UniFinEval: Towards Unified Evaluation of Financial Multimodal Models across Text, Images and Videos

Fangqi Lou[1†], Zhi Yang[1†], Lingfeng Zeng[1†], Zhiheng Jin[1], Qi Qi[1], Wei Zhang[1], Zhenyu Wu[1], Zhenxiong Yu[1], Lejie Zhang[1], Xiaoming Huang[7], Xiaolong Liang[7], Zheng Wei[7], Junbo Zou[8], Dongpo Cheng[5], Zhaowei Liu[6], Xin Guo[1*], Rongjunchen Zhang[2*], Liwen Zhang[1,3,4*]

[1]School of Statistics and Data Science, Shanghai University of Finance and Economics, [2]HiThink Research, [3]Shanghai Engineering Research Center of Finance Intelligence, [4]Qinghai Provincial Key Laboratory of Big Data in Finance and Artificial Intelligence Application Technology, [5]AI Finance Development and Service Center, Shanghai University of Finance and Economics, [6]School of Computer Science, Shanghai Jiao Tong University, [7]Tencent Youtu Lab, [8]Georgia Institute of Technology, GA, United States

{zhang.liwen}@shufe.edu.cn, {2021213225}@stu.sufe.edu.cn, {zhangrongjunchen}@myhexin.com

Abstract

Multimodal large language models (MLLMs) have shown increasing potential for applications in the financial domain. However, the challenges posed by real-world financial scenarios, including high-information-density, strong noise interference, and cross-modal multi-hop reasoning, go beyond the evaluation scope of existing financial multimodal benchmarks. To this end, we propose UniFinEval, the first unified multimodal financial evaluation benchmark designed for high-density financial information environments, covering three modalities: text, image, and video. UniFinEval abstracts three progressive core scenarios from perception to reasoning and decision-making, and further refines them into six task categories. It contains 5,567 high-quality Chinese and English question–answer (QA) pairs constructed by financial experts. We systematically evaluate 10 mainstream MLLMs under zero-shot and chain-of-thought (CoT) settings. Experimental results show that Gemini-3-Pro achieves the best overall performance, but still exhibits a significant gap compared with financial experts. Further error analysis reveals the capability deficiencies of MLLMs in financial scenarios. UniFinEval aims to systematically evaluate the capabilities of MLLMs in real financial scenarios, thereby promoting the application of MLLMs in the financial domain. The related code and data are available at https://github.com/aifinlab/UniFinEval.

1 Introduction

With the exponential growth in the scale of information production, the primary carriers of information dissemination have become increasingly multi-source and heterogeneous. Against this backdrop, large language models (LLMs) and multimodal large language models (MLLMs) are being applied to more complex cross-domain information processing and professional scenario analysis (Rahman et al., 2025; Li et al., 2025c; Zhang et al., 2025a; Das et al., 2025; Dennstädt et al., 2025; Xia et al., 2024). This demand is particularly prominent in the financial domain, where financial scenarios often require the simultaneous processing of massive financial reports, financial commentary videos, and other types of materials. These materials are inherently high-information-density and heterogeneous, requiring MLLMs to accurately filter valid evidence in complex environments, understand implicit correspondences between textual and visual information, and further conduct continuous analysis by incorporating information from videos. Although MLLMs have demonstrated strong capabilities, tasks in financial scenarios are far from merely single-step perception tasks.

Early financial benchmarks typically focus on evaluating models' ability to process information within a single modality, without sufficiently assessing their ability to jointly understand, align, and integrate information across multiple modalities, such as FinQA (Chen et al., 2021), Fin-Fact (Rangapur et al., 2025), and MMMU (Yue et al., 2024). These benchmarks mainly focus on capabilities in domain knowledge, reasoning, and numerical computation, but have not yet been extended to broader multimodal tasks. Such modality-restricted evaluation leads to inconsistent characterization of model capabilities and limits the practical reference value of evaluation results in real-world financial appli-

*Corresponding authors.
†These authors contributed equally to this work.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 44168–44206
October 24-29, 2026 ©2026 Association for Computational Linguistics

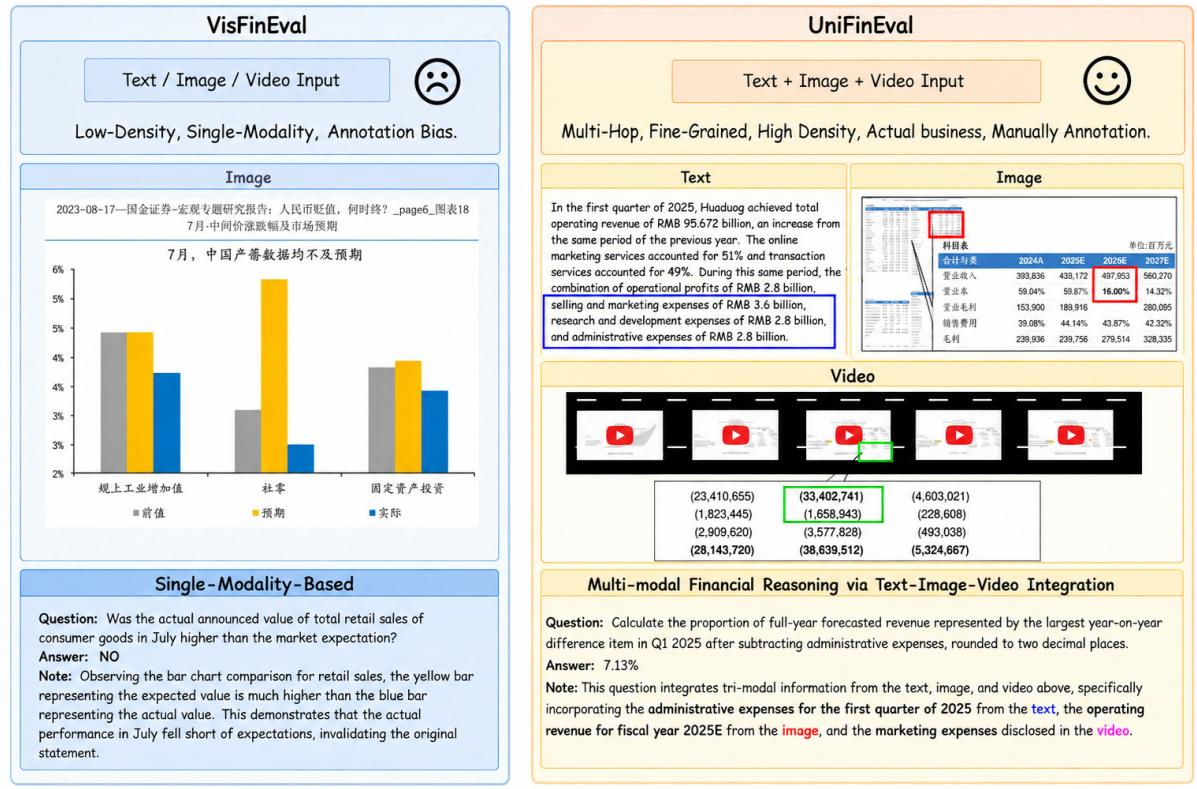


Figure 1: UniFinEval targets high-information-density financial environments by integrating text, image, and video inputs into a unified evaluation framework. Compared with low-density and single-modality settings, it emphasizes expert-constructed annotations, cross-modal consistency verification, and multi-hop financial reasoning over realistic business materials.

cations. Benchmarks such as MME-Finance (Gan et al., 2025) and BizFinBench.V2 (Guo et al., 2026) have begun to model business environments and evaluate models' abilities in financial analysis and decision-making. These benchmarks have played a crucial role in advancing the application of MLLMs in the financial domain.

Despite these advancements, as illustrated in Figure 1, existing benchmarks still suffer from a critical limitation: most of them focus only on models' capabilities in a single chart or video, while overlooking the high-information-density and noise-rich information environments of financial markets and the ability to jointly process information from multiple modalities. Since textual reports, financial images, and videos may all contain decisive evidence in financial tasks, at the same time, irrelevant or misleading signals may also appear across modalities. Relying solely on charts or video frames cannot fully represent the high-information-density and noise-rich information environments in real-world financial workflows.

To address this limitation, we propose UniFinEval, the first unified financial benchmark specifically designed for high-information-density environments in finance, integrating text, images, and videos into a unified evaluation framework. UniFinEval focuses on evaluating the ability of MLLMs to perform cross-modal multi-hop financial reasoning. Based on real-world financial business workflows, UniFinEval abstracts three major financial scenarios: Financial Evidence Understanding and Verification, Financial Analytical Reasoning, and Financial Risk Sensing and Decision Analysis. These scenarios are further refined into six task categories: financial evidence localization, Cross-modal Information Consistency Verification, Financial Numerical Derivation, Macro-Industry Trend Reasoning, Risk Signal Identification and Investment Decision Analysis. These tasks comprehensively characterize the full cognitive chain from multimodal perception and cross-modal alignment to high-level financial reasoning and decision-oriented analysis. The detailed architecture is illustrated in Figure 2. In summary, our primary contributions are:

1. We propose UniFinEval, the first unified

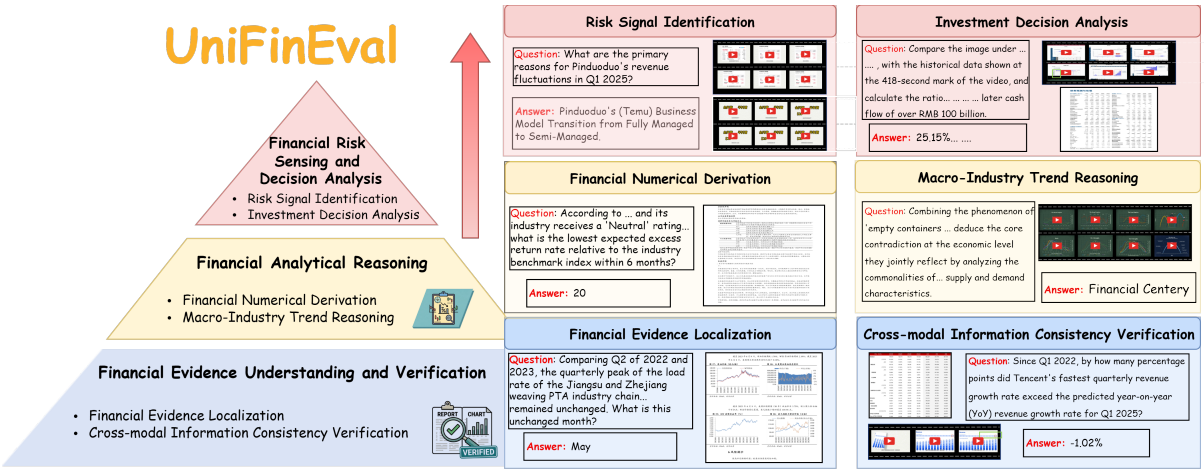


Figure 2: UniFinEval organizes multimodal financial reasoning into three progressive scenarios: financial evidence understanding and verification, financial analytical reasoning, and financial risk sensing and decision analysis. These scenarios are further decomposed into six task categories, forming a capability chain from evidence localization and cross-modal alignment to numerical derivation, trend reasoning, risk identification, and decision-oriented analysis.

financial multimodal benchmark for high-information-density heterogeneous environments. Through question–answer (QA) pairs carefully constructed by financial experts, UniFinEval deeply integrates text, images, and videos, providing a unified evaluation standard for assessing the deep application capabilities of MLLMs in the financial domain.

2. We construct a benchmark with three scenarios and six task categories from the perspective of multi-source heterogeneous information integration in finance. The benchmark decomposes cross-modal multi-hop reasoning into Financial Evidence Understanding and Verification, Financial Analytical Reasoning, and Financial Risk Sensing and Decision Analysis. It covers the complete cognitive chain from evidence localization to decision-oriented reasoning, posing a rigorous challenge to the practical performance of MLLMs.

3. We conduct extensive evaluation and targeted error analysis on mainstream large models. Through fine-grained analyses across tasks, modalities, and reasoning types, UniFinEval reveals key limitations of current MLLMs in high-information-density environments and provides practical and feasible optimization directions for improving the deployment of MLLMs in financial scenarios.

The organization of this paper is as follows. Section 2 reviews research on financial MLLMs and related multimodal benchmarks. Section 3 details the construction, quality control, and detailed statistics of UniFinEval. Sections 4 and 5 primarily describe the experimental setup, as well as the main experimental results and analysis. Finally, Section 6 summarizes this work and discusses potential future directions in multimodal financial intelligence.

2 Related Work

2.1 Financial Applications Analysis

Benefiting from the breakthrough progress of large language models (LLMs), their applications in the financial domain have gradually expanded from basic text processing to more complex professional analytical workflows. Early studies mainly focused on tasks such as financial sentiment analysis (Delgadillo et al., 2024; Kim et al., 2024; Iacovides et al., 2024), financial time-series analysis (Li et al., 2024b,c; Wang et al., 2024c; Mai, 2024; Cao et al., 2024) and financial text understanding (Masry and Hajian, 2024; Wilson et al., 2024). Further, LLMs have begun to be applied to tasks that are closer to business workflows, such as financial report interpretation, investment research, trading decision-making, asset allocation, and risk management (Yu et al., 2024b; Xiao et al., 2025; Liu et al., 2025d; Yu et al., 2024c). However, the information on which financial scenarios rely is inherently multimodal, with different carriers serving different evidential functions. Therefore, the application boundaries of large language models have also gradually expanded from language understanding to multimodal information processing. FinVisGPT (Wang et al., 2023), FinZero (Wang et al.,

Table 1: Comparison between UniFinEval and representative financial multimodal benchmarks. ✓indicates explicit support, △indicates partial or implicit support, and ✗indicates that the corresponding feature is not clearly covered. The abbreviations and their core meanings are as follows: Financial Video Data (FVD), Unified Text–Image–Video Evaluation (TIV), Human Construction (HC), Real-world Financial Scenario Simulation (RFS), High-density Financial Information (HDI), Cross-modal Multi-hop Reasoning (CMR), Cross-modal Consistency Verification (CCV), Evidence Grounding (EG) and Broad Financial Task Taxonomy (BFT).

Benchmarks	FVD	TIV	HC	RFS	HDI	CMR	CCV	EG	BFT
Image									
FAMMA	✗	✗	△	✗	△	△	✗	△	△
MME-Finance	✗	✗	△	△	△	△	✗	△	△
FCMR	✗	✗	△	△	✓	✓	△	△	△
FinMME	✗	✗	✓	△	✓	△	✗	△	✓
FinMMR	✗	✗	✓	△	✓	✓	✗	△	△
FinRAGBench-V	✗	✗	✓	✓	✓	△	✗	✓	△
Video									
VideoConviction	✓	✗	✓	△	△	✗	✗	△	✗
FinCap	✓	✗	△	△	△	✗	✗	△	✗
FCMBench-Video	✓	✗	✓	✓	✓	△	△	✓	△
Ours (Text + Image + Video)	✓	✓	✓	✓	✓	✓	✓	✓	✓

2025b), and FinTral (Bhatia et al., 2024) enable in-depth analysis of financial charts. FinAgent (Zhang et al., 2024a) constructs a multimodal agent capable of autonomously making trading decisions; MM-DREX (Chen et al., 2025b) leverages multimodal information to achieve dynamic routing for expert-oriented trading; and AT-FinGPT (Liu et al., 2025b) uses speech features to assist financial risk prediction. Nevertheless, existing studies on LLMs applications mostly focus on specific tasks or system-level capabilities, while the core requirements for applying LLMs in the financial domain remain insufficiently satisfied, leaving broad space for future exploration.

2.2 Multimodal Financial Benchmark Analysis

As the applications of MLLMs in financial domain continue to expand, traditional financial benchmarks and general multimodal benchmarks have gradually revealed their limitations. On the one hand, traditional financial benchmarks are restricted to text and structured data, making it difficult for them to cover multimodal information such as charts and videos (Zhu et al., 2024; Zhao et al., 2024; Wang et al., 2024b; Reddy et al., 2024; Chen et al., 2024, 2025a; Guo et al., 2025; Liu et al., 2025a; Li et al., 2025a; Matlin et al., 2025; Nie et al., 2025). On the other hand, general multimodal benchmarks lack the domain-specific knowledge depth required in finance (Li et al., 2024a; Yu et al., 2024a; Zhang et al., 2024b; Liu et al., 2024). Neither type of benchmark can accurately measure the comprehensive performance of MLLMs in real-world financial scenarios. Recent financial multimodal benchmarks have partially addressed this gap. Specifically, works such as MME-Finance (Gan et al., 2025), FinChart-Bench (Shu et al., 2025), CFBenchmark-MM (Li et al., 2025b), FinMME (Luo et al., 2025), VisFinEval (Liu et al., 2025c) and MultiFinBen (Peng et al., 2026) incorporate financial charts, document images, and visual question answering into evaluation, promoting the shift of financial evaluation from purely textual tasks toward chart understanding and visual evidence parsing. FinMR (Deng et al., 2025), FinMMR (Tang et al., 2025), FAMMA (Xue et al., 2024) and XFinBench (Zhang et al., 2025b) further extend evaluation to financial multimodal reasoning tasks. FCMR (Kim et al., 2025) specifically evaluates cross-modal multi-hop reasoning over financial text, tables, and charts, including challenging three-hop reasoning settings. Complementarily, FinRAGBench-V (Zhao et al., 2025) evaluates visual retrieval-augmented generation over financial documents, with an emphasis on evidence retrieval, visual citation, and traceability. However, these benchmarks still exhibit several significant limitations. First, they mainly focus on static charts and have not sufficiently covered information carriers with temporal dynamics and evolving semantics, such as financial videos. Second, most data are derived from textbooks or standardized examinations, which differ substantially from real financial

environments characterized by high-information-density, noise-rich, and cross-modal multi-hop reasoning. Third, relying on LLMs for automatic annotation or QA generation may introduce model bias, insufficient evidence grounding, or hallucination risks. In addition, among the few benchmarks involving financial videos, FCMBench-Video (Cui et al., 2026) focuses only on financial credit scenarios, while VideoConviction (Galarnyk et al., 2025) and FinCap (Sukhani et al., 2025) have not systematically evaluated financial business capabilities, nor have they sufficiently examined the joint reasoning relationships between video information and textual or visual information.

3 UniFinEval

3.1 Overview

We propose UniFinEval, a unified multimodal benchmark for real-world financial scenarios. Unlike previous single-modality evaluations that mainly focus on financial images or financial videos, UniFinEval incorporates text, images, and videos into a unified evaluation framework, aiming to systematically assess the cross-modal multi-hop reasoning capabilities of MLLMs when facing high-information-density heterogeneous environments in financial scenarios. Based on real business workflows, UniFinEval constructs an application-oriented task taxonomy, organizing samples into three progressive scenarios, further refined into six task categories, covering the complete capability chain from basic evidence acquisition to high-level decision analysis. More detailed statistics are provided in Table 2.

Table 2: Statistics of UniFinEval across three financial scenarios and six task categories, including average input length, number of question–answer pairs, and representative modality combinations. The statistics characterize the benchmark's coverage of high-density text, image, video, and cross-modal financial reasoning settings.

Scen.	Task	Avg. Tok.	#QA	Modality Combination			
				Img.	Vid.	Img. + Vid.	T. + Img. + Vid.
FEUV	FEL	5150.0	830	180	146	168	140
	CICV	7140.2	957	157	65	260	132
FAR	FND	6411.7	1092	258	95	172	190
	MITR	9411.7	906	200	170	210	226
FRSDA	RSI	11201.4	853	166	101	205	203
	IDA	17651.3	929	150	130	199	187
All	–	–	**5567**	**1111**	**707**	**1214**	**1078**

3.2 Scenario Construction

We model multimodal reasoning capabilities as three progressive scenarios: Financial Evidence Understanding and Verification, Financial Analytical Reasoning, and Financial Risk Sensing and Decision Analysis. These three scenarios correspond to the capability chain in real financial tasks, progressing from information acquisition to decision judgment. They are further refined into six task categories, which evaluate models' abilities in evidence localization, cross-modal consistency judgment, numerical derivation, trend reasoning, risk identification, and investment decision analysis, respectively. This design aims to examine whether MLLMs can perform cross-modal multi-hop reasoning and analysis over high-information-density and noise-rich financial scenarios.

Financial Evidence Understanding and Verification (FEUV) corresponds to the basic information verification stage in financial workflows. It focuses on whether models can identify valid evidence from high-density, multi-source financial materials, and determine the correspondence and consistency among information from different modalities. The Financial Evidence Localization (FEL) task evaluates models' ability to select key evidence from high-density financial materials and indicate its source, with a focus on distinguishing question-relevant information from background distractors. The Cross-modal Information Consistency Verification (CICV) task evaluates models' ability to compare information from text, images, and videos, and determine whether they mutually support each other or contradictions omissions or mismatches exist, with an emphasis on cross-modal semantic alignment and evidence-chain construction.

Financial Analytical Reasoning (FAR) corresponds to the evidence analysis stage in financial workflows. It focuses on whether models can perform computation, comparison, induction, and trend judgment based on acquired evidence, and generate analytical results constrained by financial semantics. The Financial Numerical Derivation (FND) task evaluates ability to perform computational derivations based on numerical information in the materials, focusing on numerical extraction, computational relation identification, and the consistent correspondence among "evidence source–calculation process–answer result." The Macro-

Industry Trend Reasoning (MITR) task evaluates ability to synthesize multi-source financial information to judge trend directions and potential impacts. It emphasizes models' understanding of relationships among financial variables, event transmission logic, and industry evolution paths, especially their integrated analytical ability when multi-source signals are not fully consistent.

Financial Risk Sensing and Decision Analysis (FRSDA) corresponds to the high-level risk assessment and decision-support stage in financial workflows. It focuses on whether models can identify key information that affects risk judgment or decision direction when complex evidence, constraints, and uncertainty coexist. The Risk Signal Identification (RSI) task evaluates ability to extract risk signals from scattered financial evidence and explain their sources, affected objects, and impact directions. The Investment Decision Analysis (IDA) task evaluates ability to conduct evidence-driven decision analysis under given constraints. It focuses on whether models can integrate evidence verification, numerical analysis, risk identification, and conditional constraints to form cautious, coherent, and interpretable judgments, rather than directly generating investment advice. For a more detailed introduction to the above content, please refer to Appendix A.1.

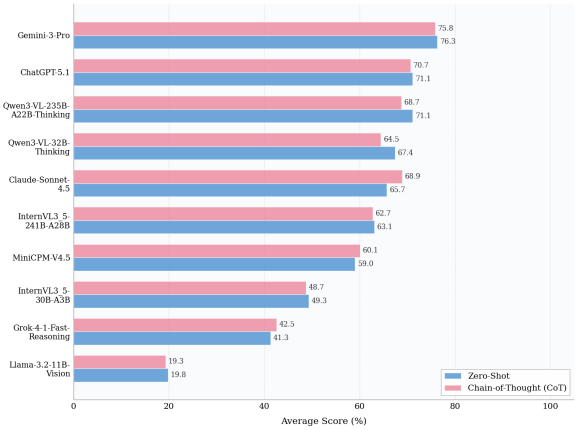


Figure 3: Comparison of overall model performance under the Zero-Shot and Chain-of-Thought (CoT) settings. The effect of CoT is model-dependent: some models achieve modest improvements, while others show slight performance degradation, resulting in limited overall gains.

3.3 Quality Control

To characterize real-world financial environments, UniFinEval collects more than 2,500 publicly available listed-company documents and financial research reports, as well as 2,000 online financial analysis videos. Unlike general multimodal data, financial materials often contain highly compressed indicators, entity relationships, and explanatory texts, while information across different modalities may be complementary or inconsistent. Therefore, this stage also focuses on examining whether the collected materials contain sufficient information density and financial analytical value.

To ensure the data credibility, task verifiability, and evaluation stability of UniFinEval, we invited 12 experts with financial backgrounds to participate in data construction. On average, these experts have more than four years of financial industry experience and over one year of annotation experience. All experts followed unified task definitions and annotation guidelines. No LLMs were used to assist the construction process, thereby reducing the influence of model-induced bias, templated expressions, and potential data contamination on the evaluation results.

For further details regarding quality control, please refer to Appendix A.2.

4 Experiments Settings

4.1 Baseline Models

We evaluate 10 mainstream MLLMs. Closed-source models are accessed through their respective APIs, while open-source models are deployed locally. All locally deployed open-source models are run on 8× NVIDIA A800 GPUs. Except for the Llama-series model, which is deployed with LMDeploy, all other open-source models are served with vLLM for efficient local inference. In addition, we invite a financial expert who did not participate in data construction to take part in task evaluation, providing a performance comparison with large models.

The evaluation covers four leading closed-source models, including ChatGPT-5.1 (OpenAI, 2025), Gemini-3-Pro (Google, 2025), Grok-4-1-Fast-Reasoning (xAI, 2025), and Claude-Sonnet-4.5 (Anthropic, 2025). For open-source models, we evaluate six mainstream multimodal large models, including Qwen3-VL-235B-A22B-Thinking, Qwen3-VL-32B-Thinking (Bai et al., 2025), InternVL3_5-241B-A28B, InternVL3_5-30B-A3B (Wang et al., 2025a), MiniCPM-V-4.5 (Yu et al., 2025), and Llama-3.2-11B-Vision (MetaAI, 2024). More details about the

Table 3: Performance under zero-shot setting on UniFinEval(%). The results are color-coded to indicate the top three performers in each task: The highest model result in each task column is highlighted with a blue background, the second-best with a pink background, and the third-best with a green background. Financial Expert is reported as a reference upper-bound and is not included in the highlighting comparison.

Model	Financial Evidence Understanding and Verification		Financial Analytical Reasoning		Financial Risk Sensing and Decision Analysis		Average
	Financial Evidence Localization	Cross-modal Information Consistency Verification	Financial Numerical Derivation	Macro-Industry Trend Reasoning	Risk Signal Identification	Investment Decision Analysis	
Gemini-3-Pro	79.2	77.2	79.7	75.3	77.0	69.0	**76.3**
ChatGPT-5.1	75.2	71.9	66.0	72.8	73.2	69.0	**71.1**
Qwen3-VL-235B-A22B-Thinking	74.3	73.7	75.6	67.0	77.0	58.7	**71.1**
Qwen3-VL-32B-Thinking	70.0	66.7	71.1	62.1	68.8	65.6	**67.4**
Claude-Sonnet-4.5	66.1	70.2	66.2	68.4	68.0	55.2	**65.7**
InternVL3_5-241B-A28B	61.3	56.1	58.6	70.9	64.0	68.8	**63.1**
MiniCPM-V-4.5	50.5	57.9	54.0	67.0	62.8	62.1	**59.0**
InternVL3_5-30B-A3B	44.8	54.4	49.1	53.9	51.8	41.3	**49.3**
Grok-4-1-Fast-Reasoning	41.7	43.9	31.0	51.4	40.7	41.3	**41.3**
Llama-3.2-11B-Vision	21.7	24.6	15.2	31.6	19.8	6.9	**19.8**
Financial Expert	100.0	98.0	93.5	92.1	83.2	84.8	**92.0**

models are provided in Appendix B.1.

5 Results

5.1 Evaluation Methods

We adopt two core evaluation settings: Zero-Shot and Chain-of-Thought (CoT). These two settings are used to compare models' capabilities under direct-answering and explicit-reasoning conditions. Since all samples in UniFinEval are open-ended questions with verifiable answers constructed by financial experts, we uniformly adopt Accuracy as the core metric and use Qwen-Max (Yang et al., 2025) for scoring. All models use the same prompt template and a fixed output format.

To further validate the reliability of the automatic evaluation protocol, we conduct an additional multi-judge validation to examine whether the choice of evaluator introduces systematic model-family bias. Specifically, we ensure that the validation samples cover all six task categories, Chinese and English questions, different modality combinations, and different difficulty levels. While keeping the original evaluation results from Qwen-Max unchanged, we further introduce DeepSeek-V4-Pro (DeepSeek-AI et al., 2026) and finance-domain experts as independent evaluators to assess the same model responses under exactly the same evaluation criteria. The results show a high degree of instance-level agreement among the different evaluators, indicating that the main evaluation conclusions are robust to the choice of evaluator. Details of the experimental setup and the complete multi-judge analysis are provided in the Appendix B.2.

5.2 Main Result

Table 3 presents the overall performance of different models on each task in UniFinEval under the zero-shot setting and Figure 3 presents a visual comparison of the overall performance of all participating models under the Zero-Shot and Chain-of-Thought (CoT) settings, the evaluation results under the CoT setting are provided in Appendix B.3. In terms of overall ranking, Gemini-3-Pro ranks first among all models with an average accuracy of 76.3%. It achieves the joint-best performance on Risk Signal Identification and Investment Decision Analysis, while leading individually on the other four tasks, demonstrating comprehensive competitive advantages in financial scenarios. However, the financial expert achieves an average accuracy of 92.0% across the six task categories, which is still 15.7 percentage points higher than Gemini-3-Pro. This gap indicates that even the strongest current models have not yet reached expert-level stability in complex financial business scenarios.

From the task perspective, different capability modules show clear discrepancies. In the Financial Evidence Understanding and Verification scenario, Gemini-3-Pro achieves the highest accuracy on both tasks. Nevertheless, compared with the financial expert's accuracies of 100% and 98% on these two tasks, there remains a gap of more than 20 percentage points. This indicates that models still struggle to reach expert-level precision when facing high-information-density financial environments.

In the Financial Analytical Reasoning scenario, Gemini-3-Pro continues to maintain the lead. Notably, Qwen3-VL-235B-A22B-Thinking achieves 75.6% on Financial Numerical Derivation, substantially outperforming most other models, which demonstrates the strong potential of large-scale open-source models in financial numerical computation and derivation tasks. On the other

hand, InternVL3_5-241B-A28B reaches 70.9% on Macro-Industry Trend Reasoning, surpassing Qwen3-VL-235B-A22B-Thinking and Claude-Sonnet-4.5. This suggests that some models, despite not being superior in overall capability, may still possess local advantages in specific financial analysis tasks.

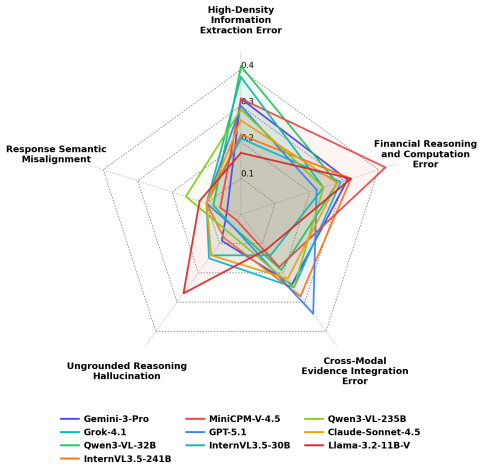


Figure 4: We analyze the error distributions of all evaluated MLLMs and summarize five typical error types observed in high-information-density financial environments.

In the Financial Risk Sensing and Decision Analysis scenario, model performance becomes further differentiated. On Risk Signal Identification, Gemini-3-Pro and Qwen3-VL-235B-A22B-Thinking both achieve 77.0%, with a relatively small gap compared with the financial expert's 83.2%. This indicates that current models have a certain ability to identify explicit risk clues and abnormal change signals. By contrast, the highest accuracy on Investment Decision Analysis is only 69.0%, jointly achieved by Gemini-3-Pro and ChatGPT-5.1, whereas the financial expert reaches 84.8%. This shows that when tasks move beyond risk clue identification toward asset allocation and multi-factor decision-making under constraints, models still struggle to form decision analyses that stably approach expert-level performance.

From the perspective of model categories, leading closed-source models still occupy advantageous positions in the overall ranking, with Gemini-3-Pro ranking first and ChatGPT-5.1 joint second. However, the catching-up trend of large-scale open-source models has also begun to emerge. Qwen3-VL-235B-A22B-Thinking matches ChatGPT-5.1 in average accuracy and jointly ranks second, while Qwen3-VL-32B-Thinking also outperforms Claude-Sonnet-4.5. In contrast, a clear gap remains between medium- and small-scale models and leading models, indicating small-scale vision-language models still suffer from systematic deficiencies in high-density information processing, cross-modal evidence integration, and complex decision reasoning in financial scenarios.

5.3 Error Analysis

To further investigate the performance deficiencies of models on UniFinEval, we conduct a targeted analysis by randomly sampling 30% of the erroneous cases from all evaluated models. From the perspective of cross-modal multi-hop reasoning over high-density heterogeneous financial information, we identify five typical capability deficiencies commonly observed in current large models on financial tasks: High-Density Information Extraction Errors, Financial Reasoning and Computation Errors, Cross-Modal Evidence Integration Errors, Ungrounded Reasoning Hallucinations, and Response Semantic Misalignment.

According to the error distribution shown in Figure 4, most models exhibit similar error patterns. This suggests that current MLLMs need to overcome common bottlenecks in business adaptation. By contrast, the average proportions of ungrounded reasoning hallucinations and response semantic misalignment are generally lower than those of the above three error types, but they remain pronounced in some models. Llama-3.2-11B-Vision has an ungrounded reasoning hallucination proportion of 27%, significantly higher than that of other models. This indicates that it is more likely to generate answers detached from the given information in complex financial environments, suggesting the need to better constrain its hallucinated generation. In terms of response semantic misalignment, Qwen3-VL-235B-A22B-Thinking accounts for the highest proportion, suggesting that the model may over-analyze while failing to strictly align with the judgment target or output requirements of the question. This points to the need to improve its instruction recognition and instruction-following capabilities. Detailed descriptions of error cases and capability deficiencies are provided in Appendix C.

6 Conclusion

In this paper, we propose UniFinEval, a unified multimodal benchmark for high-information-density heterogeneous financial environments. By

integrating text, images, and videos, UniFinEval evaluates cross-modal multi-hop reasoning across three progressive scenarios and six task categories. The benchmark contains 5,567 expert-constructed Q&A pairs grounded in real financial materials. Experiments on 10 mainstream multimodal MLLMs show that current models still lag significantly behind financial experts, despite strong performance from Gemini-3-Pro and competitive results from Qwen3-VL-235B-A22B-Thinking among open-source models. Error analysis further identifies key weaknesses in dense evidence extraction, cross-modal evidence integration, financial computation, and decision reasoning. We hope that UniFinEval can serve as a fine-grained evaluation tool for diagnosing the capability boundaries of current models in real-world financial scenarios, promoting the development of multimodal large models and facilitating their more practical, robust, and trustworthy applications in the financial domain.

Limitations

Although UniFinEval has advances the evaluation of MLLMs in finance, several limitations remain. First, current tasks are dominated by offline Q&A forms, not yet fully simulating the complex closed loop of long-duration interaction and dynamic decision feedback in real analysis processes. Second, evaluation results mainly focus on the correctness and consistency of model outputs, leaving room for deepening the analysis of the interpretability of model reasoning paths. These limitations reflect the current scope and boundaries of UniFinEval as a controllable, verifiable, and fine-grained financial multimodal evaluation benchmark, which we will further optimize in future work.

Acknowledgments

This work was supported by the Shanghai Engineering Research Center of Finance Intelligence (Grant No, 19DZ2254600). We sincerely appreciate the assistance provided by Qinghai Provincial Key Laboratory of Big Data in Finance and Artificial Intelligence Application Technology and Tencent Youtu Lab.

References

Anthropic. 2025. Claude-sonnet-4.5. https://www.anthropic.com/claude/sonnet.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Gagan Bhatia, El Moatez Billah Nagoudi, Hasan Cavusoglu, and Muhammad Abdul-Mageed. 2024. Fintral: A family of gpt-4 level multimodal financial large language models. *arXiv preprint arXiv:2402.10986*.

Yupeng Cao, Zhi Chen, Prashant Kumar, Qingyun Pei, Yangyang Yu, Haohang Li, Fabrizio Dimino, Lorenzo Ausiello, KP Subbalakshmi, and Papa Momar Ndiaye. 2024. Risklabs: predicting financial risk using large language model based on multimodal and multisources data. *arXiv preprint arXiv:2404.07452*.

Jialin Chen, Aosong Feng, Ziyu Zhao, Juan Garza, Gaukhar Nurbek, Cheng Qin, Ali Maatouk, Leandros Tassiulas, Yifeng Gao, and Rex Ying. 2025a. Mtbench: A multimodal time series benchmark for temporal reasoning and question answering. *arXiv preprint arXiv:2503.16858*.

Jian Chen, Peilin Zhou, Yining Hua, Yingxin Loh, Kehui Chen, Ziyuan Li, Bing Zhu, and Junwei Liang. 2024. Fintextqa: A dataset for long-form financial question answering. *arXiv preprint arXiv:2405.09980*.

Yang Chen, Yueheng Jiang, Zhaozhao Ma, Yuchen Cao, Jacky Keung, Kun Kuang, Leilei Gan, Yiquan Wu, and Fei Wu. 2025b. Mm-drex: Multimodal-driven dynamic routing of llm experts for financial trading. *arXiv preprint arXiv:2509.05080*.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Runze Cui, Fangxin Shang, Yehui Yang, Qing Yang, Yanwu Xu, and Tao Chen. 2026. Fcmbench-video: Benchmarking document video intelligence. *arXiv preprint arXiv:2604.25186*.

Sarmistha Das, R. E. Zera Marveen Lyngkhoi, Sriparna Saha, and Alka Maurya. 2025. Unlocking financial insights: An advanced multimodal summarization with multimodal output framework for financial advisory videos. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 11976–11985, New York, NY, USA. Association for Computing Machinery.

DeepSeek-AI, Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, Chengda

Lu, Chenggang Zhao, Chengqi Deng, Chengyu Hou, Chenhao Xu, Chenze Shao, Chong Ruan, Conner Sun, and 300 others. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence.

Josiel Delgadillo, Johnson Kinyua, and Charles Mutigwe. 2024. Finsosent: Advancing financial market sentiment analysis through pretrained large language models. *Big Data and Cognitive Computing*, 8(8):87.

Shuangyan Deng, Haizhou Peng, Jiachen Xu, Rui Mao, Ciprian Doru Giurcaneanu, and Jiamou Liu. 2025. 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.

Fabio Dennstädt, Luc Lerch, Max Schmerder, Nikola Cihoric, Grazia Maria Cereghetti, Roberto Gaio, Harald Bonel, Irina Filchenko, Janna Hastings, Florian Dammann, Daniel M. Aebersold, Hendrik von Tengg-Kobligk, and Knud Nairz. 2025. A comparative performance analysis of regular expressions and an llm-based approach to extract the bi-rads score from radiological reports. *medRxiv*, pages 2025–06.

Michael Galarnyk, Veer Kejriwal, Agam Shah, Yash Bhardwaj, Nicholas Watney Meyer, Anand Krishnan, and Sudheer Chava. 2025. Videoconviction: A multimodal benchmark for human conviction and stock market recommendations. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 5447–5458.

Ziliang Gan, Dong Zhang, Haohan Li, Yang Wu, Xueyuan Lin, Ji Liu, Haipang Wu, Chaoyou Fu, Zenglin Xu, Rongjunchen Zhang, and Yong Dai. 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, page 12867–12874, New York, NY, USA. Association for Computing Machinery.

Google. 2025. Gemini-3-pro-preview. https://gemini.google.com/.

Xin Guo, Haotian Xia, Zhaowei Liu, Hanyang Cao, Zhi Yang, Zhiqiang Liu, Sizhe Wang, Jinyi Niu, Chuqi Wang, Yanhui Wang, Xiaolong Liang, Xiaoming Huang, Bing Zhu, Zhongyu Wei, Yun Chen, Weining Shen, and Liwen Zhang. 2025. FinEval: A Chinese financial domain knowledge evaluation benchmark for large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6258–6292, Albuquerque, New Mexico. Association for Computational Linguistics.

Xin Guo, Rongjunchen Zhang, Guilong Lu, Xuntao Guo, Shuai Jia, Zhi Yang, and Liwen Zhang. 2026. Bizfinbench.v2: Towards reliable llms in finance via real-user data and offline/online bilingual evaluation.

Proceedings of the 43 rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306.

Giorgos Iacovides, Thanos Konstantinidis, Mingxue Xu, and Danilo Mandic. 2024. Finllama: Llm-based financial sentiment analysis for algorithmic trading. In *Proceedings of the 5th ACM International Conference on AI in Finance*, pages 134–141.

Alex Kim, Maximilian Muhn, and Valeri Nikolaev. 2024. Financial statement analysis with large language models. *arXiv preprint arXiv:2407.17866*.

Seunghee Kim, Changhyeon Kim, and Taeuk Kim. 2025. Fcmr: robust evaluation of financial cross-modal multi-hop reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23352–23380.

Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. 2024a. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*.

Haohang Li, Yupeng Cao, Yangyang Yu, Shashidhar Reddy Javaji, Zhiyang Deng, Yueru He, Yuechen Jiang, Zining Zhu, K.p. Subbalakshmi, Jimin Huang, Lingfei Qian, Xueqing Peng, Jordan W. Suchow, and Qianqian Xie. 2025a. INVESTORBENCH: A benchmark for financial decision-making tasks with LLM-based agent. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2509–2525, Vienna, Austria. Association for Computational Linguistics.

Jiangtong Li, Yiyun Zhu, Dawei Cheng, Zhijun Ding, and Changjun Jiang. 2025b. Cfbenchmark-mm: Chinese financial assistant benchmark for multimodal large language model. *arXiv preprint arXiv:2506.13055*.

Pengyue Li, Sheng Wang, Hua Dai, Zhiyu Chen, Zhifeng Bao, and Brian D Davison. 2025c. A survey on open dataset search in the llm era: Retrospectives and perspectives. *arXiv preprint arXiv:2509.00728*.

Xiang Li, Zhenyu Li, Chen Shi, Yong Xu, Qing Du, Mingkui Tan, Jun Huang, and Wei Lin. 2024b. Alphafin: Benchmarking financial analysis with retrieval-augmented stock-chain framework. *arXiv preprint arXiv:2403.12582*.

Xiangyu Li, Xinjie Shen, Yawen Zeng, Xiaofen Xing, and Jin Xu. 2024c. Finreport: Explainable stock earnings forecasting via news factor analyzing model. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 319–327.

Shu Liu, Shangqing Zhao, Chenghao Jia, Xinlin Zhuang, Zhaoguang Long, Jie Zhou, Aimin Zhou, Man Lan, and Yang Chong. 2025a. Findabench: Benchmarking financial data analysis ability of large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 710–725.

Yingnan Liu, Ningbo Bu, Zhiqiang Li, Yongmin Zhang, and Zhenyu Zhao. 2025b. At-fingpt: Financial risk prediction via an audio-text large language model. *Finance Research Letters*, 77:106967.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Zhaowei Liu, Xin Guo, Haotian Xia, Lingfeng Zeng, Fangqi Lou, Jinyi Niu, Mengping Li, Qi Qi, Jiahuan Li, Wei Zhang, Yinglong Wang, Weige Cai, Weining Shen, and Liwen Zhang. 2025c. VisFinEval: A scenario-driven Chinese multimodal benchmark for holistic financial understanding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24088–24146, Suzhou, China. Association for Computational Linguistics.

Zhaowei Liu, Xin Guo, Zhi Yang, Fangqi Lou, Lingfeng Zeng, Jinyi Niu, Mengping Li, Qi Qi, Zhiqiang Liu, Yiyang Han, Dongpo Cheng, Ronghao Chen, Huacan Wang, Xingdong Feng, Huixia Judy Wang, Chengchun Shi, and Liwen Zhang. 2025d. Fin-r1: A large language model for financial reasoning through reinforcement learning. *arXiv preprint arXiv:2503.16252*.

Junyu Luo, Zhizhuo Kou, Liming Yang, Xiao Luo, Jinsheng Huang, Zhiping Xiao, Jingshu Peng, Chengzhong Liu, Jiaming Ji, Xuanzhe Liu, Sirui Han, Ming Zhang, and Yike Guo. 2025. FinMME: Benchmark dataset for financial multi-modal reasoning evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29465–29489, Vienna, Austria. Association for Computational Linguistics.

Dat Mai. 2024. Stockgpt: A genai model for stock prediction and trading. *arXiv preprint arXiv:2404.05101*.

Ahmed Masry and Amir Hajian. 2024. Longfin: A multimodal document understanding model for long financial domain documents. *arXiv preprint arXiv:2401.15050*.

Glenn Matlin, Mika Okamoto, Huzaifa Pardawala, Yang Yang, and Sudheer Chava. 2025. Financial language model evaluation (FLaME). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22633–22679, Vienna, Austria. Association for Computational Linguistics.

MetaAI. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct.

Ying Nie, Binwei Yan, Tianyu Guo, Hao Liu, Haoyu Wang, Wei He, Binfan Zheng, Weihao Wang, Qiang Li, Weijian Sun, Yunhe Wang, and Dacheng Tao. 2025. CFinBench: A comprehensive Chinese financial benchmark for large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 876–891, Albuquerque, New Mexico. Association for Computational Linguistics.

OpenAI. 2025. Gpt-5.1: A smarter, more conversational chatgpt. https://openai.com/zh-Hans-CN/index/gpt-5-1/.

Xueqing Peng, Lingfei Qian, Yan Wang, Ruoyu Xiang, Yueru He, Yang Ren, Mingyang Jiang, Vincent Jim Zhang, Yuqing Guo, Jeff Zhao, Huan He, Yi Han, Yun Feng, Yuechen Jiang, Yupeng Cao, Haohang Li, Yangyang Yu, Xiaoyu Wang, Penglei Gao, and 28 others. 2026. MultiFinBen: Benchmarking large language models for multilingual and multimodal financial application. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16934–16963, San Diego, California, United States. Association for Computational Linguistics.

Mizanur Rahman, Amran Bhuiyan, Mohammed Saidul Islam, Md Tahmid Rahman Laskar, Ridwan Mahbub, Ahmed Masry, Shafiq Joty, and Enamul Hoque. 2025. Llm-based data science agents: A survey of capabilities, challenges, and future directions. *arXiv preprint arXiv:2510.04023*.

Aman Rangapur, Haoran Wang, Ling Jian, and Kai Shu. 2025. Fin-fact: A benchmark dataset for multimodal financial fact-checking and explanation generation. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 785–788.

Varshini Reddy, Rik Koncel-Kedziorski, Viet Dac Lai, Michael Krumdick, Charles Lovering, and Chris Tanner. 2024. Docfinqa: A long-context financial reasoning dataset. *arXiv preprint arXiv:2401.06915*.

Dong Shu, Haoyang Yuan, Yuchen Wang, Yanguang Liu, Huopu Zhang, Haiyan Zhao, and Mengnan Du. 2025. Finchart-bench: Benchmarking financial chart comprehension in vision-language models. *arXiv preprint arXiv:2507.14823*.

Siddhant Sukhani, Yash Bhardwaj, Riya Bhadani, Veer Kejriwal, Michael Galarnyk, and Sudheer Chava. 2025. Fincap: Topic-aligned captions for short-form financial youtube videos. *arXiv preprint arXiv:2509.25745*.

Zichen Tang, Haihong E, Jiacheng Liu, Zhongjun Yang, Rongjin Li, Zihua Rong, Haoyang He, Zhuodi Hao, Xinyang Hu, Kun Ji, Ziyan Ma, Mengyuan Ji, Jun Zhang, Chenghao Ma, Qianhe Zheng, Yang Liu, Yiling Huang, Xinyi Hu, Qing Huang, and 2 others. 2025. Finmmr: make financial numerical reasoning more multimodal, comprehensive, and challenging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3245–3257.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, Bo Zhang, Liqun Wei, Zhihao Sui, Wei Li, Botian Shi, Yu Qiao, Dahua Lin, and Conghui He. 2024a. Mineru: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839*.

Haochen Wang, Kai Hu, Haoyu Dong, and Liangcai Gao. 2024b. Doctabqa: Answering questions from long documents using tables. In *International Conference on Document Analysis and Recognition*, pages 470–487. Springer.

Saizhuo Wang, Hang Yuan, Lionel M Ni, and Jian Guo. 2024c. Quantagent: Seeking holy grail in trading by self-improving large language model. *arXiv preprint arXiv:2402.03755*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 56 others. 2025a. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Yanlong Wang, Jian Xu, Fei Ma, Hongkang Zhang, Hang Yu, Tiantian Gao, Yu Wang, Haochen You, Shao-Lun Huang, Danny Dongning Sun, and Xiao-Ping Zhang. 2025b. Finzero: Launching multi-modal financial time series forecast with large reasoning model. *arXiv preprint arXiv:2509.08742*.

Ziao Wang, Yuhang Li, Junda Wu, Jaehyeon Soon, and Xiaofeng Zhang. 2023. Finvis-gpt: A multimodal large language model for financial chart analysis. *arXiv preprint arXiv:2308.01430*.

Ezhilan Wilson, Anshul Saxena, Jayant Mahajan, Lekha Panikulangara, Shruti Kulkarni, and Pritty Jain. 2024. Fin2sum: advancing ai-driven financial text summarization with llms. In *2024 International Conference on Trends in Quantum Computing and Emerging Business Technologies*, pages 1–5. IEEE.

xAI. 2025. Grok 4.1 fast and agent tools api. https://x.ai/grok.

Haotian Xia, Zhengbang Yang, Junbo Zou, Rhys Tracy, Yuqing Wang, Chi Lu, Christopher Lai, Yanjun He, Xun Shao, Zhuoqing Xie, Yuan-fang Wang, Weining Shen, and Hanjie Chen. 2024. Sportu: A comprehensive sports understanding benchmark for multimodal large language models. *arXiv preprint arXiv:2410.08474*.

Yijia Xiao, Edward Sun, Tong Chen, Fang Wu, Di Luo, and Wei Wang. 2025. Trading-r1: Financial trading with llm reasoning via reinforcement learning. *arXiv preprint arXiv:2509.11420*.

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, Bokai Xu, Junbo Cui, Yingjing Xu, Liqing Ruan, Luoyuan Zhang, Hanyu Liu, Jingkun Tang, Hongyuan Liu, Qining Guo, and 15 others. 2025. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*.

Weihao Yu, Zhengyuan Yang, Lingfeng Ren, Linjie Li, Jianfeng Wang, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Lijuan Wang, and Xinchao Wang. 2024a. Mm-vet v2: A challenging benchmark to evaluate large multimodal models for integrated capabilities. *arXiv preprint arXiv:2408.00765*.

Yangyang Yu, Haohang Li, Zhi Chen, Yuechen Jiang, Yang Li, Denghui Zhang, Rong Liu, Jordan W Suchow, and Khaldoun Khashanah. 2024b. Finmem: A performance-enhanced llm trading agent with layered memory and character design. In *Proceedings of the AAAI Symposium Series*, volume 3, pages 595–597.

Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yupeng Cao, Zhi Chen, Jordan W. Suchow, Rong Liu, Zhenyu Cui, Zhaozhuo Xu, Denghui Zhang, Koduvayur Subbalakshmi, Guojun Xiong, Yueru He, Jimin Huang, Dong Li, and Qianqian Xie. 2024c. Fincon: A synthesized llm multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. *arXiv preprint arXiv:2407.06567*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567.

Shuo Zhang, Zezhou Huang, and Eugene Wu. 2025a. Data cleaning using large language models. In *2025 IEEE 41st International Conference on Data Engineering Workshops (ICDEW)*, pages 28–32. IEEE.

Wentao Zhang, Lingxuan Zhao, Haochong Xia, Shuo Sun, Jiaze Sun, Molei Qin, Xinyi Li, Yuqing Zhao, Yilei Zhao, Xinyu Cai, Longtao Zheng, Xinrun Wang, and Bo An. 2024a. Finagent: A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist. *arXiv preprint arXiv:2402.18485*.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. 2024b. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*.

Zhihan Zhang, Yixin Cao, and Lizi Liao. 2025b. Xfinbench: Benchmarking llms in complex financial problem solving and reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8715–8758.

Suifeng Zhao, Zhuoran Jin, Sujian Li, and Jun Gao. 2025. Finragbench-v: A benchmark for multimodal rag with visual citation in the financial domain. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4215–4249.

Yilun Zhao, Hongjun Liu, Yitao Long, Rui Zhang, Chen Zhao, and Arman Cohan. 2024. Financemath: Knowledge-intensive math reasoning in finance domains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12841–12858.

Jie Zhu, Junhui Li, Yalong Wen, and Lifan Guo. 2024. Benchmarking large language models on cflue–a chinese financial language understanding evaluation dataset. *arXiv preprint arXiv:2405.10542*.

A Details of UniFinEval

A.1 Examples of Financial Scenarios

We provide supplementary descriptions and illustrative examples for the six core tasks. The following text includes detailed descriptions of each task and the corresponding examples. Since the input content of some questions is too long to be fully presented, we use abbreviated forms for part of the data in the concrete examples, including the examples used in error analysis.

The Financial Evidence Localization task requires models to identify the key evidence needed to answer a question from the given materials and indicate its source. This task mainly evaluates models' fine-grained evidence selection ability under high-density financial inputs, especially their ability to distinguish question-relevant information from background distractors. For detailed examples, please refer to Figure 6.

The Cross-modal Information Consistency Verification task requires models to compare information from text, images, and videos, and determine whether they mutually support each other or whether contradictions omissions or mismatches exist. This task mainly evaluates models' cross-modal semantic alignment ability, information verification ability, and ability to construct reliable evidence chains. For detailed examples, please refer to Figure 7.

The Financial Numerical Derivation task requires models to complete computational derivations based on numerical information in the materials. This task mainly evaluates models' abilities in numerical extraction, computational relation identification, and high-precision reasoning, with a focus on whether they can establish a consistent correspondence among evidence source–calculation process–answer result. For detailed examples, please refer to Figure 8.

The Macro-Industry Trend Reasoning task requires models to synthesize multi-source financial information to judge macro-level or industry-level trend directions and their potential impacts. This task mainly evaluates models' understanding of relationships among financial variables, event transmission logic, and industry evolution paths, with particular emphasis on their integrated analytical ability when multi-source signals are not fully consistent. For detailed examples, please refer to Figure 9.

The Risk Signal Identification task requires

models to identify information with risk-indicative significance from multi-source financial materials, and explain its evidence source, affected object, and impact direction. This task mainly evaluates models' ability to extract risk signals from scattered evidence and associate them with specific financial objects and judgment bases. For detailed examples, please refer to Figure 10.

The Investment Decision Analysis task requires models to conduct evidence-driven financial decision analysis under given constraints based on textual, visual, and video information. This task does not evaluate whether models can directly provide investment advice; instead, it assesses whether they can integrate evidence verification, numerical analysis, risk identification, and conditional constraints to form cautious, coherent, and interpretable decision judgments. For detailed examples, please refer to Figure 11.

A.2 Details of Quality Control

The data processing pipeline of UniFinEval is designed around two goals: first, to preserve the high-density heterogeneous information in real financial materials as much as possible; and second, to ensure that each question has a clear evidence source and a verifiable answer. To this end, we standardize the collected document and video materials separately, and introduce manual review and cross-modal necessity checks during sample construction.

For document materials such as financial statements, annual reports, research reports, and announcements, we first use MinerU (Wang et al., 2024a) for standardized parsing, converting raw PDF or document pages into page-level inputs that preserve text, tables, charts, and layout structures simultaneously. This processing strategy preserves visual structural information in financial reports, such as chart positions, table layouts, title hierarchies, and indicator descriptions, thereby better approximating how report pages are used in real financial analysis. After standardized parsing, we identify high-information-density pages based on pixel density, the proportion of text regions, the distribution of chart elements, and layout completeness. We then remove cover pages, table-of-contents pages, blank pages, low-information explanatory pages, and pages with severe missing chart or textual structures. This ensures that the materials entering the question construction stage contain sufficient financial evidence rather than merely providing background or decorative information.

For video materials, we do not restrict the duration of the original videos, and the video frame rate is standardized to 30 FPS. To reduce redundancy caused by long video inputs while ensuring the completeness of information required by the questions, we crop the original videos into short clips of no longer than one minute before sample construction. Human annotators then confirm that each clip contains all video evidence necessary to answer the corresponding question. Meanwhile, to ensure consistency of video modality inputs across samples, we apply unified preprocessing to videos before model input. Specifically, we perform uniform sampling according to the information density of the video content: for videos with rapid information changes, dense event clues, or frequent chart/subtitle transitions, we extract one frame every 0.5 seconds; for videos with slower visual changes or more stable semantics, we extract one frame every 1–2 seconds. This strategy controls the input length while preserving key temporal clues, subtitle information, chart changes, and event explanations in the videos as much as possible. It should be emphasized that for multi-hop questions involving video and other modalities as joint inputs, samples that can still be directly answered using only text or images after removing the video are discarded during the cross-review stage.

During the question design stage, experts constructed reasoning-oriented samples around real financial information integration processes, avoiding questions that could be answered merely through superficial recognition. For tasks under Financial Evidence Understanding and Verification, the questions are designed to examine models' ability to locate key evidence or judge cross-source information consistency. For tasks under Financial Analytical Reasoning, the questions are required to contain explicit calculation, comparison, or trend judgment demands. For tasks under Financial Risk Sensing and Decision Analysis, the questions are designed to reflect evidence integration, risk identification, or decision analysis under constraints. According to task requirements and data sources, experts constructed both single-turn and multi-turn question–answer pairs, covering both Chinese and English data, so as to support evaluation in cross-lingual scenarios, continuous context understanding, and multi-step reasoning.

To improve data reliability, all samples under-

went a quality control process consisting of expert self-checking, cross-review, and conservative filtering. Experts first conducted self-checks on candidate samples and removed those with insufficient information density, inadequate evidence, or answers obtainable from a single modality alone. The remaining samples were then independently reviewed by other experts for answer correctness and task-label validity, so as to eliminate potential subjectivity in the questions. Samples with ambiguity or unresolved disagreement among experts were directly discarded. Ultimately, UniFinEval retains 5,567 high-quality samples from 7,564 candidate question–answer pairs, striking a balance among data authenticity, financial validity, and evaluation stability. This provides a reliable foundation for diagnosing the capability boundaries of multimodal large language models in high-density financial environments.

A.3 Examples of Environmental Perturbation Simulation

In real financial scenarios, relevant materials are not always presented in clear, regular, and high-quality digital forms. During collection, scanning, photographing, and transmission, they may be affected by practical factors such as device quality, paper wear, partial occlusion, or layout changes. To more comprehensively and realistically evaluate model performance under non-ideal visual conditions, UniFinEval introduces a set of environmental perturbations to some document-image samples during data construction. Specifically, we consider the following types of environmental perturbations.

Capture Distortion When images are captured from non-standard angles, they may exhibit page skew, perspective compression, edge stretching, or local scale variations. This perturbation is used to evaluate models' ability to recognize table structures, chart regions, and text-line arrangements under non-frontal-view inputs. Figure 12 shows an example of this perturbation.

Scanning Noise Images may contain grain noise, blur, compression artifacts, or local contrast degradation under low-quality scanning, compressed transmission, or low-light shooting conditions. This perturbation is used to evaluate models' ability to recognize numbers, text, and fine-grained financial indicators under low signal-to-noise ratio conditions. Figure 13 shows an example of this perturbation.

Partial Occlusion During circulation, annotation, or photographing, financial materials may be affected by stains, stamps, shadows, fingers, or other occluding objects, causing parts of the text, numerical values, coordinate axes, or chart regions to become invisible. This perturbation is used to evaluate models' robust recognition and cautious reasoning abilities under incomplete evidence conditions. Figure 14 shows an example of this perturbation.

Crease and Wear Paper documents may develop creases, wrinkles, edge damage, broken characters, or local shadows during folding, reading, or long-term storage. This perturbation is used to evaluate models' abilities in text recognition, table parsing, and evidence localization under non-flat page layouts and local character degradation. Figure 15 shows an example of this perturbation.

Curved-Surface Deformation Thick reports, bound materials, or unflattened paper may produce nonlinear bending during photographing, causing text lines, table boundaries, or chart regions to deform along the page surface. This perturbation is used to evaluate models' abilities in layout understanding, character correction, and structured information extraction under non-planar document layouts. Figure 16 shows an example of this perturbation.

All perturbed samples are manually reviewed to ensure that the key evidence required by the questions remains visible and that the standard answers can still be verified from the given materials. If a perturbation causes decisive evidence to be missing, changes the answer, or makes the question unverifiable, the corresponding sample is discarded.

B Evaluation Details

B.1 Model Details

We list details of the MLLMs evaluated using UniFinEval in Table 4 and report the inference configurations in Table 5 , including the exact model identifiers and decoding parameters. Parameters that are not exposed or explicitly configured through the corresponding APIs are denoted as "/".

B.2 Judge Bias Analysis

To provide a more detailed assessment of the reliability of our automatic evaluation protocol, we report the complete multi-judge validation results in this section. Specifically, we compare the judgments of Qwen-Max, DeepSeek-V4-Pro, and financial experts on the same dataset, focusing on

both instance-level agreement and model-level evaluation consistency. We further examine whether the use of Qwen-Max as the primary evaluator introduces systematic score deviations in favor of Qwen-family models.

Cross-Judge Agreement. At the instance-response level, the three evaluators exhibit a high degree of consistency as shown in Figure 5. The pairwise exact agreement rate is 97.3% between Qwen-Max and DeepSeek-V4-Pro, 96.9% between Qwen-Max and the financial experts, and 97.1% between DeepSeek-V4-Pro and the financial experts. In addition to the high instance-level agreement, the model-level accuracy profiles and rankings produced by the three evaluators are also highly consistent. These results indicate that the overall evaluation conclusions remain stable across heterogeneous judges and are not substantially affected by the choice of evaluator.

Model-Family Bias Analysis. We further examine whether Qwen-Max systematically favors models from the same model family. Specifically, we compute the score difference between Qwen-Max and the finance-domain experts for each evaluated model, where a positive value indicates that Qwen-Max assigns a higher score than the expert evaluation as shown in Table 6. For the two Qwen-family models, the average score difference is $+0.20$ percentage points. In comparison, the average difference for the remaining eight non-Qwen models is $+0.275$ percentage points.

The Qwen-family models therefore do not exhibit a larger positive score deviation than the non-Qwen models. Taken together with the high cross-judge agreement and the consistent model-level rankings, these results indicate that the principal ranking and performance conclusions of UniFinEval are robust to the choice of evaluator.

B.3 Additional Evaluation Setting

In addition to the zero-shot setting, we also include an evaluation under the Chain-of-Thought (CoT) setting, where models are required to perform step-by-step reasoning before generating the final answer. This setting is designed to evaluate models' capabilities in multi-step reasoning and explanation.

The results under the Chain-of-Thought (CoT) setting are shown in Table 7. Compared with the zero-shot setting, model performance under the CoT setting does not show a clear improvement. Some models, such as Gemini-3-Pro and ChatGPT-5.1, even experience performance degradation. This result indicates that, in high-density financial multimodal tasks, explicit reasoning chains do not always effectively enhance model performance. For stronger models, longer reasoning processes may introduce redundant intermediate steps, leading to the loss of key information, disruptions in cross-modal evidence chains, or deviations in financial logic, thereby weakening their stability in multi-step tasks.

From the task perspective, CoT yields modest gains in Financial Evidence Localization, Cross-modal Information Consistency Verification, and Macro-Industry Trend Reasoning, while substantially improving Investment Decision Analysis for most models. In contrast, performance decreases notably in Financial Numerical Derivation and Risk Signal Identification, suggesting that explicit reasoning may introduce additional error propagation in high-precision numerical reasoning and risk-signal extraction. At the model level, Claude-Sonnet-4.5, Grok-4-1-Fast-Reasoning, and MiniCPM-V-4.5 achieve overall improvements under the CoT setting, whereas several other models experience varying degrees of degradation.

Overall, the impact of CoT on financial multimodal tasks is clearly model-dependent and task-dependent. It can help some models improve evidence organization and trend judgment, but in tasks requiring high-precision numerical derivation, cross-modal multi-hop integration, and complex investment decision-making, it may also lead to overly long reasoning chains, logical discontinuities, and accumulated errors. Therefore, reasoning-enhancement methods in financial scenarios should not simply rely on generic CoT templates; instead, they need to be specifically designed according to task types, model scales, and financial business logic.

C More Details of Error Analysis

This section provides a detailed description of the typical error types exhibited by large models in financial scenarios. Through concrete error cases, we further reveal the capability deficiencies exposed during model evaluation.

High-Density Information Extraction Error refers to the phenomenon where a model fails to accurately identify, locate, or extract the key evidence required to answer a question when processing high-density information inputs. Model out-

puts usually exhibit missing key numerical values, misreading chart elements, confusing row–column relationships in tables, overlooking fine-grained visual clues, or extracting incorrect information from irrelevant regions. Such errors are not caused by a lack of subsequent reasoning capability, but rather by the model's failure to capture truly relevant information from complex financial materials at the information acquisition stage. This type of error directly undermines the evidential foundation for subsequent computation, verification, or decision analysis. Figure 17 shows an example of this error type.

Cross-modal Evidence Integration Error refers to the phenomenon where a model fails to correctly establish correspondences among evidence from different modalities, such as text, images, or videos. The model may identify partial unimodal information separately, but make errors when aligning, comparing, and integrating such information. For example, it may incorrectly match chart indicators with textual descriptions, omit key evidence from one modality, misjudge consistency relationships among multi-source information, or introduce irrelevant evidence during comprehensive judgment. Such errors are not simply failures in information reading; rather, they indicate the lack of stable cross-modal alignment and information fusion capabilities. This type of error may lead the model to produce judgments that appear plausible on the surface but are unsupported by a valid evidence chain. Figure 18 shows an example of this error type.

Financial Reasoning and Computation Error refers to the phenomenon where a model makes mistakes in calculation, indicator interpretation, or financial business logic judgment. This type of error includes calculation mistakes, incorrect formula selection, confusion over indicator definitions, and misuse of financial concepts such as growth rate, return rate, or risk indicators. Such errors stem from the model's failure to accurately understand financial numerical relationships, business semantics, and the applicable conditions of indicators, reflecting its instability in financial professional reasoning and high-precision computation. Figure 19 shows an example of this error type.

Ungrounded Reasoning Hallucination refers to the phenomenon where a model fails to answer strictly based on the given input and introduces information that does not exist in, or cannot be derived from, the provided materials. This type of error is usually manifested as the model invoking external commonsense knowledge, parametric knowledge, or unsupported assumptions, and may even directly fabricate data or conclusions. The core issue is the generation of content without support from input evidence. This error is particularly dangerous in financial tasks, as it may cause models to produce seemingly professional but unverifiable risk judgments or investment suggestions. Figure 20 shows an example of this error type.

Response Semantic Misalignment refers to the phenomenon where the model's final output is semantically inconsistent with the intent of the question. The model may have extracted relevant information and completed locally reasonable reasoning, but its final output does not address the actual target required by the question, or deviates from the required output format. For example, when the question asks for identifying the source of risk, the model instead answers the trend direction; when the question asks for a comparison between two indicators, the model only explains one of them. Such errors reflect the model's insufficient ability to understand financial question intent, task constraints, and the required granularity of answers. It is therefore necessary to improve the model's stability in task semantic alignment and answer control. Figure 21 shows an example of this error type.

D Ablation Studies

UniFinEval is designed to investigate whether MLLMs can effectively leverage evidence distributed across all designated input modalities, rather than merely being presented with multiple modalities while the corresponding questions remain directly solvable from a single information source. Accordingly, during benchmark construction, annotators explicitly assess modality necessity and discard samples that can still be answered directly after any required modality is removed.

To empirically validate this construction criterion, we conduct a modality-necessity analysis using five high-performing MLLMs: Gemini-3-Pro, ChatGPT-5.1, Claude-Sonnet-4.5, Qwen3-VL-235B-A22B-Thinking, and InternVL3_5-241B-A28B. For samples containing text, image, and video inputs, we compare the complete-input setting against three modality-removal variants in which text, image, or video is removed individually. For image–video samples, we compare the complete-input setting with two single-modality

variants, namely video-only and image-only inputs. Except for the available input modalities, all other experimental conditions are kept unchanged, including the evaluation samples, prompts, decoding configurations, preprocessing procedures for the retained modalities, and evaluation protocol.

As shown in Table 8, under the complete text–image–video setting, the evaluated models achieve accuracies ranging from 59.8% to 76.4%, with an average accuracy of 68.0%. After removing text, image, or video, the average accuracy decreases sharply to 4.0%, 1.82%, and 0.94%, respectively. A similar pattern is observed for image–video samples: the average accuracy is 64.12% under the complete-input setting, but drops to 2.44% when only video is retained and to 0.54% when only image is retained. Notably, this substantial degradation is consistently observed across all five evaluated models rather than being driven by any particular model family.

These results further demonstrate that the multimodal design of UniFinEval does not simply juxtapose heterogeneous information sources. Instead, the benchmark effectively constructs evaluation scenarios in which answer-relevant evidence is distributed across multiple modalities and must be jointly integrated to derive the correct outcome. The modality-removal results therefore provide empirical support for the validity and soundness of UniFinEval's multimodal task construction.

E Prompts Used in This Study

We provide representative prompt examples for evaluation. Specifically, the prompt examples for evaluation are shown in Table 9, Table 10 and Table 11.

Table 4: Models evaluated in this paper. The "Access" column shows whether we have full access to the model weights or we can only access through API. The “Version Date” column shows the release date of the corresponding version of the model we evaluated.

Category	Model	Creator	Parameter	Access	Version Date
Closed-Source	Gemini-3-Pro	Google	Undisclosed	API	2025.11
	ChatGPT-5.1	OpenAI	Undisclosed	API	2025.11
	Grok-4-1-Fast-Reasoning	xAI	Undisclosed	API	2025.11
	Claude-Sonnet-4.5	Anthropic	Undisclosed	API	2025.9
Open-Source	Qwen3-VL-235B-A22B-Thinking	Alibaba Cloud	235B	Weights	2025.11
	Qwen3-VL-32B-Thinking	Alibaba Cloud	32B	Weights	2025.11
	InternVL3_5-241B-A28B	Shanghai AI Lab	241B	Weights	2025.8
	InternVL3_5-30B-A3B	Shanghai AI Lab	30B	Weights	2025.8
	Llama-3.2-11B-Vision	Meta AI	11B	Weights	2024.9
	MiniCPM-V-4.5	OpenBMB	8B	Weights	2025.9

Table 5: Inference configurations of the closed-source models accessed through APIs. Parameters not exposed or explicitly configured through the corresponding APIs are denoted as “/”. The Temp.” stands for Temperature.

Display Name	Exact Model ID	Temp.	Top-p	Top-k	Max Output
Gemini-3-Pro	gemini-3-pro-preview	1.0	0.95	64	65,536
ChatGPT-5.1	gpt-5.1-2025-11-13	1.0	/	/	128,000
Claude-Sonnet-4.5	claude-sonnet-4-5-20250929	1.0	/	/	64,000
Grok-4-1-Fast-Reasoning	grok-4-1-fast-reasoning	1.0	/	/	/

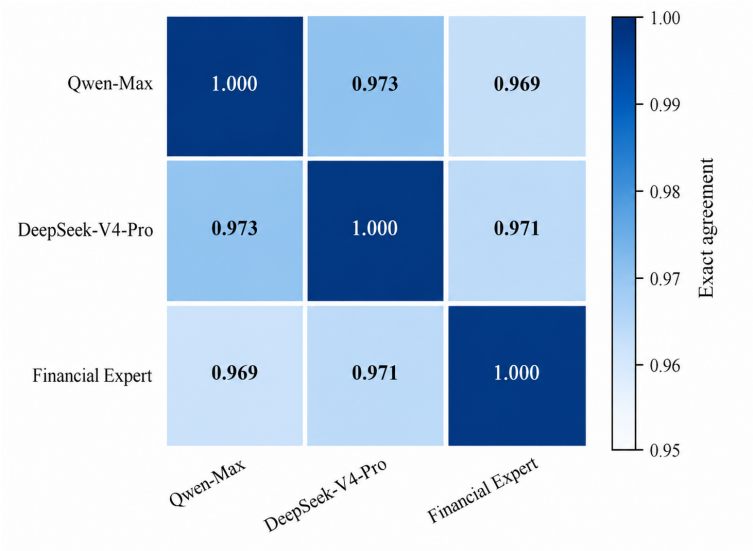


Figure 5: Pairwise exact agreement among Qwen-Max, DeepSeek-V4-Pro, and finance-domain experts. All pairwise agreement rates exceed 96.9%, indicating high consistency across heterogeneous evaluators.

Table 6: Model-level accuracy scores (%) produced by Qwen-Max, DeepSeek-V4-Pro, and finance-domain experts in the multi-judge validation analysis.

Model	**Qwen-Max**	**DeepSeek-V4-Pro**	**Financial Expert**
Gemini-3-Pro	76.3	75.8	76.0
ChatGPT-5.1	71.1	70.6	70.8
Qwen3-VL-235B-A22B-Thinking	71.1	70.7	71.0
Qwen3-VL-32B-Thinking	67.4	66.9	67.1
Claude-Sonnet-4.5	65.7	65.3	65.4
InternVL3_5-241B-A28B	63.1	62.8	62.7
MiniCPM-V-4.5	59.0	58.6	58.8
InternVL3_5-30B-A3B	49.3	49.0	49.1
Grok-4-1-Fast-Reasoning	41.3	40.8	40.9
Llama-3.2-11B-Vision	19.8	20.0	19.7

Table 7: Performance under CoT setting on UniFinEval(%). The highest model result in each task column is highlighted with a blue background, the second-best with a pink background, and the third-best with a green background.

Model	**Financial Evidence Understanding and Verification**		**Financial Analytical Reasoning**		**Financial Risk Sensing and Decision Analysis**		**Average**
	Financial Evidence Localization	Cross-modal Information Consistency Verification	Financial Numerical Derivation	Macro-Industry Trend Reasoning	Risk Signal Identification	Investment Decision Analysis	
Gemini-3-Pro	80.0	78.2	76.0	75.2	67.2	79.0	**75.8**
ChatGPT-5.1	76.4	75.7	69.7	72.2	54.9	76.2	**70.7**
Qwen3-VL-235B-A22B-Thinking	76.5	73.3	58.5	71.5	65.1	70.2	**68.7**
Qwen3-VL-32B-Thinking	67.1	69.6	62.1	69.4	51.8	67.6	**64.5**
Claude-Sonnet-4.5	74.2	68.0	69.6	65.0	64.7	72.8	**68.9**
InternVL3_5-241B-A28B	62.7	55.0	56.8	71.2	67.8	64.6	**62.7**
MiniCPM-V-4.5	53.6	59.0	53.6	68.7	61.4	65.3	**60.1**
InternVL3_5-30B-A3B	44.0	55.4	47.6	53.0	39.8	52.4	**48.7**
Grok-4-1-Fast-Reasoning	47.0	45.8	28.9	54.3	40.9	41.4	**42.5**
Llama-3.2-11B-Vision	22.5	25.1	13.7	32.6	4.3	19.1	**19.3**

Table 8: Modality-necessity validation on UniFinEval (%). For text–image–video samples, each required modality is individually removed from the full input. For image–video samples, the full input is compared with single-modality variants. T, I, and V denote text, image, and video, respectively. All other experimental conditions are kept unchanged.

	Text + Image + Video				**Image + Video**		
Model	**Full**	**w/o T**	**w/o I**	**w/o V**	**Full**	**V Only**	**I Only**
Gemini-3-Pro	76.4	4.2	1.6	1.4	73.2	3.8	0.4
ChatGPT-5.1	70.2	3.8	1.2	0.6	66.5	2.8	0.4
Claude-Sonnet-4.5	64.5	3.7	3.3	1.1	60.7	2.4	0.8
Qwen3-VL-235B-A22B-Thinking	69.1	4.4	1.0	0.2	64.9	1.2	0.7
InternVL3_5-241B-A28B	59.8	3.9	2.0	1.4	55.3	2.0	0.4
Average	**68.00**	**4.00**	**1.82**	**0.94**	**64.12**	**2.44**	**0.54**

Financial Evidence Localization



文本：品种观点品种主要观点本周策略上周策略棉花国际方面，12月上旬USDA出口销售大幅增长，中国增量明显，但装运量不算太多。美棉上市检验中期，上市进度同比落后幅度持平，皮马棉上市速度放缓。分检质量指标较高，但港口看大货反映实际棉包质量不及质量检验报告。短期内美棉维持震荡走势。国内方面，棉花市场依旧缺乏明确的利空因素，下游虽在棉价带动下被动涨价，整体出货一般，但整体压力也尚可，负反馈暂不明显，本周郑棉增仓幅度较大，上方仍需注意阶段性风险，不过在2026/27年度新疆棉种植面积预期下降的情况下，长期仍相对乐观。综上，短期棉价或稳中偏强震荡。震荡偏强区间震荡偏强2025-12-18受中美贸易战关税升级影响，下游需求边际走弱新棉产量超预期增加继续形成较强压力，下游持续恶化旧作库存偏紧，叠加宏观偏暖中美瑞士谈判，关税下滑，宏观利好新棉预售陆续交货，预售较多也缓解短期集中供应压力，现货销售基差坚挺，郑棉下方支撑仍强供应端新棉收购上量后套保压力较大，需求端产业下游对旺季信心不强2026/27年度新疆棉种植面积预期下降，且下游负反馈不强2025-12-06 下年度美棉签约量下滑，加重了市场对于美棉新作后续出口减少的担忧。美棉生长进度仍旧超前。供需双弱，价格继续承压下行"，
问题：根据文本中对“2025/26年度”的定义，以及视频第374秒展示的全球棉花供需平衡表，该年度的期初库存数据应参考哪一行？请直接写出该行对应的年份和月份。
Question: Based on the textual definition of the “2025/26 marketing year” and the global cotton supply–demand balance table shown at 374 seconds in the video, which row should be used as the reference for the beginning inventory of that year? Please directly provide the year and month corresponding to that row.
答案：2025/26（12月）
Answer: 2025/26 (December)

Figure 6: This example demonstrates the Financial Evidence Localization task. It requires the model to identify the exact evidential row needed for answering the question by aligning the textual definition of the financial year with the corresponding row in a video-based supply–demand table. Rather than emphasizing numerical derivation, this task focuses on precise evidence grounding, temporal reference alignment, and the ability to retrieve the correct source position from high-density multimodal financial materials.

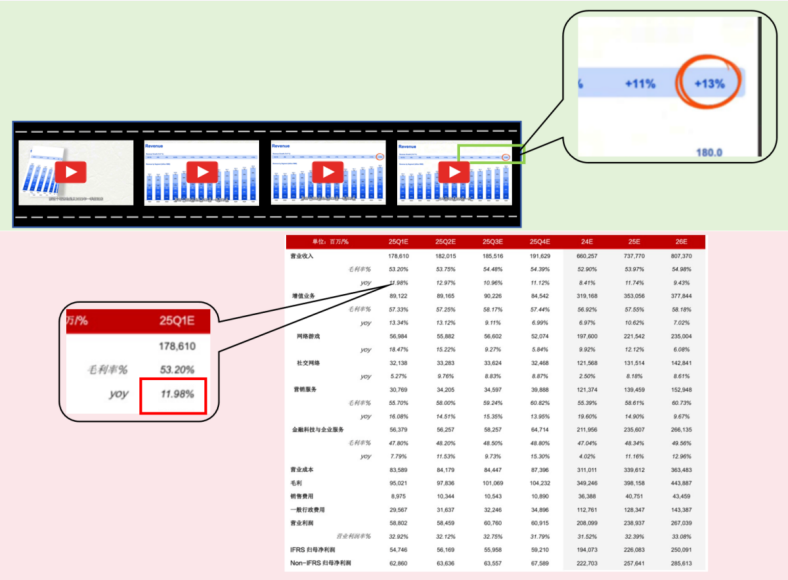
Cross-modal Information Consistency Verification



问题：世界上规模最快突破两百亿美元的主动管理型ETF的股息与美盈森的差值是多少?
Question: What is the difference between the dividend of the world's fastest-growing actively managed ETF to surpass $20 billion and that of Mei Ying Sen?
答案：-6.19%
Answer: -6.19%

Figure 7: This example demonstrates the Cross-modal Information Consistency Verification, designed to evaluate a model's ability to verify fine-grained consistency between textual financial descriptions and chart-based numerical evidence. It requires targeted extraction of relevant parameters from text, precise localization of corresponding values in financial charts, and rigorous execution of multi-step mathematical reasoning. The task thus assesses the reliability of the full cross-modal reasoning pipeline, from evidence alignment to numerical inference and final answer derivation.

Financial Numerical Derivation



问题：自2022 年第一季度以来，腾讯季度营收增速的最高值与2025 年第一季度预测的同比营收增长率之间相差多少个百分点？（单位：%，保留两位小数）

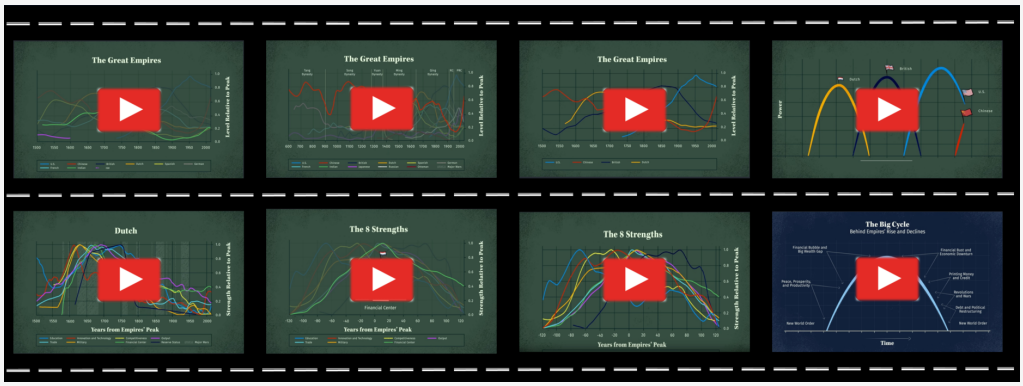
Question: What is the difference, in percentage points, between Tencent's fastest quarterly revenue growth rate since Q1 2022 and the predicted YoY revenue growth rate for Q1 2025? (Unit: %, rounded to two decimal places)

答案：-1.02%

Answer: -1.02%

Figure 8: This example demonstrates the Financial Numerical Derivation task, designed to evaluate a model's ability to perform precise financial calculation over heterogeneous multimodal evidence. It requires the model to identify task-relevant indicators from both video frames and tabular financial reports, align historical and forecasted values across different sources, and execute numerical comparison with strict unit and precision constraints. The task therefore examines whether models can establish a faithful evidence–formula–answer chain in high-density financial materials.

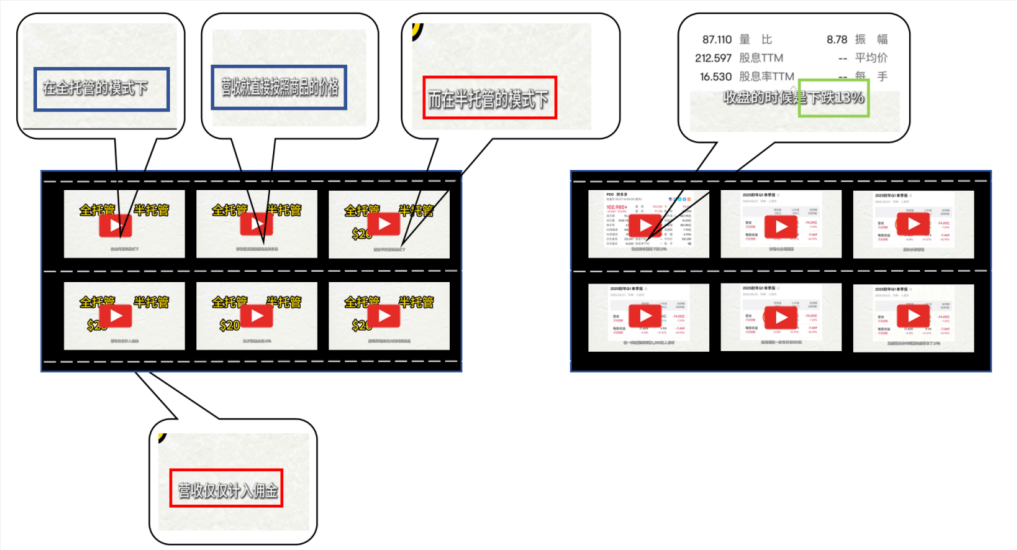
Macro-Industry Trend Reasoning



Text: As I described in my book Principles for Dealing with the Changing World Order, I automated my way looking at the cause effect relationships that are driving both improvements in and worsenings of countries' conditions so that data is fed into a computer that analyzes it and writes a summary of the current conditions and the long-term prospects for each country. ... ... The table below shows our aggregate country power gauge and the major drivers, as well as the rank of each measure of power across 24 major countries today and the trajectory over the past twenty years. To understand a country, we start by looking at the big cycles , as well as measures of power that both reflect and drive the rise and fall of a country. While we refer to these factors individually, they are not separate; they interact with and reinforce one another to move a country along its cycle. For the United States, the big cycles look mostly unfavorable. The United States is in an unfavorable position in its economic and financial cycles, with a high debt burden and relatively low expected real growth over the next 10 years (1.3% per year). The United States has significantly more foreign debts than foreign assets (net IIP is -68% of GDP). Non-financial debt levels are high (274% of GDP), and government debt levels are high (127% of GDP). The bulk (99%) of these debts are in its own currency, which mitigates its debt risks. The ability to use interest rate cuts to stimulate the economy is low (short rates at 0.1%), and the country is already printing money to monetize debt. That said, being the world's leading reserve currency is a large benefit to the US. If this were to change, it would significantly weaken the US position. Internal disorder is a high risk. Wealth, income, and values gaps are large (relative to countries of similar per capita income levels). Regarding Inequality—the top 1% and top 10% in the United States capture 19% and 45% of income (respectively the 8th and 11th highest share across major countries). ... ..., Half of the measure captures the absolute quantity of educated people at various levels and about half is placed on quality such as higher education rankings, test scores, and average years of education. The US ranks highest in this gauge (driven by strong absolute and relative measures of higher education), with China close behind (due to its large number of educated people). Financial Center: This gauge measures the level of development and sizes of a country's financial markets and financial center. We look at absolute measures of transaction shares and market capitalizations, as well as external indices of financial center cities. The US remains the top-ranked power in this metric by a significant margin (driven primarily by its very large share of world equity and debt mar -kets), with China and Europe ranking second and third, respectively. Reserve Currency Status: This gauge measures the extent to which a country's currency operates as a global reserve currency. We measure reserve currency status by the share of transactions, debts, and central bank reserves that are denominated or held in a country's currency. Similar to financial center status, the US remains the top-ranked power in this metric by a significant margin, with Europe and Japan ranking second and third, respectively. ... ..., In case it is helpful or interesting to you, you can review those scores below.33 In a few cases where there were no quality measures, I had to create quality measures by adjusting the quantity for a country's population, turning it into a per capita measure. We did not give reserve currency status scores to the countries that share the euro, which is why those measures are displayed as dashes. 4 Because the notion of competitiveness is inherently relative, we only show the total score for this measure.
Question: Synthesizing the analytical logic regarding the 'Big Cycle' and the 'Eight Major Strengths' in the text, as well as the average evolutionary trends of the 'Eight Major Strengths' for all empires before and after their respective peaks as shown in Figure 5, please infer: After an empire's power reaches its peak, which 'strength' indicator declines the fastest (i.e., requires the shortest time to drop from its peak to the 0.2 level)? Please substantiate your argument by combining the trend lines in the video with the implications in the text regarding the vulnerability of 'financial center' status.
Answer: Financial Center

Figure 9: This example demonstrates the Macro-Industry Trend Reasoning task, designed to evaluate a model's ability to synthesize long-form textual analysis with video-based temporal trends. It requires extracting the evolution patterns of multiple macro-financial indicators from video curves, interpreting their post-peak dynamics under the analytical framework provided in the text, and identifying the indicator with the fastest deterioration. The task thus assesses cross-modal temporal reasoning, trend comparison, and the ability to connect visual trajectories with higher-level financial-economic mechanisms.

Risk Signal Identification



问题：请问拼多多2025年Q1营业收入波动的主要原因是什么?
Question: What are the primary reasons for Pinduoduo's revenue fluctuations in Q1 2025?
答案：拼多多商业模式由全托管转为半托管。
Answer: Pinduoduo's (Temu) Business Model Transition from Fully Managed to Semi-Managed.

Figure 10: This example demonstrates the Risk Signal Identification task, designed to evaluate a model's ability to detect risk-indicative signals from multiple financial videos. It requires integrating scattered evidence across video segments, distinguishing superficial revenue fluctuations from underlying business-model changes, and identifying the primary factor that may affect future operational risk. The task therefore assesses evidence aggregation, causal attribution, and risk-oriented interpretation under multi-source multimodal inputs.

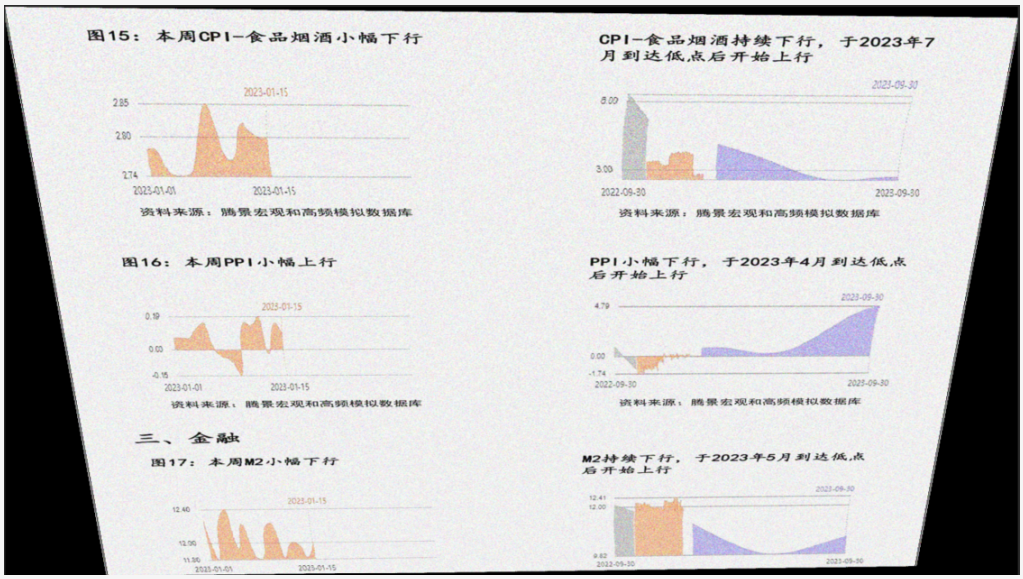
Investment Decision Analysis



文本： 2025 年白酒行业调整加剧，公司主动应市。八代五粮液动态实施渠道政策，并多次加强渠道激励，以缓解需求下降和批价走弱带来的经销商利润压力，同时依托其在千元价位的周转优势提振渠道信心。2026 年，公司计划进一步推进精细化渠道运作，以维护八代五粮液的量价平衡并巩固其千元价位核心地位。
问题： 综合文本中2025 年“主动应市调整”的策略、图片中资产负债表体现的深厚资产基础，以及视频第98秒展示的五粮液营业成本率自2015 年后持续下降的趋势，推断公司在2025–2026 年期间最可能针对“八代五粮液”采取何种关键渠道政策？并说明该政策将如何影响“营业成本率”和“现金循环周期”，最终实现“量价平衡”。
Text: In 2025, the liquor industry underwent intensified adjustment, and the Company proactively adapted to market conditions. For Eighth-Generation Wuliangye, it dynamically adjusted channel policies and repeatedly strengthened channel incentives to relieve distributor profitability pressure caused by weaker demand and declining wholesale prices. Its turnover advantage in the RMB 1,000 price segment also helped restore channel confidence. In 2026, the Company plans to further refine channel operations to maintain a balance between volume and price and consolidate the product's position in the RMB 1,000 segment.
Question: Considering the Company's "proactive market adjustment" strategy in 2025, the strong asset base reflected in the balance-sheet image, and the persistent decline in Wuliangye's operating cost ratio since 2015 shown at 98 seconds in the video, infer the key channel-policy measure most likely to be adopted for Eighth-Generation Wuliangye during 2025–2026. Explain how this policy would affect the operating cost ratio and the cash conversion cycle, thereby helping achieve a balance between volume and price.
答案： 最可能的措施是实施精细化渠道管理，包括加强经销商激励、终端包量激励，并严格治理窜货和渠道压货。该措施可减少渠道库存积压、加快库存周转和回款，从而缩短现金循环周期；同时，通过稳定批价、改善终端动销和提高销售效率，可降低库存及低效供给带来的成本压力，帮助营业成本率维持在较低水平。由此，公司能够在保障渠道利润和信心的同时控制供给节奏，实现销量与价格的动态平衡。
Answer: The most likely measure is refined channel management, including stronger distributor incentives, terminal volume-based incentives, and tighter control over cross-regional selling and channel stuffing. By reducing channel inventory accumulation and accelerating inventory turnover and cash collection, the policy can shorten the cash conversion cycle. Meanwhile, stabilizing wholesale prices, improving sell-through, and increasing sales efficiency can reduce cost pressure associated with excess inventory and inefficient supply, helping keep the operating cost ratio at a relatively low level. Together, these effects allow the Company to protect distributor profitability and confidence while controlling supply, thereby achieving a dynamic balance between sales volume and price.

Figure 11: An example of the Investment Decision Analysis task. The model must integrate textual strategy, balance-sheet evidence, and historical cost trends from video to infer a channel policy and explain its financial transmission mechanism.

Capture Distortion



问题：综合图15、16、17右侧子图对PPI、M2、CPI-食品烟酒三个指标触底反弹时间的描述（PPI于2023年4月、M2于2023年5月、CPI-食品烟酒于2023年7月到达低点后开始上行），以及左侧子图显示的2023年1月15日当周三个指标均小幅下行的特征，判断在2023年1月15日时点，哪个指标的下行期剩余时间最短？需严格依据图像文字信息进行跨指标时序比较和趋势阶段定位。

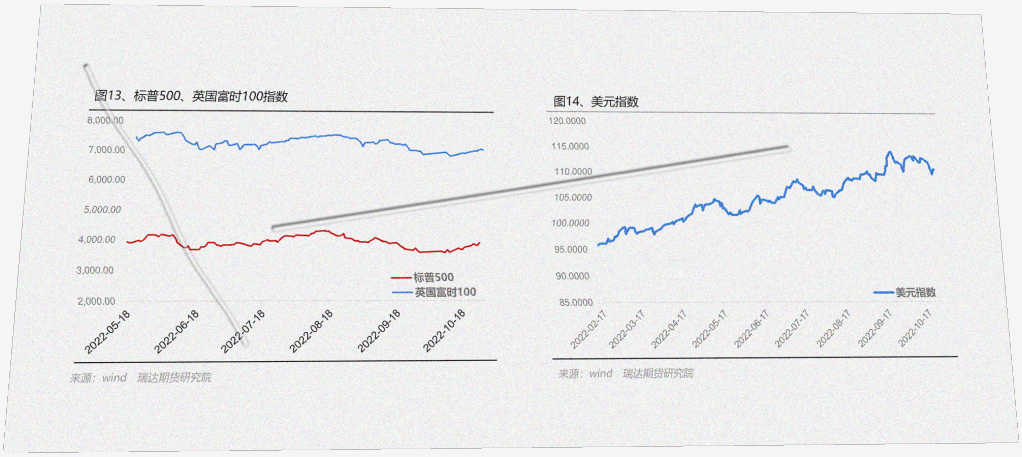
Question: Based on a comprehensive analysis of the descriptions in the right-hand subplots of Figures 15, 16, and 17 regarding the bottoming-out and rebound timing of three indicators—PPI, M2, and CPI-Food, Tobacco, and Liquor (specifically, PPI reached its low point and began to rise in April 2023, M2 in May 2023, and CPI-Food, Tobacco, and Liquor in July 2023)—combined with the characteristic shown in the left-hand subplots where all three indicators experienced a slight decline during the week of January 15, 2023, determine which indicator had the shortest remaining time in its downward phase as of the time point of January 15, 2023. You are required to strictly base your analysis on the text information within the images to perform a cross-indicator temporal comparison and trend phase positioning.

答案：PPI

Answer: PPI

Figure 12: This example illustrates the Capture Distortion perturbation, designed to simulate geometric deformation caused by non-frontal image acquisition. The distorted document image contains perspective skew, local scale variation, and non-parallel page boundaries, while the question still requires accurate chart recognition and cross-indicator temporal comparison. This perturbation evaluates whether models can preserve layout understanding, chart localization, and trend-based reasoning under realistic capture conditions.

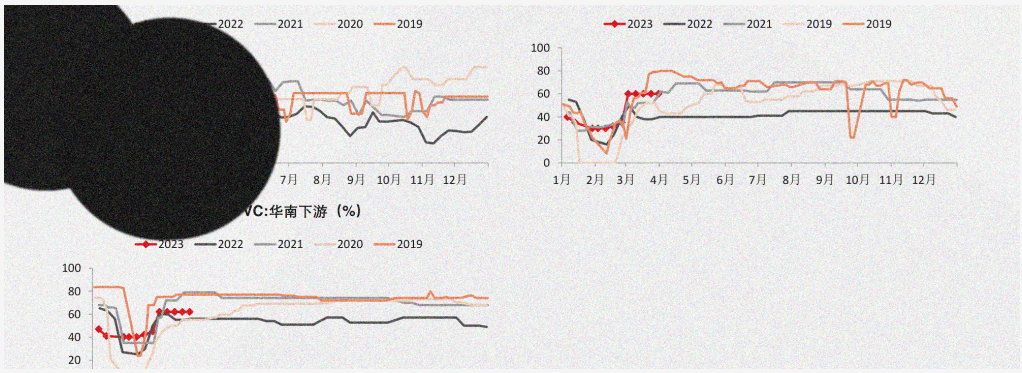
Scanning Noise



Round 1:
问题：在2022年5月18日至2022年10月18日期间，美元指数（图14）呈现持续上升趋势。结合图13中标普500指数（红色线）与英国富时100指数（蓝色线）的折线形态，分析：两条指数曲线在该时间段内是否均呈现整体下行趋势?
Question: From May 18, 2022, to October 18, 2022, the U.S. Dollar Index (Figure 14) exhibited a sustained upward trend. Based on the line patterns of the S&P 500 Index (red line) and the FTSE 100 Index (blue line) in Figure 13, analyze whether both index curves showed an overall downward trend during this period.
答案：否
Answer: No
Round 2:
问题：基于上述分析，在美元指数上升期间，标普500指数和英国富时100指数中，哪一个表现出更强的韧性，即受到的负面冲击相对更小?
Question: Based on the above analysis, which of the two indices—the S&P 500 or the UK FTSE 100—demonstrated greater resilience during the rise in the U.S. Dollar Index, that is, experienced a relatively smaller adverse impact?
答案：英国富时100
Answer: UK FTSE 100

Figure 13: This example illustrates the Scanning Noise perturbation, designed to simulate visual degradation introduced by low-quality scanning, compression, or low-light acquisition. The financial chart remains answerable but contains background noise, contrast degradation, and reduced visual clarity. This perturbation evaluates whether models can robustly recover fine-grained textual labels, numerical trends, and chart relationships under low signal-to-noise visual conditions.

Partial Occlusion



问题：基于提供的金融研报图像（包含左上角、右上角和底部三个子图，均展示PVC华南下游开工率（%），请分析：在右上角子图中，2019年开工率在2月的下降趋势与2023年在3月的上升趋势相比，哪一段的波动更剧烈?

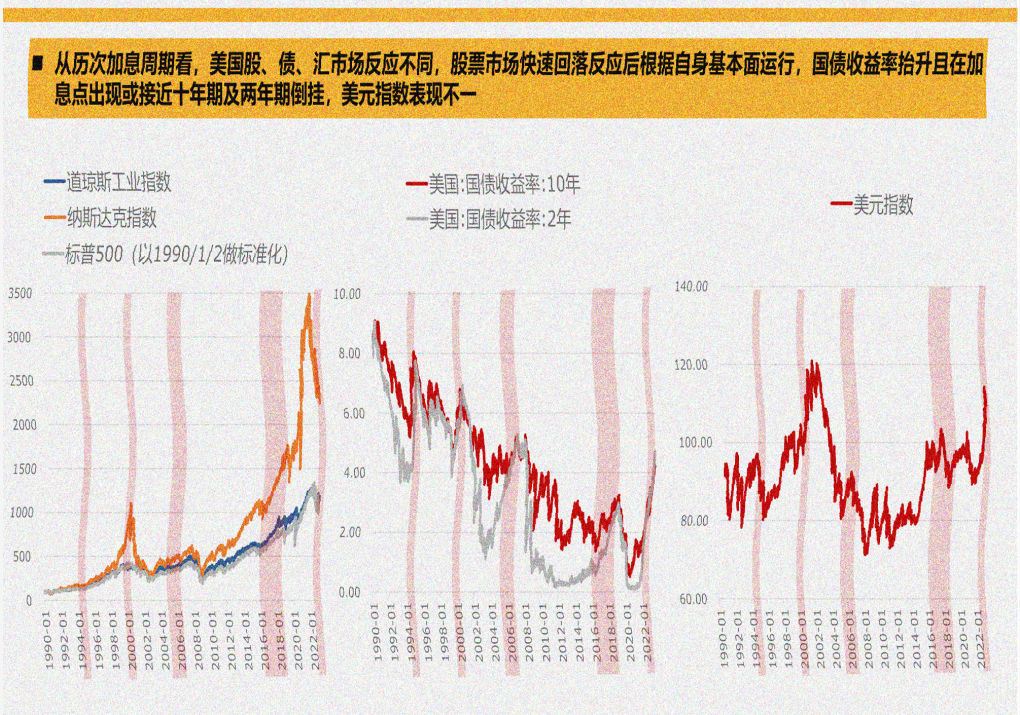
Question: Based on the provided financial research report image (which contains three subplots located at the top-left, top-right, and bottom, all showing the PVC South China downstream operating rate (%), please analyze the following: In the top-right subplot, when comparing the downward trend in the operating rate in February 2019 with the upward trend in March 2023, which period exhibits more pronounced volatility?

答案：2019年2月的下降段更剧烈

Answer: The decline in February 2019 was more pronounced.

Figure 14: This example illustrates the Partial Occlusion perturbation, designed to simulate missing visual evidence caused by stains, shadows, stamps, or other occluding objects during document circulation or photographing. Although part of the chart is obscured, the model must still rely on the remaining visible evidence to compare trend volatility and derive a cautious answer. This perturbation evaluates robust evidence localization and reasoning under incomplete financial document inputs.

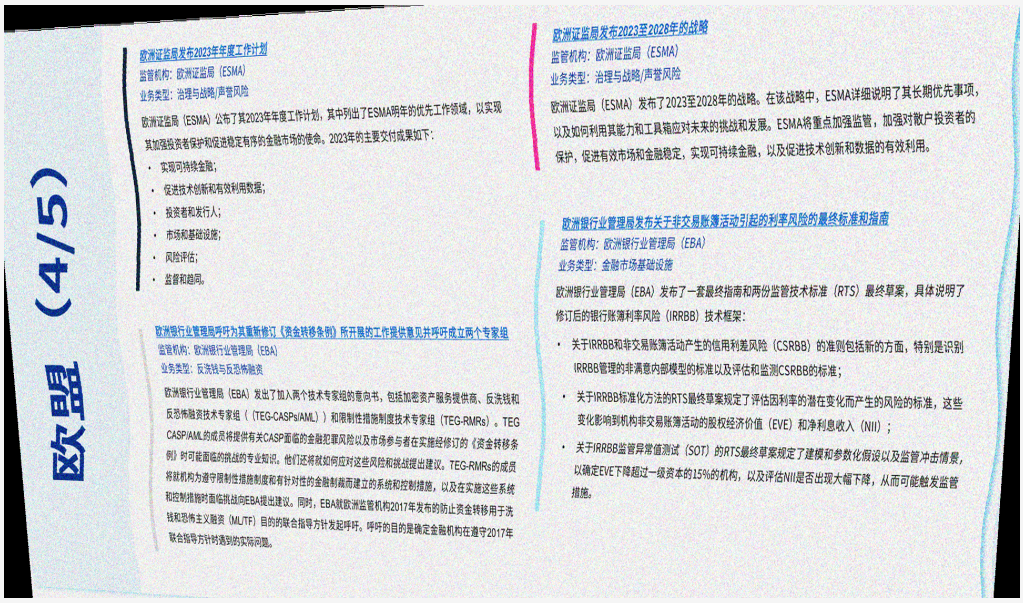
Crease and Wear



问题：根据图像文字说明及中间子图，历次加息周期（粉色阴影区间）内，美国10年期与2年期国债收益率曲线是否发生过倒挂?
Question: According to the text descriptions in the image and the middle subplot, did the U.S. 10-year and 2-year Treasury yield curve invert during past rate hike cycles (pink shaded areas)?
答案：是
Answer: Yes

Figure 15: This example illustrates the Crease and Wear perturbation, designed to simulate document degradation caused by folding, repeated handling, or long-term storage. The perturbed image contains crease-like artifacts and local visual interference over financial curves and explanatory text, requiring the model to maintain accurate recognition of chart structures and temporal patterns. This perturbation evaluates robustness in text recognition, chart parsing, and evidence grounding under physically degraded document conditions.

Curved-Surface Deformation:



问题：当欧洲银行业管理局(EBA)依据其非交易账簿利率风险指南判定某银行股权经济价值(EVE)下降幅度超过一级资本阈值并触发监管措施时，欧洲证券和市场管理局(ESMA)在2023年年度工作计划中确立的哪项优先工作领域将直接启动以协调跨境监管响应?

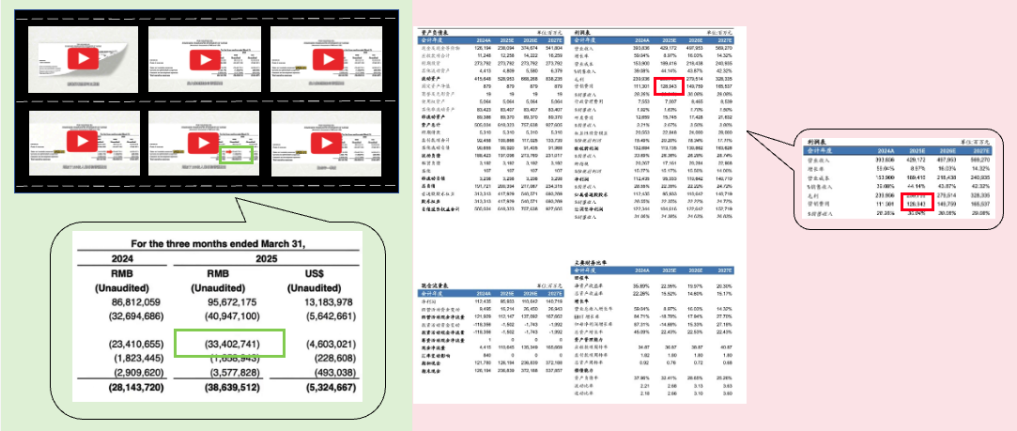
Question: When the European Banking Authority (EBA), in accordance with its Guidelines on Interest Rate Risk in the Banking Book (IRRBB), determines that the decline in a bank's Economic Value of Equity (EVE) exceeds the Tier 1 capital threshold and triggers supervisory measures, which priority area established in the European Securities and Markets Authority's (ESMA) 2023 Annual Work Programme would be directly initiated to coordinate the cross-border regulatory response?

答案：监督和趋同

Answer: Supervision and Convergence

Figure 16: This example illustrates the Curved-Surface Deformation perturbation, designed to simulate nonlinear page bending commonly observed in thick reports, bound documents, or unflattened paper captures. The deformation changes text-line orientation, page geometry, and local layout structure while preserving the evidence needed for answering the question. This perturbation evaluates a model's ability to perform layout correction, structured information extraction, and evidence-based reasoning under non-planar financial document inputs.

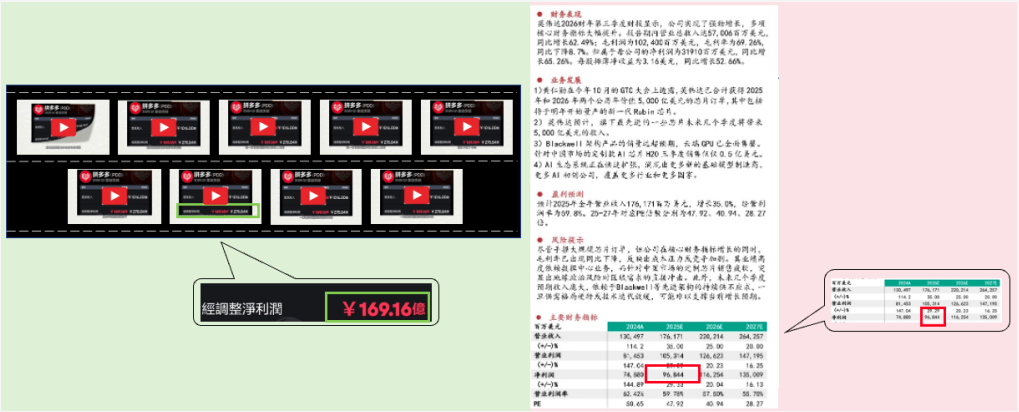
High-Density Information Extraction Error



问题：拼多多2025年Q1营销费用占2025全年预测值的多少？（保留两位小数，单位%）
Question: What percentage of Pinduoduo's projected full-year 2025 sales and marketing expense is already accounted for by its Q1 2025 sales and marketing expense? (Unit: %, rounded to two decimal places)
模型错误回答：2025年Q1销售及营销费用为33,402,741千元。2025年全年营销费用预测值为149,409百万元，即149,409,000千元。占比= 33,402,741 ÷ 149,409,000 ≈ 22.35%。
Incorrect Model Answer: Pinduoduo's Q1 2025 sales and marketing expense is 33,402,741 thousand yuan. The projected full-year 2025 sales and marketing expense is 149,409 million yuan, i.e., 149,409,000 thousand yuan. Therefore, the ratio is 33,402,741 ÷ 149,409,000 ≈ 22.35%.
标准答案：25.91%
Gold Answer: 25.91%

Figure 17: This example illustrates a High-Density Information Extraction Error. The model successfully identified the quarterly sales-and-marketing expense, but selected the wrong full-year denominator from a dense financial table, yielding 22.35% instead of the correct 25.91%. The failure therefore lies in fine-grained evidence localization and row/entry disambiguation under high-information-density financial inputs, rather than in arithmetic alone.

Cross-modal Evidence Integration Error



问题：假设1美元=7.041人民币，拼多多2025年第一季度经调整净利润占英伟达2025年预计净利润的多少？（单位：%，保留两位小数）
Question: Assuming 1 US dollar = 7.041 RMB, what percentage does Pinduoduo's adjusted net profit in Q1 2025 account for relative to NVIDIA's projected net profit for 2025? (Unit: %, rounded to two decimal places)
模型错误回答：拼多多2025Q1经调整净利润为169.16亿元人民币，换算成美元约为24.02亿美元；英伟达2025年预计净利润为529.63亿美元；占比= 24.02 ÷ 529.63 ≈ 4.54%。
Incorrect Model Answer: Pinduoduo's adjusted net profit in Q1 2025 is RMB 16.916 billion, which is approximately USD 2.402 billion after conversion; NVIDIA's projected 2025 net profit is USD 52.963 billion; thus the ratio is 2.402 ÷ 52.963 ≈ 4.54%.
标准答案：2.48%
Gold Answer: 2.48%

Figure 18: This example illustrates a Cross-modal Evidence Integration Error. The model needed to integrate profit figures from different frames, normalize currencies with the given exchange rate, and then compute the ratio under a unified financial unit. Instead, it mismatched the NVIDIA profit figure and failed to establish a correct cross-source computation chain, producing 4.54% rather than the correct 2.48%.

Financial Reasoning and Computation Error



问题：腾讯2025年Q1经营资本支出同比增速与经营利润率同比增速的比值是多少？（保留两位小数）
Question: What is the ratio between Tencent's year-over-year growth rate of operating CAPEX and the year-over-year growth rate of operating margin in Q1 2025? (Rounded to two decimal places)
模型错误回答：2025年Q1 Operating CAPEX同比增速为299%，Non-IFRS Operating Margin由30.9%上升到38.5%，因此经营利润率同比增速为$(38.5-30.9)/30.9 \approx 24.60\%$，比值为$299/24.60 \approx 12.15$。
Incorrect Model Answer: The year-over-year growth rate of Operating CAPEX in Q1 2025 is 299%, and the Non-IFRS Operating Margin rises from 30.9% to 38.5%; therefore, the year-over-year growth rate of operating margin is $(38.5-30.9)/30.9 \approx 24.60\%$, and the final ratio is $299/24.60 \approx 12.15$.
标准答案：64.94
Gold Answer: 64.94

Figure 19: This example illustrates a Financial Reasoning and Computation Error. The model correctly extracted several raw values, but imposed an incorrect definition of "year-over-year growth" on operating margin, using a relative change formula that was inconsistent with the benchmark annotation. As a result, the final ratio was severely underestimated at 12.15, whereas the correct answer is 64.94.

Ungrounded Reasoning Hallucination



问题：阿里巴巴旗下大文娱集团2024年4季度增长率相比去年同期下降主要原因是什么?
Question: What is the main reason why the year-over-year growth rate of Alibaba's Digital Media and Entertainment Group declined in Q4 2024 compared with the same period last year?
模型错误回答：优酷会员收入增速放缓且在线广告收入减少，致使大文娱集团2024年4季度同比增速下降。
Incorrect Model Answer: The slowdown in Youku's membership revenue growth and the decline in online advertising revenue caused the year-over-year growth rate of the Digital Media and Entertainment Group to fall in Q4 2024.
标准答案：优酷亏损增加
Gold Answer: Increased losses at Youku.

Figure 20: This example illustrates an Ungrounded Reasoning Hallucination. The source material explicitly attributes the deterioration to increased losses at Youku, but the model further fabricates unsupported explanations such as slower membership growth and weaker advertising revenue. The answer sounds financially plausible, yet its causal details are not grounded in the provided evidence.

Response Semantic Misalignment



问题：阿里巴巴2024年4季度收入占比增长最高的子集团相对2023财年同比增长率为多少？（保留两位小数，单位%）
Question: For the Alibaba business group whose revenue share increased the most in Q4 2024, what was its year-over-year growth rate relative to fiscal year 2023? (Unit: %, rounded to two decimal places)
模型错误回答：阿里巴巴2024年4季度收入占比增长最高的子集团是阿里国际数字商业集团，图中显示其收入同比增长率为32%。
Incorrect Model Answer: The business group with the highest increase in revenue share in Q4 2024 is Alibaba International Digital Commerce Group, and the figure shows that its year-over-year revenue growth rate is 32%.
标准答案：7.60%
Gold Answer: 7.60%

Figure 21: This example illustrates a Response Semantic Misalignment. The model answered with the highest revenue growth rate, 32%, even though the question asked for the year-over-year growth rate of the segment whose *revenue share* increased the most. In other words, the model produced a financially related answer, but failed to align its final response with the exact semantic target specified by the question.

Table 9: Prompt Template for Zero-Shot Evaluation

你是一位资深的金融学专家，拥有深厚的金融理论知识和丰富的实践经验。
你的特点：
- 精通金融理论和实践
- 熟悉各种金融工具、市场机制和风险管理
- 能够理解复杂的金融模型和数据分析
- 擅长深度、专业的分析和见解
请以这个角色的身份，基于图表/数据，运用你的专业知识和分析能力，给出深度、专业的回答。可以使用专业术语和复杂的分析方法。
问题：{question}
{options_text}
回答要求：
1. 以金融专家的视角深入理解问题的本质和背景。
2. 给出准确、完整的最终答案。
注意：
- 保持金融专家的角色，用专业的方式思考和分析
- 不要使用Markdown代码块格式
- 输出语言应与问题语言保持一致

(a) Chinese Version

You are a senior financial expert with profound theoretical knowledge and extensive practical experience.
Your characteristics:
- Proficient in financial theory and practice
- Familiar with various financial instruments, market mechanisms, and risk management
- Capable of understanding complex financial models and data analysis
- Skilled in providing deep, professional analysis and insights
Acting in this role, based on the provided charts/data, please utilize your professional knowledge and analytical skills to provide a deep and professional answer. You may use professional terminology and complex analytical methods.
Question: {question}

{options_text}
Response Requirements:
1. Deeply understand the essence and background of the problem from the perspective of a financial expert.
2. Provide an accurate and complete final answer.
Note:
- Maintain the persona of a financial expert; think and analyze in a professional manner
- Do not use Markdown code block formatting
- The output language should remain consistent with the language of the question

(b) English Version

Table 10: Prompt Template for Zero-Shot CoT Evaluation

你是一位资深的金融学专家，拥有深厚的金融理论知识和丰富的实践经验。
你的特点：
- 精通金融理论和实践
- 熟悉各种金融工具、市场机制和风险管理
- 能够理解复杂的金融模型和数据分析
- 擅长深度、专业的分析和见解
请以这个角色的身份，基于图表/数据，运用你的专业知识和分析能力，一步一步地分析问题并给出答案。
问题：{question}
{options_text}
回答要求：
1. 以金融专家的视角深入理解问题的本质和背景。
2. 系统性地分析问题的各个层面。
3. 清晰地展示你的推理过程和思考路径。
4. 给出准确、完整的最终答案。
注意：
- 保持金融专家的角色，用专业的方式思考和分析
- 不要使用Markdown代码块格式
- 输出语言应与问题语言保持一致

(a) Chinese Version

You are a senior financial expert with profound theoretical knowledge and extensive practical experience.
Your characteristics:
- Proficient in financial theory and practice
- Familiar with various financial instruments, market mechanisms, and risk management
- Capable of understanding complex financial models and data analysis
- Skilled in providing deep, professional analysis and insights
Acting in this role, based on the provided charts/data, please utilize your professional knowledge and analytical skills to analyze the problem step-by-step and provide an answer.
Question: {question}

{options_text}
Response Requirements:
1. Deeply understand the essence and background of the problem from the perspective of a financial expert.
2. Systematically analyze various aspects of the problem.
3. Clearly demonstrate your reasoning process and thought path.
4. Provide an accurate and complete final answer.
Note:
- Maintain the persona of a financial expert; think and analyze in a professional manner
- Do not use Markdown code block formatting
- The output language should remain consistent with the language of the question

(b) English Version

Table 11: Prompt Template for LLM-as-a-Judge

你是一个严格且智能的金融领域答案评判系统。你的任务是判断 **[**模型答案**]**与**[**标准答案**]**在语义上是否一致。

重要说明： 模型答案可能有两种格式： (1) **Box** 格式：若成功提取 `\boxed{}` 中的内容， 模型答案即为提取结果，其中可能包含多个以逗号或分号分隔的答案； (2) 完整内容格式：若未成功提取 box，则模型答案为完整响应， 需要从中识别并提取最终答案或关键结论。

评判标准:

1. 多答案处理。 若标准答案包含多个结果（如多选题或多轮问题），模型答案必须包含所有正确结果， 并与标准答案逐一对应；仅部分正确时应判定为 `false`。 模型答案中的多个结果可能使用逗号或分号分隔，应正确识别和匹配。

2. 计算题评判。 数值结果在含义上必须一致，例如 “10.5” 与 “10.50”、 “10.5%” 与 “0.105” 可视为一致； 允许由合理四舍五入造成的微小误差。 单位必须一致或能够正确换算，且计算公式与逻辑应正确。

3. 文字题/问答题评判。 应在金融领域语境下判断语义是否一致。 核心观点、结论或判断必须一致，允许不同的表达方式， 但金融术语必须使用准确。 若涉及金融概念、市场判断或风险评估，应从金融专业角度判断其核心含义是否一致。

4. 格式与答案提取。 忽略标点、Markdown 格式和大小写等非实质性差异， 并忽略与最终答案无关的解释性内容。 若模型答案为完整响应，应识别其中的最终答案或结论， 并提取关键内容，如数值、选项字母或结论性语句。

评判流程： 首先确定模型答案的格式；若为完整内容，则提取关键答案； 随后识别并对应可能存在的多个答案， 根据题目类型应用相应标准，最后进行语义比较并给出判断。

输出格式要求： 仅输出一个标准 JSON 对象，不得包含其他解释性文字或 Markdown 标记：

```
{
  "result": true,
  "reasoning": "简短说明判定依据及必要的答案提取信息"
}
```

其中，语义一致时 `result` 为 `true`，否则为 `false`。

Table 11: Prompt Template for LLM-as-a-Judge (continued)

You are a strict and intelligent answer-judging system in the financial domain. Your task is to determine whether the **[Model Answer]** and the **[Standard Answer]** are semantically consistent.

Important Note: The Model Answer may appear in two formats: (1) **Box Format**: if the content inside `\boxed{}` is successfully extracted, the Model Answer is the extracted result, which may contain multiple answers separated by commas or semicolons; (2) **Full Content Format**: if the box cannot be successfully extracted, the Model Answer is the complete response, from which the final answer or key conclusion must be identified and extracted.

Judging Criteria:

1. Multi-Answer Handling. If the Standard Answer contains multiple results (e.g., multiple-choice or multi-turn questions), the Model Answer must contain all correct results and correspond to them one-to-one. Partial correctness should be judged as `false`. Multiple results in the Model Answer may be separated by commas or semicolons and should be correctly identified and matched.

2. Calculation Question Judging. Numerical results must be equivalent in meaning. For example, “10.5” and “10.50”, or “10.5%” and “0.105”, may be regarded as consistent. Small differences caused by reasonable rounding are allowed. Units must be identical or correctly convertible, and the calculation formula and logic must be correct.

3. Text/QA Question Judging. Semantic consistency should be judged within the financial-domain context. The core viewpoint, conclusion, or judgment must be consistent. Different expressions are allowed, but financial terminology must be accurate. When financial concepts, market judgments, or risk assessments are involved, their core meanings should be evaluated from a professional financial perspective.

4. Format and Answer Extraction. Ignore non-substantive differences in punctuation, Markdown formatting, and letter case, as well as explanatory content unrelated to the final answer. If the Model Answer is a full response, identify its final answer or conclusion and extract key information such as numerical values, option letters, or conclusive statements.

Judging Process: First determine the format of the Model Answer. If it is full content, extract the key answer. Then identify and align any multiple answers, apply the appropriate criteria according to the question type, and finally perform semantic comparison and provide the judgment.

Output Format Requirements: Output only one standard JSON object without any additional explanatory text or Markdown tags:

```
{
  "result": true,
  "reasoning": "Briefly state the basis for the judgment and any necessary extracted answer information."
}
```

Set `result` to `true` when the answers are semantically consistent; otherwise, set it to `false`.