XFINBENCH: Benchmarking LLMs in Complex Financial Problem Solving and Reasoning

Anonymous ACL submission

Abstract

Solving financial problems demands complex reasoning, multimodal data processing, and a broad technical understanding, presenting unique challenges for current large language models (LLMs). We introduce **XFINBENCH**, a novel benchmark with 4,235 examples designed to evaluate LLM's ability in solving comple**X**, knowledge-intensive **Fin**ancial problems across diverse graduate-level finance topics with multi-modal context. We identify five core capabilities of LLMs using XFINBENCH, *i.e*, *terminology understanding*, *temporal reasoning*, *future forecasting*, *scenario planning*, and *numerical modelling*. Upon XFINBENCH, we conduct extensive experiments on 18 leading models. The result shows that o1 is the best-performing text-only model with an overall accuracy of 67.3%, but still lags significantly behind human experts with 12.5%, especially in *temporal reasoning* and *scenario planning* capabilities. We further construct a knowledge bank with 3,032 finance terms for knowledge augmentation analysis, and find that relevant knowledge to the question only brings consistent accuracy improvements to small open-source model. Additionally, our error analysis reveals that rounding errors during calculation and blindness to position and intersection of curves in the image are two primary issues leading to model's poor performance in calculating and visual-context questions, respectively. [1]

1 Introduction

Finance constitutes a critical domain, characterized by the necessity for sophisticated problem-solving skills. Beyond domain-specific knowledge, it necessitates advanced capabilities such as temporal reasoning (Su et al., 2024; Wang and Zhao, 2024), future forecasting (Jin et al., 2024; Zhou et al., 2023), scenario planning (Valmeekam et al., 2022; Geva et al., 2021), and numerical modeling (Zhao et al., 2024; Kedziorski et al., 2024). Besides, complex finance problems in real world usually involves rich multimodal information, covering time series (Yu et al., 2023), long tabular (Reddy et al., 2024) and various charts (Masry et al., 2022; Lu et al., 2024). These complexities present significant challenges for large language models (LLMs), thereby rendering finance an appropriate testbed for the evaluation of LLMs.

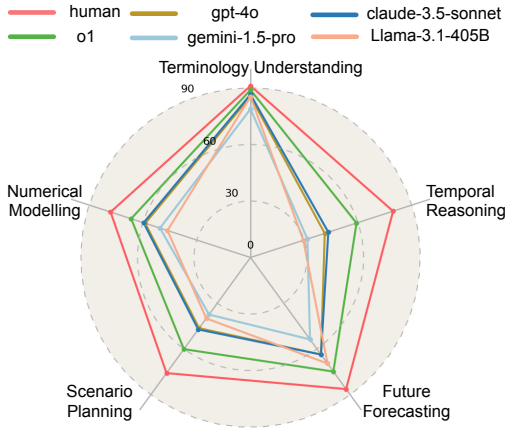


Figure 1: Evaluation result of leading LLMs and human experts on XFINBENCH across five capabilities for complex finance problem-solving. Results of o1 and Llama-3.1-405B do not cover visual-context questions.

Numerous datasets have been curated to assess the reasoning abilities of AI systems in the finance domain, with most emphasizing quantity extraction and basic mathematical reasoning (see Table 1). Existing datasets, including TAT-QA (Zhu et al., 2021), FinQA (Chen et al., 2021), Multi-Hiertt (Zhao et al., 2022), PACIFIC (Deng et al., 2022) and ConvFinQA (Chen et al., 2022), primarily focus on quantity extraction and basic numerical calculations using company's financial reports. However, they lack questions that entail extensive financial knowledge or complex reasoning

[1] Code and dataset are accessible via Anonymous GitHub: https://anonymous.4open.science/r/XFinBench-2D71.

processes. More recently, some benchmarks have been introduced to assess the performance of LLMs on knowledge-intensive finance tasks. For instance, BizBench (Kedziorski et al., 2024) collects examples from finance certificate examinations and existing datasets to test LLMs' business and financial understanding; FinanceMATH (Zhao et al., 2024) emphasizes LLMs' mathematical reasoning and code completion abilities within the finance domain; and FinEval (Zhang et al., 2023) focuses on model's understanding of finance concepts in Chinese. Nevertheless, these benchmarks still do not address the advanced capabilities necessary for solving complex financial problems like temporal reasoning, forecasting, and planning. Moreover, existing multi-modal datasets within the finance domain (Yue et al., 2024; Wang et al., 2024b) primarily assess models' visual recognition abilities, overlooking domain-specific reasoning that derives meaningful insights from financial figures (§A.3).

To bridge this gap, we introduce XFINBENCH, a novel benchmark specifically designed to evaluate LLM's ability in solving **complex**, **knowledge-intensive financial problems** across diverse financial topics with multi-modal context. XFINBENCH consists of 4,235 examples derived from graduate-level finance textbooks that ensures the complexity of financial problems in our dataset, and brings convenience to annotation of ground-truth knowledge to each problem. Different from existing datasets that only evaluate the model's grasp of specialized financial vocabulary, *i.e*, *Terminology Understanding*, XFINBENCH identifies four more advanced capabilities essential for complex finance problem-solving (§A and Figure 2): (1) *Temporal Reasoning*, involving the comprehension of time-based data and temporal relationships (§A.2.2); (2) *Future Forecasting*, testing logical reasoning in predicting financial trends based on theoretical finance models (§A.2.3); (3) *Scenario Planning*, analyzing different potential future scenarios to assess their impact on financial decisions and strategies (§A.2.4); and (4) *Numerical Modelling*, which involves constructing structured representations of companies and products' financial performance (§A.2.5). Moreover, XFINBENCH includes three tasks: *statement judging*, which evaluates the model's understanding of finance concepts; *multi-choice question answering*, which assesses strategic decision-making and predictive capabilities with visual data; and *financial calculation*, which tests mathematical reasoning in finance. To further investigate how domain-specific knowledge could boost LLM's performance on our complex financial problems, we also develop a knowledge bank with 3,032 finance terms, which is integrated with financial problems through human annotation.

We conduct extensive experiments on XFINBENCH to evaluate the complex finance problem-solving ability of 18 leading LLMs , along with knowledge augmentation analysis and error analysis. We implement Chain-of-Thought (CoT) method for all three tasks, and additionally apply Program-of-Thought (PoT) for *financial calculation*. Moreover, we establish a human performance baseline of human experts with finance degree. Our results indicate that o1 is the best-performing text-only model with an overall accuracy of 67.3%, while claude-3.5-sonnet achieves the highest accuracy of 64.0% when visual-context questions included (§3.2). Despite that LLMs achieve comparable performance with human in *terminology understanding*, they significantly lag behind human experts in more advanced capabilities for complex finance problem-solving, including *temporal reasoning* and *scenario planning*—especially when visual context is involved (Figure 1). These findings highlight that XFINBENCH represents a rigorous and challenging benchmark, offering a critical tool for advancing the development of LLMs in complex financial problem-solving and reasoning.

Our contributions are summarized as follows:

- We propose XFINBENCH, a novel benchmark designed to evaluate LLM's ability in solving complex, knowledge-intensive financial problems with multi-modal context (§2).
- We conduct extensive experiments on 18 leading LLMs and compare them with human-expert performance across five capabilities essential for complex finance problem solving (§3.2).
- We design three retrieving strategies for knowledge augmentation (§3.3), and identify multiple error types in challenging finance tasks (§3.4).

2 Dataset Construction

Our benchmark, XFINBENCH, is meticulously designed to facilitate complex reasoning in knowledge-intensive financial tasks. The dataset construction begins with the collection of questions and answers from three graduate-level finance textbooks and their solution manuals, accompanied by the creation of a knowledge bank of financial terms.

Dataset	Size	Task	Modality	Knowledge-intensive	Math-Reasoning	Complex-Problem	Source
TAT-QA (Zhu et al., 2021)	16,552	Quantity Extraction	Tabular	✗	✓	✗	Financial Report w. CrowdSource
PACIFIC (Deng et al., 2022)	2,757	Quantity Extraction	Tabular	✗	✓	✗	Existing dataset w. Automatic Pipeline
FinQA (Chen et al., 2021)	8,281	Quantity Extraction	Tabular	✗	✓	✗	Financial Report w. CrowdSource
ConvFinQA (Chen et al., 2022)	3,892	Quantity Extraction	Tabular	✗	✓	✗	Existing dataset w. CrowdSource
FinEval (Zhang et al., 2023)	4,661	Multi-choice QA	None	✓	✗	✗	Chinese Textbook
BizBench (Kedziorski et al., 2024)	19,842	Quantity Extraction Multi-choice QA	Tabular	✓	✓	✗	Existing Dataset, Certificate Exams
FinanceMATH (Zhao et al., 2024)	1,259	Financial Calculation	Tabular	✓	✓	Partial	Internet w. CrowdSource
XFINBENCH (ours)	4,235	Statement Judging Multi-choice QA Financial Calculation	Tabular, Image	✓	✓	✓	Textbook w. CrowdSource and GPT-4o

Table 1: Comparison of XFINBENCH with existing datasets.

To enrich the dataset, human experts annotate each question-answer pair with relevant financial terms and associated capabilities. Given the evaluation challenges posed by open-ended questions from textbooks, we leverage GPT-4o within a generate-then-verify framework to expand the dataset and enhance its suitability for assessing LLMs. Lastly, a rigorous quality validation process, conducted by human experts, ensures the dataset meets the highest standards of accuracy and relevance.

2.1 Initial Data Collection

Collection of Initial QA datasets. To ensure the complexity and knowledge-intensive nature of our benchmark, we extract after-class questions from three renowned graduate-level finance textbooks that cover most finance topics: *Fundamentals of Corporate Finance*, *Options Futures and Other Derivative*, and *The Economics of Money Banking and Financial Markets*. These textbooks and their solution manuals are sourced from publicly available platforms on the Internet, with strict adherence to copyright and licensing regulations. We utilize OCR techniques via the `pdfplumber` library to extract text from the downloaded PDFs. Three annotators are assigned to collect after-class questions at the end of each chapter, and capture screenshots for any accompanying visual or tabular context. Tabular data is subsequently formatted into LaTeX using GPT-4o. In total, we compile 2,018 after-class questions from textbooks, including 343 questions with visual or tabular context.

We then classify after-class questions collected from textbooks into three tasks: *statement judging*, *multi-choice question answering*, and *financial calculation*. Questions that evaluate the basic understanding of finance concepts and theoretical models are classified into *statement judging* task. Questions that focus on the application of financial strategies and models are classified into *multi-choice question answering* task. Some questions may be classified into both two tasks. For questions that involve numerical reasoning, we classify them into *financial calculation* task. Finally, 813 questions belong to the *statement judging* task, 624 to the *multi-choice question answering* task, and 858 to the *financial calculation* task (see §B.2).

Collection of Knowledge Bank. We construct a knowledge bank of finance terms and their definitions to facilitate knowledge augmentation analysis during evaluation. Using the subject index at the end of each textbook, we identify all finance terms in it along with their corresponding page ranges. Our annotators then manually extract the definitions of these terms from the specified pages. Notably, some terms share the same pages, resulting in shared definitions. Ultimately, our knowledge bank includes 3,032 finance terms with 1,766 unique definitions (see §B.3).

Bridging QA and Knowledge Bank. We so far have collected after-class question-answer pairs and finance terms in each textbook, which are initially linked through chapters. In each chapter, a collection of finance terms is introduced in the main body, followed by after-class questions in the end. Our annotators are then instructed to annotate each after-class question with 1-to-3 most relevant finance terms from the main body of the same chapter. Finally, a question is annotated with 1.3 terms on average (see §B.3).

2.2 GPT-4o Enhanced Annotation

After-class questions from textbooks are mostly open-ended or consisting of a series of sub-questions, making it difficult to evaluate the model's response. For instance, the answer to the open-ended question "*Discuss the advantages and disadvantages of options and forward contracts*"

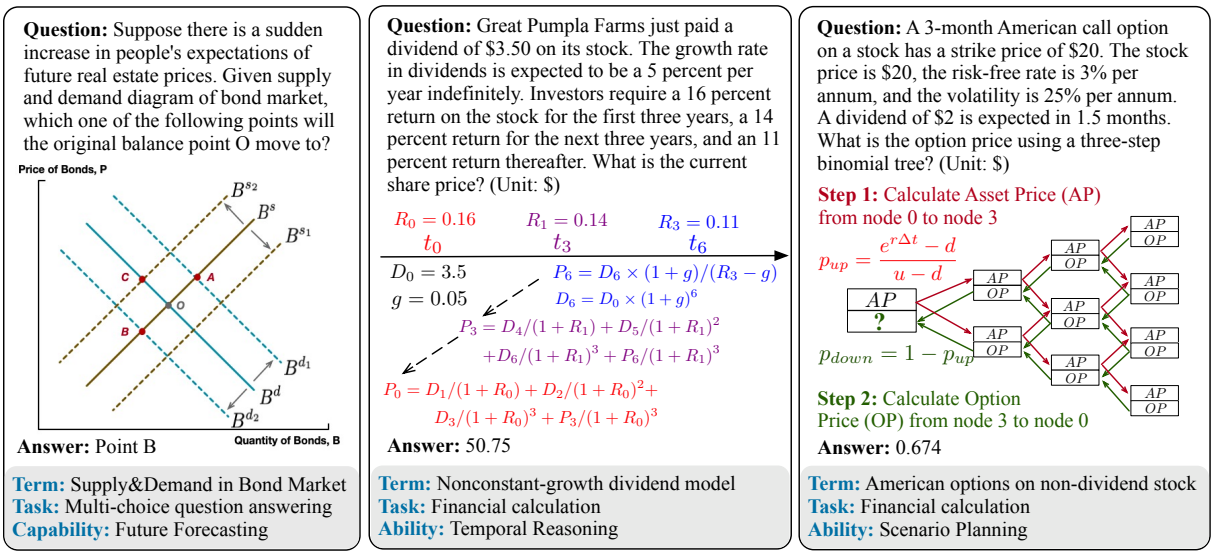


Figure 2: Examples in our dataset XFINBENCH.

includes a list of properties of options and future contracts; the calculation question "*An investment offers ... If the payment occurs for 15 years, what is its value? For 40 years? Forever?*" contains a series of sub-questions with different final answers. To ensure each question in XFINBENCH to be evaluated accurately and conveniently, we leverage GPT-4o to further process these after-class questions under a Generate-then-verify framework (Zhang et al., 2024).

Generation Stage. We employ few-shot prompts to guide GPT-4o to transform open-ended questions into those with clear final answers. For *statement judging* task, we ask GPT-4o to extract both true and false statements from each after-class question. To ensure the balance between true and false statements, we apply two prompt templates with the same after-class questions as few shots, but one with true statements and one with false statements (see §G.2.1). For *multi-choice question answering* task, we follow STARC rules (Berzak et al., 2020) to instruct GPT-4o to reformulate each after-class question and generate three candidate choices: one correct answer with evidence and two plausible but misleading distractors (see §G.2.2). For *financial calculation* task, we ask GPT-4o to decomposes the complex after-class question into a sequence of independent questions with clear final answers (see §G.2.3). Finally, 6,227 questions are generated from after-class questions in the generation stage.

Verification Stage. We then verify the quality of questions in the generation stage from multiple dimensions. We primarily evaluate *Correctness* and *Completeness* of the generated question and answer. Specifically, we evaluate whether (1) the question provides the *complete* background information to get its final answer, and (2) the final answer is *correct* to the question given the after-class question and its gold answer. Furthermore, to ensure the independence of questions in *statement judging* task, we verify if, within the same after-class question, true statements provide no evidence to support that false statement(s) is wrong. For *multi-choice question answering* task, we verify if the two misleading choices are exclusive to, but share the similar wording and length with the correct choice. For *financial calculation* task, we verify if the final answers are numerical without any text included. Finally, 35.2% questions are discarded in the verification stage (See §B.2 for details).

2.3 Human Quality Validation

We conduct a comprehensive validation protocol to ensure the high quality of all annotated examples in XFINBENCH. For each example, we assign three evaluators to validate whether: 1) the question is fluent and contains complete information to get the final answer; 2) the final answer is correct according to the gold answer of after-class question; 3) the annotated finance terms are helpful for answering the question. Each criterion is rated individually on a 1-to–5 scale. Notably, our evaluators are accessible to the corresponding after-class questions with gold answers and the knowledge bank, which is different from the close-book setting for human performance in the following Experiment section.

Statistics	Number
XFINBENCH *dataset*	
Total questions	4,235
- *statement judging*	1,795 (42.4%)
- *multi-choice question answering*	761 (18.0%)
- w. Image	146
- *financial calculation*	1,679 (39.6%)
- w. Tabular	330
Question Length (Median / Avg)	244 / 273.7
Terms per question (Median / Avg)	1.0 /1.3
Test Set Size	3,235
Validation Set Size	1,000
Knowledge Bank	
Total terms	3,032
Unique number of definition	1,766
- w. Mathematical Formula	34.3%
Definition Length (Median / Avg)	830 / 1,249

Table 2: Key statistics of XFINBENCH.

We calculate the proportions of examples with average score S $\geq$ 4: *question fluency* 97.1%, *question completeness* 96.8%, *answer correctness* 98.0%, *knowledge helpfulness* 91.2%, illustrating the high quality of XFINBENCH (See Table 6 and §B.4 for detailed results).

2.4 Data Statistics

Table 2 summarizes the key statistics of XFINBENCH, which includes 4,235 examples divided into *validation* (1,000 examples) and *test* (3,235 examples) subsets. The division is based on random sampling over the after-class questions. The *validation* set supports model development validation, while the *test* set is reserved for standard evaluation, whose answers will not be publicly released for preventing data contamination. An online evaluation platform will be maintained to facilitate test set evaluations. Our dataset also includes a knowledge bank of 3,032 finance terms and 1,766 unique definitions, covering 28 finance topics (see Figure 6 and §C for details).

3 Experiments

We conduct qualitative and quantitative studies for comprehensive evaluation of leading LLMs in knowledge-intensive finance tasks.

3.1 Experimental Setup

We evaluate the models on the test set of XFINBENCH uder two setups: 1) *Multimodal Large Language Models* (MLLMs) who allow visual input, including gpt-4o (OpenAI, 2024b), gpt-4o-mini (OpenAI, 2024a), claude-3.5-sonnet (Anthropic, 2024b), claude-3-opus, claude-3-haiku (Anthropic, 2024a), gemini-1.5-flash and gemini-1.5 pro (Team, 2024b), and Llama-3.2-Vision models (Meta, 2024c), and 2) *Text-only Large Language Models* who only allow textual input, including o1 (OpenAI, 2024c), o1-mini (OpenAI, 2024d), deepseek-chat (DeepSeek-AI, 2024), Llama-3.1 models (Meta, 2024a), Llama-3 models (Meta, 2024b), and Mixtral-7$\times$8B (Jiang et al., 2024) (§D.1). All MLLMs allow text-only input except for Llama-3.2-Vision models, which we feed with a blank image in text-only tasks.

We apply Chain-of-Thought (CoT) method (Wei et al., 2022) and evaluate performance via Accuracy (*Acc*) for three tasks. In *financial calculation* task, we additionally apply Program-of-Thought (PoT) method (Chen et al., 2023) and use $Acc_{ERR@5}$ for evaluation, which measures accuracy within a 0.5% error margin of the correct answer. [2]

We establish a human performance baseline with three graduate-level human experts over a random 1,000-example subset of the test set of XFINBENCH in a close-book setting (§D.2). None of them is involved in the dataset construction work.

3.2 Main Results

Among MLLMs, claude-3.5-sonnet achieves the best performance with 64.1% accuracy on XFINBENCH, followed by gpt-4o with 63.6% accuracy who achieve the highest accuracy in visual-context questions, *i.e.*, 65.3%. On the text-only LLM side, o1 achieves the highest accuracy in almost all tasks of XFINBENCH, with 67.3% overall accuracy; however, it still falls 12.5% short of human performance, highlighting that there is a significant scope for further improvements on our benchmark. Open-source models with large parameter size, *i.e*, Llama-3.1-405B, achieves comparable performance with o1-mini and even outperforms gpt-4o-mini in text-only tasks. However, most open-source models achieve underwhelming performance, attributed to their lack of domain knowledge and mathematical reasoning ability.

We observe that the PoT prompting method deteriorates the performance of most models in *financial calculation* task. To better analyze the reasons for these differing performance outcomes, we examine the execution rate of models under PoT prompting on XFINBENCH, measuring how

[2] Unless specified otherwise, the evaluation results for *financial calculation* task are reported using $Acc_{ERR@5}$.

Task	*Statement judging*	*Multi-choice question*		*Financial calculation*				***All***
Reasoning	CoT	CoT		CoT		PoT		CoT
Input	Q	Q	Q, I	Q	Q, T	Q	Q, T	Q, T
Multimodal Large Language Models								
gpt-4o	84.0	91.5	65.3	30.1 / 47.0	37.9 / 59.2	25.4 / 44.2	33.2 / 55.2	63.6
gpt-4o-mini	76.5	86.8	54.8	25.4 / 38.4	30.3 / 48.4	17.6 / 37.8	25.3 / 49.5	57.4
claude-3.5-sonnet	84.3	94.2	63.7	29.8 / 46.9	37.9 / 59.6	34.2 / 47.5	42.2 / 54.5	64.1
claude-3-opus	79.0	91.2	50.7	25.6 / 39.5	35.7 / 55.2	27.9 / 38.3	40.1 / 52.0	59.7
claude-3-haiku	70.0	82.9	43.6	15.1 / 22.5	23.8 / 33.9	23.3 / 28.8	34.7 / 40.4	50.1
gemini-1.5-flash	74.0	82.5	49.2	22.4 / 30.7	28.9 / 40.1	16.4 / 37.7	23,8 / 48.0	54.5
gemini-1.5-pro	76.3	86.5	50.8	25.0 / 37.4	32.5 / 44.0	24.9 / 40.7	32.9 / 50.5	57.3
Llama-3.2-90B-Vision	57.4	70.9	47.6	14.4 / 19.9	18.1 / 20.6	11.3 / 21.4	13.2 / 25.9	42.0
Llama-3.2-11B-Vision	51.8	70.3	42.0	8.3 / 12.3	11.2 / 12.6	8.2 / 15.9	14.4 / 16.0	36.9
Text-only Large Language Models								
o1	87.6	94.0		34.2 / 62.0	42.2 / 66.4	30.5 / 51.4	35.7 / 50.9	67.3
o1-mini	81.0	90.0		29.7 / 52.1	39.0 / 60.6	28.9 / 48.2	38.6 / 55.6	62.0
Llama-3.1-405B	83.6	91.9		26.2 / 39.6	34.7 / 48.7	14.1 / 28.4	22.7 / 43.7	61.9
deepseek-chat	74.4	88.2		29.2 / 44.6	37.9 / 55.2	21.8 / 45.9	28.5 / 54.9	59.6
Llama-3.1-70B	80.5	90.0		24.1 / 35.6	31.4 / 43.0	11.0 / 26.1	12.3 / 29.6	59.3
Llama-3-70B	78.2	85.9		19.9 / 27.9	28.5 / 38.3	7.2 / 18.5	13.4 / 30.7	56.1
Llama-3.1-8B	65.3	77.8		11.6 / 16.8	17.0 / 24.5	10.3 / 18.8	11.9 / 26.0	45.5
Llama-3-8B	63.0	75.9		8.3 / 12.6	14.4 / 19.1	7.0 / 12.8	10.5 / 22.4	42.9
Mixtral-8 $\times$ 7B	26.1	29.9		1.4 / 1.7	2.5 / 4.3	1.4 / 0.6	2.9 / 4.3	16.6
Human								
Human performance	90.9	92.1	81.1	63.8 / 77.6	74.6 / 83.6			79.8

Table 3: Performance of models on XFINBENCH. Input: Q: question, I: image, T: tabular. In *All* column, Q, T indicates including questions with Q and Q, T. For positions using "a / b", a refers to exact-matching accuracy and b refers to $Acc_{ERR@5}$. Dark red cells indicate the highest score within each set of models, while light red cells represent the second-highest score.

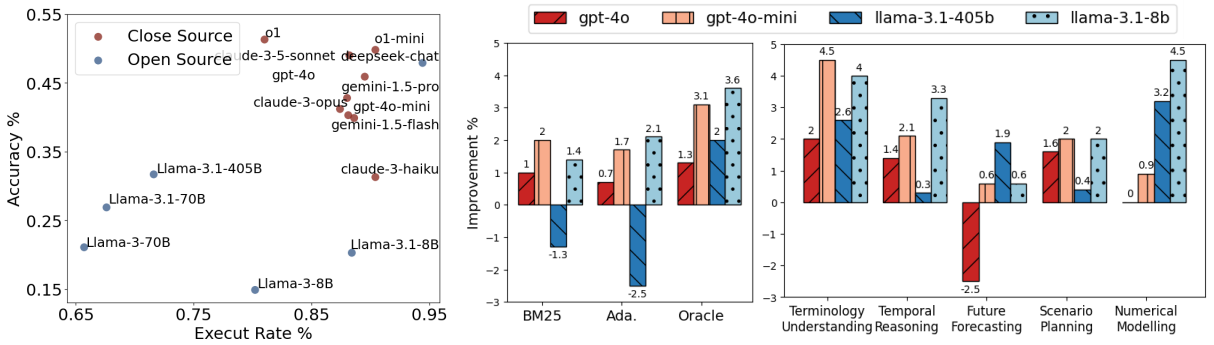
many of the generated Python programs are executable (Zhao et al., 2024). Figure 3(a) illustrates the relationship between execution rate and accuracy $Acc_{ERR@5}$ across different models, indicating that the degraded performance when applying PoT prompting is attributable to the low execution rate. For instance, while Llama-3.1-405B achieves competitive performance using CoT prompting, it struggles to consistently generate executable Python solutions, leading to lower accuracy with PoT prompting. Interestingly, while o1's execution rate lags behind most close-source models, it achieves the highest accuracy score on $Acc_{ERR@5}$, witnessing its strong and efficient reasoning ability over complex tasks. We report more fine-grained results during evaluation in §E.

3.3 Knowledge Augmentation Method

We explore the performance of models augmented with external knowledge base, and apply two types of retrievers to acquire the top-n question-relevant knowledge term from knowledge bank, *i.e. BM25* and *Ada Embed.* (OpenAI, 2022), where n is set to be 3. Recalling that we have annotated the most relevant finance terms for each question, we further design a *Oracle* setting, where models are provided with the *ground-truth* finance term(s).

We report the performance improvements of four models when augmented with a knowledge bank in Figure 3(b). For various retrieving settings, we find that the *Oracle* setting leads to the most robust improvements on most models, highlighting the high quality of our annotated dataset. *BM25* and *Ada Embed.* retrievers both improve the performance of most models; however, they result in a decline in performance for Llama-3.1-405B.

Furthermore, we report the performance improvements across five financial capabilities under *Oracle* setting in Figure 3(c). The most significant gains are observed in *terminology understanding*, while improvements in *future forecasting* are limited and even negative for GPT-4o. The smallest open-source model, *i.e*, Llama-3.1-8B, shows



(a) Relationship between accuracy and executing rate under PoT method.

(b) Performance improvement across three retrieving settings.

(c) Performance improvement across five capabilities in Oracle setting

Figure 3: Subfigure(a) shows the relationship between accuracy $Acc_{ERR@5}$ and executing rate under PoT setting. Subfigure (b) and (c) illustrates the performance improvements across three retrieving settings and five capabilities in knowledge augmentation method.

the greatest improvements across most capabilities, particularly in *numerical modelling*. We report more results of knowledge augmentation in §E.

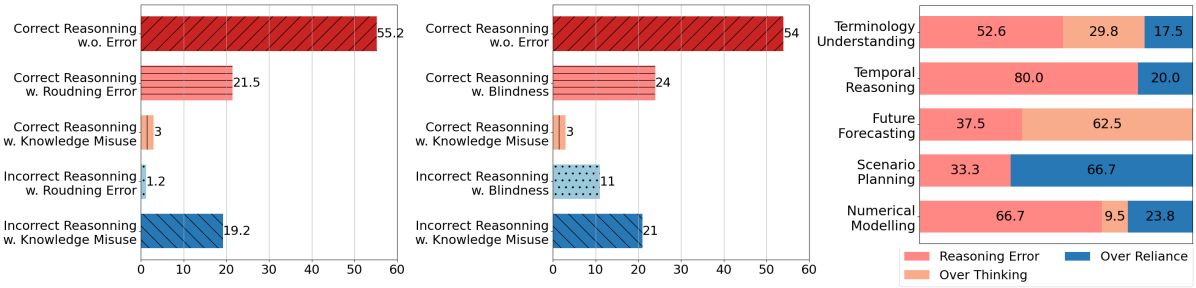
3.4 Error Analysis

We conduct error analysis on *financial calculation* task, visual-context questions and knowledge augmentation method. Human annotators are instructed for error type labeling (§F.1)

Error Analysis of Financial Calculation. We randomly select 400 samples from responses of o1 in *financial calculation* task, and observe that there are two primary reasons of incorrect responses in calculating task are: 1) Rounding Error that exists in the intermediate calculating steps, and 2) Knowledge Misuse if applying wrong or incomplete finance formulas for calculation. Figure 4(a) showcases that 55.2% of o1's response had correct reasoning path without intermediate rounding error or knowledge misuse. Knowledge misuse appears more frequently in incorrect-reasoning responses, while rounding error often exists in correct reasoning process. For better illustration, we display an example of o1's response containing both two errors in Figure 5(b). In this example, o1 fails to use the primary property of American options, *i.e.* exercising the option before expiration date for profit maximization, and hence leads to unnecessary calculation in the following nodes. It also presents a rounding error when building binomial tree, which inevitably leads to an incorrect answer in the end.

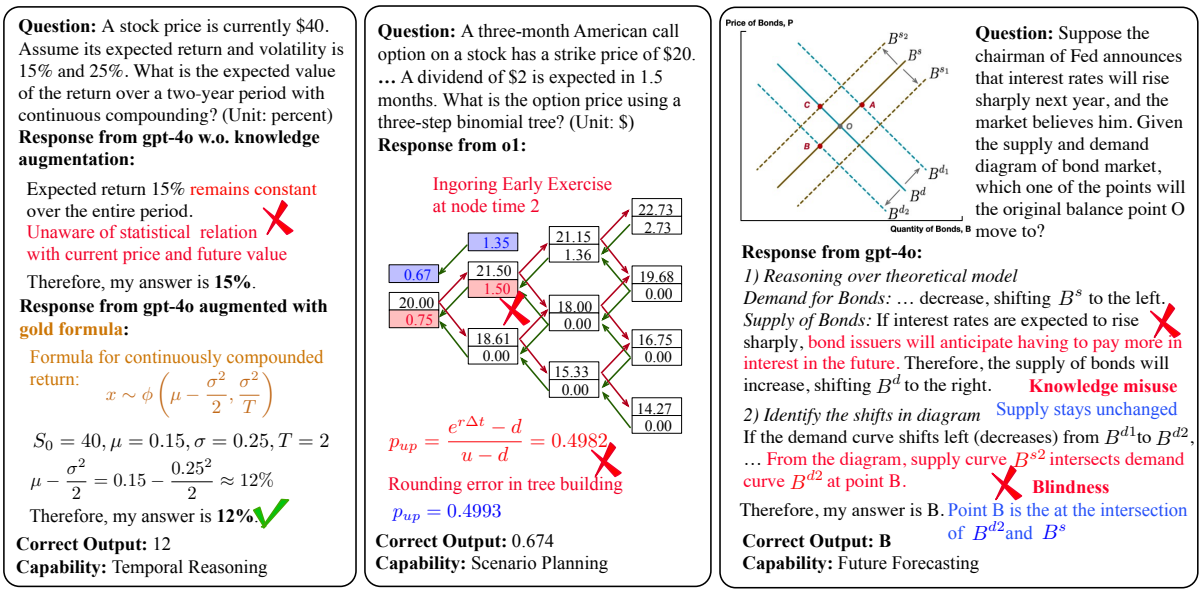
Error Analysis of Visual Context. We randomly select 100 samples from responses of GPT-4o in visual-context *multiple-choice question answering* task, and identify two primary error types: 1) Blindness (Rahmanzadehgervi et al., 2024), where the model struggles with identifying the position and/or intersection of two curves, and 2) Knowledge Misuse, occurring when irrelevant knowledge is introduced, thereby disrupting the reasoning path. Figure 4(b) showcases that the model responds with correct reasoning but either blindness (24%) or knowledge misuse (3%). It is worth noting that 35% of its responses contain blindness, highlighting that blindness is a major source of errors in the generative foundation models (Rahmanzadehgervi et al., 2024; Alayrac et al., 2022; Liu et al., 2023; Team, 2024a; Tong et al., 2024). We present an example of gpt-4o's responses to illustrate the two error types. In Figure 5(c), while GPT-4o outputs the correct final answer, its response contain the misunderstanding of supply in bond market and blindness to the intersection of R^{d2} and R^{s} curves.

Error Analysis of Knowledge Augmentation. We randomly select 100 samples from responses of GPT-4o that deliver wrong final answers under *Oracle* setting. Three error types are identified when models are augmented with *ground-truth* finance term(s) but still fail to deliver the correct final answers: 1) Reasoning Error that appears in the model's reasoning process and has no direct relation to the augmented knowledge; 2) Over Thinking, in which case augmented knowledge provides direct solutions but the model reasons further steps that go out of the question's scope; 3) Over Reliance, in which case the model's reasoning process is entirely guided by augmented knowledge, foregoing simpler approaches to answering the question. As illustrated in Figure 4(c) and 7, most of wrong final answers for calculating questions, especially those requiring *temporal reasoning* and *numerical modelling* capabilities, are caused by reasoning



(a) Error analysis of o1 in financial calculation task. (b) Error analysis of gpt-4o in visual-context questions. (c) Error analysis of gpt-4o under Oracle setting.

Figure 4: Error analysis for (a) *financial calculation*, (b) visual-context questions and (c) knowledge augmentation.



(a) Case of Knowledge misuse before knowledge augmentation

(b) Case of rounding error and knowledge misuse during calculation.

(c) Case of blindness and knowledge misuse in visual-context question.

Figure 5: Case study in error analysis. Subfigure (a) provides an example of knowledge misuse before knowledge augmentation. Subfigure (b) shows two error types in *financial calculation* task. Subfigure (c) shows two error types in visual-context questions.

error that has little to do with augmented knowledge, such as rounding error. Over thinking is most frequently observed in multiple-choice questions requiring *future forecasting* capability, suggesting that GPT-4o exhibits a tendency to engage in deeper reasoning when addressing questions involving predictions of future events. Moreover, over reliance is most commonly encountered in questions requiring *scenario planning* capability, which emphasizes the model's ability to plan rather than strictly adhering to the instructions provided in the augmented knowledge (see case studies in F.2).

4 Conclusion

In this work, we introduced XFINBENCH, a benchmark comprising 4,235 examples designed to evaluate the ability of LLMs to solve complex, knowledge-intensive financial problems across diverse topics and multi-modal contexts. Evaluation results indicate that while o1 is the best-performing text-only model with an overall accuracy of 67.3%, it falls significantly behind human experts by 12.5%, particularly in *temporal reasoning* and *scenario planning* capabilities. Further analysis revealed that integrating *ground-truth* knowledge yields only limited performance improvements in tackling complex financial problems, and that limitations in models' calculation ability and visual information recognition present significant barriers to progress in finance domain. These findings underscore the critical role of XFINBENCH in driving the development of AI agents capable of effectively solving complex, multi-modal financial problems.

Limitation

Our work evaluates the ability of large language models (LLMs) to solve complex financial problems across diverse topics and multi-modal contexts. Following a sensitivity analysis (Appendix G), our prompt template used for evaluation includes three key components: a role-play system message, the application of chain-of-thought or program-of-thought reasoning methods, and clear output requirements. This approach may impact model performance if the generated responses do not align with the specified output format. Additionally, while XFINBENCH is notable for its complexity and high-quality, it includes a comparatively smaller number of QA pairs relative to financial datasets focused on simpler tasks, such as quantity extraction.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc.

Anthropic. 2024a. The claude 3 model family: Opus, sonnet, haiku.

Anthropic. 2024b. Claude 3.5 sonnet.

Yevgeni Berzak, Jonathan Malmaud, and Roger Levy. 2020. STARC: Structured annotations for reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5726–5735, Online. Association for Computational Linguistics.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*.

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

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. ConvFinQA: Exploring the chain of numerical reasoning in conversational finance question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6279–6292, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

DeepSeek-AI. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *Preprint*, arXiv:2405.04434.

Yang Deng, Wenqiang Lei, Wenxuan Zhang, Wai Lam, and Tat-Seng Chua. 2022. PACIFIC: Towards proactive conversational question answering over tabular and textual data in finance. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6970–6984, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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 (TACL)*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. 2024. Time-LLM: Time series forecasting by reprogramming large language models. In *The Twelfth International Conference on Learning Representations*.

Rik Koncel Kedziorski, Michael Krumdick, Viet Lai, Varshini Reddy, Charles Lovering, and Chris Tanner. 2024. Bizbench: A quantitative reasoning benchmark for business and finance. *Preprint*, arXiv:2311.06602.

Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. 2024. MMC: Advancing multimodal chart understanding with large-scale instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*

(Volume 1: Long Papers), pages 1287–1310, Mexico City, Mexico. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

Meta. 2024a. Introducing llama 3.1: Our most capable models to date.

Meta. 2024b. Introducing meta llama 3: The most capable openly available llm to date.

Meta. 2024c. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models.

OpenAI. 2022. New and improved embedding model.

OpenAI. 2024a. Gpt-4o mini: advancing cost-efficient intelligence.

OpenAI. 2024b. Hello gpt-4o.

OpenAI. 2024c. Learning to reason with llms.

OpenAI. 2024d. Openai o1-mini.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. 2024. Vision language models are blind. *Preprint*, arXiv:2407.06581.

Varshini Reddy, Rik Koncel-Kedziorski, Viet Lai, Michael Krumdick, Charles Lovering, and Chris Tanner. 2024. DocFinQA: A long-context financial reasoning dataset. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 445–458, Bangkok, Thailand. Association for Computational Linguistics.

Zhaochen Su, Juntao Li, Jun Zhang, Tong Zhu, Xiaoye Qu, Pan Zhou, Yan Bowen, Yu Cheng, and Min Zhang. 2024. Living in the moment: Can large language models grasp co-temporal reasoning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13014–13033, Bangkok, Thailand. Association for Computational Linguistics.

Chameleon Team. 2024a. Chameleon: Mixed-modal early-fusion foundation models. *Preprint*, arXiv:2405.09818.

Gemini Team. 2024b. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Preprint*, arXiv:2406.16860.

Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2022. Large language models still can't plan (a benchmark for LLMs on planning and reasoning about change). In *NeurIPS 2022 Foundation Models for Decision Making Workshop*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024a. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Preprint*, arXiv:2406.01574.

Yuqing Wang and Yun Zhao. 2024. TRAM: Benchmarking temporal reasoning for large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6389–6415, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. 2024b. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *arXiv preprint arXiv:2406.18521*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Xinli Yu, Zheng Chen, and Yanbin Lu. 2023. Harnessing LLMs for temporal data - a study on explainable financial time series forecasting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 739–753, Singapore. Association for Computational Linguistics.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*.

Liwen Zhang, Weige Cai, Zhaowei Liu, Zhi Yang, Wei Dai, Yujie Liao, Qianru Qin, Yifei Li, Xingyu Liu, Zhiqiang Liu, Zhoufan Zhu, Anbo Wu, Xin Guo, and Yun Chen. 2023. Fineval: A chinese financial domain knowledge evaluation benchmark for large language models. *Preprint*, arXiv:2308.09975.

Zhihan Zhang, Yixin Cao, Chenchen Ye, Yunshan Ma, Lizi Liao, and Tat-Seng Chua. 2024. Analyzing temporal complex events with large language models? a benchmark towards temporal, long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1588–1606, Bangkok, Thailand. Association for Computational Linguistics.

Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. 2022. MultiHiertt: Numerical reasoning over multi hierarchical tabular and textual data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6588–6600, Dublin, Ireland. Association for Computational Linguistics.

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

Tian Zhou, Peisong Niu, Xue Wang, Liang Sun, and Rong Jin. 2023. One fits all: Power general time series analysis by pretrained LM. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3277–3287, Online. Association for Computational Linguistics.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. ProSA: Assessing and understanding the prompt sensitivity of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1950–1976, Miami, Florida, USA. Association for Computational Linguistics.

A Data Collection Guidelines

A.1 Financial Capability Definition and Annotation

We define five core capabilities required for tackling complex finance problems in Table 4, along with their proportions. For human annotation, we ask three human annotators to label each question in our dataset with 1-to-2 capability. A question will be labelled with one capability if at least two annotators choose this capability to label it. Specifically, questions that focus on the comprehension of financial terms and mathematical formulas are labeled as requiring *terminology understanding*. Questions necessitating the model's reasoning over time-series data, concepts, and mathematical formulas are categorized under *temporal reasoning*. When a question centers on predicting future trends, it is marked as requiring *future forecasting*. For questions that involve analyzing potential future scenarios to aid in decision-making, the label *scenario planning* is used. Lastly, questions that involve creating structured representations of a company's financial performance using financial statements and informed assumptions are identified as needing *model building*.

A.2 Examples of Financial Capability

Examples to display five capabilities for complex finance problem solving are shown in A.2.1, A.2.2, A.2.3, A.2.4, and A.2.5.

A.2.1 Examples of Terminology Understanding

Example 1 of Terminology Understanding in *Statement Judging* task

An investor holds a strip and believes that there will be a big jump in a stock price. He will earn a bigger profit when there is a large upward stock price move than a downward move.
Answer: False

Example 2 of Terminology Understanding in *Multi-choice Question* task

A bank is managing floating-rate deposits and fixed-rate loans, leading to asset-liability mismatch. Which one of the following swaps can help the bank offset risk?
A. Pay fixed and receive floating
B. Pay floating and receive fixed
C. Pay variable and receive fixed
Answer: A

A.2.2 Examples of Temporal Reasoning

Example 1 of Temporal Reasoning in *Financial Calculation* task

You own 1,000 shares of stock in Avondale Corporation. You will receive a $1.50 per share dividend in one year. In two years, Avondale will pay a liquidating dividend of $45 per share. The required return on Avondale stock is 15 percent. What would be the equal dividend per share in each of the next two years to have the same present value as the current share price? (Unit: dollar)
Answer: 21.73

Example 2 of Temporal Reasoning in *Financial Calculation* task

The price of a European call that expires in six months and has a strike price of $30 is $2. The underlying stock price is $29, and a dividend of $0.50 is expected in two months and again in five months. Interest rates (all maturities) are 10%. If the stock price is above $30 in six months, what is the present value of the profit? (Unit: dollar)
Answer: 0.49

Capability	Description
Terminology Understanding (56.1%)	It refers to the model's ability to accurately understand finance concepts, including standard financial terms, acronyms, accounting principles, various financial instruments, regulatory terminologies, and economic indicators.
Temporal Reasoning (21.7%)	It focuses on understanding temporal relations in time-based data, and making time-sensitive decisions. It involves cross-period data, like quarterly earnings reports, historical stock performance and future cash flow projections.
Future Forecasting (5.0%)	It involves predicting future values or trends of financial indicators such as output level, price level and inflation rates. It requires the model to use economic theories and quantitative methods to generate forecasts for decision-making.
Scenario Planning (7.6%)	It is the process of generating and analyzing different possible future scenarios to assess their impact on financial decisions and strategies. It requires considering various uncertainties and variables to prepare for various outcomes.
Numerical Modelling (17.2%)	It involves creating structured representations of a company or product's financial performance. Related questions typically include financial statements like income statements, balance sheets, and cash flow statements.

Table 4: Definitions of five capabilities of solving complex, knowledge-intensive finance problem.

A.2.3 Examples of Future Forecasting

Example 1 of Future Forecasting in *Multi-choice Question* task

Both Keynes' and Friedman's theories of the demand for money discuss the impact of interest rates on money demand. According to Keynes model, which one of the following outcomes happens when interest rates rise?
A. Demand for money decreases
B. Demand for money increases
C. Demand for money stays unchanged
Answer: A

Example 2 of Future Forecasting in *Multi-choice Question* task

Interest rates tend to change in response to the increase or decrease of aggregate output during economic booms and recessions. Which one of the following actions might banks take when output rises during a boom?
A. Freeze the level of their excess reserves
B. Reduce the level of their excess reserves
C. Increase the level of their excess reserves
Answer: B

A.2.4 Examples of Scenario Planning

Example 1 of Scenario Planning in *Multi-choice Question* task

A trader sells a strangle by selling a call option with a strike price of \$50 for \$3 and selling a put option with a strike price of \$40 for \$4. Within which one of the following price ranges of the underlying asset does the trader make a profit?
A. Between \$33 and \$57
B. Between \$30 and \$50
C. Between \$40 and \$60
Answer: A

Example 2 of Scenario Planning in *Financial Calculation* task

On May 8, 2013, an investor owns 100 Google shares. The share price is about \$871 and a December put option with a strike price of \$820 costs \$37.50. The investor is comparing two alternatives to limit downside risk. The first involves buying one December put option contract with a strike price of \$820. The second involves instructing a broker to sell the 100 shares as soon as Google's price reaches \$820. How much will the investor pay to buy one December put option contract with a strike price of \$820? (Unit: dollar)
Answer: 3750

A.2.5 Examples of Numerical Modelling

Example 1 of Numerical Modelling in *Financial Calculation* task

Bedrock Gravel Corp.'s 2007 income statement shows the following information: sales = $162,000; costs = $93,000; other expenses = $5,100; depreciation expense = $8,400; interest expense = $16,500; taxes = $14,820; dividends = $9,400. Additionally, the firm issued $7,350 in new equity during 2007 and redeemed $6,400 in outstanding long-term debt. What is the 2007 operating cash flow? (Unit: dollar)
Answer: 49080

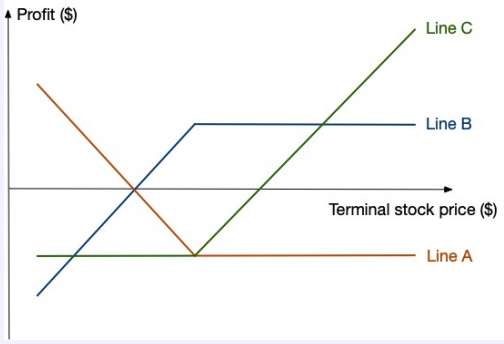
Example 2 of Numerical Modelling in *Financial Calculation* task

Winnebagel Corp. currently sells 30,000 motor homes per year at $45,000 each, and 12,000 luxury motor coaches per year at $85,000 each. The company wants to introduce a new portable camper to fill out its product line; it hopes to sell 19,000 of these campers per year at $12,000 each. An independent consultant has determined that if Winnebagel introduces the new campers, it should boost the sales of its existing motor homes by 4,500 units per year, and reduce the sales of its motor coaches by 900 units per year. What is the annual sales figure due solely to the new portable camper product line? (Unit: dollar)
Answer: 228000000

A.3 Examples of Visual-context Questions

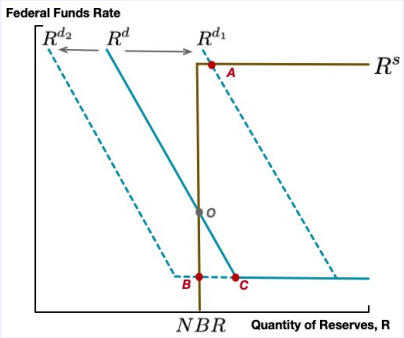
Visual-context examples in XFINBENCH align with multi-modal benchmarks involving chart understanding and reasoning but introduce a more domain-specific focus. Most related benchmarks, such as MMMU (Yue et al., 2024), MMLU-Pro (Wang et al., 2024a), and MathVista (Lu et al., 2024), emphasize descriptive questions about charts, such as identifying a bar's value or a line's trend. While these benchmarks highlight domain-specific knowledge, they primarily involve recognizing technical terms in charts rather than deeper financial reasoning. Similarly, chart-specific benchmarks like ChartQA (Masry et al., 2022), MMC (Liu et al., 2024), and CharXiv (Wang et al., 2024b) incorporate multi-step reasoning but focus on visual rather than finance-specific reasoning. In contrast, XFINBENCH's visual-context questions demand both reasoning over chart visuals and interpreting financial implications of the data. For instance, Figure 5(c) evaluates the model's ability to analyze the supply and demand curves' positional relationships and infer their shifts under changing financial conditions. This extends beyond visual reasoning to complex financial reasoning. While our visual-context questions are limited in size, XFINBENCH is the first to explore the application of finance-domain knowledge to complex chart reasoning.

Example 1 of Visual-context question in *Multi-choice Question* task that evaluates Scenario Planning capability



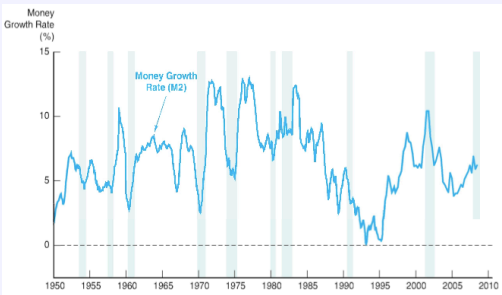
Suppose that a June put option to sell a share for $60 costs $4 and is held until June. Which line in the attached figure best describes the relationship between the option's profit and the stock price?
A. Line A
B. Line B
C. Line C
Answer: B

Example 2 of Visual-context question in *Multi-choice Question* task that evaluates Future Forecasting capability



Suppose the economy is surprisingly strong, leading to an increase in the amount of checkable deposits. Given the supply and demand diagram of reserve market, which one of the following points will the original balance point O move to?
A. Point A
B. Point B
C. Point C
Answer: A

Example 3 of Visual-context question in *Financial Calculation* task that evaluates Temporal Reasoning capability



The shaded areas in the attached figure represent recessions. What is the relationship between the rate of money growth and recessions as indicated in this figure?
A. The rate of money growth has declined before every recession;
B. The rate of money growth has little correlation with the recession periods;
C. The rate of money growth has increased before every recession.
Answer: A

Example 1 of Visual-context question in *Financial Calculation* task that evaluates Numerical Modelling capability

	2006	2007
Sales	$ 4,822	$ 5,390
Depreciation	692	723
Cost of goods sold	1,658	1,961
Other expenses	394	343
Interest	323	386
Cash	2,528	2,694
Accounts receivable	3,347	3,928
Short-term notes payable	488	478
Long-term debt	8,467	10,290
Net fixed assets	21,203	22,614
Accounts payable	2,656	2,683
Inventory	5,951	6,370
Dividends	588	674

For 2007, calculate the cash flow from assets, cash flow to creditors, and cash flow to stockholders based on financial data from the table. What is the value of total liability and equity of this firm during 2006? (Unit: dollar)
Answer: 33029

A.4 Examples of Term Definitions

Example 1 of finance term and definition

Term: Two-stage growth model for common stock valuation
If the dividend grows at rate g_1 for t periods and then grows at rate g_2 thereafter, then the price can be written as: $P_0 = \frac{D_1}{R-g_1} \times \left[1 - \left(\frac{1+g_1}{1+R}\right)^t\right] + \frac{P_t}{(1+R)^t}$, where $P_t = \frac{D_{t+1}}{R-g_2} = \frac{D_0 \times (1+g_1)^t \times (1+g_2)}{R-g_2}$, D_1 is the next dividend, and R is the required return.

Example 2 of finance term and definition

Term: Total credit cost curve of optimal credit policy
The trade-off between granting credit and not granting credit isn't hard to identify, but it is difficult to quantify precisely. As a result, $\cdots$ The sum of the carrying costs and the opportunity costs of a particular credit policy is called the total **credit cost curve**.$\cdots$ \n\n All other things being equal, for example, it is likely that firms with (1) excess capacity, (2) low variable operating costs, and (3) repeat customers will extend credit more liberally than other firms.

Example 3 of finance term and definition

Term: Open market operations for control of Monetary Base
The Federal Reserve exercises control over the monetary base through its purchases or sale of government securities in the open market, called **open market operations**, and through its extension of discount loans to banks. A purchase of bonds by the Fed is called an **open market purchase**, and a sale of bonds by the Fed is called an **open market sale**.

Example 4 of finance term and definition

Term: Exchange-rate targeting
Targeting the exchange rate is a monetary policy strategy with a long history. It can take the form of fixing the value of domestic currency to a commodity such as gold. $\cdots$ Another alternative is to adopt a crawing target or peg, in which a currency is allowed to depreciate at a steady rate so that the inflation rate in the pegging country can be higher than that of the anchor country.

Example 5 of finance term and definition

Term: American call option
Black suggests an approximate procedure for taking account of early exercise in call options. This involves calculating, the prices of European options that mature at times T and t_n, and then setting the American price equal to the greater of the two.15 This is an approximation because it in effect assumes the option holder has to decide at time zero whether the option will be exercised at time T or t_n.

Example 6 of finance term and definition

Term: Interest rates in convexity adjustment
Consider first an instrument that provides a payoff dependent on a bond yield observed at the time of the payoff. Usually the forward value of a variable S is calculated with reference to a forward contract that pays off $S_T - K$ at time T. It is the value of K that causes the contract to have zero value. $\cdots$ The relationship between the price of this bond and its yield is $G(y) = \frac{1}{1+y\tau}$ From equation (3.1), $E_T(R_T) = R_0 - \frac{1}{2}R_0^2\sigma_R^2 T\frac{G''(R_0)}{G'(R_0)}$ or $E_T(R_T) = R_0 + \frac{R_0^2\sigma_R^2\tau T}{1+R_0\tau}$ (3.2) where R_0 is the forward rate applicable to the period between T and T^* and σ_R is the volatility of the forward rate. The value of the instrument is therefore $P(0,T)L\tau\left[R_0 + \frac{R_0^2\sigma_R^2\tau T}{1+R_0\tau}\right]$.

B Detailed Data Construction

B.1 Source Data

The details of textbooks are displayed in Table 5. During data collection, annotators are instructed to adhere to copyright and license regulations, avoiding data from sites prohibiting copy and redistribution.

Textbook	Authors	Version	# Chapters
Fundamentals of Corporate Finance	Stephen A. Ross	8	22
Options, Futures and Other Derivatives	John C. Hull	9	32
The Economics of Money Banking and Financial Markets	Frederic S. Mishkin	9	25

Table 5: Details of textbooks as source data.

B.2 QA Task and Automatic Annotation

We leverage GPT-4o to process after-class questions under a generate-then-verify framework (Zhang et al., 2024). For the generation stage, examples in the prompt template illustrate the rules of transforming open-ended questions into those with clear final answers. For *statement judging* task, rules of creating false statements are: 1) antonym substitution, such as small → big; 2) object position interchange, such as "A is red and B is blue" → "B is red and A is blue"; 3) adjective modification, such as "it is possible" → "it is impossible", etc (§G.2.1). For *multi-choice question answering* task, we follow STARC (Berzak et al., 2020) rules to design two misleading choices that are mutually exclusive to but share the similar wording and length with the correct choice (§G.2.2). For *financial calculation* task, calculation questions usually have a series of sub-questions that share the same solution in the gold answer but have different final answers. In this case, GPT-4o simply split the question into independent questions with clear final answers (§G.2.3). Furthermore, to ensure that the generated question contain necessary information to get its final answer, we ask GPT-4o to extract the context in the after-class question first, and then extract the question and its final answer (see examples in prompt templates). For the verification stage, rules for discarding unqualified questions are illustrated in the prompt templates in §G.

B.3 Knowledge Bank Construction and Annotation

We collect finance terms from the subject index at the end of each textbook, and manually extract their definitions from the chapter's content. Specifically, for each term, we locate its corresponding pages indicated in the subject index, and collect the paragraphs related to this term. There are two common cases during this process: (1) the term's name is the title of a subsection, so its related paragraphs are the main content of this subsection; (2) the term's definition in the corresponding page is within a highlighted box, so we only collect the information within the box. Mathematical expressions and tabular information are also collected if any, while visual context of terms is not saved in our dataset. When retrieving relevant terms of a question, we concatenate the names of terms with their definitions for representing each term in the abstract space. It is worth noting that some terms may share the same pages, indicating that they share the same definition. Examples of term and definition are shown in A.4.

To bridge questions and finance terms, three annotators are instructed to identify 1-to-3 relevant finance terms from the knowledge bank to each question in XFINBENCH. For a question, annotators search for relevant terms from those in the same textbook and chapter with this question. A finance term would only be annotated to the question when at least two annotators agree on the high relevance. Finally, a question has 1.3 finance term on average.

B.4 Human Quality Validation

We conduct a comprehensive validation protocol to ensure the high-quality of all annotated examples in XFINBENCH. For each question, we assign three evaluators to validate whether: 1) the question contains complete information in the original question to get the final answer; 2) the final answer is correct given the original answer; 3) the associated knowledge terms are helpful for answering the question. Each criterion is rated individually on a 1-to-5 scale. During this process, human evaluators are accessible to the corresponding after-class questions with gold answers and the knowledge bank, which is different from the close-book setting for human performance during evaluation. Table 6 illustrates the result of quality validation, indicating the high quality of our dataset.

Score	*Question Fluency*	*Question Completeness*	*Answer Correctness*	*Knowledge Helpfulness*
%S = 5	92.9	95.2	96.3	94.1
%S $\geq$ 4	97.1	97.7	98.0	96.8
%S $\geq$ 3	99.4	99.3	99.6	99.8
%S $\geq$ 2	99.4	99.4	99.8	99.9
%S $\geq$ 1	100.0	100.0	100.0	100.0

Table 6: Human evaluation over the test and validation sets of XFINBENCH. Three evaluators are asked to rate the examples on a scale of 1 to 5 individually. In each dimension, we report the proportions of examples with average scores in different ranges.

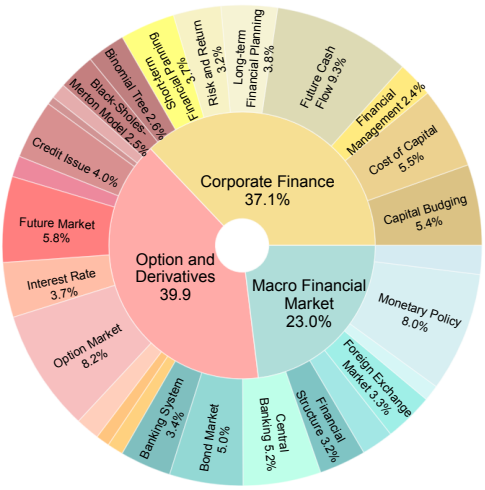


Figure 6: Finance topics distribution of XFINBENCH. Concepts with $\leq 2.5\%$ are not displayed.

C More Dataset Analysis

The distribution of question over finance topics is shown in Figure 6. The distribution of question over test and validation sets are shown in Table 7. The distribution of five capabilities for complex finance problem solving over three tasks are shown in Table 8.

Task	**Test**	**Validation**
Statement judging	1,360	436
Multi-choice question answering	592	169
Financial calculation	1,283	396
Capability	**Test**	**Validation**
Terminology understanding	1,814	582
Temporal reasoning	703	222
Future forecasting	162	44
Scenario planning	246	69
Numerical modelling	557	188

Table 7: Distribution of task and capability in the test and validation set.

Capability	Statement judging	Multi-choice question	Financial calculation
Terminology Understanding	74.7	24.3	1.0
Temporal Reasoning	3.9	6.6	89.5
Future Forecasting	22.8	45.6	31.6
Scenario Planning	3.2	8.3	88.6
Numerical Modelling	0.0	1.2	98.8

Table 8: Distribution of questions in each finance capability (row) across three tasks (column).

D More Experiment Setup

D.1 Model Hyperparamters

The hyperparameters for the experiments are set to their default values unless specified otherwise. The max_tokens is set to be 1024. Table 9 shows the source of models during evaluation.. Additionally, OpenAI Ada embedding used in knowledge augmentation analysis is `text-embedding-ada-002`.

D.2 Human Performance

We conducted a study to evaluate human performance in XFINBENCH. We randomly sampled 1,000 questions from test set of XFINBENCH, with 400 of *statement judging* task, 170 of *multi-choice question answering* task, and 430 of *financial calculation* task. Each question was then assigned to three human experts, all of whom have finance master degrees and have studied the courses covering three textbooks in our source data. None of them is involved in the dataset construction work. The human evaluation is conducted in a close-book setting, and allows standard calculators (not the financial ones). For each question in *statement judging* and *multi-choice question answering* tasks, they must complete each question within five minutes, while in *financial calculation*, the limit is ten minutes due to more reasoning process required in mathematical reasoning.

E More Experiment Results

E.1 Results across Domain Capability

We report the performance of models across five capability required by solving complex, knowledge-intensive finance problems in Table 10.

Model	Source
o1	o1-preview-2024-09-12
o1-mini	o1-mini-2024-09-12
gpt-4o	gpt-4o-2024-05-13
gpt-4o-mini	gpt-4o-mini-2024-07-18
claude-3-5-sonnet	claude-3-5-sonnet -20240620
claude-3-opus	claude-3-opus-20240229
claude-3-haiku	claude-3-haiku-20240307
gemini-1.5-flash	gemini-1.5-flash
gemini-1.5-pro	gemini-1.5-pro
deepseek-chat	deepseek-chat
Llama-3.2-90B-Vision	Meta-Llama-3.2-90B -Vision-Instruct
Llama-3.2-11B-Vision	Meta-Llama-3.2-11B -Vision-Instruct
Llama-3.1-405B	Meta-Llama-3.1-405B -Instruct
Llama-3.1-70B	Meta-Llama-3.1-70B -Instruct
Llama-3.1-8B	Meta-Llama-3.1-8B -Instruct
Llama-3-70B	Meta-Llama-3-70B -Instruct
Llama-3-8B	Meta-Llama-3-8B -Instruct
Mixtral-8 $\times$ 7B	Mixtral-8x7B -Instruct-v0.1

Table 9: Source of models during evaluation.

E.2 Results across Knowledge Augmentation Methods

We report the performance of four models with different retrieving settings in Table 11. We design an evaluation metrics of retrievers, *i.e.*, the accuracy of retrievers locating at least 1 gold terms, annotated by human experts, from the knowledge bank. Dense retriever based on Ada embedding achieve higher accuracy than sparse retriever using BM25 over all tasks, and yield better performance of models under most circumstances. This finding

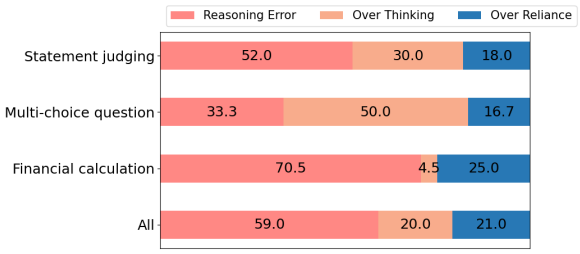


Figure 7: Knowledge augmentation rrror analysis of gpt-4o under Oracle setting across three tasks,

Model	TU	TR	FF	SP	NM
gpt-4o	85.4	41.4	63.6	46.3	58.7
gpt-4o-mini	78.4	33.3	58.6	39.8	48.7
claude-3.5-sonnet	86.4	43.4	63.6	47.2	59.6
claude-3-opus	81.5	36.6	53.1	43.1	53.5
claude-3-haiku	72.5	19.6	40.1	28.5	35.5
gemini-1.5-flash	75.6	25.5	54.3	30.5	43.6
gemini-1.5-pro	78.7	31.7	53.7	37.4	50.4
o1-preview	88.9	59.1	74.7	60.1	66.5
o1-mini	83.0	50.0	65.3	50.8	58.0
Llama-3.1-405B	85.3	29.5	69.5	39.9	46.3
Llama-3.1-8B	68.0	11.7	49.5	22.3	27.1
deepseek-chat	77.7	38.3	63.2	47.1	55.5
Llama-3-70B	79.9	18.8	61.1	32.8	43.9
Human	91.0	79.5	86.2	75.8	78.0

Table 10: Performance of models across five capabilities for complex finance problem solving. TU: Terminology Understanding, TR: Temporal Reasoning, FF: Future Forecasting, SP: Scenario Planning, NM: Numerical Modeling. Figure 1 shares the same data as this table.

illustrates that improving the question-relevance of incorporated knowledge can consistently improve the LLMs' performance. Additionally, we report their performance across five financial capability in *Oracle* setting in Table 12.

Setting	*Statement judging*					*Multi-choice question answering*				
	Retr. Acc	gpt-4o	gpt-4o -mini	Llama-3.1-405B	Llama-3.1-8B	Retr. Acc	gpt-4o	gpt-4o -mini	Llama-3.1-405B	Llama-3.1-8B
w.o. knowledge	0.0	84.0	76.5	83.6	65.3	0.0	91.5	86.8	91.9	77.8
BM25	34.6	86.5	80.7	83.9	69.2	29.7	92.3	89.7	90.8	80.8
Ada Embed.	41.2	85.9	79.6	86.0	69.6	47.9	92.1	90.0	92.0	82.3
Oracle	100.0	85.7	81.1	85.6	69.2	100.0	93.8	90.0	93.4	81.6
Setting	*Financial calculation*					*All*				
	Retr. Acc	gpt-4o	gpt-4o -mini	Llama-3.1-405B	Llama-3.1-8B	Retr. Acc	gpt-4o	gpt-4o -mini	Llama-3.1-405B	Llama-3.1-8B
w.o. knowledge	0.0	31.8	26.5	28.1	12.8	0.0	63.6	57.4	61.9	45.5
BM25	26.8	31.3	27.0	27.8	13.4	30.6	64.6	59.9	61.8	47.9
Ada Embed.	35.3	32.0	26.3	26.2	14.2	39.8	64.6	59.2	62.2	48.6
Oracle	100.0	33.0	27.1	30.3	14.5	100.0	65.2	60.2	64.0	48.5

Table 11: Performance of models augemented with knowledge bank via retrievers. *Oracle* indicates using *ground-truth* terms. Retri. Acc is short for retriever's accuracy score. Results of *financial calculation* task are evaluated using exact-match accuracy score.

Setting	*Terminology understanding*				*Temporal reasoning*			
	gpt-4o	gpt-4o -mini	Llama-3.1-405B	Llama-3.1-8B	gpt-4o	gpt-4o -mini	Llama-3.1-405B	Llama-3.1-8B
w.o. knowledge	85.4	78.4	85.3	68.0	24.6	19.9	16.1	7.9
BM25	87.5	82.4	85.3	71.7	23.9	18.5	14.4	6.1
Ada Embed.	87.3	81.6	84.8	72.2	23.9	19.2	14.3	7.4
Oracle	87.4	82.9	87.9	71.9	24.6	20.8	17.0	10.0
Setting	*Future forecasting*				*Scenario planning*			
	gpt-4o	gpt-4o -mini	Llama-3.1-405B	Llama-3.1-8B	gpt-4o	gpt-4o -mini	Llama-3.1-405B	Llama-3.1-8B
w.o. knowledge	63.6	58.6	70.5	50.5	38.6	33.7	34.5	18.9
BM25	64.8	60.5	75.8	50.5	37.8	35.4	32.4	18.5
Ada Embed.	63.6	58.0	71.6	54.7	38.2	35.8	26.5	21.0
Oracle	61.1	59.3	73.7	51.6	38.2	35.8	32.4	20.6
Setting	*modelling*				*All*			
	gpt-4o	gpt-4o -mini	Llama-3.1-405B	Llama-3.1-8B	gpt-4o	gpt-4o -mini	Llama-3.1-405B	Llama-3.1-8B
w.o. knowledge	42.0	35.7	33.8	19.2	63.6	57.4	61.9	45.5
BM25	41.3	37.3	33.5	17.5	64.6	59.9	61.8	47.9
Ada Embed.	42.0	36.4	34.4	17.2	64.6	59.2	62.2	48.6
Oracle	42.5	34.5	36.7	21.0	65.2	60.2	64	48.5

Table 12: Performance of models augemented with knowledge bank across five capabilities for complex finance problem solving. *Oracle* indicates using *ground-truth* terms. Retri. Acc is short for retriever's accuracy score. Results of *financial calculation* task are evaluated using exact-match accuracy score.

F More Error Analysis

F.1 Human Labeling Guideline

During human labeling process, annotators are provided with the gold answer of the corresponding after-class questions, which include the correct reasoning path. The result of each dimension mentioned in the following paragraphs is decided by at least two annotator's agreement.

Financial Calculation We sampled 400 responses of o1 in *financial calculation* task and assign them to three annotators. Our annotators are asked to determine 1) whether the reasoning path of o1's response coherets with the gold answer of corresponding correct answer; 2) whether there is rounding error in the intermediate calculating steps, *i.e.*, *rounding error*; and 3) whether the formula in o1's response is different from the formulas in the relevant finance terms, *i.e.*, *formula misuse*.

Visual-context question We sampled 100 responses of GPT-4o in visual-context *multi-choice question answering* task and assign them to three annotators. Our annotators are asked to determine 1) whether the reasoning path of GPT-4o's response coherets with the gold answer of corresponding correct answer; 2) if the response shows the model has difficulty identifying the positions and intersections of curves, *i.e.*, *blindness*; and 3) if the response misuses financial knowledge that leads to the error in the following reasoning steps, *i.e.*, *knowledge misuse*.

Knowledge augmentation method We sampled 100 responses of gpt-4o that give wrong final answers in *Oracle* setting, and assign them to three annotators. Our annotators are asked to determine 1) whether the first wrong reasoning step is triggered by the information in the augmented knowledge (reasoning error if no); 2) whether the augmented knowledge proposes direct solution or evidence to answer the corresponding question (over thinking if yes); and 3) whether the wrong reasoning path is led by following every detail in the augmented knowledge (over reliance if yes).

F.2 Error Cases of Knowledge Augmentation

The following three boxes present three examples that demonstrate the three error types in knowledge augmentation method, *i.e.*, Reasoning Error, Over Reliance, and Over Thinking, respectively.

Example of reasoning error in GPT-4o's response in *Oracle* setting. Red sentences are related to the wrong reasoning step.

Question (Financial Calculation task; Numerical Reasoning capability):
The Huff Co. has just gone public. Under a firm commitment agreement, Huff received $15.05 for each of the 5 million shares sold. The initial offering price was $16 per share, and the stock rose to $19.50 per share in the first few minutes of trading. Huff paid $800,000 in direct legal and other costs and $250,000 in indirect costs. What were the total indirect costs incurred by the company? (Unit: dollar)
Term and Definition:
Indirect expenses,Flotation costs. Indirect expenses are not reported on the prospectus and include the costs of management time spent working on the new issue.
Model Response:
To determine the total indirect costs incurred by Huff Co., we only need to consider the provided information about the indirect expenses. The question states that Huff Co. incurred $250,000 in indirect costs. Since no other indirect costs are mentioned, these will be the total indirect costs. Therefore, my answer is [250000].
Correct Answer: 17750000
Analysis:
The wrong reasoning path is not triggered by information from augmented knowledge.

Example of over reliance error in GPT-4o's response in *Oracle* setting. Red sentences are related to the wrong reasoning step while teal sentences enhance or correct the model's reasoning process.

Question (Financial Calculation task; Scenario Planning capability):
A stock price is currently \$40. At the end of one month, it will be either \$42 or \$38. The risk-free interest rate is 8% per annum with continuous compounding. What is the value of a one-month European call option with a strike price of \$39? (Unit: dollar, decimal)

Term and Definition:
One-step,Binomial tree. A no-arbitrage framework for pricing options on a stock is presented, where the stock price is denoted as S_0 and the current option price as f. ..... Rearranging provides: $f = S_0\Delta(1 - ue^{-rT}) + f_u e^{-rT}$. ..... This can be simplified to: $f = \frac{f_u(1-de^{-rT})+f_d(ue^{-rT}-1)}{u-d}$. We also introduce p defined as: $p = \frac{e^{rT}-d}{u-d}$.

Model Response:
To determine the value of a one-month European call option with a strike price of \$39, we will use the provided binomial tree model framework along with the relevant equations. Step-by-Step Solution ..... Determine the value of p using provided Equation: $p = \frac{e^{rT}-d}{u-d} = \frac{1.0067-0.95}{1.05-0.95} = \frac{0.0567}{0.1} = 0.567$. Calculate the value of the call option using another provided Equation: $f = \frac{f_u(1-de^{-rT})+f_d(ue^{-rT}-1)}{u-d} = 1.70595$

Correct Answer:
1.69

Analysis:
The last step calculation that leads to the wrong final answer is not necessarily so complex. The model could simply solve this problem via planning in different scenarios. The expected payoff E is given by: $E = p \cdot 3 + (1 - p) \cdot 0 = 1.701$. The value of the call option C is the discounted expected payoff: $C = E \cdot e^{-rT} = 1.701 \cdot 0.9934 \approx 1.69$.

Example of over thinking error in GPT-4o's response in *Oracle* setting. Red sentences are related to the wrong reasoning step while teal sentences enhance or correct the model's reasoning process.

Question (Statement Judging task; Future Forecasting capability):
Statement: Consider the operating procedures of the Fed in response to movements in output. If the Fed has interest-rate targets, a rise in output that raises interest rates might cause the Fed to sell bonds and bid down their price.

Term and Definition:
Traditional Interest-Rate Channels,Transmission Mechanisms of Monetary Policy. The traditional view of the monetary transmission mechanism illustrates how expansionary monetary policy impacts the economy through a schematic representation: $\Rightarrow i_t \downarrow \Rightarrow l \uparrow \Rightarrow Y \uparrow$.An expansionary monetary policy results in a decrease in real interest rates ($i_t, \downarrow$), which lowers the cost of capital, thus increasing investment spending ($l \uparrow$) and boosting aggregate demand, leading to a rise in output ($Y \uparrow$). Under interest-rate target, the government could hold back the changes on interest rate via open market operations. For example, selling bonds would cut down the interest rate due to the increasing money supply.

Model Response:
Let's break down the given statement and consider .....3. Fed's Response with Interest-Rate Targeting: If the Fed observes a rise in interest rates due to increased output but desires to maintain its interest-rate target, it would engage in open market operations to influence these rates. 4.Selling Bonds to Lower Prices and Raise Rates: The Fed sells bonds in open market operations to reduce the money supply. Selling bonds bids down their price, which inversely raises the yield or interest rates. This action can counter rising rates if the objective is to maintain higher interest rates to dampen potential inflationary pressures. Putting it together..... Therefore, my answer is true.

Correct Answer:
No

Analysis:
After the third reasoning step, the correct answer could be proposed. However, the model thinks one step further to consider other influencing factors in the long run, and finally give the wrong final answer. Note that the third reasoning step highly follows the part of augmented knowledge highlighted in blue.

G Prompt Templates

G.1 Sensitivity Analysis

We conduct sensitivity analysis on prompt templates for evaluation on XFINBENCH. ProSA (Zhuo et al., 2024) showcases four different styles of constructing prompts, *i.e.*, *simple input* (SI), *emotional support* (ES), *role player* (RP) and *output requirement* (OR). We further include two common prompting strategies, *i.e.*, *chain-of-though* (CoT) and *direct answering* (DA). Hence, we design four types of prompt templates for conducting our sensitivity analysis, as shown in Table 13. Note that *output requirement* is indispensable in our tasks for automatic evaluating the model's final answers.

We randomly sample 500 examples from the test set of XFINBENCH and use them to evaluate four models on each of prompt templates mentioned above. Results in Table 14 show that the prompt template involving CoT, RP and OR consistently brings out the best performance of most models with slight margins, while the rankings of four models hardly change over four styles of prompt templates. Hence, we design our prompt templates based on the three components, *i.e.*, *chain-of-though*, *role player* and *output requirement*.

Capability	**Task**
CoT & RP & OR	You are a financial expert. You are supposed to answer the given question.\n Question: {`after-class question`}\n Please answer the above question and output your final answer starting with 'Therefore, my answer is' at the end, where you store you final answer into '[]'.\n Let's think step by step.\n
DA & RP & OR	You are a financial expert. You are supposed to answer the given question.\n Question: {`after-class question`}\n Please answer the above question and output your final answer starting with 'Therefore, my answer is' at the end, where you store you final answer into '[]'.\n
CoT & OR	Question: {`after-class question`}\n Please answer the above question and output your final answer starting with 'Therefore, my answer is' at the end, where you store you final answer into '[]'.\n Let's think step by step.\n
DA & OR	Question: {`after-class question`}\n Please answer the above question and output your final answer starting with 'Therefore, my answer is' at the end, where you store you final answer into '[]'.\n

Table 13: Four prompt templates for sensitivity analysis during evaluation.

G.2 Prompt for Dataset Construction

We apply the generate-then-verify paradigm for constructing our dataset. Prompts used in the generate-then-verify paradigm for *statement judging*, *multi-choice question answering*, and *financial calculation* tasks, are shown in G.2.1, G.2.2, and G.2.3, respectively.

G.2.1 Prompt for Statement Judging Task

Models	gpt-4o	gpt-4o -mini	Llama-3.1-405B	Llama-3.1-8B
CoT & RP & OR	56.6	47.5	48.4	35.2
DA & RP & OR	56.0	44.8	46.8	35.0
CoT & OR	55.2	46.8	49.8	32.2
DA & OR	53.4	42.8	49.6	33.6

Table 14: Performance of models using different prompt templates during evaluation. Results in *financial calculation* task is reported in exact-match accuracy.

Prompt template for generating true statements in statement judging task.

Please rewrite a question-answer pair into one or more statement(s) which is/are true. Specifically,
1. The statement(s) should be generated from the original question-answer pair and must be true given the content of the question-answer pair.
2. None of the following expressions is allowed in the statement: (1) unclear pronoun; (2) in/given/according to the chapter/figure/table; (3) conjunctions of causality like since, because and so on.
3. You should extract the context of the original question. The context usually introduces the background of the generated statement(s). Note that: (1) the context must NOT be question; (2) there should NOT be duplicated or contradictory information between the context and the statement.
4. You are allowed to generate two or more statements from one question-answer pair. Under this case, the statements should be independent of each other, with as little overlap as possible.

Example 1:
{`example 1`}

Example 2:
{`example 2`}

Example 3:
{`example 3`}

Example 4:
{`example 4`}

Example 5:
{`example 5`}

Given the above instructions and examples, please use the following question-answer pair to generate at least one statement with a clear answer and context.
Original Question: {`after-class question`}
Original Answer: {`after-class solution`}

Example 1 in prompt template for generating true statements in statement judging task.

Original Question: Suppose that a bond portfolio with a duration of 12 years is hedged using a futures contract in which the underlying asset has a duration of four years. What is likely to be the impact on the hedge of the fact that the 12-year rate is less volatile than the four-year rate?

Original Answer: Duration-based hedging procedures assume parallel shifts in the yield curve. Since the 12-year rate tends to move by less than the 4-year rate, the portfolio manager may find that he or she is over-hedged.

Context: Suppose that a bond portfolio with a duration of 12 years is hedged using a futures contract in which the underlying asset has a duration of four years.

Statement: Considering duration-based hedging procedures assume parallel shifts in the yield curve, the portfolio manager may find that he or she is over-hedged.

Answer: True

Example 2 in prompt template for generating true statements in statement judging task.

Original Question: What is meant by the delta of a stock option?

Original Answer: The delta of a stock option measures the sensitivity of the option price to the price of the stock when small changes are considered. Specifically, it is the ratio of the change in the price of the stock option to the change in the price of the underlying stock.

Context: Consider the delta of a stock option.

Statement: The delta of a stock option measures the sensitivity of the option price to the price of the stock when small changes are considered.

Answer: True

Statement: The delta of a stock option is the ratio of the change in the price of the stock option to the change in the price of the underlying stock.

Answer: True

Example 3 in prompt template for generating true statements in statement judging task.

Original Question: The treasurer of a corporation is trying to choose between options and forward contracts to hedge the corporation's foreign exchange risk. Discuss the advantages and disadvantages of each.

Original Answer: Forward contracts lock in the exchange rate that will apply to a particular transaction in the future. Options provide insurance that the exchange rate will not be worse than some level. The advantage of a forward contract is that uncertainty is eliminated as far as possible. The disadvantage is that the outcome with hedging can be significantly worse than the outcome with no hedging. This disadvantage is not as marked with options. However, unlike forward contracts, options involve an up-front cost.

Context: The treasurer of a corporation is trying to choose between options and forward contracts to hedge the corporation's foreign exchange risk.

Statement: Options provide insurance that the exchange rate will not be worse than some level.
Answer: True

Statement: When using forward contracts for hedging, the outcome can be worse than the outcome with no hedging.
Answer: True

Statement: Using options to hedge the foreign exchange risk involve an up-front cost.
Answer: True

Example 4 in prompt template for generating true statements in statement judging task.

Original Question: The term structure of interest rates is upward sloping. Put the following in order of magnitude:
(a) The five-year zero rate
(b) The yield on a five-year coupon-bearing bond
(c) The forward rate corresponding to the period between 4.75 and 5 years in the future
What is the answer to this question when the term structure of interest rates is downward sloping?

Original Answer: When the term structure is upward sloping, $c > a > b$. When it is downward sloping, $b > a > c$.

Context: The term structure of interest rates is upward sloping.

Statement: The five-year zero rate is smaller than the forward rate corresponding to the period between 4.75 and 5 years in the future.
Answer: True

Statement: The yield on a five-year coupon-bearing bond is smaller than the forward rate corresponding to the period between 4.75 and 5 years in the future.
Answer: True

Statement: The yield on a five-year coupon-bearing bond is larger than the five-year zero rate.
Answer: True

Statement: The five-year zero rate is larger than the forward rate corresponding to the period between 4.75 and 5 years in the future.
Answer: True

Example 5 in prompt template for generating true statements in statement judging task.

Original Question: For each of the following scenarios, discuss whether profit opportunities exist from trading in the stock of the firm under the conditions that (1) the market is not weak form efficient, (2) the market is weak form but not semistrong form efficient, (3) the market is semistrong form but not strong form efficient, and (4) the market is strong form efficient. **a.** The stock price has risen steadily each day for the past 30 days. **b.** The financial statements for a company were released three days ago, and you believe you've uncovered some anomalies in the company's inventory and cost control reporting techniques that are causing the firm's true liquidity strength to be understated. **c.** You observe that the senior managers of a company have been buying a lot of the company's stock on the open market over the past week.

Original Answer:
(a). If the market is not weak form efficient, then this information could be acted on and a profit earned from following the price trend. Under (2), (3), and (4), this information is fully impounded in the current price and no abnormal profit opportunity exists.
(b). Under (2), if the market is not semi-strong form efficient, then this information could be used to buy the stock c̈heapb̈efore the rest of the market discovers the financial statement anomaly. Since (2) is stronger than (1), both imply that a profit opportunity exists; under (3) and (4), this information is fully impounded in the current price and no profit opportunity exists.
(c). Under (3), if the market is not strong form efficient, then this information could be used as a profitable trading strategy, by noting the buying activity of the insiders as a signal that the stock is underpriced or that good news is imminent. Since (1) and (2) are weaker than (3), all three imply that a profit opportunity exists. Note that this assumes the individual who sees the insider trading is the only one who sees the trading. If the information about the trades made by company management is public information, it will be discounted in the stock price and no profit opportunity exists. Under (4), this information does not signal any profit opportunity for traders; any pertinent information the manager-insiders may have is fully reflected in the current share price.

Context: Consider profit opportunities exist from trading in the stock of the firm.

Statement: In a market that is not weak form efficient, a profit could be earned from acting on the information of a stock price that has risen steadily each day for the past 30 days.
Answer: True

Statement: In a market that is not semi-strong form efficient, a profit could be earned from acting on the pertinent information the manager-insiders may have.
Answer: True

Statement: In a market that is not strong form efficient, there is no profit opportunity on the information that you observe that the senior managers of a company have been buying a lot of the company's stock on the open market over the past week.
Answer: True

Prompt template for generating false statements in statement judging task.

Please rewrite a question-answer pair into one or more statement(s) which is/are false. Specifically,
1. The statement(s) should be generated from the original question-answer pair and must be false given the content of the question-answer pair.
2. None of the following expressions is allowed in the statement: (1) unclear pronoun; (2) in/given/according to the chapter/figure/table; (3) conjunctions of causality like since, because and so on.
3. You should extract the context of the original question. The context usually introduces the background of the generated statement(s). Note that: (1) the context must NOT be question; (2) there should NOT be duplicated or contradictory information between the context and the statement.
4. You are allowed to generate two or more statements from one question-answer pair. Under this case, the statements should be independent of each other, with as little overlap as possible.

Example 1:
{`example 1`}

Example 2:
{`example 2`}

Example 3:
{`example 3`}

Example 4:
{`example 4`}

Example 5:
{`example 5`}

Given the above instructions and examples, please use the following question-answer pair to generate at least one statement with a clear answer and context.
Original Question: {`after-class question`}
Original Answer: {`after-class solution`}

Prompt template for deduplicating dependent statements in *statement judging* task.

Context of Statements: `{context}`
Statement 1: `{true statement}`
Statement 2: `{false statement}`
Please determine whether Statement 1 provides direct evidence to support that Statement 2 is false.
Your Answer (Yes or No):

Example 1 in prompt template for generating false statements in statement judging task.

Original Question: Suppose that a bond portfolio with a duration of 12 years is hedged using a futures contract in which the underlying asset has a duration of four years. What is likely to be the impact on the hedge of the fact that the 12-year rate is less volatile than the four-year rate?

Original Answer: Duration-based hedging procedures assume parallel shifts in the yield curve. Since the 12-year rate tends to move by less than the 4-year rate, the portfolio manager may find that he or she is over-hedged.

Context: Suppose that a bond portfolio with a duration of 12 years is hedged using a futures contract in which the underlying asset has a duration of four years.

Statement: Considering duration-based hedging procedures assume parallel shifts in the yield curve, the portfolio manager may find that he or she is under-hedged.
Answer: False

Example 2 in prompt template for generating false statements in statement judging task.

Original Question: What is meant by the delta of a stock option?

Original Answer: The delta of a stock option measures the sensitivity of the option price to the price of the stock when small changes are considered. Specifically, it is the ratio of the change in the price of the stock option to the change in the price of the underlying stock.

Context: Consider the delta of a stock option.

Statement: The delta of a stock option measures the sensitivity of the option price to the price of the stock when big changes are considered.
Answer: False

Example 3 in prompt template for generating false statements in statement judging task.

Original Question: The treasurer of a corporation is trying to choose between options and forward contracts to hedge the corporation's foreign exchange risk. Discuss the advantages and disadvantages of each.

Original Answer: Forward contracts lock in the exchange rate that will apply to a particular transaction in the future. Options provide insurance that the exchange rate will not be worse than some level. The advantage of a forward contract is that uncertainty is eliminated as far as possible. The disadvantage is that the outcome with hedging can be significantly worse than the outcome with no hedging. This disadvantage is not as marked with options. However, unlike forward contracts, options involve an up-front cost.

Context: The treasurer of a corporation is trying to choose between options and forward contracts to hedge the corporation's foreign exchange risk.

Statement: When using forward contracts for hedging, the outcome is definitely better than the outcome with no hedging.
Answer: False

Statement: Using forward contracts to hedge the foreign exchange risk involve an up-front cost.
Answer: False

Example 4 in prompt template for generating false statements in statement judging task.

Original Question: The term structure of interest rates is upward sloping. Put the following in order of magnitude:
(a) The five-year zero rate
(b) The yield on a five-year coupon-bearing bond
(c) The forward rate corresponding to the period between 4.75 and 5 years in the future
What is the answer to this question when the term structure of interest rates is downward sloping?

Original Answer: When the term structure is upward sloping, $c > a > b$. When it is downward sloping, $b > a > c$.

Context: The term structure of interest rates is upward sloping.

Statement: The five-year zero rate is larger than the forward rate corresponding to the period between 4.75 and 5 years in the future.
Answer: False

Statement: The yield on a five-year coupon-bearing bond is larger than the forward rate corresponding to the period between 4.75 and 5 years in the future.
Answer: False

Statement: When it is downward sloping, the yield on a five-year coupon-bearing bond is smaller than the five-year zero rate.
Answer: False

Statement: When it is downward sloping, The five-year zero rate is smaller than the forward rate corresponding to the period between 4.75 and 5 years in the future.
Answer: False

Example 5 in prompt template for generating false statements in statement judging task.

Original Question: For each of the following scenarios, discuss whether profit opportunities exist from trading in the stock of the firm under the conditions that (1) the market is not weak form efficient, (2) the market is weak form but not semistrong form efficient, (3) the market is semistrong form but not strong form efficient, and (4) the market is strong form efficient. **a.** The stock price has risen steadily each day for the past 30 days. **b.** ... **c.**...

Original Answer:
(a). If the market is not weak form efficient, then this information could be acted on and a profit earned from following the price trend. Under (2), (3), and (4), this information is fully impounded in the current price and no abnormal profit opportunity exists.
(b). ...
(c). ...

Context: Consider profit opportunities exist from trading in the firm's stock.

Statement: In a market that is weak form efficient but not semistrong form efficient, a profit could be earned from acting on the information of a stock price that has risen steadily each day for the past 30 days.
Answer: False

Statement: In a market that is strong form efficient, a profit could be earned from acting on the pertinent information the manager-insiders may have.
Answer: False

Statement: In a market that is semistrong form but not strong form efficient, there is no profit opportunity on the information that you observe that the senior managers of a company have been buying a lot of the company's stock on the open market over the past week.
Answer: False

Prompt template for verifying true statements in *statement judging* task.

Original Question: `{after-class question}`
Original Answer: `{after-class solution}`
Context of Statement: `{context}`
Statement: `{question}`
Given the above original question and answer, please answer the following two questions.
Q1: Is the statement definitely true given the original question and answer?
Q2: Does the context extract the essential background information in the original question?
Your Answer to Q1 and Q2 (Yes or No, no explanation required):

Prompt template for verifying false statements in *statement judging* task.

Original Question: `{after-class question}`
Original Answer: `{after-class solution}`
Context of Statement: `{context}`
Statement: `{question}`
Given the above original question and answer, please answer the following two questions.
Q1: Is the statement definitely false given the original question and answer?
Q2: Does the context extract the essential background information in the original question?
Your Answer to Q1 and Q2 (Yes or No, no explanation required):

G.2.2 Prompt for Multi-choice Question Answering Task

Example 3 in prompt template for generating questions in multi-choice question answering task.

Original Question: 'When a bank is negotiating currency swaps, it should try to ensure that it is receiving the lower interest rate currency from a company with a low credit risk.' Explain.

Original Answer: As time passes there is a tendency for the currency which has the lower interest rate to strengthen. This means that a swap where we are receiving this currency will tend to move in the money (i.e., have a positive value). Similarly a swap where we are paying the currency will tend to move out of the money (i.e., have a negative value). From this it follows that our expected exposure on the swap where we are receiving the low-interest currency is much greater than our expected exposure on the swap where we are receiving the high-interest currency. We should therefore look for counterparties with a low credit risk on the side of the swap where we are receiving the low-interest currency. On the other side of the swap we are far less concerned about the creditworthiness of the counterparty.

Context: A bank is negotiating currency swaps.

Generated Question: Which one of the following actions should it consider?
Choices:
(a) Seek counterparties with low credit risk where the bank is receiving the low-interest currency
(b) Seek counterparties with high credit risk where the bank is receiving the low-interest currency
(c) Seek counterparties with low credit risk where the bank is receiving the high-interest currency
Correct Answer: a

Example 4 in prompt template for generating questions in multi-choice question answering task.

Original Question: How can bank behavior and the Fed's behavior cause money supply growth to be precyclical (rising in booms and falling in recessions)?

Original Answer: Bank behavior can lead to procyclical money growth because when interest rates rise in a boom, they decrease excess reserves and increase their borrowing from the Fed, both of which lead to a higher money supply. Similarly, when interest rates fall in a recession, they increase excess reserves and decrease their borrowing from the Fed, leading to a lower money supply. The result is that the money supply will tend to grow faster in booms and slower in recessions–it is procyclical. Fed behavior also can lead to procyclical money growth because (as the answer to problem 1 indicates) an interest-rate target can lead to a slower rate of growth of the money supply during recessions and a more rapid rate of growth during booms.

Context: Bank behavior and the Fed's behavior can cause money supply growth to be precyclica.

Generated Question: Which one of the following bank and/or the Fed's behaviours would happen when interest rates rise in a boom?
Choices:
(a) Banks increase their borrowings from the Fed
(b) Banks increase excess reserves
(c) The Fed's make positive announcements
Correct Answer: a

Generated Question: Which one of the following bank and/or the Fed's behaviours would happen when interest rates rise in a recession?
Choices:
(a) Banks decrease their borrowings from the Fed
(b) Banks decrease excess reserves
(c) The Fed's make positive announcements
Correct Answer: a

Prompt template for generating questions in multi-choice question answering task.

Please rewrite a question-answer pair into one or more question(s) with three candidate choices. Specifically,
1. The question and correct answer should be generated from the question and/or answer, under a clear and concise wording style. None of the following expressions is allowed in the question: (1) unclear pronoun; (2) in/given/according to the chapter/figure/table.
2. There are three candidate choices for the question. The correct answer lies in Choice (a), and Choice (b) and (c) are both wrong to the question. Choice (a), (b) and (c), should be independent and mutually exclusive. Noising choices, i.e. (b) and (c), should share the similar wording and length with the correct answer (a). Choice (b) reflects a misunderstanding of the original question-answer pair, while Choice (c) is made up by you.
3. You should extract the context of the original question. The context usually introduces the background of the generated question(s). Note that: (1) the context must NOT be question; (2) there should NOT be duplicated or contradictory information between the context and the statement.
4. You are allowed to generate two or more questions from one original question-answer pair. Under this case, the questions should be independent of each other, with as little overlap as possible.

Example 1:
`{example 1}`

Example 2:
`{example 2}`

Example 3:
`{example 3}`

Example 4:
`{example 4}`

Example 5:
`{example 5}`

Given the above instructions and examples, please use the following question-answer pair to generate at least one question with candidate choices and context.
Original Question: `{after-class question}`
Original Answer: `{after-class solution}`

Example 1 in prompt template for generating questions in multi-choice question answering task.

Original Question: Last month, BlueSky Airline announced that it would stretch out its bill payments to 45 days from 30 days. The reason given was that the company wanted to c̈ontrol costs and optimize cash flow.T̈he increased payables period will be in effect for all of the company's 4,000 suppliers. Why don't all firms simply increase their payables periods to shorten their cash cycles?

Original Answer: They would like to! The payables period is a subject of much negotiation, and it is one aspect of the price a firm pays its suppliers. A firm will generally negotiate the best possible combination of payables period and price. Typically, suppliers provide strong financial incentives for rapid payment. This issue is discussed in detail in a later chapter on credit policy.

Context: Last month, BlueSky Airline announced that it would stretch out its bill payments to 45 days from 30 days.

Generated Question: Which one of the following choices is one of the reasons of BlueSky Airline announcement?
Choices:
(a) Optimize cash flow
(b) Increase investment in fixed costs
(c) Increase sales volume
Correct Answer: a

Example 2 in prompt template for generating questions in multi-choice question answering task.

Original Question: What are the advantages of using the DCF model for determining the cost of equity capital? What are the disadvantages? What specific piece of information do you need to find the cost of equity using this model? What are some of the ways in which you could get this estimate?

Original Answer: The primary advantage of the DCF model is its simplicity. The method is disadvantaged in that (1) the model is applicable only to firms that actually pay dividends; many do not; (2) even if a firm does pay dividends, the DCF model requires a constant dividend growth rate forever; (3) the estimated cost of equity from this method is very sensitive to changes in g, which is a very uncertain parameter; and (4) the model does not explicitly consider risk, although risk is implicitly considered to the extent that the market has impounded the relevant risk of the stock into its market price. While the share price and most recent dividend can be observed in the market, the dividend growth rate must be estimated. Two common methods of estimating g are to use analysts' earnings and payout forecasts or to determine some appropriate average historical g from the firm's available data.

Context: The DCF model have advantages and disadvantages for determining the cost of equity capital.

Generated Question: Which one of the following advantages do the DCF model have?
Choices:
(a) Simple calculation
(b) Applicable for firms that do not pay dividends
(c) Insensitivity to the financial environment
Correct Answer: a

Example 5 in prompt template for generating questions in multi-choice question answering task.

Original Question: Which regulatory agency has the primary responsibility for supervising the following categories of commercial banks? a. National banks; b. Bank holding companies; c. Non-Federal Reserve member state banks; d. Federal Reserve member state banks

Original Answer: (a) Office of the Comptroller of the Currency; (b) the Federal Reserve; (c) state banking authorities and the FDIC; (d) the Federal Reserve

Context: Regulatory agencies have the primary responsibility for supervising commercial banks.

Generated Question: Which one of the following agencies has the primary responsibility for supervising national banks?
Choices:
(a) Office of the Comptroller of the Currency
(b) state banking authorities
(c) the Bank of Settlement
Correct Answer: a

Generated Question: Which one of the following agencies has the primary responsibility for supervising non-Federal Reserve member state banks?
Choices:
(a) state banking authorities and the FDIC
(b) the Federal Reserve
(c) the National Credit Union Administration
Correct Answer: a

Generated Question: Which one of the following agencies has the primary responsibility for supervising Federal Reserve member state banks?
Choices:
(a) the Federal Reserve
(b) the FDIC
(c) Financial Stability Oversight Council
Correct Answer: a

Prompt template for verifying questions in *multi-choice question answering* task.

Original Question: `{after-class question}`
Original Answer: `{after-class solution}`
Context of Generated Question: `{context}`
Generated Question: `{question}`
Candidate Choices:`{choices}`
Correct Answer: `{answer}`

Given the above original question and answer, please answer the following two questions.
Q1: Is the correct answer definitely true to the generated question?
Q2: Are the other two misleading answers within candidate choices definitely false to the generated question?
Q3: Are the three candidate choices mutually exclusive but sharing the similar wording and length with each other?
Q4: Does the context extract the essential background information in the original question?
Your Answer to Q1, Q2, Q3 and Q4 (Yes or No, no explanation required):

G.2.3 Prompt for Financial Calculation Task

Example 3 in prompt template for generating questions in financial calculation task.

Original Question: You want to buy a new sports coupe for $61,800, and the finance office at the dealership has quoted you a 7.4 percent APR loan for 60 months to buy the car. What will your monthly payments be? What is the effective annual rate on this loan?

Original Answer: We first need to find the annuity payment. We have the PVA, the length of the annuity, and the interest rate. Using the PVA equation:

$$PVA = C([1 - [1/(1 + r)]^t]r)$$

$$\$61,800 = C[1-[1/[1+(.074/12)]^{60}]](.074/12)]$$

Solving for the payment, we get:

$$C = \$61,800/50.02385 = \$1,235.41$$

To find the EAR, we use the EAR equation:

$$EAR = [1 + (APR/m)]^m - 1$$

$$EAR = [1 + (.074/12)]^{12} - 1 = .0766$$

Context: You want to buy a new sports coupe for $61,800, and the finance office at the dealership has quoted you a 7.4 percent APR loan for 60 months to buy the car.

Generated Question: What will your monthly payments be? (Unit: dollar)
Answer: 1235.41

Generated Question: What is the effective annual rate on this loan?
Answer: 0.0766

Example 4 in prompt template for generating questions in financial calculation task.

Original Question: What is the value of an investment that pays $7,500 every other year forever, if the first payment occurs one year from today and the discount rate is 11 percent compounded daily? What is the value today if the first payment occurs four years from today?

Original Answer: The cash flows in this problem occur every two years, so we need to find the effective two year rate. ... So, the effective two-year interest rate is: Effective 2-year rate $= [1 + (.11/365)]^{[} 365(2)] - 1 = .2460$
We can use this interest rate to find the PV of the perpetuity. Doing so, we find: PV $= \$7,500/.2460 = \$30,483.41$
Remember that the PV equation for a perpetuity (and an ordinary annuity) tells you the PV one period before the first cash flow. In this problem, since the cash flows are two years apart, we have found the value of the perpetuity one period (two years) before the first payment, which is one year ago. We need to compound this value for one year to find the value today. The value of the cash flows today is: PV $= \$30,483.41(1 + .11/365)^{365} = \$34,027.40$ The second part of the question assumes the perpetuity cash flows begin in four years. In this case, when we use the PV of a perpetuity equation, we find the value of the perpetuity two years from today. So, the value of these cash flows today is: PV $= \$30,483.41/(1 + .11/365)^{2(365)} = \$24,464.32$ Context: An investment pays $7,500 every other year forever. The discount rate is 11 percent compounded daily.

Generated Question: What is the value of the investment if the first payment occurs one year from today? (Unit: dollar)
Answer: 34027.40

Generated Question: What is the value of the investment if the first payment occurs four year from today? (Unit: dollar)
Answer: 24464.32

Prompt template for generating questions in financial calculation task.

Please rewrite a question-answer pair into one or more question(s) with clear answer(s). Specifically,
1. The question should be generated from the original question-answer pair and written in a clear and concise wording style. The question should clarify the unit for its answer at the end if any.
2. The answer MUST be pure numbers from the original answer without any symbol attached. Specifically, it should be in decimal form and have no special symbols like percent sign and currency symbols.
3. You should extract the context of the original question. The context usually contains the necessary details for calculation, and serves as the background of the generated question(s). Note that: (1) the context must NOT be question; (2) there should NOT be duplicated or contradictory information between the context and the statement.
4. You are allowed to generate two or more questions from one question-answer pair, each with a answer. Under this case, the questions should be independent of each other. It is not allowed that the answer to any questions is an intermediate step to other questions.

Example 1:
`{example 1}`

Example 2:
`{example 2}`

Example 3:
`{example 3}`

Example 4:
`{example 4}`

Example 5:
`{example 5}`

Given the above instructions and examples, please use the following question-answer pair to generate at least one question with a clear answer and context.
Original Question: `{after-class question}`
Original Answer: `{after-class solution}`

Example 5 in prompt template for generating questions in financial calculation task.

Original Question: An investment offers $4,600 per year for 15 years, with the first payment occurring one year from now. If the required return is 8 percent, what is the value of the investment? What would the value be if the payments occurred for 40 years? For 75 years? Forever?

Original Answer: To find the PVA, we use the equation:
$PVA = C([1\breve{}[1/(1+r)]^t]/r)$
PVA@15 yrs:
$\$4,600[[1\breve{}(1/1.08)^{15}]/.08]$
$= \$39,373.60$
PVA@40 yrs:
$\$4,600[[1\breve{}(1/1.08)^{40}]/.08]$
$= \$54,853.22$
PVA@75 yrs:
$\$4,600[[1\breve{}(1/1.08)^{75}]/.08]$
$= \$57,320.99$
To find the PV of a perpetuity, we use the equation:
$PV = C/r$
$PV = \$4,600/.08 = \$57,500.00$
Notice that ... all perpetuity payments beyond 75 years is only $179.01.

Context: An investment offers $4,600 per year for 15 years, with the first payment occurring one year from now. The required return is 8 percent

Generated Question: What is the value of the investment? (Unit: dollar)
Answer: 39373.60

Generated Question: If the payments occurred for 40 years, what is the value of the investment? (Unit: dollar)
Answer: 54853.22

Generated Question: If the payments occurred for 75 years, what is the value of the investment? (Unit: dollar)
Answer: 57320.99

Generated Question: If the payments occurred forever, what is the value of the investment? (Unit: dollar)
Answer: 57500.00

Prompt template for verifying questions in *financial calculation* task.

Original Question: `{after-class question}`
Original Answer: `{after-class solution}`
Context of Generated Question: `{context}`
Generated Question: `{question}`
Correct Answer: `{answer}`

Given the above original question and answer, please answer the following two questions.
Q1: Is the correct answer definitely true to the generated question?
Q2: Does the context provide the necessary information for the calculation to answer the generated question?

Your Answer to Q1 and Q2 (Yes or No, no explanation required):

Example 1 in prompt template for generating questions in financial calculation task.

Original Question: A credit default swap requires a semiannual payment at the rate of 60 basis points per year. The principal is \$300 million and the credit default swap is settled in cash. A default occurs after four years and two months, and the calculation agent estimates that the price of the cheapest deliverable bond is 40% of its face value shortly after the default. List the cash flows and their timing for the seller of the credit default swap.

Original Answer: The seller receives

$$300,000,000 \times 0.0060 \times 0.5 = \$900,000$$

at times 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, and 4.0 years. The seller also receives a final accrual payment of about

$$\$300,000 (= \$300,000,000 \times 0.060 \times 2/12)$$

at the time of the default (4 years and two months). The seller pays

$$300,000,000 \times 0.6 = \$180,000,000$$

at the time of the default. (This does not consider day count conventions.)

Context: A credit default swap requires a semiannual payment at the rate of 60 basis points per year. The principal is \$300 million and the credit default swap is settled in cash. A default occurs after four years and two months, and the calculation agent estimates that the price of the cheapest deliverable bond is 40% of its face value shortly after the default.

Generated Question: What is the cash paid by the seller at the time of the default? (Unit: dollar)
Answer: 180000000.00

Example 2 in prompt template for generating questions in financial calculation task.

Original Question: Calculate the price of a three-month American put option on a non-dividend-paying stock when the stock price is \$60, the strike price is \$60, the risk-free interest rate is 10% per annum, and the volatility is 45% per annum. Use a binomial tree with a time interval of one month.

Original Answer: In this case, $S_0 = 60$, $K = 60$, $r = 0.1$, $\sigma = 0.45$, $T = 0.25$, and $\Delta t = 0.0833$. Also

$$u = e^{\sigma/\Delta t} = e^{0.45\sqrt{0.0833}} = 1.1387$$

$$d = \frac{1}{u} = 0.8782$$

$$a = e^{r\Delta t} = e^{0.1 \cdot 0.0833} = 1.0084$$

$$p = \frac{a-d}{u-d} = 0.4998$$

$$1 - p = 0.5002$$

The output from DerivaGem for this example is shown in the Figure S21.1. The calculated price of the option is \$5.16.
Figure S21.1: Tree for Problem 21.2
Context: Here is a three-month American put option on a non-dividend-paying stock. Suppose the stock price is \$60, the strike price is \$60, the risk-free interest rate is 10% per annum, and the volatility is 45% per annum.

Generated Question: What is the price of this put option using a binomial tree with a time interval of one month?
Answer: 5.16

G.3 Prompt for Evaluating Baselines

Chain-of-thought prompt templates for evaluating baselines are shown in G.3.1. The program-of-thought prompt template for financial calculation task is shown in G.3.2.

G.3.1 Prompt for Chain-of-Thought Method

Prompt template for evaluation in *statement judging* task using CoT prompting. `knowledge` is an empty string when no finance term is provided.

```
{knowledge}

Statement: {question}
Is the above statement true or false? Please output your answer starting with 'Therefore, my answer is' at the end.
Let's think step by step.
```

Prompt template for evaluation in *multi-choice question answering* task using CoT prompting. `knowledge` is an empty string when no finance term is provided.

```
{knowledge}

Question: {question}
Choices: {choices}
Which one of the above choices is the most appropriate to answer the question? Please output your answer starting with 'Therefore, my answer is' at the end.
Let's think step by step.
```

Prompt template for evaluation in *financial calculation* task using CoT prompting. `knowledge` is an empty string when no finance term is provided.

```
{knowledge}

Question: {question}
Please answer the above question and output your final answer starting with 'Therefore, my answer is' at the end, where you store you final answer into '[]'.
Let's think step by step.
```

G.3.2 Prompt for Program-of-Thought Method

Prompt template for evaluation in *financial calculation* task using PoT prompting. `knowledge` is an empty string when no finance term is provided.

```
{knowledge}

Question: {question}
Please generate a Python program to answer the given question.
```python
def solution():
```
```

H Ethics and Societal Impact

We envision XFINBENCH as a comprehensive benchmark designed to assist researchers in evaluating the performance of their models within the finance domain. By offering a robust evaluation framework, XFINBENCH aims to drive advancements in foundational models for the research community, providing valuable insights into critical model capabilities such as temporal reasoning, future forecasting, scenario planning, numerical modeling, and cross-modal reasoning.

For constructing examples in XFINBENCH and finance terms for the knowledge bank, we primarily rely on textbooks that are openly available on the internet. Our annotators strictly adhere to copyright and licensing regulations, ensuring that data from sources prohibiting copying or redistribution is excluded. Furthermore, during the automated annotation and human quality validation processes for examples in XFINBENCH, we implement rigorous ethical guidelines to prevent biased content and safeguard against the inclusion of private data.