial reasoning displays semantic clustering, where interdependent charts required for a query typically appear within the

Table 11: Evaluation results for the **Chinese subset**. Abbreviations are question types: Single-Choice (**Single**), Multiple-Choice (**Multi**), Numerical-Precise (**N(p)**), Numerical-Approximate (**N(a)**), and Open-Ended (**Open**). **Overall** score are the weighted average.

Model	Single	Multi	N(p)	N(a)	Open	Overall
<i>Proprietary Models</i>						
Gemini-3.1 Pro	76.44	54.51	58.91	33.45	3.95 (79.0)	60.46
Doubao-1.6 Vision	73.77	59.69	53.81	27.92	4.22 (84.4)	59.92
GPT-5.4	72.92	54.09	51.36	27.01	4.29 (85.8)	58.24
GPT-5	70.25	49.84	46.98	26.10	4.30 (86.0)	55.83
Gemini-2.5 Pro	69.93	43.94	34.70	21.06	3.65 (73.0)	48.53
Gemini-2.5 Flash	60.90	39.17	22.99	20.78	3.88 (77.6)	44.29
Qwen3-Max	45.25	26.22	1.80	10.01	3.32 (66.4)	29.94
Grok-4.1 Fast	47.01	21.04	2.01	8.61	2.98 (59.6)	27.65
<i>Open-Source Models</i>						
Qwen3-VL-235B	73.93	55.03	54.89	25.61	3.94 (78.8)	57.65
Qwen3-VL-30B	67.79	49.95	44.47	21.62	3.73 (74.6)	51.69
Qwen2.5-VL-72B	53.53	25.39	8.76	14.00	2.98 (59.6)	32.26

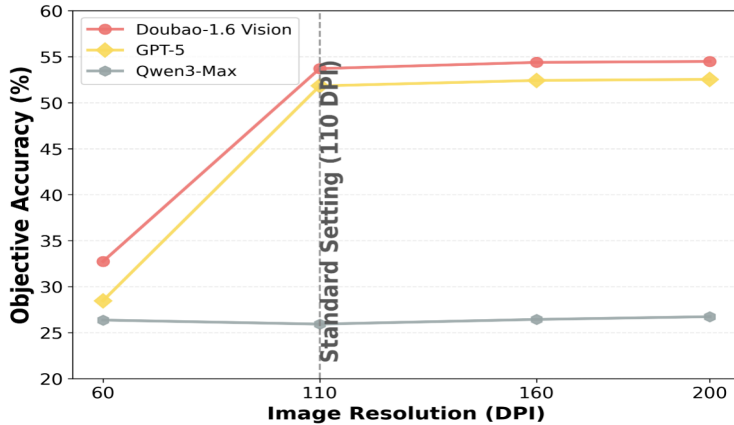


Figure 7: **Impact of Image Resolution (DPI)**. Advanced models (*e.g.*, Doubao, GPT-5) show **minimal gains** beyond 110 DPI, supporting the experimental setting. Other models (*e.g.*, Qwen3-Max) remain limited by reasoning capabilities regardless of resolution.

same thematic section (*e.g.*, adjacent pages in the MD&A chapter). These characteristics enables models to process local information clusters without grounding disparate data across full document.

## D Extra Case Study

### D.1 Visual Bias in Comparative Analysis

In Fig. 8, we present a task requiring cross-referencing charts to identify structural descriptors. GPT-5 correctly selects options A and B by calculating the exact parity (14%) between sector weights, confirming “Consumer-Tech Convergence.” In contrast, Qwen3-VL-30B exhibits *Visual Impression Bias*. Although it correctly identifies the timeframe, it incorrectly selects “Tech Weight Expansion” (Option D), citing the visual addition of new rows as evidence of growth. The model confuses *visual layout expansion* with value increase, overlooking the numerical decrease from 18% to 14%. This suggests that smaller models may prioritize visual heuristics (*e.g.*, “more rows implies growth”) over numerical extraction when visual signals conflict with quantitative data.

### D.2 Domain Cognition in Strategic Synthesis

In Fig. 9, we evaluate strategic synthesis capabilities using an LLM-as-a-Judge scoring system. Qwen3-VL-30B scores 3; while factually accurate, it provides literal chart descriptions, limiting

implications to generic consumer concepts like data plan sales. In contrast, GPT-5 scores 5 by demonstrating domain alignment. It infers second-order economic effects, connecting low data costs to “digital-first fintech” and the “financialization of household savings” concepts derived logically rather than visually. This comparison illustrates that while visual perception is comparable, significant disparities remain in the ability to contextualize evidence within a professional framework.

## E Broader Impacts

Our work introduces FinDocMRE to advance document-level reasoning in financial AI, which can enhance the efficiency and transparency of financial information processing. However, we emphasize that this benchmark is for research purposes only and should not be used to generate automated investment advice or replace professional human judgment in financial decision-making.

Table 12: Performance breakdown (Part 1/2): Business Segments to Industry Analysis. Metrics: **Accuracy (%) / Score (0-5)**.

Model	Bus. Seg.	ESG	Fin. Stat.	Fintech	Fund Perf.	Ind. Anal.
<i>Proprietary Models</i>						
Gemini-3.1 Pro	62.9 / 4.1	59.9 / 4.2	63.2 / 4.0	66.1 / 4.3	58.2 / 4.1	62.2 / 4.2
GPT-5.4	60.6 / 4.3	58.3 / 4.4	58.2 / 4.4	63.9 / 4.4	55.3 / 4.2	59.8 / 4.5
Doubao-1.6 Vision	59.8 / 4.2	52.6 / 4.2	55.6 / 3.9	59.2 / 4.5	53.0 / 3.9	57.6 / 4.4
GPT-5	56.8 / 4.3	56.0 / 4.3	50.9 / 4.2	61.2 / 4.4	44.9 / 3.8	55.9 / 4.5
Gemini-2.5 Pro	52.2 / 3.9	49.1 / 4.2	38.0 / 3.4	59.0 / 4.3	37.2 / 3.1	52.5 / 4.0
Gemini-2.5 Flash	45.8 / 3.8	38.3 / 4.0	28.8 / 3.6	51.8 / 4.4	26.2 / 3.2	44.0 / 4.2
Qwen3-Max	32.6 / 3.4	24.1 / 3.7	9.4 / 2.7	39.1 / 3.9	12.1 / 2.5	27.7 / 3.6
Grok-4.1 Fast	31.1 / 3.3	26.5 / 3.4	10.1 / 2.4	39.1 / 4.0	14.1 / 2.1	26.8 / 3.4
<i>Open-Source Models</i>						
Qwen3-VL-235B	58.9 / 3.9	55.9 / 4.2	51.8 / 3.7	62.0 / 4.5	52.8 / 3.7	57.9 / 4.3
Qwen3-VL-30B	51.5 / 3.6	42.7 / 3.8	47.8 / 3.6	55.0 / 4.1	45.7 / 3.3	48.1 / 4.1
Qwen2.5-VL-72B	38.6 / 3.4	33.3 / 3.3	15.4 / 2.7	47.5 / 3.6	19.1 / 2.7	37.0 / 3.3

Table 13: Performance breakdown (Part 2/2): M&A to Valuation. Metrics: **Accuracy (%) / Score (0-5)**.

Model	M&A	Macro Mkt.	Macro Ind.	Port. Mgmt.	Risk Mgmt.	Valuation
<i>Proprietary Models</i>						
Gemini-3.1 Pro	66.3 / 4.1	50.2 / 4.0	49.1 / 4.2	58.2 / 4.0	68.4 / 4.2	75.5 / 4.2
GPT-5.4	60.5 / 4.4	47.1 / 4.4	46.8 / 4.5	55.9 / 4.3	66.1 / 4.5	72.3 / 4.5
Doubao-1.6 Vision	57.4 / 4.4	45.3 / 4.0	41.5 / 4.3	52.4 / 4.0	62.3 / 4.1	68.7 / 4.2
GPT-5	50.0 / 4.3	43.5 / 4.2	44.5 / 4.3	43.8 / 4.1	63.0 / 4.3	66.3 / 4.5
Gemini-2.5 Pro	52.1 / 4.0	39.5 / 3.9	39.4 / 4.0	41.5 / 3.6	55.9 / 4.0	58.2 / 4.0
Gemini-2.5 Flash	40.5 / 4.1	38.2 / 3.8	37.4 / 4.1	36.9 / 3.6	50.0 / 4.0	51.7 / 3.7
Qwen3-Max	15.8 / 3.8	24.0 / 3.5	27.0 / 3.8	23.6 / 3.1	35.1 / 3.7	24.5 / 3.5
Grok-4.1 Fast	22.6 / 3.5	25.7 / 3.2	27.6 / 3.6	22.5 / 2.8	35.3 / 3.3	24.0 / 3.0
<i>Open-Source Models</i>						
Qwen3-VL-235B	63.2 / 4.2	45.5 / 3.9	40.2 / 4.2	53.9 / 3.9	64.2 / 4.1	71.2 / 4.3
Qwen3-VL-30B	53.2 / 3.6	38.7 / 3.5	35.9 / 3.8	42.4 / 3.4	54.4 / 3.8	61.4 / 4.0
Qwen2.5-VL-72B	26.3 / 3.2	30.2 / 3.1	32.3 / 3.3	29.2 / 2.8	45.7 / 3.4	27.9 / 3.4

---

## Prompt 1: Standard Model Inference Prompt

---

**You are a professional financial domain expert.** I have uploaded a complete financial report PDF as page screenshots, along with a series of financial questions. These questions are guaranteed to be related to one or more charts within the PDF. Please answer all questions based on the relevant charts.

**The question types are:** `single_choice`, `multiple_choice`, `numerical_precise`, `numerical_approximate`, and `open_ended`.

- For `single_choice`, the answer is a single option letter.
- For `multiple_choice`, the answer is multiple option letters.
- For `numerical_precise` and `numerical_approximate`, the answer is a pure number without any units.
- For `open_ended`, the answer is a string of text.

**Your task is to determine the correct answer for each question.**

I will provide all questions at once. Please answer them in order. The response **MUST** be in JSON format.

### Output Requirements:

1. Output **ONLY** the required JSON content, with no other explanatory text or information.
2. The format for each question's answer is a dictionary with the key "answer" and the value as a string.
  - Example for `single_choice`: {"answer": "C"}
  - Example for `multiple_choice`: {"answer": "ABD"}
  - Example for `numerical_precise/approximate`: {"answer": "12.3"} (No units. Round or format decimals as the question requires).
  - Example for `open_ended`: {"answer": "This is a text answer."}
3. The outermost structure must be a list of these dictionary results, in the same order as the questions.

A strict reference for the output format is as follows (do not reference the content, only the format):

```
1 [
2   {"answer": "C" },
3   {"answer": "ABD" },
4   {"answer": "12.3" },
5   {"answer": "This is a text answer" }
6 ]
```

Here is the set of financial questions to be answered: *[Insert Questions List]*

---

Table 14: Full evaluation prompts for Model Inference.

## Prompt 2: LLM-as-a-Judge Scoring Prompt

You are a professional financial domain expert. I will provide an open-ended financial question, its reference answer, and an answer to be evaluated. Please provide a score for the evaluated answer based on the question and the reference answer.

### Scoring Criteria:

Scoring must be an integer between 0 and 5, where 0 is the lowest and 5 is the highest. The reference answer is for guidance only and does not necessarily represent a 5-point (perfect) answer.

### Output Format Requirement:

Provide only a single integer between 0 and 5. Do not include any punctuation or other content.

### [Input Template]

The Question is: {question}

The reference answer is: {context}

The answer to be evaluated is: {model\_answer}

Table 15: Full evaluation prompts for LLM-based Scoring.

Model	Reasoning Type	Domain Type	Associated Chart	Question	Options	Correct Answer	Model Answer	LMM Eval Score
Qwen3-VL-30B	Multiple Choice	Macro Markets	[page_16_image_3.png, page_19_table_2.png]	Comparing the projected index restructuring in India with the historical structural transition of China, which two descriptors accurately capture the empirical market evolution?	A.Back-loaded Transition B.Consumer-Tech Convergence C.Early Digital Hegemony D.Tech Weight Expansion	AB	<think>The chart in Exhibit 32 visually depicts a convex curve where the 'New China' growth is concentrated in the late cycle, which clearly supports the 'Back-loaded Transition' narrative (A) rather than any early dominance (C). Regarding the table in Exhibit 41, the massive injection of 'Internet Retail' into the pro-forma index implies a broad structural shift towards digital assets; although the specific tech line varies, the aggregate 'New Economy' ecosystem appears to grow significantly, suggesting a net 'Tech Weight Expansion' (D). Option B is less likely because Consumer and Tech sectors remain listed as distinct separate rows in the table, contradicting the idea of a fundamental convergence.</think><answer>AD</answer>	3
GPT-5	Multiple Choice	Macro Markets	[page_16_image_3.png, page_19_table_2.png]	Comparing the projected index restructuring in India with the historical structural transition of China, which two descriptors accurately capture the empirical market evolution?	A.Back-loaded Transition B.Consumer-Tech Convergence C.Early Digital Hegemony D.Tech Weight Expansion	AB	<think>Analyzing the temporal trajectory in Exhibit 32, the 'New China' share remains structurally low until 2014 and only breaches the 50% threshold post-2017, a pattern that definitively confirms 'Back-loaded Transition' (A) while simultaneously refuting 'Early Digital Hegemony' (C). Turning to the quantitative adjustments in Exhibit 41, I observe that while 'Consumer Discretionary' rises to 14%, the 'Info Tech' sector actually contracts from 18% to 14%; this precise numerical parity at 14% mathematically validates 'Consumer-Tech Convergence' (B), and the explicit reduction in Tech's weight directly falsifies the hypothesis of 'Tech Weight Expansion' (D).</think><answer> AB</answer>	5

Figure 8: Comparison of GPT-5 and Qwen3-VL-30B on Multiple Choice task.

Model	Reasoning Type	Domain Type	Associated Chart	Question	Reference Answer	Model Answer	LMM Eval Score
Qwen3-VL-30B	Open Ended	Industry & Competitive Analysis	[page_7_image_0.png, page_8_table_1.png]	Comparing Based on the information regarding average mobile data usage per user and the growth of internet users, describe the key characteristics of the digital consumption landscape in India and its potential implications for businesses. Provide specific data points to support your analysis.	...	India has high mobile data usage... over 14 GB, which is more than countries like the US or China... and the number of users is growing fast... This shows that Indian consumers are addicted to their phones... The main strategic implication is for telecom operators to sell more data plans... and for smartphone manufacturers to push expensive devices... because high data consumption means people need better hardware to watch videos... making it a hardware-driven market...	3
GPT-5	Open Ended	Industry & Competitive Analysis	[page_7_image_0.png, page_8_table_1.png]	Comparing Based on the information regarding average mobile data usage per user and the growth of internet users, describe the key characteristics of the digital consumption landscape in India and its potential implications for businesses. Provide specific data points to support your analysis.	...	India's digital landscape is characterized by a massive 800-million strong internet user base... supported by globally lowest data costs (~\$0.2/GB) and rapid adoption of the UPI payment interface... This structural shift creates a fertile environment where digital-first (fintech and SaaS businesses) can scale rapidly by leveraging the India Stack infrastructure... compels incumbents to aggressively invest in digital platforms and last-mile logistics... and fundamentally accelerates the financialization of household savings through rising retail equity participation...	5

Figure 9: Comparison of GPT-5 and Qwen3-VL-30B on Open Ended task.