VisFinEval: A Scenario-Driven Chinese Multimodal Benchmark for Holistic Financial Understanding

Zhaowei Liu[1†], Xin Guo[1†], Haotian Xia[4†], Lingfeng Zeng[1], Fangqi Lou[1], Jinyi Niu[5], Mengping Li[1], Qi Qi[1], Jiahuan Li[7], Wei Zhang[1], Yinglong Wang[6], Weige Cai[1], Weining Shen[4*], Liwen Zhang[1,2,3*]

[1]School of Statistics and Data Science, Shanghai University of Finance and Economics, [2]Shanghai Financial Intelligent Engineering Technology Research Center, Shanghai University of Finance and Economics, [3]Qinghai Provincial Key Laboratory of Big Data in Finance and Artificial Intelligence Application Technology, [4]University of California, Irvine, [5]Fudan University, [6]Johns Hopkins University, [7]Shanghai University

{zhang.liwen}@shufe.edu.cn, {weinings}@uci.edu

Abstract

Multimodal large language models (MLLMs) hold great promise for automating complex financial analysis. To comprehensively evaluate their capabilities, we introduce **VisFinEval**, the first large-scale Chinese benchmark that spans the full front-middle-back office lifecycle of financial tasks. VisFinEval comprises 15,848 annotated question–answer pairs drawn from eight common financial image modalities (e.g., K-line charts, financial statements, official seals), organized into three hierarchical scenario depths: Financial Knowledge & Data Analysis, Financial Analysis & Decision Support, and Financial Risk Control & Asset Optimization. We evaluate 21 state-of-the-art MLLMs in a zero-shot setting. The top model, Qwen-VL-max, achieves an overall accuracy of 76.3%, outperforming non-expert humans but trailing financial experts by over 14 percentage points. Our error analysis uncovers six recurring failure modes—including cross-modal misalignment, hallucinations, and lapses in business-process reasoning—that highlight critical avenues for future research. VisFinEval aims to accelerate the development of robust, domain-tailored MLLMs capable of seamlessly integrating textual and visual financial information. The data and the code are available at https://github.com/SUFE-AIFLM-Lab/VisFinEval.

1 Introduction

The advent of multimodal large language models (MLLMs) has dramatically broadened the scope of AI applications beyond pure text understanding to encompass tasks that require joint reasoning over images and text, including web navigation, sports analytics, and visual quality assessment (Deng et al., 2023; Koh et al., 2024; Xia et al., 2024; Ku et al., 2023). In the financial domain, practitioners routinely encounter richly formatted visual data—charts, tables, official seals—and yet existing benchmarks predominantly target textual comprehension, leaving a critical gap in the evaluation of MLLMs' ability to integrate and reason over financial visuals. Text–only financial benchmarks such as FinEval (Guo et al., 2024) and CFBenchmark (Zhu et al., 2024) capture important language understanding skills but ignore chart- and document-based information that drives real-world decision making.

Recent works, FinVQA (Bhatia et al., 2024), FIN-FACT (Zhang et al., 2024), MMMU (Wang et al., 2023b), have begun to address multimodal finance, yet they suffer from limited scale, shallow question designs, or narrow coverage of business workflows. In practice, financial analysts progress through front-office data ingestion, mid-office analysis and decision support, and back-office policy and risk control. No existing benchmark systematically evaluates MLLMs across this full process, with tasks that range from basic chart reading to multi-step numerical calculations and counterfactual inferences under realistic perturbations.

To bridge these gaps, we present **VisFinEval**, the first large-scale Chinese benchmark for multimodal financial evaluation that mirrors end-to-end business scenarios. VisFinEval comprises 15,848 rigorously annotated QA pairs drawn from eight common financial image types (e.g., relationship graphs, K-line charts, official seals) and organized into three cascading scenario depths:

- **Financial Knowledge and Data Analysis (Front-Office)** tests foundational chart interpretation and basic numerical reasoning.
- **Financial Analysis and Decision Support (Mid-Office)** challenges models with multi-

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

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 24088–24146
November 4-9, 2025

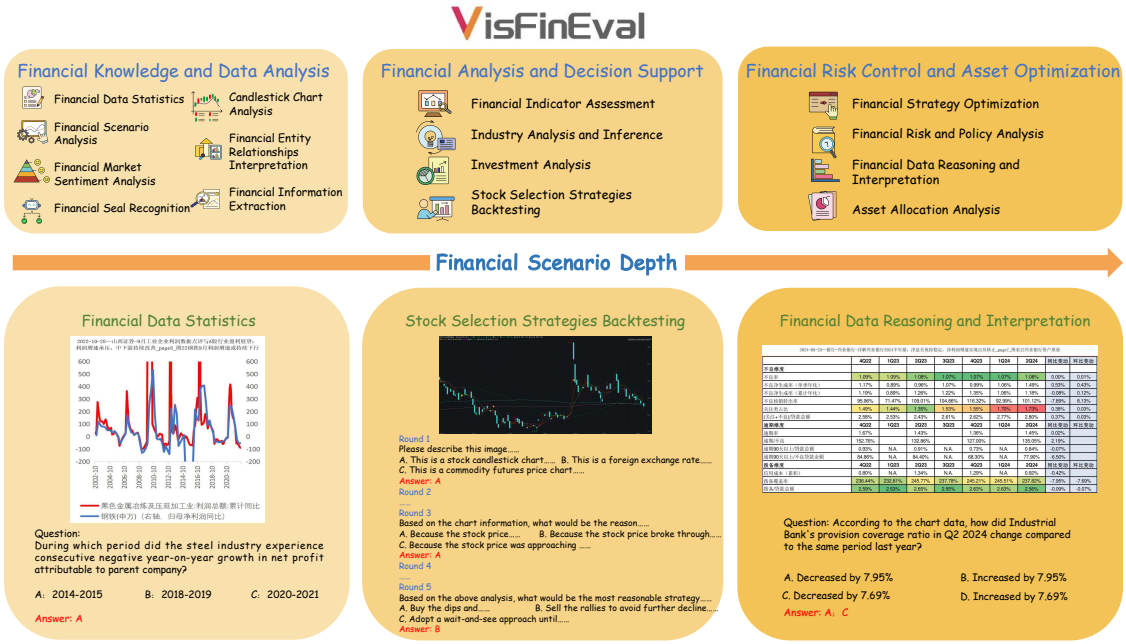


Figure 1: VisFinEval provides a multimodal evaluation framework for full-process financial operations. Starting from the perspective of business depth, it designs 3 major scenarios: Financial Knowledge and Data Analysis, Financial Analysis and Decision Support, and Financial Risk Control and Asset Optimization. Each major scenario corresponds to 7, 4, and 4 sub-scenarios, respectively, accurately reflecting the diverse business scenarios within the financial industry. Furthermore, it has constructed 15,848 multiple-choice and question-answering items based on 8 common types of images in the financial domain.The upper part of the image illustrates the overall structure of VisFinEval, where the business depth increases with the change in color. Concurrently, the demands on the model's understanding and analytical capabilities regarding financial business operations progressively increase. The lower part shows a specific example corresponding to the sub-scenarios.

image reasoning, metric computation, and investment backtesting.

- **Financial Risk Control and Asset Optimization (Back-Office)** probes advanced capabilities in strategy optimization, policy impact analysis, and complex data extrapolation, including professional-exam-level questions.

By simulating real-world document perturbations and multi-turn dialogues, VisFinEval captures the full complexity of financial workflows.

Our contributions are threefold:

- **Comprehensive Multimodal Benchmark.** We construct VisFinEval with 15,848 QA pairs spanning eight types of financial images and three hierarchical scenario depths, filling a gap in financial MLLM evaluation.
- **Full-Process Business Workflow.** We align tasks with front-, mid-, and back-office functions, such as data perception, analytical decision support, and strategic optimization—thereby providing a practical, process-aware assessment framework.
- **Extensive Zero-Shot Evaluation.** We benchmark 21 state-of-the-art MLLMs in a zero-shot setting, analyze failure modes across six error categories, and compare model performance against non-expert and expert human baselines to highlight remaining challenges.

This paper is organized as follows. Section 2 provides a review of related work in financial MLLMs and multimodal benchmarks. Section 3 details the construction of VisFinEval, including data collection, question design, and quality control procedures. Section 4 and Section 5 presents our experimental setup and results across different difficulty levels, followed by error analysis. Finally, Section 6 concludes the paper and discusses potential future directions in multimodal financial intelligence.

2 Related Work

Financial Scenario Analysis Under the accelerating digital transformation in the financial sector, the groundbreaking advancements in Large Language Models (LLMs) have injected new momentum into the integration of artificial intelligence

Table 1: Comparison of various benchmarks across multiple dimensions. Abbreviations in the header are: QT(Question Type), MC (Multiple-Choice questions), OE (Open-Ended questions), T/F (True/False questions), MLD(Multi-level Difficulty), SD(Scenario Depth), RES (Realistic Environment Simulation), OS (Official Seal), FRG (Financial Relationship Graph), NoFFT (Number of Financial Figure Type), NoFS (Number of Financial Scenarios), NoQ (Number of Questions), and NoM (Number of Models). To better simulate real-world environments during the question-answering process, we introduced RES, which refers to simulating unexpected situations that may occur in real-world financial business scenarios.

Benchmarks	QT	MLD	SD	RES	OS	FRG	NoFFT	NoFS	NoQ	NoM
Text										
FinDABench	OE	✓	-	-	-	-	-	5	2400	40
SuperCLUE-Fin	OE	✗	-	-	-	-	-	6	1000	11
CFBenchmark	OE	✗	-	-	-	-	-	8	3917	22
FinEval	MC+OE	✗	-	-	-	-	-	9	8351	19
MultiModal										
SEEDBENCH	MC	✗	✗	✗	✗	-	-	-	19000	18
MMMU	MC	✓	✗	✗	✗	✗	-	-	11500	30
FinVQA	OE	✗	✗	✗	✗	✗	2	2	1025	9
FIN-FACT	T/F	✗	✗	✗	✗	✗	2	5	3369	4
FAMMA	MC+OE	✓	✗	✗	✗	✗	3	8	1758	4
MME-Finance	OE	✓	✗	✗	✗	✗	6	11	2274	19
VisFinEval (Ours)	MC+T/F+OE	✓	✓	✓	✓	✓	8	15	15848	21

(AI) and finance. Early research primarily focused on unimodal technical applications, including text understanding (Masry and Hajian, 2024; Wilson et al., 2024), sentiment analysis (Delgadillo et al., 2024; Zhang et al., 2023), financial time-series forecasting (Li et al., 2024a,b; Wang et al., 2024b; Mai, 2024), and decision support (Yu et al., 2024b; Wang et al., 2023a; Yu et al., 2024a; Yang et al., 2023; Li et al., 2023b). However, these studies largely overlooked the critical value of chart-based data in financial contexts. Financial charts often encapsulate pivotal decision-making insights that penetrate beyond superficial data representations, only through accurate interpretation of such data can the core logic of financial decision-making be unveiled. This limitation was not alleviated until the emergence of multimodal large language models (MLLMs) (Bhatia et al., 2024; Zhang et al., 2024; Wang et al., 2023b) , which reconstructed MLLMs' comprehensive cognitive framework for the financial domain.

While existing benchmarks (Nie et al., 2024; Zhu et al., 2024; Koncel-Kedziorski et al., 2023; Zhu et al., 2021; Zhao et al., 2024; Wang et al., 2024a; Reddy et al., 2024; Chen et al., 2021, 2022, 2024a, 2025) effectively evaluate models' financial text comprehension capabilities, they remain inadequate for assessing models' understanding of complex financial operations and multimodal chart data. Consequently, there is an urgent need to extend existing benchmarks to comprehensively evaluate models' multimodal financial data comprehension and reasoning abilities, thereby more authentically reflecting their real-world applicability in financial scenarios.

Multimodal Financial Benchmark Up to today, the availability of dedicated benchmarks for multimodal financial scenarios remains limited. General multimodal benchmarks (Li et al., 2023a; Liu et al., 2024b; Guthaus et al., 2001; Yue et al., 2024; Mathew et al., 2021) predominantly fail to adequately encompass domain-specific financial tasks, making it challenging to accurately assess models' professional capabilities in financial contexts. Furthermore, existing studies on financial multimodal benchmarks are confined to knowledge-level validation and lack systematic evaluation of models' operational depth and workflow integration in financial scenarios, thereby failing to holistically reflect their practical efficacy in real-world financial applications.

FAMMA (Xue et al., 2024) provides financial knowledge-related question-answering tasks, but its data primarily originates from university textbooks and examination questions, limiting its evaluation scope to knowledge verification rather than complex financial operational scenarios. FinTMMBench (Zhu et al., 2025a) incorporates images of limited diversity, and its question design lacks explicit mapping to concrete financial business contexts. While MME-Finance (Gan et al., 2024) addresses operational scenarios, its narrow business

scope, limited question volume, and absence of task difficulty stratification aligned with real-world complexities restrict its generalizability and result in insufficient task depth, creating a disconnect from practical realities. Additionally, current research predominantly evaluates MLLMs' performance in controlled environments while neglecting the inherent complexity of real-world financial scenarios, thereby impeding accurate assessment of large models' true capabilities in financial applications.

To address these gaps in prior studies, we introduce VisFinEval, the first large - scale benchmark specifically designed for multimodal large language models in finance. This benchmark integrates diverse real-world financial scenarios and potential edge cases, employing hierarchical evaluation criteria to comprehensively cover tasks ranging from foundational knowledge to complex operational workflows. VisFinEval effectively bridges the capability gaps in existing evaluation frameworks, enabling rigorous and realistic assessment of models' financial multimodal intelligence.

3 VisFinEval Benchmark

3.1 Overview

We propose VisFinEval, a multimodal benchmark designed for the Chinese financial domain, which aims to evaluate the capabilities of MLLMs in processing and reasoning across the entire financial business workflow. As the first large-scale evaluation framework that deeply integrates multimodal tasks with end-to-end financial business scenarios, VisFinEval is constructed based on the actual operational flow of the financial industry. It establishes evaluation dimensions aligned with real-world needs, spanning from foundational front-office financial data perception, to mid-office analytical decision-making, and ultimately to high-level back-office strategic planning. This structure reflects not only the high-frequency demands observed in practical financial contexts, but also follows a hierarchical and process-aware financial decision-making process. This enables the construction of a comprehensive evaluation framework that covers the entire financial business process. Therefore, VisFinEval provides a multimodal financial evaluation benchmark that is both professionally grounded and practically valuable. The overall framework is illustrated in Figure 1.

VisFinEval is designed based on the front–mid–back office collaborative architecture commonly adopted in real-world financial systems, and establishes a three-tier evaluation framework that spans the entire financial business lifecycle. By integrating multimodal data with scenario-driven financial tasks, this benchmark systematically assesses the domain-specific capabilities of MLLMs within the vertical of financial scenarios.

This evaluation system is the first to achieve full-chain coverage of front-, mid-, and back-office financial functions. The front-office layer establishes a foundation of multimodal data perception, the mid-office layer constructs the core for analytical decision-making, and the back-office layer forms a closed loop for strategic optimization. Through modular decomposition and reorganization of financial workflows, VisFinEval ensures both the professional depth of evaluation tasks within each layer and the assessment of holistic model performance in cross-functional collaboration scenarios.

The evaluation tasks adopt a wide range of objective formats, including single-choice, multiple-choice, true/false, and numerical reasoning questions, while also introducing dimensions such as multi-turn dialogue, counterfactual inference, multimodal consistency evaluation, and complex perturbation. By combining different question types and aligning them with specific sub-scenarios, the evaluation ensures a comprehensive assessment of MLLMs' capabilities in real-world financial tasks. Details on the task scenarios and dataset distributions are shown in Table 4, and representative examples of question types are provided in the Appendix A.

3.2 Data Generation and Quality Control

During the data construction process of VisFinEval, most visual inputs are collected from PDF documents within the financial domain. These images are sourced from financial research reports, annual reports, and professional examinations such as the Chinese CPA and actuary exams. The dataset includes a diverse set of image types, including financial relationship graphs, line charts, histograms, candlestick (K-line) charts, pie charts, official seals, financial statements, and supporting data tables. Specifically, line charts, histograms, pie charts, and relationship graphs are primarily extracted from financial research reports; financial statements and supporting tables are collected from annual reports and exam questions; seal images are sourced from open-source datasets such as (Gmgge, 2024); and

K-line charts are obtained from publicly accessible financial websites. All image materials are verified to be free from copyright restrictions.

The core of the data generation process lies in constructing scenario-specific prompts aligned with financial experts to ensure the domain relevance and consistency of question-answer (QA) pairs. These prompts guided the use of the Qwen-VL-Plus-latest (Yang et al., 2024) to generate QA pairs based on the input images. The reliability of vision-language models in such generation tasks has been validated by prior work such as InstructBLIP (Dai et al., 2023). Furthermore, we used Qwen-max to classify the generated QA pairs into appropriate financial business scenarios spanning the full process. Detailed prompts used for the data generation and classification are provided in the Appendix C.

The QA data underwent a three-stage rigorous review process, including MLLM-based automated filtering based on multi-dimensional evaluation criteria, fine-grained annotation by trained and qualified undergraduate students, and cyclical validation by financial experts with ten years of work experience. This multi-layered processing pipeline ensures that we obtain high-quality QA pairs that meet standards such as accuracy, domain relevance, and consistency. Detailed review procedures and example prompts are provided in the Appendix A.2.

Through this rigorous data generation and validation pipeline, VisFinEval offers a high-quality multimodal QA dataset tailored for evaluating the domain-specific capabilities of large multimodal models in the financial sector.

3.3 VisFinEval Question Architecture

VisFinEval, with financial business scenarios as its core starting point, has constructed a three-level hierarchical evaluation framework. This framework aims to systematically assess the comprehensive capability performance of Multimodal Large Language Models (MLLMs) in financial multimodal business, encompassing 15,848 high-quality QA pairs. It is further divided into the following three major real-world financial business scenarios based on scenario depth:

Front-Office: Financial Knowledge and Data Analysis Front-office operations in the financial domain are more oriented towards customer needs, focusing on the basic understanding of financial knowledge and data analysis-related capabilities. In this scenario, we have designed tasks covering financial cognition and data processing, primarily evaluating MLLMs' ability to understand customer needs and intentions in financial business and to process and analyze financial business data. This includes seven sub-scenarios: Financial Data Statistics, Candlestick Chart Analysis, Financial Indicator Assessment, Financial Entity Relationships Interpretation, Stock Selection Strategies Backtesting, Financial Information Extraction, and Financial Seal Recognition, corresponding to the real-world demands of financial front-office business activities.

Mid-Office: Financial Analysis and Decision Support The mid-office operations in the financial domain emphasize large-scale business facing the company or enterprise, requiring business personnel to have a deeper understanding and analysis of actual business, and to make clear and effective decisions on specific business issues under the influence of various factors in the real environment. Therefore, the design of this scenario aims to test the MLLM's comprehensive analysis and reasoning decision-making capabilities in a financial context. Tasks involve interpreting complex financial data structures and business logic, such as equity research and financial metric evaluation. Consequently, we have designed four core business sub-scenarios to evaluate the MLLM's information integration and systematic decision-making abilities: Financial Scenario Analysis, Industry Analysis and Inference, Investment Analysis, and Financial Market Sentiment Analysis, focusing on the analytical and decision-support functions typically undertaken by the mid-office.

Back-Office: Financial Risk Control and Asset Optimization The back-office operations in the financial domain lean more towards strategic research and risk control. Business personnel need to possess strong domain expertise, mathematical calculation, and reasoning abilities to support front-office and mid-office operations through various internal decision-making processes, thereby ensuring the overall efficiency and effectiveness of financial business. To maintain the business authenticity of the evaluation, we have designed a series of highly complex financial tasks, including statistical inference, audit analysis, and expert-level reasoning, to assess whether MLLMs meet the requirements of actual business. Some of these tasks are adapted from challenging professional qualification exams, such as the Chinese CPA exam. Consequently, we

have identified four sub-scenarios: Financial Strategy Optimization, Financial Risk and Policy Analysis, Financial Data Reasoning and Interpretation, and Asset Allocation Analysis, designed to simulate the strategic decision-making and optimization processes of the financial back office.

Meticulously designed based on the entire financial business process system, VisFinEval boasts advantages in terms of its systematic nature, practicality, and comprehensiveness. It can provide a professional and challenging benchmark for evaluating the real-world applicability of MLLMs in financial business.

4 Experiments Settings

4.1 Models

We tested 21 multimodal large language models , with close-source models accessed through their respective APIs and open-source models deployed locally. All inference tasks were run on NVIDIA A800 GPUs. For more details on the models please refer to Appendix B.

Closed-source models: For close-source models, we evaluated 9 models, including Qwen-VL-max-lastest (Bai et al., 2023), Qwen-VL-max (Bai et al., 2023), Doubao-1.5-vision-pro-32k (Doubao, 2025), Step-1o-vision-32k (StepStar, 2025), Gemini-2.5-pro-exp-03-25 (Google, 2025), GPT-4o-2024-11-20 (OpenAI, 2024), Moonshot-V1-32k-vision-preview (MoonshotAI, 2024), Claude-3-7-Sonnet-20250219 (Anthropic, 2025) and GLM-4v-Plus-20250111 (ZhipuAI, 2025).

Open-source models: For open-source models, we evaluated 12 models from several mainstream MLLMs, including Qwen2.5-VL-3B, Qwen2.5-VL-7B and Qwen2.5-VL-72B from the Qwen series (Yang et al., 2024); InternVL3-8B (Zhu et al., 2025b), InternVL2.5-78B (Chen et al., 2024b) and InternVL3-78B (Zhu et al., 2025b) from the InternVL series (Chen et al., 2024b); LLaVA-v1.6-Mistral-7B (Liu et al., 2023), LLaVA-NeXT-13B and LLaVA-NeXT-34B from the LLaVA series (Liu et al., 2024a); as well as MiniCPM-V-2.6 (Yao et al., 2024), Molmo-7B-D-0924 (Deitke et al., 2024), and Llama-3.2-11B-Vision-Instruct (Touvron et al., 2023).

4.2 Evaluation Methods

Despite our efforts to optimize prompts to improve model output, some models exhibit poor instruction following capabilities, making their output unsuitable for evaluation via rule-based extraction. To address this challenge, MMBench (Xu et al., 2023) proposed leveraging LLMs as selection extractors, which significantly improved evaluation accuracy. Following a similar approach, we designed specific prompts and employed Qwen-max-latest as the judge model to evaluate the outputs of various models. To validate the judge model's evaluations, we conducted a manual review of all the results it provided for each model and task. The review showed that the accuracy of the judge model's judgments exceeded 98%.

5 Results

5.1 Main Results

We evaluated 21 mainstream MLLMs, as shown in Table 2. Due to a few limitations such as context length or multi-image support, certain questions were excluded from evaluation for some models; their results are provided separately for reference.

Among all the results, Qwen-VL-max achieved the best overall performance, with an average accuarcy of 76.3%. It ranked highest among all evaluated models in 10 out of 15 sub-scenarios, strongly indicating Qwen-VL-max's stable and powerful capabilities across diverse and in-depth multimodal financial scenarios. Closely following was Qwen-VL-max-latest, with only a 2.5% difference, also demonstrating outstanding performance in FMSA and IA. Together, these results highlight the Qwen series' excellence in the financial multimodal domain. Ranked third to sixth were InternVL3-78B, Doubao-1.5-vision-pro-32k, InternVL2.5-78B, and Qwen2.5-VL-72B, with relatively close scores. InternVL3-78B tied for the top score in FSO with Qwen-VL-max, reflecting its ability to optimize strategies in response to various challenges in financial business. Doubao-1.5-vision-pro-32k performed well in FIE , demonstrating strong visual information extraction capabilities in multimodal settings, and its high score in AAA further underscores its competence in asset allocation and financial analysis tasks. It is worth noting that Step-1o-vision-32k achieved an exceptionally high accuracy of 98.0% in FSR significantly outperforming all other models. This suggests strong capabilities in this sub-scenario. In stark contrast, Claude-3-7-Sonnet-20250219 scored only 34.7% in the same scenario, often failing to recognize seals correctly and sometimes even producing incorrect responses despite correct recognition. We

Table 2: Main Results. The higher the value in the table, the higher the accuracy of the surface model. The Financial Analysis and Decision Support assesses models with Financial Data Statistics (FDS), Candlestick Chart Analysis (CCA), Financial Indicator Assessment (FIA), Financial Entity Relationships Interpretation (FERI), Stock Selection Strategies Backtesting (SSSB), Financial Information Extraction (FIE), and Financial Seal Recognition (FSR). The Financial Analysis and Decision Support tests with Financial Scenario Analysis (FSA), Industry Analysis and Inference (IAI), Investment Analysis (IA), and Financial Market Sentiment Analysis (FMSA). The Financial Risk Control and Asset Optimization evaluates Financial Strategy Optimization (FSO), Financial Risk and Policy Analysis (FRPA), Financial Data Reasoning and Interpretation (FDRI), and Asset Allocation Analysis (AAA), concluding with the calculation of the Weighted Average (WA) score for each model. The table also indicates operational constraints encountered by certain models in multi-image tasks, such as Multi-image Limit and Context Window Limit.

Model	Size	Limit	Financial Knowledge and Data Analysis							Financial Analysis and Decision Support				Financial Risk Control and Asset Optimization				WA
			FDS	CCA	FIA	FERI	SSSB	FIE	FSR	FSA	IAI	IA	FMSA	FSO	FRPA	FDRI	AAA	
Qwen-VL-max	Unknown	/	**78.8**	**90.5**	**87.4**	**89.2**	**86.2**	90.6	77.9	**65.3**	**83.1**	82.3	76.8	**49.1**	**58.2**	**58.2**	71.0	**76.3**
Qwen-VL-max-latest	Unknown	/	76.0	84.5	86.1	87.1	79.3	88.6	84.4	59.6	82.6	**82.8**	**79.3**	44.0	52.2	48.9	71.8	73.8
InternVL3-78B	78B	/	71.2	83.5	71.4	86.7	79.5	87.8	87.4	64.3	82.1	80.4	78.7	**49.1**	52.8	46.6	66.5	72.5
Doubao-1.5-vision-pro-32k	Unknown	/	75.6	79.0	84.2	85.5	76.8	**91.7**	74.4	56.7	80.2	79.8	77.3	30.0	54.5	54.5	**75.6**	71.7
InternVL2.5-78B	78B	/	73.3	77.9	72.3	84.2	84.0	88.4	82.9	63.3	81.5	80.1	75.2	41.0	53.1	47.6	68.4	71.5
Qwen2.5-VL-72B	72B	/	75.9	77.0	72.8	85.4	81.5	88.3	80.4	57.4	82.4	80.3	74.5	41.4	53.4	42.6	71.9	71.0
GPT-4o-2024-11-20	Unknown	/	72.0	76.8	74.9	81.7	71.8	83.8	83.9	61.9	77.9	78.5	73.2	41.0	40.5	41.6	67.9	68.5
Step-1o-vision-32k	Unknown	/	48.9	78.4	80.2	84.1	75.3	88.2	**98.0**	40.3	78.8	78.6	76.1	39.2	45.2	49.0	65.8	68.4
Moonshot-V1-32k-vision-preview	Unknown	/	56.2	82.8	73.4	80.5	73.9	87.6	68.3	61.9	77.7	77.0	72.3	39.2	55.8	53.6	64.0	68.3
Qwen2.5-VL-7B	7B	/	71.4	75.9	69.2	80.9	74.0	85.5	69.9	53.4	79.7	76.5	70.7	37.2	37.6	35.4	63.2	65.4
InternVL3-8B	8B	/	68.2	78.0	62.8	87.0	74.1	84.0	77.4	56.5	76.1	76.8	71.7	29.7	46.2	36.8	55.3	65.4
Gemini-2.5-pro-exp-03-25	Unknown	/	73.6	76.7	72.6	81.0	73.0	89.4	87.4	53.2	72.4	70.8	75.5	28.4	38.0	41.5	37.7	64.7
Claude-3-7-Sonnet-20250219	Unknown	/	70.5	73.4	80.3	71.1	77.5	83.2	34.7	48.0	76.1	75.5	64.0	26.8	50.3	48.6	64.4	62.9
Qwen2.5-VL-3B	3B	/	69.5	81.1	65.9	76.6	73.6	83.4	72.4	50.0	75.4	74.7	66.6	22.9	34.8	35.9	53.8	62.4
MiniCPM-V-2.6	8B	/	61.3	83.5	56.9	76.7	75.2	73.4	80.9	48.3	69.7	70.7	69.1	20.6	35.5	26.8	52.7	60.1
Llama-3.2-11B-Vision-Instruct	11B	/	56.9	40.8	59.3	63.9	62.9	73.1	70.4	45.3	69.7	67.1	63.4	18.0	22.1	19.9	31.1	50.9
Molmo-7B-D-0924	7B	/	60.1	74.8	54.5	62.2	59.1	60.5	42.2	39.7	64.4	62.8	63.4	23.4	31.7	21.9	26.5	49.8
GLM-4v-Plus-20250111	Unknown	Multi-image Limit	73.8	86.6	87.9	87.5	81.2	89.3	72.7	56.5	78.1	74.9	74.6	45.1	54.1	45.3	73.2	72.0
LLaVA-NEXT-34B	34B	Context Window Limit	55.3	79.8	92.3	63.2	87.8	55.0	58.8	54.3	88.2	88.1	66.9	13.1	17.5	12.7	7.7	56.0
LLaVA-v1.6-Mistral-7B	7B	Context Window Limit	54.6	73.4	65.9	62.1	47.4	47.0	62.3	42.3	58.3	56.4	63.7	10.2	16.3	35.9	21.1	47.8
LLaVA-NEXT-13B	13B	Context Window Limit	50.2	64.8	43.9	57.2	62.5	50.2	38.7	34.7	59.2	59.0	52.9	14.7	10.8	15.8	31.1	43.0

attribute this primarily to its poor semantic alignment with Chinese, resulting in hallucinations.

From a pattern perspective, the performance gap between the open-source model InternVL3-78B and the closed-source model Qwen-VL-max is only 3.8%, suggesting that as MLLMs continue to evolve, the performance disparity among top-tier models in financial tasks will gradually diminish. Regarding model size, both the Qwen and InternVL series show a clear trend where larger parameter models exhibit stronger capabilities. A similar phenomenon is also observed in Molmo-7B-D-0924, a topic further explored in detail within FinEval (Guo et al., 2024). As task complexity increases, all models exhibit a noticeable decline in performance. This also demonstrates that VisFinEval effectively tests the boundaries of MLLMs' financial business capabilities, reflecting its authenticity and effectiveness.

5.2 Comparative Analysis

To better compare the capabilities of MLLMs and make a meaningful contribution to model research, we randomly selected 2% of the questions from VisFinEval (approximately 300 questions) to conduct a competition among models, non-experts, and financial experts. Considering differences in domain knowledge and in order to better reflect the current stage of MLLM development, we selected the top two performing models from both open-source and closed-source categories for comparison. On the human side, we included a undergraduate students with no background in finance as representatives of non-experts, while the financial expert role was filled by a PhD candidate majoring in finance. All participants were uninvolved in any data annotation or review processes related to this study, and all responses were collected under closed-book conditions.

Unlike Table 2, here we calculate the average results for each of the three major scenarios as well as the overall average result to compare with human performance. As shown in Table 3, the top-performing MLLMs have already outperformed the non-experts in all three major scenarios as well as in terms of overall average score. However, there remains a performance gap of over 14% between the MLLMs and the financial expert, indicating that further iterations and improvements are still needed. A detailed analysis of the three major scenarios can be found in Appendix B.1.

5.3 Error Analysis

In all the incorrect answers from the evaluated MLLMs, we conducted a stratified sampling of 10% of the questions for error analysis to investigate the issues MLLMs face in financial domain business capabilities. Based on the various types of errors made by MLLMs, we summarized six major problems in the financial domain: Lack of cross-

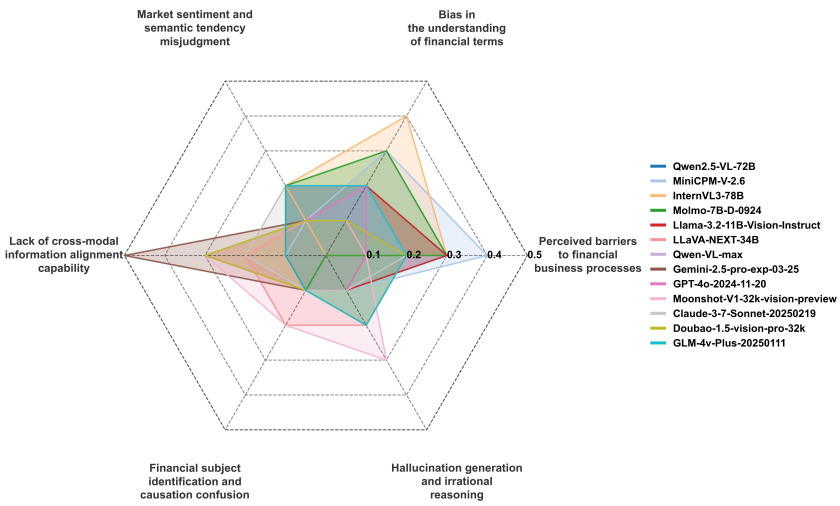


Figure 2: Error types' distribution across different MLLMs in VisFinEval tasks.

Table 3: Performance comparison across non-experts, experts and MLLMs. FKDA represented Financial Knowledge and Data Analysis. FADS refers to Financial Analysis and Decision Support. FRCAO means Financial Risk Control and Asset Optimization.

Source	Category	FKDA	FADS	FRCAO	Average
Human	Non-experts	72.1	57.0	40.1	56.4
	Experts	93.3	88.0	82.8	88.0
Closed-Source	Qwen-VL-max	85.8	76.9	59.1	73.9
	GPT-4o-2024-11-20	77.8	72.9	47.7	66.1
Open-Source	InternVL3-78B	81.1	76.4	53.7	70.4
	Qwen2.5-VL-72B	80.2	73.6	52.3	68.7

modal information alignment capability (Evaluating the model's cross-modal alignment capability is one of the innovations of this paper, and it is also crucial in real-world applications.), Market sentiment and semantic tendency misjudgment, Bias in the understanding of financial terms, perceived barriers to financial business processes, Hallucination generation and irrational reasoning, Financial subject identification and causation confusion. These issues limit the MLLMs' performance in specific financial scenarios.

From Figure 2, most MLLMs exhibit a relatively even distribution of errors, such as Qwen-VL-max and MoonShot-V1-32k-vision-preview, while a few models, such as Gemini-2.5-pro-exp-03-25, InternVL3-78B, and MiniCPM-V-2.6, show a higher concentration of errors in areas such as cross-modal consistency, understanding of financial terms, and financial business process, respectively. For more detailed analysis and related error examples, please refer to Appendix B.2.

6 Conclusion

This paper introduces VisFinEval, a benchmark designed to evaluate MLLMs' based on the full-process business system of the financial domain. t assesses MLLMs' understanding and application abilities in real-world financial tasks through three major business scenarios, which together comprise fifteen sub-scenarios. Through comprehensive analysis of these scenarios, VisFinEval identifies eight commonly used chart types that cover a wide range of financial applications, enabling a performance evaluation grounded in actual business competencies. The results indicate that Qwen-VL-max performs the best overall; however, it still experiences some performance degradation in scenarios with the highest business complexity. Compared to humans, most current MLLMs have already outperformed non-expert individuals without a financial background, but a substantial gap remains when compared to financial experts. In addition, the error analysis highlights six major capability deficiencies that MLLMs exhibit in real-world financial applications. As a benchmark rooted in the full-process business workflows of the financial industry, VisFinEval provides a structured framework for measuring MLLMs' practical capabilities in finance. We hope it will drive progress in MLLM research and contribute to enabling a deeper understanding of real-world financial scenarios.

Limitations

While VisFinEval has made significant progress in evaluating multimodal large language models in

the financial domain, it still has some limitations. Although VisFinEval includes some analysis of dynamic trend changes, it lacks in-depth research on the more dynamic micro and macro financial markets, which are closely related to time. Future work needs to consider designing an evaluation framework that can better assess the performance of MLLMs in more dynamic scenarios. The current work's evaluation is mainly focused on zero-shot performance, and it is necessary to further consider the potential of MLLMs to adapt through few-shot learning. Finally, although VisFinEval includes various financial image types, the distribution of these types and their relative importance in real-world financial analysis could be further refined, as the importance varies across different business scenarios. It is necessary to design more appropriate business scenario weights to evaluate the true performance of MLLMs in the financial domain.

Acknowledgments

This work was supported jointly by the National Social Science Fund of China (Grant No. 22BTJ031, Liwen Zhang) and the Qinghai Provincial Key Laboratory of Big Data in Finance and Artificial Intelligence Application Technology.

References

Anthropic. 2025. Claude-3.7-sonnet. https://www.anthropic.com/.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

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*.

Jialin Chen, Aosong Feng, Ziyu Zhao, Juan Garza, Gaukhar Nurbek, Cheng Qin, Ali Maatouk, Leandros Tassiulas, Yifeng Gao, and Rex Ying. 2025. 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. 2024a. Fintextqa: A dataset for long-form financial question answering. *arXiv preprint arXiv:2405.09980*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, et al. 2021. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering. *arXiv preprint arXiv:2210.03849*.

Wenliang Dai, Junnan Li, Dongxu Li, et al. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*. Accepted to ECCV 2023.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.

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.

Zihao Deng, Yinghao Ma, Yudong Liu, Rongchen Guo, Ge Zhang, Wenhu Chen, Wenhao Huang, and Emmanouil Benetos. 2023. Musilingo: Bridging music and text with pre-trained language models for music captioning and query response. *arXiv preprint arXiv:2309.08730*.

Doubao. 2025. Doubao-1.5-vision-pro-32k. https://console.volcengine.com.

Ziliang Gan, Yu Lu, Dong Zhang, Haohan Li, Che Liu, Jian Liu, Ji Liu, Haipang Wu, Chaoyou Fu, Zenglin Xu, et al. 2024. Mme-finance: A multimodal finance benchmark for expert-level understanding and reasoning. *arXiv preprint arXiv:2411.03314*.

Gmgge. 2024. Trocr-seal-recognition. https://github.com/Gmgge/TrOCR-Seal-Recognition/blob/main/DataSet.md.

Google. 2025. Gemini-2.5-pro-exp-03-25. 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. 2024. Fineval: A chinese financial domain knowledge evaluation

benchmark for large language models. *arXiv preprint arXiv:2308.09975v2*.

Matthew R Guthaus, Jeffrey S Ringenberg, Dan Ernst, Todd M Austin, Trevor Mudge, and Richard B Brown. 2001. Mibench: A free, commercially representative embedded benchmark suite. In *Proceedings of the fourth annual IEEE international workshop on workload characterization. WWC-4 (Cat. No. 01EX538)*, pages 3–14. IEEE.

Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Chong Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Ruslan Salakhutdinov, and Daniel Fried. 2024. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. *arXiv preprint arXiv:2401.13649*.

Rik Koncel-Kedziorski, Michael Krumdick, Viet Lai, Varshini Reddy, Charles Lovering, and Chris Tanner. 2023. Bizbench: A quantitative reasoning benchmark for business and finance. *arXiv preprint arXiv:2311.06602*.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. 2023. Viescore: Towards explainable metrics for conditional image synthesis evaluation. *arXiv preprint arXiv:2312.14867*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Lezhi Li, Ting-Yu Chang, and Hai Wang. 2023b. Multimodal gen-ai for fundamental investment research. *arXiv preprint arXiv:2401.06164*.

Xiang Li, Zhenyu Li, Chen Shi, Yong Xu, Qing Du, Mingkui Tan, Jun Huang, and Wei Lin. 2024a. 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. 2024b. Finreport: Explainable stock earnings forecasting via news factor analyzing model. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 319–327.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

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

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*.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

MoonshotAI. 2024. Kimi. https://www.moonshot.cn/.

Ying Nie, Binwei Yan, Tianyu Guo, Hao Liu, Haoyu Wang, Wei He, Binfan Zheng, Weihao Wang, Qiang Li, Weijian Sun, et al. 2024. Cfinbench: A comprehensive chinese financial benchmark for large language models. *arXiv preprint arXiv:2407.02301*.

OpenAI. 2024. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/.

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*.

StepStar. 2025. Step-1o-vision-32k. https://www.stepfun.com/.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Haochen Wang, Kai Hu, Haoyu Dong, and Liangcai Gao. 2024a. 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. 2024b. Quantagent: Seeking holy grail in trading by self-improving large language model. *arXiv preprint arXiv:2402.03755*.

Saizhuo Wang, Hang Yuan, Leon Zhou, Lionel M Ni, Heung-Yeung Shum, and Jian Guo. 2023a. Alphagpt: Human-ai interactive alpha mining for quantitative investment. *arXiv preprint arXiv:2308.00016*.

Ziao Wang, Yuhang Li, Junda Wu, Jaehyeon Soon, and Xiaofeng Zhang. 2023b. 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.

Haotian Xia, Zhengbang Yang, Junbo Zou, Rhys Tracy, Yuqing Wang, Chi Lu, Christopher Lai, Yanjun He, Xun Shao, Zhuoqing Xie, et al. 2024. Sportu: A comprehensive sports understanding benchmark for multimodal large language models. *arXiv preprint arXiv:2410.08474*.

Cheng Xu, Xiaofeng Hou, Jiacheng Liu, Chao Li, Tianhao Huang, Xiaozhi Zhu, Mo Niu, Lingyu Sun, Peng Tang, Tongqiao Xu, et al. 2023. Mmbench: Benchmarking end-to-end multi-modal dnns and understanding their hardware-software implications. In *2023 IEEE International Symposium on Workload Characterization (IISWC)*, pages 154–166. IEEE.

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, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Yi Yang, Yixuan Tang, and Kar Yan Tam. 2023. Investlm: A large language model for investment using financial domain instruction tuning. *arXiv preprint arXiv:2309.13064*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Yangyang Yu, Haohang Li, Zhi Chen, Yuechen Jiang, Yang Li, Denghui Zhang, Rong Liu, Jordan W Suchow, and Khaldoun Khashanah. 2024a. 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, et al. 2024b. 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, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu. 2023. Instruct-fingpt: Financial sentiment analysis by instruction tuning of general-purpose large language models. *arXiv preprint arXiv:2306.12659*.

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

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.

ZhipuAI. 2025. Glm-4v-plus. https://www.zhipuai.cn/.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. *arXiv preprint arXiv:2105.07624*.

Fengbin Zhu, Junfeng Li, Liangming Pan, Wenjie Wang, Fuli Feng, Chao Wang, Huanbo Luan, and Tat-Seng Chua. 2025a. Fintmmbench: Benchmarking temporal-aware multi-modal rag in finance. *arXiv preprint arXiv:2503.05185*.

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*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2025b. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.

A Details of VisFinEval

A.1 Design and Examples of Financial Business Scenarios

We list the detailed information of VisFinEval data in Table 4. Since we adopt a three-tier structure of financial business scenarios, comprising front-office, mid-office, and back-office. The detailed information of the financial business cenarios are presented below.

Financial Knowledge and Data Analysis includes the following seven financial business scenarios:

Financial Data Statistics Organizing and analyzing enterprise or market financial data to support subsequent modeling and decision-making.

Financial Information Extraction Extracting key information from structured or unstructured data using NLP and computer vision techniques.

Financial Indicator Assessment Interpreting key financial indicators to assess operational capabilities and profitability of enterprises.

Financial Entity Relationships Interpretation Analyzing the logical and business relationships among institutions, individuals, and products presented in texts or images.

Stock Selection Strategies Backtesting Designing and backtesting quantitative stock selection strategies to evaluate historical performance and robustness.

Candlestick Chart Analysis Interpreting price trends and technical signals embedded in candlestick charts to support technical trading analysis.

Financial Seal Recognition Automatically detecting and verifying seals in financial documents (e.g., bills, contracts) to assist in compliance auditing.

Financial Analysis and Decision includes four core business scenarios:

Industry Analysis and Inference Leveraging industry data and trends to analyze industrial chain structures, competitive landscapes, and development trajectories.

Investment Analysis Evaluating asset allocation, valuation levels, and market outlooks to support investment decisions with quantitative insights.

Financial Market Sentiment Analysis Mining sentiment signals from sources such as news and social media to inform market forecasting and risk control.

Financial Scenario Analysis Identifying and modeling typical financial events, market behaviors, or trading contexts to assist in complex scene understanding.

Financial Risk Control and Asset Optimization includes four high-level financial task scenarios:

Financial Strategy Optimization Optimizing trading, investment, or risk management strategies under given constraints to improve the risk-return profile.

Financial Risk and Policy Analysis Identifying and quantifying systemic financial risks, and assessing the impact of macroeconomic and regulatory policy shifts on market stability and institutional behavior.

Financial Data Reasoning and Interpretation Building predictive models and causal inference frameworks from multi-source financial data to support strategic decision-making.

Asset Allocation Analysis Optimizing allocations across multiple asset classes to balance risk and return, aiming to construct optimal investment portfolios.

Figure 4, 5, 6, 7, 8, 9, 10 are the examples of Financial Knowledge and Data Analysis. Figure 11, 12, 13, 14 are the examples of Financial Analysis and Decision Support. Figure 15, 16, 17, 18 are the examples of Financial Risk Control and Asset Optimization.

A.2 Details of Quality Control

The similarity heatmap for scene classification of the QA pairs is shown in the figure 3, with (b) corresponding to this task. The results show that the similarity between Qwen-max's performance on this task and human performance is relatively high. The QA data underwent a three-stage quality control process to ensure accuracy and domain relevance:

Automated Filtering Based on Multi-Dimensional Evaluation Metrics We develop an automatic screening process driven by a set of prompts and scoring criteria, including image information density, semantic validity of the QAs, data diversity, objectivity, and computational complexity. Qwen-VL-Plus-latest is employed to score and filter the generated QA pairs. The similarity heatmap for quality filtering of the QA pairs is shown in (a) of Figure 3. This stage focuses on removing incorrect answers, highly ambiguous data, and constructing a clean base dataset. Prompt examples used in this phase are shown in Table 13 and Table 14.

Manual Annotation In the second stage, all QA

pairs are manually annotated by six trained undergraduate students majoring in finance. These annotators have a strong background in financial knowledge and undergo a rigorous training process, including tests to ensure their competence in evaluating the correctness, domain specificity, verifiability of answers, completeness of visual elements, contextual alignment with financial scenarios, and logical coherence of question design. Only after passing the evaluation phase are annotators allowed to proceed with batch annotation, ensuring the overall consistency and accuracy of the labeled data.

Final Review by Financial Experts The third stage involves a comprehensive review by three financial experts, each with over a decade of experience in finance (the same domain experts who contribute to prompt design). The review focuses on several critical aspects, including logical rigor, policy compliance, decisional determinacy, accuracy of terminology, and completeness of scenario coverage. Each QA pair had to be unanimously approved by all three experts to pass this stage. This final review guarantees that each QA item is well-designed, has a unique correct answer, and faithfully reflects real-world financial business logic.

A.3 Examples of Environmental Perturbation Simulations

In real-world financial applications, environmental perturbations often arise from practical factors such as document quality degradation, scanning errors, complex layout structures, or missing information. To simulate these common sources of noise and disturbance in financial document processing, VisFinEval incorporates a set of environmental perturbation tasks. These simulations provide a more realistic assessment of model performance under non-ideal visual conditions. We categorize four representative types of perturbations as follows:

Key Information Occlusion Critical regions of the image—such as data tables, axis labels, or seal texts—are partially obscured or blurred. Figure 19 presents an example of this perturbation.

Redundant Image Perturbation The original image is overlaid or mixed with visually similar but irrelevant graphical content, such as unrelated charts. Figure 20 illustrates this type of perturbation.

Missing Relevant Information The image lacks the information explicitly referenced in the question, simulating scenarios where relevant content is omitted due to formatting issues or cross-page references. Figure 21 provides an example of this case.

Irrelevant Information Perturbation Unrelated content is added to the image without removing the original task-relevant information, resulting in semantic interference. Figure 22 demonstrates this perturbation type.

These four types of perturbations reflect common visual complexities in real financial scenarios and enable a systematic evaluation of multimodal large language models under environmentally degraded conditions.

B Details of MLLMs

We list details of the MLLMs evaluated using VisFinEval in Table 8.

B.1 Details of Evaluation Results

In our experiments, we conducted an in-depth comparative analysis of the performance of different models in specific financial business scenarios. The experimental results show that the models exhibited significant performance differentiation across scenarios of varying difficulty levels. Notably, the Qwen-VL-max ranked first across all three difficulty levels, demonstrating outstanding performance and strong adaptability.

Financial Knowledge and Data Analysis The comparative results of different models under the Financial Knowledge and Data Analysis scenario are detailed in Table 5, Qwen-VL-max secured first place in multiple tasks, ultimately achieving the top accuracy of 85.8. It is noteworthy that Moonshot-V1-32k-vision-preview far outperformed other models in the FSR(Financial Seal Recognition) task with the accuracy of 98.0, only 2 points behind human experts.

Financial Analysis and Decision Support Table 6 shows how each model performed in the Financial Analysis and Decision Support scenario, InternVL3-78B, which ranked first in the FSA (Financial Scenario Analysis) task, secured the second position overall with the accuracy of 76.4. This surpassed Qwen-VL-max-latest, which performed excellently in the IA (Investment Analysis) and FMASA (Financial Market Anomaly Sentiment Analysis) tasks. Furthermore, InternVL3-78B was only 0.5 points behind the top-ranked Qwen-VL-max, making it the most powerful open-source model.

Financial Risk Control and Asset Optimization

Table 7 summarizes the performance of the models in the more complex scenario, there was a considerable gap between the models' performance and that of human experts. Although the Qwen-VL-max model ranked first with the accuracy of 59.1, it was still 23.7 points lower than human experts. This indicates that there is still significant room for improvement in model performance in complex financial business scenarios.

Overall, while some models can approach human experts in specific scenarios, a significant disparity remains between models and human experts in tasks of higher complexity and difficulty.

B.2 Examples for Error Analysis

In this section, we explain in detail the meaning of six types of errors of MLLMs in financial business scenarios and provide examples and related error analysis.

Market sentiment and semantic tendency misjudgment The model is unable to accurately capture the front and back logic and key nodes of business operations, resulting in output results that are out of sync with real business processes or do not conform to real business thinking logic. An example of this can be seen in Figure 23.

Financial subject identification and causation confusion The model has insufficient understanding of the definitions and calculation logic of specialized terms or financial indicators, which may easily lead to conceptual confusion or numerical calculation errors. An example of this can be seen in Figure 24.

Hallucination generation and irrational reasoning The model in the parsing of financial texts, public opinion, research reports, etc., misjudges emotional tendency and semantic intensity, ignoring or misinterpreting the policy signals and industry atmosphere. An example of this can be seen in Figure 25.

Bias in the understanding of financial terms and indicators Difficulty for models to effectively integrate charts, tables and contextual textual information, leading to biased understanding of trends, data correlations, or visualization content. An example of this can be seen in Figure 26.

Lack of cross-modal information alignment capability The model is unable to accurately discern the actual causal relationship between companies, industries, or indicators, and is prone to treating correlation as causation or confusing the roles of different subjects. An example of this can be seen in Figure 27.

Perceived barriers to financial business processes The model may "make up" facts or make illogical inferences when information is missing or ambiguous, and the output does not match the reality. An example of this can be seen in Figure 28.

C Prompts Used in This Study

We provide representative prompt examples for question generation, image or question quality verification. Specifically, the prompt examples for question generation are shown in Table 9, Table 10, Table 11, Table 12. Prompt examples for quality verification are shown in Table 13, Table 14.

Prompts for financial scenario classification are shown in Table 15, Table 16, Table 17.

D Analysis of CoT Results

D.1 Evaluation results for CoT

To provide a comprehensive assessment of model performance under the Chain-of-Thought (CoT) reasoning paradigm, we present detailed evaluation results across multiple financial task domains and a direct comparison with Zero-shot performance. Table 18 summarizes the CoT performance of various models, covering three core evaluation dimensions: Financial Knowledge and Data Analysis (encompassing metrics such as FDS, CCA, and FIE), Financial Analysis and Decision Support (including FSA, IAI, and FMSA), and Financial Risk Control and Asset Optimization (featuring FSO, FRPA, and AAA). Specifically, InternVL3-VL-78B and Qwen-VL-max show relatively strong performance with high WA scores, while models like GLM-4-Vision-0924 and LLaVA-NEXT-13B have lower WA scores and face operational constraints, indicating poorer overall performance under the CoT paradigm in financial tasks. Additionally, the weighted average (WA) score for each model is calculated to reflect overall performance, with operational constraints (e.g., Multi-image Limit and Context Window Limit) noted for relevant models.

To further investigate the impact of reasoning strategies on model performance, Table19presents a head-to-head comparison between Chain-of-Thought (CoT) and Zero-shot results. This table includes the Weighted Average (WA) scores of each model under both reasoning settings and quantifies the performance difference (CoT minus Zero-shot), enabling a clear identification

Table 4: Financial Scenario Data Distribution. This table systematically presents the distribution of financial scenarios across the three progressive depths of the VisFinEval dataset, as follows: financial literacy and data analytics covering 8,700 questions, financial analytics and decision support covering 4,650 questions, and financial risk control and asset optimization covering 2,498 questions, culminating in 15,848 questions that have been rigorously manually annotated. This structured presentation accurately assesses the benchmark's ability to simulate real-world financial complexity through increasing difficulty.

Scenario Depth	Financial Scenario	Questions
Financial Knowledge and Data Analysis	Financial Data Statistics	3655
	Candlestick Chart Analysis	1124
	Financial Indicator Assessment	1160
	Financial Entity Relationships Interpretation	919
	Stock Selection Strategies Backtesting	719
	Financial Information Extraction	924
	Financial Seal Recognition	199
	All	**8700**
Financial Analysis and Decision Support	Financial Scenario Analysis	2040
	Industry Analysis and Inference	1361
	Investment Analysis	933
	Financial Market Sentiment Analysis	316
	All	**4650**
Financial Risk Control and Asset Optimization	Financial Strategy Optimization	111
	Financial Risk and Policy Analysis	181
	Financial Data Reasoning and Interpretation	1839
	Asset Allocation Analysis	367
	All	**2498**
VisFinEval	All	**15848**

of whether CoT enhances or diminishes performance relative to the Zero-shot approach for each model.Specifically, 6 models, namely Gemini-2.5-pro-exp-03-25, Claude-3-7-Sonnet-20250219, Step-Io-vision-32k, InternVL3-8B, Qwen2.5-VL-7B, and Llama-3.2-11B-Vision-Instruct, show improved performance with CoT (positive differences). The majority of the remaining models (15 in total as counted) exhibit a declining performance trend (negative differences). Overall, only a few models achieve performance optimization through CoT, indicating that the general positive effect of CoT on model performance is limited in this scenario.

D.2 CoT case study

To more intuitively demonstrate the error cases of models under the Chain-of-Thought (CoT) setting, we strategically selected representative error examples from two prominent models, namely Qwen2.5-VL-72B and Claude-3-7-Sonnet-20250219. For each of these two models, we curated two distinct types of errors: one being the scenario of "incorrect analysis accompanied by an incorrect answer" and the other representing "correct analysis yet an ultimately incorrect answer". It's noteworthy that these cases are all sourced from questions that models were capable of answering accurately under the zero-shot setting, which makes the performance discrepancies under the CoT framework even more striking.

These error cases are presented in the form of diagrams. Specifically, the two error cases corresponding to Qwen2.5-VL-72B can be found in Figure 29 and Figure 30, while those for Claude-3-7-Sonnet-20250219 are showcased in Figure 31 and Figure 32. Through these visual representations, we can clearly observe the performance differences of the models when employing different reasoning methods. In turn, this provides tangible and concrete examples that effectively underpin the subsequent analysis of the factors.

E Source of data

In the VisFinEval data generation process, image data is sourced primarily from PDF files within the financial sector. Specifically, line charts, histograms, pie charts, and financial relationship diagrams are mainly derived from financial research reports, while certain questions originate from professional examinations such as the CPA and Chinese Actuary examinations. Financial statements and supporting data tables are sourced from annual

Table 5: Evaluation Results of Financial Knowledge and Data Analysis. This table presents comparative evaluation results of various LLMs in Financial Knowledge and Data Analysis scenario. "Human" refers to the scores of human experts in the test, and the last column shows the average scores for each respective model.

			Financial Knowledge and Data Analysis							
Model	**Size**	**Limit**	**FDS**	**CCA**	**FIA**	**FERI**	**SSSB**	**FIE**	**FSR**	**WA**
Qwen-VL-max	Unknown	/	**78.8**	**90.5**	**87.4**	**89.2**	**86.2**	90.6	77.9	**85.8**
Qwen-VL-max-latest	Unknown	/	76.0	84.5	86.1	87.1	79.3	88.6	84.4	83.7
InternVL3-78B	78B	/	71.2	83.5	71.4	86.7	79.5	87.8	87.4	81.1
Doubao-1.5-vision-pro-32k	Unknown	/	75.6	79.0	84.2	85.5	76.8	**91.7**	74.4	81.0
InternVL2.5-78B	78B	/	73.3	77.9	72.3	84.2	84.0	88.4	82.9	80.4
Qwen2.5-VL-72B	72B	/	75.9	77.0	72.8	85.4	81.5	88.3	80.4	80.2
GPT-4o-2024-11-20	Unknown	/	72.0	76.8	74.9	81.7	71.8	83.8	83.9	77.8
Step-1o-vision-32k	Unknown	/	48.9	78.4	80.2	84.1	75.3	88.2	**98.0**	79.0
Moonshot-V1-32k-vision-preview	Unknown	/	56.2	82.8	73.4	80.5	73.9	87.6	68.3	74.7
Qwen2.5-VL-7B	7B	/	71.4	75.9	69.2	80.9	74.0	85.5	69.9	75.3
InternVL3-8B	8B	/	68.2	78.0	62.8	87.0	74.1	84.0	77.4	75.9
Gemini-2.5-pro-exp-03-25	Unknown	/	73.6	76.7	72.6	81.0	73.0	89.4	87.4	79.1
Claude-3-7-Sonnet-20250219	Unknown	/	70.5	73.4	80.3	71.1	77.5	83.2	34.7	70.1
Qwen2.5-VL-3B	3B	/	69.5	81.1	65.9	76.6	73.6	83.4	72.4	74.6
MiniCPM-V-2.6	8B	/	61.3	83.5	56.9	76.7	75.2	73.4	80.9	72.5
Llama-3.2-11B-Vision-Instruct	11B	/	56.9	40.8	59.3	63.9	62.9	73.1	70.4	61.0
Molmo-7B-D-0924	7B	/	60.1	74.8	54.5	62.2	59.1	60.5	42.2	59.1
GLM-4v-Plus-20250111	Unknown	Multi-image Limit	73.8	86.6	87.9	87.5	81.2	89.3	72.7	82.7
LLaVA-NEXT-34B	34B	Context Window Limit	55.3	79.8	92.3	63.2	87.8	55.0	58.8	70.3
LLaVA-v1.6-Mistral-7B	7B	Context Window Limit	54.6	73.4	65.9	62.1	47.4	47.0	62.3	58.9
LLaVA-NEXT-13B	13B	Context Window Limit	50.2	64.8	43.9	57.2	62.5	50.2	38.7	52.5
Human	/	/	95.8	92.5	83.7	96.8	91.8	92.4	100	93.3

reports, such as those from EastMoney and Hibor. seal images are sourced from open-source datasets such as (Gmgge, 2024), while candlestick charts are sourced from publicly accessible financial websites like SinaFinance and 10jqka. Specific website information is presented in Table 20.

F Explanation of Financial Business Scenarios and Category Design

To address reviewers' concerns regarding the financial scenario classifications and jargon usage, we elaborate on the design logic and rationale here.

Rationale for Scenario Selection and Design The classification of financial business scenarios, including "Financial Scenario Analysis, Industry Analysis and Inference, Investment Analysis, and Financial Market Sentiment Analysis", is deeply rooted in industry expertise rather than arbitrary LLM-generated labels. Specifically, these scenarios were collaboratively defined by financial experts involved in data quality assurance for this study. Their design (encompassing front-office, middle-office, and back-office financial business logics) directly reflects experts' insights into real-world financial workflows — a perspective rarely prioritized in prior MLLM evaluation research.This meticulous design stems from a critical need: most existing studies lack depth in financial business processes. By embedding scenario classifications into the full financial workflow, our benchmark aims to rigorously test MLLMs' true applicability in finance — an objective that simpler, more generic category systems cannot achieve.

Role of Human Oversight in Classification While LLMs assisted in classifying QA pairs (via carefully constructed prompts to improve efficiency), human expertise remained central to quality control. After LLM-driven classification, financial experts sampled and evaluated results to verify alignment between question categorization and real financial business logic. This hybrid process (LLM for efficiency + human for accuracy) ensured our scenario labels are both operationally feasible and substantively valid, avoiding "verbosity" or irrelevant jargon that purely LLM-generated categories might introduce.

G Details on Undergraduate Annotators' Recruitment, Training, and Annotation Tools

To address the reviewer's concern regarding the qualification verification and training of the six undergraduate students involved in data quality control, we provide detailed information on their recruitment criteria, training process, and the annotation tool used, as follows.

Table 6: Evaluation Results of Financial Analysis and Decision Support. This table presents comparative evaluation results of various LLMs in Financial Analysis and Decision Support scenario. "Human" refers to the scores of human experts in the test, and the last column shows the average scores for each respective model.

			Financial Analysis and Decision Support				
Model	**Size**	**Limit**	**FSA**	**IAI**	**IA**	**FMASA**	**WA**
Qwen-VL-max	Unknown	/	65.3	**83.1**	82.3	76.8	**76.9**
Qwen-VL-max-latest	Unknown	/	59.6	82.6	**82.8**	**79.3**	76.1
InternVL3-78B	78B	/	**64.3**	82.1	80.4	78.7	76.4
Doubao-1.5-vision-pro-32k	Unknown	/	56.7	80.2	79.8	77.3	73.5
InternVL2.5-78B	78B	/	63.3	81.5	80.1	75.2	75.0
Qwen2.5-VL-72B	72B	/	57.4	82.4	80.3	74.5	73.6
GPT-4o-2024-11-20	Unknown	/	61.9	77.9	78.5	73.2	72.9
Step-1o-vision-32k	Unknown	/	40.3	78.8	78.6	76.1	68.4
Moonshot-V1-32k-vision-preview	Unknown	/	61.9	77.7	77.0	72.3	72.2
Qwen2.5-VL-7B	7B	/	53.4	79.7	76.5	70.7	70.1
InternVL3-8B	8B	/	56.5	76.1	76.8	71.7	70.3
Gemini-2.5-pro-exp-03-25	Unknown	/	53.2	72.4	70.8	75.5	68.0
Claude-3-7-Sonnet-20250219	Unknown	/	48.0	76.1	75.5	64.0	65.9
Qwen2.5-VL-3B	3B	/	50.0	75.4	74.7	66.6	66.7
MiniCPM-V-2.6	8B	/	48.3	69.7	70.7	69.1	64.5
Llama-3.2-11B-Vision-Instruct	11B	/	45.3	69.7	67.1	63.4	61.4
Molmo-7B-D-0924	7B	/	39.7	64.4	62.8	63.4	57.5
GLM-4v-Plus-20250111	Unknown	Multi-image Limit	56.5	78.1	74.9	74.6	71.0
LLaVA-NEXT-34B	34B	Context Window Limit	54.3	88.2	88.1	66.9	74.4
LLaVA-v1.6-Mistral-7B	7B	Context Window Limit	42.3	58.3	56.4	63.7	55.2
LLaVA-NEXT-13B	13B	Context Window Limit	34.7	59.2	59.0	52.9	51.4
Human	/	/	88.8	90.6	87.1	85.3	88.0

Recruitment Requirements for Undergraduate Annotators To ensure the professionalism and reliability of data annotation (consistent with the high-quality standards of subsequent expert review), we established strict recruitment criteria. Only candidates meeting all the following basic requirements were selected:

I. Academic Background: Have completed core financial courses, including Financial Engineering, Investment, Financial Management, Financial Markets, Econometrics, and Insurance; maintain an average GPA of 3.5 or above (on a 4.0 scale) to demonstrate solid academic performance in financial knowledge.

II. Practical Experience: Possess no less than 3 months of internship experience in front-line business departments of financial institutions (e.g., securities companies, commercial banks), to ensure familiarity with real-world financial business scenarios and data contexts.

III. Certification Preference: Priority was given to candidates holding internationally or domestically recognized financial certifications, such as ACCA (Association of Chartered Certified Accountants), FRM (Financial Risk Manager), CFA (Chartered Financial Analyst), or CPA (Certified Public Accountant), to further validate their domain expertise.

Key Points of Annotator Training After recruitment, all undergraduate annotators underwent systematic training (led by the research team and financial experts) to standardize their annotation operations and ensure alignment with the study's quality requirements. I. Familiarize with relevant annotation documents and learn annotation requirements and operations. II. Ensure image clarity, removing VQA data with issues such as blurriness or occlusion. III. Ensure consistency between image content and question-answer pairs, and accuracy of answers. IV. Align business depth and question difficulty with financial experts to ensure differentiation in question-answer pair difficulty and business alignment.

Note: I. For annotation convenience, considering the computer proficiency of undergraduate finance students, we provided an Excel spreadsheet as an annotation tool. Annotators select questions via column headers and judge/filter question quality. From left to right, the columns are Index Number, Image Address, Image Name, Image, Question, Options, Answer, and Retain/Discard. Annotators need to be familiar with the annotation format and operation methods. II. Expert-related training for the fourth point in the training key points is handled

Table 7: Evaluation Results of Financial Risk Control and Asset Optimization. This table presents comparative evaluation results of various LLMs in Financial Risk Control and Asset Optimization. "Human" refers to the scores of human experts in the test, and the last column shows the average scores for each respective model.

			Financial Risk Control and Asset Optimization				
Model	**Size**	**Limit**	**FSO**	**FRPA**	**FDRI**	**AAA**	**WA**
Qwen-VL-max	Unknown	/	**49.1**	**58.2**	**58.2**	71.0	**59.1**
Qwen-VL-max-lastest	Unknown	/	44.0	52.2	48.9	71.8	54.2
InternVL3-78B	78B	/	**49.1**	52.8	46.6	66.5	53.7
Doubao-1.5-vision-pro-32k	Unknown	/	30.0	54.5	54.5	**75.6**	53.7
InternVL2.5-78B	78B	/	41	53.1	47.6	68.4	52.5
Qwen2.5-VL-72B	72B	/	41.4	53.4	42.6	71.9	52.3
GPT-4o-2024-11-20	Unknown	/	41	40.5	41.6	67.9	47.7
Step-1o-vision-32k	Unknown	/	39.2	45.2	49	65.8	49.8
Moonshot-V1-32k-vision-preview	Unknown	/	39.2	55.8	53.6	64.0	53.1
Qwen2.5-VL-7B	7B	/	37.2	37.6	35.4	63.2	43.4
InternVL3-8B	8B	/	29.7	46.2	36.8	55.3	42.0
Gemini-2.5-pro-exp-03-25	Unknown	/	28.4	38	41.5	37.7	36.4
Claude-3-7-Sonnet-20250219	Unknown	/	26.8	50.3	48.6	64.4	47.5
Qwen2.5-VL-3B	3B	/	22.9	34.8	35.9	53.8	36.9
MiniCPM-V-2.6	8B	/	20.6	35.5	26.8	52.7	33.9
Llama-3.2-11B-Vision-Instruct	11B	/	18.0	22.1	19.9	31.1	22.8
Molmo-7B-D-0924	7B	/	23.4	31.7	21.9	26.5	25.9
GLM-4v-Plus-20250111	Unknown	Multi-image Limit	45.1	54.1	45.3	73.2	54.4
LLaVA-NEXT-34B	34B	Context Window Limit	13.1	17.5	12.7	7.7	12.7
LLaVA-v1.6-Mistral-7B	7B	Context Window Limit	10.2	16.3	35.9	21.1	20.9
LLaVA-NEXT-13B	13B	Context Window Limit	14.7	10.8	15.8	31.1	18.1
Human	/	/	84.4	80.4	81.1	85.2	82.8

by experts; due to personal reasons of the experts, specific individual experiences in the financial industry cannot be publicly shared.

Table 8: 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
Close-Source	Qwen-VL-max-latest	Alibaba Cloud	Undisclosed	API	2025.1
	Qwen-VL-max	Alibaba Cloud	Undisclosed	API	2025.1
	Step-1o-vision-32k	StepStar	Undisclosed	API	2025.1
	Gemini-2.5-pro-exp-03-25	Google	Undisclosed	API	2025.3
	GPT-4o-2024-11-20	OpenAI	Undisclosed	API	2024.11
	Moonshot-V1-32k-vision-preview	MoonshotAI	Undisclosed	API	2025.1
	Claude-3-7-Sonnet-20250219	Anthropic	Undisclosed	API	2024.10
	Doubao-1.5-vision-pro-32k	ByteDance	Undisclosed	API	2025.1
	GLM-4v-Plus-20250111	Zhipu.AI	Undisclosed	API	2025.1
Open-Source	Qwen2.5-VL-3B	Alibaba Cloud	3B	Weights	2025.1
	Qwen2.5-VL-7B	Alibaba Cloud	7B	Weights	2025.1
	Qwen2.5-VL-72B	Alibaba Cloud	72B	Weights	2025.1
	MiniCPM-V-2.6	OpenBMB	8B	Weights	2025.1
	InternVL3-8B	Shanghai AI Lab	8B	Weights	2025.4
	InternVL2.5-78B	Shanghai AI Lab	78B	Weights	2024.12
	InternVL3-78B	Shanghai AI Lab	78B	Weights	2025.4
	Molmo-7B-D-0924	Allen Institute for AI	7B	Weights	2024.9
	Llama-3.2-11B-Vision-Instruct	Meta AI	11B	Weights	2024.9
	LLaVA-v1.6-Mistral-7B	Liu et.al	7B	Weights	2024.5
	LLaVA-NeXT-13B	LLaVA-VL	13B	Weights	2024.1
	LLLaVA-NeXT-34B	LLaVA-VL	34B	Weights	2024.1

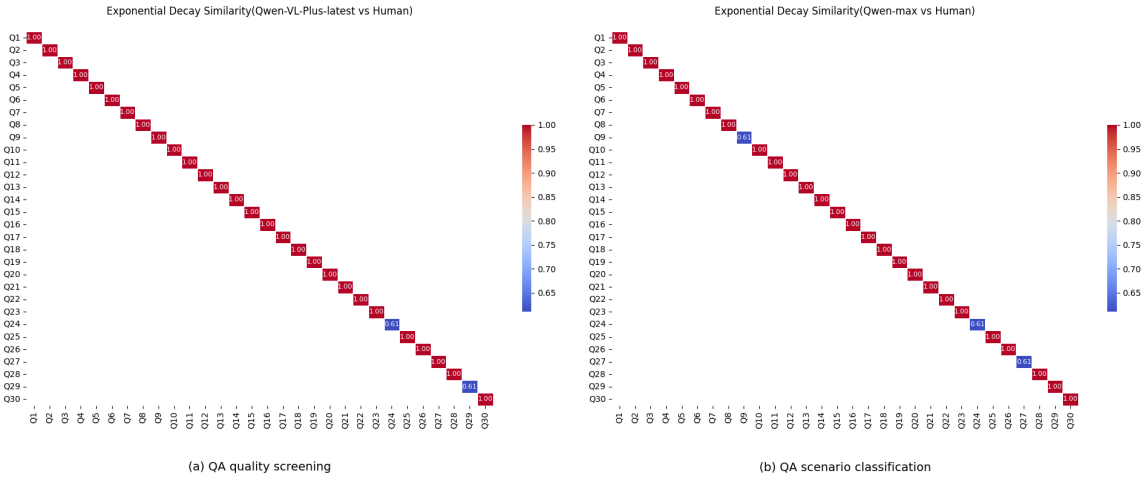
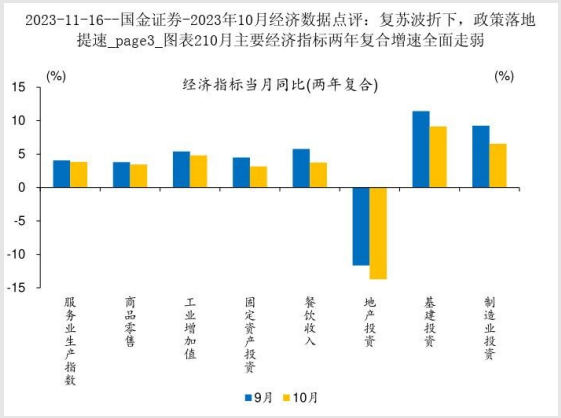


Figure 3: The graphs show the exponential decay similarity between LLMs and human evaluation. (a) depicts the similarity between LLM used for quality screening (Qwen-VL-Plus-latest) and human evaluation, where a value of 1 indicates complete similarity, and a value of 0.61 represents a non-ideal match. (b) illustrates the similarity between LLM used as a classifier (Qwen-max) and human evaluation, with the same similarity scale: 1 for complete similarity and 0.61 for a lower match.



问题：哪一项经济指标在9月和10月均出现了负增长?

Question：Which economic indicator showed negative growth in both September and October?

A：餐饮收入

A：Catering revenue

B：地产投资

B：Real estate investment

C：固定资产投资

C：Fixed asset investment

答案：B

Answer：B

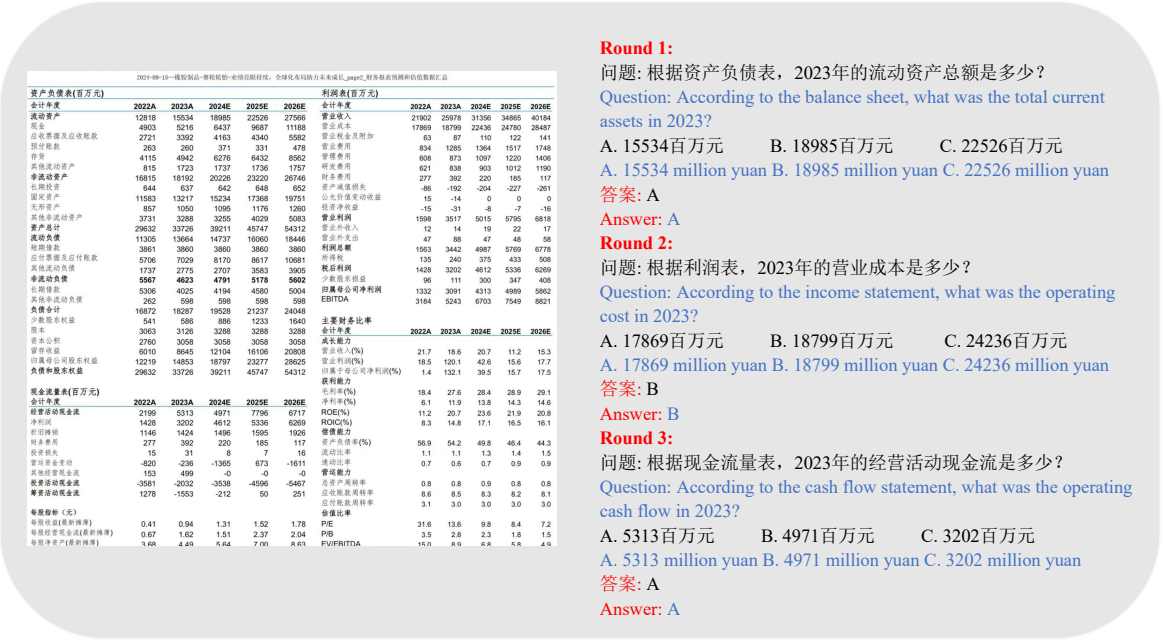
Figure 4: This is a three-option single-choice question related to Financial Data Statistics scenario. To answer this question accurately, the model must read the chart, identify the monthly growth rates of various economic indicators, and determine whether they showed negative growth in both September and October. This requires cross-temporal comparison and judgment of negative trends. The question assesses the model's ability to extract consistent temporal trends from time series indicators, testing its precision in structured understanding and numerical reasoning over financial data.

问题：根据 K 线图和 MACD 指标，当前股票的走势是?
Question：Based on the candlestick chart and the MACD indicator, what is the current trend of the stock?

A：上涨趋势
A：Uptrend
B：下跌趋势
B：Downtrend
C：震荡整理
C：Sideways consolidation
D：无法判断
D：Unable to determine

答案：B; C
Answer：B; C

Figure 5: This is a four-option multiple-choice question focused on Candlestick Chart Analysis scenario. By identifying the price patterns in the candlestick chart and determining whether the MACD indicator has formed a "death cross" or "golden cross," the model must judge whether the stock is in an upward, downward, or sideways trend. Accurate answering requires interpretation of price action and understanding of MACD crossovers. The question evaluates the model's ability to jointly reason over technical charts and financial indicators, testing its multimodal chart understanding and trend analysis capabilities in financial contexts.



资产负债表(百万元)

会计年度	2022A	2023A	2024E	2025E	2026E
流动资产	12818	15534	18985	22526	27566
现金	4903	5216	6437	9687	11188
应收票据及应收账款	2721	3392	4163	4340	5582
预付账款	263	260	371	331	478
存货	4115	4942	6276	6432	8562
其他流动资产	815	1723	1737	1736	1757
非流动资产	16815	18192	20226	23220	26746
长期投资	644	637	642	648	652
固定资产	11583	13217	15234	17368	19751
无形资产	857	1050	1095	1176	1260
其他非流动资产	3731	3288	3255	4029	5083
资产总计	29632	33726	39211	45747	54312
流动负债	11305	13664	14737	16060	18446
短期借款	3861	3860	3860	3860	3860
应付票据及应付账款	5706	7029	8170	8617	10681
其他流动负债	1737	2775	2707	3583	3905
非流动负债	**5567**	**4623**	**4791**	**5178**	**5602**
长期借款	5306	4025	4194	4580	5004
其他非流动负债	262	598	598	598	598
负债合计	16872	18287	19528	21237	24048
少数股东权益	541	586	886	1233	1640
股本	3063	3126	3288	3288	3288
资本公积	2760	3058	3058	3058	3058
留存收益	6010	8645	12104	16106	20808
归属母公司股东权益	12219	14853	18797	23277	28625
负债和股东权益	29632	33726	39211	45747	54312

现金流量表(百万元)

会计年度	2022A	2023A	2024E	2025E	2026E
经营活动现金流	2199	5313	4971	7796	6717
净利润	1428	3202	4612	5336	6269
折旧摊销	1146	1424	1496	1595	1926
财务费用	277	392	220	185	117
投资损失	15	31	8	7	16
营运资金变动	-820	-236	-1365	673	-1611
其他经营现金流	153	499	-0	-0	-0
投资活动现金流	-3581	-2032	-3538	-4596	-5467
筹资活动现金流	1278	-1553	-212	50	251
每股指标（元）					
每股收益(最新摊薄)	0.41	0.94	1.31	1.52	1.78
每股经营现金流(最新摊薄)	0.67	1.62	1.51	2.37	2.04
每股净资产(最新摊薄)	3.68	4.49	5.64	7.00	8.63

利润表(百万元)

会计年度	2022A	2023A	2024E	2025E	2026E
营业收入	21902	25978	31356	34865	40184
营业成本	17869	18799	22436	24780	28487
营业税金及附加	63	87	110	122	141
营业费用	834	1285	1364	1517	1748
管理费用	608	873	1097	1220	1406
研发费用	621	838	903	1012	1190
财务费用	277	392	220	185	117
资产减值损失	-86	-192	-204	-227	-261
公允价值变动收益	15	-14	0	0	0
投资净收益	-15	-31	-8	-7	-16
营业利润	1598	3517	5015	5795	6818
营业外收入	12	14	19	22	17
营业外支出	47	88	47	48	58
利润总额	1563	3442	4987	5769	6778
所得税	135	240	375	433	508
税后利润	1428	3202	4612	5336	6269
少数股东损益	96	111	300	347	408
归属母公司净利润	1332	3091	4313	4989	5862
EBITDA	3184	5243	6703	7549	8821

主要财务比率

会计年度	2022A	2023A	2024E	2025E	2026E
成长能力					
营业收入(%)	21.7	18.6	20.7	11.2	15.3
营业利润(%)	18.5	120.1	42.6	15.6	17.7
归属于母公司净利润(%)	1.4	132.1	39.5	15.7	17.5
获利能力					
毛利率(%)	18.4	27.6	28.4	28.9	29.1
净利率(%)	6.1	11.9	13.8	14.3	14.6
ROE(%)	11.2	20.7	23.6	21.9	20.8
ROIC(%)	8.3	14.8	17.1	16.5	16.1
偿债能力					
资产负债率(%)	56.9	54.2	49.8	46.4	44.3
流动比率	1.1	1.1	1.3	1.4	1.5
速动比率	0.7	0.6	0.7	0.9	0.9
营运能力					
总资产周转率	0.8	0.8	0.9	0.8	0.8
应收账款周转率	8.6	8.5	8.3	8.2	8.1
应付账款周转率	3.1	3.0	3.0	3.0	3.0
估值比率					
P/E	31.6	13.6	9.8	8.4	7.2
P/B	3.5	2.8	2.3	1.8	1.5
EV/EBITDA	15.0	8.9	6.8	5.8	4.9

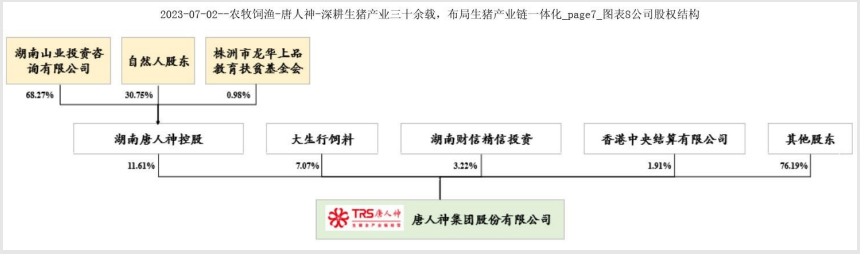
Round 1:
问题: 根据资产负债表，2023年的流动资产总额是多少？
Question: According to the balance sheet, what was the total current assets in 2023?
A. 15534百万元　　B. 18985百万元　　C. 22526百万元
A. 15534 million yuan B. 18985 million yuan C. 22526 million yuan
答案: A
Answer: A
Round 2:
问题: 根据利润表，2023年的营业成本是多少？
Question: According to the income statement, what was the operating cost in 2023?
A. 17869百万元　　B. 18799百万元　　C. 24236百万元
A. 17869 million yuan B. 18799 million yuan C. 24236 million yuan
答案: B
Answer: B
Round 3:
问题: 根据现金流量表，2023年的经营活动现金流是多少？
Question: According to the cash flow statement, what was the operating cash flow in 2023?
A. 5313百万元　　B. 4971百万元　　C. 3202百万元
A. 5313 million yuan B. 4971 million yuan C. 3202 million yuan
答案: A
Answer: A

Figure 6: This is a single-choice question involving Financial Indicator Assessment, serving as an example of such analysis. Accurately answering this question requires the large model to identify the specified data year and specific accounting items in the question, locate the corresponding line items in the financial statements on the left, and then verify the data units and numerical precision while excluding distractors in the options. The accurately extracted financial data must then be compared with each option one by one. By examining key data from the three core financial statements (balance sheet, income statement, and cash flow statement), the question assesses the large model's fundamental ability to interpret a company's financial condition.For better readability, the English translation is displayed below the corresponding Chinese text.

Table 9: Prompt Template for Constructing Four-Option Multiple-Choice Questions Based on Line Charts

```
你是一名金融分析师，请根据提供的折线图，生成三道四选题。
题目应基于折线图中的数据趋势、关键点或特征。
要求：
    1. 每道问题必须清晰明确，选项应具有区分度。
    2. 每道题的选项 A、B、C、D 应涵盖不同的可能性，避免过于简单或明显。
    3. 每道题的答案必须是 A、B、C 或 D 中的一个。
    4. 每道题的问题长度不少于 10 字。
    5. 三道题目必须完全不同，且每道题需要编号为 1、2、3。
    6. **只输出 JSON 格式的内容，不要包含任何额外的描述性文本。**
你可以参考的示例：
{random_few_shots}
输出格式为：
[
    {"id": "1", "q": "问题1", "A": "选项A", "B": "选项B", "C": "选项C", "Answer": "正确答案"},
    {"id": "2", "q": "问题2", "A": "选项A", "B": "选项B", "C": "选项C", "Answer": "正确答案"},
    {"id": "3", "q": "问题3", "A": "选项A", "B": "选项B", "C": "选项C", "Answer": "正确答案"}
]
```

(a) Chinese Version



问题：湖南唐人神控股通过子公司间接持有唐人神集团股份有限公司的比例大于其直接持股比例。
Question：Hunan Tangrenshen Holdings holds a larger equity stake in Tangrenshen Group Co., Ltd. indirectly through its subsidiaries than it does directly.

答案：是
Answer：True

Figure 7: This is a true/false question in the scenario of Financial Entity Relationships Interpretation. To answer correctly, the model needs to analyze a corporate structure diagram or table, extract both direct and indirect shareholding paths, and compute the aggregate stake to determine whether indirect holdings exceed direct ones. The main challenge lies in multi-layered structural parsing and path aggregation. This question assesses the model's ability to parse corporate ownership structures and reason over control paths, testing its accuracy in abstracting and comparing hierarchical entity relationships.

问题：从K线图中可以看出，该股票在 2024年12月6日出现了明显的长下影线。请问这种形态通常预示着什么?
Question：From the K-line chart, it can be seen that this stock had a clear long lower shadow on December 6, 2024. What does this pattern usually indicate?"

A：股价将继续下跌
A：The stock price will continue to decline.
B：股价将出现反弹
B：The stock price will rebound.
C：股价将横盘整理
C：The stock price will fluctuate sideways.

答案：B
Answer：B

Figure 8: This is an example of Stock Selection Strategies Backtesting. Accurately answering this question requires the large model to identify the characteristics of the candlestick pattern and analyze the market implications of a long lower shadow. By recognizing the candlestick pattern (long lower shadow), this question tests the large model's quantitative application ability regarding technical analysis indicators.For better readability, the English translation is displayed below the corresponding Chinese text.

2024-08-22--银行-常熟银行-详解常熟银行2024年半年报：净利润同比+19.6%，业绩高增韧性强_page7_图表12常熟银行十大股东持股情况1H24

股东名称	方向	持股数量(股)	持股数量变动(股)	占总股本比例(%)	持股比例变动(%)
交通银行股份有限公司	不变	271,586,361	24689669	9.01	
香港中央結算有限公司	增持	111,029,056	30,747,084	3.68	0.75
全国社保基金110组合	不变	99,888,312	9080756	3.31	
常熟市发展投资有限公司	不变	92,875,077	8443189	3.08	
江苏江南商贸集团有限责任公司	不变	84,365,429	7,669,584	2.80	
全国社保基金16042组合	不变	82,155,796	7,468,709	2.72	
全国社保基金413组合	减持	70,474,418	4,006,765	2.34	-0.09
全国社保基金401组合	不变	37,506,415	3,409,674	1.24	
中证500交易型开放式指数证券投资基金	增持	34,428,731	4218266	1.14	0.04
富兰克林国海中小盘股票型证券投资基金	减持	31,053,119	-176,900	1.03	-0.11
合　计		915,362,714		30.35	

问题：根据图表，交通银行股份有限公司的持股数量为多少?
Question：According to the chart, what is the number of shares held by Bank of Communications Co., Ltd.?

A：271,586,361
B：24,689,669
C：9.01

答案：A
Answer：A

Figure 9: This is a three-option single-choice question in the scenario of Financial Information Extraction. To answer correctly, the model must locate the relevant row and column associated with the Bank of Communications in a tabular or graphical chart and extract the corresponding numerical value. The key lies in precise visual localization and accurate data extraction. This question tests the model's ability to locate and extract key information from structured visual content, evaluating its accuracy in structured vision-language understanding and entity-value alignment.

问题：图中所示公章属于哪家公司？
Question：To which company does the official seal in the image belong?

A：中审（深圳）认证有限公司
A：China Audit (Shenzhen) Certification Co., Ltd.
B：中审（广州）认证有限公司
B：China Audit (Guangzhou) Certification Co., Ltd.
C：中审（北京）认证有限公司
C：China Audit (Beijing) Certification Co., Ltd.

答案：A
Answer：A

Figure 10: This is a three-option single-choice question on Intelligent Seal Recognition. To answer this question accurately, the model must analyze the textual structure and formatting of the seal shown in the image and compare it against the names of candidate institutions. The focus is on recognizing key terms in the seal and determining hierarchical or departmental alignment. This question assesses the model's capacity to understand textual structures within stamp images and judge visual-semantic similarity, testing its fine-grained multimodal entity recognition and image-text alignment skills.

Round 1:

问题：结合金融理论，如果全球大豆播种面积持续增加，可能会对大豆期货市场产生什么影响？

Question: According to financial theory, if global soybean planting area continues to increase, what impact might this have on the soybean futures market?

A. 价格上涨　　B. 价格下跌　　C. 价格波动不大

A. Price increase　　B. Price decrease　　C. Little price fluctuation

答案: A

Answer: A

Round 2:

问题：假设全球小麦播种面积在未来几年内继续稳定增长，这对相关农业公司的股价有何潜在影响？

Question: Assuming global wheat planting area continues to grow steadily in the coming years, what potential impact would this have on the stock prices of related agricultural companies?

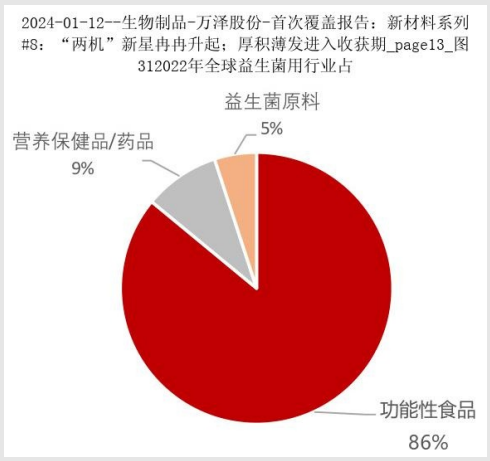
A. 股价上涨　　B. 股价下跌　　C. 股价波动不大

A. Stock price increase　　B. Stock price decrease　　C. Little stock price fluctuation

答案: A

Answer: A

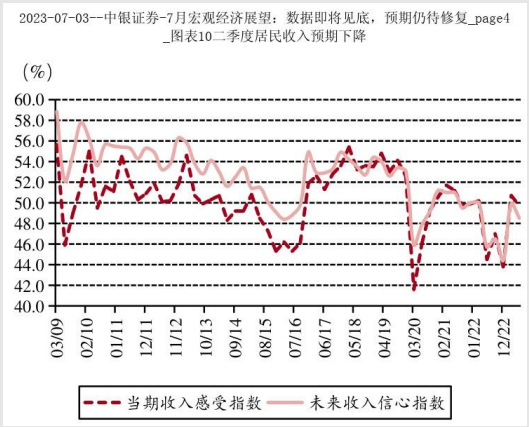
Figure 11: This is a single-choice question involving Industry Analysis and Inference. Accurately answering this question requires the large model to analyze the trend of chart data, observe the changes in the bar chart of global major crop planting areas from 2014 to 2023, and combine the question to interpret the market impact mechanism and summarize the transmission logic. By examining the transmission impact of changes in global crop planting areas on futures markets and the stock prices of listed companies, the question tests the large model's comprehensive analytical capabilities regarding the supply-demand relationship in the agricultural industry chain and investment logic.For better readability, the English translation is displayed below the corresponding Chinese text.



问题：如果功能性食品的市场份额减少5%，其仍然会是最大的市场部分。
Question：If the market share of functional foods decreases by 5%, it would still remain the largest market segment.

答案：是
Answer：True

Figure 12: This is a counterfactual inference question within the Financial Scenario Analysis. A correct answer requires the model to perform hypothetical adjustments to the original market share data and determine whether functional foods would still hold the largest market share after a 5% decrease. The key lies in constructing a counterfactual scenario and comparing adjusted values. The question examines the model's sensitivity to causal changes among variables and the rigor of its reasoning process, testing its ability in numerical inference and logical reasoning under hypothetical financial settings.

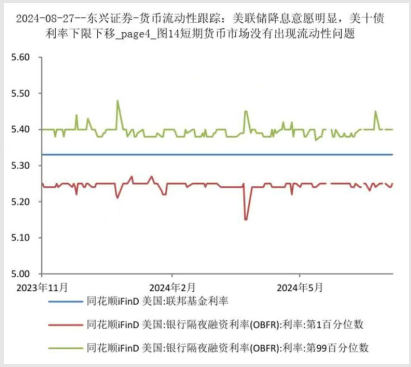


问题：如果未来收入信心指数在2018年后持续上升，当期收入感受指数会如何反应？
Question：If the income confidence index continues to rise after 2018, how will the current income perception index respond?

A：同步上升
A：Rise simultaneously
B：先上升后下降
B：Rise first, then decline
C：保持稳定
C：Remain stable
D：持续下降
D：Continue to decline

答案：A
Answer：A

Figure 13: This is a example for Financial Market Sentiment Analysis. To answer this question accurately, the model must understand the logical relationship between the income confidence index and the income perception index, typically assuming that rising confidence leads to a rise in perception. The key is grasping trend co-movement and the economic implications of sentiment indicators. This question evaluates the model's understanding of dynamic relationships among macro sentiment variables, testing its capacity in predictive reasoning and sentiment-driven analyzing in financial psychology contexts.



Round 1:
问题：如果联邦基金利率持续低于银行隔夜融资利率，对金融市场会产生什么影响？
Question: If the federal funds rate remains consistently lower than the overnight bank financing rate, what impact would this have on financial markets?
A. 金融市场流动性紧张　　B. 金融市场流动性宽松　　C. 金融市场风险增加
A. Tight financial market liquidity B. Loose financial market liquidity C. Increased financial market risk
答案: B
Answer: B

Round 2:
问题：从图中可以看出，银行隔夜融资利率的波动性如何？
Question: From the chart, what can be observed about the volatility of the overnight bank financing rate?
A. 波动性较小　　B. 波动性较大　　C. 波动性适中
A. Low volatility　　B. High volatility　　C. Moderate volatility
答案: B
Answer: B

Round 3:
问题：银行隔夜融资利率的波动性对金融市场有何影响？
Question: What impact does the volatility of the overnight bank financing rate have on financial markets?
A. 对金融市场影响不大
A. Minimal impact on financial markets
B. 可能引发金融市场的动荡
B. May cause turbulence in financial markets
C. 对金融市场影响较小
C. Small impact on financial markets
答案: B
Answer: B

Figure 14: This is a standard Investment Analysis question. To respond accurately, the model must analyze the trends of three interest rate curves in the chart, interpret liquidity easing signals based on the question's context, and assess OBFR rate volatility's impact on financial markets to evaluate short-term money market liquidity risks and stability. The question tests the model's ability to analyze monetary policy rate differentials, market rate volatility, and their transmission effects, assessing its comprehensive understanding of money market liquidity and systemic risk mechanisms.

2024-05-07--化学制品-聚胶股份-卫材热熔胶全球破局者，海外扩张稳步推进_page28_表9假设主要原材料价格以5的幅度变动下公司毛利率敏感性分析

矿物油/石油树脂	0.89	0.95	1.00	1.05	1.10	1.16	1.21
1.06	28%	25%	23%	21%	18%	16%	14%
1.13	27%	24%	22%	20%	17%	15%	13%
1.19	26%	23%	21%	19%	16%	14%	12%
1.25	25%	22%	20%	18%	15%	13%	11%
1.31	24%	21%	19%	17%	14%	12%	10%
1.38	23%	20%	18%	16%	13%	11%	9%
1.44	21%	19%	17%	15%	12%	10%	8%

Image

2024-05-07--化学制品-聚胶股份-卫材热熔胶全球破局者，海外扩张稳步推进_page28_表9假设主要原材料价格以5的幅度变动下公司毛利率敏感性分析

矿物油/石油树脂	0.89	0.95	1.00	1.05	1.10	1.16	1.21
1.06	28%	25%	23%	21%	18%	16%	14%
1.13	27%	24%	22%	20%	17%	15%	13%
1.19	26%	23%	21%	19%	16%	14%	12%
1.25	25%	22%	20%	18%	15%	13%	11%
1.31	24%	21%	19%	17%	14%	12%	10%
1.38	23%	20%	18%	16%	13%	11%	9%
1.44	21%	19%	17%	15%	12%	10%	8%

MarkDown Document

问题：假设公司当前的毛利率为25%，根据表格数据，如果矿物油/石油树脂的价格从0.95上升到1.10，毛利率将降至多少？

Question：Assuming that the company's current gross margin is 25%, based on the data in the table, if the price of mineral oil/petroleum resins rises from 0.95 to 1.10, how much will the gross margin fall to?

A：0.17

B：0.18

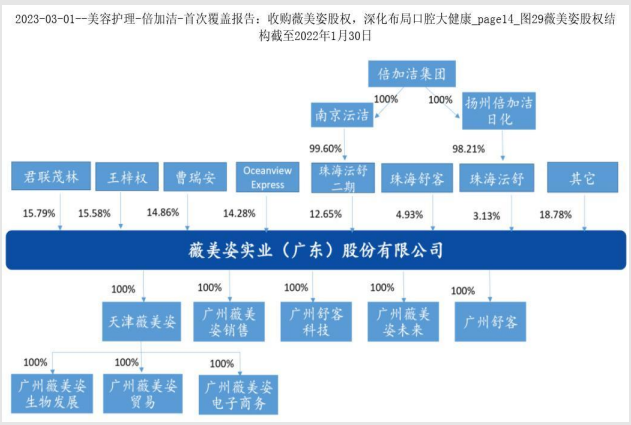
C：0.19

D：0.2

答案：B

Answer：B

Figure 15: This is an example of for Financial Strategy Optimization. First, the large model needs to locate the row and column in the table where the price of mineral oil/petroleum resin is 0.95 and the corresponding gross margin is 25%, then find the corresponding gross margin value when the price rises to 1.10 in the row, and finally match the value with the options. This question tests the model's ability to accurately find and locate the data in the table and analyze it according to the correlation between the data in financial scenarios. In addition, it also verifies the model's ability to extract consistent key data from images (visual) and structured text (Markdown table) to get the correct answer.



问题：在薇美姿实业（广东）股份有限公司的股权结构中，哪些股东持股比例低于20%？

Question：Which shareholders hold less than 20% of the shares in the shareholding structure of WEIMEIZI Industrial (Guangdong) Co., Ltd?

A：君联茂林

A：Beijing Legend Maolin Equity Investment Partnership Enterprise (Limited Partnership)

B：王梓权

B：Ziquan Wang

C：曹瑞安

C：Ruian Cao

D：Oceanview Express

D：Oceanview Express

答案：A、B、C、D

Answer：A、B、C、D

Figure 16: This is a question for Asset Allocation Analysis. Answering this question requires extracting information about the identity of shareholders in the equity structure diagram, obtaining their shareholdings, and comparing the 20% threshold. The question tests the large model's ability to understand and extract data from the mapping of financial relationships, equity penetration analysis, and the ability to make judgments about specific conditions.

问题：在本周人民币兑主要货币汇率波动中，哪些货币的波动幅度超过2%？

Question：Which currencies fluctuated by more than 2% during the week's fluctuations in the yuan against major currencies?

A：日元

A：Japanese yen （JPY,）

B：英镑

B：Great Britain Pound（GBP）

C：加元

C：Canadian dollar（CAD）

D：韩元

D：South Korean Won（KRW）

答案：A、C、D

Answer：A、C、D

Figure 17: This question is a multiple-choice item centered on Financial Risk and Policy Analysis. The critical task is to accurately extract the percentage fluctuations in the exchange-rate chart and benchmark them against a 2 percent threshold. The item evaluates the model's proficiency in recognizing financial data, assessing risk thresholds, and providing decision support.

2024-09-18--电子化学品-安集科技-业绩稳健增长，先进制程产品持续上量_page1_图表1公司2024年上半年财务报告摘要

（百万元）	2024 年上半年	2023 年上半年	同比增长（%）
一、营业总收入	797.27	574.91	38.68
二、营业总成本	555.29	397.67	39.64
其中：营业成本	337.10	258.18	30.57
营业税金及附加	1.02	0.65	57.16
销售费用	23.76	20.86	13.87
管理费用	50.35	30.83	63.33
研发费用	144.57	101.30	42.72
财务费用	(7.93)	(12.44)	-
资产减值损失	(2.78)	(0.65)	-
三、其他经营收益	18.70	83.29	(77.54)
公允价值变动收益	(7.24)	7.47	-
投资收益	(2.84)	6.23	(145.67)
四、营业利润	250.60	273.53	(8.38)
加：营业外收入	0.02	0.00	-
减：营业外支出	0.52	0.00	-
五、利润总额	250.10	273.53	(8.57)
减：所得税	16.10	38.52	(58.20)
六、净利润	234.00	235.01	(0.43)
减：少数股东损益	0.00	0.00	-
七、归属母公司净利润	234.00	235.01	(0.43)

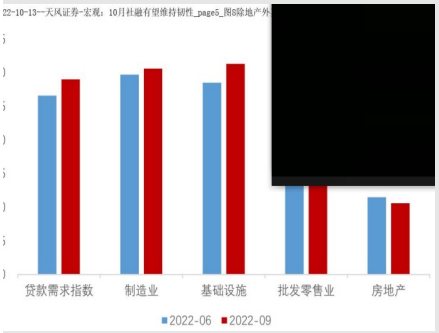
问题：如果2024年上半年的管理费用占总成本的比例与2023年上半年相同，那么2024年上半年的管理费用应该是多少百万元？

Question：If administrative expenses as a percentage of total costs in the first half of 2024 are the same as in the first half of 2023, how many millions of dollars should administrative expenses be in the first half of 2024?

A：50.35
B：30.83
C：40.09
D：60

答案：C
Answer：C

Figure 18: This is a multiple choice question involving Financial Data Reasoning and Interpretation. First, the large model needs to identify the overhead and total cost data extracted from the first half of 2023 and calculate its ratio; use this ratio to extrapolate with the total cost data obtained in 2024, and finally calculate the theoretical value of overhead in 2024. The question tests the ability of the large model to extract, calculate, and logically extrapolate financial statement data.



问题：从2022年6月到2022年9月，哪个行业的贷款需求指数增长最多？

Question: From June 2022 to September 2022, which industry experienced the highest increase in the loan demand index?

A: 制造业
A: Manufacturing
B: 基础设施
B: Infrastructure
C: 批发零售业
C: Wholesale and Retail Trade
D: 无法判断
D: Can not determine

答案：D
Answer: D

Figure 19: This is an example of Key Information Occlusion. The upper-right corner of the histogram contains loan demand index data for the wholesale and retail sector, as well as the real estate industry. However, this critical region is obscured, preventing the model from accurately extracting the necessary information. To generate correct responses under such conditions, the model must have the ability to detect occlusion or missing information.

在2024年的财务报告中，春风动力和北极星两家公司在全地形车及中大排摩托车市场上的表现引人注目。从财务指标来看，两家公司均展现了强劲的增长势头，尤其是在营收、利润、ROE、ROIC等关键指标上。春风动力通过推出多款高性能车型，如2025 CFORCE 1000 TOURING和2025 UFORCE 1000，成功吸引了大量消费者，推动了销售额的显著增长。同时，公司采取了一系列成本控制措施，优化了供应链管理，降低了生产成本，从而提高了利润率。此外，春风动力还积极拓展海外市场，特别是在东南亚和欧洲地区取得了显著进展，进一步扩大了市场份额。

北极星则通过并购和战略合作，增强了其产品线的竞争力。例如，公司收购了一家专注于高端SSV（侧边驾驶车辆）制造的企业，推出了RZR XP 1000 Sport等高性价比车型，迅速占领了市场。这些战略举措不仅提升了公司的品牌影响力，还为其带来了可观的收入增长。然而，由于市场竞争加剧，北极星在某些季度的利润受到了一定影响，但总体而言，公司的财务状况依然稳健。

宏观经济环境对两家公司的影响不容忽视。全球经济形势的不确定性导致市场需求波动，政策变化也对行业产生了重要影响。例如，政府对环保标准的提高促使企业加大研发投入，以开发更环保的产品。此外，原材料价格的波动也增加了企业的成本压力。面对这些挑战，春风动力和北极星都采取了相应的应对策略，如加强技术创新、优化产品结构等，以保持竞争优势。

从股东回报与估值角度来看，两家公司的P/E比率相对较高，表明市场对其未来增长潜力持乐观态度。尽管如此，考虑到当前的经济环境和行业竞争态势，投资者仍需谨慎评估公司的估值水平。未来股价的增长潜力将取决于公司能否持续创新并有效应对市场变化。

风险因素方面，市场竞争激烈、价格波动以及政策变化等因素可能对公司财务预期产生不利影响。为此，春风动力和北极星都在积极调整策略，如加强品牌建设、拓展多元化市场等，以降低潜在风险

短期与长期财务趋势显示，两家公司在接下来的几个季度可能会面临一定的波动，尤其是在新产品推出初期和市场适应阶段。为了应对这些变化，公司需要灵活调整经营策略，确保业务稳定增长。

问题：根据春风动力的图片信息，春风动力的2025 CFORCE 1000 TOURING和北极星的Sportsman XP 1000 Ultimate相比，哪个车型的马力更大？

In the 2024 financial report, the performance of Chunfeng Power and Beiji Xing in the all-terrain vehicle and mid-to-large displacement motorcycle markets is noteworthy. Both companies have shown strong growth in key financial metrics—such as revenue, profit, ROE, and ROIC—thanks to their efforts in cost control and market expansion. Chunfeng Power boosted sales by launching several high-performance models (e.g., the 2025 CFORCE 1000 TOURING and 2025 UFORCE 1000), implemented strict cost control measures, optimized supply chain management, and reduced production costs, thereby increasing profit margins. In addition, it expanded into overseas markets, especially in Southeast Asia and Europe, further increasing its market share.

Beiji Xing, meanwhile, enhanced its product competitiveness through acquisitions and strategic partnerships. For instance, by acquiring a company specializing in high-end SSV (side-by-side vehicles) manufacturing and launching competitively priced models like the RZR XP 1000 Sport, it quickly captured market share. These strategies not only bolstered the company's brand influence but also drove significant revenue growth. Although intensified competition affected its quarterly profits, Beiji Xing's overall financial condition remains robust.

The macroeconomic environment has also played a critical role. Global economic uncertainty has led to fluctuating market demand, while policy changes—such as stricter environmental standards prompting increased R&D for greener products—and raw material price volatility have added cost pressures. In response, both companies have adopted measures like enhancing technological innovation and optimizing product structures to maintain their competitive edge.

From the perspective of shareholder returns and valuation, both companies currently exhibit relatively high P/E ratios, suggesting that the market is optimistic about their future growth potential. However, given the current economic climate and fierce industry competition, investors should evaluate their valuations cautiously. The potential for future stock price growth will depend on their ability to continue innovating and effectively respond to market changes.

Risk factors such as intense market competition, price fluctuations, and policy changes could negatively impact financial expectations. To mitigate these risks, Chunfeng Power and Beiji Xing are actively adjusting their strategies—strengthening brand building and expanding into diversified markets—to reduce potential vulnerabilities.

Short-term and long-term financial trends indicate that the companies may experience some volatility in the coming quarters, especially during new product launches and market adaptation phases. To manage these fluctuations, they will need to flexibly adjust their business strategies to ensure stable, sustainable growth.

Question：Based on the image information from CF Moto, regarding the comparison between CF Moto's 2025 CFORCE 1000 TOURING and Polaris's Sportsman XP 1000 Ultimate, which model has greater horsepower?

A. 春风动力的2025 CFORCE 1000 TOURING

A. CF Moto's 2025 CFORCE 1000 TOURING

B. 北极星的Sportsman XP 1000 Ultimate

B. Polaris Sportsman XP 1000 Ultimate

C. 两者相同

C. Both are the same

C. 无法确定

C. Cannot be determined

答案：B

Answer: B

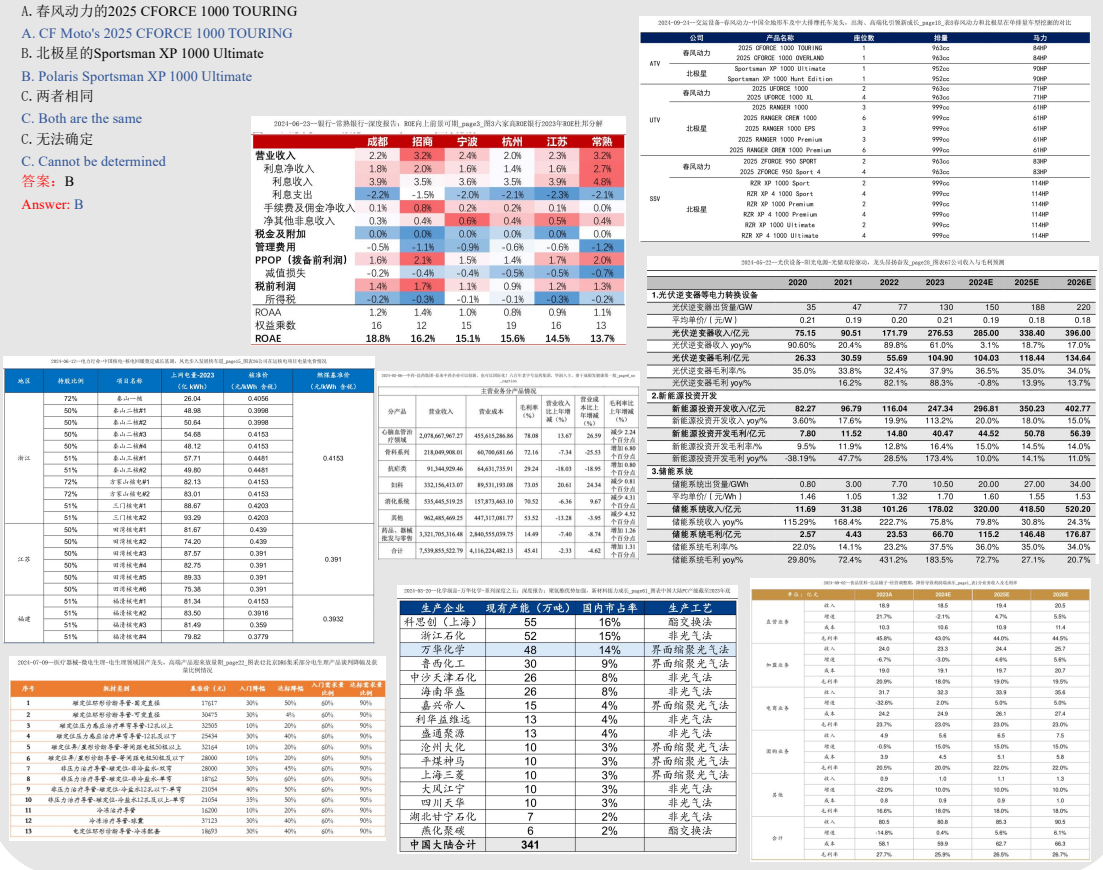
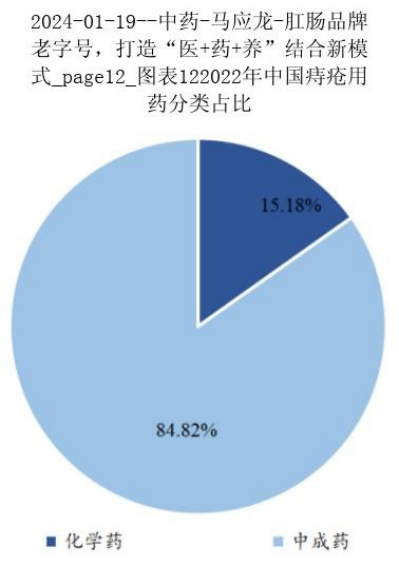


Figure 20: This is an example of Redundant Image Perturbation. As multiple images are used for perturbation in this case, the resolution has been compressed for display purposes. This example is for illustration only; the resolution of the original images remains unchanged in the actual dataset. The image contains multiple financial tables with similar formats, most of which are unrelated to the question. The model must possess effective vision-language alignment and contextual matching capabilities to accurately locate the one table that is relevant to the question in order to answer correctly.



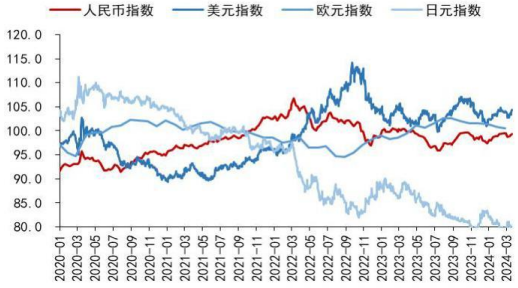
问题：根据饼图，乳膏剂在中国痔疮用药市场细分药品类型中占比多少?
Question: According to the pie chart, what is the market share of cream agents in the Chinese hemorrhoids medication market among the segmented pharmaceutical types?

A：62%
B：18%
C：无法确定
C：Cannot be determined
D：4%

答案：C
Answer：C

Figure 21: This is an example of Missing Relevant Visual Information. These types of questions often ask about content that doesn't exist in the image. Therefore, the model needs to determine whether relevant information is present in the image. This scenario simulates situations where business personnel might encounter incomplete customer information in real-world operations.

2024-03-26--国信证券-货币政策与流动性观察：资金利率温和回升_page11
_图29主要货币指数走势



美联储主席鲍威尔在JacksonHole会议暗示9月美联储有望开启降息。联邦资金利率期货隐含9月美联储FOMC降息25bp的可能性为63.5%，降息50bp的可能性为36.5%。国内流动性方面，随着税期和政府债券缴款压力度过，央行上周大幅净回笼流动性，因此资金面均衡偏紧。尽管上周隔夜和7天资金利率中枢较前一周回落，但从上周全周的资金走势看，总体“前低后高”，反映周中流动性压力有所上升。国信货币政策力度指数较前一周（8月12日-8月18日）下降0.1，位于101.42。其中价格指标上升贡献-95.6%，数量指标下降贡献195.6%。价格指数上升（宽松）主要由上周全周标准化后R加权利率偏离度、DR加权利率偏离度（相较同期限OMO）以及R001导致……

问题：图中哪个货币指数在2022年波动最大？

Fed Chair Powell hinted at the JacksonHole conference that the Fed is expected to kick off a rate cut in September. Federal funds rate futures implied a 63.5% chance of a 25bp Fed FOMC rate cut in September and a 36.5% chance of a 50bp rate cut. On the domestic liquidity front, funding was balanced tight as the central bank made a significant net liquidity repatriation last week as the tax period and government bond payment pressures passed. Although the overnight and 7-day funds rate pivot fell last week compared with the previous week, but from the last week of the whole week's funding trend, the overall “low before and high after”, reflecting the mid-week liquidity pressures have risen. The Guosen Monetary Policy Strength Index declined 0.1 from the previous week (August 12-August 18) to 101.42, with the price index rising by -95.6% and the quantity index falling by 195.6%. The rise in the price index (easing) was mainly caused by last week's full-week standardized R-weighted interest rate deviation, DR-weighted interest rate deviation (compared to OMOs of the same maturity), and R001 .....

Question: Which currency index in the chart is the most volatile in 2022?

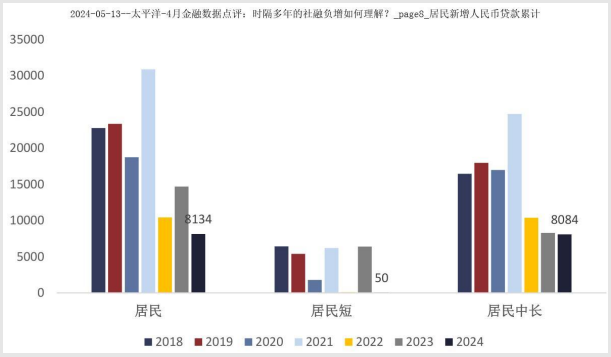
A：人民币指数　B：美元指数　C：欧元指数　D：日元指数

A：Renminbi Index　B：US dollar index　C：Euro Index　D：Japanese yen index

答案：C

Answer：C

Figure 22: This is an example of Irrelevant Information Perturbation. These types of questions add text that is similar to the question during input but actually provides no assistance for answering. This introduces semantic noise, thereby increasing the difficulty for the model to answer and simulating the real-world scenario where business personnel need to conduct business operations under the interference of a large amount of irrelevant information.



问题：如果2024年的居民中长期贷款在2023年之前达到峰值，那么2024年的居民中长期贷款会如何变化？

Question: How will residential medium- and long-term lending change in 2024 if it peaks before 2023?

A. 增加　B. 减少　C. 保持不变　D. 无法确定

A. Increased　B. Decreased　C. Remained the same　D. Unable to determine

正确答案：B

Correct Answer：B

模型输出：A

Model Output：A

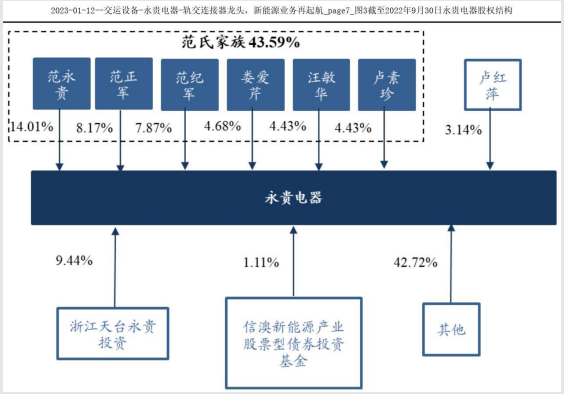
思考过程：用户根据所提供的柱形图，就 2024 年 “居民中长期贷款 ”的趋势提出一个假设问题。问题提出了一个条件：“如果 2024 年的居民中长期贷款在 2023 年之前达到峰值”。我需要查看图表中的 “居民中长期 ”部分。所示年份为 2018、2019、2020、2021、2022、2023 和 2024 年。如果直接从图中解读，问题就很简单了。如果条款是数据实际满足的条件，因此它不是脱离图表的纯粹假设情景。这更像是鉴于[图表中的事实]，[图表中的事实]是什么？“居民中长”：2023（灰条）在 2024（黑条，8084）的左边。2024 条确实明显比 2023 条高。2023 条形图的顶部低于 2024 条形图的顶部。因此，2024的值 > 2023的值。

Thinking Process：The user is asking a hypothetical question about the trend of “Resident Medium and Long-term Loans” in 2024, based on the provided bar chart.The question states a condition: "if the 2024 Resident Medium and Long-term Loans peaked before 2023".I need to look at the “Resident Medium and Long-term Loans” section of the chart.The years shown are 2018, 2019, 2020, 2021, 2022, 2023, and 2024.The question is simple if interpreted directly from the graph. The "if" clause is a condition that is actually met by the data, so it's not a purely hypothetical scenario divorced from the chart. It's more like "Given that [fact from chart], what is [other fact from chart]?"Final check of the numbers and bars:“Resident Medium and Long--term Loans”:2023 (grey bar) is to the left of 2024 (black bar, 8084).The 2024 bar is indeed visibly taller than the 2023 bar. The top of the 2023 bar is below the top of the 2024 bar.Therefore, 2024_value > 2023_value.Therefore, an increase.No other interpretation seems to make sense or lead to a different definitive answer from the options.If we had to assume the condition about the peak was counterfactual (e.g., if the peak was not before 2023 in the chart), then we might need external financial knowledge. But the condition is met, so we use the chart data.The question is about how 2024 will change, implying an observation of the 2024 data point relative to its predecessor.

错误原因：尽管图中清晰显示自2021年起该类贷款数据逐年下降，2024年继续呈现下行趋势，但模型在分析过程中忽视了图中所体现的政策收紧背景和房地产相关信贷持续承压的整体氛围，进而误导地得出了“中长期贷款增长”的结论。此外，模型在推理中对题干中“峰值”概念进行了过度解释，推理过程中出现了关于峰值的多个讨论，也误判了相关语义强度，导致最终选择了错误的选项。

Reason for Error：Although the graph clearly shows that the data of this type of loan has been declining year by year since 2021, and will continue to show a downward trend in 2024, the model ignores the background of policy tightening and the overall atmosphere of continuous pressure on real estate-related credit reflected in the graph, and then misleadingly concludes that “medium- and long-term loans are growing”. In addition, the model over-interpreted the concept of “peak” in the question in its reasoning, which led to multiple discussions of peaks in the reasoning process and misjudged the related semantic strength, resulting in the selection of the wrong option.

Figure 23: An example of Errors in misjudging Market Sentiment and Semantic Tendency Misjudgment. The large model over-interpreted the concept of “peak” in the question stem and ignored the overall atmosphere of continued pressure on real estate-related credit presented by the data trends in the chart. This reflects the fact that the model was too obsessed with a particular detail in the question and answer and lacked a big-picture view of the overall trend behind the data.



问题：范氏家族的持股比例为43.59%，高于其他股东的总持股比例。

Question: The Fan family's shareholding is 43.59%, which is higher than the total shareholding of other shareholders.

正确答案：是

Correct Answer：True

模型输出：否

Model Output：False

思考过程：用户要求根据所提供的股东结构图验证一个语句。语句是 “范氏家族的持股比例为 43.59%，高于其他股东的总持股比例。

1. **指出范氏家族的持股比例：** 图片清楚地标明 “范氏家族 43.59%”。
2. **指明 “其他股东 ”及其所占百分比：** ,* 卢红萍： 3.14%,* 浙江天台永贵投资： 9.44%,* 信澳新能源产业股票型证券投资基金： 1.11%,* 其他： 42.72%
3. **计算其他股东的总持股比例：**3.14% (卢红萍) + 9.44% (浙江天台永贵投资) + 1.11% (信澳新能源产业股票型证券投资基金) + 42.72% (其他) = 3.14 + 9.44 + 1.11 + 42.72= 12.58 + 1.11 + 42.72 = 13.69 + 42.72= 56.41%
4. **将范氏家族的持股比例与其他股东的持股比例相比较：**， * 范氏家族：43.59， * 其他股东合计：56.41% 5.
5. **评价语句：** “范氏家族的持股比例为 43.59%，高于其他股东的总持股比例”。43.59% > 56.41%？不，不是。

因此，这种说法是错误的。

Thinking Process：The user is asking to validate a statement based on the provided image, which is a shareholder structure diagram.The statement is: The Fan family's shareholding ratio is 43.59%, which is higher than the total shareholding ratio of other shareholders.

1. **Identify the Fan family's shareholding percentage:** The image clearly states the Fan family's shareholding ratio is 43.59%,
2. **Identify the "other shareholders" and their percentages:**, * Hongping Lu: 3.14%, * Zhejiang Tiantai Yonggui Investment: 9.44%, * Xinao New Energy Industry Equity Investment Fund: 1.11%, * Others: 42.72%
3. **Calculate the total shareholding ratio of other shareholders:**

3.14% (Lu) + 9.44% (Zhejiang Tiantai Yonggu Investment:) + 1.11% (Xinao New Energy Industry Equity Investment Fund) + 42.72% (Others) = 3.14 + 9.44 + 1.11 + 42.72 = 12.58 + 1.11 + 42.72 = 13.69 + 42.72= 56.41%

4. **Compare the Fan family's shareholding with the total of other shareholders:**, * Fan family: 43.59%, * Other shareholders total: 56.41%.
5. **Evaluate the statement:** "The Fan family's shareholding ratio is 43.59%, which is higher than the total shareholding ratio of other shareholders."Is 43.59% > 56.41%? No, it is not.

Therefore, the statement is false.

错误原因：模型在识别语义关系时，未能准确区分“其他股东”作为独立统计口径的主体，而是错误地将其理解为“所有非家族股东”或“全部股东之和”，从而导致对股权结构的误判。这种错误源于模型在中文语义解析中的歧义处理不足，特别是在面对“其他”这类模糊指代词时，未能结合上下文或表格信息明确其边界含义，最终影响了对核心判断项的逻辑推理与选择。

Reason for Error：The model fails to accurately distinguish “other shareholders” as the subject of independent statistics when identifying semantic relationships, but incorrectly interprets it as “all non-family shareholders” or “the sum of all shareholders”, leading to a misinterpretation of the equity structure. Instead, it is wrongly interpreted as “all non-family shareholders” or “the sum of all shareholders”, which leads to the misjudgment of the equity structure. This kind of error stems from the model's insufficient handling of ambiguity in Chinese semantic parsing, especially in the face of such vague pronouns as “other”, which fails to clarify their boundary meanings with contextual or tabular information, and ultimately affects the logical reasoning and selection of the core judgment items.

Figure 24: An example of Financial subject identification and causation confusion. Failure to accurately distinguish “other shareholders” as a subject of independent statistical caliber resulted in an incorrect answer. This reflects the model’s lack of ability to handle ambiguity in Chinese semantic parsing.

问题：图中所示公章属于哪家公司？
Question: To which company does the official seal shown in the picture belong?
A. 广西南方水泥有限公司　　B. 广西南方水泥集团有限公司
A. Guangxi Southern Cement Co., Ltd.　　B. Guangxi Southern Cement Group Co., Ltd.
C. 广西南方水泥股份有限公司
C. Guangxi Southern Cement Company Limited by Shares
正确答案：A
Correct Answer：A
模型输出：C
Model Output：C

思考过程：用户提供了一张公司印章的图片和一道多选题。我需要从印章中识别出公司名称，并选择正确的选项。看一下印章：
-中间星星周围的字似乎是“财务专用章”，意思是“财务专用章”。这在中国公司印章上很常见。
- 外圈的字符是公司名称。
- 从最左边的字符开始，顺时针方向，我依稀可以辨认出
广 、西、南、方、水 、泥 、股 （这个字是关键字）、份 （这个字也是关键字）、有、限、公、司
因此，这个名字似乎是 “广西南方水泥股份有限公司”。
现在让我们将其与选项进行比较：
A. 广西南方水泥有限公司——缺少 “股份”（股份/股票）。
B. 广西南方水泥集团有限公司—有 “集团”（无 “股份”）。
C. 广西南方水泥股份有限公司--这与我从印章上解读的相符。
印章上的 “股份”字清晰可见（“有限公 ”前的两个字）。因此，选项 C 是正确的。

Thinking Process：The user has provided an image of a company seal and a multiple-choice question. I need to identify the company name from the seal and select the correct option.Looking at the seal:
- The characters around the star in the center appear to be "Financial Special Seal". This is common on Chinese company seals.
- The characters in the outer ring are the company name.
- Starting from the leftmost character and going clockwise, I can vaguely make out:
Guǎng, Xī, Nan, Fang, Shui, Ni,Gu(This character is key.), Fen(This character is key.),You, Xian, Gong Si
So the name appears to beGuangxi Southern Cement Company Limited by Shares .
Now let's compare this to the options:
A. Guangxi Southern Cement Co., Ltd. - Missing shares/stock.
B.Guangxi Southern Cement Group Co., Ltd.- Has group instead of Share
C. Guangxi Southern Cement Company Limited by Shares - This matches what I deciphered from the seal.
The key differentiator is “Limited by Shares ” or "Joint Stock Limited Company".
Option A is a "Limited Liability Company".
Option B is a "Group Limited Liability Company".
The characters“share” are clearly visible on the seal. Therefore, option C is the correct one.

错误原因：模型的视觉感知过程在处理图像信息时出现了故障，未能正确识别公章中并不存在“股份”二字的实际情况。相反，它错误地生成原图不存在的“股”“份”两个字并将其作为关键判别依据，从而在推理过程中引入了错误的视觉输入，导致最终选择了与图像内容不符的错误选项。这种误判说明模型在图文结合判断任务中存在感知偏差，对图像中细节文字的识别准确性仍需提升。

Reason for Error：The model's visual perception process malfunctioned in processing the image information, failing to correctly recognize the fact that the word “share” did not exist in the official seal. Instead, it incorrectly generates the word “share”, which does not exist in the original image, and uses it as the key discriminative basis, thus introducing erroneous visual inputs into the reasoning process, leading to the selection of the wrong option that does not match the content of the image. This kind of misjudgment indicates that the model has perceptual bias in the judgment task of combining graphics and text, and the accuracy of recognizing the detailed text in the image still needs to be improved.

Figure 25: An example of Hallucination generation and irrational reasoning. Although the model has correctly recognized the words on the official seal, there is no information or words about “shares” on the official seal, so the model creates an illusion in the output process and uses the shares as the basis for judgment, resulting in an incorrect answer.

问题：从 K 线图中可以看出，该股票在 12 月 6 日出现了哪种形态？

Question: Which pattern can be seen on the K-line chart for the stock on December 6?

A. 早晨之星　　B. 黄昏之星　　C. 十字星

A. Morning star　　B. Evening star　　C. Doji

正确答案：C

Correct Answer：C

模型输出：A

Model Output：A

思考过程：用户提供了一张 K 线图（蜡烛图），要识别出 12 月 6 日当日出现的特定形态。我需要定位到 12 月 6 日，并分析当天的蜡烛图走势。定位 12 月 6 日： 横轴显示日期，可以看到“12.6 09:35”和“12.6 13:05”。分析 12 月 6 日的形态：在“12.6 09:35”附近，先是一根较长的蓝色（空头）蜡烛，随后是一根小实体且下影线很长的蜡烛，再接着又是一根小实体蜡烛。这一序列如果前面是下跌趋势、后面又有上涨确认，就类似“早晨之星”。但题目问的是 12 月 6 日“出现了”什么形态，而不是包含 12 月 6 日的多日形态。再看 12 月 6 日的单根关键蜡烛：大约在“12.6 09:35”形成当日最低点的那根蜡烛，实体非常小，下影线很长，上影线几乎没有，这种小实体长下影的形态可以视为“星形”蜡烛。当天其他蜡烛多为小实体，但最显著的是这根“星形”蜡烛及其后的反转信号。评估选项：A. 早晨之星（Morning Star）：是一个三根蜡烛组成的多头反转形态：①下跌趋势中的大阴线；②一根小实体或十字星，其交易区间位于第一根阴线下方；③紧接着的一根大阳线，其收盘需回补第一根阴线实体的一半以上。图中在 12 月 6 日 09:35 附近确有这样三根 5 分钟蜡烛的组合：一根长阴、一根小实体“星”、随后一根明显的长阳，形成典型的“早晨之星”反转序列。B. 黄昏之星（Evening Star）：是三根蜡烛的空头反转形态，与此处的底部反转不符。C. 十字星（Doji）：虽然当天有多根小实体蜡烛符合“十字星”特征，但整体更具意义的是形成了完整的“早晨之星”多头反转信号。综合来看，12 月 6 日当日的最显著形态是那三根蜡烛构成的早晨之星。

Thinking Process: The user has provided a K-line chart (candlestick chart) to identify a specific pattern that occurred on December 6, the same day. I need to locate December 6 and analyze the candlestick movement for that day. Locating December 6: The horizontal axis shows the date and you can see “12.6 09:35” and “12.6 13:05”. Analyzing the pattern of December 6: near “12.6 09:35”, there was a long blue (short) candle, followed by a small solid candle with a long lower shadow, and then another small solid candle. This sequence is similar to a “Morning Star” if it is preceded by a downtrend and followed by an uptrend confirmation. But the title asks what pattern “appeared” on December 6, not the multi-day pattern that encompassed it. Look at the single key candle of December 6: the one that formed the low of the day around “12.6 09:35”, the entity is very small, the lower shadow is very long, and the upper shadow is almost non-existent. Candle. The other candles of the day were mostly small, but the most notable was this “star” candle and its subsequent reversal signal. Evaluating the options: A. Morning Star: A long reversal pattern consisting of three candles: (1) a large negative line in a downtrend; (2) a small solid or doji star with a trading range below the first negative line; and (3) a large positive line immediately afterward, with the close required to make up more than half of the solidity of the first negative line. The chart at 09:35 on December 6, near the combination of three 5-minute candles: a long negative, a small solid “star”, followed by a clear long positive, forming a typical “morning star” reversal sequence. b. Twilight Star (Evening Star: A three-candle short reversal pattern that does not match the bottom reversal here. C. Doji: Although there were multiple small solid candles on the day that fit the Doji profile, the overall more significant feature was the formation of a full Morning Star .The Doji: Although there were a number of small candles on that day that fit the Doji profile, the overall significance was the formation of a complete Morning Star. Taken together, the most notable pattern of the day on December 6 was the three candles that formed the Morning Star.

错误原因：误将局部“星形”蜡烛认定为早晨之星中心烛，大模型将12月6日早盘09:35那根小实体长下影蜡烛视作早晨之星的第二根“星形”烛，并强行寻找前后“第一烛”“第三烛”以验证其三烛模式。同时混淆“十字星”与“早晨之星”的本质，大模型在详述“早晨之星”后，又承认当日存在多根小实体蜡烛，可归为“十字星”，但最终仍坚持选A，忽视C选项“十字星”才是最直接、唯一符合当日形态特征的答案。

Reason for Error: Mistakenly identified the local “star-shaped” candle as the morning star center candle, the big model will be the morning session of December 6, 09:35 that small entity long lower shadow candle as the morning star of the second “star” candle, and forced to look for before and after the The “first candle” and “third candle” to verify its three-candle pattern. At the same time confuse the nature of the “cross” and “morning star”, the big model in the details of the “morning star”, and then recognized the existence of a number of small solid candles that day, can be categorized as “The big model, after detailing the Morning Star, recognizes the existence of multiple small candles on the day, which can be classified as a Doji, but ultimately insists on choosing A, ignoring the fact that option C, the Doji, is the most direct and only answer that matches the pattern of the day.

Figure 26: An example of Bias in the understanding of financial terms and indicators. Although the Model correctly extracted information such as the time point and the shape of the K-line chart, it confused the difference between different K-line patterns and eventually chose the wrong answer. This reflects the Big Model's lack of ability to discriminate between financial terminology and indicators.

2024-06-25--中药-白云山-公司深度报告：华南医药消费龙头稳健增长，盈利能力持续优化_page15_表1大健康板块同行企业估值对比

股票代码	公司简称	收盘价（元）	总市值（亿）	EPS（元/股）			PE		
				2023A	2024E	2025E	2023A	2024E	2025E
000848.SZ	承德露露	7.88	82.94	6.38	6.84	7.52	13.00	12.13	11.04
603156.SH	养元饮品	21.39	270.69	14.67	16.70	18.92	18.45	16.21	14.31
600887.SH	伊利股份	25.97	1653.28	104.29	130.12	125.23	15.85	12.71	13.20
605499.SH	东鹏饮料	217.80	871.22	20.40	26.55	33.33	42.71	32.81	26.14
605337.SH	李子园	9.76	38.50	2.37	2.77	3.25	16.25	13.90	11.86
603711.SH	香飘飘	13.54	55.61	2.80	3.47	4.33	19.84	16.02	12.83
可比公司平均 PE							21.02	17.30	14.90

问题：如果以2023年的市盈率（PE）为基准，哪家公司相对于可比公司平均PE的折价幅度最大？

Question: If the price-to-earnings (PE) ratio for 2023 is used as a benchmark, which company has the largest discount to the average PE of comparable companies?

A. 承德露露

A. Cheng De Lolo Co,.Ltd.

B.养元饮品

B. HebeiYangyuanZhiHui Beverage Co., Ltd.

C. 伊利股份

C. Inner Mongolia Yili Industrial Group Co.,Ltd.

D.李子园

D. Zhe Jiang Li Zi Yuan Food CO.,Ltd.

正确答案：A

Correct Answer：A

模型输出：C

Model Output：C

股票代码	公司简称	收盘价(元)	总市值(亿)	EPS(2023A)(元/股)	PE(2023A)	EPS(2024E)(元/股)	PE(2024E)	EPS(2025E)(元/股)	PE(2025E)
000848.SZ	承德露露	7.88	82.94	6.38	13	6.84	12.13	7.52	11.04
603156.SH	养元饮品	21.39	270.69	14.67	18.92	16.7	18.45	18.92	16.21
600887.SH	伊利股份	25.97	1653.28	104.29	12.71	130.12	15.85	125.23	13.2
605499.SH	东鹏饮料	217.8	871.22	20.4	42.71	26.55	32.81	33.33	26.14
605337.SH	李子园	9.76	38.5	2.37	16.25	2.77	13.9	3.25	11.86
603711.SH	香飘飘	13.54	55.61	2.8	19.84	3.47	16.02	4.33	12.83
可比公司平均PE	-	-	-	-	21.02	-	17.3	-	14.9

问题：如果以2023年的市盈率（PE）为基准，哪家公司相对于可比公司平均PE的折价幅度最大？

Question: If the price-to-earnings (PE) ratio for 2023 is used as a benchmark, which company has the largest discount to the average PE of comparable companies?

A. 承德露露

A. Cheng De Lolo Co,.Ltd.

B.养元饮品

B. HebeiYangyuanZhiHui Beverage Co., Ltd.

C. 伊利股份

C. Inner Mongolia Yili Industrial Group Co.,Ltd.

D.李子园

D. Zhe Jiang Li Zi Yuan Food CO.,Ltd.

正确答案：A

Correct Answer：A

模型输出：A

Model Output：A

错误原因：大模型在无法从包含相同信息的图片和文本（markdown格式）中根据问题提取到相同的关键信息，导致两次回答不一致。

Reason for Error：The large model was unable to extract the same key information based on the question from images and text (markdown format) that contained the same information, leading to inconsistencies in the two responses.

Figure 27: An example of Lack of cross-modal information alignment capability. This error type is generally a failure of the model to effectively combine image, chart, table, and text semantics, resulting in an incorrect trend determination or a numerical reading error. Although the large model found the key information and answered the question correctly in the text format of markdown, it answered the question incorrectly in the image format, containing the same information, which reflects the large model's lack of cross-modal information alignment ability.

单位：万元

资产负债项目	2016 年末
货币资金	600
应收账款	1600
存货	1500
固定资产	8300
资产总计	12000
应付账款	1000
其他流动负债	2000
长期借款	3000
股东权益	6000
负债及股东权益总计	12000
利润表项目	2016 年度
营业收入	16000
减：营业成本	10000
税金及附加	560
销售费用	1000
管理费用	2000
财务费用	240
利润总额	2200
减：所得税费用	550
净利润	1650

背景信息：甲公司是一家新型建筑材料生产企业，为做好 2017 年财务计划，拟进行财务报表分析和预测。相关资料如下：（1）甲公司 2016 年主要财务数据如图所示。（2）公司没有优先股且没有外部股权融资计划，股东权益变动均来自留存收益。公司采用固定股利支付率政策，股利支付率 60%。（3）销售部门预测 2017 年公司营业收入增长率 10%。（4）甲公司的企业所得税税率 25%。

Backgroud information：Company A is a new building material manufacturer, and it is going to analyze and forecast the financial statements for the purpose of making a good financial plan in 2017. The relevant information is as follows: (1) Company A's key financial data for 2016 are shown in the chart. (2) The company has no preferred stock and no external equity financing plan, and all changes in stockholders' equity come from retained earnings. The company adopts a fixed dividend payout ratio policy with a dividend payout ratio of 60%. (3) The sales department forecasts a 10% growth rate in the company's operating income for 2017. (4) Company A has a corporate income tax rate of 25%.

问题：假设 2017 年甲公司除长期借款外所有资产和负债与营业收入保持 2016 年的百分比关系，所有成本费用与营业收入的占比关系维持 2016 年水平，用销售百分比法初步测算公司 2017 年融资总需求和外部融资需求。

Question: Assuming that in 2017, all assets and liabilities of Company A, except for long-term borrowings, maintain the percentage relationship with operating revenues as in 2016, and that the percentage relationship of all costs and expenses with operating revenues maintains the level of 2016, the percentage of sales method is used to preliminarily measure the Company's total financing needs and external financing needs in 2017..

正确答案：900, 174　　模型输出：1200,174

Correct Answer：900,174　　Model Output：1200,174

思考过程：1.分析2016年财务数据及计算相关比率：营业收入 (S_{2016})： 16000，总资产 (A_{2016})： 12000，资产与营业收入比率 (A/S) = 12000 / 16000 = 0.75……2. 预测2017年营业收入及利润：2017年预计营业收入 (S_{2017})：$S_{2017} = S_{2016} \times (1 + \text{营业收入增长率}) = 16000 \times (1 + 0.10) = 17600$,……3. 预测2017年资产和自发性负债：2017年预计总资产 (A_{2017})：$A_{2017} = (A/S) \times S_{2017} = 0.75 \times 17600 = 13200$；2017年预计自发性负债 ($L^*_{2017}$)：$L_{2017} = (L/S) \times S_{2017} = 0.1875 \times 17600 = 3300$。4. 计算2017年融资总需求融资总需求通常指为支持销售增长所需增加的总资产额。资产增加额 (ΔA) = $A_{2017} - A_{2016} = 13200 - 12000 = 1200$，融资总需求 = 1200 万元。5. 计算2017年外部融资需求外部融资需求（AFN）是指在考虑了自发性负债增长和留存收益增加后，仍需从外部筹集的资金。自发性负债增加额 (ΔL^*) = $L_{2017} - L_{2016} = 3300 - 3000 = 300$，外部融资需求 (AFN) = 资产增加额 ($\Delta A$) - 自发性负债增加额 ($\Delta L^*$) - 预计增加的留存收益 ($\Delta RE_{2017}$)AFN = 1200 - 300 - 726 = 174，外部融资需求 = 174 万元。由于公司没有外部股权融资计划，这174万元的外部融资需求将通过增加负债（如增加长期借款）来满足。最终结论：根据销售百分比法测算：公司2017年融资总需求为 1200 万元。公司2017年外部融资需求为 174 万元。

Thinking Process：1. Analyze 2016 Financial Data and Calculate Relevant Ratios:**Operating Revenue (S_{2016}):** 16,000，**Total Assets (A_{2016}):** 12,000 Ratio of Assets to Operating Revenue (A/S) = 12,000 / 16,000 = 0.75…… 2. Forecast 2017 Operating Revenue and Profit**:Forecasted 2017 Operating Revenue (S_{2017}):** $S_{2017} = S_{2016} \times (1 + \text{Operating Revenue Growth Rate}) = 16,000 \times (1 + 0.10) = 17,600$, ……3. Forecast 2017 Assets and Spontaneous Liabilities**:Forecasted 2017 Total Assets (A_{2017}):**.$A_{2017} = (A/S) \times S_{2017} = 0.75 \times 17,600 = 13,200$, Forecasted 2017 Spontaneous Liabilities (L^*_{2017}):** $L^*_{2017} = (L^*/S) \times S_{2017} = 0.1875 \times 17,600 = 3,300$. 4. Calculate 2017 Total Financing Needs**Total financing needs typically refer to the increase in total assets required to support sales growth. Increase in Assets (ΔA) = $A_{2017} - A_{2016} = 13,200 - 12,000 = 1,200$Total Financing Needs = 1,200 Ten Thousand Yuan**. 5. Calculate 2017 External Financing Needs (AFN - Additional Funds Needed)**External financing needs refer to the funds that must be raised externally after considering the increase in spontaneous liabilities and the increase in retained earnings.Increase in Spontaneous Liabilities (ΔL^*) = $L^*_{2017} - L^*_{2016} = 3,300 - 3,000 = 300$. External Financing Needs (AFN) = Increase in Assets (ΔA) - Increase in Spontaneous Liabilities (ΔL^*) - Forecasted Increase in Retained Earnings (ΔRE_{2017}) AFN = 1,200 - 300 - 726 = 174. External Financing Needs = 174 Ten Thousand Yuan**.Since the company has no plans for external equity financing, the external financing needs of 1.74 million yuan will be met by increasing debt (such as increasing long-term loans). Final Conclusion:**:According to the percentage of sales method:The company's total financing needs for 2017 are \mathbf{1,200 Ten Thousand Yuan}.The company's external financing needs for 2017 are \mathbf{174 Ten Thousand Yuan}

错误原因：模型在推理过程中未能充分理解销售百分比法的完整应用逻辑，导致对“总融资需求”和“外部融资需求”的定义和计算公式应用不当。这道题是多轮计算，第一轮的错误导致后面计算全部出错。这反映模型在处理多步骤的财务预测任务时，未能正确理解并应用业务流程中的关键逻辑和计算步骤，导致最终结果出现偏差。

Reason for Error：The model failed to fully understand the full logic of the percentage of sales method in its reasoning, resulting in an inappropriate application of the definitions and formulas for Total Financing Requirement and External Financing Requirement. This question is a multi-round calculation, and an error in the first round leads to errors in all subsequent calculations. This reflects the model's failure to properly understand and apply the key logic and calculation steps in the business process when dealing with a multi-step financial forecasting task, resulting in a biased final result.

Figure 28: An example of Perceived barriers to financial business processes. This is a multi-round Q&A, and since the big model has already answered the first round of questions incorrectly, resulting in incorrect answers to the subsequent questions based on this incorrect answer, only the first round of Q&A is shown here as an example. Here, the model has successfully simulated the finance staff to identify the subjects and corresponding data to be calculated, but the model failed to fully understand the complete application logic of the Percentage of Sales method in the reasoning process, resulting in the improper application of the definitions and formulas of Total Financing Requirement and External Financing Requirement. This leads to improper application of the definitions and formulas of “total financing needs” and “external financing needs”. This error reflects the model’s inadequate understanding of the dependencies between the steps and the logic of calculation when dealing with complex financial business processes.

(continued)Prompt Template for Constructing Four-Option Multiple-Choice Questions Based on Line Charts (English Versions)

You are a financial analyst. Based on the provided line chart, generate three four-option multiple-choice questions.
The questions should be grounded in the data trends, key points, or features shown in the chart.
Requirements:

1. Each question must be clearly stated, and the options should be meaningfully differentiated.
2. Options A, B, C, and D for each question should represent distinct possibilities and avoid being overly obvious or simplistic.
3. The answer to each question must be one of A, B, C, or D.
4. Each question should be no fewer than 10 Chinese characters in length.
5. All three questions must be entirely different, and each should be labeled as 1, 2, and 3.
6. **Only output the content in JSON format. Do not include any additional descriptive text.**

You may refer to the following examples:
{random_few_shots}
Output format:

```
[
  {"id": "1", "q": "Question 1", "A": "Option A", "B": "Option B", "C": "Option C", "Answer": "Correct Answer"},
  {"id": "2", "q": "Question 2", "A": "Option A", "B": "Option B", "C": "Option C", "Answer": "Correct Answer"},
  {"id": "3", "q": "Question 3", "A": "Option A", "B": "Option B", "C": "Option C", "Answer": "Correct Answer"}
]
```

Table 10: Prompt Template for Constructing Counterfactual Inference Questions Based on Histograms (Chinese Versions)

你是一名资深数据分析师，请根据提供的直方图，生成三道反事实推断的单选题，一般形式就是如果某个不是事实的事情发生，会有什么结果。
题目应基于你对直方图的理解，参考直方图的分布、峰值、偏态、异常值等特征。要求：
1. 每道问题必须使用中文语言，清晰明确，选项应具有区分度。
2. 每道题的选项 A、B、C、D 应涵盖不同的可能性，避免过于简单或明显。
3. 每道题的答案必须是 A、B、C 或 D 中的一个，不能是其他内容。
4. 三道题目必须完全不同，且每道题需要编号为 1、2、3。
5. **只输出 JSON 格式的内容，不要包含任何额外的描述性文本。**
你可以参考的示例：
{random_few_shots}
输出格式为：

```
[
  {"id": "1", "q": "问题1", "A": "选项A", "B": "选项B", "C": "选项C", "D": "选项D", "Answer": "A/B/C/D"},
  {"id": "2", "q": "问题1", "A": "选项A", "B": "选项B", "C": "选项C", "D": "选项D", "Answer": "A/B/C/D"},
  {"id": "3", "q": "问题1", "A": "选项A", "B": "选项B", "C": "选项C", "D": "选项D", "Answer": "A/B/C/D"}
]
```

(a) Chinese Version

(continued)Prompt Template for Constructing Counterfactual Inference Questions Based on Histograms (English Versions)

You are a senior data analyst. Based on the provided histogram, generate three single-choice counterfactual inference questions. These questions should generally take the form: if something that did not actually happen were to occur, what would be the result?
Questions should be based on your understanding of the histogram, referencing features such as distribution, peaks, skewness, and outliers. Requirements:

1. Each question must be written in Chinese, clearly stated, and the options should be distinguishable.

2. Each question should have four options A, B, C, and D representing different possibilities. Avoid options that are too simple or obvious.

3. The answer to each question must be one of A, B, C, or D and nothing else.

4. The three questions must be completely different and should be numbered as 1, 2, and 3.

5. **Only output the content in JSON format. Do not include any additional descriptive text.**

You may refer to the following examples:
{random_few_shots}
Output format:

```
[
  {"id": "1", "q": "Question 1", "A": "Option A", "B": "Option B", "C": "Option C", "D": "Option D", "Answer": "A/B/C/D"},
  {"id": "2", "q": "Question 2", "A": "Option A", "B": "Option B", "C": "Option C", "D": "Option D", "Answer": "A/B/C/D"},
  {"id": "3", "q": "Question 3", "A": "Option A", "B": "Option B", "C": "Option C", "D": "Option D", "Answer": "A/B/C/D"}
]
```

(b) English Version

Table 11: Prompt Template for Constructing Multi-turn Dialogue Tasks Based on Candlestick Charts (Chinese Versions)

你是一名专业的金融分析师，擅长分析K线图。
现在给你的是不同的股票的几张含有不同参数的K线图，请根据提供的多张K线图生成三道专业金融题目，要求：
1. 题目类型包括趋势分析、数据比较、计算题等，尽量是客观题，尽量丰富题型，保证正确答案是客观的。
2. 每道题四个选项并标注正确答案，正确的答案只能有一个，即单选题。
3. 必须基于所有图片内容，即每个题目都必须用到图对的所有图片的内容，每个题目前面可以用一两句话描述下图片和题目，最后再次重申：使用到所有图片的内容，题目尽量长一点。
你可以参考的示例:
{random_few_shots}
输出格式（每个问题一个 JSON 对象）：

```
[
  {
    "q": "问题描述",
    "A": "选项A",
    "B": "选项B",
    "C": "选项C",
    "D": "选项D",
    "Answer": "正确答案"},
  },
  {
    "q": "问题描述",
    "A": "选项A",
    "B": "选项B",
    "C": "选项C",
    "D": "选项D",
    "Answer": "正确答案"},
  },
  {
    "q": "问题描述",
    "A": "选项A",
    "B": "选项B",
    "C": "选项C",
    "D": "选项D",
    "Answer": "正确答案"},
  }
]
```

(continued)Prompt Template for Constructing Multi-turn Dialogue Tasks Based on Candlestick Charts (English Versions)

You are a professional financial analyst skilled in analyzing candlestick charts.
You are provided with several candlestick charts of different stocks with varying parameters. Please generate three professional financial questions based on these charts, with the following requirements:

1. Question types should include trend analysis, data comparison, calculation problems, etc. Prioritize objective questions with diverse formats, ensuring answers are fact-based.

2. Each question must have four options with one clearly marked correct answer (single-choice format).

3. All questions must incorporate content from every provided image. Each question may be preceded by 1-2 sentences describing the relevant chart elements. Remember: every question must utilize all images' content, and questions should be sufficiently detailed.

Reference examples:
{random_few_shots}
Output format (one JSON object per question):

```
[
    {
      "q": "Question description",
      "A": "Option A",
      "B": "Option B",
      "C": "Option C",
      "D": "Option D",
      "Answer": "Correct option"},
    },
    {
      "q": "Question description",
      "A": "Option A",
      "B": "Option B",
      "C": "Option C",
      "D": "Option D",
      "Answer": "Correct option"},
    },
    {
      "q": "Question description",
      "A": "Option A",
      "B": "Option B",
      "C": "Option C",
      "D": "Option D",
      "Answer": "Correct option"},
    }
]
```

Table 12: Prompt Template for Constructing True/False Judgment Tasks Based on Pie Charts (Chinese and English Versions)

你作为一名专业的金融分析师，擅长分析饼图，请帮我根据我给你提供的饼图及图片的caption，给我出3道专业且带有难度的金融判断题目。
要求如下：
1. 只给我输出我对应的格式信息，不要给我其他信息。
2. 问题尽可能的多样化、复杂化，所有的问题请基于我的图片内容。
3. 你最后的正确答案应该同时符合你的问题逻辑和图片内容。
4. 请确保问题基于图片内容生成。
5. **只输出 JSON 格式的内容，不要包含任何额外的描述性文本。**
你可以参考的示例:
{random_few_shots}
输出格式为：

```
[
  {{ "Question1":"","Answer":"True" }},
  {{ "Question2":"","Answer":"True" }},
  {{ "Question3":"","Answer":"True" }}
]
```

(a) Chinese Version

As a professional financial analyst with expertise in interpreting pie charts, please generate 3 professional and challenging true/false financial questions based on the pie chart and its caption that I provide.
Requirements:
1. Only return the formatted information I requested. Do not include any additional content.
2. The questions should be as diverse and complex as possible, and must be based entirely on the content of the image.
3. The correct answers must logically align with both the question structure and the image content.
4. Please ensure that the questions are generated based on the image.
5. **Only output content in JSON format. Do not include any descriptive or explanatory text.**
You may refer to the following examples:
{random_few_shots}
Output format:

```
[
  {{ "Question1":"","Answer":"True" }},
  {{ "Question2":"","Answer":"True" }},
  {{ "Question3":"","Answer":"True" }}
]
```

(b) English Version

Table 13: Prompt Template for Verifying the Content Quality of Images (Chinese and English Versions)

```
你是一个专业的图像分析助手。请根据以下标准筛选高质量的折线图：
  1. **数据多样性**：折线图中应展示至少2条不同的折线，每条折线代表一个独立的数据类别或指标，且数据变化趋势应具有一定的多样性（如上升、下降、波动等）。
  2. **数据清晰可辨**：折线图中的数据点、坐标轴、图例等应清晰可辨，避免模糊或难以解读的图表。
  3. **具备问答意义**：图表中的数据应能够产生有效的问答对，且问题应具有一定的计算意义或挑战性。确保问答可以基于这些数据进行推理、计算或者对比。
  4. **去除低质量图表**：如果折线图中只展示了一条折线，或者数据变化趋势过于简单（如单调上升或下降），则不符合要求。
"""
   messages=[{"role": "user", "content": [{"type": "image_url", "image_url": {"url": f"{img_base}"}}, {"type": "text", "text": f """
以下是针对这张图片生成的三个问题：{questions_text}
你是一名专业的金融分析师，擅长分析折线图。请根据问题的专业性和难度，从中选择一个最好的问题，返回问题的索引（从1开始）。
请只返回一个数字，例如：1、2 或 3。
```

(a) Chinese Version

```
You are a professional image analysis assistant. Please filter high-quality line charts based on the following criteria:
  1. **Data Diversity**: The line chart should display at least two distinct lines, each representing an independent data category or metric. The trends should exhibit diversity (e.g., increase, decrease, fluctuation).
  2. **Clarity of Data**: Data points, axes, legends, and other elements in the chart should be clearly distinguishable, avoiding any blurry or unreadable visuals.
  3. **Question-Answer Relevance**: The chart should enable the generation of meaningful QA pairs. The questions should involve some degree of calculation or reasoning. Ensure that the data in the chart supports logical inference, computation, or comparison.
  4. **Exclude Low-Quality Charts**: If the chart contains only one line, or if the data trend is overly simplistic (e.g., strictly increasing or decreasing), it should be excluded.
"""
   messages=[{"role": "user", "content": [{"type": "image_url", "image_url": {"url": f"{img_base}"}}, {"type": "text", "text": f """
Here are three questions generated based on this image: {questions_text}
You are a professional financial analyst skilled in interpreting line charts. Based on the professionalism and difficulty of the questions, select the best one and return its index (starting from 1).
Only return a single number, such as: 1, 2, or 3.
```

(b) English Version

Table 14: Prompt Template for Validating the Consistency and Correctness of QA Pairs (Chinese and English Versions)

请验证以下问答对的质量：
[问题] {question.get('query', '')}
[参考答案] {question.get('answer', '')}
验证标准：
1. 答案准确性（基于图表数据）
2. 问题复杂度（需两步以上推理）
3. 问题客观程度，需是客观题或者计算题
4. 选项合理性（如为选择题）
验证结论格式：通过/不通过|理由

(a) Chinese Version

Please validate the quality of the following QA pair:
[Question] {question.get('query', '')}
[Reference Answer] {question.get('answer', '')}
Validation Criteria:
1. Answer accuracy (based on the chart data)
2. Question complexity (requires more than two steps of reasoning)
3. Objectivity of the question (must be objective or computational)
4. Option quality (if multiple-choice)
Validation Output Format: Pass / Fail | Reason

(b) English Version

Table 15: Prompt Template for Classifying Scenarios in the Financial Knowledge and Data Analysis Category (Chinese Version)

请根据以下内容为问题进行分类：{combined_text}
请将问题分类到以下7个金融场景之一：
1. 股票K线解读，智能验印，金融信息识别，财务数据统计，金融实体关系解读，金融市场情绪洞察，金融情景分析
其中
2. 股票K线解读：该场景通过解读股票K线图及相关技术指标（如MACD、RSI、成交量等），分析股价的历史走势、当前状态及未来趋势。
3. 智能验印：该场景需要对金融或行政文档中的印章进行识别、验证与比对，以判断其真伪、归属及合规性。
4. 金融信息识别：该场景关注金融、经济、投资领域中金融信息的识别和解读任务，识别其所表达的金融含义。
财务数据统计：该场景关注对具体财务或经济数据的整理、趋势分析和对比评估，例如地方政府债券发行量、资本项目差额、财政收支变动等。
5. 金融实体关系解读：该场景聚焦于经济主体（如公司、政府、部门）之间的关系分析和经济影响链条解读，例如“财政扩张如何影响居民消费”或“资本流入对汇率的影响”。
6. 金融市场情绪洞察：该场景侧重从投资者行为、舆情或市场表现中提取市场情绪趋势，例如通过新闻、评论、价格行为等数据推测市场预期。
7. 金融情景分析：该场景包含假设性问题和反事实推理，例如“如果2022年债券发行没有增加，将可能发生什么？”此类问题需要基于对金融机制的理解推测可能后果。
请仅回答类别名称，不要解释。

(continued) Prompt Template for Classifying Scenarios in the Financial Knowledge and Data Analysis Category (English Version)

Please classify the question based on the following content: {combined_text}
Assign the question to one of the following seven financial scenarios:

1. Candlestick Chart Analysis, Intelligent Seal Recognition, Financial Information Extraction, Statistical Analysis of Financial Data, Interpretation of Financial Entity Relationships, Financial Market Sentiment Analysis, Financial Scenario Analysis
Descriptions:

2. Candlestick Chart Analysis: This scenario involves interpreting candlestick charts and related technical indicators (e.g., MACD, RSI, trading volume) to analyze historical price trends, current states, and potential future movements.

3. Intelligent Seal Recognition: This scenario requires identifying, verifying, and matching seals in financial or administrative documents to determine their authenticity, origin, and compliance.

4. Financial Information Extraction: This scenario focuses on identifying and interpreting financial concepts and information in the fields of finance, economics, and investment.
Statistical Analysis of Financial Data: This scenario focuses on organizing, analyzing trends, and comparing financial or economic data—such as bond issuance volume by local governments, capital account balance, and fiscal revenue/expenditure changes.

5. Interpretation of Financial Entity Relationships: This scenario centers on analyzing the relationships among economic entities (e.g., firms, governments, departments) and tracing economic impact chains, such as "How does fiscal expansion affect household consumption?" or "What is the impact of capital inflows on exchange rates?"

6. Financial Market Sentiment Analysis: This scenario emphasizes extracting market sentiment from investor behavior, public opinion, or market movements—e.g., inferring expectations through news, comments, or price behavior.

7. Financial Scenario Analysis: This scenario involves hypothetical and counterfactual reasoning, such as "What would have happened if bond issuance had not increased in 2022?" These tasks require understanding financial mechanisms to infer potential outcomes.

Please return only the scenario category name. Do not include any explanations.

Table 16: Prompt Template for Classifying Scenarios in the Financial Analysis and Decision Support Category (Chinese and English Versions)

【金融题目智能分类任务】
请基于以下信息，选择最合适的分类（仅返回类别名称）：
【背景上下文】
{row['background']}
{problem_presentation}
【分类标准】（四选一）：
1. 产业分析推断 - 行业趋势、政策影响类问题
2. 财务指标分析 - 涉及财务比率、指标计算
3. 金融报表分析 - 资产负债表/利润表等解读
4. 投资分析 - 综合投资决策评估
判断要求：
1. 单轮问题直接根据问题内容分类
2. 多轮问题需综合分析各轮次的关联性
3. 背景信息可帮助理解问题场景
4. 只需返回最匹配的中文类别名称

(a) Chinese Version

[Financial Question Scenario Classification Task]
Based on the following information, select the most appropriate category (return only the category name):
[Background Context]
{row['background']}
{problem_presentation}
[Classification Criteria] (Choose one):
1. Industry Analysis and Inference – questions related to industry trends or policy impacts
2. Financial Performance Indicator Analysis – questions involving financial ratios or indicator calculations
3. Financial Statement Analysis – interpretation of balance sheets, income statements, etc.
4. Investment Analysis – comprehensive evaluation of investment decisions
Classification Guidelines:
1. For single-turn questions, classify based on the question content alone
2. For multi-turn questions, consider the relationship between all turns
3. Background context may assist in understanding the question
4. Only return the best-matching category name in Chinese

(b) English Version

Table 17: Prompt Template for Classifying Scenarios in the Financial Risk Control and Asset Optimization Category (Chinese Version)

你现在是一位在金融领域的从业专家，请判断下列问题：
题目
{question}
{options_text}
属于哪个【金融业务场景】（请从以下给定场景中严格选择一个）。
金融业务场景：
1.资产配置分析—涉及投资组合结构、资产比例调整、风险收益平衡、企业股权架构设计等等。
2.金融策略优化—聚焦企业财务策略调整（如定价/成本/营销策略）及其对盈利的影响等等。
3.金融数据推演与解释—依赖数据计算、财务指标预测、数据间逻辑关系推导等等。
4.金融风险与政策解读—汇率/利率波动风险识别、政策对金融市场（如股市、债市）或企业的影响分析、市场风险信号判断（如资产价格大幅波动）、政策导向解读（如货币政策调整对信贷的影响）等等。
补充说明：
以上4个场景仅做了简单的描述，但这些描述不足以囊括该场景的所有情况，因此，如果某一问题并不符合上述描述，此时你可以基于自身对四个场景的理解，自行判断该问题应该属于哪一类金融场景。
冲突处理：
- 若同时涉及数据+策略，优先选择金融策略优化
- 若同时涉及数据+资产配置，优先选择资产配置分析
- 若同时涉及数据+风险或政策，优先选择金融风险与政策解读
输出格式：
场景分类：XXX
判定依据：YYY
禁止行为：
添加额外解释
脱离给定场景列表分类
修改预设输出格式

(continued) Prompt Template for Classifying Scenarios in the Financial Risk Control and Asset Optimization Category (English Version)

You are now a financial domain expert. Please classify the following question:
Question:
{question}
{options_text}
Determine which of the following **financial business scenarios** it belongs to (please strictly choose one from the list below).
Financial Business Scenarios:

1. Asset Allocation Analysis — related to portfolio structure, asset proportion adjustment, risk-return balancing, equity structure design, etc.

2. Financial Strategy Optimization — focuses on corporate financial strategy adjustments (e.g., pricing/cost/marketing strategies) and their impact on profitability.

3. Financial Data Reasoning and Interpretation — relies on numerical computation, financial indicator forecasting, inference of logical relationships between data, etc.

4. Financial Risk and Policy Analysis — includes identification of risks from exchange rate/interest rate fluctuations, analysis of policy impacts on financial markets (e.g., stock/bond markets) or firms, judgment of market risk signals (e.g., significant asset price volatility), and interpretation of policy directions (e.g., how monetary policy adjustments affect credit).

Supplementary Note:
The above descriptions are simplified and do not fully capture all cases under each scenario. If a question does not clearly match any description, you may rely on your own understanding of the four scenarios to make a reasoned judgment.
Conflict Resolution:

- If the question involves both data and strategy, prioritize Financial Strategy Optimization.
- If it involves both data and asset allocation, prioritize Asset Allocation Analysis.
- If it involves both data and risk or policy, prioritize Financial Risk and Policy Analysis.

Output Format:
Scenario Classification: XXX
Justification: YYY
Prohibited Actions:
Adding extra explanation
Classifying outside the given list
Modifying the preset output format

Table 18: Chain-of-Thought(CoT) Results. The higher the value in the table, the higher the accuracy of the surface model. The Financial Analysis and Decision Support assesses models with Financial Data Statistics (FDS), Candlestick Chart Analysis (CCA), Financial Indicator Assessment (FIA), Financial Entity Relationships Interpretation (FERI), Stock Selection Strategies Backtesting (SSSB), Financial Information Extraction (FIE), and Financial Seal Recognition (FSR). The Financial Analysis and Decision Support tests with Financial Scenario Analysis (FSA), Industry Analysis and Inference (IAI), Investment Analysis (IA), and Financial Market Sentiment Analysis (FMSA). The Financial Risk Control and Asset Optimization evaluates Financial Strategy Optimization (FSO), Financial Risk and Policy Analysis (FRPA), Financial Data Reasoning and Interpretation (FDRI), and Asset Allocation Analysis (AAA), concluding with the calculation of the Weighted Average (WA) score for each model. The table also indicates operational constraints encountered by certain models in multi-image tasks, such as Multi-image Limit and Context Window Limit.

Model	Size	Limit	Financial Knowledge and Data Analysis							Financial Analysis and Decision Support				Financial Risk Control and Asset Optimization				WA
			FDS	CCA	FIA	FERI	SSSB	FIE	FSR	FSA	IAI	IA	FMSA	FSO	FRPA	FDRI	AAA	
InternVL3-78B	78B	/	74.2	84.8	68.4	88.2	77.9	90.6	87.4	57.6	82.8	76.2	75.2	42.8	52.1	42.2	68.7	71.3
Qwen-VL-max	Unknown	/	72.4	81.2	68.8	80.1	80.5	86.8	80.8	55.9	82.7	78.8	78.7	48.1	50.2	44.0	73.5	70.8
InternVL2.5-78B	78B	/	72.0	84.9	66.7	85.4	77.4	89.2	82.8	54.7	81.9	76.9	80.0	44.3	48.5	43.2	72.5	70.7
Doubao-1.5-vision-pro-32k	Unknown	/	75.9	81.1	72.2	84.5	76.8	91.5	77.8	56.0	82.7	79.1	80.7	45.7	43.1	44.1	66.3	70.5
Step-1o-vision-32k	Unknown	/	73.0	78.9	67.1	84.7	76.4	87.4	97.0	58.8	78.8	77.3	73.6	51.9	41.3	42.4	68.3	70.5
Qwen-VL-max-latest	Unknown	/	72.7	80.3	69.2	80.6	79.4	85.9	80.8	54.8	83.3	78.7	78.7	43.4	45.3	44.3	72.7	70.0
Qwen2.5-VL-72B	72B	/	69.3	80.6	67.2	85.5	78.7	86.4	66.7	51.8	80.6	80.1	74.4	43.5	44.2	46.2	71.1	68.4
Gemini-2.5-pro-exp-03-25	Unknown	/	70.6	67.7	64.9	65.7	75.1	91.0	88.9	51.4	80.0	71.9	76.5	47.2	50.4	43.3	75.2	68.0
Qwen2.5-VL-7B	7B	/	72.2	81.0	66.5	81.7	75.9	87.3	84.9	51.0	61.2	74.7	67.8	35.6	42.4	37.7	70.7	66.0
Claude-3-7-Sonnet-20250219	Unknown	/	66.9	74.2	71.1	80.5	75.4	89.4	43.4	44.9	81.6	78.1	73.6	47.1	46.7	41.6	73.7	65.9
InternVL3-8B	8B	/	68.1	67.1	63.3	77.7	70.3	84.3	89.9	57.7	77.5	75.2	67.8	35.6	43.3	40.0	64.9	65.5
GPT-4o-all	Unknown	/	68.6	76.7	64.7	77.1	72.8	84.4	79.8	51.8	79.0	75.8	71.6	34.9	43.6	37.5	64.4	65.5
Moonshot-V1-32k-vision-preview	Unknown	/	70.2	73.5	62.4	75.3	74.8	86.3	60.6	50.9	80.1	75.5	63.2	37.7	45.6	40.4	67.6	64.3
MiniCPM-V-2.6	8B	/	60.6	80.1	51.2	74.7	67.8	75.9	85.9	47.5	73.6	68.7	68.4	23.6	34.1	27.6	56.0	59.7
Qwen2.5-VL-3B	3B	/	61.1	79.6	65.6	72.3	72.5	83.3	62.3	48.2	75.0	74.3	64.6	22.1	28.7	30.5	53.9	59.6
Llama-3.2-11B-Vision-Instruct	11B	/	59.1	64.6	52.8	71.1	55.5	77.4	60.0	65.6	70.2	65.5	65.8	38.7	34.8	26.5	45.2	56.8
Molmo-7B-D-0924	7B	/	55.9	72.1	45.6	65.3	54.2	59.2	22.2	42.3	60.3	59.6	57.4	21.0	28.2	23.3	28.6	46.3
GLM-4v-Plus-20250111	Unknown	Multi-image Limit	73.6	81.9	71.5	83.8	78.2	88.9	81.8	54.0	81.0	76.1	75.5	40.6	44.5	40.7	65.7	69.2
LLaVA-NEXT-34B	34B	Context Window Limit	52.8	51.0	56.9	50.7	59.0	53.0	20.0	40.2	64.4	61.5	59.5	24.0	24.1	18.6	28.6	44.3
LLaVA-v1.6-Mistral-7B	7B	Context Window Limit	52.7	68.7	47.3	55.6	45.8	51.2	30.3	27.8	54.0	54.5	62.3	21.7	10.2	11.5	22.5	41.1
LLaVA-NEXT-13B	13B	Context Window Limit	46.5	63.2	38.9	60.3	51.4	45.6	24.2	27.6	57.9	56.3	51.0	11.3	15.9	12.0	29.7	39.4

Table 19: Comparison of CoT and Zero-shot Results. The table shows the performance of different models using Chain-of-Thought (CoT) and Zero-shot approaches, along with the difference between the two methods (CoT - Zero-shot). A positive difference indicates better performance with CoT, while a negative difference indicates better performance with Zero-shot.

Model	Size	Limit	CoT	Zero-shot	Difference
Qwen-VL-max-latest	Unknown	/	70.0	73.8	-3.8
Qwen-VL-max	Unknown	/	70.8	76.3	-5.5
Doubao-1.5-vision-pro-32k	Unknown	/	70.5	71.7	-1.2
Gemini-2.5-pro-exp-03-25	Unknown	/	68.0	64.7	3.2
Moonshot-V1-32k-vision-preview	Unknown	/	64.3	68.3	-4.0
GPT-4o-2024-11-20	Unknown	/	65.5	68.5	-3.0
Claude-3-7-Sonnet-20250219	Unknown	/	65.9	62.9	2.9
Step-1o-vision-32k	Unknown	/	70.5	68.4	2.1
InternVL3-78B	78B	/	71.3	72.5	-1.3
InternVL3-8B	8B	/	65.5	65.4	0.2
MiniCPM-V-2.6	8B	/	59.7	60.1	-0.4
InternVL2.5-78B	78B	/	70.7	71.5	-0.8
Qwen2.5-VL-7B	7B	/	66.0	65.4	0.7
Qwen2.5-VL-3B	3B	/	59.6	62.4	-2.8
Qwen2.5-VL-72B	72B	/	68.4	71.0	-2.6
Llama-3.2-11B-Vision-Instruct	11B	/	56.8	50.9	5.9
Molmo-7B-D-0924	7B	/	46.3	49.8	-3.5
GLM-4v-Plus-20250111	Unknown	Multi-image Limit	69.2	72.0	-2.9
LLaVA-v1.6-Mistral-7B	7B	Context Window Limit	41.1	47.8	-6.7
LLaVA-NEXT-13B	13B	Context Window Limit	39.4	43.0	-3.6
LLaVA-NEXT-34B	34B	Context Window Limit	44.3	54.7	-10.5

2024-08-04--半导体-海光信息-稀缺CPU+AI双料厂商，深算系列有望迎来质变突破_page16_图33海光7285与Intel同期铂金系列产品性能对比

厂商	产品名称	发布时间	4路测试结果		双路测试结果		性能差异（Intel 数据/海光数据-1）	
			Speccpu_INT	Speccpu_FP	Speccpu_INT	Speccpu_FP	Speccpu_INT	Speccpu_FP
intel	Intel8380HL（铂金）	2020 年第二季度	784	657	392	329	12.64%	6.66%
	Intel8380H（铂金）	2020 年第二季度	784	653	392	327	12.64%	6.01%
	Intel8376HL（铂金）	2020 年第二季度	765	641	383	321	9.91%	4.06%
	Intel8376H（铂金）	2020 年第二季度	756	643	378	322	8.62%	4.38%
HYGON	海光 7285	2020 年第一季度	-	-	348	308	-	-
intel	Intel8360HL（铂金）	2020 年第三季度	690	599	345	300	-0.86%	-2.76%
	Intel8360H（铂金）	2020 年第三季度	688	597	344	299	-1.15%	-3.08%

问题：根据图中数据，Intel8380HL（铂金）与Intel8376HL（铂金）在双路测试中的Speccpu_INT和Speccpu_FP性能差异分别是多少？

Question: According to the data in the figure, what are the performance differences in Speccpu_INT and Speccpu_FP for Intel 8380HL (Platinum) and Intel 8376HL (Platinum) in dual-socket tests, respectively?

Question: Which shareholders hold less than 20% of the shares in the shareholding structure of Vimax Industrial (Guangdong) Co?

A. -1.15%; -3.08% B. -0.86%; -2.76% C. -1.15%; -2.76% D. -0.86%; -3.08%

Model Answer: To answer this question, we need to calculate the performance differences in Speccpu_INT and Speccpu_FP for Intel 8380HL (Platinum) and Intel 8376HL (Platinum) in dual-socket tests based on the data in the chart.\n\n### Calculation Steps:\n\n1. Speccpu_INT Performance Difference:\n - Speccpu_INT for Intel 8380HL (Platinum): 392\n - Speccpu_INT for Intel 8376HL (Platinum): 383\n - Difference calculation: (383 - 392) / 392 * 100% = -2.298% ≈ -2.30%\n\n2. Speccpu_FP Performance Difference:\n - Speccpu_FP for Intel 8380HL (Platinum): 329\n - Speccpu_FP for Intel 8376HL (Platinum): 321\n - Difference calculation: (321 - 329) / 329 * 100% = -2.432% ≈ -2.43%\n\n### Option Analysis:\n- A. -1.15%; -3.08%\n- B. -0.86%; -2.76%\n- C. -1.15%; -2.76%\n- D. -0.86%; -3.08%\n\nAccording to the calculation results, none of the options completely match the calculated difference values. Therefore, based on the chart data and calculation results, none of the options are entirely correct.\n\nFinal Conclusion: No correct option.

答案：A;B

Answer: A;B

Figure 29: This is an example where the Qwen2.5-VL-72B model with incorrect analysis and an incorrect answer. This is a question involving the calculation of performance differences between processor models. In response to the question asking for the performance differences in Speccpu_INT and Speccpu_FP between Intel 8380HL (Platinum) and Intel 8376HL (Platinum) in dual-socket tests, the model attempted to calculate the differences independently but used an improper formula. It neither aligned with the predefined "performance difference" data in the table nor matched the standard comparison logic implied by the options. Furthermore, the model failed to recognize that the question intended to directly reference the existing results in the "performance difference" column of the table rather than requiring recalculation, and it neglected to cross-verify its calculation results with the provided options. This error reflects the model's inadequate understanding of question intent, improper application of calculation methods, and lack of awareness regarding result verification when handling data-driven comparison tasks.

Table 20: List of financial information platforms and their URLs

Institution Name	URL
EastMoney	https://data.eastmoney.com/report/
CNINFO	https://www.cninfo.com.cn/
10jqka	https://data.10jqka.com.cn/
SinaFinance	https://vip.stock.finance.sina.com.cn/
Shanghai Stock Exchange	https://www.sse.com.cn/
Shenzhen Stock Exchange	https://www.szse.cn/
Ducaibao	https://www.ducaibao.com.cn/
Hibor	https://www.hibor.com.cn/
Jianwei Data	https://www.jianweidata.com/
KPMG	https://kpmg.com/cn/zh/home.html
Robo DataYes	https://robo.datayes.com/

附：综合所得个人所得税税率表（部分）

级数	全年应纳税所得额	税率（%）	速算扣除数
1	不超过 36000 元的部分	3	0
2	超过 36000 元至 144000 元的部分	10	2520
3	超过 144000 元至 300000 元的部分	20	16920

附：按月换算后的综合所得个人所得税税率表（部分）

级数	月应纳税所得额	税率（%）	速算扣除数
1	不超过 3000 元的部分	3	0
2	超过 3000 元至 12000 元的部分	10	210
3	超过 12000 元至 25000 元的部分	20	1410

题目及背景信息：[补充信息] 居民个人王某，其独生子年满 5 周岁并正在接受学前教育，父母均已年满 65 周岁。2022年王某的收入情况如下：（1）每月工资 15000 元，含按国家标准缴纳的“三险一金”3500 元。（2）4 月委托某拍卖公司拍卖收藏的祖传字画一幅，最终拍卖成交价为 50000 元，不能提供该字画原值凭证。（3）12 月从单位领取年终奖 72000 元，选择单独计税。（其他相关资料：王某为独生子，相关专项附加扣除均由王某 100%扣除，王某已向单位报送其专项附加扣除信息；不考虑字画拍卖过程中缴纳的税费。）计算王某拍卖所得应缴纳的个人所得税。

Question and Background Information: [Supplementary Information] Resident individual Wang, whose only child is 5 years old and receiving pre-school education, and whose parents are both over 65 years old. Wang's income in 2022 is as follows: (1) Monthly salary of 15,000 yuan, including "three insurances and one housing fund" paid according to national standards, totaling 3,500 yuan. (2) In April, a family heirloom calligraphy and painting was consigned to an auction company and sold for 50,000 yuan; no original value certificate for the calligraphy and painting could be provided. (3) In December, an annual bonus of 72,000 yuan was received from the employer, choosing separate taxation. (Other relevant information: Wang is an only child, all relevant special additional deductions are 100% deducted by Wang, and Wang has submitted his special additional deduction information to the employer; tax fees paid during the calligraphy and painting auction process are not considered.) Calculate the individual income tax payable on Wang's auction income.

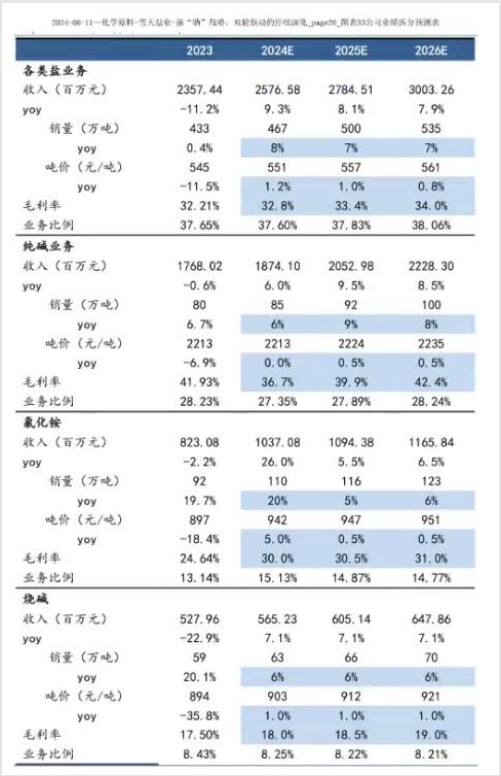
模型回答：要计算王某拍卖所得应缴纳的个人所得税，我们需要按照以下步骤进行分析和计算： 1. 确定拍卖所得的应纳税所得额\n根据税法规定，拍卖财产所得的应纳税所得额为拍卖收入减去财产原值和合理费用。由于王某不能提供该字画的原值凭证，因此无法确定财产原值。在这种情况下，根据税法规定，应纳税所得额为拍卖收入的全额。拍卖收入：50000元。2. 确定适用税率和速算扣除数根据《中华人民共和国个人所得税法》及其实施条例，财产转让所得适用20%的税率。因此，王某拍卖所得应适用20%的税率。3. 计算应纳税额 应纳税额 = 应纳税所得额 × 税率- 应纳税所得额：50000元 税率：20%应纳税额 = 50000元 × 20% = 10000元 最终结论\n王某拍卖所得应缴纳的个人所得税为 \\boxed{10000元}。

Model Response: To calculate the individual income tax payable on Wang's auction income, we need to analyze and calculate according to the following steps: 1. Determine the taxable income for auction proceeds\nAccording to tax law, the taxable income from property auction is the auction income minus the original value of the property and reasonable expenses. Since Wang cannot provide the original value certificate for the calligraphy and painting, the original value cannot be determined. In this case, according to tax law, the taxable income is the full amount of the auction income. Auction income: 50,000 yuan. 2. Determine the applicable tax rate and quick deduction amount According to the "Individual Income Tax Law of the People's Republic of China" and its implementing regulations, a 20% tax rate applies to income from property transfers. Therefore, a 20% tax rate should apply to Wang's auction income. 3. Calculate the tax payable Tax payable = Taxable income × Tax rate - Taxable income: 50,000 yuan Tax rate: 20% Tax payable = 50,000 yuan × 20% = 10,000 yuan Final Conclusion\nThe individual income tax payable on Wang's auction income is \boxed{10000 yuan}.

正确答案：1500

Correct Answer: 1500

Figure 30: This is an example where the Qwen2.5-VL-72B model with correct reasoning but incorrect answer: the model correctly structured the reasoning process and applied basic tax calculation steps (identifying taxable income as full auction proceeds due to lack of original value certificates and using the 20% tax rate) but failed to consider regional policies or industry practices where personal income tax on self-owned calligraphy and painting auctions without original value certificates is actually withheld at 3% of the transaction price, resulting in an incorrect final answer.



	2023	2024E	2025E	2026E
各类盐业务				
收入（百万元）	2357.44	2576.58	2784.51	3003.26
yoy	-11.2%	9.3%	8.1%	7.9%
销量（万吨）	433	467	500	535
yoy	0.4%	8%	7%	7%
吨价（元/吨）	545	551	557	561
yoy	-11.5%	1.2%	1.0%	0.8%
毛利率	32.21%	32.8%	33.4%	34.0%
业务比例	37.65%	37.60%	37.83%	38.06%
纯碱业务				
收入（百万元）	1768.02	1874.10	2052.98	2228.30
yoy	-0.6%	6.0%	9.5%	8.5%
销量（万吨）	80	85	92	100
yoy	6.7%	6%	9%	8%
吨价（元/吨）	2213	2213	2224	2235
yoy	-6.9%	0.0%	0.5%	0.5%
毛利率	41.93%	36.7%	39.9%	42.4%
业务比例	28.23%	27.35%	27.89%	28.24%
氯化铵				
收入（百万元）	823.08	1037.08	1094.38	1165.84
yoy	-2.2%	26.0%	5.5%	6.5%
销量（万吨）	92	110	116	123
yoy	19.7%	20%	5%	6%
吨价（元/吨）	897	942	947	951
yoy	-18.4%	5.0%	0.5%	0.5%
毛利率	24.64%	30.0%	30.5%	31.0%
业务比例	13.14%	15.13%	14.87%	14.77%
烧碱				
收入（百万元）	527.96	565.23	605.14	647.86
yoy	-22.9%	7.1%	7.1%	7.1%
销量（万吨）	59	63	66	70
yoy	20.1%	6%	6%	6%
吨价（元/吨）	894	903	912	921
yoy	-35.8%	1.0%	1.0%	1.0%
毛利率	17.50%	18.0%	18.5%	19.0%
业务比例	8.43%	8.25%	8.22%	8.21%

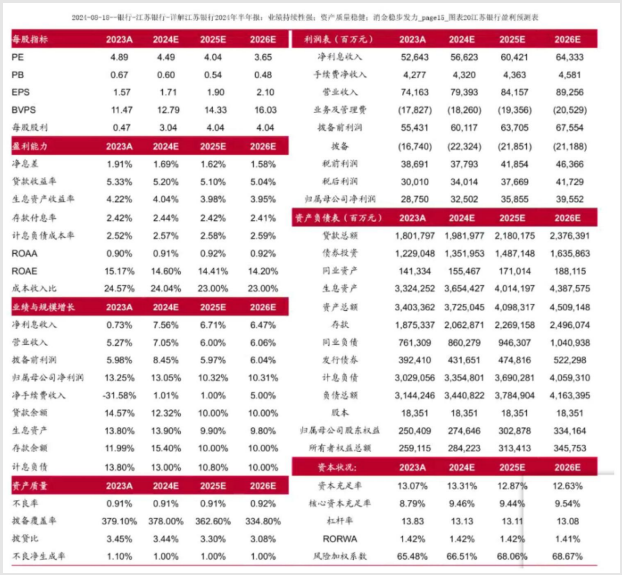
问题：假设2023年纯碱业务的毛利率为40%，且2024年收入增长率为10%，那么2024年的毛利润是多少？（单位：亿元）
Question: Assuming that the gross profit margin for the soda ash business in 2023 is 40% and the revenue growth rate in 2024 is 10%, what will the gross profit be in 2024?（Unit: 100 million RMB）

A. 754.4　　B. 713.7
C. 774.4　　D. 784.4

模型回答：我将逐步分析这个问题。\n\n首先，从题目中获取的关键信息：\n- 2023年纯碱业务的毛利率假设为40%\n- 2024年收入增长率假设为10%\n- 需要计算2024年的毛利润\n\n从图表中，我可以看到2023年纯碱业务的收入为1768.02百万元。\n\n计算步骤：\n1. 计算2024年纯碱业务的收入：\n 2024年收入 = 2023年收入 × (1 + 增长率)\n = 1768.02 × (1 + 10%)\n = 1768.02 × 1.1\n = 1944.822百万元\n\n2. 假设毛利率保持不变（题目未说明变化），计算2024年毛利润：\n 2024年毛利润 = 2024年收入 × 毛利率\n = 1944.822 × 40%\n = 777.9288百万元\n\n3. 比较选项：\n A. 754.4百万元\n B. 713.7百万元\n C. 774.4百万元（最接近计算结果）\n D. 784.4百万元\n\nC"
Model Response: I will analyze this question step by step. \n\nFirst, key information obtained from the question: \n- The gross profit margin for the soda ash business in 2023 is assumed to be 40%. \n- The revenue growth rate for 2024 is assumed to be 10%. \n- The gross profit for 2024 needs to be calculated. \n\nFrom the chart, I can see that the revenue for the soda ash business in 2023 is 1,768.02 million yuan. \n\nCalculation steps: \n1. Calculate the revenue for the soda ash business in 2024:\n 2024 revenue = 2023 revenue × (1 + growth rate)\n = 1,768.02 × (1 + 10%)\n = 1,768.02 × 1.1\n = 1,944.822 million yuan\n\n2. Assuming the gross margin remains unchanged (the question does not specify any changes), calculate the gross profit for 2024: \n 2024 Gross Profit = 2024 Revenue × Gross Margin \n = 1,944.822 × 40% \n = 777.9288 million \n\n3. Compare the options: \n A. 754.4 million \n B. 713.7million \n C. 774.4 million (closest to the calculated result) \n D. 784.4 million \n\nC"

答案：B
Answer: B

Figure 31: This is an example of the Claude-3-7-Sonnet-20250219's "incorrect analysis and incorrect answer." This is a calculation problem where the large model needs to extract key information from the question: the 2023 soda ash business gross margin is assumed to be 40%, and the 2024 revenue growth rate is 10%. It must then identify the 2023 soda ash business revenue data from the chart, calculate the 2024 revenue, and derive the gross profit result. During its reasoning process, the large model incorrectly applied the 40% gross margin assumption for 2023 from the question. The correct approach would have been to calculate using the actual 36.7% gross margin for 2024 from the chart. The model's use of erroneous data ultimately led to an incorrect result.



每股指标	2023A	2024E	2025E	2026E
PE	4.89	4.49	4.04	3.65
PB	0.67	0.60	0.54	0.48
EPS	1.57	1.71	1.90	2.10
BVPS	11.47	12.79	14.33	16.03
每股股利	0.47	3.04	4.04	4.04
盈利能力	**2023A**	**2024E**	**2025E**	**2026E**
净息差	1.91%	1.69%	1.62%	1.58%
贷款收益率	5.33%	5.20%	5.10%	5.04%
生息资产收益率	4.22%	4.04%	3.98%	3.95%
存款付息率	2.42%	2.44%	2.42%	2.41%
计息负债成本率	2.52%	2.57%	2.58%	2.59%
ROAA	0.90%	0.91%	0.92%	0.92%
ROAE	15.17%	14.60%	14.41%	14.20%
成本收入比	24.57%	24.04%	23.00%	23.00%
业绩与规模增长	**2023A**	**2024E**	**2025E**	**2026E**
净利息收入	0.73%	7.56%	6.71%	6.47%
营业收入	5.27%	7.05%	6.00%	6.06%
拨备前利润	5.98%	8.45%	5.97%	6.04%
归属母公司净利润	13.25%	13.05%	10.32%	10.31%
净手续费收入	-31.58%	1.01%	1.00%	5.00%
贷款余额	14.57%	12.32%	10.00%	10.00%
生息资产	13.80%	13.90%	9.90%	9.80%
存款余额	11.99%	15.40%	10.00%	10.00%
计息负债	13.80%	13.00%	10.80%	10.00%
资产质量	**2023A**	**2024E**	**2025E**	**2026E**
不良率	0.91%	0.91%	0.91%	0.92%
拨备覆盖率	379.10%	378.00%	362.60%	334.80%
拨贷比	3.45%	3.44%	3.30%	3.08%
不良净生成率	1.10%	1.00%	1.00%	1.00%

利润表（百万元）	2023A	2024E	2025E	2026E
净利息收入	52,643	56,623	60,421	64,333
手续费净收入	4,277	4,320	4,363	4,581
营业收入	74,163	79,393	84,157	89,256
业务及管理费	(17,827)	(18,260)	(19,356)	(20,529)
拨备前利润	55,431	60,117	63,705	67,554
拨备	(16,740)	(22,324)	(21,851)	(21,188)
税前利润	38,691	37,793	41,854	46,366
税后利润	30,010	34,014	37,669	41,729
归属母公司净利润	28,750	32,502	35,855	39,552
资产负债表（百万元）	**2023A**	**2024E**	**2025E**	**2026E**
贷款总额	1,801,797	1,981,977	2,180,175	2,376,391
债券投资	1,229,048	1,351,953	1,487,148	1,635,863
同业资产	141,334	155,467	171,014	188,115
生息资产	3,324,252	3,654,427	4,014,197	4,387,575
资产总额	3,403,362	3,725,045	4,098,317	4,509,148
存款	1,875,337	2,062,871	2,269,158	2,496,074
同业负债	761,309	860,279	946,307	1,040,938
发行债券	392,410	431,651	474,816	522,298
计息负债	3,029,056	3,354,801	3,690,281	4,059,310
负债总额	3,144,246	3,440,822	3,784,904	4,163,395
股本	18,351	18,351	18,351	18,351
归属母公司股东权益	250,409	274,646	302,878	334,164
所有者权益总额	259,115	284,223	313,413	345,753
资本状况：	**2023A**	**2024E**	**2025E**	**2026E**
资本充足率	13.07%	13.31%	12.87%	12.63%
核心资本充足率	8.79%	9.46%	9.44%	9.54%
杠杆率	13.83	13.13	13.11	13.08
RORWA	1.42%	1.42%	1.42%	1.41%
风险加权系数	65.48%	66.51%	68.06%	68.67%

问题：2023年与2025E年相比，江苏银行的每股收益（EPS）增长了多少倍？
Question: How many times did Jiangsu Bank's earnings per share (EPS) increase between 2023 and 2025E?

A. 1.35　　B. 1.97　　C. 1.48

模型回答：我需要计算江苏银行2023年至2025E年EPS的增长倍数。\n\n步骤1：从图表中找出相关数据\n- 2023年EPS = 1.57\n- 2025E年EPS = 2.1\n\n步骤2：计算增长倍数\n增长倍数 = 2025E年EPS ÷ 2023年EPS\n增长倍数 = 2.1 ÷ 1.57\n增长倍数 = 1.3375...倍 ≈ 1.34倍\n\n步骤3：比较计算结果与选项\n计算得到的结果约为1.34倍，最接近的选项是1.35倍。A
Model Response: I need to calculate the growth multiple for Jiangsu Bank's EPS from 2023 to 2025E. \n\nStep 1: Find the relevant data from the chart \n - 2023 EPS = 1.57\n- 2025E EPS = 2.10\n\nStep 2: Calculate the growth multiple\nGrowth multiple = 2025E EPS ÷ 2023 EPS\nGrowth multiple = 2.10 ÷ 1.57\nGrowth multiple = 1.3375... times ≈ 1.34 times\n\nStep 3: Compare the calculated result with the options\nThe calculated result is approximately 1.34 times, and the closest option is 1.35 times. A

答案：B
Answer: B

Figure 32: This is a classic case of Claude-3-7-Sonnet-20250219 producing "logically correct reasoning but incorrect answers." Although the model's reasoning process and calculation formula itself are correct—i.e., Growth Multiple = 2025E EPS ÷ 2023 EPS—the large model made a critical error in data selection. While correctly identifying the 2023 EPS data, it erroneously used the 2026E EPS figure when referencing 2025E EPS, representing a clear year confusion. This serves as a reminder that large models must exercise particular caution when handling financial data, ensuring the correct year is used and verifying that calculations rely on accurate data sources.