

RealFin: How Well Do LLMs Reason About Finance When Users Leave Things Unsaid?

Yuyang Dai^{1*}, Yan Lin^{1,2*}, Zhuohan Xie³, Yuxia Wang¹

¹INSAIT, Sofia University “St. Kliment Ohridski”

²Newcastle University ³MBZUAI

y9657422@gmail.com, y.lin64@ncl.ac.uk

zhuohan.xie@mbzuai.ac.ae, yuxia.wang@insait.ai

Abstract

Reliable financial reasoning requires knowing not only how to answer, but also when an answer cannot be justified. In real financial practice, problems often rely on implicit assumptions that are taken for granted rather than stated explicitly, causing problems to appear solvable while lacking enough information for a definite answer. We introduce REALFIN, a bilingual benchmark that evaluates financial reasoning by systematically removing essential premises from exam-style questions while keeping them linguistically plausible. Based on this, we evaluate models under three formulations that test answering, recognizing missing information, and rejecting unjustified options, and find consistent performance drops when key conditions are absent. General-purpose models tend to over-commit and guess, while most finance-specialized models fail to clearly identify missing premises. These results highlight a critical gap in current evaluations and show that reliable financial models must know when a question should not be answered. The dataset and code are available at <https://github.com/xnlp-lab/RealFin>.

1 Introduction

A central requirement of reliable reasoning in high-stakes domains is not only answering correctly when a task is well-posed, but also recognizing when the available information does not justify a determinate conclusion. In real-world financial problems, key premises needed for decision-making, such as assumptions, constraints, time horizons, or applicable regulatory and accounting standards, are often missing or unspecified in users’ initial requests (Ferson, 2025). When this happens, the difficulty is not computation, but assessing whether any conditions are missing to forward a valid reasoning process. This situation is common in complex financial practice and professional training, where

problems are rarely well defined from beginning. They appear solvable but are logically underdetermined (Diebold and Yilmaz, 2015). In such cases, committing to a specific assumption is unjustified, since multiple interpretations lead to different conclusions. A rational professional response is to withhold commitment until the missing conditions are clarified (Ferson, 2025).

Despite its practical importance, examination of such capabilities is largely absent in current financial evaluations. Most financial benchmarks adopt an implicit closed-world assumption that every question is fully specified and admits a single correct answer (Jiang et al., 2025). Consequently, benchmark design focuses on whether a model can select the correct option given complete premises, while excluding cases where missing conditions invalidate the act of answering, or no correct answers among the given options. Systems such as FinEval, FinBen, FinMaster, and the Open FinLLM Leaderboard follow this paradigm, limiting their ability to assess whether models can identify under-specified questions or confidently answer unsolvable ones (Islam et al., 2023; Wu et al., 2025). While abstention and unanswerable question detection have been studied in general domains (Wen et al., 2025; Kirichenko et al., 2025), their intersection with professional financial reasoning remains largely unexplored.

To address this gap, we introduce REALFIN, a bilingual financial reasoning benchmark that explicitly evaluates models with incomplete problems. Starting from professional, well-posed exam-style questions, we construct paired counterparts by removing logically necessary premises while preserving coherence and realism (Fish et al., 2024). The resulting condition-missing questions remain linguistically plausible but underdetermined, so any concrete solution requires clarifying unstated assumptions.

Overall, we design three task formulations: (i)

*These authors contributed equally to this work.




Original	Revised
<p>A 5-year EUR government bond is expected to experience a 25 bp yield decline over the next 6 months. Which statement about the rolldown return is correct?</p> <p>A. The rolldown return equals the difference between the price of the 5-year bond and that of a 4.5-year bond at the lower yield-to-maturity.</p> <p>B. The rolldown return consists of the 5-year bond's basis point value multiplied by the expected 25 bp yield-to-maturity change over the next six months.</p> <p>C. The rolldown return will be negative if the 5-year bond has a zero coupon and is trading at a premium.</p>	<p>A 5-year EUR government bond is expected to experience a yield decline over the next 6 months. The bond is a zero-coupon bond currently trading at a premium.</p> <p>Over the holding period, the bond's total return is expected to be close to zero. The most likely magnitude of the yields-to-maturity to decline by ? bps</p> <p>A. 10 B. 20 C. 25 D. 30</p>
<p> A Option A correctly describes this calculation. Option B incorrectly uses BPV, which relates to parallel shifts. Option C is incorrect as premium/discount status doesn't determine rolldown return direction.</p> <p>Conf: 95</p>	<p> C Conf: 85 Smaller yield declines (10 or 20 bps) are not sufficient, while 30 bps would be more than enough.</p> <p style="text-align: center;">NOTA</p> <p style="background-color: #ffe0e0;">A 5-year ... If the rolldown... to decline by ? bps</p> <p>A. 10 B. 20 C. 30 D. None of the Above</p> <p> B Conf: 80 I can't answer this question. 20 seems the right answer.</p>

Figure 1: Original question and its revision with GPT-5.1-mini predictions and confidence, **bold** correct option.

Original: full-condition questions with a unique justified answer; (ii) *Revised*: condition-missing questions require further clarification; and (iii) *None-of-the-above*: replacing the correct answer in multi-choice question answering with none-of-the-above (NOTA), which forces a model to genuinely infer correct answer, rather than pattern recognition. Our contributions are as follows:

- We propose to evaluate models' financial reasoning with under-specified questions, assessing whether a model will over-commit and confidently answer an unsolvable question.
- We manually curate a bilingual financial reasoning dataset in English and Chinese, with in total of 2,020 questions spanning full-condition and condition-missing variants.
- We evaluated 15 LLMs, including five commercial, five finance-specialized, and five reasoning-enhanced LLMs, revealing systematic performance degradation when conditions are missing or no options are correct for open-source models.

2 Related Work

Financial Benchmarks Prior evaluations in the financial domain focused on information extraction and sentiment analysis (Araci, 2019), as well as short-horizon prediction (Yang et al., 2023a), where evidence is typically assumed to be local and fully specified. More recent work extends financial QA toward numerical and long-form reasoning over reports, as in FinQA (Chen et al., 2021) and TAT-QA (Zhu et al., 2021), and toward multi-task settings that integrate retrieval and multi-step reasoning, including PIXIU (Xie et al., 2023), FinBen (Xie et al., 2024), EMS (Hu et al., 2025) and FinChain (Xie et al., 2025). Beyond

QA, simulation-based evaluations further explore strategic behavior in trading (Fish et al., 2024), auctions (Henning et al., 2025), and pricing games (Biancotti et al., 2025). Despite advances, these benchmarks assume that financial problems are fully specified and admit a unique correct answer, offering limited insight into *how models behave when critical conditions are missing*.

What still Challenges LLMs in Finance? Recent studies show that current language models still struggle with many core financial tasks. For problems that require precise numerical prediction or strict rule application (e.g., credit rating, risk assessment), traditional statistical and rule-based methods often outperform LLMs (Drinkall et al., 2025). Many models remain difficult to interpret and control in ways that satisfy regulatory expectations (Tatsat and Shater, 2025). Fish et al. (2025) shows that model performance drops when financial problems become uncertain or change over time. LLMs are also sensitive to how questions are phrased. They change their behaviors when prompts are slightly modified, especially in strategic or incentive-driven settings (Fish et al., 2024). Moreover, requests combining text with tables, charts, or time-series data remain challenging for most models (Hao et al., 2025).

Research that treats LLMs as decision-making agents further finds that models can behave inconsistently or make choices that do not align with basic economic reasoning when incentives are involved (Zhou et al., 2025b; Ouyang et al., 2025). These hard cases guide the scope of our data collection: selecting challenging scenarios and samples.

Specialised Financial Language Models The rise of financial AI has led to a proliferation of specialist models. Early pioneering work, such as BloombergGPT (Wu et al., 2023) and Pixiu (Xie

et al., 2023), established the effectiveness of large-scale domain-specific pre-training and multi-task instruction tailoring. Several open-source models, including FinGPT (Yang et al., 2023a), DISC-FinLLM (Chen et al., 2023), and CFGPT (Li et al., 2023), as well as decision-orientated systems like InvestLM (Yang et al., 2023b), have since emerged.

Recently, the focus has switched to reasoning-enhanced architectures. New models, such as Fin-o1 (Qian et al., 2025), Fin-R1 (Liu et al., 2025), and Dianjin-R1 (Zhu et al., 2025), use large-scale RL and Chain-of-Thought to solve difficult financial logic. While these models represent the cutting edge of financial issue solver, they are still generally assessed on “closed-system” standards, which presume that a valid answer exists at all times. Leaderboards, such as the Open FinLLM Leaderboard (Lin et al., 2025), test performance on fully stated tasks, but rarely evaluate on condition-missing problems.

Unanswerable Questions and Abstention

Knowing *when not to answer* has been studied along two branches: (i) model safety rejects harmful queries (Wang et al., 2024), and (ii) model clarification identifies under-specified queries (Wen et al., 2025) or questions that cannot be answered with the given context (Rajpurkar et al., 2018). Our work belongs to the second line.

Empirically, models tend to be systematically overconfident and over-commit by making implicit assumptions (Xiong et al., 2023), a pattern that is especially pronounced on tasks requiring professional knowledge (Wen et al., 2025). Even recent reasoning-enhanced models struggle with questions that are not fully specified (Kirichenko et al., 2025). This failure persists even when critical context is explicitly removed or perturbed (Wen et al., 2024; Madhusudhan et al., 2025).

Theoretically, selective prediction frameworks formalize the coverage-accuracy trade-off that arises from allowing abstention (Geifman and El-Yaniv, 2017). Reiter’s default logic explains the behavior that models silently substitute missing conditions with learned defaults, producing answers that are coherent but unjustified (Reiter, 1980). LLMs must be explicitly trained to resist this tendency (Kapoor et al., 2024). Beyond passive rejection, reliable agents should proactively acquire missing information by asking clarifying questions (Li et al., 2025).

REALFIN targets *logical underdetermination*

in professional financial reasoning, where missing regulatory standards or valuation assumptions make any committed answer unjustified, a form of under-specification that general-domain benchmarks do not capture.

3 Dataset

We introduce a financial reasoning dataset designed to capture both the breadth and difficulty of real-world tasks.

3.1 Taxonomy

Question Types Six question types follow the hierarchical logic of Bloom’s Revised Taxonomy (Anderson and Krathwohl, 2001), where each question is assigned to the category that best reflects its primary *reasoning demand*. Question types are not required to be mutually exclusive. For example, a complex calculation question may also involve conceptual understanding, but its dominant challenge is multi-step numerical integration. This approach is consistent with the CFA Institute’s Learning Outcome Statement command words (CFA Institute, 2024), which similarly stratify tasks by cognitive complexity.

(i) *Financial Term Explanation and Conceptual Understanding* questions require understanding the meaning of financial terminology, e.g., identifying what “rolldown return” or “cost basis” means in a bond or tax context.

(ii) *Simple Calculation* questions typically require first understanding the problem, and then performing basic arithmetic. For example, computing interest income given a coupon rate and face value.

(iii) *Complex Calculation* involves multiple steps computations using more complex formulas. For instance, compute a bond’s total return by jointly considering price change, coupon income, and yield movements over time.

(iv) *Summary and Comprehensive Judgment* involve consolidating information from multiple sources and aspects, and then making a informed conclusion based on the information. For example, decide whether a firm’s financial position has improved based on changes in cash flow, leverage, and profitability.

(v) *Knowledge Transfer and Application* questions require models to transfer understanding from theory to practical situations, e.g., applying depreciation rules learned in accounting theory to a specific asset purchase case.

(vi) *Statistical and Econometric Methods* use statistical or econometric approaches to analyze data. A case like interpreting regression coefficients, hypothesis test results, or volatility estimates in a financial context.

Financial Topics From the perspective of fine-grained financial topics, we classify English questions into nine sub-domains including equity investment, fixed income investment, quantitative analysis, derivatives investment, financial statement analysis, economics, corporate finance, portfolio management, and other investments. Chinese questions follow five institutionally grounded CPA categories: auditing, accounting, tax law, economic law, and wealth management.

3.2 Full-condition Data

Data Source and Collection The full-condition dataset consists of 1,178 financial reasoning questions in English and Chinese. The two raw data sources are released under permissive licenses that allow their use for academic research. English questions are curated from publicly available CFA-style preparatory materials, excluding real examination content. Chinese questions are collected from publicly accessible CPA-style instructional materials, also excluding real exam items.

During questions selection, we prioritize challenging cases identified in prior work (Section 2). Most selected questions typically require a combination of conceptual understanding, multi-step reasoning, information integration. We choose these question types because they expose model behaviors that purely factual questions cannot reveal, such as whether a model truly understands the problem structure, can reason across multiple constraints, or can recognize when the given information is insufficient to support a valid answer.

Afterwards, each question is manually reviewed to ensure quality, and tagged with both a sub-domain label (e.g., *equity investment*) and a question type label indicating what capabilities are evaluated by this question (e.g., *knowledge transfer and application*).

Quality Control To ensure cross-lingual consistency and accuracy, we recruited two volunteer annotators who have education background in finance and AI research experience, with one female PhD and one male undergraduate student. Both are native Chinese speakers with study experience abroad and are fluent in English. Annotators were

first trained and then asked to review and label using the guideline that defines the scope of each category.

3.3 Condition-Missing Data

Each condition-missing question is derived from a full-condition instance by manually removing one or more assumptions that are required to determine a unique answer. As a result, the revised question still appears well-formed, but no longer provides enough information to justify a single correct conclusion. In practice, we removed four types of conditions.

(i) *Initial assumptions*: missing the high-level economic backdrop that the system takes for granted. (macro assumptions: interest-rate regime, inflation expectations, central bank forward guidance, and going-concern assumption).

(ii) *Intermediate linking conditions*: missing the rules/models that turn the given inputs into the required output. (linking method: valuation model such as CAPM/DCF; definition of valuation multiples; hedging logic; depreciation/amortisation method)

(iii) *Constraint-defining information*: missing safety guardrails and system boundaries within financial contracts. (red lines: financial covenants, seniority/recourse attributes, and related-party identifiers)

(iv) *Standard-selection cues*: missing the governing rulebook that determines how to interpret and record the same transaction. (standards: IFRS vs GAAP; regulatory framework; revenue recognition timing; AML/CFT rating rules; taxable temporary differences)

To ensure that the revised questions remain valid and realistic, two annotators first revise the questions by removing some conditions independently. Then, they cross-validate the revised question by assessing (i) whether the questions still make sense within the financial context and (ii) whether they lead to situations where no unique answer can be derived. Discrepancies are resolved through discussions to maintain the integrity of the task design.

Statistics Table 1 summarizes the distribution of questions across six categories for both languages and two task formulations. There are a total of 1,062 English examples including 604 original and 457 revised, and 959 Chinese cases with 574 original and 385 revised. *Summary and Comprehensive Judgment* constitutes the largest category, reflect-

Reasoning Category	English		Chinese		Total
	Ori.	Rev.	Ori.	Rev.	
Financial Concepts	62	43	119	110	334
Simple Calculation	130	79	141	53	403
Complex Calculation	85	64	52	40	241
Summary & Judgment	161	136	127	99	523
Knowledge Transfer	96	88	87	73	344
Statistical / Econometric	70	47	48	10	175
Total	604	457	574	385	2020

Table 1: Statistics of dataset across six question types in two languages and two task formulations: full- vs. missing- conditions (**O**riinal vs. **R**evised).

ing our emphasis on context-dependent reasoning. *Simple Calculation* also accounts for a substantial portion, aiming to evaluate model’s math application in financial domain.

4 Experiments

We evaluated 15 language models: five general-purpose models (*GPT-5.1-mini*, *Gemini-2.5-Flash*, *Claude-Sonnet-3.5*, *DeepSeek-V3*, *Qwen3-Max*), five finance-specific models (*XuanYuan3-70B*, *Fin-R1-7B*, *CFGPT2-7B*, *DISC-FinLLM-13B*, *FinGPT-7B*), and five reasoning-enhanced models (*DeepSeek-R1*, *GPT-5.2-series*, *GPT-OSS-20B*, *GPT-OSS-120B*, *DianJin-R1-32B*). Table 10 in Appendix C.1 describes all models’ features in detail.

4.1 Experimental Setups

All evaluations are performed in a zero-shot setting, without any in-context demonstration examples.

Prompts For both full- and missing-condition questions, all models are instructed to answer using the same prompt: “*You are an expert in financial problem solving. You will be given a single- or multi-choice question. Please return a STRICT JSON object with three keys: reason, answer, and confidence.*” Figures 3 and 4 show the specific prompt for Chinese and English respectively.

None of the Above To assess whether models genuinely reason about missing conditions rather than relying on guessing, we replace the correct answer option with “none of the above”, a contrastive setting performed on the condition-missing subset.

Decoding Configuration All experiments are conducted with a fixed temperature of 0 (`do_sample=False`) to eliminate randomness. For all offline models, we set `max_new_tokens=1024` and use each model’s native chat template. See more details in Appendix C.2 and C.4 to C.8.

Model	Scale	En CFA		Zh CPA	
		O	R	O	R
DeepSeek-R1	–	76.50	72.81	79.27	77.54
GPT-5.2-series	–	83.80	79.23	80.24	75.39
GPT-OSS-20B	20B	69.67	86.65	8.88	67.43
GPT-OSS-120B	120B	78.42	89.10	10.53	84.57
DianJin-R1-32B	32B	72.95	88.28	13.82	77.71

Table 2: Accuracy (%) of reasoning-enhanced models on English CFA and Chinese CPA. O = original; R = revised. For each column, best is in **Best-O / Best-R** and worst is in **Worst-O / Worst-R**; the best is **bolded**.

4.2 Results

Who Wins? Unexpectedly, five general-purpose models overall perform the best, when financial models are the least (Tables 3, 4), with reasoning-enhanced models in the middle (Table 2).

The English CFA is led by Claude-Sonnet-3.5 under the missing-condition (89.62%), followed by GPT-OSS-120B (89.10%), GPT-5.1-mini (89.01%), Qwen3-Max (88.83%), DianJin-R1-32B (88.28%), and Gemini-2.5-Flash (88.19%) at the same level. In the Chinese CPA, Qwen3-Max establishes a significant lead by 92.00%. Gemini-2.5-Flash (84.80%) and GPT-OSS-120B (84.57%) show comparable performance, as GPT-5.1-mini (80.81%), Claude-Sonnet-3.5 (80.46%) and DeepSeek-V3 (81.06%) are similar.

General Models: analyzing by question types in Table 3, GPT-5.1-mini sweeps all original categories in the CFA and DeepSeek-V3 in CPA. Revised tasks favor Claude-Sonnet-3.5 for CFA and Qwen3-Max for CPA.

Across topics in Table 4, general models strikingly share identical high scores, with green/blue cells clustered instead of scattered. All general models gain 81.82% for original and 90.00% for revised in *Economics*, 90.91% (O) and 100.00% (R) in *Portfolio Management*, 79.66% (O) in *FSA*, and 71.43% (O) in *Fixed Income*. Such sameness may imply that (i) general pre-training corpora cover these topics and (ii) a common performance ceiling exists among state-of-the-art models.

What are Easy and What are Hard? Across question types for both O and R in Table 3, complex calculation is the hardest for all general models. Simple calculation and knowledge transfer & application are comparably challenging. Notably, GPT-5.1-mini exhibits a systemic weakness in CPA, particularly in revised complex calculations.

Regarding topic-specific performance in Table 4,

Question Type	General Models										Financial Models									
	GPT-5.1-m		Gem-2.5		CI-3.5		DS-V3		Qwen3		XY-70B		Fin-R1		CFGPT2		DISC		FinGPT	
	O	R	O	R	O	R	O	R	O	R	O	R	O	R	O	R	O	R	O	R
English CFA																				
Conceptual	82.14	90.12	80.23	89.45	79.87	91.23	75.12	87.98	80.45	90.56	77.42	79.07	69.35	76.74	56.45	76.74	8.06	18.60	1.61	6.98
Simple Calc.	78.56	87.23	77.89	86.12	76.34	88.76	72.45	85.34	77.23	87.89	36.67	82.12	50.00	82.12	30.00	68.16	23.33	20.67	6.67	2.79
Complex Calc.	75.23	86.45	74.12	85.78	73.89	87.12	70.67	84.23	74.56	86.34	33.33	100.0	26.67	100.0	40.00	100.0	6.67	25.00	0.00	0.00
Comp. Judg.	81.45	89.78	79.67	88.23	78.23	90.12	74.56	86.89	79.12	89.45	62.38	88.37	57.43	84.88	60.40	80.23	22.77	22.09	1.98	4.65
Knowledge App.	79.87	88.45	78.12	87.12	77.56	89.34	73.23	85.67	78.45	88.78	40.70	75.00	38.37	75.00	31.40	75.00	10.47	12.50	0.00	0.00
Stats Methods	83.45	91.67	81.23	90.12	80.56	92.34	76.89	88.45	81.67	91.23	71.43	93.62	61.43	93.62	48.57	76.60	2.86	14.89	4.29	2.13
All	80.59	89.01	78.99	88.19	77.81	89.62	73.55	86.34	78.42	88.83	58.47	84.74	54.10	83.65	47.54	73.57	13.11	19.89	2.19	3.54
Chinese CPA																				
Conceptual	82.34	69.12	75.23	86.45	71.12	82.34	82.89	81.67	78.12	93.89	10.92	80.00	7.56	90.00	2.52	40.00	4.20	30.00	-	-
Simple Calc.	80.67	67.89	73.45	84.12	69.34	80.12	81.23	79.89	76.45	91.23	9.76	64.15	7.32	71.70	4.88	49.06	4.88	15.09	-	-
Complex Calc.	79.23	66.45	72.12	82.78	68.12	78.89	80.45	78.23	75.23	89.67	0.00	-	0.00	-	0.00	-	0.00	-	-	-
Comp. Judg.	81.45	68.56	74.23	85.12	70.23	81.23	81.89	80.56	77.12	92.45	10.34	71.72	5.75	72.73	2.30	49.49	5.75	20.20	-	-
Knowledge App.	80.12	67.23	73.12	83.45	69.12	79.67	81.12	79.12	76.23	90.89	0.00	66.67	14.29	66.67	0.00	0.00	14.29	0.00	-	-
Stats Methods	82.67	69.89	75.67	86.89	71.45	82.89	83.12	82.12	78.56	94.23	16.67	60.00	10.42	70.00	4.17	80.00	2.08	30.00	-	-
All	80.81	68.45	73.42	84.80	69.37	80.46	81.06	80.00	76.32	92.00	11.18	69.14	7.57	73.14	2.96	49.71	4.61	19.43	-	-

Table 3: **Accuracy (%) across six question types** for English CFA (top) and Chinese CPA (bottom). *O* = original; *R* = revised. For each row, best original is in **Best-O**, **Best-R**, **Worst-O**, **Worst-R**; best per row is **bolded**.

Topic	General Models										Financial Models									
	GPT-5.1-m		Gem-2.5		CI-3.5		DS-V3		Qwen3		XY-70B		Fin-R1		CFGPT2		DISC		FinGPT	
	O	R	O	R	O	R	O	R	O	R	O	R	O	R	O	R	O	R	O	R
English CFA																				
Corp. Finance	82.50	90.00	91.25	100.0	81.88	90.00	82.19	90.00	83.75	91.00	60.00	90.00	50.00	90.00	60.00	90.00	30.00	40.00	0.00	0.00
Derivatives	80.00	88.00	87.20	96.00	83.68	92.00	80.08	88.00	82.80	90.00	53.33	92.00	53.33	88.00	31.11	76.00	20.00	28.00	6.67	4.00
Economics	81.82	90.00	81.82	90.00	81.82	90.00	81.82	90.00	81.82	89.09	53.33	80.00	53.33	60.00	53.33	60.00	13.33	0.00	0.00	0.00
Equity Inv.	79.49	88.46	79.49	88.46	79.49	88.46	76.28	84.62	78.85	87.18	57.63	86.54	62.71	86.54	59.32	78.85	20.34	13.46	1.69	3.85
FSA	79.66	88.89	79.66	88.89	79.66	88.89	79.66	88.89	79.66	89.63	65.12	88.89	39.53	81.48	53.49	66.67	11.63	22.22	2.33	3.70
Fixed Income	71.43	78.57	71.43	78.57	71.43	78.57	71.43	78.57	71.43	76.79	42.86	78.57	32.65	85.71	36.73	57.14	14.29	14.29	0.00	0.00
Other Inv.	79.75	88.89	77.78	86.42	79.75	88.89	76.54	85.19	78.02	86.42	62.77	86.42	57.45	87.65	50.00	77.78	10.64	23.46	3.19	3.70
Portfolio Mgt	90.91	100.0	90.91	100.0	90.91	100.0	90.91	100.0	90.91	100.0	75.00	100.0	75.00	100.0	75.00	100.0	0.00	33.33	0.00	33.33
Quant Analysis	80.00	88.89	80.00	88.89	82.09	91.11	78.21	86.67	78.61	87.78	65.96	86.67	72.34	88.89	42.55	73.33	0.00	6.67	0.00	2.22
All	80.59	89.01	78.99	88.19	77.81	89.62	73.55	86.34	78.42	88.83	58.47	84.74	54.10	83.65	47.54	73.57	13.11	19.89	2.19	3.54
Chinese CPA																				
Accounting	33.33	36.67	75.76	83.33	72.73	80.00	66.67	73.33	69.32	85.00	4.49	63.33	6.74	66.67	2.25	60.00	4.49	26.67	-	-
Auditing	84.85	93.33	90.91	100.0	87.88	96.67	84.85	93.33	88.64	97.50	25.76	96.67	13.64	83.33	7.58	60.00	4.55	26.67	-	-
Economic Law	66.28	72.97	78.57	86.49	76.19	83.78	81.08	89.19	73.81	90.63	6.67	78.38	6.67	91.89	1.67	45.95	6.67	18.92	-	-
Tax Law	67.11	73.81	75.76	83.33	69.23	76.19	80.10	88.10	71.43	85.94	15.79	66.67	7.02	71.43	1.75	40.48	5.26	14.29	-	-
Wealth Mgt.	45.45	50.00	58.06	63.89	60.61	66.67	50.51	55.56	54.55	67.50	0.00	44.44	0.00	52.78	0.00	47.22	0.00	13.89	-	-
All	80.81	68.45	73.42	84.80	69.37	80.46	81.06	80.00	76.32	92.00	11.18	69.14	7.57	73.14	2.96	49.71	4.61	19.43	-	-

Table 4: **Accuracy (%) across financial topics** for English CFA (top) and Chinese CPA (bottom). *O* = original; *R* = revised. For each row, best original is in **Best-O**, **Best-R**, **Worst-O**, **Worst-R**; best per row is **bolded**.

Fin-R1-7B achieves the highest scores in *fixed income*, *portfolio management*, and *economic law*. In particular, XuanYuan3-70B, Fin-R1-7B, and CFGPT2-7B achieve a perfect precision of 100.0% in *portfolio management*, showing its easiness. Reasoning models in Table 7 show similar trends.

Financial Models: In the Chinese CPA, 4 times 0% in *knowledge application* and *complex calculation* represent systemic “dead zones” for financial models. Under the wealth management topic, the finance-specific models also stand out by 4 times 0%. These failures indicate that domain-specific supervised fine-tuning (SFT), though enhancing factual recall capabilities, has failed to cultivate the multi-step reasoning abilities that require for

synthesising complex regulatory rules.

However, on the English CFA, XuanYuan3-70B, Fin-R1-7B, and CFGPT2-7B surprisingly attain their peak results in *complex calculation* by 100% and *statistical methods* for revised variants.

Overall, question difficulty is consistent across model architectures, which is determined by reasoning demands and information integration capabilities, rather than specific types or topics. Our correlation measurements also reflect the same finding: whether the questions is challenging is driven by information integration density ($r = 0.78$) rather than surface features like question length ($\rho = 0.43$) or numerical frequency ($\rho = 0.52$). See measurement details in Appendix B.1, B.2, B.3.

Why Do Original and Revised Questions Behave Differently?

The performance gap between original (*O*) and revised (*R*) settings reveals how models respond to missing information. We examine accuracy changes across question types (Table 3) and financial topics (Table 4). While the general models maintain a near-total monopoly on Best-*O* results, a significant shift occurs in the revised setting, where approximately 1/5 of the Best-*R* are captured by specialized financial models.

Among the general models, they tend to speculate rather than detecting missing information. Changing from original to revised, English CFA accuracies improve by 8% to 12%. Specifically, the accuracy of GPT-5.1-mini increased from 80.59% to 89.01% (+8.42%), and DeepSeek-V3's accuracy rose from 73.55% to 86.34% (+12.79%). However, the Chinese performance declines by 12% (the accuracy of GPT-5.1-mini decreased from 80.81% to 68.45%), and the Worst-*R* is 66.45% with GPT-5.1-mini in the Chinese CPA *complex calculation*.

A natural question rises: *why these models can recognize the lack of information in English but not in Chinese?* This difference is likely driven by how questions are expressed in the two languages. In Chinese, many words are highly context-dependent and can take on different meanings depending on surrounding cues. For example, words like 走 (*to walk*) can mean *walk* or *leave* in different contexts. Therefore, when a key condition is removed, the remaining words may still form a seemingly coherent sentence, but their meaning becomes ambiguous. Models are then easily misled by these remaining cues and tend to interpret the question in a plausible but unsupported way. In English, although context also matters, missing information more often triggers a different response. This can sometimes lead to higher accuracy, not because the model correctly recognizes the lack of information, but because the guessed assumption happens to align with the reference answer.

Table 8 presents paired bilingual examples that illustrate how the same model (GPT-5.1-mini) exhibits distinct failure modes in two languages.

The evaluation results on reasoning models further shows that the dead-zone pattern on Chinese CPA original questions is architecture-agnostic. GPT-OSS-120B (10.53%), DianJin-R1-32B (13.82%), and GPT-OSS-20B (8.88%) all exhibit the same systematic failure, implying that the bottleneck is conceptual rather than computational.

By contrast, reasoning-focused financial models

such as Fin-R1-7B show much larger gains under missing conditions (e.g., +65.57% on Chinese CPA), while earlier financial models such as DISC-FinLLM-13B and FinGPT-7B remain stagnant or decline.

Specifically, zooming into English revised settings across question types in Table 3, XuanYuan3-70B, Fin-R1-7B, and CFGPT2-7B achieve 100.0% accuracy in both *complex calculation* and *statistical methods*. These models similarly reach perfect scores in revised *portfolio management* in Table 4, with Fin-R1-7B securing five Best-*R* rankings across diverse topics. The jump from near-zero scores in the original *complex calculations* to perfect performance in the revised setting shows that the issue is not math ability, but unclear concepts.

To further explore the two points above, we introduce an additional evaluation setting: None of the Above, testing whether models can genuinely identify which specific condition is missing.

Why Do LLMs Make Errors? The additional evaluation reveals that many LLM errors stem from an inability to reason about which specific condition is missing.

General-purpose models such as GPT-5.1-mini and DeepSeek-V3 often respond with high confidence even when key information is missing. When models are asked not only to choose an answer but to identify what information is absent, general models often fail to give a concrete explanation. Instead, they fall back on vague assumptions or restate parts of the question. This behavior indicates that the error is not simply a wrong final choice, but defaults to guessing based on learned patterns.

This behavior is confirmed quantitatively by the substantial accuracy drop from the Revised to the None of the Above (NOTA) setting shown in Table 5. If models were reasoning strictly from the given information, performance should remain stable when the correct option is replaced by *None of the Above*. However, drops occur across general models, indicating that their correct answers in the Revised setting often come from guessing by filling in missing information, rather than reasoning from what is actually given.

In other words, models prioritize producing a plausible completion over verifying whether the question contains sufficient information. When a key condition is missing, they substitute it with a common scenario learned from training data. This makes the answer appear coherent, even though it

General Models				Financial Models				Reasoning-enhanced Models			
Model	Chi	Eng	Avg	Model	Chi	Eng	Avg	Model	Chi	Eng	Avg
Gemini 2.5 Flash	48.78	72.38	60.58	Fin-R1-7B	77.10	84.50	81.10	DeepSeek-R1	73.31	66.93	70.12
Claude 3.5 Sonnet	37.36	70.11	53.73	XuanYuan3-70B	61.70	84.70	73.20	GPT-5.2-series	71.16	73.21	72.19
Qwen3-Max	44.85	59.94	52.39	CFGPT2-7B	54.90	75.20	65.05	GPT-OSS-20B	67.43	86.65	77.04
GPT-5.1-mini	30.64	68.23	49.44	DISC-FinLLM-13B	20.00	12.26	16.13	GPT-OSS-120B	84.57	89.10	86.84
DeepSeek V3	36.21	56.87	46.54	FinGPT-7B	–	3.50	3.50	DianJin-R1-32B	77.71	88.28	83.00
Avg.	39.57	65.51	52.54	Avg.	53.43	52.03	47.80	Avg.	74.84	80.83	77.84

Table 5: Cross-comparison of model accuracy (%) on None-of-the-Above questions versus Revised accuracy, across three groups of models. Color coding: Green: $\text{NOTA} > \text{Revised}$; Red: $\text{NOTA} < \text{Revised}$; Yellow: $\text{NOTA} = \text{Revised}$ (model treats NOTA options identically to standard distractors).

is no longer supported by the problem statement.

Financial models such as DISC-FinLLM-13B and FinGPT-7B perform extremely poorly, with accuracy below 20.00%. This outcome is counter-intuitive, as these models are explicitly fine-tuned for financial tasks and possess substantial domain knowledge. A closer inspection suggests that this poor performance is not due to a lack of financial knowledge. Instead, domain-specific terminology in the questions, such as "Processing Agreement" or "Consumption Tax," appears to trigger aggressive knowledge retrieval. Once this retrieval process is activated, the models tend to commit to forced calculations or rule-based procedures, even when the information provided is insufficient to support any option. As a result, these models are more likely to select an answer rather than recognize that no option can be justified, which leads to particularly low accuracy in all settings.

In other words, early financial models such as DISC-FinLLM-13B and FinGPT-7B are trained on financial data, they mainly learn to recognize financial terms, and this helps them reduce errors caused by unfamiliar terminology and better identify the meaning of a question. However, these models are not trained to recognize when the information provided is insufficient, hence, still lack the ability to detect when a question itself cannot be answered.

Reasoning-focused models such as Fin-R1-7B also face a new challenge. Although Fin-R1-7B has learned to reject answers when information is insufficient, it sometimes struggles to decide where to stop.

We observe that Fin-R1-7B frequently identifies the missing condition correctly in its reasoning trace but continues generating for **200–300 additional tokens** exploring hypothetical scenarios not implied by the question. Its reasoning trajectory shows that after the model correctly recognizes the

core principles, it still continues reasoning, such as considering unlikely financial cases that are not implied by the question.

Take-Away: The ability to say *No* is essential for handling missing conditions, but without a clear stopping criterion, it can also trigger unnecessary complexity. This also mirrors a common challenge in human decision-making, where caution and thoroughness sometimes lead to overthinking.

4.3 How Do Models Behave When No Answer is Justified

Based on the revised subset, we replace the correct option with “none of the above” as the final choice, to assess whether models can truly identify missing information when no answer is justified.

Accuracy Decreases for Most Models When Models Are Forced to Choose NOTA More apparent declines happen among general-purpose models. Financial-specific models witness majority of rises in Chinese questions in Table 5.¹

Most Models Still Try to Answer Even When No Answer Is Justified Both finance-specific models and general frontier models struggle when *None of the above* is the correct answer. In Table 5, models such as GPT-5.1-mini and DeepSeek V3 show large accuracy drops, especially on Chinese questions. These errors do not stem from a lack of reasoning ability, but from a tendency to keep selecting an answer even when the question cannot be resolved with the given information.

Empirically, general-purpose models tend to over-commit by selecting an answer even when the question lacks a clear logical basis. This behavior likely comes from training objectives that reward producing a response rather than recognizing when

¹The financial models and the reasoning-enhanced models in Table 5 were evaluated on different GPU hardware, which may account for minor numerical variances in the accuracies.

Model	English CFA (367)			Chinese CPA (175)		
	Ans.	Reas.	Conf.	Ans.	Reas.	Conf.
XuanYuan3-70B	84.7	8.4	92.9	61.7	5.1	87.9
Fin-R1-7B	84.5	42.5	93.8	77.7	41.1	90.3
CFGPT2-7B	75.2	8.5	82.6	54.9	11.4	81.0
DISC-13B	12.3	2.3	24.2	20.0	18.9	65.2
FinGPT-7B	3.5	0.0	50.3	–	–	–
GPT-OSS-20B	86.6	12.3	90.7	67.4	4.0	79.2
GPT-OSS-120B	89.1	20.7	95.7	84.6	1.1	91.5
DianJin-R1-32B	88.3	29.2	93.7	77.7	5.7	77.8

Table 6: Open-source models’ NOTA results: Answer Acc vs. Reasoning Acc gap, along with confidence.

a problem cannot be solved. As a result, these models fail not because they cannot reason, but because they do not know when to stop reasoning.

However, the strong performance of Fin-R1-7B shows that this limitation is not shared by all models. Unlike other models, Fin-R1-7B does not treat the *None of the above* option as a fallback choice. Instead, it consistently checks whether the information provided in the question is sufficient to justify any of the listed options. Its reasoning outputs indicate that the model actively looks for missing inputs before committing to an answer. For example, in tax-related questions, Fin-R1-7B explicitly notes the absence of required quantities, such as a comparable market price, and refrains from selecting an option when such information is missing. As Fin-R1-7B is designed to prioritize condition checking before answer selection, which directly counteracts the tendency to over-answer encouraged by RLHF. This design also prevents the retrieval-heavy calculation mode triggered by financial terminology from turning into forced commitments, so the model can choose NOTA when the question is underspecified.

Why Accuracy Overestimates Reliability in NOTA In NOTA settings, selecting the correct option alone is insufficient to indicate reliable decision-making. When no answer is justified, accuracy can be artificially inflated by models that avoid making a clear choice without understanding why the problem cannot be resolved.

To address this, we examine whether a model’s reasoning supports the same conclusion as its final answer. To distinguish principled abstentions from lucky guesses, we evaluate every NOTA response against two complementary metrics: *Answer Accuracy* (whether NOTA is selected) and *Reasoning Accuracy* (whether the explanation correctly identi-

fies the specific missing premise).

In Table 6, eight LLMs consistently exhibit a significant gap between answer accuracy and reasoning accuracy. The smallest gap occurs in reasoning-focused financial model Fin-R1-7B. It maintains relatively higher consistency between answers and explanations, identifying which required condition is missing before selecting NOTA.

By contrast, some financial models and reasoning models appear to perform reasonably well based on answer accuracy alone, but their explanations fail to support the chosen answer. For example, XuanYuan3-70B, GPT-OSS-120B/20B and DianJin-R1 drop from >85% answer accuracy to 8-30% for reasoning on English NOTA questions. In these cases, the model selects NOTA simply because it is a safe option, not because it understands that the question cannot be answered.

We further observe a degenerate response pattern in some models, where uncertainty is expressed by selecting multiple options at once (*e.g.*, *AB or ABCD*). Although this behavior may seem cautious, it does not represent a valid financial decision because it offers no clear conclusion. Taken together, financial applications require models to reject an answer for clear and traceable reasons, rather than by chance.

5 Conclusions

In this paper, we show that evaluating financial LLMs solely by answer accuracy misses a key aspect of reliability: knowing when not to answer. Models often answer despite missing information, especially in Chinese, where terminology hides missing conditions and triggers more assumptions. Financial models reduce blind guessing but still struggle to identify what is missing, while reasoning-focused models improve condition detection but may over-reason. Scaling alone does not solve this: larger models achieve high answer accuracy but low reasoning accuracy, increasing over-commitment without improving condition checking. We therefore argue that reliable financial LLMs must satisfy three criteria: answer correctly when information is sufficient, identify missing conditions when it is not, and abstain when no valid answer exists. Only models that perform consistently across these settings in both languages are dependable in practice.

Limitations and Future Work

A main limitation of this work is that we do not modify or train any models. Our study can show where failures come from, but we do not directly demonstrate how much fine-tuning can fix them. Because our evaluation uses multiple-choice questions, the same behavior may show up in different ways when models respond free-form questions or interact with users.

These limits point to a clear next step. As our results suggest that models either guess, or they pick a safe option without being able to explain what is missing, or they reason too far. A practical solution is a unified training pipeline that separates what should be learned from what should be controlled. First, specialized SFT should teach professional knowledge and stable reasoning formats, so the model learns what variables and rules matter in common financial tasks. Then logic-driven RL should train the model's behavior when conditions are missing. At this stage, the model should be rewarded for explicitly pointing out what information is missing, and penalized for guessing, making up assumptions, or reasoning too far beyond what the question supports.

We also plan to move beyond passive rejection. When a question lacks a key variable, the model should not only say "This question has no answer", but also state what information it needs and ask for it. This would build a interactive workflow for users. We will also broaden the dataset beyond accounting and tax law to cover risk management, securities regulation, and quantitative finance, so that the same evaluation logic can test whether models stay reliable across the full range of real financial work.

Ethical Statement This work is intended to improve the safe and responsible use of large language models in financial applications. All datasets used in this study are constructed from publicly accessible materials or manually rewritten questions and do not include proprietary data, confidential information, or verbatim content from protected examinations. Our evaluation highlights a specific risk relevant to high-stakes domains: confident model outputs in situations where information is insufficient to justify a determinate answer. By explicitly identifying and measuring this behaviour, our work aims to reduce the likelihood of uncritical deployment of LLMs in financial decision-making, rather than to encourage automation of such decisions.

We emphasise that the proposed benchmark and any derivative artefacts, such as distilled small-parameter models, are intended for research and risk-assessment purposes only. Any deployment of language models in real financial, regulatory, or advisory contexts must prioritize the auditability of reasoning chains, involve appropriate human oversight, and comply with applicable legal and ethical standards. We aim to support the development of "logic-first" assistance tools rather than fully autonomous systems for high-stakes financial judgments.

References

- Lorin W Anderson and David R Krathwohl. 2001. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition*. Addison Wesley Longman, Inc.
- Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.
- Claudia Biancotti, Carolina Camassa, Andrea Coletta, Oliver Giudice, and Aldo Glielmo. 2025. Chat bankman-fried: an exploration of llm alignment in finance. In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFin-Legal)*, pages 1–22.
- CFA Institute. 2024. **CFA and CIPM LOS command words**.
- Wei Chen, Qiushi Wang, Zefei Long, Xianyin Zhang, Zhongtian Lu, Bingxuan Li, Siyuan Wang, Jiarong Xu, Xiang Bai, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-finllm: A chinese financial large language model based on multiple experts fine-tuning. *arXiv preprint arXiv:2310.15205*.
- Zhiyu Chen, Wenhui 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.
- Francis X Diebold and Kamil Yilmaz. 2015. *Financial and macroeconomic connectedness: A network approach to measurement and monitoring*. Oxford University Press.
- Felix Drinkall, Janet Pierrehumbert, and Stefan Zohren. 2025. Forecasting credit ratings: A case study where traditional methods outperform generative llms. In

- Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 118–133.
- Wayne Ferson. 2025. Empirical asset pricing: Models and methods.
- Sara Fish, Yannai A Gonczarowski, and Ran I Shorrer. 2024. Algorithmic collusion by large language models. *arXiv preprint arXiv:2404.00806*, 7(2):5.
- Sara Fish, Julia Shephard, Minkai Li, Ran I Shorrer, and Yannai A Gonczarowski. 2025. Econevals: Benchmarks and litmus tests for llm agents in unknown environments. In *2nd Workshop on Models of Human Feedback for AI Alignment*.
- Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30.
- Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. 2025. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. *arXiv preprint arXiv:2501.05444*.
- Thomas Henning, Siddhartha M Ojha, Ross Spoon, Jiantong Han, and Colin F Camerer. 2025. Llm trading: Analysis of llm agent behavior in experimental asset markets. *arXiv e-prints*, pages arXiv–2502.
- Bo Hu, Han Yuan, Vlad Pandealea, Wuqiong Luo, Yingzhu Zhao, and Zheng Ma. 2025. Extract, match, and score: An evaluation paradigm for long question-context-answer triplets in financial analysis. *arXiv preprint arXiv:2503.16575*.
- Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. Financebench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*.
- Junzhe Jiang, Chang Yang, Aixin Cui, Sihan Jin, Ruiyu Wang, Bo Li, Xiao Huang, Dongning Sun, and Xinrun Wang. 2025. Finmaster: A holistic benchmark for mastering full-pipeline financial workflows with llms. *arXiv preprint arXiv:2505.13533*.
- Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew G Wilson. 2024. Large language models must be taught to know what they don’t know. *Advances in Neural Information Processing Systems*, 37:85932–85972.
- Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J Bell. 2025. Abstentionbench: Reasoning llms fail on unanswerable questions. *arXiv preprint arXiv:2506.09038*.
- Belinda Z Li, Been Kim, and Zi Wang. 2025. Quest-bench: Can llms ask the right question to acquire information in reasoning tasks? *arXiv preprint arXiv:2503.22674*.
- Jiangtong Li, Yuxuan Bian, Guoxuan Wang, Yang Lei, Dawei Cheng, Zhijun Ding, and Changjun Jiang. 2023. Cfgpt: Chinese financial assistant with large language model. *arXiv preprint arXiv:2309.10654*.
- Shengyuan Colin Lin, Felix Tian, Keyi Wang, Xingjian Zhao, Jimin Huang, Qianqian Xie, Luca Borella, Matt White, Christina Dan Wang, Kairong Xiao, Xiao-Yang Liu Yanglet, and Li Deng. 2025. Open finllm leaderboard: Towards financial ai readiness. *arXiv preprint arXiv:2501.10963*.
- Zhaowei Liu, Xin Guo, Fangqi Lou, Lingfeng Zeng, Jinyi Niu, Zixuan Wang, Jiajie Xu, Weige Cai, Ziwei Yang, Xueqian Zhao, Chao Li, Sheng Xu, Dezhi Chen, Yun Chen, Zuo Bai, and Liwen Zhang. 2025. Fin-r1: A large language model for financial reasoning through reinforcement learning. *arXiv preprint arXiv:2503.16252*.
- Nishanth Madhusudhan, Sathwik Tejaswi Madhusudhan, Vikas Yadav, and Masoud Hashemi. 2025. Do llms know when to not answer? investigating abstention abilities of large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9329–9345.
- Shumiao Ouyang, Hayong Yun, and Xingjian Zheng. 2025. Ai as decision-maker: ethics and risk preferences of llms. *Preprint at http://arxiv.org/abs/2406.01168*.
- Lingfei Qian, Weipeng Zhou, Yan Wang, Xueqing Peng, Jimin Huang, and Qianqian Xie. 2025. Fino1: On the transferability of reasoning enhanced llms to finance. *arXiv e-prints*, pages arXiv–2502.
- Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don’t know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.
- Raymond Reiter. 1980. A logic for default reasoning. *Artificial intelligence*, 13(1-2):81–132.
- Hariom Tatsat and Ariye Shater. 2025. Beyond the black box: Interpretability of llms in finance. *arXiv preprint arXiv:2505.24650*.
- Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. 2024. Do-not-answer: Evaluating safeguards in llms. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 896–911.
- Bingbing Wen, Bill Howe, and Lucy Lu Wang. 2024. Characterizing llm abstention behavior in science qa with context perturbations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3437–3450.

- Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. 2025. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 13:529–556.
- Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabrovolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kam-badur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.
- Yingqian Wu, Qiushi Wang, Zefei Long, Rong Ye, Zhongtian Lu, Xianyin Zhang, Bingxuan Li, Wei Chen, Liwen Zhang, and Zhongyu Wei. 2025. Fin-team: A multi-agent collaborative intelligence system for comprehensive financial scenarios. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 443–455. Springer.
- Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, Yijing Xu, Haoqiang Kang, Ziyang Kuang, Chenhan Yuan, Kailai Yang, Zheheng Luo, Tianlin Zhang, Zhiwei Liu, Guojun Xiong, Zhiyang Deng, Yuechen Jiang, Zhiyuan Yao, Haohang Li, Yangyang Yu, Gang Hu, Jiajia Huang, Xiao-Yang Liu, Alejandro Lopez-Lira, Benyou Wang, Yanzhao Lai, Hao Wang, Min Peng, Sophia Ananiadou, and Jimin Huang. 2024. Finben: A holistic financial benchmark for large language models. *Advances in Neural Information Processing Systems*, 37:95716–95743.
- Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: A comprehensive benchmark, instruction dataset and large language model for finance. *Advances in Neural Information Processing Systems*, 36:33469–33484.
- Zhuohan Xie, Daniil Orel, Rushil Thareja, Dhruv Sahnan, Hachem Madmoun, Fan Zhang, Debopriyo Banerjee, Georgi Georgiev, Xueqing Peng, Lingfei Qian, Jimin Huang, Jinyan Su, Aaryamonvikram Singh, Rui Xing, Rania Elbadry, Chen Xu, Haonan Li, Fajri Koto, Ivan Koychev, Tanmoy Chakraborty, Yuxia Wang, Salem Lahlou, Veselin Stoyanov, Sophia Ananiadou, and Preslav Nakov. 2025. FINCHAIN: A symbolic benchmark for verifiable chain-of-thought financial reasoning. *arXiv preprint arXiv:2506.02515*.
- Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*.
- Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023a. Fingpt: Open-source financial large language models. *arXiv preprint arXiv:2306.06031*.
- Yi Yang, Yixuan Tang, and Kar Yan Tam. 2023b. Investlm: A large language model for investment using financial domain instruction tuning. *arXiv preprint arXiv:2309.13064*.
- Hanyu Zhang, Boyu Qiu, Yuhao Feng, Shuqi Li, Qian Ma, Xiyuan Zhang, Qiang Ju, Dong Yan, and Jian Xie. 2024a. Baichuan4-finance technical report. *arXiv preprint arXiv:2412.15270*.
- Wentao Zhang, Lingxuan Zhao, Haochong Xia, Shuo Sun, Jiase Sun, Molei Qin, Xinyi Li, Yuqing Zhao, Yilei Zhao, Xinyu Cai, Longtao Zheng, Xinrun Wang, and Bo An. 2024b. A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist. In *Proceedings of the 30th acm sigkdd conference on knowledge discovery and data mining*, pages 4314–4325.
- Xuanyu Zhang and Qing Yang. 2023. Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, pages 4435–4439.
- Yuanchen Zhou, Shuo Jiang, Jie Zhu, Junhui Li, Lifan Guo, Feng Chen, and Chi Zhang. 2025a. Fin-prm: A domain-specialized process reward model for financial reasoning in large language models. *arXiv preprint arXiv:2508.15202*.
- Yuhang Zhou, Yuchen Ni, Zhiheng Xi, Zhangyue Yin, Yu He, Gan Yunhui, Xiang Liu, Zhang Jian, Sen Liu, Xipeng Qiu, Yixin Cao, Guangnan Ye, and Hongfeng Chai. 2025b. Are llms rational investors? a study on the financial bias in llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24139–24173.
- 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.
- Jie Zhu, Qian Chen, Huaixia Dou, Junhui Li, Lifan Guo, Feng Chen, and Chi Zhang. 2025. Dianjin-r1: Evaluating and enhancing financial reasoning in large language models. *arXiv preprint arXiv:2504.15716*.

Category	Reasoning-enhanced Models									
	DeepSeek-R1		GPT-5.2		GPT-OSS-20B		GPT-OSS-120B		DianJin-32B	
	O	R	O	R	O	R	O	R	O	R
English CFA										
<i>By Question Type</i>										
Conceptual	79.13	71.34	81.45	74.42	83.87	83.72	85.48	86.05	80.65	86.05
Simple Calc.	71.67	68.53	75.71	71.26	63.33	84.92	93.33	87.71	66.67	87.15
Complex Calc.	73.49	76.56	78.13	81.25	60.00	100.00	86.67	100.00	60.00	100.00
Comp. Judg.	76.65	70.47	80.12	74.83	68.32	88.37	78.22	89.53	84.16	89.53
Knowledge App.	72.09	67.38	76.04	71.59	48.84	87.50	56.98	100.00	44.19	87.50
Stats Methods	80.00	78.43	84.29	82.13	88.57	91.49	90.00	93.62	90.00	91.49
<i>By Topic</i>										
Corp. Finance	77.50	71.76	81.25	75.34	80.00	90.00	80.00	90.00	80.00	90.00
Derivatives	68.89	63.27	73.63	67.35	60.00	88.00	71.11	88.00	62.22	96.00
Economics	72.22	67.69	76.87	70.38	66.67	90.00	66.67	90.00	60.00	90.00
Equity Inv.	77.97	72.43	81.33	75.96	74.58	88.46	83.05	88.46	76.27	88.46
FSA	76.74	71.48	80.24	74.81	74.42	85.19	74.42	88.89	76.74	88.89
Fixed Income	63.27	58.16	67.35	61.22	53.06	78.57	67.35	78.57	46.94	78.57
Other Inv.	75.53	70.37	78.59	73.46	68.09	88.89	81.91	90.12	79.79	88.89
Portfolio Mgt	90.91	88.24	95.45	92.71	100.00	100.00	100.00	100.00	100.00	100.00
Quant Analysis	78.72	73.47	82.98	76.83	85.11	88.89	89.36	91.11	89.36	88.89
All	76.50	72.81	83.80	79.23	69.67	86.65	78.42	89.10	72.95	88.28
Chinese CPA										
<i>By Question Type</i>										
Conceptual	80.34	76.82	82.18	78.64	10.08	90.00	8.40	100.00	15.13	90.00
Simple Calc.	78.57	74.29	80.14	75.96	2.44	49.06	4.88	69.81	4.88	62.26
Complex Calc.	76.92	72.73	78.85	75.19	0.00	–	0.00	–	0.00	–
Comp. Judg.	78.16	74.58	80.23	76.34	4.60	77.78	9.20	91.92	13.79	84.85
Knowledge App.	76.34	71.21	78.57	73.28	14.29	33.33	28.57	66.67	0.00	100.00
Stats Methods	79.13	75.45	81.25	77.91	18.75	50.00	20.83	80.00	20.83	70.00
<i>By Topic</i>										
Accounting	78.41	73.67	80.23	75.42	6.74	36.67	5.62	66.67	7.87	60.00
Auditing	80.76	91.52	83.94	93.18	16.67	93.33	25.76	93.33	36.36	93.33
Economic Law	79.39	81.43	81.67	83.47	3.33	70.27	6.67	91.89	8.33	89.19
Tax Law	79.82	75.64	82.46	77.14	7.02	83.33	7.02	92.86	8.77	88.10
Wealth Mgt.	77.65	73.16	78.29	74.38	12.50	50.00	6.25	75.00	3.12	55.56
All	79.27	77.54	80.24	75.39	8.88	67.43	10.53	84.57	13.82	77.71

Table 7: Accuracy (%) of reasoning-enhanced models across question types and financial topics for English CFA (top) and Chinese CPA (bottom). O = original; R = revised. For each row, best original is in **Best-O**, best revised is in **Best-R**, worst original is in **Worst-O**, and worst revised is in **Worst-R**; best per row is **bolded**. Ties are all highlighted.

	English Version	Chinese Version
Question	When evaluating audit evidence, if a CPA discovers a material misstatement caused by fraud, which of the following measures should be taken? A. Issue a modified audit report. B. Re-evaluate the risk of material misstatement due to fraud. C. Reconsider the reliability of previously obtained audit evidence. D. Expand the scope of audit procedures.	在评价审计证据时，注册会计师发现由于舞弊导致的重大错报，应当采取的措施包括（）。A. 出具非无保留意见的审计报告B. 重新评估舞弊导致的重大错报风险C. 重新考虑此前获取的审计证据的可靠性D. 扩大审计程序的范围
Correct answer	BC	BC
Missing condition	The applicable auditing standard (e.g., ISA 240 vs. PCAOB AS 2401) is unspecified, leaving the scope of required responses underdetermined.	未指明适用的审计准则体系（如中国注册会计师审计准则第1141号），导致所要求的应对措施范围无法唯一确定。
Model output (GPT-5.1-mini)	<i>Reason:</i> “When fraud is detected, the auditor must re-evaluate the fraud risk assessment and reconsider the reliability of prior evidence. Issuing a modified report (A) is premature at this stage, and expanding audit scope (D) is not automatically required.” <i>Answer:</i> BC <i>Conf:</i> 85	<i>Reason:</i> 注册会计师发现舞弊导致的重大错报时，应当重新评估相关风险、重新考虑审计证据可靠性、扩大审计程序范围，并在必要时出具非无保留意见的审计报告。 <i>Answer:</i> ABCD <i>Conf:</i> 91
Failure mode	Assumption filling: model silently assumes ISA 240 scope and excludes A and D as conditionally required, arriving at BC by coincidence rather than by recognizing the missing standard.	Aggressive retrieval: terminology (“舞弊”、“重大错报”) triggers exhaustive rule recall, causing the model to include all options regardless of conditionality, resulting in ABCD.

Table 8: Paired bilingual example illustrating cross-lingual failure modes of GPT-5.1-mini on the same audit question. English : model arrives at correct answer BC by silently assuming a default auditing standard, not by detecting the missing condition. Chinese : professional terminology triggers exhaustive knowledge retrieval, producing the overcomplete answer ABCD.

A Dataset Curation

A.1 Formal Definition of Condition-Missing Construction

We formalize the construction of condition-missing questions to clarify how missing information is introduced in a controlled manner.

Full-condition formulation. Each full-condition financial reasoning question can be abstracted as a logical implication

$$(P_1 \wedge P_2 \wedge \dots \wedge P_n) \Rightarrow C, \quad (1)$$

where $\{P_i\}$ denotes the set of premises explicitly stated in the question, such as numerical assumptions, regulatory standards, time horizons, or accounting rules, and C denotes the unique justified conclusion. Under full conditions, the premise set is sufficient to determine C without ambiguity.

Condition removal. To construct a condition-missing question, we remove one or more logically necessary premises from the original set. Let $\mathcal{P}^* \subset \{P_1, \dots, P_n\}$ denote the subset of premises identified as necessary for determining the conclusion. The revised question is defined by the reduced premise set

$$\{P_i\} \setminus \mathcal{P}^*. \quad (2)$$

Coherence constraint. After removal, the remaining premises must still admit at least one coherent interpretation. Formally, we require

$$\exists \mathcal{M} \text{ s.t. } \mathcal{M} \models \bigwedge_{P_i \in \{P\} \setminus \mathcal{P}^*} P_i, \quad (3)$$

ensuring that the revised question remains grammatically and semantically well-formed, rather than logically inconsistent.

Non-uniqueness constraint. At the same time, the remaining premises must no longer uniquely determine the conclusion. Specifically, we require the existence of multiple internally consistent interpretations

$$\mathcal{M}_j \models \bigwedge_{P_i \in \{P\} \setminus \mathcal{P}^*} P_i \text{ and } \mathcal{M}_j \models C_j, \quad (4)$$

with

$$C_j \neq C_{j'}, \quad (5)$$

for some $j \neq j'$. This condition guarantees that no single answer can be justified given the available information.

Functional tagging. Each removed premise $P_k \in \mathcal{P}^*$ is annotated by its functional role

$$\tau(P_k) \in \{\text{numerical, regulatory, temporal, standard}\}, \quad (6)$$

to ensure that only logically necessary information is removed, rather than incidental details.

Implications for accuracy and model behavior.

In original questions, accuracy directly reflects whether the model successfully derives the unique justified conclusion C . In revised questions, however, accuracy depends on whether the model’s implicit assumptions happen to align with the reference answer retained from the original task. Formally, a correct prediction on a revised question satisfies

$$\hat{C} = C_{\text{ref}}, \quad (7)$$

even though

$$\{P_i\} \setminus \mathcal{P}^* \not\models C_{\text{ref}}. \quad (8)$$

This explains why some models improve after condition removal while others degrade: performance changes reflect differences in fallback assumptions and inductive biases, rather than differences in reasoning ability. Consequently, accuracy ceases to be a reliable indicator of reasoning quality once questions are no longer well-posed. Together, these constraints ensure that condition-missing questions differ from full-condition questions not in computational difficulty, but in epistemic status: the task shifts from computing a correct answer to recognizing that no uniquely justified answer can be derived.

A.2 Formalization of Abstention Under Missing Correct Answers

This appendix provides a formal explanation of the experiment in Section 4, where the correct option in condition-missing questions is replaced by “None of the above” (English) or “以上都不是” (Chinese).

Problem setup. A full-condition financial reasoning question can be abstracted as a set of premises and a target conclusion,

$$(P_1 \wedge P_2 \wedge \dots \wedge P_n) \models C, \quad (9)$$

where the premises jointly support a unique, justified conclusion. In the condition-missing setting,

one or more logically necessary premises are removed. The resulting task is characterized by

$$(P_1 \wedge \dots \wedge P_{k-1} \wedge P_{k+1} \wedge \dots \wedge P_n) \not\models C, \quad (10)$$

meaning that the remaining information is insufficient to justify any single concrete answer.

Implicit assumption completion. Despite this underdetermination, models often still select a concrete option. This behavior can be understood as implicitly completing the missing information with a default assumption. Formally, the model selects an internal completion \mathcal{M}_j such that

$$\mathcal{M}_j \models \bigwedge_{i \neq k} P_i \quad \text{and} \quad \mathcal{M}_j \models C_j, \quad (11)$$

where C_j corresponds to one of the provided answer options. If C_j happens to match the reference answer retained from the original full-condition question, the prediction is counted as correct, even though the conclusion is not logically supported by the observed premises alone.

Abstention as the rational strategy. When the correct option is explicitly set to “None of the above”, this shortcut is no longer available. In this setting, a rational response requires recognizing that none of the concrete options is justified. Formally, the correct decision rule is

$$\forall j, \{P_i\} \not\models C_j \Rightarrow \text{select NOTA}. \quad (12)$$

This formulation makes explicit that the task is no longer to identify the best answer, but to assess whether any answer can be supported at all. Empirically, however, models frequently violate this condition by selecting a specific option even when no option is entailed.

Interpretation. The sharp accuracy drop observed in the *None-of-the-above* setting therefore indicates that many correct answers in the revised setting were achieved through implicit assumption alignment rather than principled reasoning. This experiment isolates a structural limitation of current LLMs: they lack a robust mechanism for recognizing when available information is insufficient and reasoning should be suspended rather than completed.

B Statistical Analysis Methodology

This appendix details the computational procedures and formulas used for all statistical analyses reported in Section 4.

B.1 Difficulty Ranking Consistency (Spearman’s ρ)

Purpose. To measure whether different models exhibit consistent difficulty rankings across question classes.

Data Preparation. For each model and each question class, we compute the average accuracy as the percentage of correctly answered questions. For model m on question class c :

$$\text{Accuracy}(m, c) = \frac{\text{Number of correct answers}}{\text{Total questions in class } c} \times 100\% \quad (13)$$

Ranking Assignment. For each model m , we rank the question classes from hardest (rank 1) to easiest (rank C) based on accuracy:

$$R_m(c) = \text{rank of class } c \text{ for model } m \quad (14)$$

where lower accuracy yields lower (harder) rank.

Pairwise Spearman Correlation. For each pair of models (m_1, m_2) , we compute Spearman’s rank correlation coefficient:

$$\rho_{m_1, m_2} = 1 - \frac{6 \sum_{c=1}^C d_c^2}{C(C^2 - 1)} \quad (15)$$

where $d_c = R_{m_1}(c) - R_{m_2}(c)$ is the difference in ranks for class c .

Average Correlation. The reported average Spearman’s $\rho = 0.91$ is computed as:

$$\bar{\rho} = \frac{2}{M(M-1)} \sum_{m_1=1}^{M-1} \sum_{m_2=m_1+1}^M \rho_{m_1, m_2} \quad (16)$$

where $M = 10$ models.

B.2 Surface Feature Correlations

Question Length Analysis. For each question class c , we compute the average token count:

$$L_c = \frac{1}{N_c} \sum_{i=1}^{N_c} \text{tokens}(q_i) \quad (17)$$

where $\text{tokens}(q_i)$ is the number of tokens in question i .

Dataset	Type	Content
Original (Chinese)	Topic	Auditing
	Q	下列关于管理层编制财务报告的说法中，正确的有（）。
	Opts	A. 根据财务报告编制基础编制，不能运用自身判断 B. 根据相关法律法规的规定确定适用的财务报告编制基础 C. 根据适用的财务报告编制基础编制财务报表 D. 在财务报表中对使用的财务报告编制基础作出恰当的说明
	Gold	BCD
	Rsn	管理层在编制财务报表时应严格按照财务报告编制基础，但是也存在需要根据使用的财务报告编制基础运用判断做出合理会计估计的情况，不是完全不运用自身判断。
Missing (Chinese)	Topic	Accounting
	Q	若要计算委托加工业务中受托方应代收代缴的消费税额，应补充下列哪项关键信息（）。
	Opts	A. 该应税物资在受托方是否存在同类消费品的销售价格 B. 该批原材料是否属于自产产品 C. 加工费是否以现金方式结算 D. 委托方是否具有一般纳税人资格
	Gold	A
	Rsn	需要知道是否有同类消费品销售价格来确定计税依据。
Original (English)	Topic	Financial Statement Analysis
	Q	Which of the following is most likely to have been included in Sea Ltd's SEC registration statement?
	Opts	A. Underwriters' fairness opinion B. Assessment of risk factors C. Projected cash flows and earnings
	Gold	B
	Rsn	SEC registration statements must include risk factor assessments.
Missing (English)	Topic	Derivatives Investment
	Q	For partnerships to achieve pass-through taxation, which structural characteristic must be present?
	Opts	A. Entity level taxation B. Individual partner level only C. Both entity and individual D. No taxation
	Gold	B
	Rsn	Pass-through taxation means income is taxed only at individual partner level.

Table 9: Question examples from Original and Missing datasets. Gold answers and reasons are highlighted in yellow and displayed in **bold italic**. “Q” = Question, “Opts” = Options, “Gold” = Gold Answer, “Rsn” = Reason.

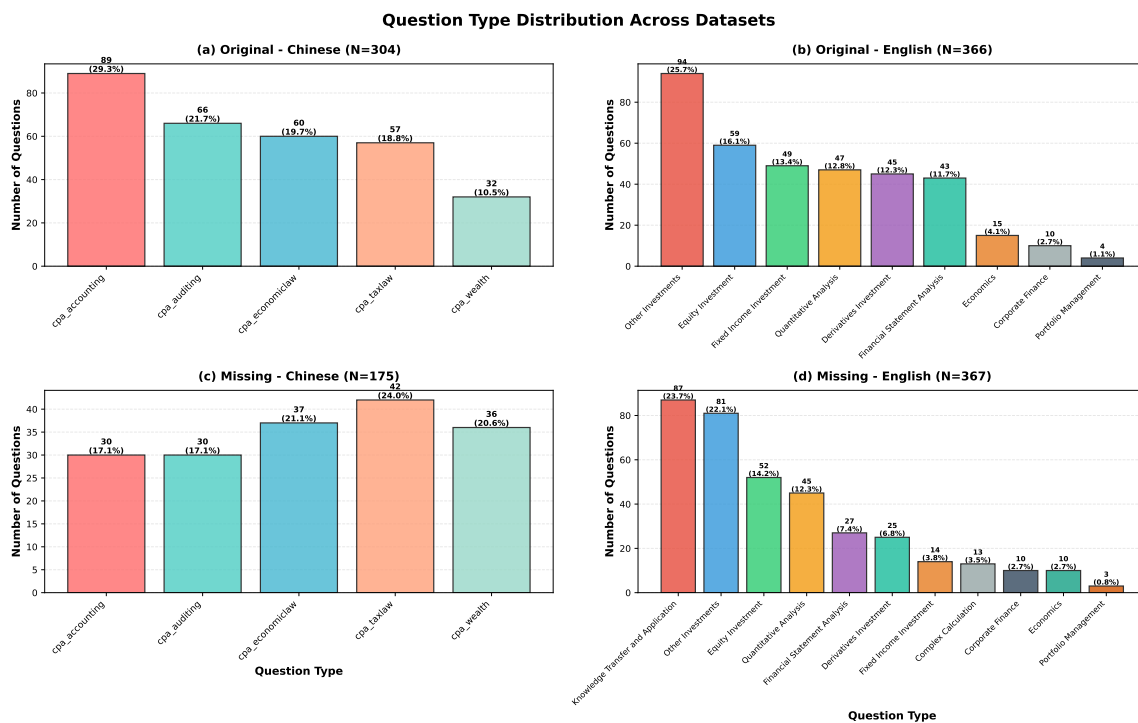


Figure 2: Question Type Distribution Across Datasets

Independent Samples t -test. To test whether token length differs between the hardest (Knowledge Transfer, $L_{KT} = 127$) and easiest (Term Explanation, $L_{TE} = 134$) categories:

$$t = \frac{\bar{L}_{KT} - \bar{L}_{TE}}{\sqrt{\frac{s_{KT}^2}{n_{KT}} + \frac{s_{TE}^2}{n_{TE}}}} \quad (18)$$

where s_c^2 is the sample variance of token counts in class c , and n_c is the sample size.

Numerical Density Correlation. We compute numerical density as:

$$D_c = \frac{1}{N_c} \sum_{i=1}^{N_c} \frac{\text{count of numbers in } q_i}{\text{tokens}(q_i)} \times 100 \quad (19)$$

Pearson correlation between D_c and error rate $E_c = 100 - \text{Acc}_{\text{avg}}(c)$:

$$r_D = \frac{\sum_{c=1}^C (D_c - \bar{D})(E_c - \bar{E})}{\sqrt{\sum_{c=1}^C (D_c - \bar{D})^2} \sqrt{\sum_{c=1}^C (E_c - \bar{E})^2}} \quad (20)$$

B.3 Information Integration Metrics

Feature Annotation. Each question was manually annotated by three expert annotators (Fleiss’ $\kappa = 0.81$, substantial agreement) for:

- **Interacting Premises (P_i):** Number of distinct information pieces that must be combined.
- **Cross-domain Dependencies (C_i):** Rated 1–5, whether reasoning spans multiple financial subdomains.
- **Constraint Reconciliation (R_i):** Rated 1–5, complexity of reconciling conflicting requirements.

Aggregation. For each question class c :

$$\bar{P}_c = \frac{1}{N_c} \sum_{i=1}^{N_c} P_i \quad (21)$$

$$\bar{C}_c = \frac{1}{N_c} \sum_{i=1}^{N_c} C_i \quad (22)$$

$$\bar{R}_c = \frac{1}{N_c} \sum_{i=1}^{N_c} R_i \quad (23)$$

Pearson Correlation with Error Rate. For feature $X \in \{P, C, R\}$:

$$r_X = \frac{\sum_{c=1}^C (\bar{X}_c - \bar{X})(E_c - \bar{E})}{\sqrt{\sum_{c=1}^C (\bar{X}_c - \bar{X})^2} \sqrt{\sum_{c=1}^C (E_c - \bar{E})^2}} \quad (24)$$

C LLMs Implementation Details

C.1 LLMs Overview and Selection

We set up a 5v5 comparison framework to make sure that the cross-paradigm examination was thorough and fair. This includes five cutting-edge general-purpose models that can be accessed through standard APIs: GPT-5.1-mini, Gemini-2.5-Flash, Claude-3.5-Sonnet, DeepSeek-V3, and Qwen3-Max. These models were compared to five representative open-weight financial models that have been successfully used in our local inference environment: XuanYuan3-70B(Zhang and Yang, 2023), Fin-R1-7B(Liu et al., 2025), CFGPT2-7B(Li et al., 2023), DISC-FinLLM-13B(Chen et al., 2023), and FinGPT-7B(Yang et al., 2023a).

We started with a poll of possible candidates and then narrowed it down to this curated list based on strong standards of operational viability and technical transparency. Some intriguing models were left out because of administrative or proprietary issues. For example, InvestLM (HKUST)(Yang et al., 2023b) was still not available since the licensing approval process did not respond to multiple academic enquiries. Industrial systems like BloombergGPT(Wu et al., 2023) and Baichuan4-Finance(Zhang et al., 2024a) were also left out because their weights are strictly closed-source. Zhi Xiaobao, on the other hand, is part of a consumer-facing app that doesn’t have a research-grade API. Also, ICBC and Hundsun’s² internal infrastructure models can only be used in private banking settings, which makes it impossible to verify them independently.

Commercial API limits and architectural restrictions also required certain exclusions. HithinkGPT (Tonghuashun)³ has a conversational interface, but its rigorous token-level quota of only 1,000 tokens with no institutional top-up option made it unsuitable for high-throughput reasoning audits. FinGPT-7B(Yang et al., 2023a) was successfully deployed, but its tokeniser doesn’t work well with Chinese

²<https://www.hs.net/lightgpt/>

³<https://aimiai.com/>

characters; hence, Chinese tasks produce garbled outputs.

To protect the integrity of the data and make sure the comparison was fair, FinGPT-7B’s evaluation was limited to the English sub-benchmarks. Furthermore, across all question types in the Chinese CPA, all financial-specific models face collapse when encountering complex calculation problems, whether original or revised questions.

To directly address concerns about the coverage of recent reasoning-focused large models, we extend the core 5v5 suite with five additional reasoning-enhanced models spanning different families, scales, and training strategies: DeepSeek-R1⁴ and GPT-5.2-series, two frontier reasoning specialists evaluated via their official APIs; GPT-OSS-20B and GPT-OSS-120B⁵, OpenAI’s open-weight Mixture-of-Experts reasoning models at 20B and 120B scale, deployed locally with native MXFP4 quantisation; and DianJin-R1-32B⁶ (Zhu et al., 2025), a Qwen2.5-base financial reasoning model trained with GRPO-style reinforcement learning. These five models were selected to cover the three strategies highlighted as underrepresented in the initial evaluation: reinforcement-learned reasoning (DeepSeek-R1, DianJin-R1-32B), frontier closed-source scaling (GPT-5.2-series), and open-weight reasoning at different parameter scales (GPT-OSS-20B/120B). Per-model configurations and deployment details are given in Appendix C.4–C.10.

Table 10 summarises representative commercial and financial LLMs, with all 15 evaluated models marked with *. Looking ahead, we intend to broaden this research by including a wider spectrum of emerging financial LLMs, such as Fin-o1-8B⁷ (Qian et al., 2025), Finance-Llama⁸, Finance-Qwen⁹, and Fin-PRM (Zhou et al., 2025a). Future editions of this study will look into additional specialised open-source financial models as they become available in order to better benchmark the growing environment of financial AI.

Beyond static model evaluation, our future work will pivot towards autonomous financial intelligence. We aim to explore and develop frameworks

⁴<https://huggingface.co/deepseek-ai/DeepSeek-R1>

⁵<https://huggingface.co/openai/gpt-oss-20b>; <https://huggingface.co/openai/gpt-oss-120b>

⁶<https://huggingface.co/DianJin/DianJin-R1-32B>

⁷<https://huggingface.co/TheFinAI/Fin-o1-8B>

⁸<https://huggingface.co/tarun7r/Finance-Llama-8B>

⁹<https://huggingface.co/WiroAI/WiroAI-Finance-Qwen-1.5B>; <https://huggingface.co/WiroAI/WiroAI-Finance-Qwen-7B>

similar to FinChain (Xie et al., 2025), focusing on structured financial reasoning chains, and FinAgent (Zhang et al., 2024b), which emphasizes multi-agent collaboration for complex task automation. These directions will allow us to transition from benchmarking individual models to evaluating integrated financial AI systems in real-world operational environments.

C.2 Hardware Configuration

The hardware assignment for each locally deployed model followed its memory footprint. XuanYuan3-70B, the largest dense model in our suite (70B parameters, ~140 GB at FP16), was deployed on 2× NVIDIA H20 (96 GB each) with automatic tensor parallelism. DianJin-R1-32B (32B, Qwen2.5 base) fit on a single H20 (96 GB). The two GPT-OSS Mixture-of-Experts models, GPT-OSS-20B (21B total, ~14 GB MXFP4) and GPT-OSS-120B (117B total, ~63 GB MXFP4). They were each deployed on a single NVIDIA H100 (80 GB), since their native MXFP4 quantization and Triton MXFP4 kernel require a newer software environment than the rest of the suite (detailed below). DISC-FinLLM-13B (Baichuan-13B base) ran on a 32 GB vGPU instance. The remaining 7B-scale models (CFGPT2, FinGPT, Fin-R1) each fit on a single NVIDIA RTX 4090 (24 GB) at FP16.

In contrast, the seven frontier general-purpose and reasoning-specialist models, GPT-5.1-mini, GPT-5.2-series, Gemini-2.5-Flash, Claude-3.5-Sonnet, DeepSeek-V3, DeepSeek-R1, and Qwen3-Max. They were evaluated via their official APIs, which eliminated local GPU requirements while ensuring access to the latest model iterations.

All locally deployed experiments shared a unified software stack: PyTorch 2.1.0, Python 3.10 (Ubuntu 22.04), CUDA 12.1, and Transformers 4.38.2, with greedy decoding (`do_sample=False`, `max_new_tokens=1024`, `repetition_penalty=1.05`) to ensure deterministic outputs. The two GPT-OSS models are the only exception: because they rely on native MXFP4 quantization and the Triton `matmul_ogs` MoE kernel unavailable in Transformers 4.38, they were run in a separate environment (Transformers ≥ 4.55 , triton ≥ 3.4 , and the kernels package) while retaining identical decoding settings.

C.3 Prompts and Structured Output Protocol

All models are assessed under a unified protocol tailored to their specific deployment:

Category	Organization	Model	Size	Key Features / Notes	Source / Reference
General Purpose	Closed-source	GPT-5.1-mini*	–	Frontier general-purpose performance	openai/gpt-5.1-mini
		GPT-5.2-series*	–	Reasoning-enhanced frontier model	openai/gpt-5.2
		Claude-3.5-Sonnet*	–	Strong reasoning and stability	anthropic/claude-3.5
		Gemini-2.5-Flash*	–	High-efficiency reasoning model	google/gemini-2.5
		Qwen3-Max*	~1T	Proprietary MoE, leading multilingual	qwen/qwen3-max
	Open-source	DeepSeek-V3*	37B/671B	Advanced MoE reasoning focus	deepseek/deepseek-v3
		DeepSeek-R1*	37B/671B	RL-trained reasoning specialist	deepseek-ai/DeepSeek-R1
		GPT-OSS-20B*	20B	Open-weight MoE, native MXFP4	openai/gpt-oss-20b
GPT-OSS-120B*		120B	Open-weight MoE, native MXFP4	openai/gpt-oss-120b	
Academic	Financial	FinGPT-7B*	7B	Multi-source data curriculum	AI4Finance/fin-gpt-7b
		InvestLM	65B	Expert-level CFA-style reasoning	HKUST/investlm
		DISC-FinLLM-13B*	13B	Multi-turn advisory optimization	Fudan/disc-finllm-13b
		CFGPT2-7B*	7B	Knowledge graph integration	Tongji/cfgpt2-7b
		Fin-R1-7B*	7B	GRPO reinforcement learning	SUFE/fin-r1-7b
		DianJin-R1-32B*	32B	GRPO RL on Qwen2.5 base	DianJin/DianJin-R1-32B
		FinMA-7B	7B	Multi-task instruction tuning	Pixiu/finma-7b
Industrial	Financial	BloombergGPT	50B	Proprietary domain pre-training	bloomberg/bloomberggpt
		XuanYuan3-70B*	70B	Long-report synthesis focus	duxiaoman/xuanyuan3-70b
		Baichuan4-Fin	–	Professional logic alignment	baichuan/baichuan4-fin
		HithinkGPT	–	Real-time API integration	hithink/hithinkgpt
		Zhi Xiaobao	–	Consumer wealth management	antgroup/zhixiaobao
Infra-structure	Financial	ICBC Zhiyong	–	Risk control and compliance	icbc/zhiyong
		Light GPT	–	Securities trading backend	lightgpt/lightgpt

Table 10: Comprehensive Taxonomy of Large Language Models in the Financial Domain. Models marked with (*) are those specifically evaluated in our empirical study, including 10 core-suite models and 5 extended reasoning-enhanced models added during revision.

locally implemented financial models utilize their native chat templates via `tokenizer.apply_chat_template()` to maintain strict alignment with original training paradigms, whereas general-purpose commercial models are accessed directly via standard API endpoints.

All experiments, which include the *Original*, *Revised*, and *NOTA* settings, employ bilingual prompts mandating a structured JSON output with three fields: *reason*, *answer*, and *confidence*. To ensure robust data extraction and minimize evaluation bias, we develop a cascaded four-stage parsing strategy: the system first attempts to parse the entire output string using standard library decoders;

if this fails, it proceeds to extract content within `json` or `markdown` tags; if the output remains unparsable, regular expressions are applied to identify the outermost balanced curly braces `{...}` to isolate the JSON object from potential conversational filler; finally, in cases of severe syntax corruption, we utilize field-level regex fallback to recover the *reason*, *answer*, and *confidence* fields independently, thereby maximizing the recovery of partial but valid reasoning. For multiple-choice questions, extracted answers are normalized to uppercase and sorted alphabetically.

One major thing we saw when we looked at the five open-weight financial LLMs is the rise of

a "formatting tax". This means that using strict JSON syntax usually lowers accuracy by 3% to 4% compared to free-form output. We hypothesize that the cognitive overhead of constraint satisfaction competes with the neural resources required for complex multi-step financial reasoning. However, we prioritize the structured protocol to ensure Epistemic Awareness can be explicitly audited through the reason field, decoupling logical deduction from linguistic fluency.

Chinese Prompt Template (CPA)

System: 您是一位金融问题解决专家。您将收到一道单选或多选题。请返回一个严格的JSON数据，包含三个键："reason" 和 "answer" 和 "confidence"。

User:

问题: {question}

请严格按照以下格式输出JSON数据。请勿在JSON对象之外包含任何文本。

```
{
  "reason": "简要解释这些选项正确的原因。",
  "answer": "仅输出大写字母(A-Z)。对于多选题，请返回按字母顺序排列并连接起来的字母集合（例如，ACD）。对于单选题，请返回字母，例如A或B或C或D。",
  "confidence": "你对本题你所回答的答案的信心程度是多少，在0-100里选一个。"
}
```

Figure 3: The Chinese prompt template used for CPA-style evaluation. Applied consistently across all test variants.

C.4 XuanYuan3-70B Configuration

Model **Source.** [Duxiaoman-DI/Llama3-XuanYuan3-70B-Chat](https://huggingface.co/Duxiaoman-DI/Llama3-XuanYuan3-70B-Chat)¹⁰ (Zhang and Yang, 2023)

XuanYuan3-70B is a state-of-the-art Chinese large language financial model developed by Du Xiaoman. Building upon the Llama-3-70B architecture, it extends the foundational capabilities established in its predecessor, XuanYuan 2.0, through advanced post-training and financial-domain-specific instruction tuning. It is designed to handle sophisticated financial analysis, regulatory compliance, and cross-lingual financial reasoning with high parameter efficiency and domain expertise. We deploy XuanYuan3-70B using 2× NVIDIA H20 GPUs (96GB each) with automatic tensor parallelism, FP16 precision, greedy decoding, and

¹⁰<https://huggingface.co/Duxiaoman-DI/Llama3-XuanYuan3-70B>

English Prompt Template (CFA)

System: You are an expert in financial problem solving. You will be given a single- or multi-choice question. Please return STRICT JSON with three keys: "reason", "answer" and "confidence".

User:

Question: {question}

Output JSON EXACTLY in the following format. Do NOT include any text outside the JSON object.

```
{
  "reason": "a concise explanation of why these options are correct.",
  "answer": "output uppercase letters only (A-Z). For multi-choice, return the set of letters sorted alphabetically and concatenated (e.g., ACD). For single-choice, return letter like A or B or C or D.",
  "confidence": "How confident are you in your answer to this question? Choose a score between 0 and 100."
}
```

Figure 4: The English prompt template used for CFA-style evaluation. Applied consistently across all test variants.

repetition_penalty=1.05.

Chat Template. XuanYuan3-70B follows the native Llama-3 chat template, in which system, user, and assistant roles are explicitly encoded at the tokenizer level. We apply the template via `tokenizer.apply_chat_template()` with `add_generation_prompt=True` (Figure 5).

```
<|begin_of_text|>
<|start_header_id|>system<|end_header_id|>

{system}<|eot_id|>
<|start_header_id|>user<|end_header_id|>

{user}<|eot_id|>
<|start_header_id|>assistant<|end_header_id|>
```

Figure 5: Llama-3 chat template used by XuanYuan3-70B.

Detailed Results. Table 11 demonstrates that the new strategy performs considerably better in all areas of finance. The model achieves nearly perfect results in Portfolio Management (100%) and Derivatives Investment (92.00%) on the English CFA track. The model shows a significant improvement in the Chinese CPA track, which is more crucial. Accuracy increased by 70.91% in Auditing and 71.71% in Economic Law. This shows that the

optimization was effective in assisting the model in dealing with complex laws and regulations unique to distinct places.

Table 3 illustrates the proficiency gains across specific task categories. We observe a significant advancement in mathematical reasoning; within the English CFA, accuracy in Complex Calculation reaches 100.0% under revised settings. Similarly, in the Chinese CPA, performance in Simple Calculation increases by more than 50.0%. Most notably, XuanYuan3-70B effectively addresses the zero-shot performance collapse in CPA Knowledge Application, elevating accuracy from 0.00% to 66.67%. These results indicate that the model maintains professional-level proficiency across both conceptual frameworks and quantitative analysis.

Robustness Analysis under Adversarial Masking. The performance of XuanYuan3-70B within the NOTA-masked adversarial environment reveals a significant logical calibration challenge inherent to large-scale parameter models (Table 6). The model exhibits a pronounced decoupling between reasoning trajectories and final label selection. Specifically, qualitative analysis shows that while the internal reasoning often converges on a specific numerical option, the model ultimately selects the “None of the Above” (NOTA) label due to surface-level pattern matching of the adversarial prompt structure. Furthermore, the model maintains exceptionally high confidence levels ($\mu > 90.0$) despite logical validity falling below 10%, indicating a severe overconfidence bias. This suggests that while the model successfully identifies adversarial markers with high frequency in the English CFA (exceeding 80%), it fails to ground its selection in a valid deductive process.

In summary, whilst XuanYuan3-70B utilizes a parameter count ten times that of Fin-R1-7B, it exhibits significantly lower logical consistency. This disparity suggests that the structural reasoning constraints established through reinforcement learning, specifically GRPO, represent a fundamental capability that cannot be replaced by the brute-force aesthetics of parameter expansion. These findings underscore that specialized reasoning paths are not emergent properties of scale alone but require targeted algorithmic constraints to ensure deductive integrity.

C.5 Fin-R1-7B Configuration

Model Source. SUFE-AIFLM-Lab/Fin-R1¹¹ (Liu et al., 2025)

Fin-R1 is developed by Shanghai University of Finance and Economics (SUFE), based on Qwen2.5-7B. It achieves state-of-the-art reasoning performance via GRPO-style reinforcement learning for complex financial problems. As of December 2025, it nearly matched the results of XuanYuan3-70B. We use a single NVIDIA RTX 4090 (24GB) with FP16 precision, greedy decoding, and `repetition_penalty=1.05`.

Chat Template. Fin-R1 inherits the Qwen2.5 chat template, accessible via `tokenizer.apply_chat_template()`. We use the standard message format with system and user roles:

```
messages = [
    {"role": "system", "content": system_prompt},
    {"role": "user", "content": user_prompt},
]
prompt = tokenizer.apply_chat_template(
    messages,
    tokenize=False,
    add_generation_prompt=True
)
```

Figure 6: Fin-R1 inference using official Qwen2.5 chat template.

Note: The official Fin-R1 repository recommends deployment via vLLM with an OpenAI-compatible API and a specific system prompt using `<think>...</think><answer>...</answer>` format for chain-of-thought reasoning. For consistency with our unified evaluation framework, we use direct inference with our standard JSON-output prompt instead.

Detailed Results. The evaluation results for Fin-R1-7B (Table 13 and 14) present several significant insights. Despite possessing merely 7 billion parameters, Fin-R1-7B attains an overall accuracy of **83.65%** in English CFA and **73.14%** in Chinese CPA, closely approaching the performance of the significantly larger XuanYuan3-70B. This illustrates that GRPO-style reinforcement learning markedly improves the parameter efficacy of the base model, allowing it to internalise complex financial reasoning processes that generally necessitate higher-capacity architectures.

¹¹<https://huggingface.co/SUFE-AIFLM-Lab/Fin-R1>

Question Topic (CFA)	Orig.	Rev.	Δ	n	Question Topic (CPA)	Orig.	Rev.	Δ	n
Corporate Finance	60.00	90.00	+30.00	10	Accounting	4.49	63.33	+58.84	89
Derivatives Investment	53.33	92.00	+38.67	45	Auditing	25.76	96.67	+70.91	66
Economics	53.33	80.00	+26.67	15	Economic Law	6.67	78.38	+71.71	60
Equity Investment	57.63	86.54	+28.91	59	Tax Law	15.79	66.67	+50.88	57
Financial Statement Analysis	65.12	88.89	+23.77	43	Wealth Management	0.00	44.44	+44.44	32
Fixed Income Investment	42.86	78.57	+35.71	49					
Other Investments	62.77	86.42	+23.65	94					
Portfolio Management	75.00	100.00	+25.00	4					
Quantitative Analysis	65.96	86.67	+20.71	47					
Overall	58.47	84.74	+26.27	366	Overall	11.18	69.14	+57.96	304

Table 11: Accuracy (%) of XuanYuan3-70B by question topic. Left: English CFA (9 subjects). Right: Chinese CPA (5 subjects). n denotes the number of original questions.

Question Type	English CFA				Chinese CPA			
	Orig.	Rev.	Δ	n	Orig.	Rev.	Δ	n
Conceptual Understanding	77.42	79.07	+1.65	62	10.92	80.00	+69.08	119
Simple Calculation	36.67	82.12	+45.46	30	9.76	64.15	+54.39	41
Complex Calculation	33.33	100.00	+66.67	15	0.00	–	–	2
Comprehensive Judgment	62.38	88.37	+26.00	101	10.34	71.72	+61.37	87
Knowledge Application	40.70	75.00	+34.30	86	0.00	66.67	+66.67	7
Statistical Methods	71.43	93.62	+22.19	70	16.67	60.00	+43.33	48
Overall	58.47	84.74	+26.27	364	11.18	69.14	+57.96	304

Table 12: Accuracy (%) of XuanYuan3-70B by question type across English CFA and Chinese CPA examinations.

Question Topic (CFA)	Orig.	Rev.	Δ	n	Question Topic (CPA)	Orig.	Rev.	Δ	n
Corporate Finance	50.00	90.00	+40.00	10	Accounting	6.74	66.67	+59.93	89
Derivatives Investment	53.33	88.00	+34.67	45	Auditing	13.64	83.33	+69.70	66
Economics	53.33	60.00	+6.67	15	Economic Law	6.67	91.89	+85.22	60
Equity Investment	62.71	86.54	+23.83	59	Tax Law	7.02	71.43	+64.41	57
Financial Statement Analysis	39.53	81.48	+41.95	43	Wealth Management	0.00	52.78	+52.78	32
Fixed Income Investment	32.65	85.71	+53.06	49					
Other Investments	57.45	87.65	+30.21	94					
Portfolio Management	75.00	100.00	+25.00	4					
Quantitative Analysis	72.34	88.89	+16.55	47					
Overall	54.10	83.65	+29.55	366	Overall	7.57	73.14	+65.57	304

Table 13: Accuracy (%) of Fin-R1-7B by question topic. Left: English CFA (9 subjects). Right: Chinese CPA (5 subjects). n denotes the number of original questions.

Fin-R1-7B exhibits a transformative accuracy surge within the Chinese CPA benchmark, rising from a near-random baseline of 7.57% to 73.14% ($\Delta+65.57\%$). Most notably, in Economic Law, the model achieves a peak accuracy of 91.89%, surpassing several significantly larger architectures. This trajectory suggests that reasoning-oriented fine-tuning generalizes effectively across linguistic boundaries, even when navigating localized regulatory and legal frameworks.

As detailed in Table 3, Fin-R1-7B demonstrates substantial advancements in quantitative reasoning. Performance in Complex Calculation jumped from 26.67% to 100.0%, while Simple Calculation

witnessed a gain exceeding 32%. This improvement represents a hallmark of the reasoning model (R1) lineage, where the integration of structured chain-of-thought (CoT) processes—even when constrained to a JSON output format—significantly mitigates arithmetic and logical hallucinations.

The model maintains balanced proficiency across diverse cognitive tasks, specifically within Conceptual Understanding (90.00% in CPA) and Statistical Methods (93.62% in CFA). The consistent improvement (Δ) across all task categories indicates that the reinforcement learning process does not merely optimize for specific keyword triggers but instead fosters a systematic enhancement

in professional judgment and deductive integrity.

Robustness Analysis under Adversarial Masking.

Fin-R1-7B demonstrates significant resilience in adversarial settings where the correct response is “None of the Above” (NOTA), achieving accuracies of 84.5% in the CFA and 77.7% in the CPA. Unlike general-purpose models that often exhibit a nearest-neighbor bias—the tendency to select the numerical option closest to an erroneous internal calculation—Fin-R1-7B maintains high reasoning consistency by actively rejecting plausible distractors through self-verification. The strong alignment between its high mean confidence ($\mu > 90.0$) and actual accuracy indicates a well-calibrated professional judgment likely fostered by GRPO reinforcement learning. This training process appears to effectively penalize logical shortcuts, enabling the model to identify specific flaws in distractor logic, such as the failure to account for tax-exempt income deductions. These findings highlight the parameter efficiency of specialized reasoning models, as the 7B architecture develops robust, logic-driven heuristics that rival the performance of much larger general architectures.

C.6 CFGPT2-7B Configuration

Model Source. TongjiFinLab/CFGPT2-7B¹² (Li et al., 2023)

CFGPT2 is a Chinese financial assistant developed by Tongji University, built upon the InternLM-7B base. It is pre-trained and fine-tuned on a massive corpus of Chinese financial data. We use a single NVIDIA RTX 4090 (24GB) with BF16 precision.

Chat Teampate. CFGPT2 uses the InternLM-style `model.chat()` method. The official example uses an empty `meta_instruction`, but the interface supports passing system-level instructions:

```
response, history = model.chat(
    tokenizer=tokenizer,
    query=user_prompt,
    history=[],
    max_new_tokens=1024,
    do_sample=False,
    repetition_penalty=1.1,
    meta_instruction=system_prompt
)
```

Figure 7: CFGPT2-7B inference code (InternLM-style).

¹²<https://huggingface.co/TongjiFinLab/CFGPT2-7B>

Detailed Results. The evaluation of CFGPT2-7B (Table 15 and 16) highlights the significant impact of [Optimization Method] on a domain-specific 7B model.

The original model (*Orig.*) did almost nothing in the Chinese CPA track (2.96%), which shows that it was very out of line with professional Chinese financial norms. But the amended version (*Rev.*) jumped to 49.71%, showing that fine-tuning can quickly bring out hidden financial information.

Like bigger reasoning models, CFGPT2-7B got a 100% accuracy in English *Complex Calculation* after revision. The increases in *Simple Calculation* were more than 38% and 44% across both benchmarks.

The model is quite good in CFA’s *Corporate Finance* (90.00%) and *Portfolio Management* (100%). In CPA, it scored 80.00% in *Statistical Methods*, making it a great candidate for quantitative financial jobs even though it has fewer parameters.

Robustness Analysis under Adversarial Masking.

CFGPT2-7B exhibits a critical *reasoning-selection gap*, where its moderate *Answer Accuracy* (75.2% in CFA; 54.9% in CPA) is starkly contradicted by its near-complete collapse in *Reasoning Accuracy* (~10%). This suggests that the model’s correct outputs are largely driven by stochastic output heuristics rather than deterministic derivation. Unlike Fin-R1-7B, CFGPT2-7B lacks a functional verification closed-loop (i.e., the logical gate: “Calculated Result \neq Candidates \rightarrow Select NOTA”), often leading to a correct label despite a fallacious reasoning path.

The performance of CFGPT2-7B is characterized by unstable hallucination-driven hits. Qualitative traces indicate that even when the model correctly identifies the NOTA mask, its CoT frequently contains factual distortions or logical non-sequiturs. This implies that financial domain knowledge has not been internalized into a robust reasoning faculty but remains at the level of probabilistic surface matching.

Despite its logical invalidity, CFGPT2-7B maintains an average confidence score of $\mu > 81.0$. This severe miscalibration represents a significant risk in financial applications; the model exhibits a “blind certainty,” providing authoritative yet erroneous justifications.

Question Type	English CFA				Chinese CPA			
	Orig.	Rev.	Δ	n	Orig.	Rev.	Δ	n
Conceptual Understanding	69.35	76.74	+7.39	62	7.56	90.00	+82.44	119
Simple Calculation	50.00	82.12	+32.12	30	7.32	71.70	+64.38	41
Complex Calculation	26.67	100.00	+73.33	15	0.00	–	–	2
Comprehensive Judgment	57.43	84.88	+27.45	101	5.75	72.73	+66.98	87
Knowledge Application	38.37	75.00	+36.63	86	14.29	66.67	+52.38	7
Statistical Methods	61.43	93.62	+32.19	70	10.42	70.00	+59.58	48
Overall	54.10	83.65	+29.55	364	7.57	73.14	+65.57	304

Table 14: Accuracy (%) of Fin-R1-7B by question type across English CFA and Chinese CPA examinations.

Question Topic (CFA)	Orig.	Rev.	Δ	n	Question Topic (CPA)	Orig.	Rev.	Δ	n
Corporate Finance	60.00	90.00	+30.00	10	Accounting	2.25	60.00	+57.75	89
Derivatives Investment	31.11	76.00	+44.89	45	Auditing	7.58	60.00	+52.42	66
Economics	53.33	60.00	+6.67	15	Economic Law	1.67	45.95	+44.28	60
Equity Investment	59.32	78.85	+19.53	59	Tax Law	1.75	40.48	+38.73	57
Financial Statement Analysis	53.49	66.67	+13.18	43	Wealth Management	0.00	47.22	+47.22	32
Fixed Income Investment	36.73	57.14	+20.41	49					
Other Investments	50.00	77.78	+27.78	94					
Portfolio Management	75.00	100.00	+25.00	4					
Quantitative Analysis	42.55	73.33	+30.78	47					
Overall	47.54	73.57	+26.03	366	Overall	2.96	49.71	+46.75	304

Table 15: Accuracy (%) of CFGPT2-7B by question topic. Left: English CFA (9 subjects). Right: Chinese CPA (5 subjects). n denotes the number of original questions.

Question Type	English CFA				Chinese CPA			
	Orig.	Rev.	Δ	n	Orig.	Rev.	Δ	n
Conceptual Understanding	56.45	76.74	+20.29	62	2.52	40.00	+37.48	119
Simple Calculation	30.00	68.16	+38.16	30	4.88	49.06	+44.18	41
Complex Calculation	40.00	100.00	+60.00	15	0.00	–	–	2
Comprehensive Judgment	60.40	80.23	+19.83	101	2.30	49.49	+47.19	87
Knowledge Application	31.40	75.00	+43.60	86	0.00	0.00	0.00	7
Statistical Methods	48.57	76.60	+28.03	70	4.17	80.00	+75.83	48
Overall	47.54	73.57	+26.03	364	2.96	49.71	+46.75	304

Table 16: Accuracy (%) of CFGPT2-7B by question type across English CFA and Chinese CPA examinations.

C.7 DISC-FinLLM-13B Configuration

Model Source. Go4miii/DISC-FinLLM¹³ (Chen et al., 2023)

DISC-FinLLM is a specialised financial large language model fine-tuned on the Baichuan-13B base. It is worth noting that DISC-FinLLM is primarily engineered for multi-turn financial consulting dialogues rather than single-turn MCQ tasks. We use a virtual GPU (vGPU-32GB) with FP16 precision and greedy decoding.

Chat Template. DISC-FinLLM uses the Baichuan-style `model.chat()` method, which does not support a separate system role. To incorporate our evaluation instructions, we concatenate

the system prompt with the user query into a single user message:

```
full_content = f"{system_prompt}\n\n{user_content}"
messages = [{"role": "user", "content": full_content}]
response = model.chat(tokenizer, messages)
```

Figure 8: DISC-FinLLM inference with merged system prompt.

Detailed Results. The evaluation results for DISC-FinLLM (Table 17 and 18) provide insights into the cross-task generalisation of dialogue-

¹³<https://huggingface.co/Go4miii/DISC-FinLLM>

specialised financial models

As a model primarily optimised for multi-turn financial consulting, DISC-FinLLM exhibited relatively low baseline accuracy in both the English CFA (13.11%) and Chinese CPA (4.61%) tracks. This suggests that the instruction-tuning for conversational equity does not naturally translate into the rigorous logical deduction required for professional MCQs.

Although the revised version (*Rev.*) showed improvement, climbing to 19.89% in CFA and 19.43% in CPA. The gains are significantly more modest compared to reasoning-dense models like Fin-R1. Notably, in several CFA subjects such as *Economics* and *Equity Investment*, the model even experienced performance regression ($\Delta < 0$), indicating a potential instruction interference where revised prompts might conflict with its conversational priors.

The question-type breakdown reveals that DISC-FinLLM-13B struggles with Statistical Methods and Quantitative Analysis, often scoring near or below the expected value of random guessing. While it showed a minor breakthrough in CPA *Conceptual Understanding* (rising from 4.20% to 30.00%), its overall performance trajectory confirms that without specific reinforcement learning for reasoning paths, dialogue-centric models remain insufficient for autonomous high-stakes financial examination tasks.

Robustness Analysis under Adversarial Masking. The erratic performance and diminished confidence ratings, particularly in English CFA, suggest that DISC-FinLLM-13B is acutely susceptible to hostile prompt alterations, lacking the ability to generalise outside its conversational tuning distribution. The deterioration of reasoning accuracy indicates that the model lacks the structural logic necessary to recognise absent information, instead depending on language templates that are readily undermined in NOTA situations.

C.8 FinGPT Configuration

Model **Source.** *AI4Finance-Foundation/FinGPT* (Yang et al., 2023a).¹⁴ We utilised the **Qwen-7B-based variant** of FinGPT to evaluate the English CFA tracks (Original and Revised). Among the FinGPT variant family

¹⁴Base and LoRA adapter: <https://huggingface.co/Qwen/Qwen-7B> and https://huggingface.co/FinGPT/fin-gpt-mt-qwen-7b_lora respectively.

```
# FinGPT (Qwen-7B base + LoRA) inference logic
prompt = f"Instruction: {system_prompt}\n
Input: {user_prompt}\n
Answer: "
```

Figure 9: FinGPT (Qwen-7B base + LoRA) inference template for English CFA tasks. All FinGPT variants share this flat Instruction/Input/Answer format; no chat template is applied, since FinGPT’s LoRA was fine-tuned directly on the non-chat base checkpoint.

(Llama-2, Falcon, Bloom, MPT, ChatGLM2, Qwen), we selected the Qwen variant because the Qwen-7B tokeniser natively handles Chinese characters, whereas Llama-2- and Falcon-based variants corrupt Chinese inputs at the tokenisation stage. Preliminary experiments, however, revealed a second, subtler limitation specific to this variant, which we term **Reasoning Dissociation**: the LoRA adapter, fine-tuned primarily on English financial sentiment data, interferes with the Qwen-7B base’s generative consistency in zero-shot professional reasoning. On English CFA items we observed fragmented, largely uninformative outputs (e.g., truncated phrases such as `pass_through` or `Acorp` in place of structured reasoning), which precluded meaningful accuracy gains from the adapter. To preserve evaluation integrity across our model suite, we therefore limited FinGPT’s evaluation to the English CFA sub-benchmark.

Chat Template. FinGPT utilises an explicit Instruction/Input/Answer format, inherited from its multi-task fine-tuning paradigm in which the non-chat base checkpoint is adapted via LoRA directly on plain-text instructions. To maintain consistency with this training distribution, we mapped the evaluation system prompt to the *Instruction* field and the user prompt to the *Input* field (Figure 9). Critically, we do *not* invoke `tokenizer.apply_chat_template()` here, since the Qwen-7B *base* checkpoint (as distinct from *Qwen-7B-Chat*) does not ship with a chat template, and FinGPT’s adapter was not trained under any chat-formatted input distribution.

Detailed Results. As shown in Table 19, FinGPT-7B achieved an overall accuracy of only **2.19%** (Original) and 3.54% (Revised), significantly below random guessing. Qualitative inspection revealed that the model frequently outputs sentiment markers (e.g., “positive/negative”) or jumbled char-

Question Topic (CFA)	Orig.	Rev.	Δ	n	Question Topic (CPA)	Orig.	Rev.	Δ	n
Corporate Finance	30.00	40.00	+10.00	10	Accounting	4.49	26.67	+22.18	89
Derivatives Investment	20.00	28.00	+8.00	45	Auditing	4.55	26.67	+22.12	66
Economics	13.33	0.00	-13.33	15	Economic Law	6.67	18.92	+12.25	60
Equity Investment	20.34	13.46	-6.88	59	Tax Law	5.26	14.29	+9.03	57
Financial Statement Analysis	11.63	22.22	+10.59	43	Wealth Management	0.00	13.89	+13.89	32
Fixed Income Investment	14.29	14.29	0.00	49					
Other Investments	10.64	23.46	+12.82	94					
Portfolio Management	0.00	33.33	+33.33	4					
Quantitative Analysis	0.00	6.67	+6.67	47					
Overall	13.11	19.89	+6.78	366	Overall	4.61	19.43	+14.82	304

Table 17: Accuracy (%) of DISC-FinLLM-13B by question topic. Left: English CFA (9 subjects). Right: Chinese CPA (5 subjects). n denotes the number of original questions.

Question Type	English CFA				Chinese CPA			
	Orig.	Rev.	Δ	n	Orig.	Rev.	Δ	n
Conceptual Understanding	8.06	18.60	+10.54	62	4.20	30.00	+25.80	119
Simple Calculation	23.33	20.67	-2.66	30	4.88	15.09	+10.21	41
Complex Calculation	6.67	25.00	+18.33	15	0.00	-	-	2
Comprehensive Judgment	22.77	22.09	-0.68	101	5.75	20.20	+14.45	87
Knowledge Application	10.47	12.50	+2.03	86	14.29	0.00	-14.29	7
Statistical Methods	2.86	14.89	+12.03	70	2.08	30.00	+27.92	48
Overall	13.11	19.89	+6.78	364	4.61	19.43	+14.82	304

Table 18: Accuracy (%) of DISC-FinLLM-13B by question type across English CFA and Chinese CPA examinations.

Topic	Orig.	Rev.	Δ	n
Corp. Fin.	0.00	0.00	0.00	10
Derivatives	6.67	4.00	-2.67	45
Economics	0.00	0.00	0.00	15
Equity	1.69	3.85	+2.16	59
FSA	2.33	3.70	+1.37	43
Fixed Inc.	0.00	0.00	0.00	49
Other Inv.	3.19	3.70	+0.51	94
Portfolio	0.00	33.33	+33.33	4
Quant.	0.00	2.22	+2.22	47
Overall	2.19	3.54	+1.35	366
Type	Orig.	Rev.	Δ	n
Conceptual	1.61	6.98	+5.37	62
Simple Calc.	6.67	2.79	-3.88	30
Complex Calc.	0.00	0.00	0.00	15
Comp. Judgment	1.98	4.65	+2.67	101
Know. App.	0.00	0.00	0.00	86
Stat. Methods	4.29	2.13	-2.16	70
Overall	2.19	3.54	+1.35	364

Table 19: Performance of FinGPT-7B on English CFA Benchmark.

acters instead of logical steps. This confirms that its fine-tuning distribution—highly specialized for sentiment classification—severely compromises its capacity for multi-step financial deduction.

Robustness Analysis under Adversarial Masking. FinGPT-7B undergoes a total functional col-

lapse in NOTA-masked scenarios (Table 6). The near-zero reasoning accuracy and stagnant confidence score (~ 50.0) reveal an absolute lack of *logical plasticity*. Rather than identifying missing information, the model reverts to fixed templates, proving that sentiment-optimized models remain inherently fragile in adversarial professional environments.

C.9 GPT-OSS-20B and GPT-OSS-120B Configuration

Model Sources. openai/gpt-oss-20b¹⁵ and openai/gpt-oss-120b¹⁶

The two GPT-OSS models are open-weight reasoning Mixture-of-Experts (MoE) models released by OpenAI under the Apache 2.0 licence. GPT-OSS-20B has 21B total parameters with approximately 3.6B active parameters per forward pass (32 experts, top-4 routing). GPT-OSS-120B has 117B total parameters with approximately 5.1B active parameters (128 experts, top-4 routing). Both ship with native MXFP4 quantisation of the MoE weights and BF16 storage for all other tensors. The MXFP4 footprint is approximately 14 GB for the 20B variant and 63 GB for the 120B variant,

¹⁵<https://huggingface.co/openai/gpt-oss-20b>

¹⁶<https://huggingface.co/openai/gpt-oss-120b>

reduced from ~ 240 GB at BF16. Both models were deployed on a single NVIDIA H100 (80 GB), which motivates their joint treatment in this subsection. Including both variants lets us isolate the effect of parameter scaling under a fixed training paradigm and chat template.

Chat Template. Both models use the built-in Harmony response format, accessed through the HuggingFace transformers pipeline at `transformers \geq 4.55`, together with the Triton `matmul_ogs MXFP4` kernel (`triton \geq 3.4`, kernels). The chat template and decoding settings are identical across the two models:

```
messages = [
    {"role": "system", "content": system_prompt},
    {"role": "user", "content": user_prompt},
]
prompt = tokenizer.apply_chat_template(
    messages, tokenize=False,
    add_generation_prompt=True)

# greedy decoding, consistent with the suite
outputs = model.generate(
    **tokenizer(prompt, return_tensors="pt"),
    do_sample=False,
    max_new_tokens=1024,
    repetition_penalty=1.05,
)
```

Figure 10: GPT-OSS inference code (Harmony chat template). The same code applies to both GPT-OSS-20B and GPT-OSS-120B.

Detailed Results: GPT-OSS-20B. Table 20 shows a profile typical of a general-purpose reasoning model rather than a finance-tuned one. English CFA *Orig.* accuracy is already 78.22% without any financial fine-tuning, and *Rev.* accuracy reaches 88.58% (+10.36%). Chinese CPA reveals a sharper pattern. *Orig.* accuracy collapses to 12.00%, the same dead-zone signature observed in Chinese-financial-specialized models, yet *Rev.* accuracy rises to 77.63% (+65.63%). A 65-point Chinese jump in the absence of Chinese-CPA training data is consistent with our main-paper account of *implicit assumption alignment*: the model fills the missing premise with a plausible default and commits to it.

English performance is uniformly strong, with 100% on *Portfolio Management* and above 85% on *Quantitative Analysis*, *Equity Investment*, and *Economics*. The weakest topic is *Fixed Income Investment* at 63.41%. On Chinese CPA, *Rev.* accuracy

exceeds 80% on the three rule-prescriptive topics (*Auditing*, *Economic Law*, *Tax Law*) but remains below 65% on *Wealth Management*, the cross-product reasoning topic.

Detailed Results: GPT-OSS-120B. Table 22 allows a direct scaling comparison at fixed training paradigm. A roughly $6\times$ increase in total parameters yields less than three points of improvement in either English setting: *Orig.* rises from 78.22% to 80.39% and *Rev.* from 88.58% to 89.10%. On Chinese CPA, *Orig.* accuracy actually declines from 12.00% to 10.53%. The Chinese dead-zone is therefore structural rather than capacity-bound. The Chinese *Rev.* gain is larger for the 120B model (+74.05%) than for the 20B (+65.63%), indicating that the larger model commits to its default assumptions with greater confidence, not with greater caution.

Scaling compresses the dynamic range across English topics. *Fixed Income Investment* rises from 63.41% to 70.21%, while *Quantitative Analysis* is essentially unchanged. On Chinese CPA, the two rule-prescriptive topics most amenable to default substitution (*Economic Law*, *Tax Law*) exceed 91% under *Rev.*, while *Wealth Management* still lags at 75.00%. This asymmetry supports the main-paper claim that cross-product reasoning resists implicit assumption alignment more effectively than single-rule lookup does.

Robustness Analysis under Adversarial Masking. The NOTA-masked results for the GPT-OSS pair (Table 6) provide the clearest evidence in our evaluation that parameter scaling alone does not resolve the structural failure we study.

GPT-OSS-20B exhibits a severe *reasoning-selection gap*. *Answer Accuracy* is 86.65% on English and 67.43% on Chinese, while *Reasoning Accuracy* collapses to 12.26% and 4.00%. Qualitative inspection of the Harmony-format chains of thought confirms that the model seldom identifies the missing premise. Instead it commits to a specific numerical or regulatory default and proceeds as if the question were fully specified. The 20B variant also reports mean confidence $\mu = 90.66$ on the English NOTA subset despite only 12.26% reasoning accuracy, a clearer instance of the *blind certainty* pattern first named for CFGPT2-7B (Appendix C.6).

Scaling to 120B raises *Answer Accuracy* to 89.10% on English and 84.57% on Chinese. At face value this suggests substantial improvement.

Question Topic (CFA)	Orig.	Rev.	Δ	n	Question Topic (CPA)	Orig.	Rev.	Δ	n
Corporate Finance	80.00	90.00	+10.00	10	Accounting	10.00	45.83	+35.83	89
Derivatives Investment	79.41	91.67	+12.25	45	Auditing	19.30	93.33	+74.04	66
Economics	90.91	90.00	-0.91	15	Economic Law	4.88	81.25	+76.37	60
Equity Investment	83.02	88.46	+5.44	59	Tax Law	10.26	94.59	+84.34	57
Financial Statement Analysis	76.19	85.19	+8.99	43	Wealth Management	14.29	62.07	+47.78	32
Fixed Income Investment	63.41	78.57	+15.16	49					
Other Investments	76.19	88.89	+12.70	94					
Portfolio Management	100.00	100.00	+0.00	4					
Quantitative Analysis	85.11	88.89	+3.78	47					
Overall	78.22	88.58	+10.36	366	Overall	12.00	77.63	+65.63	304

Table 20: Accuracy (%) of GPT-OSS-20B by question topic. Left: English CFA (9 subjects). Right: Chinese CPA (5 subjects). n denotes the number of original questions.

Question Type	English CFA				Chinese CPA			
	Orig.	Rev.	Δ	n	Orig.	Rev.	Δ	n
Conceptual Understanding	85.25	83.72	-1.52	62	12.12	90.00	+77.88	119
Simple Calculation	86.36	88.37	+2.01	30	4.76	65.00	+60.24	41
Complex Calculation	75.00	100.00	+25.00	15	0.00	-	-	2
Comprehensive Judgment	74.19	89.41	+15.22	101	6.56	84.62	+78.06	87
Knowledge Application	63.64	87.50	+23.86	86	16.67	33.33	+16.67	7
Statistical Methods	88.57	91.49	+2.92	70	24.32	62.50	+38.18	48
Overall	78.22	88.58	+10.36	364	12.00	77.63	+65.63	304

Table 21: Accuracy (%) of GPT-OSS-20B by question type across English CFA and Chinese CPA examinations.

Question Topic (CFA)	Orig.	Rev.	Δ	n	Question Topic (CPA)	Orig.	Rev.	Δ	n
Corporate Finance	80.00	90.00	+10.00	10	Accounting	5.62	66.67	+61.05	89
Derivatives Investment	74.42	88.00	+13.58	45	Auditing	25.76	93.33	+67.58	66
Economics	76.92	90.00	+13.08	15	Economic Law	6.67	91.89	+85.23	60
Equity Investment	85.96	88.46	+2.50	59	Tax Law	7.02	92.86	+85.84	57
Financial Statement Analysis	74.42	88.89	+14.47	43	Wealth Management	6.25	75.00	+68.75	32
Fixed Income Investment	70.21	78.57	+8.36	49					
Other Investments	82.80	90.12	+7.33	94					
Portfolio Management	100.00	100.00	+0.00	4					
Quantitative Analysis	89.36	91.11	+1.75	47					
Overall	80.39	89.10	+8.71	366	Overall	10.53	84.57	+74.05	304

Table 22: Accuracy (%) of GPT-OSS-120B by question topic. Left: English CFA (9 subjects). Right: Chinese CPA (5 subjects). n denotes the number of original questions.

Question Type	English CFA				Chinese CPA			
	Orig.	Rev.	Δ	n	Orig.	Rev.	Δ	n
Conceptual Understanding	85.48	86.05	+0.56	62	8.40	100.00	+91.60	119
Simple Calculation	93.33	87.71	-5.62	30	4.88	69.81	+64.93	41
Complex Calculation	86.67	100.00	+13.33	15	0.00	-	-	2
Comprehensive Judgment	78.22	89.53	+11.32	101	9.20	91.92	+82.72	87
Knowledge Application	63.64	100.00	+36.36	86	28.57	66.67	+38.10	7
Statistical Methods	90.00	93.62	+3.62	70	20.83	80.00	+59.17	48
Overall	80.39	89.10	+8.71	366	10.53	84.57	+74.05	304

Table 23: Accuracy (%) of GPT-OSS-120B by question type across English CFA and Chinese CPA examinations.

The *Reasoning Accuracy* column disagrees. On English it rises modestly from 12.26% to 20.71%, still well below Fin-R1-7B’s 42.50% on the same subset. On Chinese it decreases from 4.00% to 1.14%.

The answer–reasoning gap therefore widens with scale on Chinese, from 63.4 to 83.4 percentage points, and narrows only slightly on English, from 74.4 to 68.4 percentage points.

Confidence amplifies the concern. Mean confidence rises from 90.66 (20B) to 95.66 (120B) on English, and from 79.17 to 91.46 on Chinese. Both 120B values are the highest observed in our entire evaluation. The largest and most confident general-purpose reasoning model in our suite simultaneously shows the widest answer–reasoning gap. This is the central empirical argument against treating answer accuracy alone as a proxy for reliability in professional financial reasoning.

C.10 DianJin-R1-32B Configuration

Model Source. DianJin/DianJin-R1-32B¹⁷ (Zhu et al., 2025)

DianJin-R1-32B is a financial reasoning model developed by Alibaba Cloud, built on the Qwen2.5-32B base and trained with GRPO-style reinforcement learning on large-scale financial question-answering data. It is included in our evaluation as a larger and more recent finance-specialised reasoning model than Fin-R1-7B (Appendix C.5). Its inclusion tests whether the epistemic-caution gains observed at 7B scale under the same GRPO paradigm persist at 32B scale. We use a single NVIDIA H20 (96 GB HBM3) with BF16 precision. At this precision the model occupies approximately 64 GB and fits on a single GPU without tensor parallelism.

Chat Template. DianJin-R1-32B inherits the native Qwen-style chat template, applied through `tokenizer.apply_chat_template()` with `trust_remote_code=True`. We use the standard system/user message structure of our unified protocol (Appendix C.3):

Detailed Results. Table 24 reflects the profile of a high-capacity finance-specialised reasoning model. English CFA *Orig.* accuracy reaches 74.79% and *Rev.* reaches 89.26% (+14.47%), placing DianJin-R1-32B among the strongest open-weight finance models in our suite. Chinese CPA shows a much sharper contrast. *Orig.* accuracy of 16.87% is an order of magnitude above CFGPT2-7B’s 2.96% on the same benchmark. *Rev.* accuracy rises to 93.79% (+76.93%), the highest Chinese *Rev.* value in our entire evaluation. The size of the Chinese jump is consistent with the implicit assumption alignment pattern described in Section 4: a model trained on Chinese financial data aligns

```
tokenizer = AutoTokenizer.from_pretrained(
    model_dir, trust_remote_code=True)
model = AutoModelForCausalLM.from_pretrained(
    model_dir,
    device_map="auto",
    torch_dtype="auto",
    trust_remote_code=True)
model.eval()

messages = [
    {"role": "system", "content": system_prompt},
    {"role": "user", "content": user_prompt},
]
prompt = tokenizer.apply_chat_template(
    messages, tokenize=False,
    add_generation_prompt=True)

inputs = tokenizer(prompt, return_tensors="pt").
to(device)
out = model.generate(
    **inputs,
    max_new_tokens=1024,
    do_sample=False,
    repetition_penalty=1.05,
    eos_token_id=tokenizer.eos_token_id,
    pad_token_id=tokenizer.eos_token_id)
```

Figure 11: DianJin-R1-32B inference code (Qwen-style chat template, greedy decoding consistent with the rest of the suite).

its regulatory and accounting defaults with the removed premises almost perfectly.

The Chinese *Orig.* profile is non-trivial in its own right. *Auditing* reaches 42.11% and *Accounting* 11.29%, reflecting Chinese-financial knowledge acquired during continued pre-training. Under the *Rev.* setting, the three rule-prescriptive topics (*Economic Law*, *Tax Law*, *Auditing*) all exceed 96%. The less rule-prescriptive *Wealth Management* reaches 86.96%, higher than for every other finance-specialised model in our evaluation. *Rev.* peaks at 97.37% on *Tax Law*, just short of the ceiling.

Robustness Analysis under Adversarial Masking. DianJin-R1-32B exhibits the most pronounced English-Chinese asymmetry among all evaluated models. On the English NOTA subset *Answer Accuracy* reaches 88.28% and *Reasoning Accuracy* reaches 29.16%, a 59.1-point gap. This gap is substantially narrower than GPT-OSS-120B’s 68.4-point gap on the same subset and is broadly consistent with the behaviour of the smaller Fin-R1-7B (Appendix C.5). GRPO-style reasoning RL, when transferred to a larger 32B Qwen2.5 base, therefore preserves a non-trivial share of English condition-detection capability.

¹⁷<https://huggingface.co/DianJin/DianJin-R1-32B>

Question Topic (CFA)	Orig.	Rev.	Δ	n	Question Topic (CPA)	Orig.	Rev.	Δ	n
Corporate Finance	80.00	90.00	+10.00	10	Accounting	11.29	85.71	+74.42	89
Derivatives Investment	63.64	96.00	+32.36	45	Auditing	42.11	96.55	+54.45	66
Economics	75.00	90.00	+15.00	15	Economic Law	9.43	97.06	+87.62	60
Equity Investment	76.27	88.46	+12.19	59	Tax Law	9.80	97.37	+87.56	57
Financial Statement Analysis	76.74	88.89	+12.14	43	Wealth Management	3.85	86.96	+83.11	32
Fixed Income Investment	51.11	78.57	+27.46	49					
Other Investments	80.65	88.89	+8.24	94					
Portfolio Management	100.00	100.00	+0.00	4					
Quantitative Analysis	89.36	88.89	-0.47	47					
Overall	74.79	89.26	+14.47	366	Overall	16.87	93.79	+76.93	304

Table 24: Accuracy (%) of DianJin-R1-32B by question topic. Left: English CFA (9 subjects). Right: Chinese CPA (5 subjects). n denotes the number of original questions.

Question Type	English CFA				Chinese CPA			
	Orig.	Rev.	Δ	n	Orig.	Rev.	Δ	n
Conceptual Understanding	80.65	88.10	+7.45	62	16.22	100.00	+83.78	119
Simple Calculation	71.43	88.64	+17.21	30	9.09	94.29	+85.19	41
Complex Calculation	64.29	100.00	+35.71	15	0.00	-	-	2
Comprehensive Judgment	84.16	89.53	+5.38	101	16.90	92.31	+75.41	87
Knowledge Application	47.50	87.50	+40.00	86	0.00	100.00	+100.00	7
Statistical Methods	90.00	91.49	+1.49	70	25.64	100.00	+74.36	48
Overall	74.79	89.26	+14.47	364	16.87	93.79	+76.93	304

Table 25: Accuracy (%) of DianJin-R1-32B by question type across English CFA and Chinese CPA examinations.

The Chinese picture is sharply different. *Answer Accuracy* remains high at 77.71%, in line with the 93.79% Chinese *Rev.* accuracy reported above. *Reasoning Accuracy* drops to 5.71%, a 72.0-point gap. Qualitative inspection shows that Chinese explanations rarely articulate that a required premise is missing. The model commits to a specific regulatory or accounting default and returns the option consistent with that default. The 93.79% Chinese *Rev.* value therefore reflects successful *alignment* of training-time defaults with the benchmark’s removed premises, not recognition of under-specification.

Confidence is moderate by the standards of the extended suite. Mean confidence is 93.71 on English and 77.80 on Chinese, both below GPT-OSS-120B (95.66 and 91.46) and broadly comparable to Fin-R1-7B. DianJin-R1-32B thus exhibits the signature GRPO-trained profile: high answer accuracy, measurable English reasoning consistency relative to non-RL baselines, and moderate confidence. It does not escape the core Chinese-CPA failure mode that persists across every model family in our evaluation.