FinEval-KR: A Financial Domain Evaluation Framework for Large Language Models' Knowledge and Reasoning

Shaoyu Dou[*]
Ant Group

Yutian Shen[*†] **, Mofan Chen**[*†]
Zixuan Wang[†] **and Jiajie Xu**[†]
Shanghai University of Finance and Economics

Qi Guo, Kailai Shao
Chao Chen and Haixiang Hu
Ant Group

Haibo Shi
Min Min and Liwen Zhang[‡]
Shanghai University of Finance and Economics

Abstract

Large Language Models (LLMs) demonstrate significant potential but face challenges in complex financial reasoning tasks requiring both domain knowledge and sophisticated reasoning. Current evaluation benchmarks often fall short by not decoupling these capabilities indicators from single task performance and lack root cause analysis for task failure. To address this, we introduce FinEval-KR, a novel evaluation framework for decoupling and quantifying LLMs' knowledge and reasoning abilities independently, proposing distinct knowledge score and reasoning score metrics. Inspired by cognitive science, we further propose a cognitive score based on Bloom's taxonomy to analyze capabilities in reasoning tasks across different cognitive levels. We also release a new open-source **Chinese** financial reasoning dataset covering 22 subfields to support reproducible research and further advancements in financial reasoning. Our experimental results reveal that LLM reasoning ability and higher-order cognitive ability are the core factors influencing reasoning accuracy. We also specifically find that even top models still face a bottleneck with knowledge application. Furthermore, our analysis shows that specialized financial LLMs generally lag behind the top general large models across multiple metrics.

1 Introduction

In recent years, rapid LLM development has led the transformation of artificial intelligence. LLMs demonstrate strong natural language processing capabilities and inspire application innovations across various fields, including scientific research (Zhang et al., 2024), financial services (Nie et al., 2024b), content creation (Betker et al., 2023), etc. However, achieving satisfactory performance for complex tasks, such as financial decision making, proves difficult when relying solely on knowledge introduced during training or instructions (Wang and Brorsson, 2025; Liu et al., 2025; Wang et al., 2024). Reasoning ability, that is, the ability of logical deduction, problem solving, and abstract thought, stands as a core hallmark of advanced intelligence and becomes crucial for assessing LLM intelligibility and applicability (Wang and Song, 2024; Li et al., 2024b; Valmeekam et al., 2023). Therefore, it is critical to accurately assess the knowledge and reasoning abilities of LLMs. This helps to understand the shortcomings of the model to support targeted optimization.

Although several LLM evaluation benchmarks have been proposed, they still have several limitations in evaluating reasoning capabilities.

Insufficient capability decoupling. Mainstream benchmarks typically evaluate a model's capabilities based on its performance across various tasks. Their performance represents either the model's knowledge capacity (Nie et al., 2024a; Liu and Jin, 2024) or its reasoning ability (Saparov and He, 2023; Geva et al., 2021). However, our experiments demonstrate that LLM performance in reasoning tasks is influenced by both knowledge and reasoning ability. Therefore, it is essential to decouple and quantify them separately to achieve a more accurate capability characterization.

Lack of root cause analysis. Current reasoning evaluation frameworks mostly focus on the correctness of the reasoning processes and results, while neglecting the diagnosis of the erroneous result. Specifically, there is not yet an effective way to distinguish whether a reasoning failure stems from knowledge gaps, such as unclear concept comprehension, or from flaws in the reasoning processes, like missing reasoning steps. This limits the potential for specifically optimizing the model.

Neglect of the cognitive science perspective in financial LLM benchmark. While some benchmarks

[*] Equal contribution.

[†] This work was completed during the internship at Ant Group.

[‡] Corresponding author.

have begun to assess reasoning ability, they generally lack a design grounded in cognitive science–a critical omission for the financial domain. Unlike the deductive reasoning of law or the diagnostic processes of medicine, the essence of financial decision-making is a *quantitative game against future uncertainty*. This requires a spectrum of higher-order cognitive abilities that extend far beyond simple knowledge application (Zhang et al., 2025). For instance, evaluating the impact of a central bank's interest rate policy requires not only *applying* knowledge of rate-to-exchange dynamics but also *evaluating* the complex interplay between market sentiment and strategic policy expectations. This distinction underscores the need for a framework like Bloom's Taxonomy as an essential diagnostic tool to pinpoint the specific cognitive deficiencies that hinder advanced financial reasoning in LLMs.

The above limitations motivate us to explore the following questions:

Q1: How do knowledge and reasoning ability jointly determine the performance of LLMs in domain reasoning tasks?

Q2: Can we develop an evaluation framework that decouples and independently quantifies knowledge and reasoning ability from task performance?

We begin by empirically verifying the fundamental role of knowledge in domain-specific reasoning through a preliminary experiment in the finance sector. Building on these initial findings, we make the following contributions:

- **A novel evaluation framework**. We propose a novel LLM evaluation framework that disentangles the assessment of domain knowledge and reasoning ability from task performance metrics. This allows us to introduce two distinct metrics: a *Knowledge Score* and a *Reasoning Score*. Furthermore, inspired by cognitive science, we posit that complex reasoning relies on a hierarchy of cognitive abilities. This motivates our third metric, the *Cognitive Score*, which leverages Bloom's Taxonomy to provide a fine-grained analysis of the cognitive processes employed by LLMs during reasoning.

- **A new open-source Chinese-language dataset for financial reasoning**[1]. The released dataset encompasses 22 key financial subfields, its primary contribution lies in its multi-layered annotations, where each sample includes knowledge point labels, step-by-step reasoning chains, and the required cognitive skills. As a comprehensive and deeply annotated resource, it is designed to serve as a specialized benchmark to advance research in both financial and cognitive reasoning.

- **Evaluation results of mainstream LLMs**. Based on the proposed framework and dataset, we conduct a comprehensive evaluation of the current mainstream LLMs, which verifies the effectiveness of the proposed evaluation methodology and yields a series of insightful conclusions (see Section 5.3 and Appendix I for details).

2 Related Work

Recent studies highlight a paradigm shift in LLM evaluation, moving from task-specific benchmarks to capability-based assessments of core competencies like knowledge and reasoning (Cao et al., 2025). Financial LLM benchmarks have followed a similar trajectory, evolving through three main phases. Early benchmarks adapted general NLP tasks to the financial domain (e.g., FinGPT (Wang et al., 2023), CFBenchmark (Lei et al., 2023)). As tasks grew more complex, the focus shifted to specialized knowledge evaluation (e.g., FinEval (Zhang et al., 2023), FinTruthQA (Xu et al., 2024)). Contemporary benchmarks now incorporate complex decision-making tasks requiring integrative reasoning, such as market analysis and risk assessment (e.g., InvestorBench (Li et al., 2024a), FinBen (Xie et al., 2024)).

Despite this progress, financial reasoning poses unique challenges that current evaluation methods struggle to address. It demands (1) comprehension of complex, multi-modal data (e.g., time-series and unstructured text), (2) deep domain-specific knowledge, and (3) advanced computational and deductive skills. This complexity creates a critical limitation in existing benchmarks: overall performance metrics entangle knowledge mastery with reasoning ability, making it impossible to diagnose the true source of a model's failure. This fundamental challenge motivates our primary contribution: a decoupled assessment framework. Furthermore, these unique demands directly guided the design of our financial reasoning dataset (see Section 4.1).

[1] https://github.com/SUFE-AIFLM-Lab/FinEval-KR

Beyond the issue of entangled evaluation, a second key limitation exists: the lack of fine-grained cognitive analysis. While cognitive science perspectives offer crucial insights for LLM optimization (Huber and Niklaus, 2025; Adams, 2015), they are largely overlooked in financial benchmarks. This gap hinders targeted model improvement, as engineers cannot pinpoint specific cognitive deficiencies (e.g., analysis vs. evaluation) to address during fine-tuning. Our work fills this gap by introducing a cognitive evaluation dimension, enabling a more diagnostic approach to model development.

3 Preliminary Experiment

To further illustrate our research motivation, we design a set of preliminary experiments focusing on a simple reasoning task in the financial domain – a financial calculation problem that requires only a single formula. For the dataset and prompts used in this experiments, please refer to Appendix A.

3.1 Experiment Settings

In order to correctly solve such problems, LLMs need to complete the following three sub-tasks: (1) Recall the calculation formula for the variable to be solved. (2) Identify the variable names and their values from the problem statement. (3) Substitute the variable values into the formula and calculate the target variable. According to Bloom's Taxonomy, these three sub-tasks correspond to remembering, understanding and applying/analyzing respectively. Obviously, the model can only answer the questions correctly if all these subtasks are done correctly.

We design the following three comparative experiments to evaluate the impact of knowledge on reasoning task: *Experiment 1* (E1). The LLM is directly prompted with the original question and asked to complete the entire reasoning process independently. *Experiment 2* (E2). Based on experiment 1, inject variables and their values into the prompt. *Experiment 3* (E3). Further provide formulas for calculation based on experiment 2.

To ensure the fairness, we add an equal amount of irrelevant information as distractors to the control groups, while providing the key knowledge in the experimental groups. The experiments are first conducted on the Qwen2.5-7B_Instruct, and the generalizability of the findings is verified using GPT-4o. The experimental results are presented in Table 1.

Model/Settings	E1	E2	E3
Qwen2.5-7B_Instruct	58.0%	72.4%	85.9%
GPT-4o	64.5%	80.5%	92.5%

Table 1: Cumulative correctness rate of reasoning task in three experimental settings.

3.2 Results Analysis

Experiment 3 to 1 can be regarded as knowledge stripping experiments. As key knowledge is progressively removed, the problem-solving rate decreases significantly, suggesting that the lack of knowledge is often the root cause of reasoning failure in reasoning tasks. This leads to the following conclusion.

In complex domain reasoning tasks, knowledge is necessary for successful reasoning.

On the contrary, Experiments 1 to 3 can be regarded as knowledge enhancement experiments, and the results show that the reasoning success rate of LLM is significantly improved after the introduction of key knowledge in the prompts. However, even knowledge is sufficiently injected into the prompt, GPT-4o still persists with an error rate of about 7.5%, suggesting that reasoning ability may become a performance bottleneck for such tasks. Combined with the previous conclusion, we infer that:

Knowledge is a necessary but not sufficient condition for successful reasoning.

It is noteworthy that both models exhibit significant knowledge dependence in all experimental settings, while the performance gap between them persists. This performance discrepancy suggests that there may be significant differences in the knowledge and reasoning capabilities of different models. Therefore, the decoupled evaluation framework can help us identify more clearly the shortcomings of the models in terms of knowledge and reasoning capabilities.

From a cognitive science perspective, knowledge stripping experiments revealed the damaging effects of lower-order cognitive deficits on higher-order reasoning, as evidenced by a 27.94% and 28% decrease in reasoning accuracy in qwen2.5-7b and GPT-4o, respectively. This change also supports the progressive dependence between cognitive levels. Furthermore, the performance difference between the two models in the same experiment settings suggests that there is a significant gap between them, at least in the remembering layer.

4 FinEval-KR

In this section, we present a **Fin**ancial domain **Eval**uation framework for assessing **K**nowledge and **R**easoning abilities (**FinEval-KR**). We first describe the methodology used to construct the evaluation dataset. Next, we introduce a multi-stage evaluation framework that performs root cause analysis of reasoning errors via knowledge-augmented question answering, enabling a decoupled assessment of a model's knowledge mastery and reasoning ability. Finally, we define a series of evaluation metrics: a *knowledge score* based on domain knowledge coverage; a *reasoning score* and a *cognitive score* which is based on Bloom's taxonomy. This approach improves the interpretability of LLM evaluations in financial scenarios and provides clear directions for targeted model improvement.

4.1 Benchmark Dataset Construction

The FinEval-KR benchmark dataset is constructed through four steps: data collection and processing, automated question generation, answer generation, and dataset annotation. This framework ensures comprehensive coverage of financial knowledge domains while maintaining academic rigor. All prompt templates for the dataset generation are detailed in Appendix B.

Corpus Collection To ensure the dataset is both authoritative and relevant, we selected nine canonical textbooks from major financial disciplines. These sources provide a comprehensive and up-to-date overview of modern finance. This process yielded a financial corpus totaling 8,460 pages. A detailed list of the textbooks and the rationale for their selection is available in Appendix C.

Question Generation We generate financial problems from the obtained corpus using a two-stage process. First, we use a custom-designed prompt to instruct OpenAI o1, a state-of-the-art model at the time of our research that was specially enhanced for reasoning capabilities, to create a computational question based on a given text segment. The prompt's design ensures the question meets predefined standards (see Figure 6 in the appendix for details). Second, we subject each candidate question to a three-step automated validation: it is checked for logical coherence, consistency with the source material, and overall quality. Questions that fail any check are discarded, ensuring the high fidelity of the final dataset. The details of validation please refer to Appendix D.

Answer Generation We generate and validate the ground-truth answers using a structured three-stage pipeline: (1) Unconstrained generation: We first prompt OpenAI o1 to solve each problem, guided by the prompt shown in Figure 7. Crucially, we impose no initial format constraints on the output, a strategy designed to capture the model's most natural and diverse problem-solving pathways. (2) Standardized formatting: The resulting raw solutions are then systematically parsed and reformatted according to a predefined template to ensure consistent and uniform presentation across the dataset. (3) Rigorous validation: Finally, each formatted answer undergoes a triple-validation protocol, which mirrors the question validation process (see Appendix D for details).

Dataset Annotation For each question, we use OpenAI o1 guided by the prompt in Figure 8 to identify all requisite knowledge points. These include, but are not limited to, core financial concepts, regulatory frameworks and mathematical operations.

For each step in the answer, we annotate the corresponding cognitive level using Bloom's Taxonomy. To mitigate the inherent subjectivity of this task, we developed a constrained, keyword-driven methodology. We guide the OpenAI o1 using a predefined set of keywords strongly associated with each cognitive level, derived from Anderson and Krathwohl (2001) (see Figure 9 for prompt template). This process maps each reasoning step to one or more levels: Remembering, Understanding, Applying, Analyzing, and Evaluating. We intentionally exclude the Creating level. This decision ensures objectivity, as our dataset comprises problems with determinate solutions, a principle that aligns with standardized financial certification exams.

Dataset Characteristics and Exemplary Samples The final evaluation set contains a total of 9,782 question-answer pairs and their associated annotations. The questions and answers in the dataset are verified by human experts, and a sample check showed that the accuracy of the dataset is above 90%. Further details on the human annotators, the quality control mechanisms, and the validation process are provided in the Appendix J. An exemplary sample from the constructed dataset is given in Figure 1 on a yellow background. (see Figure 10 for the complete example). The complete statistical characterization of the dataset is detailed in Appendix E.

4.2 Evaluation Framework

As shown in Figure 1, the FinEval-KR framework comprises three evaluation stages. First, a model attempts a problem, and an LLM-as-a-judge identifies any errors, extracting the specific knowledge points needed for a correct solution. In the second stage, we provide the model with this missing knowledge and have it re-answer the question. The final stage performs a comparative analysis for attributing the initial error to knowledge deficiency or reasoning ability deficiency. This mechanism allows for the independent quantification of a model's knowledge and reasoning capabilities.

4.2.1 Stage 1: Question Answering

Unconstrained Solution Generation In this initial stage, the model under evaluation generates a solution without any format constraints. This design is crucial for preventing judgment errors based on superficial format-matching. It ensures, for example, that a logically sound reasoning path is not unfairly penalized simply for deviating from the step-order of the reference answer. The prompt for this stage is detailed in Appendix F, Figure 15.

Structured Judgment and Error Analysis Next, we employ an LLM-as-a-judge to analyze the generated solution against a reference answer. To ensure the judgment is objective and reproducible, the judge is guided by a highly structured, Chain-of-Thought (CoT) prompt that enforces a rigorous step-by-step analysis (see Figure 14). Additionally, we address the potential bias challenge of the judge model in Appendix G.

If an error is detected, the judge's output, termed the *review result*, pinpoints the first incorrect step, identifies its root cause, and lists the corresponding knowledge deficiencies (an example is shown in Figure 2). If the final answer is correct, we consider the entire reasoning process valid. This assumption is grounded in the novelty and multi-step complexity of our dataset, which makes correct answers via guessing or exploiting artifacts (the "Clever Hans" effect) highly improbable.

4.2.2 Stage 2: Knowledge-augmented Answering

If the evaluated model makes a error in Stage 1, the framework proceeds to this second stage. Here, we provide the model with the exact knowledge points that the judge identified as missing in Stage 1. This knowledge is integrated into a new prompt (see Appendix F, Figure 16), instructing the model to re-attempt the problem. The core purpose of this stage is to isolate the reasoning variable. By explicitly providing the necessary knowledge – a prerequisite for correct reasoning as validated in our preliminary study – we can now assess if the model can reason correctly when its knowledge gaps are filled. The judge then re-evaluates the new solution using the same protocol as in Stage 1. The outcome of this assessment is termed the *augmented review result*.

4.2.3 Stage 3: Error Diagnosis

The final stage performs a comparative analysis between the outcomes of Stage 1 and Stage 2 to determine the root cause of the initial error. Our approach is grounded in the principle that *LLM's preference to external information reveals its internal knowledge gaps* (Wu et al., 2024). Specifically, the judge model compares the *review result* with the *augmented review result*, following the principles below to determine the root cause of reasoning errors:

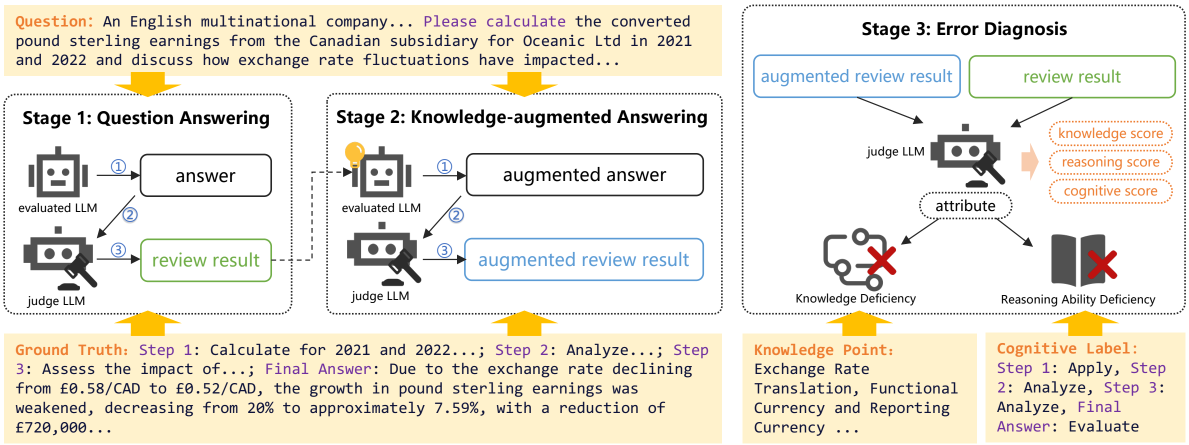
- *Knowledge Deficiency*. If the augmented review result shows that the model reasons correctly in Stage 2, or the erroneous step occurs later than in Stage 1. This indicates that the evaluated model preferred the augmented knowledge in the second stage. This proves that the initial error is caused by knowledge deficiency.

- *Reasoning Ability Deficiency*. If the evaluated model still makes a reasoning error in Stage 2, and the erroneous step is consistent with Stage 1, this indicates that the evaluated model still preferred its internal prior knowledge. This proves that the initial error is rooted in poor reasoning ability.

Through this attribution method, our framework successfully decouples a model's knowledge and reasoning abilities from its overall task performance.

4.2.4 Evaluation Metrics

We propose three core metrics: knowledge score, reasoning score, and cognitive score, while also retaining accuracy to measure the model's overall task performance.



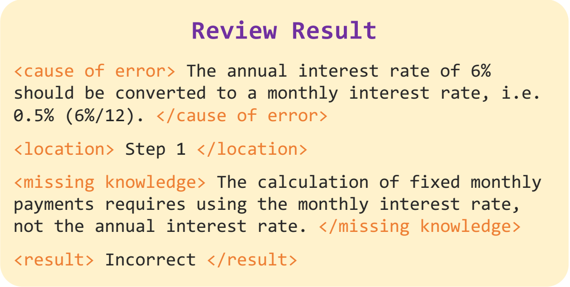
Question: An English multinational company... Please calculate the converted pound sterling earnings from the Canadian subsidiary for Oceanic Ltd in 2021 and 2022 and discuss how exchange rate fluctuations have impacted...

Ground Truth: Step 1: Calculate for 2021 and 2022...; Step 2: Analyze...; Step 3: Assess the impact of...; Final Answer: Due to the exchange rate declining from £0.58/CAD to £0.52/CAD, the growth in pound sterling earnings was weakened, decreasing from 20% to approximately 7.59%, with a reduction of £720,000...

Knowledge Point: Exchange Rate Translation, Functional Currency and Reporting Currency ...

Cognitive Label: Step 1: Apply, Step 2: Analyze, Step 3: Analyze, Final Answer: Evaluate

Figure 1: Three-stage evaluation framework of FinEval-KR, and an exemplary sample of the dataset. Note that the original dataset is in Chinese, the figure provides an English translation for readability.



Review Result

<cause of error> The annual interest rate of 6% should be converted to a monthly interest rate, i.e. 0.5% (6%/12). </cause of error>

<location> Step 1 </location>

<missing knowledge> The calculation of fixed monthly payments requires using the monthly interest rate, not the annual interest rate. </missing knowledge>

<result> Incorrect </result>

Figure 2: Example of review result generated by the judge model (original in Chinese, with English translation).

Knowledge Score (KS) This metric quantifies the evaluated model's knowledge coverage in the financial domain.

$$KS = 1 - \frac{\left|\bigcup_{i=1}^{M} K_i'\right|}{\left|\bigcup_{i=1}^{N} K_i\right|}, \tag{1}$$

where M is the number of errors attributed to knowledge deficiency during evaluation. K_i' denotes the set of knowledge points involved in the erroneous reasoning steps for the i-th question whose error was attributed to a knowledge deficiency. N is the total number of evaluation samples. K_i denotes the set of knowledge points involved in the i-th evaluation question. The denominator in the Eq. (1) is the total number of knowledge points across all evaluation questions in the dataset.

Reasoning Score (RS) This metric measures the evaluated model's reasoning ability in the financial domain.

$$RS = \frac{\sum_{i=1}^{N} \mathbb{I}(a_i = a_i^{\text{ref}})}{N - \sum_{i=1}^{N} \mathbb{I}(a_i \neq a_i^{\text{ref}} \wedge r(a_i) = \text{K})}, \tag{2}$$

where a_i is the answer of the evaluated model for the i-th question, and a_i^{ref} is the corresponding reference answer. Thus, the numerator of Eq. (2) is the total number of questions with correct reasoning. $r(a_i)$ represents the root cause for the erroneous result a_i, which can be either K or R, representing knowledge and reasoning ability deficiency, respectively. In addition, since each reasoning step in the dataset is annotated with a cognitive label, the reasoning ability deficit can be further subdivided into the ability deficit at a certain cognitive level, i.e., R^j, $j \in \{1, 2, 3, 4, 5\}$, where j is the index of the level in Bloom's taxonomy, and $1, 2, 3, 4, 5$ denote remembering, understanding, applying, analyzing, and evaluating respectively. The denominator in Eq. (2) is the total number of questions whose reasoning errors are not attributed to knowledge deficiencies.

Cognitive Score (CS) Furthermore, to explore the level of cognition exhibited by LLMs when solving reasoning tasks, we expanded a series of fine-grained metrics drawing on RS. The j-th cognitive level score of the LLM is defined as,

$$CS_j = RS \times (1- \alpha_j \times \frac{\sum_{i=1}^{N} \mathbb{I}(r(a_i) = \text{R}^j)}{N - \sum_{i=1}^{N} \mathbb{I}(a_i \neq a_i^{\text{ref}} \wedge r(a_i) = \text{K})}). \tag{3}$$

$\alpha_j \in (0, 1)$ is a penalty coefficient, and it is designed to have negative correlation with the cognitive level j. This weighting scheme is designed to heavily penalize errors made by the model during lower-level cognitive reasoning steps. In our experiments, we empirically set $\alpha_1, \cdots, \alpha_5$ to a linearly decreasing sequence 0.9, 0.8, 0.7, 0.6 and 0.5.

Accuracy (Acc) Finally, the success rate for all reasoning tasks is defined as,

$$Acc = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(a_i = a_i^{\text{ref}}). \tag{4}$$

5 Experiments

This section first validates the FinEval-KR framework's effectiveness in identifying potential knowledge weaknesses and performing root cause analysis. Subsequently, we selected a range of LLMs, and evaluated them using the proposed FinEval-KR framework and our open-sourced dataset.

5.1 Alignment Analysis of Judge Model in FinEval-KR

The core of FinEval-KR lies in the design of the judge model for identifying initial reasoning step, recalling potential knowledge weaknesses and pinpointing root causes, which is fundamental for achieving decoupled evaluation. To this end, we design a set of comparative experiments to assess the alignment between the judge model and human evaluation.

The methods participating in the comparison include: (a) *Direct Prompting*: We use advanced models like OpenAI o1, directly prompting them to execute knowledge identification and root cause localization tasks. (b) *Task Decomposition*: Decompose the above two tasks into multiple subtasks and clearly define the logical dependencies between these subtasks (c) *Ours*: Based on task decomposition, this method requires the model to self-reflect and explicitly output reasoning processes step-by-step before generating the final conclusion (i.e., adding the `<Inner Thought>` as shown in Figure 14).

Given that our benchmark dataset is composed in Chinese, we select Qwen2.5-72B_Instruct, which is specifically optimized through extensive pretraining on Chinese corpora, as backbone LLM in the latter two methods. We use the following three metrics to evaluate the performance of these methods: (a) *Accuracy of error identification*: This refers to the proportion of correctly localized initial reasoning step in the Stage 1. (b) *Accuracy of recalled knowledge points*: The percentage of missing knowledge correctly recalled in Stage 1. (c) *Accuracy of error attribution* : The proportion of correctly attributed reasoning errors to knowledge or reasoning ability deficits.

Methods	Error Identification	Knowledge Recall	Error Attribution
Direct Prompting	0.56	0.24	0.20
Task Decomposition	0.85	0.60	0.53
Ours	**0.92**	**0.94**	**0.93**

Table 2: Accuracy for three evaluation tasks.

All metrics are calculated after a manual review of the model outputs by financial domain experts. Experimental results are presented in Table 2. It can be observed that FinEval-KR outperforms the comparison methods across all metrics, demonstrating its superior performance in error attribution and knowledge identification tasks.

Additionally, we discuss the limitations of adopting Qwen2.5-72B_Instruct as the judge model in the Limitations section.

5.2 Evaluation

For our evaluation, we select 18 leading and representative LLMs, referencing prominent leaderboards like the Chatbot Arena[2]. This selection spans a diverse range of models, including open-source and closed-source systems, models of varying parameter scales, and different architectures such as dense and Mixture-of-Experts (MoE) (see Appendix H for details).

In our analysis, we focus on the relative performance rankings (i.e., performance tiers) of these models rather than their absolute scores. This approach is designed to ensure the robustness of our findings. While an LLM-as-a-judge may have inherent systematic biases, such biases have a smaller impact on the relative ordering of models than on their absolute scores.

5.3 Results and Core Findings

Table 3 lists the average metrics for all 18 models from three independent runs, conducted with a model temperature of 1. The complete results, including standard deviations, are presented in Table 7.

[2] https://openlm.ai/chatbot-arena/

Model/Metrics	Acc	KS	RS	CS$_1$ (remember)	CS$_2$ (understand)	CS$_3$ (apply)	CS$_4$ (analyze)	CS$_5$ (evaluate)
Qwen2.5-14B_Instruct	0.5473	0.8490	0.6863	0.6547	0.6603	0.3893	0.6863	0.6820
QwQ-32B-preview	0.7380	0.9073	0.8627	0.8450	0.8503	0.6987	0.8510	0.8597
DeepSeek-v3	0.8270	0.9427	0.9077	0.8963	0.8993	0.7963	0.8943	0.9057
DeepSeek-R1	0.8700	0.9517	**0.9347**	**0.9377**	**0.9397**	**0.8810**	**0.9380**	**0.9433**
Doubao-pro-32k	0.7825	0.9195	0.8750	0.8560	0.8600	0.7340	0.8565	0.8720
Moonshot-v1-128k	0.4533	0.8340	0.6020	0.5620	0.5670	0.2763	0.5653	0.5973
Ernie-bot-4.0	0.5733	0.8627	0.7053	0.6680	0.6753	0.4383	0.6847	0.6927
Qwen-max-latest	0.6467	0.8797	0.7733	0.7507	0.7547	0.5340	0.7440	0.7703
GPT-3.5-turbo	0.2830	0.7527	0.3973	0.3527	0.3603	0.0900	0.3893	0.3970
GPT-4o	0.6853	0.9020	0.8067	0.7847	0.7890	0.5930	0.7870	0.8030
GPT-4.1	0.8263	0.9520	0.9063	0.8957	0.8977	0.7890	0.8927	0.9050
o1-mini	0.7503	0.8997	0.8453	0.8340	0.8363	0.6983	0.8477	0.8450
o3-mini	0.8207	0.9260	0.9070	0.9047	0.9073	0.8127	0.9023	0.9120
Gemini-2.5-pro	**0.8750**	**0.9627**	0.9233	0.9123	0.9163	0.8403	0.9050	0.9120
Gemini-2.5-flash	0.8440	0.9540	0.9203	0.9103	0.9133	0.8307	0.9100	0.9177
Claude-3.7-sonnet	0.7923	0.9390	0.8823	0.8663	0.8703	0.7433	0.8653	0.8803
Xuanyuan-FinX1-preview	0.5890	0.8687	0.7323	0.7063	0.7130	0.4610	0.7323	0.7300
Fin-R1-7B	0.4153	0.7510	0.5570	0.5190	0.5277	0.2170	0.5570	0.5527

Table 3: Reasoning Accuracy (Acc), Knowledge Score (KS), Reasoning Score (RS), and Cognitive Score (CS) of evaluated LLMs on the FinEval-KR. The complete results please refer to Table 7 in the appendix.

The analysis of the model performance echelons and discussion of the results are detailed in in Appendix I. In summary, the comprehensive analysis of all evaluation metrics identifies the current Tier 1 models as DeepSeek-R1, Gemini-2.5-pro and Gemini-2.5-flash. These models typically have parameters in excess of a hundred billion and use the MoE architecture to optimize computational resources. Furthermore, they specifically optimize reasoning capabilities through methods like reinforcement learning, and demonstrate outstanding performance in knowledge coverage and the completeness of reasoning paths.

Bottleneck in Knowledge Applying Abilities Our analysis reveals that it is reasoning and specific cognitive skills, not merely knowledge, that truly drive performance in advanced LLMs. While the Knowledge Score (KS) and Reasoning Score (RS) are positively correlated, KS scores converge among top-tier models, indicating that sheer knowledge is no longer the primary performance bottleneck. Instead, RS shows a strong correlation with accuracy, establishing reasoning ability as crucial for success. A deeper cognitive analysis pinpoints the ability to apply knowledge (CS$_3$) as the critical differentiator, evidenced by a sharp drop in this metric, which directly degrades their reasoning and accuracy.

Crucially, this weakness is not confined to lower-tier models. Even top-tier LLMs exhibit a significant deficit in applying knowledge (CS$_3$) compared to their abilities in analyzing (CS$_4$) or evaluating (CS$_5$). For instance, GPT-4.1 scores 0.7890 in CS$_3$ versus 0.8927 in CS$_4$. This universal shortcoming underscores a fundamental limitation of current models: a profound difficulty in transferring theoretical knowledge to practical, real-world application.

The Dilemma of Financial LLMs Our evaluation establishes Xuanyuan-FinX1-preview as the leading specialized financial LLM, consistently outperforming counterparts like Fin-R1-7B. However, a more critical finding emerges when comparing it to state-of-the-art general LLMs. Despite its domain leadership, Xuanyuan-FinX1-preview exhibits a significant performance gap of above 20%, a deficit that spans across financial complex reasoning, and higher-order cognitive skills. We attribute this gap to the superior generalization capabilities of leading general LLMs, which are developed through pre-training on vast, multi-domain datasets and benefit from more rapid iteration cycles. This advantage allows them to achieve strong performance even in financial fields, highlighting the limitations of current financial LLMs in terms of data diversity and development velocity.

Consequently, we predict a dual-track future for developing high-performance financial LLMs. The first track involves building upon state-of-the-art general foundation models to leverage their vast world knowledge and robust generalization. The second requires employing advanced fine-tuning techniques to distill and align these models for specialized financial reasoning, moving beyond a simple reliance on domain data.

6 Conclusion

This paper introduces FinEval-KR, a novel evaluation framework designed to decouple and assess the knowledge and reasoning abilities of LLMs in the financial domain, supplemented by a cognitive perspective and a new dataset. Our evaluation results indicate that reasoning and higher-order cognitive abilities are crucial for reasoning accuracy. Even top models encounter a bottleneck in knowledge application, and specialized financial models generally lag behind top general LLMs.

Limitations

A potential limitation of this study lies in the choice of the judge model. Our primary experiments were conducted using Qwen-2.5-72B_Instruct, which represented the state-of-the-art among publicly available models with strong Chinese support during our experimental phase in late 2024. With the rapid evolution of large language models, even more capable reasoning models like DeepSeek-R1 have since emerged.

To investigate the impact of this evolution, we performed a small-scale evaluation using DeepSeek-R1 as the judge model. The results revealed a clear performance-efficiency trade-off: while DeepSeek-R1 yielded a marginal accuracy improvement of approximately 3%, it nearly doubled the inference time, posing significant challenges for large-scale evaluation. Crucially, we found that although the absolute scores of the evaluated models slightly increased, their relative rankings and the performance gaps between them remained highly consistent. Since our research focuses on the comparative performance of different methods, this consistency confirms the robustness of our conclusions.

Our future work will focus on enhancing the framework's robustness by incorporating reasoning LLMs and more diverse evaluation paradigms. Specifically, we plan to employ multiple, heterogeneous models for dataset generation and implement a cross-validated, multi-judge evaluation pipeline to minimize potential biases, raise the evaluation ceiling, and bolster the benchmark's overall reliability.

References

Nancy E Adams. 2015. Bloom 's taxonomy of cognitive learning objectives. *Journal of the Medical Library Association: JMLA*, 103(3):152.

Lorin W Anderson and David R Krathwohl. 2001. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition*. Addison Wesley Longman, Inc.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, and 1 others. 2023. Improving image generation with better captions. *Computer Science. https://cdn.openai.com/papers/dall-e-3.pdf*, 2(3):8.

Yixin Cao, Shibo Hong, Xinze Li, Jiahao Ying, Yubo Ma, Haiyuan Liang, Yantao Liu, Zijun Yao, Xiaozhi Wang, Dan Huang, and 1 others. 2025. Toward generalizable evaluation in the LLM era: A survey beyond benchmarks. *arXiv preprint arXiv:2504.18838*.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Thomas Huber and Christina Niklaus. 2025. LLMs meet Bloom 's taxonomy: A cognitive view on large language model evaluations. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 5211–5246. Association for Computational Linguistics.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2024. Benchmarking cognitive biases in large language models as evaluators. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 517–545.

Yang Lei, Jiangtong Li, Ming Jiang, Junjie Hu, Dawei Cheng, Zhijun Ding, and Changjun Jiang. 2023. Cfbenchmark: Chinese financial assistant benchmark for large language model. *arXiv preprint arXiv:2311.05812*.

Haohang Li, Yupeng Cao, Yangyang Yu, Shashidhar Reddy Javaji, Zhiyang Deng, Yueru He, Yuechen Jiang, Zining Zhu, Koduvayur Subbalakshmi, Guojun Xiong, and 1 others. 2024a. InvestorBench: A benchmark for financial decision-making tasks with LLM-based agent. *arXiv preprint arXiv:2412.18174*.

Zhiming Li, Yushi Cao, Xiufeng Xu, Junzhe Jiang, Xu Liu, Yon Shin Teo, Shang-Wei Lin, and Yang Liu. 2024b. LLMs for relational reasoning: How far are we? In *Proceedings of the 1st International Workshop on Large Language Models for Code*, pages 119–126.

Xinyu Liu and Ke Jin. 2024. MTFinEval: A multi-domain chinese financial benchmark with eurypalynous questions. *arXiv preprint arXiv:2408.10921*.

Zhaowei Liu, Xin Guo, Fangqi Lou, Lingfeng Zeng, Jinyi Niu, Zixuan Wang, Jiajie Xu, Weige Cai, Ziwei Yang, Xueqian Zhao, and 1 others. 2025. Fin-r1: A large language model for financial reasoning through reinforcement learning. *arXiv preprint arXiv:2503.16252*.

Ying Nie, Binwei Yan, Tianyu Guo, Hao Liu, Haoyu Wang, Wei He, Binfan Zheng, Weihao Wang, Qiang Li, Weijian Sun, and 1 others. 2024a. Cfinbench: A comprehensive chinese financial benchmark for large language models. *arXiv preprint arXiv:2407.02301*.

Yuqi Nie, Yaxuan Kong, Xiaowen Dong, John M Mulvey, H Vincent Poor, Qingsong Wen, and Stefan Zohren. 2024b. A survey of large language models for financial applications: Progress, prospects and challenges. *arXiv preprint arXiv:2406.11903*.

Abulhair Saparov and He He. 2023. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2024. Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-judges. *arXiv preprint arXiv:2406.12624*.

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2023. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. *Advances in Neural Information Processing Systems*, 36:38975–38987.

Neng Wang, Hongyang Yang, and Christina Dan Wang. 2023. FinGPT: Instruction tuning benchmark for open-source large language models in financial datasets. *arXiv preprint arXiv:2310.04793*.

Weiqi Wang and Yangqiu Song. 2024. MARS: benchmarking the metaphysical reasoning abilities of language models with a multi-task evaluation dataset. *arXiv preprint arXiv:2406.02106*.

Xinlin Wang and Mats Brorsson. 2025. Can large language model analyze financial statements well? In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 196–206.

Zhihu Wang, Shiwan Zhao, Yu Wang, Heyuan Huang, Sitao Xie, Yubo Zhang, Jiaxin Shi, Zhixing Wang, Hongyan Li, and Junchi Yan. 2024. Re-task: Revisiting LLM tasks from capability, skill, and knowledge perspectives. *arXiv preprint arXiv:2408.06904*.

Kevin Wu, Eric Wu, and James Zou. 2024. How faithful are rag models? quantifying the tug-of-war between rag and llms' internal prior. *CoRR*, abs/2404.10198.

Minghao Wu and Alham Fikri Aji. 2025. Style over substance: Evaluation biases for large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 297–312.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, and 1 others. 2024. FinBen: A holistic financial benchmark for large language models. *Advances in Neural Information Processing Systems*, 37:95716–95743.

Ziyue Xu, Peilin Zhou, Xinyu Shi, Jiageng Wu, Yikang Jiang, Dading Chong, Bin Ke, and Jie Yang. 2024. FintruthQA: A benchmark dataset for evaluating the quality of financial information disclosure. *arXiv preprint arXiv:2406.12009*.

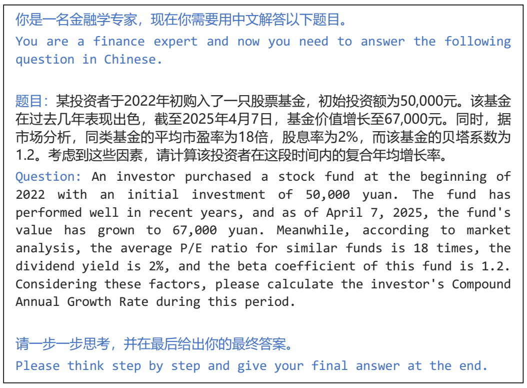
Liwen Zhang, Weige Cai, Zhaowei Liu, Zhi Yang, Wei Dai, Yujie Liao, Qianru Qin, Yifei Li, Xingyu Liu, Zhiqiang Liu, and 1 others. 2023. FinEval: A chinese financial domain knowledge evaluation benchmark for large language models. *arXiv preprint arXiv:2308.09975*.

Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. 2024. A comprehensive survey of scientific large language models and their applications in scientific discovery. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8783–8817.

Zhihan Zhang, Yixin Cao, and Lizi Liao. 2025. XFIN-BENCH: Benchmarking llms in complex financial problem solving and reasoning. *arXiv preprint arXiv:2508.15861*.

A The Dataset and Prompts Used in Preliminary Experiments

We manually construct a dataset consisting of 200 samples. Each sample in this dataset includes the question, the financial formula being examined, the mapping between the formula's variables and the specific numerical values, and the ground truth to the question. Figure 3, 4, and 5 show the prompt templates and example samples used in experiments 1, 2, and 3, respectively. In these figures, the content of the prompt template is shown in blue text, while the test samples are shown in black text. The ground truth for this problem is 0.1024 or 10.24%.



你是一名金融学专家，现在你需要用中文解答以下题目。

You are a finance expert and now you need to answer the following question in Chinese.

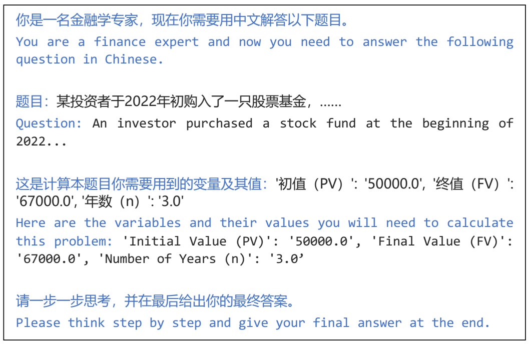
题目：某投资者于2022年初购入了一只股票基金，初始投资额为50,000元。该基金在过去几年表现出色，截至2025年4月7日，基金价值增长至67,000元。同时，据市场分析，同类基金的平均市盈率为18倍，股息率为2%，而该基金的贝塔系数为1.2。考虑到这些因素，请计算该投资者在这段时间内的复合年均增长率。

Question: An investor purchased a stock fund at the beginning of 2022 with an initial investment of 50,000 yuan. The fund has performed well in recent years, and as of April 7, 2025, the fund's value has grown to 67,000 yuan. Meanwhile, according to market analysis, the average P/E ratio for similar funds is 18 times, the dividend yield is 2%, and the beta coefficient of this fund is 1.2. Considering these factors, please calculate the investor's Compound Annual Growth Rate during this period.

请一步一步思考，并在最后给出你的最终答案。

Please think step by step and give your final answer at the end.

Figure 3: The prompt for experiment 1 and an exemplary sample (original in Chinese, with English translation).



你是一名金融学专家，现在你需要用中文解答以下题目。

You are a finance expert and now you need to answer the following question in Chinese.

题目：某投资者于2022年初购入了一只股票基金，......

Question: An investor purchased a stock fund at the beginning of 2022...

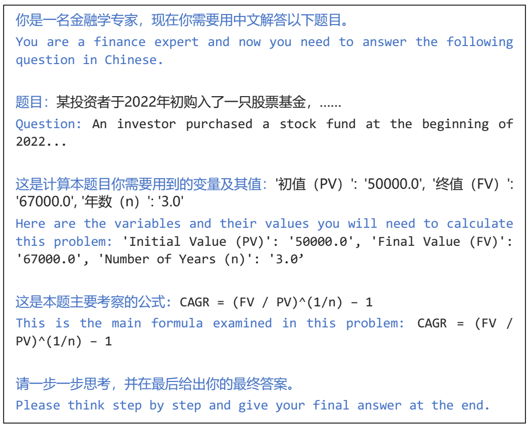
这是计算本题目你需要用到的变量及其值：'初值（PV）': '50000.0', '终值（FV）': '67000.0', '年数（n）': '3.0'

Here are the variables and their values you will need to calculate this problem: 'Initial Value (PV)': '50000.0', 'Final Value (FV)': '67000.0', 'Number of Years (n)': '3.0’

请一步一步思考，并在最后给出你的最终答案。

Please think step by step and give your final answer at the end.

Figure 4: The prompt for experiment 2 and an exemplary sample (original in Chinese, with English translation).



你是一名金融学专家，现在你需要用中文解答以下题目。

You are a finance expert and now you need to answer the following question in Chinese.

题目：某投资者于2022年初购入了一只股票基金，......

Question: An investor purchased a stock fund at the beginning of 2022...

这是计算本题目你需要用到的变量及其值：'初值（PV）': '50000.0', '终值（FV）': '67000.0', '年数（n）': '3.0'

Here are the variables and their values you will need to calculate this problem: 'Initial Value (PV)': '50000.0', 'Final Value (FV)': '67000.0', 'Number of Years (n)': '3.0’

这是本题主要考察的公式：CAGR = (FV / PV)^(1/n) - 1

This is the main formula examined in this problem: CAGR = (FV / PV)^(1/n) - 1

请一步一步思考，并在最后给出你的最终答案。

Please think step by step and give your final answer at the end.

Figure 5: The prompt for experiment 3 and an exemplary sample (original in Chinese, with English translation).

B The Prompt Templates for the Dataset Generation

Figure 6 shows a prompt template for generating questions for a given subfield based on a given piece of corpus. Figure 7 shows a prompt template that generates an solution to a given question. Figures 8 and 9 show the prompt templates for labeling knowledge points and step-level cognitive abilities, respectively.

C Data Sources for the FinEval-KR Dataset

To ensure our benchmark is both authoritative and comprehensive, we constructed the source corpus from nine classic textbooks in modern finance. This selection provides extensive coverage across key subfields, including corporate finance, investments, financial markets, risk management, and monetary policy. These foundational texts supply a rich combination of core theoretical principles and practical case studies, forming a robust basis for evaluating financial knowledge and reasoning.

We processed the corpus using a three-stage pipeline: extraction, cleaning, and standardization. (1) Extraction: We used OCR to convert all text and mathematical equations from the source materials into a machine-readable Markdown format. (2) Cleaning: We then manually curated the extracted content, removing non-essential sections (e.g., prefaces, appendices) and performing quality assurance checks. (3) Standardization: Finally, we transformed the cleaned content into a structured format suitable for automated processing. This rigorous process ensures the final dataset is of high quality, integrity, and utility.

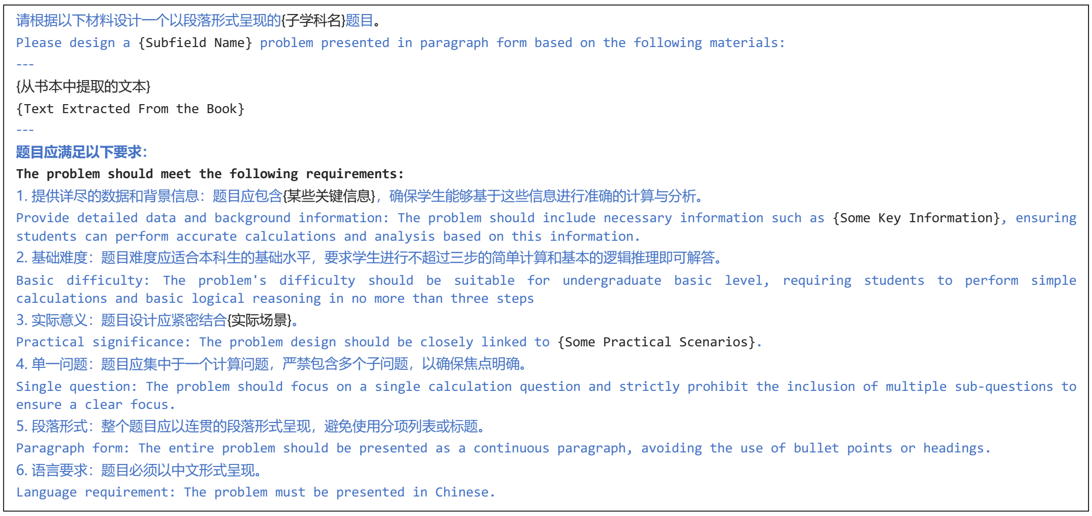
Corporate Finance

- Selected Textbook: *Corporate Finance* (13th edition, 2021) by Stephen A. Ross, Randolph W. Westerfield, Jeffrey Jaffe, and Bradford D. Jordan.

- Rationale: This textbook is widely used in MBA and undergraduate finance courses. It systematically explains core concepts of modern corporate finance, such as arbitrage theory, net present value (NPV), the efficient market hypothesis, agency theory, and the risk-return tradeoff.

- Covered Financial Subfields: Corporate financing, capital structure, investment decisions, dividend policy, firm valuation, etc.

- Role in the Benchmark Dataset: Provides a solid theoretical foundation and abundant practical examples for reasoning and computational questions in corporate finance.

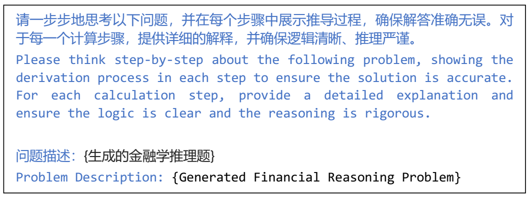


请根据以下材料设计一个以段落形式呈现的{子学科名}题目。
Please design a {Subfield Name} problem presented in paragraph form based on the following materials:

{从书本中提取的文本}
{Text Extracted From the Book}

题目应满足以下要求：
The problem should meet the following requirements:
1. 提供详尽的数据和背景信息：题目应包含{某些关键信息}，确保学生能够基于这些信息进行准确的计算与分析。
Provide detailed data and background information: The problem should include necessary information such as {Some Key Information}, ensuring students can perform accurate calculations and analysis based on this information.
2. 基础难度：题目难度应适合本科生的基础水平，要求学生进行不超过三步的简单计算和基本的逻辑推理即可解答。
Basic difficulty: The problem's difficulty should be suitable for undergraduate basic level, requiring students to perform simple calculations and basic logical reasoning in no more than three steps
3. 实际意义：题目设计应紧密结合{实际场景}。
Practical significance: The problem design should be closely linked to {Some Practical Scenarios}.
4. 单一问题：题目应集中于一个计算问题，严禁包含多个子问题，以确保焦点明确。
Single question: The problem should focus on a single calculation question and strictly prohibit the inclusion of multiple sub-questions to ensure a clear focus.
5. 段落形式：整个题目应以连贯的段落形式呈现，避免使用分项列表或标题。
Paragraph form: The entire problem should be presented as a continuous paragraph, avoiding the use of bullet points or headings.
6. 语言要求：题目必须以中文形式呈现。
Language requirement: The problem must be presented in Chinese.

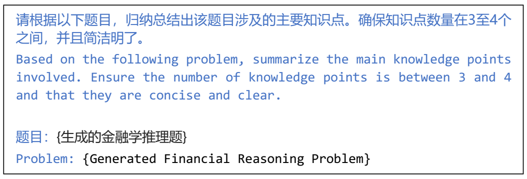
Figure 6: Prompt template for generating questions for a given subfield based on a given piece of corpus (original in Chinese, with English translation).



请一步步地思考以下问题，并在每个步骤中展示推导过程，确保解答准确无误。对于每一个计算步骤，提供详细的解释，并确保逻辑清晰、推理严谨。
Please think step-by-step about the following problem, showing the derivation process in each step to ensure the solution is accurate. For each calculation step, provide a detailed explanation and ensure the logic is clear and the reasoning is rigorous.

问题描述：{生成的金融学推理题}
Problem Description: {Generated Financial Reasoning Problem}

Figure 7: Prompt template that generates an solution to a given question (original in Chinese, with English translation).



请根据以下题目，归纳总结出该题目涉及的主要知识点。确保知识点数量在3至4个之间，并且简洁明了。
Based on the following problem, summarize the main knowledge points involved. Ensure the number of knowledge points is between 3 and 4 and that they are concise and clear.

题目：{生成的金融学推理题}
Problem: {Generated Financial Reasoning Problem}

Figure 8: Prompt template for labeling knowledge points for a given question (original in Chinese, with English translation).

Investments

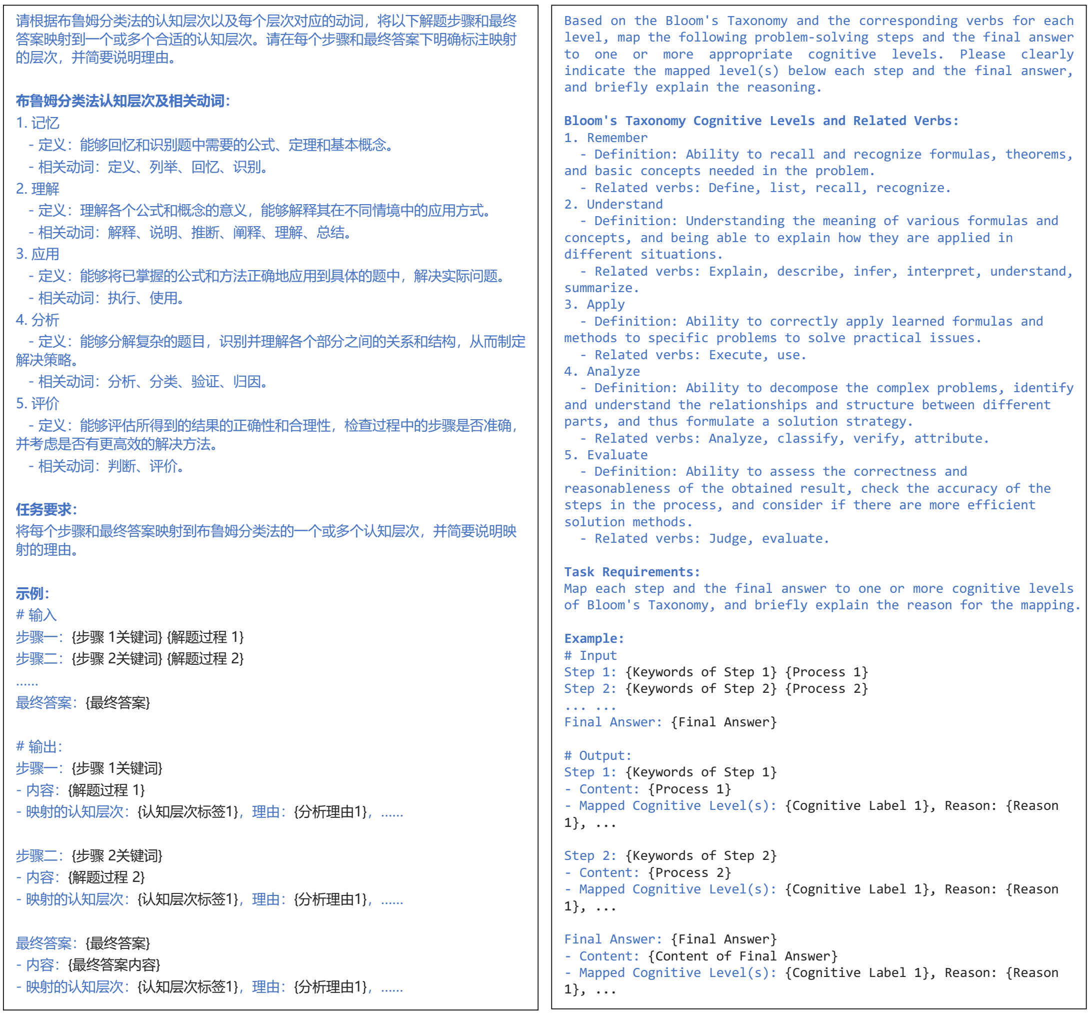
- Selected Textbook: *Investments* (13th edition, 2023) by Zvi Bodie, Alex Kane, and Alan J. Marcus.
- Rationale: This book deeply explores securities market efficiency, risk-return relationships, and asset allocation strategies. Its content is highly aligned with the CFA (Chartered Financial Analyst) exam syllabus.
- Covered Financial Subfields: Securities markets, asset pricing, portfolio theory, behavioral finance, derivatives, etc.
- Role in the Benchmark Dataset: Offers authoritative theoretical support and practical guidance for reasoning and computational questions in investments.

Financial Institutions and Markets

- Selected Textbook: *Financial Markets & Institutions* (13th edition, 2020) by Jeff Madura.
- Rationale: This book comprehensively analyzes the operational mechanisms and regulatory frameworks of financial institutions like commercial and investment banks. It also provides empirical and case analyses on contemporary hot topics such as stock valuation and market microstructure.
- Covered Financial Subfields: Financial institutions, financial markets, central banking, monetary policy, market regulation, etc.
- Role in the Benchmark Dataset: Supplies a systematic theoretical framework and practical examples for reasoning and computational questions concerning financial institutions and markets.

Money and Banking

- Selected Textbook: *The Economics of Money, Banking, and Financial Markets* (13th edition, 2021) by Frederic S. Mishkin.
- Rationale: This book offers an in-depth analysis, from both theoretical and empirical perspectives, of money demand and supply, commercial banking operations and regulation,



```
请根据布鲁姆分类法的认知层次以及每个层次对应的动词，将以下解题步骤和最终答案映射到一个或多个合适的认知层次。请在每个步骤和最终答案下明确标注映射的层次，并简要说明理由。

布鲁姆分类法认知层次及相关动词：
1. 记忆
  - 定义：能够回忆和识别题中需要的公式、定理和基本概念。
  - 相关动词：定义、列举、回忆、识别。
2. 理解
  - 定义：理解各个公式和概念的意义，能够解释其在不同情境中的应用方式。
  - 相关动词：解释、说明、推断、阐释、理解、总结。
3. 应用
  - 定义：能够将已掌握的公式和方法正确地应用到具体的题中，解决实际问题。
  - 相关动词：执行、使用。
4. 分析
  - 定义：能够分解复杂的题目，识别并理解各个部分之间的关系和结构，从而制定解决策略。
  - 相关动词：分析、分类、验证、归因。
5. 评价
  - 定义：能够评估所得到的结果的正确性和合理性，检查过程中的步骤是否准确，并考虑是否有更高效的解决方法。
  - 相关动词：判断、评价。

任务要求：
将每个步骤和最终答案映射到布鲁姆分类法的一个或多个认知层次，并简要说明映射的理由。

示例：
# 输入
步骤一：{步骤 1关键词} {解题过程 1}
步骤二：{步骤 2关键词} {解题过程 2}
......
最终答案：{最终答案}

# 输出：
步骤一：{步骤 1关键词}
- 内容：{解题过程 1}
- 映射的认知层次：{认知层次标签1}，理由：{分析理由1}，......

步骤二：{步骤 2关键词}
- 内容：{解题过程 2}
- 映射的认知层次：{认知层次标签1}，理由：{分析理由1}，......

最终答案：{最终答案}
- 内容：{最终答案内容}
- 映射的认知层次：{认知层次标签1}，理由：{分析理由1}，......
```

```
Based on the Bloom's Taxonomy and the corresponding verbs for each level, map the following problem-solving steps and the final answer to one or more appropriate cognitive levels. Please clearly indicate the mapped level(s) below each step and the final answer, and briefly explain the reasoning.

Bloom's Taxonomy Cognitive Levels and Related Verbs:
1. Remember
  - Definition: Ability to recall and recognize formulas, theorems, and basic concepts needed in the problem.
  - Related verbs: Define, list, recall, recognize.
2. Understand
  - Definition: Understanding the meaning of various formulas and concepts, and being able to explain how they are applied in different situations.
  - Related verbs: Explain, describe, infer, interpret, understand, summarize.
3. Apply
  - Definition: Ability to correctly apply learned formulas and methods to specific problems to solve practical issues.
  - Related verbs: Execute, use.
4. Analyze
  - Definition: Ability to decompose the complex problems, identify and understand the relationships and structure between different parts, and thus formulate a solution strategy.
  - Related verbs: Analyze, classify, verify, attribute.
5. Evaluate
  - Definition: Ability to assess the correctness and reasonableness of the obtained result, check the accuracy of the steps in the process, and consider if there are more efficient solution methods.
  - Related verbs: Judge, evaluate.

Task Requirements:
Map each step and the final answer to one or more cognitive levels of Bloom's Taxonomy, and briefly explain the reason for the mapping.

Example:
# Input
Step 1: {Keywords of Step 1} {Process 1}
Step 2: {Keywords of Step 2} {Process 2}
... ...
Final Answer: {Final Answer}

# Output:
Step 1: {Keywords of Step 1}
- Content: {Process 1}
- Mapped Cognitive Level(s): {Cognitive Label 1}, Reason: {Reason 1}, ...

Step 2: {Keywords of Step 2}
- Content: {Process 2}
- Mapped Cognitive Level(s): {Cognitive Label 1}, Reason: {Reason 1}, ...

Final Answer: {Final Answer}
- Content: {Content of Final Answer}
- Mapped Cognitive Level(s): {Cognitive Label 1}, Reason: {Reason 1}, ...
```

Figure 9: Prompt template for labeling step-level cognitive labels for a given answer (original in Chinese, with English translation).

and the interaction mechanisms between monetary policy tools and financial markets.

- Covered Financial Subfields: Monetary theory, banking systems, monetary policy, financial markets, etc.
- Role in the Benchmark Dataset: Delivers in-depth theoretical analysis and empirical support for reasoning and computational questions in money and banking.

International Finance

- Selected Textbook: *International Financial Management* (6th edition, 2023) by Jeff Madura and Roland Fox.
- Rationale: In the context of globalization, this textbook discusses cross-border capital flows, exchange rate volatility, and risk management strategies. It uses numerous case studies to examine practical operations in the international financial environment.
- Covered Financial Subfields: International capital flows, exchange rate theory, foreign exchange markets, international investment, multinational corporate financial management, etc.
- Role in the Benchmark Dataset: Provides a global perspective and practical examples for reasoning and computational questions in international finance.

Financial Risk Management

- Selected Textbook: *Risk Management and Financial Institutions* (6th edition, 2022) by John C. Hull.
- Rationale: This book comprehensively reviews methods for measuring and hedging

market risk, credit risk, and operational risk. It places particular emphasis on the application of financial derivatives in risk management.

- Covered Financial Subfields: Risk management, financial derivatives, financial institution regulation, risk measurement and hedging, etc.

- Role in the Benchmark Dataset: Offers a systematic risk analysis framework and practical guidance for reasoning and computational questions in financial risk management.

Fixed Income Securities

- Selected Textbook: *Fixed Income Securities: Tools for Today's Markets* (4th edition, 2022) by Bruce Tuckman and Angel Serrat.

- Rationale: This book provides detailed discussions on the pricing principles and trading strategies for fixed income products such as government bonds, interest rate swaps, and credit default swaps.

- Covered Financial Subfields: Fixed income securities, bond pricing, interest rate derivatives, credit risk, etc.

- Role in the Benchmark Dataset: Provides authoritative pricing models and practical examples for reasoning and computational questions related to fixed income securities.

Financial Engineering and Derivatives

- Selected Textbook: *Options, Futures, and Other Derivatives* (10th edition, 2018) by John C. Hull and Basu Sankarshan.

- Rationale: This textbook comprehensively covers core topics in financial engineering, including option pricing models, futures contract structures, and the pricing of interest rate and credit derivatives.

- Covered Financial Subfields: Derivatives markets, option pricing, futures contracts, interest rate derivatives, credit derivatives, etc.

- Role in the Benchmark Dataset: Offers in-depth theoretical analysis and practical guidance for reasoning and computational questions in financial engineering and derivatives.

Monetary Theory and Policy

- Selected Textbook: *Monetary Theory and Policy* (4th edition, 2017) by Carl E. Walsh.

- Rationale: This book systematically explains the framework of modern monetary theory, focusing on the transmission mechanisms of various monetary policy tools and their effectiveness in low-interest-rate environments.

- Covered Financial Subfields: Monetary theory, monetary policy, macroeconomic models, policy transmission mechanisms, etc.

- Role in the Benchmark Dataset: Provides a macroeconomic perspective and policy analysis framework for reasoning and computational questions in monetary theory and policy.

In summary, these nine textbooks are not only authoritative and reliable but also closely aligned with current academic frontiers. They lay a comprehensive and in-depth academic foundation for the financial reasoning and computation benchmark dataset constructed in this study. This ensures that the test questions possess both professional depth and practical relevance.

D Validation in Question and Answer Generation

For both the question and answer generation phases, we adopted a three-stage verification process. The verification focus for each stage is detailed in Table 4 and Table 5, respectively.

E Statistical Characterization of the FinEval-KR Dataset

A complete sample from the constructed dataset is shown in Figure 10. The sample size and its distribution for each subdiscipline in FinEval-KR Dataset is shown in Figure 11, where the subdiscipline categorization methodology refers to Zhang et al. (2023). The top-50 knowledge points in the dataset are shown in Figure 12. In each subdiscipline, the distribution of cognitive labels is shown in Figure 13.

F Prompt Template Adopted by the FinEval-KR Evaluation Framework

Figure 15 shows the prompt template used in Stage 1 where the evaluated model answers the questions

Stage	Aspect	Criteria
Logical Validation	Clarity and Completeness Check	(1) Is the question description clear and unambiguous? (2) Is there any misuse of terminology? (3) Are all necessary conditions and data for calculation provided?
	Plausibility Check	(1) Are numerical values (e.g., interest rates, returns, prices) within a plausible range? (2) Is the scenario self-contradictory or unrealistic?
	Solvability Check	Assuming the data is complete and plausible, does the question have a deterministic solution that can be calculated using financial models?
Consistency Assessment	Relevance Check	Do the core concepts in the question align with the title or keywords of the corresponding textbook chapter?
Final Validation	Samples that fail any of the above checks are marked as "unqualified" and removed from the final dataset.	

Table 4: Validation in question generation stage.

Stages	Aspect	Criteria
Logical Validation	Formula/Model Selection Check	(1) Is the selected formula a standard method for this type of problem? (2) Does the variable required by the question match the variable solved for in the answer?
	Parameter Substitution Check	Do the numerical values from the question correctly correspond to the variables in the formula during calculation?
Consistency Assessment	Calculation Validation	Execute the calculation code output by OpenAI o1 to verify the answer's correctness.
Final Validation	Samples that fail any of the above checks are marked as "unqualified" and removed from the final dataset.	

Table 5: Validation in answer generation stage.

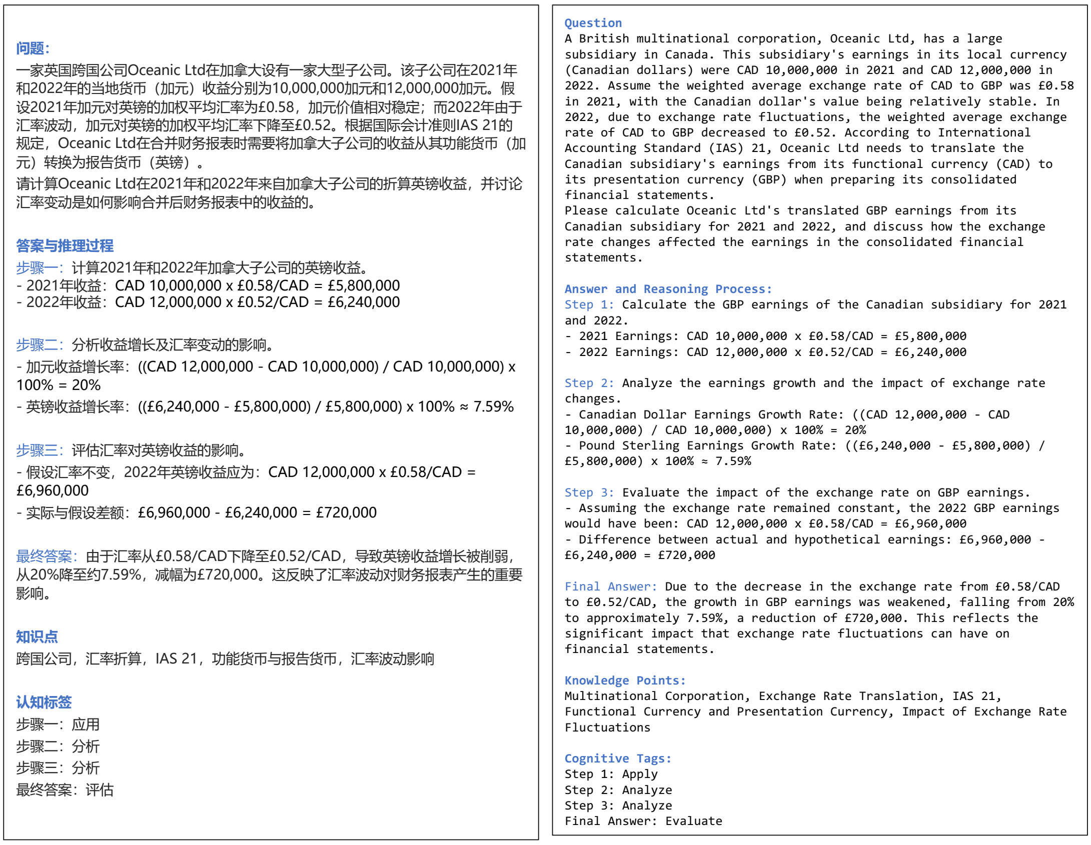
in a free-form format. Figure 16 shows the prompt template used in Stage 2, in which the evaluated model re-answer the question with knowledge point augmented. Figure 14 shows the prompt templates for the judge model to generate review results and augmented review results based on the reference answers in Stage 1 and 2.

G Bias Challenges in Judge Model

Previous research has shown that using LLMs as judges to evaluate the output of other models inevitably introduces certain evaluation biases, which may lead to unfair comparison results. Therefore, this section will discuss the main evaluation biases discovered during the experimental process of this study and their corresponding mitigation methods.

- *Style Bias*: This bias refers to the tendency of the judge LLM to give higher scores to content with a more appealing text style (e.g., clear structure, moderate length), even if the answers contain reasoning errors (Wu and Aji, 2025). To reduce the impact of this type of bias, we did not restrict the output format of the models being evaluated in Stage 1 and Stage 2, encouraging them to reason freely. Subsequently, we used methods such as regular expressions to unify the original output of each model into a format consistent with the reference answer. This processing method effectively reduced evaluation biases caused by differences in text style.

- *Cognitive Bias*: This bias refers to the self-bias that LLMs may exhibit during the evaluation, i.e., a tendency to give higher scores to content they generated themselves, thereby affecting the fairness of the evaluation (Koo et al., 2024). To avoid this type of cognitive bias, we excluded OpenAI o1 from the scope of evaluated models in our experiment, as it had already been used in the dataset construction and validation. Furthermore, during the preliminary experiment, we tested whether Qwen2.5-72B_Instruct, used as the judging model, exhibited significant cognitive bias in the root cause localization and knowledge gap identification tasks. The experimental results showed that this model did not exhibit a significant self-bias tendency in the aforementioned two tasks. We believe this may be because these two tasks are more specific and objective compared to result comparison tasks without sub-task decomposition, and are further aided by the model introspection prompting shown in Figure 14, which helps enhance the



问题:

一家英国跨国公司Oceanic Ltd在加拿大设有一家大型子公司。该子公司在2021年和2022年的当地货币（加元）收益分别为10,000,000加元和12,000,000加元。假设2021年加元对英镑的加权平均汇率为£0.58，加元价值相对稳定；而2022年由于汇率波动，加元对英镑的加权平均汇率下降至£0.52。根据国际会计准则IAS 21的规定，Oceanic Ltd在合并财务报表时需要将加拿大子公司的收益从其功能货币（加元）转换为报告货币（英镑）。

请计算Oceanic Ltd在2021年和2022年来自加拿大子公司的折算英镑收益，并讨论汇率变动是如何影响合并后财务报表中的收益的。

答案与推理过程

步骤一：计算2021年和2022年加拿大子公司的英镑收益。
- 2021年收益：CAD 10,000,000 x £0.58/CAD = £5,800,000
- 2022年收益：CAD 12,000,000 x £0.52/CAD = £6,240,000

步骤二：分析收益增长及汇率变动的影响。
- 加元收益增长率：((CAD 12,000,000 - CAD 10,000,000) / CAD 10,000,000) x 100% = 20%
- 英镑收益增长率：((£6,240,000 - £5,800,000) / £5,800,000) x 100% ≈ 7.59%

步骤三：评估汇率对英镑收益的影响。
- 假设汇率不变，2022年英镑收益应为：CAD 12,000,000 x £0.58/CAD = £6,960,000
- 实际与假设差额：£6,960,000 - £6,240,000 = £720,000

最终答案：由于汇率从£0.58/CAD下降至£0.52/CAD，导致英镑收益增长被削弱，从20%降至约7.59%，减幅为£720,000。这反映了汇率波动对财务报表产生的重要影响。

知识点

跨国公司，汇率折算，IAS 21，功能货币与报告货币，汇率波动影响

认知标签

步骤一：应用
步骤二：分析
步骤三：分析
最终答案：评估

Question

A British multinational corporation, Oceanic Ltd, has a large subsidiary in Canada. This subsidiary's earnings in its local currency (Canadian dollars) were CAD 10,000,000 in 2021 and CAD 12,000,000 in 2022. Assume the weighted average exchange rate of CAD to GBP was £0.58 in 2021, with the Canadian dollar's value being relatively stable. In 2022, due to exchange rate fluctuations, the weighted average exchange rate of CAD to GBP decreased to £0.52. According to International Accounting Standard (IAS) 21, Oceanic Ltd needs to translate the Canadian subsidiary's earnings from its functional currency (CAD) to its presentation currency (GBP) when preparing its consolidated financial statements.
Please calculate Oceanic Ltd's translated GBP earnings from its Canadian subsidiary for 2021 and 2022, and discuss how the exchange rate changes affected the earnings in the consolidated financial statements.

Answer and Reasoning Process:

Step 1: Calculate the GBP earnings of the Canadian subsidiary for 2021 and 2022.
- 2021 Earnings: CAD 10,000,000 x £0.58/CAD = £5,800,000
- 2022 Earnings: CAD 12,000,000 x £0.52/CAD = £6,240,000

Step 2: Analyze the earnings growth and the impact of exchange rate changes.
- Canadian Dollar Earnings Growth Rate: ((CAD 12,000,000 - CAD 10,000,000) / CAD 10,000,000) x 100% = 20%
- Pound Sterling Earnings Growth Rate: ((£6,240,000 - £5,800,000) / £5,800,000) x 100% ≈ 7.59%

Step 3: Evaluate the impact of the exchange rate on GBP earnings.
- Assuming the exchange rate remained constant, the 2022 GBP earnings would have been: CAD 12,000,000 x £0.58/CAD = £6,960,000
- Difference between actual and hypothetical earnings: £6,960,000 - £6,240,000 = £720,000

Final Answer: Due to the decrease in the exchange rate from £0.58/CAD to £0.52/CAD, the growth in GBP earnings was weakened, falling from 20% to approximately 7.59%, a reduction of £720,000. This reflects the significant impact that exchange rate fluctuations can have on financial statements.

Knowledge Points:

Multinational Corporation, Exchange Rate Translation, IAS 21, Functional Currency and Presentation Currency, Impact of Exchange Rate Fluctuations

Cognitive Tags:

Step 1: Apply
Step 2: Analyze
Step 3: Analyze
Final Answer: Evaluate

Figure 10: A complete sample from FinEval-KR dataset (original in Chinese, with English translation).

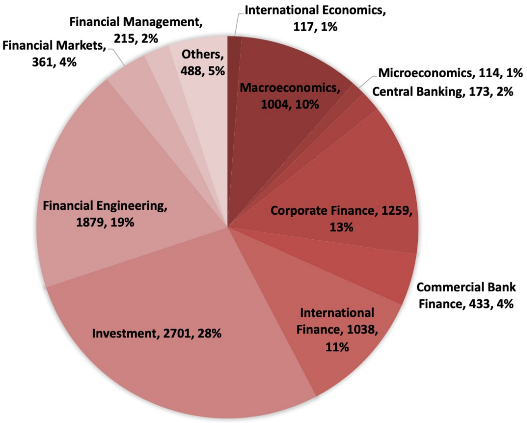


Figure 11: The number of samples in each subdiscipline in the FinEval-KR dataset and their percentage, and "others" in the pie chart includes: econometrics, public finance, insurance, monetary economics, managerial accounting, intermediate financial accounting, corporate strategy and risk management, auditing, cost accounting, taxation and advanced financial accounting.

objectivity of Qwen2.5-72B_Instruct during the evaluation.

- To prevent the judge model from the disturbance of "simple deception" (Thakur et al., 2024), we filter out meaningless content in the generated answers, such as isolated affirmative words like "yes" or "of course", ensuring that the evaluation focuses on the substantive reasoning process rather than superficial linguistic features.

H Details of Experiment

Evaluated Open-source Models We select several popular LLMs, including DeepSeek-R1, DeepSeek-V3, and QwQ-32B-preview. To study the effect of model size on performance, we also include smaller models, such as Qwen2.5-14B_Instruct. In total, this study includes four open-source LLMs.

Evaluated Close-source Models We include a selection of prominent models. This selection covers the latest reasoning models, such as Claude-3.7-sonnet, Gemini-2.5-flash, Gemini-2.5-pro, o3-mini, and o1-mini. We also include the current top non-

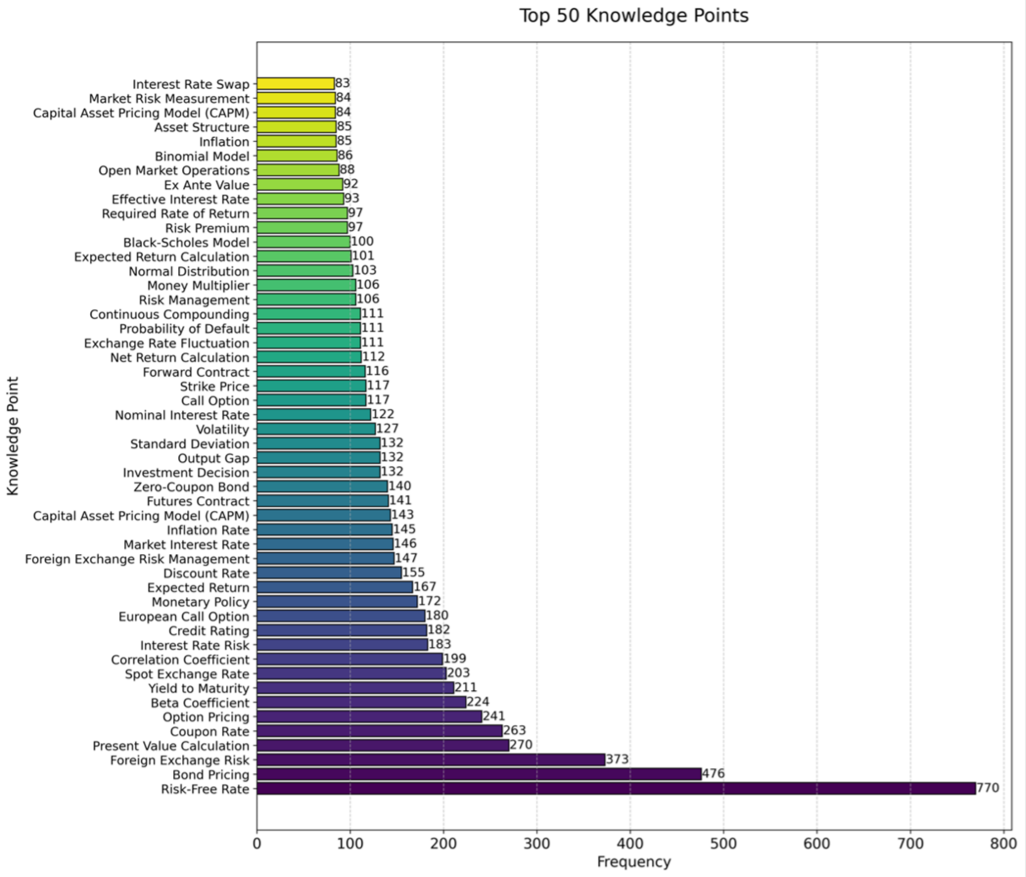


Figure 12: Top 50 knowledge points in the dataset.

reasoning models: GPT-4.1, GPT-4o, and Qwen-max-latest. Furthermore, we add some models released in 2024 and 2023, including Moonshot-v1-128k, Doubao-pro-32k, Ernie-Bot-4.0, and GPT-3.5-turbo.

Financial LLMs Furthermore, we specifically select two financial reasoning LLMs for evaluation. The first is Xuanyuan-FinX1-preview from Duxiaoman AI-Lab, a Chinese financial dialogue and reasoning model designed specifically for the financial domain. It is also the first o1-like model in the financial industry. The second is Fin-R1, a financial reasoning LLM jointly developed by Shanghai University of Finance and Economics and StepFun Technology. This model is trained on Qwen2.5-7B_Instruct and designed for complex financial reasoning tasks, balancing high performance with low deployment cost.

Implement During evaluation, all closed-source models are accessed through the official APIs provided by their respective developers. In contrast, open-source models are accessed using the service provided by either Bailian[3]or ModelScope[4].

Model	Version
Qwen2.5-14B_Instruct	2024-09-19
QwQ-32B-preview	2025-03-06
DeepSeek-V3	2025-03-24
DeepSeek-R1	2025-01-20
Doubao-pro-32k	2024-06-15
Moonshot-v1-128k	2024-01-31
Ernie-Bot-4.0	2023-11-17
Qwen-max-latest	2025-01-25
GPT-3.5-turbo	2024-01-25
GPT-4o	2024-11-20
GPT-4.1	2025-04-14
Gemini-2.5-pro	2025-05-06
Claude-3.7-sonnet	2025-02-19
o1-mini	2024-09-12
o3-mini	2025-01-31
Gemini-2.5-flash	2025-04-17
Xuanyuan-FinX1-preview	2024-12-27
Fin-R1	2025-03-22

Table 6: Version of the model being evaluated.

[3] https://bailian.console.aliyun.com

[4] https://www.modelscope.cn

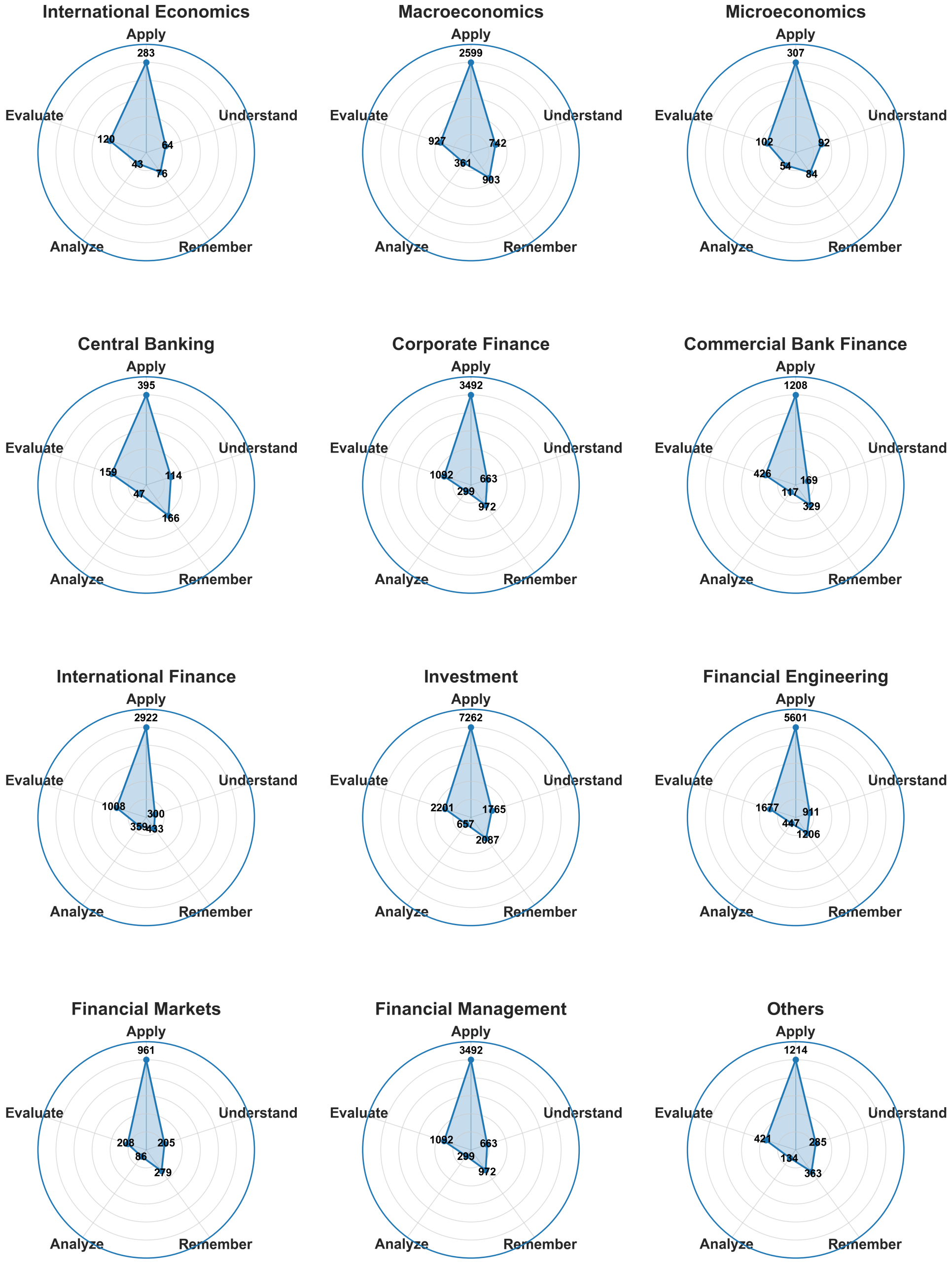


Figure 13: The distribution of cognitive labels in each subdiscipline.

I Discussion of Experimental Results

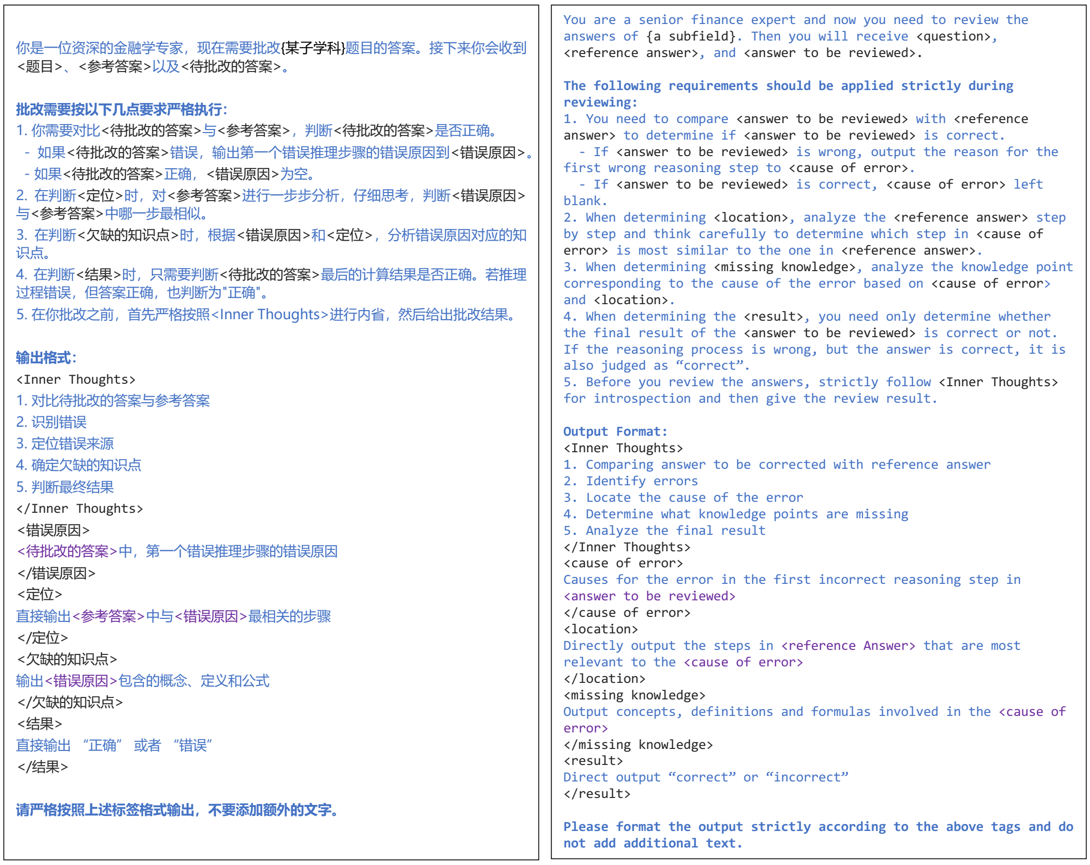
We evaluate 18 LLMs listed in Appendix H using our proposed financial reasoning dataset. Table 7 presents the complete evaluation results across several metrics: Knowledge Score (KS), Reasoning Score (RS), Cognitive Scores (CS_1 to CS_5), and Task Accuracy (Acc). Results in the table are from three runs of each model. The CS_1 to CS_5 correspond to remembering, understanding, applying, analyzing, and evaluating in Bloom's taxonomy, respectively.

For all subsequent analyses, our focus is on performance tiers instead of absolute scores, which helps alleviate assessment errors caused by the systematic bias and randomness of the judge model. Additionally, we set the distribution of models across the tiers to 3:4:5:6.

I.1 Analysis of Knowledge Score

The KS measures the breadth of a LLM's knowledge coverage in the financial domain. Based on the evaluation results, models fall into four tiers, as Figure 17 shows.

Tier 1 is exclusively composed of closed-source models that exhibit exceptionally high financial



你是一位资深的金融学专家，现在需要批改{某子学科}题目的答案。接下来你会收到<题目>、<参考答案>以及<待批改的答案>。

批改需要按以下几点要求严格执行：

1. 你需要对比<待批改的答案>与<参考答案>，判断<待批改的答案>是否正确。
   - 如果<待批改的答案>错误，输出第一个错误推理步骤的错误原因到<错误原因>。
   - 如果<待批改的答案>正确，<错误原因>为空。
2. 在判断<定位>时，对<参考答案>进行一步步分析，仔细思考，判断<错误原因>与<参考答案>中哪一步最相似。
3. 在判断<欠缺的知识点>时，根据<错误原因>和<定位>，分析错误原因对应的知识点。
4. 在判断<结果>时，只需要判断<待批改的答案>最后的计算结果是否正确。若推理过程错误，但答案正确，也判断为"正确"。
5. 在你批改之前，首先严格按照<Inner Thoughts>进行内省，然后给出批改结果。

输出格式：

<Inner Thoughts>
1. 对比待批改的答案与参考答案
2. 识别错误
3. 定位错误来源
4. 确定欠缺的知识点
5. 判断最终结果
</Inner Thoughts>
<错误原因>
<待批改的答案>中，第一个错误推理步骤的错误原因
</错误原因>
<定位>
直接输出<参考答案>中与<错误原因>最相关的步骤
</定位>
<欠缺的知识点>
输出<错误原因>包含的概念、定义和公式
</欠缺的知识点>
<结果>
直接输出 “正确” 或者 “错误”
</结果>

请严格按照上述标签格式输出，不要添加额外的文字。

You are a senior finance expert and now you need to review the answers of {a subfield}. Then you will receive <question>, <reference answer>, and <answer to be reviewed>.

The following requirements should be applied strictly during reviewing:

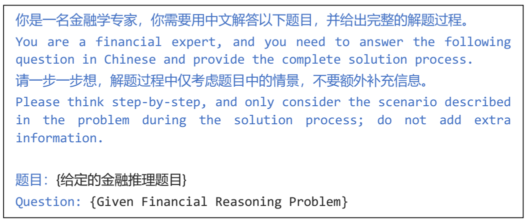
1. You need to compare <answer to be reviewed> with <reference answer> to determine if <answer to be reviewed> is correct.
   - If <answer to be reviewed> is wrong, output the reason for the first wrong reasoning step to <cause of error>.
   - If <answer to be reviewed> is correct, <cause of error> left blank.
2. When determining <location>, analyze the <reference answer> step by step and think carefully to determine which step in <cause of error> is most similar to the one in <reference answer>.
3. When determining <missing knowledge>, analyze the knowledge point corresponding to the cause of the error based on <cause of error> and <location>.
4. When determining the <result>, you need only determine whether the final result of the <answer to be reviewed> is correct or not. If the reasoning process is wrong, but the answer is correct, it is also judged as “correct”.
5. Before you review the answers, strictly follow <Inner Thoughts> for introspection and then give the review result.

Output Format:

<Inner Thoughts>
1. Comparing answer to be corrected with reference answer
2. Identify errors
3. Locate the cause of the error
4. Determine what knowledge points are missing
5. Analyze the final result
</Inner Thoughts>
<cause of error>
Causes for the error in the first incorrect reasoning step in <answer to be reviewed>
</cause of error>
<location>
Directly output the steps in <reference Answer> that are most relevant to the <cause of error>
</location>
<missing knowledge>
Output concepts, definitions and formulas involved in the <cause of error>
</missing knowledge>
<result>
Direct output “correct” or “incorrect”
</result>

Please format the output strictly according to the above tags and do not add additional text.

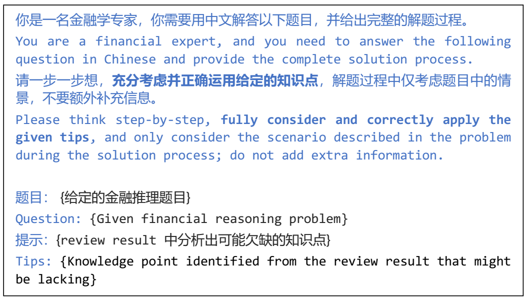
Figure 14: Prompt templates for the judge model (original in Chinese, with English translation).



你是一名金融学专家，你需要用中文解答以下题目，并给出完整的解题过程。
You are a financial expert, and you need to answer the following question in Chinese and provide the complete solution process.
请一步一步想，解题过程中仅考虑题目中的情景，不要额外补充信息。
Please think step-by-step, and only consider the scenario described in the problem during the solution process; do not add extra information.

题目：{给定的金融推理题目}
Question: {Given Financial Reasoning Problem}

Figure 15: Prompt template for Stage 1 where the evaluated model answers the questions in a free-form format (original in Chinese, with English translation).



你是一名金融学专家，你需要用中文解答以下题目，并给出完整的解题过程。
You are a financial expert, and you need to answer the following question in Chinese and provide the complete solution process.
请一步一步想，**充分考虑并正确运用给定的知识点**，解题过程中仅考虑题目中的情景，不要额外补充信息。
Please think step-by-step, **fully consider and correctly apply the given tips**, and only consider the scenario described in the problem during the solution process; do not add extra information.

题目： {给定的金融推理题目}
Question: {Given financial reasoning problem}
提示：{review result 中分析出可能欠缺的知识点}
Tips: {Knowledge point identified from the review result that might be lacking}

Figure 16: Prompt template for Stage 2 where the evaluated model re-answer the question with knowledge point augmented (original in Chinese, with English translation).

knowledge coverage. Tier 2 is dominated by the top-performing open-source models. Although they rank just below Tier 1, the absolute score difference is marginal, indicating that their financial knowledge coverage is nearly on par with the leading closed-source models.

A significant performance drop-off occurs in the lower tiers. In Tiers 3 and 4, the older GPT-3.5-turbo notably outperforms other models within this bracket. At the bottom of the ranking is the specialized financial model, Fin-R1-7B, whose lower performance is primarily attributed to its significantly smaller parameter scale.

In summary, leading closed-source and top open-source reasoning models demonstrate the strongest performance in financial knowledge coverage, which is significantly influenced by model scale. While financial knowledge is a mature capability in most mainstream LLMs and no longer the primary differentiator among top models, it remains a fundamental prerequisite for high-quality reasoning.

Model/Metrics	Acc	Acc.std	KS	KS.std	RS	RS.std	CS_1 (remember)	CS_1.std	CS_2 (understand)	CS_2.std	CS_3 (apply)	CS_3.std	CS_4 (analyze)	CS_4.std	CS_5 (evaluate)	CS_5.std
Open-source lightweight LLMs without reasoning																
Qwen2.5-14B_Instruct	0.5473	0.0006	0.8490	0.0010	0.6863	0.0038	0.6547	0.0015	0.6603	0.0023	0.3893	0.0064	0.6863	0.0038	0.6820	0.0046
Open-source lightweight LLMs with reasoning																
QwQ-32B-preview	0.7380	0.0061	0.9073	0.0057	0.8627	0.0136	0.8450	0.0141	0.8503	0.0143	0.6987	0.0267	0.8510	0.0075	0.8597	0.0136
Open-source LLMs without reasoning																
DeepSeek-v3	0.8270	0.0062	0.9427	0.0050	0.9077	0.0050	0.8963	0.0059	0.8993	0.0059	0.7963	0.0125	0.8943	0.0075	0.9057	0.0057
Open-source LLMs with reasoning																
DeepSeek-R1	0.8700	0.0165	0.9517	0.0171	**0.9347**	0.0153	**0.9377**	0.0186	**0.9397**	0.0179	**0.8810**	0.0358	**0.9380**	0.0190	**0.9433**	0.0158
Closed-source LLMs without reasoning																
Doubao-pro-32k	0.7825	0.0007	0.9195	0.0007	0.8750	0.0057	0.8560	0.0071	0.8600	0.0085	0.7340	0.0113	0.8565	0.0064	0.8720	0.0057
Moonshot-v1-128k	0.4533	0.0015	0.8340	0.0061	0.6020	0.0082	0.5620	0.0108	0.5670	0.0087	0.2763	0.0064	0.5653	0.0074	0.5973	0.0074
Ernie-bot-4.0	0.5733	0.0025	0.8627	0.0081	0.7053	0.0091	0.6680	0.0089	0.6753	0.0091	0.4383	0.0146	0.6847	0.0074	0.6927	0.0251
Qwen-max-latest	0.6467	0.0015	0.8797	0.0050	0.7733	0.0042	0.7507	0.0050	0.7547	0.0057	0.5340	0.0040	0.7440	0.0026	0.7703	0.0042
GPT-3.5-turbo	0.2830	0.0040	0.7527	0.0038	0.3973	0.0040	0.3527	0.0021	0.3603	0.0021	0.0900	0.0036	0.3893	0.0081	0.3970	0.0036
GPT-4o	0.6853	0.0142	0.9020	0.0159	0.8067	0.0080	0.7847	0.0081	0.7890	0.0090	0.5930	0.0145	0.7870	0.0110	0.8030	0.0085
GPT-4.1	0.8263	0.0025	0.9520	0.0040	0.9063	0.0015	0.8957	0.0021	0.8977	0.0025	0.7890	0.0036	0.8927	0.0015	0.9050	0.0017
Closed-source LLMs with reasoning																
o1-mini	0.7503	0.0031	0.8997	0.0076	0.8453	0.0067	0.8340	0.0066	0.8363	0.0031	0.6983	0.0081	0.8477	0.0070	0.8450	0.0060
o3-mini	0.8207	0.0095	0.9260	0.0106	0.9070	0.0052	0.9047	0.0102	0.9073	0.0099	0.8127	0.0110	0.9023	0.0107	0.9120	0.0113
Gemini-2.5-pro	**0.8750**	0.0079	**0.9627**	0.0134	0.9233	0.0238	0.9123	0.0272	0.9163	0.0290	0.8403	0.0291	0.9050	0.0260	0.9120	0.0243
Gemini-2.5-flash	0.8440	0.0061	0.9540	0.0020	0.9203	0.0091	0.9103	0.0100	0.9133	0.0108	0.8307	0.0061	0.9100	0.0104	0.9177	0.0110
Claude-3.7-sonnet	0.7923	0.0040	0.9390	0.0030	0.8823	0.0086	0.8663	0.0100	0.8703	0.0096	0.7433	0.0120	0.8653	0.0093	0.8803	0.0086
Financial LLMs with reasoning																
Xuanyuan-FinX1-preview	0.5890	0.0026	0.8687	0.0032	0.7323	0.0042	0.7063	0.0032	0.7130	0.0044	0.4610	0.0066	0.7323	0.0042	0.7300	0.0035
Fin-R1-7B	0.4153	0.0031	0.7510	0.0346	0.5570	0.0040	0.5190	0.0046	0.5277	0.0065	0.2170	0.0036	0.5570	0.0040	0.5527	0.0045

Table 7: The complete evaluation results across several metrics: Knowledge Score (KS), Reasoning Score (RS), Cognitive Scores (CS_1 to CS_5), and Task Accuracy (Acc).

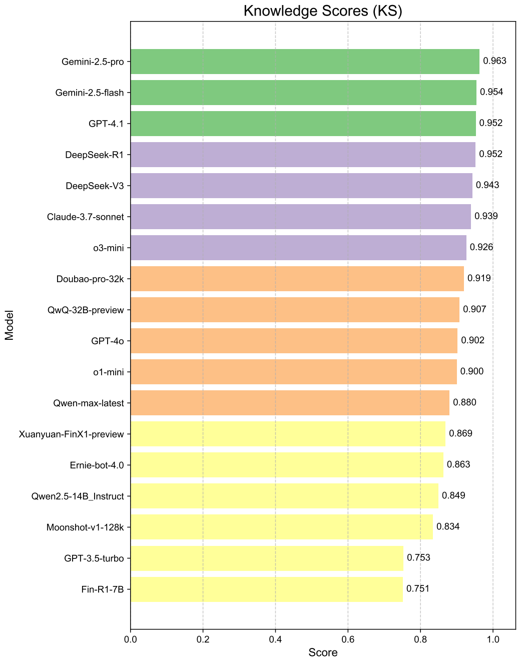


Figure 17: The knowledge score of the models.

I.2 Analysis of Reasoning Score

The RS is inversely related to the proportion of failures caused by incorrect reasoning steps. It reflects a model's reasoning ability. Based on this metric, the evaluated models are also divided into four tiers, as Figure 18 shows.

Tier 1 represents the pinnacle of performance, comprising reasoning-optimized models that demonstrate outstanding accuracy and logical completeness. Tier 2 includes some high-performing, non-reasoning models like GPT-4.1 and DeepSeek-V3. A significant performance gap separates the top two tiers from the bottom two. This clear stratification underscores the need for future model development to prioritize the design and optimization of the reasoning pipeline, which is crucial for enhancing the reliability and stability of complex reasoning tasks.

I.3 Analysis of Cognitive Score

The CS provides a systematic evaluation of models' cognitive abilities based on Bloom's Taxonomy, across five dimensions, that is remembering (CS_1), understanding (CS_2), applying (CS_3), analyzing (CS_4), and evaluating (CS_5). As Table 7 shows, CS scores generally exhibit a positive correlation with the KS and the RS.

While most models achieve high scores at lower cognitive levels (CS_1: Remembering, CS_2: Understanding), their performance diverges significantly on higher-order tasks. These more demanding abilities—Applying (CS_3), Analyzing (CS_4), and Evaluating (CS_5)—reveal clear distinctions among the models. Consequently, our analysis focuses on these three dimensions. We define a primary metric, CS_{avg}, as the average score across these higher-order skills, and stratify the models into four performance tiers based on this metric (see Figure 19).

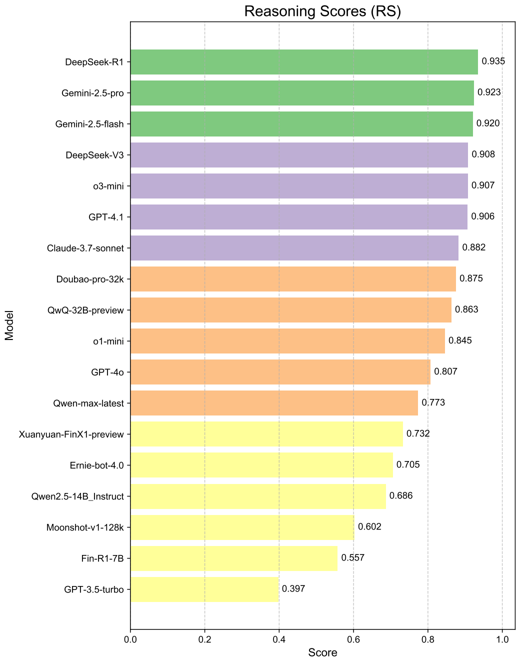


Figure 18: The reasoning score of the models.

Tier 1 models excel across all cognitive levels, demonstrating a distinct advantage in higher-order abilities. This tier is led by DeepSeek-R1, followed by Gemini-2.5-flash and pro. Tier 2 models also exhibit strong higher-order cognitive skills, with performance slightly below that of Tier 1. This tier includes most general-purpose reasoning models as well as the top-performing non-reasoning models, DeepSeek-V3 and GPT-4.1. Tiers 3 and 4 primarily consist of non-reasoning or smaller-scale models.

I.4 Analysis of Task Accuracy

Task Accuracy measures a model's direct success rate in executing reasoning tasks. Achieving high accuracy requires a synthesis of a broad knowledge base, robust reasoning capabilities, and advanced cognitive skills—particularly in application and analysis. Consequently, the performance gradient observed in Task Accuracy closely mirrors those of the RS and CS. The tiers of models based on this metric are shown in Figure 20.

I.5 Variance Analysis

We evaluate a model's performance stability by the standard deviation of its scores across multiple test runs. We classify stability into two categories: *High Stability* (a standard deviation on the order of 10^{-4} to 10^{-3}), indicating highly consistent and reproducible outputs, and *Low Stability* (an order of 10^{-2} to 10^{-1}), which suggests significant performance fluctuations.

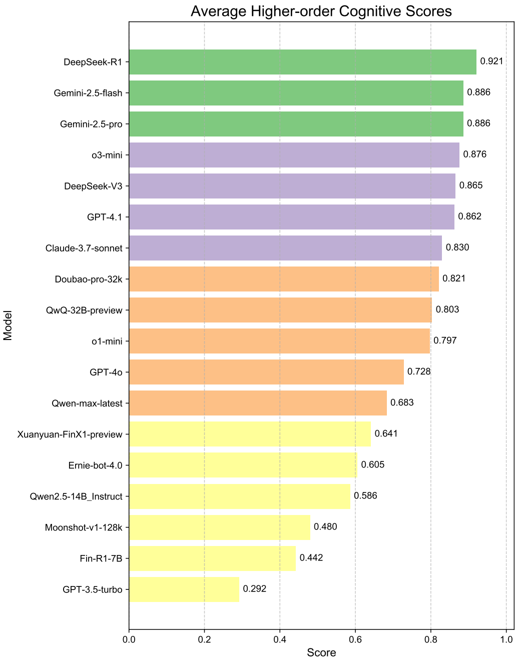


Figure 19: The average higher-order cognitive scores (CS_3 to CS_5) for the model.

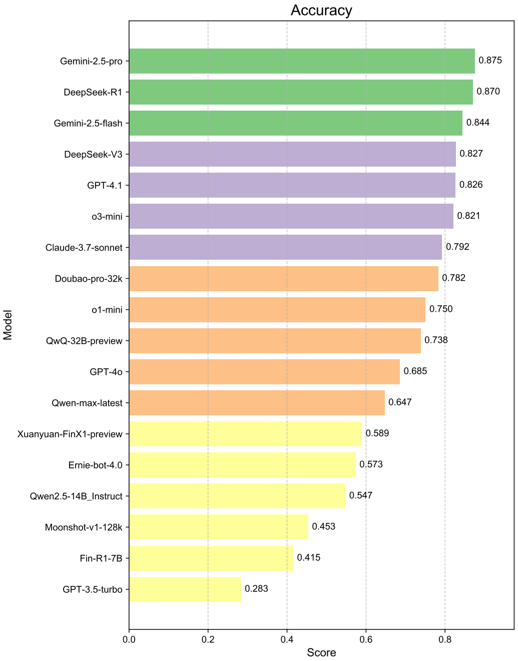


Figure 20: The accuracy of the models.

Our key findings are as follows:

- **Knowledge Retrieval is More Stable Than Reasoning.** For most models, the KS is consistently more stable than the RS. This is intuitive, as retrieving a stored fact is a more deterministic process for a well-trained model than performing a complex, multi-step logical deduction, which allows for greater variability.

- **GPT-4o is a Unique Outlier.** GPT-4o defies the general trend. Its reasoning process is remarkably stable, with an RS standard deviation of 8×10^{-3}, which is significantly more stable than its knowledge retrieval (KS standard deviation of 2×10^{-2}). We hypothesize that GPT-4o may possess a highly consistent, almost programmatic reasoning structure, while its knowledge function exhibits greater variance to adapt to diverse queries. This unusual stability profile warrants further investigation.

J Details of Human Evaluators and Validation Process

J.1 Evaluator Qualifications and Number

A total of 30 human experts participated in our validation effort. All experts are postgraduate students with academic backgrounds in finance, economics, or statistics, ensuring they possess an accurate understanding of the relevant professional terminology, fundamental concepts, and practical scenarios.

The entire validation process was conducted on a professional annotation platform provided by a leading technology company to ensure procedural standardization and data security.

J.2 Quality Control Mechanism

To guarantee the reliability of our validation results, we implemented a multi-stage quality control process. First, we randomly sampled 10% of the dataset. Each sample was independently validated by 2 experts to ensure consistency. Following this cross-validation, we organized a team of 3 senior experts to conduct a final quality check on a random 10% of the already-validated sample (amounting to a final check on 1% of the total dataset). This final step was designed to ensure the quality and uniformity of the standards applied during the cross-validation stage.

J.3 The Validation Process

The experts' validation work was divided into three strict, sequential stages:

Stage 1: Question and Knowledge Point Validation In this initial stage, experts were only shown the question and its associated knowledge points. They were tasked with the following checks:

- **Question Validity**: Is the question relevant to a realistic financial scenario? Is professional terminology used correctly? Are the numerical values within a reasonable range? Does the question have a single, definitive answer?

- **Knowledge Point Relevance**: Are the tagged knowledge points accurate and comprehensive? Is the naming of the knowledge points consistent with standard terminology in mainstream textbooks?

Independent Answering After confirming the quality of the question, experts were required to solve the problem independently, without reference to any provided solution. The goal of this step was to obtain a high-quality, unbiased human answer to serve as a benchmark for subsequent comparisons.

Stage 3: Reasoning Steps and Cognitive Labels Validation Finally, the system presented the experts with the answer, the step-by-step solution, and the Bloom's Taxonomy cognitive label from our dataset. The experts were required to perform the following checks:

- **Answer and Solution Process Verification**: First, they compared their own answer to the one in the dataset. If the answers did not match, the sample was immediately flagged as "unqualified". If the answers matched, they proceeded to meticulously review the solution steps provided in the dataset, assessing whether the logic was clear, the steps were reasonable, and the calculations were correct.

- Cognitive Label Accuracy Check: Based on the predefined verb list corresponding to each cognitive level (as defined in Figure 9), the experts had to judge whether the cognitive label assigned to the question was accurate.

K License and Usage Constraints

The released dataset is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

The dataset is only for for evaluating LLMs in non-commercial academic research. The dataset is explicitly not authorized for use in training or fine-tuning machine learning models , including pre-training, instruction-tuning, or reinforcement learning stages. Access conditions ensure that all derived data products remain confined to research contexts, with no transfer or application permitted in industrial, governmental, or other operational domains.

L AI Assistants Usage Disclosure

This study did not employ any AI assistants during the research design, data analysis, or coding phases. During manuscript preparation, the authors exclusively utilized Google Gemini for grammatical refinement and stylistic polishing. No AI-generated content was incorporated into the methodology and results, ensuring the work's originality and human-driven intellectual integrity.