Beyond Getting the Right Answer: A Knowledge-Grounded Benchmark for Evaluating Financial Reasoning

Anonymous ACL submission

Abstract

Existing financial LLM benchmarks rely exclusively on outcome-level metrics, conflating correct answers with genuine understanding. This is particularly problematic in finance, where flawed reasoning that coincidentally yields correct answers poses substantial deployment risks, and where static question banks are increasingly susceptible to data contamination. We introduce **FinDepth**, a Chinese financial benchmark with 13 subjects, 6,740 questions, four types, and three capability layers, featuring: (1) FinProbe, a knowledge graph-driven CoT evaluation framework measuring Knowledge F1 and structural alignment; (2) a dynamic question generation mechanism using parameterized numerical generation and answer-preserving text augmentation. Experiments on 12 LLMs reveal a clear accuracy–reasoning dissociation: DeepSeek-R1 achieves a CoT score of 22.3% yet only 61.4% answer accuracy, while InternLM3-8B reaches 64.7% accuracy despite the lowest CoT score of 12.2%. Models also exhibit an average 3.1% accuracy drop on dynamically generated questions, confirming contamination risks in static benchmarks. These findings establish that outcome-level metrics alone are insufficient for assessing financial reasoning competence.

1 Introduction

The financial domain is not merely knowledge-intensive; it is simultaneously rule-bound, numerically constrained, and consequential—where a superficially correct answer can mask fundamentally flawed reasoning. A model may arrive at the correct tax liability through offsetting errors in rate selection and deduction eligibility, or produce a valid net present value from incompatible cash-flow assumptions. In real financial workflows, such failures correspond to misapplied regulation or distorted valuation, with consequences far more serious than a mere explanation error. As LLMs are increasingly deployed in high-stakes financial applications such as automated advisory and regulatory compliance analysis (Xie et al.; Zhang et al., 2025), this disconnect between *getting the right answer* and *reasoning correctly* poses a critical but largely unexamined risk.

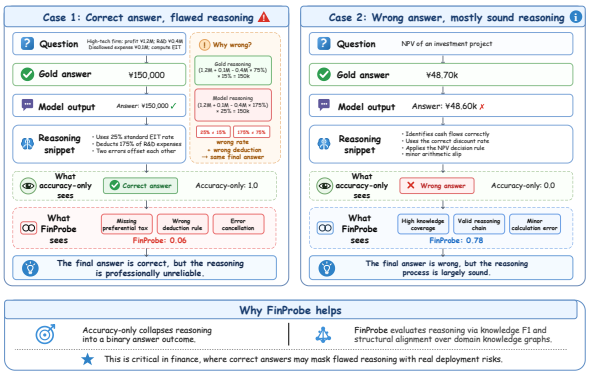

Figure 1: Two financial reasoning cases contrasting accuracy-only evaluation with FinProbe. Left: correct answer from flawed reasoning. Right: wrong answer despite sound reasoning.

Existing financial benchmarks—FinEval (Guo et al., 2025b), CFLUE (Zhu et al., 2024a), CFinBench (Nie et al., 2025), FinBen (Xie et al.), XFinBench (Zhang et al., 2025), among others—have made valuable contributions by assembling large-scale question banks covering diverse financial subjects. Yet they share a common blind spot: evaluation is *exclusively outcome-level*, measuring only whether the final answer is correct via accuracy, F1, or ROUGE scores. Under this paradigm, a model that selects the correct option through surface-level keyword matching scores identically to one that arrives at the same answer through rigorous domain reasoning. In finance, where flawed reasoning that coincidentally yields correct answers can mask misapplied regulations or distorted valuations, this blind spot is especially dangerous—it creates a false sense of reliability (Figure 1). This problem

is further exacerbated by the *static* nature of existing benchmarks. As LLM pre-training corpora grow exponentially, static test items are increasingly likely to leak into training data (Deng et al., 2024; Golchin and Surdeanu, 2024), inflating accuracy scores through memorization rather than genuine understanding. When answer-level metrics are already unreliable indicators of model competence, and data contamination further erodes their trustworthiness, evaluating the *reasoning process* itself becomes essential for obtaining a faithful picture of financial LLM capabilities.

However, existing reasoning evaluation methods are developed independently of financial domain knowledge and are insufficient for this purpose. ROSCOE (Golovneva et al., 2023) scores step-level semantic coherence via pretrained embeddings, RECEval (Prasad et al., 2023) assesses correctness and informativeness through NLI models, PRM (Lightman et al., 2024) trains step-level verifiers with binary correct/incorrect labels, and ReasonEval (Xia et al., 2025) measures validity and redundancy of individual steps. These methods evaluate general logical coherence or textual consistency without understanding the structured knowledge dependencies specific to finance: whether a tax-rate application presupposes correct taxpayer-category identification, or whether an audit conclusion demands a complete evidence chain from risk assessment to substantive procedures. Financial reasoning quality hinges not merely on whether each step is locally valid, but on whether the reasoning *invokes the correct domain knowledge points* and *traverses valid paths within the disciplinary knowledge structure*—dimensions that no existing method captures.

To address both limitations, we introduce FINDEPTH, a Chinese financial benchmark that evaluates not only *what* a model answers but *how* it reasons. Built on 13 core financial subjects with 6,740 questions across four types (multiple-choice, true/false, open-ended QA, and numerical computation) organized into three capability layers, FINDEPTH introduces two key innovations: The first is **FinProbe**, a knowledge graph-driven CoT evaluation framework. We construct subject-specific knowledge graphs for each of the 13 financial subjects and assess reasoning quality along two complementary dimensions: *knowledge coverage and precision* (Knowledge F1), which measures whether the reasoning invokes the essential domain concepts without introducing irrelevant ones, and *structural alignment*, which evaluates whether reasoning steps follow valid paths within the knowledge graph. This grounds CoT evaluation in structured domain knowledge rather than relying on surface-level text matching or generic logical coherence. The second is a **dynamic question generation** mechanism designed to resist data contamination. For numerical questions, parameterized generation functions produce fresh instances by randomizing values within domain-valid ranges; for text-based questions, answer-preserving semantic augmentation rewrites question stems while keeping correct answers and knowledge points fixed. Together, these two innovations enable FINDEPTH to provide both a faithful assessment of reasoning quality and a contamination-resistant evaluation protocol.

Experiments on 12 LLMs reveal a striking *accuracy–reasoning dissociation*: DeepSeek-R1 achieves the second-highest CoT score of 22.3% among open models yet attains only 61.4% answer accuracy, while InternLM3-8B reaches 64.7% accuracy despite the lowest CoT score of 12.2%. Models also show an average accuracy drop of 3.1% on dynamically generated questions, confirming contamination risks in static benchmarks. These findings confirm that outcome-level metrics alone are insufficient, and potentially misleading, for assessing financial reasoning competence.

In summary, our contributions are:

1. We propose **FinProbe**, a knowledge graph-driven CoT evaluation framework that assesses financial reasoning quality through Knowledge F1 and structural alignment against domain-specific knowledge graphs.
2. We design a **dynamic question generation** mechanism with parameterized functions and domain-specific plausibility constraints, producing fresh instances to counter data contamination.
3. We construct **FINDEPTH**, a comprehensive Chinese financial benchmark covering 13 subjects, 6,740 questions, 4 question types, 3 capability layers, and subject-specific knowledge graphs.
4. Through experiments on 12 LLMs, we demonstrate that accuracy and reasoning quality can diverge substantially, establishing the necessity of process-level evaluation in finance.

2 Related Work

Financial LLM Benchmarks. A growing body of work has established benchmarks for evaluating LLMs in the financial domain. For Chinese,

FinEval (Guo et al., 2025b) and CFLUE (Zhu et al., 2024a) pioneer financial knowledge evaluation with multiple-choice questions, CFinBench (Nie et al., 2025) scales to 99,100 questions across 43 categories, and FLAME (Guo et al., 2025a) combines certification exams with business scenarios. For English, FinBen (Xie et al.) provides panoramic evaluation across 36 datasets, XFinBench (Zhang et al., 2025) targets complex financial reasoning, and FinanceQA (Mateega et al., 2025) focuses on analyst workflows. Specialized efforts such as OmniEval (Wang et al., 2025) and FinMTEB (Tang and Yang, 2025) further address financial RAG and embedding evaluation. Despite this progress, all existing benchmarks share a common limitation: they exclusively employ outcome-level metrics and static question banks, leaving reasoning quality unexamined and contamination risks unaddressed.

Reasoning Evaluation and Data Contamination. CoT prompting (Wei et al., 2022) elicits intermediate reasoning steps, yet evaluating their quality remains challenging. ROSCOE (Golovneva et al., 2023) introduces reference-based semantic metrics for step-level assessment, PRM (Lightman et al., 2024) trains verifiers for step-level correctness, ReCEval (Prasad et al., 2023) evaluates reasoning chains via correctness and informativeness, ReasonEval (Xia et al., 2025) measures validity and redundancy of reasoning steps, and LLM-as-Judge (Zheng et al., 2023) leverages strong models as evaluators. Knowledge graphs have been used for grounding in NLP (Liang et al., 2024), but not for CoT quality assessment in domain-specific benchmarks. Orthogonally, data contamination has emerged as a critical concern (Deng et al., 2024; Golchin and Surdeanu, 2024), with mitigation approaches including temporal holdouts (White et al., 2025) and dynamic generation such as DyVal (Zhu et al., 2024b). Our work bridges both gaps: leveraging financial knowledge graphs for structured reasoning assessment, and extending dynamic evaluation to the financial domain through parameterized computation problems.

3 The FINDEPTH Benchmark

We present the design and construction of FINDEPTH, structured around three pillars: a comprehensive financial dataset (§3.1), domain knowledge graphs (§3.3), and a multi-dimensional evaluation system (§3.4). FINDEPTH is defined as a triple $\mathcal{B} = (\mathcal{D}_{\text{dyn}}, \mathcal{G}, \mathcal{F})$, where $\mathcal{D}_{\text{dyn}}$ is the evaluation dataset dynamically generated from an expert-curated static question bank $\mathcal{D}_{\text{sta}}$ (§3.1–§3.2), $\mathcal{G} = \{\mathcal{G}_s\}_{s \in \mathcal{S}}$ is a collection of subject-level knowledge graphs encoding structural dependencies among financial concepts (§3.3), and $\mathcal{F}$ is a multi-dimensional evaluation system that jointly assesses answer correctness and reasoning quality (§3.4). Figure 2 illustrates the overall pipeline and Table 1 compares FINDEPTH with existing benchmarks.

3.1 Static Question Bank

The static question bank $\mathcal{D}_{\text{sta}}$ covers 13 core financial subjects organized into three capability layers: *Foundational Knowledge* (definitions, institutional rules, basic formulas), *Disciplinary Application* (single-subject calculation and judgment), and *Professional Reasoning* (multi-step, cross-subject scenario analysis). Questions are curated from professional certification examinations, university-level textbook exercises, and expert-authored scenario-based problems. See Appendix A.2 for detailed layer definitions and examples.

Each instance is represented as a tuple $d = (s, l, t, \mathcal{K}, q, a)$, where $s \in \mathcal{S}$ denotes the subject, $l \in \{\text{Foundational}, \text{Disciplinary}, \text{Professional}\}$ the capability layer, $t \in \{\text{MC}, \text{TF}, \text{QA}, \text{Num}\}$ the question type, $\mathcal{K} \subseteq \mathcal{V}_s$ the set of associated knowledge points linked to the subject knowledge graph, q the question text, and a the reference answer. The four question types target complementary competencies: multiple-choice and true/false questions assess factual recall and conceptual discrimination, open-ended QA requires multi-step analysis of financial principles, and numerical computation evaluates precise formula application under domain constraints. Table 2 presents the distribution across subjects and types.

3.2 Dynamic Question Generation

To mitigate data contamination, we transform the static bank $\mathcal{D}_{\text{sta}}$ into a dynamically generated evaluation set $\mathcal{D}_{\text{dyn}}$. For each original instance $d = (s, l, t, \mathcal{K}, q, a) \in \mathcal{D}_{\text{sta}}$, we produce a new instance $d' = (s, l, t, \mathcal{K}, q', a')$ where subject, capability layer, question type, and knowledge points remain invariant, while the question text and answer are regenerated.

Numerical questions. For computational problems, we design a parameterized generation

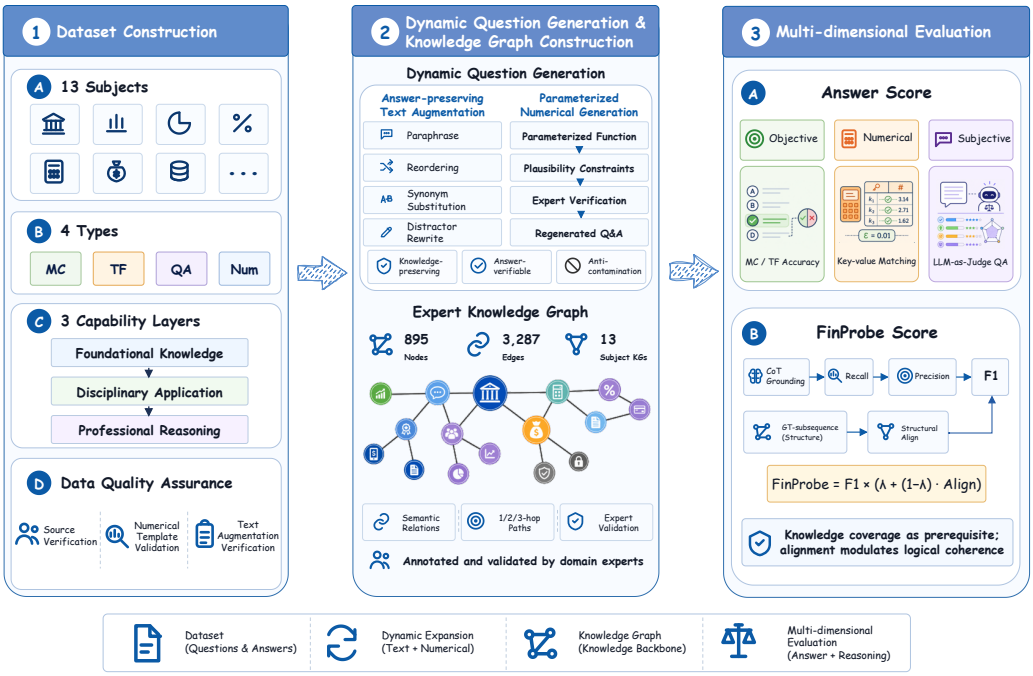

Figure 2: Overview of FINDEPTH, consisting of three pillars: ❶ dataset construction (13 subjects, 4 question types, 3 capability layers), ❷ dynamic question generation and expert knowledge graph construction, and ❸ multi-dimensional evaluation combining answer scoring with FinProbe reasoning assessment.

pipeline that converts each original question into a reusable generation function. The pipeline, summarized in table 3, operates in four stages: an LLM first produces a candidate generation function encoding the solution logic; automated testing validates computational correctness across 100 random seeds; domain experts verify solving logic and plausibility constraints; and functions failing validation are iteratively revised.

Formally, each validated function g_q accepts a parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ drawn from predefined distributions and produces:

$$g_q(\boldsymbol{\theta}) = \big(\text{question}(\boldsymbol{\theta}),\ \text{answer}(\boldsymbol{\theta})\big) \qquad (1)$$

The answer is deterministic given $\boldsymbol{\theta}$, enabling fully automated and exact evaluation. Financial dynamic generation must satisfy *professional plausibility constraints* across four categories: regulatory, accounting consistency, economic and valuation, and answer-feasibility (detailed in Appendix A.5). We maintain 484 validated generation templates covering 10 subjects. The complete pipeline is described in Appendix A.4.

Text-based questions. For MC, TF, and QA questions, we employ *answer-preserving semantic augmentation*. Using the original question, reference answer, and knowledge-point annotations as anchors, we apply controlled perturbations—syntactic reordering, sentence restructuring, synonym substitution, and terminology-preserving paraphrasing—to generate semantically equivalent variants while keeping correct answers and knowledge-point labels fixed. For MC questions, distractors are additionally modified through rewriting or substitution while the correct option remains unchanged.

Quality assurance. Three domain experts independently review all static questions (inter-annotator agreement: 94.3%). For numerical templates, experts verify solving logic and constraints for all generated functions; for text augmentations, they confirm answer preservation and absence of ambiguity. Questions failing any stage are discarded and regenerated (details in Appendix A.3).

3.3 Knowledge Graph Construction

Each subject-level knowledge graph is defined as $\mathcal{G}_s = (\mathcal{V}_s, \mathcal{E}_s)$, where nodes $v \in \mathcal{V}_s$ represent fine-grained knowledge points and edges $e = (v_i, v_j) \in \mathcal{E}_s$ capture directed dependencies between them. The knowledge-point set $\mathcal{K} \subseteq \mathcal{V}_s$ associated with each question instance corresponds to a subset of nodes in the subject graph, grounding each question within the knowledge structure.

Construction follows a three-stage pipeline.

Benchmark	Venue	Lang	#Samples	#Subjects	MC	TF	QA	Num	Dyn.Gen.	CoT Eval
FinEval	NAACL'25	ZH	8,351	34	✓	✗	✗	✗	✗	✗
CFLUE	ACL'24	ZH	38,000+	15	✓	✗	✗	✗	✗	✗
CFinBench	NAACL'25	ZH	99,100	43	✓	✓	✗	✗	✗	✗
FinBen	NeurIPS'24	EN	–	24	✗	✗	✓	✗	✗	✗
XFinBench	ACL'25	EN	4,235	5	✓	✓	✗	✓	✗	✗
OmniEval	EMNLP'25	ZH	13,100	16	✗	✗	✓	✗	✗	✗
FLaME	ACL'25	EN	–	20	✗	✗	✓	✗	✗	✗
FINDEPTH	–	**ZH**	**6,740**	**13**	✓	✓	✓	✓	✓	✓

Table 1: Comparison of FINDEPTH with existing financial LLM benchmarks. **#Samples** = number of evaluation instances; **#Subjects** = number of subject areas or task categories; **MC** = multiple-choice; **TF** = true/false; **QA** = open-ended question answering; **Num** = numerical computation; **Dyn.Gen.** = dynamic question generation; **CoT Eval** = chain-of-thought quality evaluation.

Subject	#Q	MC	TF	QA	Num
Internal Control	1,611	604	180	827	–
Investment	1,446	247	58	926	215
Tax Law	812	245	280	212	75
Microeconomics	572	127	–	429	16
Auditing	467	–	–	451	16
Cost & Mgmt Acct.	464	–	–	421	43
Economic Law	436	200	–	236	–
Fin. Stmt. Analysis	252	25	4	223	–
Macroeconomics	201	14	–	181	6
Adv. Fin. Acct.	179	–	–	149	30
Accounting	100	7	–	72	21
Interm. Fin. Acct.	100	2	–	60	38
Financial Mgmt.	100	–	4	72	24
Total	**6,740**	**1,471**	**526**	**4,259**	**484**

Table 2: Dataset statistics across 13 financial subjects. MC = multiple-choice, TF = true/false, QA = open-ended, Num = numerical computation (dynamic). "–" indicates zero.

Algorithm 1: Numerical Question Generation

Input: Static instance $d = (s, l, t, \mathcal{K}, q, a)$ with $t = \text{Num}$
Output: Validated generation function g_q

Stage 1: Function Generation
Prompt LLM with $(q, a, \mathcal{K})$ to produce $g_q(\boldsymbol{\theta})$
g_q encodes solution logic + plausibility constraints

Stage 2: Automated Testing
for seed $= 1, \ldots, 100$ **do**
Sample $\boldsymbol{\theta} \sim P(\boldsymbol{\theta})$; verify constraints satisfied
Check: answers finite, non-degenerate, consistent

Stage 3: Expert Verification
Verify solving logic matches $\mathcal{K}$
Verify constraints: regulatory, accounting, economic/valuation, answer-feasibility

Stage 4: Iterative Refinement
if failed $\rightarrow$ return to Stage 1 with feedback

At evaluation time:
Sample $\boldsymbol{\theta}$; $(q', a') = g_q(\boldsymbol{\theta})$

Table 3: Pipeline for generating parameterized numerical question functions from static templates.

First, domain concepts are extracted from authoritative Chinese financial textbooks aligned with each subject's curriculum, including terminology, formulas, rules, and analytical frameworks. Second, experts annotate edges using two relation types: *prerequisite* (concept A must be mastered before B) and *related* (A and B are topically related without a directional dependency). Edges are annotated independently by three experts and resolved via majority vote (inter-annotator agreement: 89.1%, Cohen's $\kappa = 0.82$). Third, each graph undergoes structural validation to ensure all nodes are reachable and the graph diameter remains within 8 hops. Table 4 summarizes the statistics across all 13 subjects. These knowledge graphs serve a dual role: organizing the question bank by grounding each question to specific knowledge points, and providing the structural reference against which model reasoning chains are evaluated in the FinProbe framework (§3.4).

3.4 Multi-Dimensional Evaluation System

The evaluation system $\mathcal{F}$ operates on two complementary layers: answer evaluation for outcome correctness and FinProbe for reasoning quality.

Answer Evaluation. For objective questions (MC, TF), we extract the model's selected option via rule-based pattern matching with an LLM fallback, scoring binary correctness. For numerical questions, we apply key-value matching with tolerance $\epsilon = 0.01$: $\text{Acc}_{\text{num}} = |\{j : |a_j - \hat{a}_j| < \epsilon\}|/|J|$, where J is the set of answer fields. For open-ended QA, an LLM judge scores responses across seven dimensions (Zheng et al., 2023) including

Subject	#Nodes	#Edges	Avg.Deg.
Microeconomics	76	389	10.2
Macroeconomics	61	181	5.9
Accounting	50	180	7.2
Interm. Fin. Acct.	86	220	5.1
Adv. Fin. Acct.	38	108	5.7
Cost & Mgmt Acct.	55	236	8.6
Fin. Stmt. Analysis	44	121	5.5
Financial Mgmt.	40	64	3.2
Investment	143	634	8.9
Auditing	83	332	8.0
Internal Control	61	326	10.7
Tax Law	88	189	4.3
Economic Law	70	307	8.8
Total / Avg.	**895**	**3,287**	**7.3**

Table 4: Knowledge graph statistics for all 13 financial subjects. Avg.Deg. denotes the average node degree ($2|\mathcal{E}|/|\mathcal{V}|$).

relevance, consistency, faithfulness, and a subject-specific dimension (see Appendix C.3).

FinProbe: Knowledge Graph-Driven CoT Evaluation. The core innovation of FINDEPTH is FinProbe, which assesses reasoning quality by grounding a model's chain-of-thought against the domain knowledge graph along two dimensions: whether the reasoning *invokes the right knowledge* and whether it *connects knowledge in a valid structure*. Given a question q with ground-truth knowledge points $\mathcal{K}_q \subseteq \mathcal{V}_s$ and the model's reasoning chain $\mathcal{C}$, evaluation proceeds as follows.

Step 1: Concept grounding. We employ a two-stage extraction-then-grounding process: an LLM first extracts candidate financial concepts from $\mathcal{C}$ in order of appearance, then grounds each candidate against $\mathcal{V}_s$ based on the full reasoning context. This yields a deduplicated, order-preserving sequence $\hat{\mathcal{K}}_{\mathcal{C}}^{\text{seq}} = (\hat{v}_1, \hat{v}_2, \dots)$ and its set form $\hat{\mathcal{K}}_{\mathcal{C}}$.

Step 2: Knowledge F1. We define Knowledge Recall $= |\hat{\mathcal{K}}_{\mathcal{C}} \cap \mathcal{K}_q|/|\mathcal{K}_q|$, measuring coverage of essential concepts, and Knowledge Precision $= |\hat{\mathcal{K}}_{\mathcal{C}} \cap \mathcal{K}_q|/|\hat{\mathcal{K}}_{\mathcal{C}}|$, penalizing reasoning bloat with irrelevant concepts:

$$F_1 = \frac{2 \cdot \text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}} \tag{2}$$

Step 3: Structural Alignment. Coverage alone is insufficient—a model may invoke correct concepts in an incoherent order. We extract from $\hat{\mathcal{K}}_{\mathcal{C}}^{\text{seq}}$ only those nodes in $\mathcal{K}_q$, preserving their order, and score each consecutive pair by shortest-path distance in $\mathcal{G}_s$: We extract from $\hat{\mathcal{K}}_{\mathcal{C}}^{\text{seq}}$ only those nodes belonging to $\mathcal{K}_q$, preserving the model's reasoning order, and form the set of consecutive pairs $\mathcal{P}_{\text{seq}}$. Each pair is scored by shortest-path distance in $\mathcal{G}_s$:

$$\text{Align}(q, \mathcal{C}) = \frac{1}{|\mathcal{P}_{\text{seq}}|} \sum_{(v_i, v_j) \in \mathcal{P}_{\text{seq}}} \gamma^{d(v_i, v_j) - 1} \mathbb{1}[\text{path}(v_i, v_j) \neq \emptyset] \tag{3}$$

where $\mathcal{P}_{\text{seq}}$ is the set of consecutive ground-truth node pairs as ordered by the model's reasoning chain, $d(v_i, v_j)$ is the undirected shortest-path distance in $\mathcal{G}_s$ (max depth 4), and $\gamma \in (0, 1)$ is a decay factor. where $\mathcal{P}_{\text{gt}}$ is the set of consecutive ground-truth node pairs, $d(v_i, v_j)$ is the shortest-path distance in $\mathcal{G}_s$ (max depth 4), and $\gamma \in (0, 1)$ is a decay factor. Directly connected concepts score $\gamma^0 = 1$; more distant transitions are exponentially penalized. Boundary rules are detailed in Appendix C.2.

Step 4: FinProbe Score. The final score combines coverage and structural coherence multiplicatively:

$$\text{FinProbe}(q, \mathcal{C}) = F_1 \cdot \big(\lambda + (1 - \lambda) \cdot \text{Align}(q, \mathcal{C})\big) \tag{4}$$

where $\lambda \in (0, 1)$ is a floor factor. This formulation ensures F_1 serves as a prerequisite—without adequate knowledge coverage, structural alignment alone cannot yield a high score—while alignment modulates the final score to reward logically coherent reasoning progressions. Hyperparameters γ and λ are selected via grid search maximizing Spearman correlation with expert annotations (Appendix C.1).

4 Experiments

4.1 Experimental Setup

Models. We evaluate 12 representative LLMs spanning two categories: (1) *Closed-source API models*: Claude Opus 4.6, Qwen3.6-Plus, Gemini 3.1 Pro, GPT-5.4 (the OpenAI Team, 2026), and GPT-4o (the OpenAI Team, 2024); (2) *Open-weight models*: DeepSeek-V4-Pro (DeepSeek-AI, 2026), DeepSeek-R1 (DeepSeek-AI, 2025), Qwen3.5-397B-A17B and Qwen3.5-35B-A3B (Team, 2026), Qwen3.6-35B-A3B (Qwen Team, 2026), Qwen3-8B-Instruct (Yang et al., 2025), and InternLM3-8B-Instruct (Cai et al., 2024). This selection covers diverse model scales, architectures, and design philosophies.

Evaluation settings. All models are evaluated in a zero-shot setting with standardized prompts

instructing chain-of-thought reasoning. For numerical questions, we generate one dynamic instance per template using a fixed random seed for reproducibility. All evaluation components (FinProbe concept grounding, subjective QA scoring, and objective question fallback) uniformly use `gpt-4-0409-global` as the judge model. All experiments are conducted on a unified evaluation pipeline with identical prompts and post-processing logic to ensure fair comparison.

4.2 Main Results

Table 5 presents the comprehensive evaluation results.

Overall performance. Closed-source models generally outperform open-weight models on both answer accuracy and reasoning quality. Among closed-source models, Claude Opus 4.6 achieves the highest Avg Ans. of 82.4% and CoT score of 24.2%, followed by Qwen3.6-Plus at 79.3% and 24.0% respectively. In the open-weight category, DeepSeek-V4-Pro rivals top closed-source systems with 80.3% Avg Ans. and 23.0% CoT, while the best numerical accuracy across all models is achieved by GPT-5.4 at 61.4%. Performance varies substantially across question types: TF questions are the easiest with the top score reaching 90.4%, followed by MC at 82.1%, while numerical computation remains the most challenging with even the best model achieving only 61.4%.

Accuracy–reasoning dissociation. The most significant finding from our evaluation is that answer accuracy and reasoning quality do not always align, with notable individual deviations that challenge the assumption that correct answers imply sound reasoning. DeepSeek-R1 exemplifies this phenomenon most clearly: it ranks second among open models in both CoT score at 22.3% and knowledge F1 at 24.7%, demonstrating thorough engagement with domain concepts, yet achieves only 61.4% Avg Ans. with notably weak numerical accuracy of 26.3%. This shows that strong reasoning does not guarantee correct answers when precise computation is additionally required. In the opposite direction, InternLM3-8B attains 64.7% Avg Ans. and 77.8% TF accuracy while recording the lowest CoT score of 12.2% across all evaluated models, suggesting that its accuracy derives primarily from surface-level pattern matching rather than grounded domain reasoning. The Spearman correlation between Avg Ans. and CoT across all 12 models is $\rho = 0.72$, indicating moderate positive alignment overall; however, the notable individual deviations confirm that answer accuracy and reasoning quality capture partially independent competency dimensions, validating the necessity of FinProbe as a complementary evaluation axis.

4.3 Analysis

Validation of FinProbe. We validate FinProbe against human expert judgment by sampling 200 questions from $\mathcal{D}_{\text{dyn}}$ across 8 subjects and collecting CoT outputs from Claude Opus 4.6, DeepSeek-R1, and Qwen3-8B-Instruct, yielding 600 instances. Three financial domain experts independently rate each CoT on four dimensions using a 1–5 Likert scale (Appendix C.4), with inter-annotator agreement of Krippendorff's $\alpha = 0.72$. We compare FinProbe against five alternative methods: ReasonEval (Xia et al., 2025), LLM-Judge, ReCEval (Prasad et al., 2023), PRM-style step-level judgment, and ROSCOE (Golovneva et al., 2023) (Appendix C.5). As shown in Table 6, FinProbe achieves the highest Spearman ρ of 0.71 on Knowledge Completeness and 0.68 on Overall Quality, outperforming the second-best baseline by 0.14 and 0.13 respectively. LLM-Judge leads on Knowledge Correctness at 0.61, which is expected as correctness verification relies on the evaluator's internal knowledge rather than structural grounding.

Performance by Capability Layer. Table 7 reveals an inverse relationship between task difficulty and the two evaluation dimensions for Claude Opus 4.6. Answer accuracy drops from 84.1% at Foundational Knowledge to 71.7% at Disciplinary Application and 37.8% at Professional Reasoning. FinProbe scores exhibit the opposite trend: CoT scores *increase* from 20.7% to 29.5% and 35.9% across the same progression. Harder tasks requiring multi-step reasoning and cross-disciplinary integration elicit more thorough reasoning chains, even when the final answers are incorrect. This further underscores the limitation of accuracy-only evaluation: a model scoring poorly on professional-level questions may nonetheless demonstrate strong domain reasoning.

Static vs. Dynamic Evaluation. A central motivation for dynamic question generation is to mitigate data contamination—the risk that models have memorized answers from publicly available question banks during pre-training. To quantify this effect, we sample 500 questions from $\mathcal{D}_{\text{sta}}$ to form $\mathcal{D}'_{\text{sta}}$, and generate a matched set $\mathcal{D}'_{\text{dyn}}$ of 500 questions via our dynamic generation mechanism, cov-

Category	Model	Size	COT		FinProbe	Answer				Avg Ans.
			F1	Align		MC	TF	Num	Sub	
Closed	Claude Opus 4.6	–	**27.5**	**76.3**	**24.2**	**82.1**	90.0	56.6	**85.5**	**82.4**
	Qwen3.6-Plus	–	26.2	74.8	24.0	82.0	**90.4**	34.9	81.8	79.3
	Gemini 3.1 Pro	–	24.2	75.4	21.2	79.5	87.2	51.3	74.6	76.2
	GPT-5.4	–	24.8	71.6	21.3	74.4	84.0	**61.4**	73.1	73.9
	GPT-4o	–	22.2	67.1	19.9	58.9	74.6	54.2	75.1	66.8
Open	DeepSeek-V4-Pro	1.6T	**25.1**	**71.8**	**23.0**	**80.4**	**88.4**	50.4	**83.8**	**80.3**
	DeepSeek-R1	671B	24.7	71.7	22.3	65.4	77.8	26.3	59.5	61.4
	Qwen3.5-397B-A17B	397B	20.6	65.7	18.4	76.0	85.3	**55.9**	83.7	78.7
	Qwen3.5-35B-A3B	35B	20.7	64.7	18.3	41.7	71.1	48.4	68.1	55.9
	Qwen3.6-35B-A3B	35B	18.8	69.4	15.9	44.0	53.0	12.4	45.2	43.1
	Qwen3-8B-Instruct	8B	19.2	26.6	13.2	66.8	75.6	40.4	61.4	65.1
	InternLM3-8B	8B	16.9	25.0	12.2	64.6	77.8	20.3	71.1	64.7

Table 5: Main evaluation results on FINDEPTH. **COT**: F1 = knowledge F1 (%), Align = structural alignment (%). **FinProbe** = overall FinProbe score combining F1 and Align (%). **Answer**: MC = multiple-choice (%), TF = true/false (%), Num = numerical (%), Sub = subjective QA (%). **Avg Ans.** = micro-averaged answer score. Best per category in **bold**.

Method	Compl.	Corr.	Coher.	Overall
FinProbe (Ours)	**0.71**	0.58	**0.66**	**0.68**
ReasonEval	0.57	0.51	0.48	0.54
LLM-Judge	0.49	**0.61**	0.62	0.55
ReCEval	0.41	0.53	0.56	0.48
PRM-style	0.38	0.42	0.44	0.41
ROSCOE	0.32	0.34	0.47	0.36

Table 6: Spearman ρ between automated methods and human expert ratings on 600 CoT instances. Best per dimension in **bold**.

Capability Layer	#Q	Acc.(%)	CoT(%)
Foundational Knowledge	4,203	84.1	20.7
Disciplinary Application	2,348	71.7	29.5
Professional Reasoning	189	37.8	35.9

Table 7: Performance of Claude Opus 4.6 by capability layer. Accuracy decreases as capability demands increase, while CoT quality increases—harder tasks elicit more thorough reasoning.

Model	MC (%)			TF (%)			Num (%)			Sub (%)		
	Sta	Dyn	Δ	Sta	Dyn	Δ	Sta	Dyn	Δ	Sta	Dyn	Δ
Claude Opus 4.6	86.4	82.1	-4.3	92.7	90.0	-2.7	58.0	56.6	-1.4	87.2	85.5	-1.7
DeepSeek-V4-Pro	84.8	80.4	-4.4	91.3	88.4	-2.9	53.1	50.4	-2.7	85.6	83.8	-1.8
GPT-4o	65.2	58.9	-6.3	79.8	74.6	-5.2	56.1	54.2	-1.9	78.4	75.1	-3.3
DeepSeek-R1	68.5	65.4	-3.1	80.6	77.8	-2.8	28.6	26.3	-2.3	62.1	59.5	-2.6
Average	76.2	71.7	-4.5	86.1	82.7	-3.4	49.0	46.9	-2.1	78.3	76.0	-2.4

Table 8: Performance on static $\mathcal{D}_{\text{sta}}$ vs. dynamic $\mathcal{D}_{\text{dyn}}$ questions. $\Delta = \text{Dyn} - \text{Sta}$. Larger drops suggest greater reliance on memorization.

ering the same knowledge points and difficulty distribution but with randomized numerical parameters and rephrased stems. Table 8 reports the comparison results. All models exhibit performance drops on $\mathcal{D}'_{\text{dyn}}$, with an average decline of 3.1% across all question types, confirming that static benchmarks overestimate capabilities due to contamination. The effect is type-dependent: MC and TF drop 4.5% and 3.4% on average, while numerical and subjective questions drop only 2.1% and 2.4%—consistent with the fact that MC/TF answers can be memorized verbatim whereas computation and reasoning require genuine understanding. DeepSeek-R1 shows the smallest average drop of 2.7%, while GPT-4o exhibits the largest at 4.2%, suggesting that reasoning-enhanced training yields greater robustness to contamination. These results validate dynamic generation as an effective contamination-resistant evaluation strategy.

5 Conclusion

We presented FINDEPTH, a Chinese financial benchmark that evaluates both answer correctness and reasoning quality. Our FinProbe framework grounds CoT assessment against domain knowledge graphs, while a dynamic generation mechanism mitigates data contamination. Experiments on 12 LLMs reveal a clear accuracy–reasoning dissociation: models can produce correct answers through flawed reasoning, confirming that accuracy alone is an insufficient measure of financial competence. We will open-source FINDEPTH to facilitate future research.

Limitations

Our work has several limitations. First, FINDEPTH focuses exclusively on the Chinese financial domain; its knowledge graph and question bank are grounded in China's regulatory and institutional context, limiting direct generalizability to other jurisdictions. Second, the knowledge graph is constructed semi-automatically with expert curation, which constrains scalability to new subjects—extending coverage requires additional domain expertise and annotation effort. Third, FinProbe relies on an LLM to extract knowledge concepts from CoT outputs, introducing potential extraction errors that propagate to the final scores. Finally, our dynamic generation mechanism currently supports parameterized regeneration for numerical questions and answer-preserving augmentation for text questions, but does not yet generate entirely novel question structures beyond those in the static bank.

References

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, and 81 others. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

DeepSeek-AI. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2024. Investigating data contamination in modern benchmarks for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8706–8719, Mexico City, Mexico. Association for Computational Linguistics.

Shahriar Golchin and Mihai Surdeanu. 2024. Time travel in LLMs: Tracing data contamination in large language models. In *The Twelfth International Conference on Learning Representations*.

Olga Golovneva, Moya Peng Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. ROSCOE: A suite of metrics for scoring step-by-step reasoning. In *The Eleventh International Conference on Learning Representations*.

Jiayu Guo, Yu Guo, Martha Li, and Songtao Tan. 2025a. Flame: Financial large-language model assessment and metrics evaluation. *Preprint*, arXiv:2501.06211.

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

Ke Liang, Lingyuan Meng, Meng Liu, Yue Liu, Wenxuan Tu, Siwei Wang, Sihang Zhou, Xinwang Liu, Fuchun Sun, and Kunlun He. 2024. A survey of knowledge graph reasoning on graph types: Static, dynamic, and multi-modal. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(12):9456–9478.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Spencer Mateega, Carlos Georgescu, and Danny Tang. 2025. Financeqa: A benchmark for evaluating financial analysis capabilities of large language models. *Preprint*, arXiv:2501.18062.

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

Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. 2023. ReCEval: Evaluating reasoning chains via correctness and informativeness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10066–10086, Singapore. Association for Computational Linguistics.

Qwen Team. 2026. Qwen3.6-35B-A3B: Agentic coding power, now open to all.

Yixuan Tang and Yi Yang. 2025. FinMTEB: Finance massive text embedding benchmark. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3620–3638, Suzhou, China. Association for Computational Linguistics.

Qwen Team. 2026. Qwen3.5: Accelerating productivity with native multimodal agents.

the OpenAI Team. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

the OpenAI Team. 2026. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Shuting Wang, Jiejun Tan, Zhicheng Dou, and Ji-Rong Wen. 2025. OmniEval: An omnidirectional and automatic RAG evaluation benchmark in financial domain. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5726–5751, Suzhou, China. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Venkat Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. 2025. Livebench: A challenging, contamination-limited LLM benchmark. In *The Thirteenth International Conference on Learning Representations*.

Shijie Xia, Xuefeng Li, Yixin Liu, Tongshuang Wu, and Pengfei Liu. 2025. Evaluating mathematical reasoning beyond accuracy. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(26):27723–27730.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, Yijing Xu, Haoqiang Kang, Ziyan Kuang, Chenhan Yuan, Kailai Yang, Zheheng Luo, Tianlin Zhang, Zhiwei Liu, Guojun Xiong, and 15 others.

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

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Jie Zhu, Junhui Li, Yalong Wen, and Lifan Guo. 2024a. Benchmarking large language models on CFLUE - a Chinese financial language understanding evaluation dataset. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5673–5693, Bangkok, Thailand. Association for Computational Linguistics.

Kaijie Zhu, Jiaao Chen, Jindong Wang, Neil Zhenqiang Gong, Diyi Yang, and Xing Xie. 2024b. Dyval: Dynamic evaluation of large language models for reasoning tasks. In *The Twelfth International Conference on Learning Representations*.

A Dataset Details

A.1 Subject Descriptions

Table 9 provides detailed descriptions for all 13 financial subjects covered by FINDEPTH.

A.2 Capability-Layer Definitions

Table 10 provides the detailed definition of each capability layer with typical reasoning demands and examples.

A.3 Data Quality Assurance Procedures

We implement a comprehensive multi-stage quality assurance pipeline:

Source data verification. All original questions undergo expert review by graduate students or professionals with relevant financial backgrounds. Reviewers verify: (1) the correctness of reference answers, (2) the appropriateness of knowledge-point annotations relative to the subject knowledge graph, and (3) the accuracy of capability-layer assignments. We randomly sample 10% of questions for independent verification by a second annotator, achieving an inter-annotator agreement rate of 94.3%.

Dynamic numerical template validation. Each numerical generation template is validated through both automated testing and expert audit:

- *Automated testing*: Each template is executed across 100 random seeds to verify computational correctness and constraint satisfaction. We check that all generated answers are finite, non-negative where required, and consistent with the analytical solution.

Subject	Category	Description
Microeconomics	Economics	Explains how individuals, firms, and markets make decisions under scarcity, providing the basis for pricing, competition, and resource allocation.
Macroeconomics	Economics	Places financial decisions in the broader economic environment shaped by policy, inflation, growth, and external shocks.
Accounting	Accounting	Translates business transactions into standardized financial information for recording, reporting, and interpretation.
Interm. Fin. Acct.	Accounting	Trains the application of accounting standards to core statement elements and routine reporting judgments.
Adv. Fin. Acct.	Accounting	Addresses complex reporting settings involving consolidation, ownership structures, and special accounting treatments.
Cost & Mgmt Acct.	Accounting	Uses cost information to support planning, control, performance evaluation, and managerial decision-making.
Fin. Stmt. Analysis	Accounting	Infers firm performance, risk, and earnings quality from reported accounting information.
Financial Mgmt.	Finance	Studies how firms make investment, financing, valuation, and payout decisions under risk.
Investment	Finance	Examines how financial assets are priced, selected, and managed under uncertainty.
Auditing	Auditing & Control	Evaluates the credibility of financial information through evidence-based assurance and professional judgment.
Internal Control	Auditing & Control	Studies organizational mechanisms for risk prevention, compliance, operational reliability, and governance.
Tax Law	Law & Taxation	Applies statutory tax rules to business transactions, compliance decisions, and financial consequences.
Economic Law	Law & Taxation	Provides the legal basis for evaluating rights, obligations, liabilities, and regulatory constraints in commerce.

Table 9: Detailed descriptions of the 13 financial subjects in FINDEPTH.

Layer	Core Content	Typical Reasoning Demand	Examples
Foundational Knowledge	Concepts, definitions, institutional rules, basic formulas	Recall or recognition of core knowledge points	Definition of assets; statutory tax rate; CAPM formula
Disciplinary Application	Single-subject calculation, judgment, or analysis	Apply a rule or formula within one disciplinary framework	Compute NPV; classify an audit procedure; determine tax applicability
Professional Reasoning	Multi-knowledge, multi-step, scenario-based reasoning	Integrate assumptions, constraints, and professional judgment	Evaluate valuation under changing cash flows; link accounting to tax adjustment and audit evidence

Table 10: Capability-layer taxonomy with definitions, reasoning demands, and representative examples.

- *Expert audit*: Domain experts verify that generated questions remain faithful to intended knowledge points, that intermediate variables retain clear financial meanings, and that solution paths are invariant to value substitution. For templates involving accounting statements, we verify statement-level consistency and subtotal relations. For valuation templates, we check compatibility of cash-flow timing, discount rates, and growth assumptions. For tax templates, we ensure generated values do not inadvertently change applicable institutional conditions.

This step is necessary because dynamic generation can introduce hidden label noise: a small parameter change may alter the applicable rule or make multiple answers defensible.

Text augmentation verification. For dynamically rewritten text-based questions, we conduct targeted quality checks:

- Verifying that the augmented question preserves the same correct answer as the original.
- Confirming that the core knowledge points being assessed remain unchanged.
- Ensuring that no unintended ambiguity or factual error is introduced during paraphrasing.
- For multiple-choice questions, verifying that distractors remain plausible but clearly incorrect.

Questions failing any criterion are discarded and regenerated. Approximately 12% of initially generated text augmentations are rejected through this process.

A.4 Numerical Question Generation Pipeline

The parameterized generation functions for numerical questions are produced through the following pipeline:

1. **LLM-based function generation**: Given an original numerical question, its reference answer, and associated knowledge-point annotations, we prompt GPT-4 to generate a Python function that encodes the solution logic. The function accepts a parameter vector $\boldsymbol{\theta}$ and returns a new question text and its deterministic answer. The LLM is also instructed to embed domain-specific plausibility constraints within the function to ensure that randomized parameters yield financially meaningful instances.
2. **Automated testing**: Each generated function is executed across 100 random seeds to verify computational correctness, constraint satisfaction, and that all produced answers are finite and non-degenerate.
3. **Expert verification**: Domain experts review each function to verify that (a) the solving logic faithfully reflects the intended knowledge points, (b) the plausibility constraints are reasonable and complete, and (c) the generated question text remains clear and unambiguous. Experts test the function with representative parameter values and compare outputs against manually computed solutions.
4. **Iteration**: Functions failing automated testing or expert review are returned to the LLM for revision with error-specific feedback, repeating until convergence. Approximately 18% of initially generated functions require at least one revision cycle.

We maintain 484 validated generation templates covering 10 subjects through this pipeline.

A.5 Financial Plausibility Constraints

Dynamic numerical question generation enforces four categories of domain-specific constraints to ensure professional validity:

1. **Regulatory and institutional constraints.** Tax rates, preferential-policy conditions, legal thresholds, and reporting assumptions must remain consistent with the intended financial context. For example, the enterprise income tax rate should be 25% (standard) or 15% (high-tech preferential), not arbitrary values.
2. **Accounting consistency constraints.** Generated instances must preserve accounting identities (assets = liabilities + equity), subtotal-to-total relations, and coherence between accrual items and cash-flow items. Statement-level balance checks are applied to all accounting templates.
3. **Economic and valuation constraints.** Parameters such as discount rates, growth rates, leverage ratios, profit margins, and risk premiums are restricted to economically interpretable ranges. Key rules include: terminal growth rate $<$ discount rate (for perpetuity calculations); risk-adjusted discount rate compatible with risk-free rate, β, and market premium (for CAPM-based templates); positive real interest rates in standard scenarios.
4. **Answer-feasibility constraints.** Generated parameters must not produce degenerate cases: division by zero, multiple defensible correct answers, economically meaningless negative quantities, or numerical values that make the intended solution path ambiguous.

These constraints ensure that dynamic generation functions as a *constrained financial simulation* rather than purely random numerical substitution, preserving the same underlying knowledge structure while changing surface values.

B Knowledge Graph Construction Details

The knowledge graph construction follows a three-stage pipeline:

Stage 1: Concept Extraction. We extract domain concepts from authoritative Chinese financial textbooks aligned with each subject's curriculum. Concepts include terminology, formulas, rules, institutional provisions, and analytical frameworks. Each concept is validated for uniqueness and appropriate granularity.

Stage 2: Relation Annotation. Domain experts annotate directed edges between concepts using two relation types: (1) *prerequisite*: concept A must be understood before concept B; (2) *related*: concept A and concept B are topically related without a directional dependency. Edges are annotated independently by three experts and resolved via majority vote.

Stage 3: Quality Control. Each completed graph undergoes structural validation: we verify that all nodes are reachable, that no orphan clusters exist, and that the graph diameter remains within a reasonable range (typically $\leq$8 hops) to ensure

meaningful structural alignment computation. Average inter-annotator agreement on edge annotation is 89.1% (Cohen's $\kappa = 0.82$).

C Evaluation Details

C.1 FinProbe Hyperparameter Selection

The FinProbe scoring formula involves two hyperparameters: the distance-decay factor γ and the floor factor λ. We determine their optimal values through a grid search that maximizes Spearman rank correlation (ρ) between the automated FinProbe scores and expert annotations on a separate calibration set $\mathcal{D}_{\text{cal}}$ of 200 reasoning chains sampled across subjects.

Three financial domain experts independently rated each reasoning chain on a 1–5 scale for overall reasoning quality. We computed the average expert score as the gold standard and evaluated all combinations of $\gamma \in \{0.6, 0.7, 0.8, 0.9\}$ and $\lambda \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$.

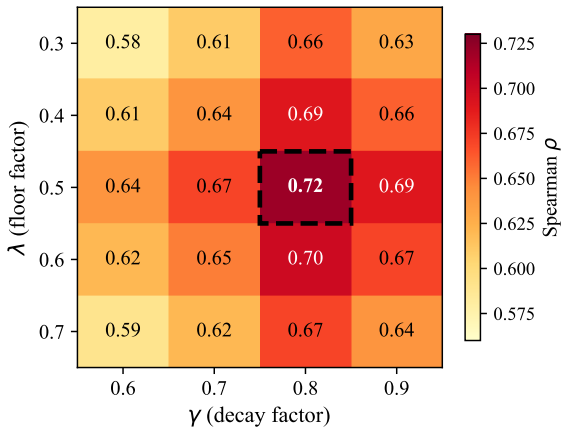

Figure 3: Spearman ρ between FinProbe scores and expert annotations on $\mathcal{D}_{\text{cal}}$. The dashed box marks the optimal setting ($\gamma = 0.8$, $\lambda = 0.5$, $\rho = 0.72$).

The selected configuration ($\gamma = 0.8$, $\lambda = 0.5$) achieves $\rho = 0.72$, indicating strong agreement with expert evaluation. The decay factor $\gamma = 0.8$ appropriately penalizes distant concept transitions while remaining tolerant of indirect but valid reasoning paths. The floor factor $\lambda = 0.5$ balances the contribution of knowledge quality and structural coherence, ensuring that models with strong concept coverage are not overly penalized when structural evaluation is limited.

C.2 FinProbe Boundary Rules

The structural alignment computation applies the following boundary rules for edge cases:

1. When $|\mathcal{K}_q| = 1$ (only one ground-truth knowledge point), structural alignment is undefined and excluded from scoring. The FinProbe score reduces to $\lambda \cdot F_1$.
2. When $|\mathcal{K}_q| \geq 2$ but fewer than 2 GT nodes are successfully grounded in the reasoning chain, Align $= 0$, indicating the model failed to connect essential concepts.
3. Self-loops (consecutive identical nodes in the GT-filtered subsequence) are skipped and do not contribute to the alignment score.
4. The shortest-path search is performed bidirectionally with a maximum depth of 4. If no path exists between two consecutive GT nodes within this depth, the pair contributes 0 to the alignment score.

C.3 Subjective Evaluation Dimensions

Open-ended QA questions are evaluated across six common dimensions plus one or more subject-specific dimensions, as detailed in Figure 4.

C.4 Human Evaluation Rubric

The four dimensions used in the human evaluation study are defined in Figure 5 (1–5 Likert scale).

C.5 Baseline Methods for Human Evaluation

We compare FinProbe against five representative reasoning evaluation methods spanning reference-free metrics, LLM-based judgment, and process supervision.

ReasonEval (Xia et al., 2025). ReasonEval evaluates mathematical reasoning quality beyond final-answer accuracy by assessing each reasoning step along two dimensions: *validity*, whether each step logically follows from previous steps and the problem statement, and *redundancy*, whether a step contributes meaningfully to the solution. It trains dedicated evaluator LLMs on human-annotated step-level labels and aggregates step scores into an overall reasoning quality metric. We adapt its step-level scoring to financial reasoning chains.

LLM-Judge. We prompt a strong LLM (GPT-4) to directly assess the overall quality of each reasoning chain on the same four dimensions used by human annotators: knowledge completeness, correctness, coherence, and overall quality. The LLM receives the question, reference answer, and model-generated CoT, and outputs a 1–5 score per dimension. This serves as a straightforward base-

Subjective QA Evaluation Dimensions

Part A: Common Dimensions (applied to all subjects)

Dimension	Scale	Weight	Description
Relevance	1–5	0.15	Whether the answer directly addresses the question without extraneous content.
Helpfulness	1–5	0.25	Whether the answer provides sufficient depth, data, and analysis to solve the problem.
Fluency	1–5	0.10	Grammatical correctness and naturalness of expression.
Coherence	1–5	0.10	Logical structure, organization, and argumentation consistency.
Consistency	1–5	0.20	Factual and numerical agreement with the reference answer.
Faithfulness	1–5	0.10	Absence of hallucinated facts or unsupported claims.

Part B: Subject-Specific Dimensions (one or more per subject; uniformly scored 1–5, weight = 0.10 each)

Subject	Dimension	Description
Microeconomics	Concept Precision	Accuracy of economic concepts, symbols, model assumptions, and curve relationships.
Macroeconomics	Concept Precision	Accuracy of macroeconomic variables, symbols, and model assumptions.
	Policy Applicability	Depth of policy implication and real-economy connection analysis.
Accounting	Standards Compliance	Accuracy of accounting standards, journal entries, and account titles.
Interm. Fin. Acct.	Accounting Standards	Compliance with standards in measurement and recognition.
	Calculation Accuracy	Correctness of numerical computation and data handling.
Adv. Fin. Acct.	Standards Application	Accuracy in applying standards to complex transactions.
	Step Completeness	Whether all critical processing steps are covered.
Cost & Mgmt Acct.	Cost Computation	Accuracy of cost classification, allocation, and budgeting.
Fin. Stmt. Analysis	Data Interpretation	Accuracy of ratio calculation, trend analysis, and data interpretation.
Financial Mgmt.	Calculation Accuracy	Correctness of formulas, data substitution, and computation.
	Decision Rationality	Whether financial decisions are data-driven and scenario-aware.
Investment	Closed-loop Consistency	Whether model selection, assumptions, derivation, and conclusion form a coherent chain.
Auditing	Standards Compliance	Adherence to auditing standards and completeness of key procedures.
Internal Control	Regulatory Compliance	Conformity with COSO framework and internal control regulations.
Tax Law	Legal Provision Accuracy	Accuracy of tax law provisions, rates, and administration procedures.
Economic Law	Legal Article Accuracy	Accuracy of legal articles, principles, and legal relationships.

Figure 4: Complete evaluation dimensions for subjective QA scoring. The final score is computed as a weighted sum of normalized dimension scores: Score $= \sum_i (s_i/s_i^{\max}) \cdot w_i \,/\, \sum_i w_i$.

line representing the common practice of using LLMs as evaluators without structured grounding.

ReCEval (Prasad et al., 2023). ReCEval (Reasoning Chain Evaluation) evaluates reasoning chains via two complementary properties: *correctness*, whether each reasoning step is factually and logically sound, and *informativeness*, whether each step contributes new and relevant information toward the final answer. It decomposes reasoning chains into atomic steps and scores each step independently using NLI-based models, then aggregates step-level scores into chain-level metrics.

PRM-style (Lightman et al., 2024). Inspired by Process Reward Models (PRMs), this baseline performs step-by-step verification of reasoning chains. We prompt an LLM to label each reasoning step as *correct*, *neutral*, or *incorrect*, following the methodology of outcome-based and process-based reward modeling. The chain-level score is computed as the proportion of correct steps. Unlike FinProbe, this approach does not leverage domain knowledge graphs and relies solely on the evaluator LLM's internal knowledge.

ROSCOE (Golovneva et al., 2023). ROSCOE (Reasoning Over Semantic COhErence) provides a suite of interpretable, unsupervised metrics for scoring step-by-step reasoning. It evaluates reasoning chains across four categories: *semantic alignment* between steps and the source, *semantic similarity* between consecutive steps, *logical inference* quality, and *language coherence*. The metrics are computed using pretrained sentence embeddings and NLI models without requiring reference reasoning chains.

Human Evaluation Rubric (1–5 Likert Scale)

Dim. 1: Knowledge Completeness — *Does the CoT cover all essential domain concepts?*

5 All key knowledge points, definitions, and formulas are explicitly mentioned and correctly applied.
4 Most key concepts present; one minor point omitted without affecting the reasoning.
3 Core concepts covered but several supporting points relevant to a complete answer are missing.
2 Only partially addresses required knowledge; multiple essential concepts absent.
1 Critical domain concepts entirely missing; lacks fundamental knowledge coverage.

Dim. 2: Knowledge Correctness — *Is the invoked domain knowledge factually accurate?*

5 All stated facts, rules, formulas, and regulatory references are correct.
4 One minor inaccuracy that does not materially affect the conclusion.
3 One or two noticeable errors but overall direction remains partially valid.
2 Multiple factual errors that significantly undermine the reasoning validity.
1 Fundamental misconceptions or fabricated knowledge; factually unreliable.

Dim. 3: Logical Coherence — *Do the reasoning steps follow a logically sound progression?*

5 Argument flows naturally with clear causal connections between steps.
4 Generally coherent with one minor logical gap that does not break the argument.
3 Discernible structure but contains noticeable gaps or unjustified jumps.
2 Largely disorganized; key logical connections missing or contradictory.
1 No discernible logical structure; steps disconnected or self-contradictory.

Dim. 4: Overall Quality — *Holistic assessment of completeness, correctness, coherence, and relevance.*

5 Exemplary reasoning that a financial professional would consider thorough and reliable.
4 Good reasoning with minor imperfections that do not affect practical utility.
3 Acceptable reasoning that captures the main idea but lacks rigor in details or structure.
2 Below-average reasoning with significant gaps in knowledge, correctness, or logic.
1 Entirely inadequate; provides no meaningful value for answering the question.

Figure 5: Human evaluation rubric used by three financial domain experts to rate CoT quality on $\mathcal{D}_{\text{human}}$.

D Additional Analysis

D.1 Performance by Question Type

We observe substantial variation in both accuracy and reasoning quality across question types. Across all 12 models, true/false questions yield the highest average accuracy at 79.6%, followed by multiple-choice at 68.0% and numerical computation at 42.7%. The gap between TF and Num exceeds 36 percentage points, indicating that precise multi-step calculation remains the most challenging task for current LLMs. Notably, GPT-5.4 leads all models in numerical accuracy (61.4%) yet ranks only fourth in MC (74.4%), while Qwen3.6-Plus achieves the highest TF accuracy (90.4%) but drops to 34.9% on numerical questions—revealing highly uneven capability profiles across question types.

For FinProbe evaluation, numerical questions produce the richest reasoning chains with average CoT scores nearly 2× higher than multiple-choice and true/false questions. This is because numerical computation inherently requires explicit formula invocation, parameter identification, and multi-step calculation, all of which are captured by our FinProbe framework. Open-ended QA questions achieve intermediate CoT scores, as these require conceptual explanation but not necessarily the precise formula application that numerical questions demand.

D.2 Model Scale and Reasoning Quality

Comparing models within the Qwen family provides insight into the relationship between scale and evaluation dimensions. Qwen3.5-397B-A17B (17B active) achieves 78.7% Avg Ans. with 18.4% CoT (F1 = 20.6%), while the smaller Qwen3.5-35B-A3B (3B active) achieves only 55.9% Avg Ans. yet attains nearly identical CoT of 18.3% (F1 = 20.7%). This suggests that scaling primarily benefits answer accuracy through broader knowledge memorization, while reasoning quality—as measured by Knowledge F1 and structural alignment—does not improve proportionally with scale. The newer Qwen3.6-35B-A3B further illustrates this: despite architectural improvements yielding higher structural alignment (69.4% vs. 64.7%), its overall Avg Ans. drops to 43.1% due to extremely weak numerical accuracy (12.4%).

At the 8B scale, Qwen3-8B-Instruct and InternLM3-8B show markedly lower structural alignment (26.6% and 25.0%) compared to larger models, while maintaining moderate answer accuracy (65.1% and 64.7%). This indicates that small models can achieve reasonable accuracy through pattern matching but lack the capacity to produce structurally coherent reasoning chains.

DeepSeek-R1, which employs explicit reasoning-enhanced training via reinforcement learning from process rewards, achieves the highest knowledge F1 of 24.7% and structural alignment of 71.7% among open models despite its moderate Avg Ans. of 61.4%. Its weak answer performance stems primarily from numerical (26.3%) and subjective (59.5%) tasks, suggesting that reasoning-enhanced training improves conceptual depth but does not guarantee computational precision. This demonstrates that targeted process-level training can substantially improve reasoning quality independent of answer accuracy.

E Prompt Templates

This section presents the key prompts used in our evaluation pipeline, organized by their role in the workflow: inference (Section E.1), knowledge extraction and grounding (Section E.2), and answer evaluation (Section E.3).

E.1 Inference Prompts

All inference prompts share a common system role and inject the subject-level KG node list to simulate realistic examination conditions where the candidate knows the syllabus but not the specific questions.

E.2 Knowledge Extraction and Grounding Prompts

FinProbe uses a two-stage process: first extracting knowledge concepts from the model's CoT, then grounding them against the subject KG.

E.3 Subjective Answer Evaluation Prompt

For subjective QA questions, we use an LLM-based multi-dimensional scoring approach.

F Ethical Considerations and Artifact Information

F.1 Licenses

FINDEPTH will be released under the CC BY-NC-SA 4.0 license for research purposes. The knowledge graph and evaluation scripts are released under the Apache 2.0 license. All source questions in the static question bank are derived from publicly available Chinese financial professional examination materials, which are in the public domain. The models evaluated in this work are used via their official APIs or publicly released weights under their respective licenses (e.g., DeepSeek models under the DeepSeek License, Qwen models under the Qwen License, InternLM under the Apache 2.0 License).

F.2 Intended Use

FINDEPTH is intended solely for research purposes: evaluating and comparing the financial reasoning capabilities of large language models. It should not be used as a substitute for professional financial advice, certification assessment, or any commercial decision-making. Our use of existing model APIs and open-weight models is consistent with their intended research and evaluation purposes as specified in their respective terms of service and model cards.

F.3 Human Annotators

Recruitment and qualifications. All annotators are graduate students majoring in finance-related disciplines at Chinese universities. They possess professional knowledge in financial accounting, securities, banking, and taxation, ensuring the quality of knowledge graph annotation and human evaluation tasks.

Compensation. Annotators are compensated at a fixed monthly rate of 2,000 RMB (approximately 275 USD), which is above the typical graduate research assistant stipend in the annotators' region of residence.

Consent and data use. All annotators were informed of the research purpose and data usage prior to participation. They provided written consent for their annotations to be used in academic publications and open-source dataset releases. No personally identifiable information is collected or retained in the released artifacts.

System Role (Shared Across All Question Types)

You are a senior financial domain expert with CPA, CFA, and PhD-level expertise. You are proficient in: macroeconomics and microeconomics theory; corporate finance, financial statement analysis and valuation; financial markets, portfolio theory and risk management; accounting standards (Chinese GAAP / IFRS / US GAAP); tax law, auditing, and economic law; banking, insurance, and securities business.
Your core capability: accurately identify relevant knowledge points, apply rigorous logical reasoning chains, and integrate theory with practice to produce professional, accurate, and insightful answers.

Inference Prompt: Multiple-Choice / True-False Questions

`[System Role]`

Subject Knowledge System ({n_nodes} knowledge points for reference)
{knowledge_points}

Requirements:
1. **Knowledge Anchoring**: Identify the core knowledge points tested, citing their definitions, properties, or criteria.
2. **Item-by-Item Analysis**: For MC: analyze each option against identified knowledge points with specific reasoning. For TF: state the judgment basis and whether the proposition matches professional definitions.
3. **Elimination and Confirmation**: Through rigorous reasoning, eliminate incorrect options and confirm the answer.

Question: {question}

Output: Place reasoning in `<reasoning>` tags. Place only the final answer letter(s) or "Correct"/"Incorrect" in `<answer>` tags.

Inference Prompt: Numerical Computation Questions

`[System Role]`

Subject Knowledge System ({n_nodes} knowledge points for reference)
{knowledge_points}

Requirements:
1. **Formula Sourcing**: Identify the core formulas or computational models from the knowledge system.
2. **Parameter Extraction**: Accurately extract all given conditions and data from the problem.
3. **Step-by-Step Computation**: Show each calculation step with the formula used and values substituted.
4. **Cross-Validation**: If possible, verify the result using an alternative method.

Question: {question}

Output: Place reasoning in `<reasoning>` tags. Place the final numerical result in `<answer>` tags.

FinProbe Stage 1: Knowledge Concept Extraction

You are a financial domain knowledge extraction expert. Extract all professional financial knowledge concepts from the following reasoning process, in order of appearance.

Requirements:
1. Only extract professional financial concepts (e.g., "corporate income tax", "net present value", "marginal cost").
2. Do not extract generic words (e.g., "calculate", "analyze", "result").
3. Express each concept as a concise term.
4. Strictly follow the order of appearance in the reasoning.
5. If the same concept appears multiple times, keep only its first occurrence.

Reasoning Process: {thinking}

Output: A JSON list, e.g., ["Concept1", "Concept2", "Concept3"]. Output only the JSON list.

FinProbe Stage 2: Context-Aware Knowledge Grounding

You are a financial domain knowledge matching expert. Given a model's **reasoning process**, the **candidate concepts** extracted from it, and the subject's **standard KG node list**, determine which KG node each candidate should map to (or NULL).

Matching Criteria:
1. Judge semantic mapping based on the full reasoning context, not surface form alone.
2. If a concept is only mentioned in passing without substantive discussion, or has no semantically matching KG node → map to "NULL".
3. Multiple candidates may map to the same KG node.
4. Matching is based on semantic equivalence or close relatedness, not exact string match.
5. Output KG node strings must exactly match the node list (character-for-character).

KG Node List ({n_nodes} nodes): {kg_nodes}
Candidate Concepts: {raw_concepts}
Reasoning Process: {thinking}

Output: A JSON object mapping each candidate to a KG node string or "NULL".
Example: {"Concept A": "KG Node X", "Concept B": "NULL", "Concept C": "KG Node Y"}

Subjective Answer Evaluation Dimensions

Rate the model's answer on the following dimensions (1–5 scale each):

Relevance (weight 0.15): Does the answer address the question directly?
Usefulness (weight 0.25): Does the answer provide valuable information that satisfies the user's needs?
Fluency (weight 0.10): Is the language grammatically correct and natural?
Coherence (weight 0.10): Is the logical structure clear and well-organized?
Consistency (weight 0.20): Does the answer align with the reference answer in conclusions and data?
Faithfulness (weight 0.10): Is the answer free of hallucinations and fabricated information?
Domain Specificity (weight 0.10): Does the answer demonstrate subject-specific expertise?

Given: Question, Reference Answer, Model Answer
Output: JSON object with dimension scores and brief justifications.