KNOW THYSELF: CONCEPT-LEVEL DIAGNOSIS FOR FINANCIAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly used in finance, but existing benchmarks provide limited evidence of what financial capabilities a model has actually mastered. As a result, models with similar benchmark scores may differ sharply in their underlying strengths and weaknesses. We introduce FinCDM, a framework for diagnosis at the concept level that combines the official CPA and CFA concept taxonomies, mappings between questions and concepts verified by experts, model response patterns, and a replaceable capability estimator. We construct a bilingual benchmark suite derived from the CPA and CFA examinations, consisting of CPA-KQA (Chinese, 1,260 questions, 70 concepts) and CFA-KQA (English, 738 questions, 41 concepts). Across 20 representative LLMs, we find that score-equivalent models are often not capability-equivalent, and that benchmark imbalance and linguistic limitations can further distort apparent competence. We further propose SkillTune, a diagnosis-guided data selection strategy for fine-tuning that targets each model's weak concepts. SkillTune improves multiple LLMs and, in a controlled comparison, outperforms alternative data selection strategies. These results support financial LLM evaluation based on concept-level profiles and diagnosis-guided improvement. Datasets and scripts will be publicly available.

1 INTRODUCTION

What financial capabilities does a large language model actually master? Large language models (LLMs) are increasingly used for financial tasks Nie et al. (2024). Yet similar task-level performance may arise from very different underlying abilities, while apparently strong results can still mask important weaknesses in reasoning, domain understanding, and robustness. What remains unresolved is what the model has actually learned to do reliably, making it difficult to assess the current state of capability development, and determine where improvement should be targeted, with downstream consequences for validation and trustworthy adoption in practice.

However, existing financial benchmarks mostly provide statistical evidence of task success, rather than diagnostic evidence of the financial capabilities. The community has proposed benchmarks spanning from comprehensive suites such as FinBen Xie et al. (2024) and CFinBench Nie et al. (2025) to more targeted benchmarks such as FinEval Guo et al. (2025) and InvestorBench Li et al. (2025). Yet across these benchmarks, evaluation is still based mainly on whether a model gets questions right and on its final score. As a result, two models can receive similar scores while mastering very different concepts, and a high score can sometimes come simply from doing well on a few over-tested topics or on the language of the benchmark.

To address this gap, we introduce **FinCDM**, a concept-level diagnosis framework for financial LLMs grounded in cognitive diagnosis from educational assessment. Cognitive diagnosis is the general paradigm of inferring fine-grained capabilities from observable response patterns; concept-level diagnosis is the output that FinCDM produces for financial LLMs. The framework connects the official CPA and CFA concept taxonomies, a Q matrix containing mappings between questions and concepts verified by experts, and a response matrix for models and questions to a capability profile for individual concepts. To enable such diagnosis, we construct a rigorously curated bilingual benchmark suite with broad financial knowledge coverage, including CPA-KQA, a Chinese dataset of 1,260 multiple-choice questions spanning 70 key concepts from the Certified Public Accountant

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

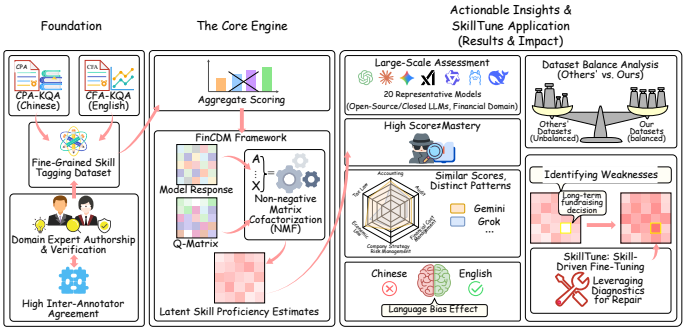


Figure 1: FinCDM proceeds from benchmark construction to diagnosis at the concept level and its application. The official CPA and CFA curricula define the concept taxonomies. Domain experts author questions and verify the mappings between questions and concepts. Model responses form the response matrix, a capability estimator produces profiles for individual concepts, and SkillTune uses diagnosed weak concepts to select targeted supervision.

(CPA) exam, and CFA-KQA, an English dataset of 738 questions covering 41 key concepts from the Chartered Financial Analyst (CFA) Level II exam. Both components support concept diagnosis and concept-driven improvement, and all questions are created and verified through a strict two-stage expert annotation process, yielding high-quality fine-grained concept labels. Finally, we propose SkillTune, a diagnosis-guided data selection strategy that uses each model's diagnosed weak concepts to select fine-tuning data.

Across 20 representative LLMs, we find that score-equivalent financial models are often not capability-equivalent. On CPA-KQA, Grok 3 and GPT-4 obtain the same overall accuracy (0.62) yet specialize in different legal subdomains, with Grok 3 stronger in *Economic Law* and GPT-4 in *Tax Law*. On CFA-KQA, GPT-4o and Llama 3 70B also tie at 0.72 accuracy while showing distinct strengths in *Derivative Investments* and *Alternative Investments*, respectively. The opposite contrasts of Llama 3 70B and DeepSeek V3 between CPA-KQA and CFA-KQA further reveal how linguistic limitations can distort apparent financial competence. With SkillTune, we use each model's diagnosed weak concepts to select fine-tuning data for Qwen 3 0.6B, Qwen 3 8B, and Llama 3.1 8B, fine-tuning each model separately on CPA-KQA and CFA-KQA. All six fine-tuned models improve on the benchmark used for fine-tuning, and the models fine-tuned on CPA-KQA also improve on the disjoint FinEval-KQA. In a controlled comparison, SkillTune also outperforms alternative data selection strategies.

In summary, our major contributions are: (1) We formulate financial LLM evaluation as concept-level diagnosis and introduce FinCDM (Figure 1), a comprehensive framework that integrates the official CPA and CFA concept taxonomies, mappings between questions and concepts verified by experts, and concept-level model capability profiling. (2) We construct two expert-annotated and concept-balanced benchmarks: CPA-KQA, which contains 1,260 Chinese questions covering 70 CPA concepts, and CFA-KQA, which contains 738 English questions covering 41 CFA Level II concepts. (3) We propose SkillTune, a diagnosis-guided data selection strategy for fine-tuning, and show that SkillTune-guided fine-tuning improves model capability and that SkillTune outperforms alternative data selection strategies. (4) Across 20 LLMs, we show that aggregate scores conceal distinct concept-level capability patterns.

2 Related Work

Financial evaluation. Financial benchmarks such as FinBen (Xie et al., 2024), CFinBench (Nie et al., 2025), and FinEval (Guo et al., 2025) evaluate diverse tasks and languages. FinQA (Chen et al., 2021) and TAT-QA (Zhu et al., 2021) emphasize numerical reasoning over financial reports, while FinTagging (Wang et al., 2025) provides taxonomy-aligned extraction labels. FinCDM focuses on estimating model capabilities over explicit CPA/CFA concepts, using balanced question coverage and expert-verified item–concept mappings.

Cognitive diagnosis. DINA (De La Torre, 2009) uses a parametric response function, whereas SNMCF (Yu et al., 2023), NeuralCD (Wang et al., 2022), and RCD (Gao et al., 2021) learn latent representations. FinCDM applies cognitive diagnosis to financial LLMs through matrix co-factorization; SkillTune uses the resulting concept profiles to select training data. Appendix A provides an extended review.

3 PRELIMINARIES

Cognitive diagnosis in educational assessment infers a learner's fine-grained capabilities from response patterns over items associated with explicit knowledge attributes Leighton & Gierl (2007). It estimates each capability from performance across multiple related items, while individual responses may also reflect guessing, memorization, or surface matching. FinCDM applies this framework to financial LLMs and produces concept-level capability profiles from their item responses.

Formally, let $\boldsymbol{\alpha} \in \{0,1\}^K$ denote the latent mastery profile, where $\alpha_k = 1$ indicates mastery of the k-th concept. Each item i is associated with a subset of required concepts through a binary Q-matrix $\mathbf{Q} \in \{0,1\}^{M\times K}$, where $q_{ik} = 1$ means that item i depends on concept k. Given an observed response vector $\mathbf{X} \in \{0,1\}^M$, CDMs model the probability of a correct response as

$$P(X_i = 1 \mid \boldsymbol{\alpha}, \mathbf{q}_i, \boldsymbol{\theta}) = f_i(\boldsymbol{\alpha}, \mathbf{q}_i; \boldsymbol{\theta}), \tag{1}$$

where f_i is the item response function and $\boldsymbol{\theta}$ denotes model parameters. FinCDM keeps this diagnostic goal but, instead of a parametric response function over discrete mastery profiles, estimates continuous concept-level scores through matrix co-factorization.

4 FINCDM

4.1 DATASET CURATION

Existing benchmarks often rely on aggregate scores and cover limited financial concepts. To address this, we introduce a bilingual expert-annotated data suite including CPA-KQA and CFA-KQA datasets, with comprehensive financial knowledge coverage, to support concept-level diagnosis and model training. The official CPA and CFA curricula define the subject structures, learning objectives, and concept inventories used in our benchmarks. Our experts author an equal number of original questions for each concept and independently verify every mapping between a question and its concept. Throughout this paper, *concept* refers to one of these official fine grained units, represented as a column of the Q matrix.

4.1.1 CPA-KQA

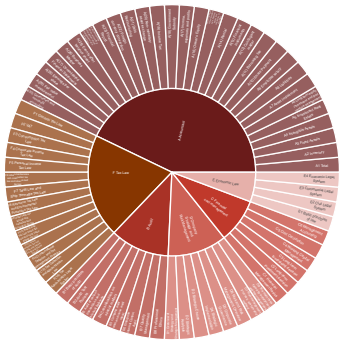
CPA-KQA is referenced against the Certified Public Accountant (CPA) examination specification[1] and spans six core areas: accounting, auditing, cost management, strategy, economic law, and tax law. It is organized around **70** CPA-defined concepts (see Figure 2a), with **18** multiple-choice items per concept (**1,260** items total).

Expert Annotation and Quality Validation. Three domain experts[2] constructed the dataset through authoring and independent review. Each expert wrote questions for assigned concepts, and the other two independently assessed each question's correctness, clarity, and alignment with the intended concept. Table 1 reports agreement before adjudication. The experts discussed disagreements, revised the affected items, and reviewed them again. All retained items reached post-adjudication consensus. For CPA-KQA, pre-adjudication agreement was $\alpha = 0.937$, $\sigma = 0.9904$, and the KS-based score was 1.0 Braylan et al. (2022).

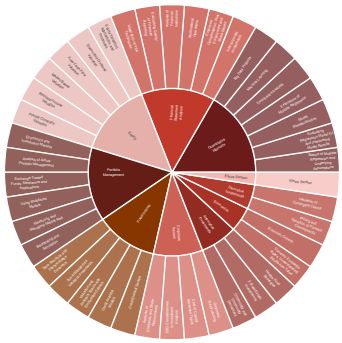
Strategic Data Partitioning. To support the complete "diagnosis-to-improvement" loop, we partitioned the full dataset into a test set and a training set using a 1:5 ratio. We reserved 3 items for each concept (210 total) to serve as the test benchmark for FinCDM. The remaining 15 items per

[1] https://www.cicpa.org.cn/xxfb/tzgg/202502/W020250228535255887805.pdf

[2] An undergraduate, a master's graduate, and an associate professor in finance, all proficient in Chinese and English financial terminology.



(a) 70 financial concepts covered by CPA-KQA.



(b) 41 financial concepts covered by CFA-KQA.

Figure 2: Concepts covered by CPA-KQA and CFA-KQA

	Dataset		
Metric	**CPA-KQA**	**CFA-KQA**	**FinEval-KQA**
Krippendorff's α	0.937	0.965	-
σ	0.9904	0.9974	0.9208
KS-based measure	1	1	1

Table 1: Inter-annotator agreement before adjudication for CPA-KQA, CFA-KQA, and FinEval-KQA. All retained items reached consensus after adjudication. Krippendorff's α measures agreement on item usability and is not reported for FinEval-KQA, whose questions are taken from an existing benchmark and require no usability review.

concept (1,050 total) form a supplementary training corpus, supporting the concept-driven fine-tuning experiments described in later sections.

4.1.2 CFA-KQA

To verify FinCDM's cross-lingual generalizability, we constructed CFA-KQA using the English-language Chartered Financial Analyst (CFA) Level II examination. We chose Level II because its vignette-based multiple-choice questions assess applied valuation and financial reasoning while supporting objective scoring and consistent item–concept mappings. Level I emphasizes foundational knowledge, while Level III includes constructed-response questions that require additional scoring procedures. CFA-KQA covers all 41 concepts in the CFA Level II curriculum, spanning core modules including ethical standards, quantitative methods, economics, and financial reporting, as shown in Figure 2b.

To ensure data quality, we applied the same rigorous, expert-driven pipeline to construct a pool of 738 high-quality questions, with 18 questions for each of the 41 concepts. As in CPA-KQA, three questions per concept (123 in total) form the test benchmark, and the remaining 15 per concept (615 in total) form the training corpus (Table 6).

4.1.3 FINEVAL-KQA

To examine knowledge coverage in existing benchmarks, we introduce **FinEval-KQA**. Specifically, we extend FinEval Guo et al. (2025), a widely used Chinese financial evaluation suite with over 2,000 questions spanning diverse sub-domains, by applying the **CPA-KQA** taxonomy to a curated subset of 101 questions spanning multiple CPA subjects. This extension equips a public benchmark with fine-grained, concept-level diagnostic functionality. Because these questions come from an existing benchmark, no item authoring or usability review is needed. One expert assigned each question to a CPA-KQA concept, and two other experts independently judged whether they agreed with the assignment. We assess annotation quality using the same metrics as in the CPA-KQA, namely σ and the KS-based measure. Agreement before adjudication is strong ($\sigma = 0.9208$, KS $= 1$).

4.2 Evaluation Framework

FinCDM organizes model responses into capability profiles for individual financial concepts using the official CPA and CFA concept taxonomies and mappings between questions and concepts verified by experts. The official taxonomies define the capabilities to evaluate, the binary Q matrix links each question to its concept, and the response matrix records each model's performance on every question. A replaceable capability estimator combines these inputs to produce a profile for each model. SkillTune uses the diagnosed weak concepts to select training examples for that model.

Let M, N, and K denote the numbers of items, models, and concepts, respectively. For item i in dataset d, let s_i denote the question stem and $\mathcal{O}_i^d$ the candidate options. On sampling run t, model j predicts an answer label $\hat{y}_{ij}^{(t)} = f_{\theta_j}(s_i, \mathcal{O}_i^d)$. The numbers of options are $L_{\text{CFA}} = 3$ and $L_{\text{CPA}} = L_{\text{FinEval}} = 4$.

For each item–model pair, we sample ten responses and score each response with the binary correctness function $\mathcal{V}(\hat{y}_{ij}^{(t)}, y_i)$: a response receives 1 if its predicted label matches the reference answer, and 0 otherwise. The corresponding entry in the response matrix X is the mean of these ten binary scores:

$$X_{ij} = \frac{1}{10} \sum_{t=1}^{10} \mathbf{1}[\hat{y}_{ij}^{(t)} = y_i], \qquad X \in [0,1]^{M \times N}. \tag{2}$$

Thus, X records each model's mean correctness on each item. We average the binary scores before fitting the estimator to the resulting matrix.

CPA-KQA and CFA-KQA support the core diagnostic experiments. FinEval-KQA serves two purposes: analyzing concept coverage in an existing benchmark, and testing whether models fine-tuned on CPA-KQA transfer to an independent set of CPA questions.

FinCDM separates the diagnostic pipeline from the capability estimator. Our primary implementation uses non-negative matrix co-factorization (MCF), drawing on SNMCF (Yu et al., 2023) and the attribute-level perspective of cognitive diagnosis (De La Torre, 2009).

We jointly factorize the response matrix X and the expert-verified item–concept matrix $Q \in \{0,1\}^{M \times K}$ through a shared item representation E:

$$X \approx EU, \qquad Q \approx EV, \qquad E, U, V \geq 0. \tag{3}$$

$E \in \mathbb{R}_{\geq 0}^{M \times r}$ contains item representations in an r-dimensional non-negative latent space. Each column of $U \in \mathbb{R}_{\geq 0}^{r \times N}$ represents a model's latent capabilities in this space, while $V \in \mathbb{R}_{\geq 0}^{r \times K}$ links the latent factors to explicit financial concepts. We set the latent rank according to the number of top-level subject areas represented in each assessment: $r = 6$ for CPA-KQA, $r = 10$ for CFA-KQA, and $r = 5$ for FinEval-KQA. These dataset-specific ranks are used for the capability profiles, thresholded concept counts, and held-out response-prediction results in Table 4, as well as for MCF-guided training-data selection.

We estimate E, U, and V by minimizing the reconstruction errors of X and Q:

$$\min_{E,U,V \geq 0} \quad \|W \circ (X - EU)\|_F^2 + \|Q - EV\|_F^2. \tag{4}$$

The first term reconstructs the response matrix, and the second reconstructs the mappings between questions and concepts. Both terms use unit weight.

The operator $\circ$ denotes elementwise multiplication. The mask $W \in \{0,1\}^{M \times N}$ identifies response entries used for fitting. We set $W = \mathbf{1}$ when fitting capability profiles to a fully observed response matrix and set $W_{ij} = 0$ for held-out entries in the estimator comparison. Appendix F provides the multiplicative update rules. The fitted factors link item responses to concept annotations through the shared latent representation.

We define the model–concept score matrix as

$$F = U^\top V, \qquad F \in \mathbb{R}_{\geq 0}^{N \times K}. \tag{5}$$

Each entry combines model and concept representations in the shared latent space. We fix the initialization scheme and random seed of the factorization (Appendix F), so F is determined by the

response matrix. For a fine-tuned model, we fix the item representation E and concept mapping V fitted before fine-tuning, estimate only the new model representation U_{ft} from its responses, and compute $F_{\text{ft}} = U_{\text{ft}}^{\top} V$. Concept scores before and after fine-tuning are therefore compared using the same item representations and concept mapping.

F_{jk} denotes the nonnegative latent score of model j on concept k. We set $\tau = 0.9$ as an empirical operating threshold; the model ranking by Con remains stable for τ between 0.85 and 0.95 (Appendix E.2). The number of concepts above the threshold is

$$\text{Con}_j(\tau) = \sum_{k=1}^{K} \mathbf{1}[F_{jk} > \tau]. \tag{6}$$

4.3 SKILLTUNE

SkillTune converts concept-level capability profiles into model-specific training supervision. Given $F \in \mathbb{R}_{\geq 0}^{N \times K}$, it selects training examples associated with concepts whose latent scores fall at or below the operating threshold. For model $j \in \{1, \ldots, N\}$, the selected set of weak concepts is

$$\mathcal{C}_j^{-}(\tau) = \{c \in \{1, \ldots, K\} \mid F_{j,c} \leq \tau\}. \tag{7}$$

We use the operating threshold $\tau = 0.9$. Let $\mathcal{D}_{\text{train}} = \{(x_i, y_i, c_i)\}_{i=1}^{|\mathcal{D}_{\text{train}}|}$ denote the training pool, where x_i is the instruction input, y_i is the target response, and c_i is the expert-verified concept label from the official CPA or CFA taxonomy. For model j, SkillTune selects the training subset

$$\mathcal{D}_j^{-} = \{(x_i, y_i) \mid (x_i, y_i, c_i) \in \mathcal{D}_{\text{train}},\ c_i \in \mathcal{C}_j^{-}(\tau)\}. \tag{8}$$

The target concepts are determined from each model's responses to the test items of the corresponding benchmark, and training examples are drawn only from the training pool, which shares no items with the test items. FinEval-KQA is not used for target-concept selection, hyperparameter tuning, or checkpoint selection, so it provides an evaluation that is independent of these choices.

We fine-tune each model on $\mathcal{D}_j^{-}$ through LoRA-based instruction tuning. Let θ denote the trainable LoRA parameters. We optimize

$$\theta' = \arg\min_{\theta} \sum_{(x,y) \in \mathcal{D}_j^{-}} \ell(f_\theta(x), y), \tag{9}$$

where ℓ denotes the instruction-tuning loss for target response y conditioned on input x. Appendix K reports the LoRA configuration and training settings.

For the controlled Qwen 3 8B study, we compare SkillTune with four alternative data selection strategies. Random-question randomly samples the same number of training examples as SkillTune from the full training pool without using concept labels. Random-concept randomly selects the same number of concepts as SkillTune and matches its total training-example budget. All-concept uses the complete training pool. Per-concept guided uses the model's empirical accuracy on the diagnostic questions. A question counts as answered incorrectly when at most half of the model's ten sampled responses are correct, that is, when its mean correctness is at most 0.5. A concept is selected as weak when at least two of its three diagnostic questions are answered incorrectly, and the same fine-tuning procedure is then applied.

We evaluate SkillTune by fine-tuning Qwen 3 0.6B, Qwen 3 8B, and Llama 3.1 8B separately on CPA-KQA and CFA-KQA. This produces six fine-tuned models. Each model is evaluated on the test set of the benchmark used for its fine-tuning, and the three models fine-tuned on CPA-KQA are additionally evaluated on FinEval-KQA.

To compare training-data selection strategies and evaluate cross-dataset transfer, we conduct an additional controlled study with Qwen 3 8B in the CPA setting. We use one backbone for this repeated comparison to keep the computational cost tractable. All methods use the same optimizer and LoRA configuration. We report mean accuracy over five data-selection and fine-tuning runs for Random-question and Random-concept, and one run for SkillTune, All-concept, and Per-concept guided. The models are fine-tuned on CPA-KQA and evaluated on both CPA-KQA and FinEval-KQA. FinEval-KQA contains all 101 CPA questions provided by FinEval and therefore serves as an external CPA test set for examining whether improvements obtained on CPA-KQA transfer across datasets.

4.4 BENCHMARKING

We benchmark 20 LLMs spanning proprietary general models, open general models, and financial models. The proprietary models include GPT-4 and GPT-4o Achiam et al. (2023), Claude 3.5 Sonnet and Claude 3.7 Sonnet Anthropic (2024), Gemini 1.5 Pro and Gemini 2.5 Pro Exp Team et al. (2024), Grok 3 xAI (2025), and Qwen Max Team (2024a). The open general models include Baichuan2 13B Chat Baichuan Inc. (2023), Falcon 7B Almazrouei et al. (2023), GLM-4 9B GLM et al. (2024), Qwen 2.5 7B Instruct, Qwen 3 0.6B, Qwen 3 8B, Qwen 3 235B Yang et al. (2025a); Team (2024b), Llama 3 70B, Llama 3.1 8B Dubey et al. (2024), and DeepSeek V3 Liu et al. (2024). The financial models are FinMA 7B Full Xie et al. (2023) and Fin-o1 8B Qian et al. (2025). Appendix H provides detailed model information.

We use the same prompting and evaluation protocol for all models. Each model generates 10 responses per question, and we report the mean score across these responses. The prompts are provided in Appendix I.

5 EXPERIMENTS

To evaluate the effectiveness of FinCDM, we address the following research questions: **RQ1:** Can FinCDM reveal fine-grained knowledge gaps and behavioral patterns that are obscured by aggregate scoring metrics? **RQ2:** Do the proposed CPA-KQA and CFA-KQA benchmarks provide broader and more balanced coverage of financial concepts than existing benchmarks? **RQ3:** Can FinCDM's diagnostic signals effectively guide targeted model improvement, e.g., via concept-driven fine-tuning?

5.1 MAIN RESULTS

CFA-KQA			**CPA-KQA**			**FinEval-KQA**		
Acc	Model	Con	Acc	Model	Con	Acc	Model	Con
0.86	Claude 3.7 Sonnet	**32**	0.73	Claude 3.5 Sonnet	**43**	**0.87**	Gemini 2.5 Pro Exp	**9**
0.77	Grok 3	31	**0.83**	Gemini 2.5 Pro Exp	42	0.76	Claude 3.7 Sonnet	7
0.77	Qwen 3 235B	25	0.75	Claude 3.7 Sonnet	40	0.70	Claude 3.5 Sonnet	6
0.77	Gemini 2.5 Pro Exp	24	0.74	Qwen Max	37	0.75	Qwen Max	6
0.77	Gemini 1.5 Pro	22	0.66	DeepSeek V3	36	0.80	Qwen 3 235B	6
0.72	GPT-4o	22	0.74	Qwen 3 235B	35	0.68	Qwen 2.5 7B Instruct	6
0.72	Llama 3 70B	22	0.62	GPT-4	33	0.49	GPT-4o	5
0.71	Qwen Max	22	0.62	Grok 3	29	0.74	DeepSeek V3	5
0.72	Claude 3.5 Sonnet	21	0.58	Gemini 1.5 Pro	25	0.39	Baichuan2 13B Chat	5
0.53	Fin-o1 8B	21	0.55	GLM-4 9B	22	0.65	GPT-4	4
0.69	GPT-4	17	0.61	Qwen 2.5 7B Instruct	20	0.54	Gemini 1.5 Pro	3
0.56	GLM-4 9B	17	0.53	GPT-4o	17	0.60	Grok 3	3
0.37	Qwen 3 0.6B	17	0.45	Llama 3 70B	17	0.30	Qwen 3 0.6B	3
0.54	Qwen 3 8B	15	0.24	Qwen 3 0.6B	14	0.47	Llama 3 70B	3
0.32	Finma 7B Full	12	0.42	Fin-o1 8B	12	0.56	GLM-4 9B	3
0.43	Baichuan2 13B Chat	11	0.35	Baichuan2 13B Chat	7	0.29	Qwen 3 8B	2
0.52	DeepSeek V3	9	0.33	Qwen 3 8B	6	0.23	Llama 3.1 8B	2
0.27	Falcon 7B	9	0.23	Llama 3.1 8B	5	0.38	Fin-o1 8B	2
0.63	Qwen 2.5 7B Instruct	8	0.21	Finma 7B Full	0	0.21	Falcon 7B	2
0.31	Llama 3.1 8B	6	0.15	Falcon 7B	0	0.19	Finma 7B Full	2

Table 2: Results on CFA-KQA-test, CPA-KQA-test and FinEval-KQA. "Con" denotes the number of concepts with a latent score above $\tau = 0.9$. Con on FinEval-KQA has no diagnostic role and is shown only to illustrate the uneven concept distribution of this benchmark. Rows are sorted by Con in descending order.

Table 2 summarizes aggregate accuracy across CPA-KQA, CFA-KQA, and FinEval-KQA, and Figure 3 shows the concept-level latent scores on CPA-KQA and CFA-KQA. We discuss the main findings below.

The Role of FinCDM. FinCDM provides finer-grained information than aggregate metrics by organizing performance across individual financial concepts. As shown in Figure 3 and Table 2, Claude 3.5 Sonnet attains high average accuracy on CFA-KQA while its concept-level profile remains uneven. This profile identifies where the model performs strongly and where targeted evaluation or training is needed.

Divergent Knowledge Specialization across High-Performing Models. Identical aggregate accuracy often implies uniform competence, but our analysis reveals substantial divergence in subdomain

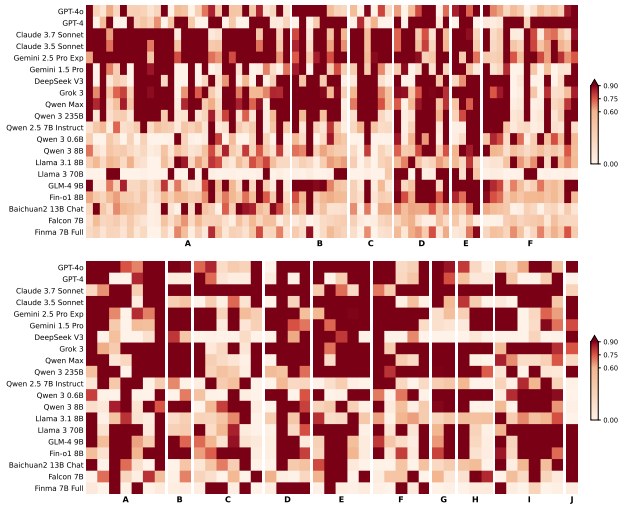


Figure 3: Concept-level latent-score heatmaps for CPA-KQA and CFA-KQA.

expertise. For example, in CPA-KQA, see Figure 3 and Table 2, Grok 3 and GPT-4 share a 0.62 accuracy yet display opposing strengths. GPT-4 excels in *Tax Law*, whereas Grok 3 demonstrates superior proficiency in *Economic Law*. Similarly, within the CFA-KQA dataset, GPT-4o and Llama 3 70B tie at 0.72 accuracy but exhibit distinct profiles: GPT-4o is stronger in *Derivative Investments*, while Llama 3 70B is stronger in *Alternative Investments*. These findings highlight that "equivalent" models are rarely interchangeable, reinforcing the importance of modular evaluation to isolate specific capabilities.

Influence of Linguistic Resources on Model Performance. Model performance may also depend on the benchmark language. Llama 3 70B reaches 0.72 accuracy on English CFA-KQA but 0.45 on Chinese CPA-KQA, whereas DeepSeek V3 reaches 0.66 on CPA-KQA but 0.52 on CFA-KQA. The concept-level profiles show a similar contrast in *Tax Law*, whose concepts, such as value-added tax, consumption tax, and vehicle and vessel tax, are defined by Chinese statutes and Chinese-specific terminology: Llama 3 70B has only 2 of 14 *Tax Law* concepts above τ, compared with 8 for DeepSeek V3. This opposite pattern is consistent with their training corpora (Dubey et al., 2024; Liu et al., 2024).

Evaluating Datasets on Knowledge Mastery. To examine how dataset structure affects evaluation, we additionally test LLMs on FINEVAL-KQA, which we annotate at the concept level. As shown in Figure 8 (Appendix G), its questions are concentrated in a few concepts, including *Financial Instruments* and *Commercial Law*, with the latter appearing up to 13 times. Aggregate accuracy on FinEval-KQA therefore weights these frequent concepts more heavily, whereas CPA-KQA assigns the same number of questions to every concept, so each concept contributes equally to its aggregate accuracy.

Effectiveness Analysis of SkillTune. We apply model-specific weak-concept selection to Qwen 3 0.6B, Qwen 3 8B, and Llama 3.1 8B on CPA-KQA and CFA-KQA (Table 3a and Figure 4). On CPA-KQA, accuracy increases from 0.24 to 0.39, from 0.33 to 0.58, and from 0.23 to 0.34, respectively. On CFA-KQA, the corresponding increases are from 0.37 to 0.44, from 0.54 to 0.62, and from 0.31 to 0.58. Con also increases for every model on both benchmarks. The models fine-tuned on CPA-KQA also improve on FinEval-KQA, from 0.30 to 0.51, from 0.29 to 0.58, and from 0.23 to 0.41. Figure 4 shows that Qwen 3 8B and Llama 3.1 8B retain nearly all previously mastered concepts while gaining new ones, with no sign of catastrophic forgetting; Qwen 3 0.6B retains about half, likely due to its limited capacity.

Comparison with data selection baselines. Table 3b compares SkillTune with four alternative data selection strategies on Qwen 3 8B fine-tuned on CPA-KQA. All strategies improve over the untuned model on both CPA-KQA and FinEval-KQA, and SkillTune achieves the highest accuracy on both (0.58 and 0.58). It exceeds the strongest alternative, All-concept, by 0.03 on CPA-KQA and 0.05 on

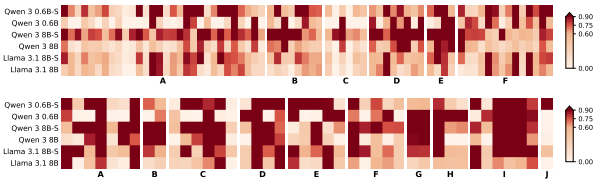


Figure 4: Concept-level latent-score heatmaps for CPA-KQA and CFA-KQA before and after SkillTune. Darker cells indicate higher latent scores; suffix -S denotes a SkillTune model.

Table 3: SkillTune results. CPA, CFA, and FinEval denote CPA-KQA, CFA-KQA, and FinEval-KQA. (a) Accuracy before → after tuning, with Con in parentheses. (b) Qwen 3 8B data-selection controls on CPA-KQA. FinEval-KQA evaluates CPA-trained models. Con is not reported on FinEval-KQA because its uneven concept distribution gives the count no diagnostic role. Random controls average five runs; other fine-tuned results use one.

(a) Across models.

Model	CPA	CFA	FinEval
Qwen 3 0.6B	$0.24 \rightarrow 0.39$ $(14 \rightarrow 20)$	$0.37 \rightarrow 0.44$ $(17 \rightarrow 19)$	$0.30 \rightarrow 0.51$ (–)
Qwen 3 8B	$0.33 \rightarrow 0.58$ $(6 \rightarrow 24)$	$0.54 \rightarrow 0.62$ $(15 \rightarrow 22)$	$0.29 \rightarrow 0.58$ (–)
Llama 3.1 8B	$0.23 \rightarrow 0.34$ $(5 \rightarrow 8)$	$0.31 \rightarrow 0.58$ $(6 \rightarrow 17)$	$0.23 \rightarrow 0.41$ (–)

(b) Data selection.

Method	CPA	FinEval
Untuned	0.33	0.29
SkillTune	**0.58**	**0.58**
Per-concept guided	0.51	0.49
Random-question	0.51	0.50
Random-concept	0.52	0.51
All-concept	0.55	0.53

FinEval-KQA. Per-concept guided selects concepts from the same diagnostic responses using raw accuracy without the matrix co-factorization, and SkillTune exceeds it by 0.07 on CPA-KQA and 0.09 on FinEval-KQA. All-concept trains on the complete training pool, so SkillTune's advantage comes from which examples it selects rather than from using more data. The larger margin on the external FinEval-KQA indicates that the gain transfers beyond the fine-tuning benchmark.

6 Conclusion

We introduced FinCDM, a framework that diagnoses financial LLMs at the level of individual professional concepts, together with the concept-balanced CPA-KQA and CFA-KQA benchmarks. Our analysis shows that aggregate accuracy is an incomplete summary of financial competence: models with similar scores can differ systematically in what they know, and benchmark composition can shape the scores themselves. The diagnosis is also actionable, as SkillTune turns diagnosed weak concepts into more effective fine-tuning data. These results suggest that evaluation in professional domains should report where a model is competent in addition to how often it is correct. Future work can combine FinCDM with Computerized Adaptive Testing and extend concept-level diagnosis to other expert-curriculum domains such as law and medicine.

7 Limitations

FinCDM currently diagnoses LLMs through text-based multiple-choice questions. Financial practice also involves open-ended tasks, such as report generation and investment recommendation, and multimodal inputs, such as balance sheets, charts, and regulatory filings. Extending the diagnosis to these settings requires response formats and scoring schemes beyond binary correctness. In addition, CPA-KQA and CFA-KQA follow the concept-level organization of the two examinations, so each item maps to one concept and the Q matrix is taken as given from expert annotation. Settings in which items draw on several concepts would allow FinCDM to model concept interactions and to refine the Q matrix from response data. Finally, the controlled comparison of data selection strategies is conducted on Qwen 3 8B, and repeating it on other model families and scales would further test how SkillTune generalizes.

AI USE STATEMENT

In this work, we used generative AI tools to improve the readability of the manuscript, identify potential factual errors, and assist in checking references selected by the authors. The authors reviewed all AI-assisted suggestions and incorporated manuscript changes only after human assessment. The authors take responsibility for the final manuscript, references, and results.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*, 2023.

Anthropic. Introducing the claude 3 model family, 2024. URL https://www.anthropic.com/news/claude-3-family. Accessed: 2025-07-30.

Baichuan Inc. Baichuan2: Open-source large language models, 2023. URL https://github.com/baichuan-inc/Baichuan2. Accessed: 2025-07-30.

Alexander Braylan, Omar Alonso, and Matthew Lease. Measuring annotator agreement generally across complex structured, multi-object, and free-text annotation tasks. In *Proceedings of the ACM Web Conference 2022*, pp. 1720–1730, 2022.

Yupeng Cao, Haohang Li, Yangyang Yu, Shashidhar Reddy Javaji, Yueru He, Jimin Huang, Zining Zhu, Qianqian Xie, Xiao-yang Liu, Koduvayur Subbalakshmi, et al. Finaudio: A benchmark for audio large language models in financial applications. *arXiv preprint arXiv:2503.20990*, 2025.

Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. Overview of the ntcir-14 finnum task: Fine-grained numeral understanding in financial social media data. In *Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies*, pp. 19–27, 2019.

Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. Overview of the NTCIR-15 FinNum-2 task: Numeral attachment in financial tweets. In *Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies*, 2020. URL https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings15/pdf/ntcir/01-NTCIR15-OV-FINNUM-ChenC.pdf.

Jian Chen, Peilin Zhou, Yining Hua, Loh Xin, Kehui Chen, Ziyuan Li, Bing Zhu, and Junwei Liang. FinTextQA: A dataset for long-form financial question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6025–6047. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.328. URL https://aclanthology.org/2024.acl-long.328/.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3697–3711. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.300. URL https://aclanthology.org/2021.emnlp-main.300/.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. ConvFinQA: Exploring the chain of numerical reasoning in conversational finance question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6279–6292. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.421. URL https://aclanthology.org/2022.emnlp-main.421/.

Chanyeol Choi, Jihoon Kwon, Jaeseon Ha, Hojun Choi, Chaewoon Kim, Yongjae Lee, Jy-yong Sohn, and Alejandro Lopez-Lira. FinDER: Financial dataset for question answering and evaluating retrieval-augmented generation. In *Proceedings of the 6th ACM International Conference on AI in Finance*, pp. 638–646. Association for Computing Machinery, 2025. doi: 10.1145/3768292.3770361. URL https://doi.org/10.1145/3768292.3770361.

Jimmy De La Torre. Dina model and parameter estimation: A didactic. *Journal of educational and behavioral statistics*, 34(1):115–130, 2009.

Jimmy De La Torre. The generalized dina model framework. *Psychometrika*, 76(2):179–199, 2011.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akshay Mathur, ..., and the Llama 3 Team. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783. Meta AI.

Mahmoud El-Haj, Marina Litvak, Nikiforos Pittaras, George Giannakopoulos, et al. The financial narrative summarisation shared task (fns 2020). In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pp. 1–12, 2020.

Weibo Gao, Qi Liu, Zhenya Huang, Yu Yin, Haoyang Bi, Mu-Chun Wang, Jianhui Ma, Shijin Wang, and Yu Su. Rcd: Relation map driven cognitive diagnosis for intelligent education systems. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pp. 501–510, 2021.

Sohom Ghosh and Sudip Kumar Naskar. Fincat: Financial numeral claim analysis tool. In *Companion Proceedings of the Web Conference 2022*, pp. 583–585, 2022.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

Yuqi Gu and Gongjun Xu. The sufficient and necessary condition for the identifiability and estimability of the dina model. *Psychometrika*, 84(2):468–483, 2019.

Xin Guo, Haotian Xia, Zhaowei Liu, Hanyang Cao, Zhi Yang, Zhiqiang Liu, Sizhe Wang, Jinyi Niu, Chuqi Wang, Yanhui Wang, et al. 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)*, pp. 6258–6292, 2025.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. Financebench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*, 2023.

Simerjot Kaur, Charese Smiley, Akshat Gupta, Joy Sain, Dongsheng Wang, Suchetha Siddagangappa, Toyin Aguda, and Sameena Shah. Refind: Relation extraction financial dataset. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pp. 3054–3063, 2023.

Mateusz Klimaszewski, Pinzhen Chen, Liane Guillou, Ioannis Papaioannou, Barry Haddow, and Alexandra Birch. Avenibench: Accessible and versatile evaluation of finance intelligence. 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)*, pp. 111–117, 2025.

Michael Krumdick, Rik Koncel-Kedziorski, Viet Dac Lai, Varshini Reddy, Charles Lovering, and Chris Tanner. Bizbench: A quantitative reasoning benchmark for business and finance. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8309–8332, 2024.

Viet Dac Lai, Michael Krumdick, Charles Lovering, Varshini Reddy, Craig Schmidt, and Chris Tanner. SEC-QA: A systematic evaluation corpus for financial QA. In *Proceedings of the 10th Workshop on Financial Technology and Natural Language Processing*, pp. 221–236. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.finnlp-2.15. URL https://aclanthology.org/2025.finnlp-2.15/.

Kai Lan, Jiayong Zhu, Jiangtong Li, Dawei Cheng, Guang Chen, and Changjun Jiang. Finlmm-r1: Enhancing financial reasoning in lmm through scalable data and reward design. *arXiv preprint arXiv:2506.13066*, 2025.

Jacqueline Leighton and Mark Gierl. *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press, 2007.

Haohang Li, Yupeng Cao, Yangyang Yu, Shashidhar Reddy Javaji, Zhiyang Deng, Yueru He, Yuechen Jiang, Zining Zhu, K. P. Subbalakshmi, Jimin Huang, Lingfei Qian, Xueqing Peng, Jordan W. Suchow, and Qianqian Xie. INVESTORBENCH: A benchmark for financial decision-making tasks with LLM-based agent. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2509–2525. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.126. URL https://aclanthology.org/2025.acl-long.126/.

Xiang Li, Zhenyu Li, Chen Shi, Yong Xu, Qing Du, Mingkui Tan, and Jun Huang. AlphaFin: Benchmarking financial analysis with retrieval-augmented stock-chain framework. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 773–783. ELRA and ICCL, 2024. URL https://aclanthology.org/2024.lrec-main.69/.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Shu Liu, Shangqing Zhao, Chenghao Jia, Xinlin Zhuang, Zhaoguang Long, Jie Zhou, Aimin Zhou, Man Lan, and Yang Chong. FinDABench: Benchmarking financial data analysis ability of large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 710–725. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.coling-main.48/.

Lefteris Loukas, Manos Fergadiotis, Ilias Chalkidis, Eirini Spyropoulou, Prodromos Malakasiotis, Ion Androutsopoulos, and Georgios Paliouras. FiNER: Financial numeric entity recognition for XBRL tagging. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4419–4431. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.303. URL https://aclanthology.org/2022.acl-long.303/.

Guilong Lu, Xuntao Guo, Rongjunchen Zhang, Wenqiao Zhu, and Ji Liu. Bizfinbench: A business-driven real-world financial benchmark for evaluating llms. *arXiv preprint arXiv:2505.19457*, 2025.

Junyu Luo, Zhizhuo Kou, Liming Yang, Xiao Luo, Jinsheng Huang, Zhiping Xiao, Jingshu Peng, Chengzhong Liu, Jiaming Ji, Xuanzhe Liu, Sirui Han, Ming Zhang, and Yike Guo. FinMME: Benchmark dataset for financial multi-modal reasoning evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29465–29489. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1426. URL https://aclanthology.org/2025.acl-long.1426/.

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

Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, Manjunath Hegde, Afreen Shaikh, Shivani Shrivastava, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh, and Pawan Goyal. ECTSum: A new benchmark dataset for bullet point summarization of long earnings call transcripts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 10893–10906. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.748. URL https://aclanthology.org/2022.emnlp-main.748/.

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. 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), pp. 876–891. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.40. URL https://aclanthology.org/2025.naacl-long.40/.

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

Xueqing Peng, Triantafillos Papadopoulos, Efstathia Soufleri, Polydoros Giannouris, Ruoyu Xiang, Yan Wang, Lingfei Qian, Jimin Huang, Qianqian Xie, and Sophia Ananiadou. Plutus: Benchmarking large language models in low-resource greek finance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30176–30202. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1535. URL https://aclanthology.org/2025.emnlp-main.1535/.

Xueqing Peng, Lingfei Qian, Yan Wang, Ruoyu Xiang, Yueru He, Yang Ren, Mingyang Jiang, Vincent Jim Zhang, Yuqing Guo, Jeff Zhao, Huan He, Yi Han, Yun Feng, Yuechen Jiang, Yupeng Cao, Haohang Li, Yangyang Yu, Xiaoyu Wang, Penglei Gao, Shengyuan Lin, Keyi Wang, Shanshan Yang, Yilun Zhao, Zhiwei Liu, Peng Lu, Jerry Huang, Suyuchen Wang, Triantafillos Papadopoulos, Polydoros Giannouris, Efstathia Soufleri, Nuo Chen, Zhiyang Deng, Heming Fu, Yijia Zhao, Mingquan Lin, Meikang Qiu, Kaleb E. Smith, Arman Cohan, Xiao-Yang Liu, Jimin Huang, Guojun Xiong, Alejandro Lopez-Lira, Xi Chen, Junichi Tsujii, Jian-Yun Nie, Sophia Ananiadou, and Qianqian Xie. MultiFinBen: Benchmarking large language models for multilingual and multimodal financial application. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16934–16963. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.770. URL https://aclanthology.org/2026.acl-long.770/.

Lingfei Qian, Weipeng Zhou, Yan Wang, Xueqing Peng, Jimin Huang, and Qianqian Xie. Fino1: On the transferability of reasoning enhanced LLMs to finance, 2025. URL https://arxiv.org/abs/2502.08127.

Varshini Reddy, Rik Koncel-Kedziorski, Viet Dac Lai, Michael Krumdick, Charles Lovering, and Chris Tanner. DocFinQA: A long-context financial reasoning dataset. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 445–458. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-short.42. URL https://aclanthology.org/2024.acl-short.42/.

Agam Shah, Abhinav Gullapalli, Ruchit Vithani, Michael Galarnyk, and Sudheer Chava. Finer-ord: Financial named entity recognition open research dataset. *arXiv preprint arXiv:2302.11157*, 2023.

Soumya Sharma, Tapas Nayak, Arusarka Bose, Ajay Kumar Meena, Koustuv Dasgupta, Niloy Ganguly, and Pawan Goyal. Finred: A dataset for relation extraction in financial domain. In *Companion Proceedings of the Web Conference 2022*, pp. 595–597, 2022.

Yixuan Tang and Yi Yang. FinMTEB: Finance massive text embedding benchmark. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3620–3638. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.179. URL https://aclanthology.org/2025.emnlp-main.179/.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Qwen Team. Qwen2.5-technical-report. *arXiv preprint arXiv:2412.15115*, 2024a.

Qwen Team. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024b.

Jonathan L Templin and Robert A Henson. Measurement of psychological disorders using cognitive diagnosis models. *Psychological methods*, 11(3):287, 2006.

Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang, Yu Yin, Shijin Wang, and Yu Su. Neuralcd: a general framework for cognitive diagnosis. *IEEE Transactions on Knowledge and Data Engineering*, 35 (8):8312–8327, 2022.

Yan Wang, Yang Ren, Lingfei Qian, Xueqing Peng, Keyi Wang, Yi Han, Dongji Feng, Xiao-Yang Liu, Jimin Huang, and Qianqian Xie. Fintagging: An llm-ready benchmark for extracting and structuring financial information. *arXiv preprint arXiv:2505.20650*, 2025.

Xiaojun Wu, Junxi Liu, Huan-Yi Su, Zhouchi Lin, Yiyan Qi, Chengjin Xu, Jiajun Su, Jiajie Zhong, Fuwei Wang, Saizhuo Wang, Fengrui Hua, Jia Li, and Jian Guo. Golden touchstone: A comprehensive bilingual benchmark for evaluating financial large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 22544–22560. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1227. URL https://aclanthology.org/2025.findings-emnlp.1227/.

xAI. Introducing grok-3, 2025. URL https://x.ai/news/grok-3. Accessed: 2025-07-30.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. Pixiu: A comprehensive benchmark, instruction dataset and large language model for finance. *Advances in Neural Information Processing Systems*, 36:33469–33484, 2023.

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

Zhuohan Xie, Daniil Orel, Rushil Thareja, Dhruv Sahnan, Hachem Madmoun, Fan Zhang, Debopriyo Banerjee, Georgi Nenkov 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. FinChain: A symbolic benchmark for verifiable chain-of-thought financial reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14529–14553. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.662. URL https://aclanthology.org/2026.acl-long.662/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Yuzhe Yang, Yifei Zhang, Yan Hu, Yilin Guo, Ruoli Gan, Yueru He, Mingcong Lei, Xiao Zhang, Haining Wang, Qianqian Xie, Jimin Huang, Honghai Yu, and Benyou Wang. UCFE: A user-centric financial expertise benchmark for large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5444–5463. Association for Computational Linguistics, 2025b. doi: 10.18653/v1/2025.findings-naacl.300. URL https://aclanthology.org/2025.findings-naacl.300/.

Shenbao Yu, Yifeng Zeng, Yinghui Pan, and Fan Yang. Snmcf: a scalable non-negative matrix co-factorization for student cognitive modeling. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3486–3500, 2023.

Ziqiang Yuan, Kaiyuan Wang, Shoutai Zhu, Ye Yuan, Jingya Zhou, Yanlin Zhu, and Wenqi Wei. FinLLMs: A framework for financial reasoning dataset generation with large language models. *IEEE Transactions on Big Data*, 11(5):2264–2277, 2025. doi: 10.1109/TBDATA.2024.3524083. URL https://doi.org/10.1109/TBDATA.2024.3524083.

Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. MultiHiertt: Numerical reasoning over multi hierarchical tabular and textual data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6588–6600. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.454. URL https://aclanthology.org/2022.acl-long.454/.

Peilin Zhou, Ziyue Xu, Xinyu Shi, Jiageng Wu, Yikang Jiang, Dading Chong, Wang Dong, Jun Chen, Bin Ke, and Jie Yang. FinTruthQA: A benchmark for AI-driven financial disclosure quality assessment in investor-firm interactions. *Proceedings of the AAAI Symposium Series*, 9(1):204–212, 2026. doi: 10.1609/aaaiss.v9i1.42926. URL https://doi.org/10.1609/aaaiss.v9i1.42926.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 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)*, pp. 3277–3287. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.254. URL https://aclanthology.org/2021.acl-long.254/.

Jie Zhu, Junhui Li, Yalong Wen, and Lifan Guo. Benchmarking large language models on CFLUE: A chinese financial language understanding evaluation dataset. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5673–5693. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.337. URL https://aclanthology.org/2024.findings-acl.337/.

Jie Zhu, Junhui Li, Yalong Wen, Xiandong Li, Lifan Guo, and Feng Chen. MFinMeeting: A multilingual, multi-sector, and multi-task financial meeting understanding evaluation dataset. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 244–266. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.14. URL https://aclanthology.org/2025.findings-acl.14/.

A Related Work

A.1 Financial Benchmark

Numerous benchmarks have been developed to evaluate LLMs in financial domains, covering tasks such as extraction, QA, reasoning, simulation, and retrieval. Early efforts like PIXIU Xie et al. (2023) and FinBen Xie et al. (2024) offer broad coverage but rely on aggregate metrics. Benchmarks like FinanceBench Islam et al. (2023), BizBench Krumdick et al. (2024), and BizFinBench Lu et al. (2025) focus on QA and business reasoning, while multilingual and low-resource evaluation is addressed by CFinBench Nie et al. (2025), CFLUE Zhu et al. (2024), FinEval Guo et al. (2025), Golden Touchstone Wu et al. (2025), Plutus-ben Peng et al. (2025), and the more recent MultiFinBen Peng et al. (2026). FinMTEB Tang & Yang (2025) covers embedding-based retrieval and classification. Structured reasoning is evaluated in FinDABench Liu et al. (2025), Fino1 Qian et al. (2025), FinChain Xie et al. (2026), and agentic decision-making in InvestorBench Li et al. (2025), AlphaFin Li et al. (2024), and AveniBench Klimaszewski et al. (2025). Multimodal understanding is addressed by M³FinMeeting Zhu et al. (2025) and FinAudio Cao et al. (2025). FinTagging Wang et al. (2025) and FinDER Choi et al. (2025) focus on fine-grained financial concept extraction and retrieval-augmented QA from 10-K filings and XBRL reports. Despite their breadth, these benchmarks rely on task-level aggregate metrics and lack concept-aware diagnostics.

A.2 Financial Dataset Design

A wide range of data sets have been developed to evaluate financial LLM in reasoning, QA, information extraction, summarization, and multimodal tasks. Numerical reasoning datasets include FinQA Chen et al. (2021), TAT-QA Zhu et al. (2021), ConvFinQA Chen et al. (2022), DocFinQA Reddy et al. (2024), and FinanceQA Mateega et al. (2025), which require multistep or conversational numerical inference over financial reports. FinTextQA Chen et al. (2024), FinLLMs Yuan et al. (2025), and FinTruthQA Zhou et al. (2026) focus on long-form or formula-based QA, while SEC-QA Lai et al. (2025) and MultiHiertt Zhao et al. (2022) introduce multidocument and table-text hybrid challenges. Structured extraction is addressed by FinTagging Wang et al. (2025), FiNER-ORD Shah et al. (2023), FiNER-139 Loukas et al. (2022), FinRED Sharma et al. (2022), REFinD Kaur et al. (2023), which allow evaluation of concept and relationship levels aligned with financial taxonomies. For numeral understanding, FinNum Chen et al. (2019), FinNum-2 Chen et al. (2020), and FiNCAT Ghosh & Naskar (2022) annotate fine-grained semantic types or numeral claims in finance text. The summarization is covered by ECTSum Mukherjee et al. (2022), FNS-2020 El-Haj et al. (2020), focusing on earnings calls and annual reports. Chinese and multilingual resources include FinEval Guo et al. (2025), CFLUE Zhu et al. (2024), and UCFE Yang et al. (2025b), enabling QA and NLU in non-English contexts. Multimodal datasets such as FinMME Luo et al. (2025) and FinLMM-R1 Lan et al. (2025) support chart, image, and document–text alignment. Despite covering diverse formats and tasks, most existing datasets are task-driven with limited concept coverage.

A.3 Cognitive Diagnosis Model

Cognitive diagnosis models (CDMs), rooted in educational assessment, aim to infer mastery of latent knowledge attributes from observed response behaviors. Early interpretable models such as the Deterministic Inputs, Noisy 'And' gate model (DINA) De La Torre (2009), the Deterministic Inputs, Noisy 'Or' gate model (DINO) Templin & Henson (2006), and the Generalized DINA (GDINA) model De La Torre (2011) employ binary latent attributes and probabilistic response functions; while valuable for interpretation, they depend heavily on accurate Q matrix specification and strict parametric assumptions Gu & Xu (2019). To mitigate these limitations, matrix factorization-based approaches, such as SNMCF Yu et al. (2023), embed examinees and attributes into low-dimensional spaces, enabling more flexible and robust modeling of item-attribute interactions. Neural network-based CDMs such as NeuralCD Wang et al. (2022) and recent graph-based models such as RCD Gao et al. (2021) use deep architectures and disentangled graph learning to model nonlinear and noise-robust concept, exercise, and student representations. These neural approaches offer superior predictive power on large datasets but can suffer from overfitting and reduced interpretability when evaluated with few entities and many attributes.

B Estimator Comparison

Effectiveness of Different CDMs. Table 4 compares MCF, NeuralCD, and RCD on held-out response prediction. We hold out 20% of the entries of each response matrix, fit each estimator on the remaining 80%, and predict the held-out entries. MCF achieves the highest accuracy and AUC and the lowest RMSE on both benchmarks. We do not include classical parametric CDMs such as DINA (De La Torre, 2009), DINO (Templin & Henson, 2006), and G-DINA (De La Torre, 2011). These models estimate a distribution over 2^K latent attribute profiles, which amounts to 2^{70} and 2^{41} profiles for CPA-KQA and CFA-KQA, while our setting provides only 20 examinees (the evaluated LLMs). Their parameters therefore cannot be reliably estimated, and a comparison would not be informative. NeuralCD and RCD represent proficiency with continuous embeddings and can be fitted in this setting.

Table 4: Held-out response prediction of the capability estimators. For each benchmark, 20% of the response entries are held out, each estimator is fitted on the remaining 80%, and the metrics are computed on the held-out entries.

	CPA-KQA			CFA-KQA		
Model	**Acc**	**AUC**	**RMSE**	**Acc**	**AUC**	**RMSE**
Matrix Co-Factorization (MCF)	**0.8142**	**0.8949**	**0.3622**	**0.8002**	**0.8653**	**0.3894**
NeuralCD Wang et al. (2022)	0.7035	0.7717	0.4420	0.7080	0.7817	0.4344
RCD Gao et al. (2021)	0.7527	0.8356	0.4083	0.7632	0.8440	0.4019

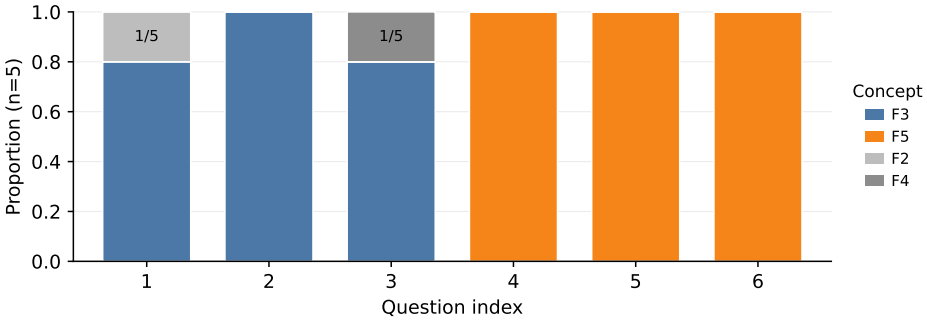


Figure 5: Expert interpretation of six model responses. Five experts independently inferred the capability deficiency reflected in each response, with 0.80 agreement with the F3/F5 profile.

C Qualitative Case Study

We examine Claude 3.5 Sonnet on concepts F3 and F5, using all three test questions associated with each concept. The five case-study experts were independent of the three dataset constructors and held academic backgrounds in finance or auditing. Each expert reviewed a question together with the model's complete response, without access to the correctness label, original concept annotation, or FinCDM output. They independently determined whether the response reflected a capability deficiency and identified the financial knowledge involved. Their inferred deficiencies agree with the F3/F5 profile at a rate of 0.80, with deviations to F2 or F4 on two questions (Figure 5). These six responses are part of the response matrix used by FinCDM, so this case study provides a focused in-sample interpretation of the profile.

D Guidelines for Creating and Annotating CPA-KQA and CFA-KQA Items

D.1 Principles and Requirements for Item Creation

Consistency with true Exam Style. Each assessment item must align closely with real-world CPA and CFA examinations in terms of wording, question format, and option setting, reflecting their rigorous standards for professional knowledge.

Originality and Accuracy. All items must be original creations. Direct copying or quoting from existing CPA and CFA exam items or publicly available materials is strictly prohibited. Ensure accuracy, clarity, and absence of ambiguity in each item.

Specificity and Clarity. Each item should clearly assess its corresponding financial concept. Avoid incorporating content irrelevant to that concept to ensure accurate and effective evaluation.

Avoidance of Politically Sensitive Content. Strictly avoid incorporating politically sensitive topics, events, or individuals in any assessment items. Content that may be viewed as politically controversial or sensitive is explicitly prohibited.

Item Formats. Create 18 items per concept in Single-choice questions. Provide clearly distinguishable options according to the specific exam standards: four options (A, B, C, D) for CPA and three options (A, B, C) for CFA, with exactly one correct answer.

D.2 Required Format for Item Submission

Submit each item strictly following this format:

- **Concept**: *Name of the Concept*
- **Question Stem**: *Clearly and completely stated options:*
  For CPA: A. ... B. ... C. ... D. ...
  For CFA: A. ... B. ... C. ...
- **Correct Answer**: *Option letter.*

D.3 Annotation and Quality Review Procedures

Experts will create and annotate items as follows:

- Each expert independently creates items for their assigned concepts.
- Upon completion, two other experts annotate each item on two dimensions:
  Quality Annotation: Mark as either "Usable" or "Unusable". Criteria for "Unusable" include unclear meaning, inaccurate wording, calculation errors, ambiguous answers, or controversial content.
  Concept Consistency Annotation: Verify if the item accurately corresponds to the specified concept.
- Items marked "Unusable" or "Inconsistent" by both reviewers are deleted directly.
- Items marked by only one reviewer require one-time revision and must pass re-review.
- Deleted items must be replaced until each concept has 18 approved items, which are then split into three test items and 15 training items.

D.4 Annotation Process

Figure 6 shows the Label Studio interface used in this process, where reviewers read each item and record their quality judgment.

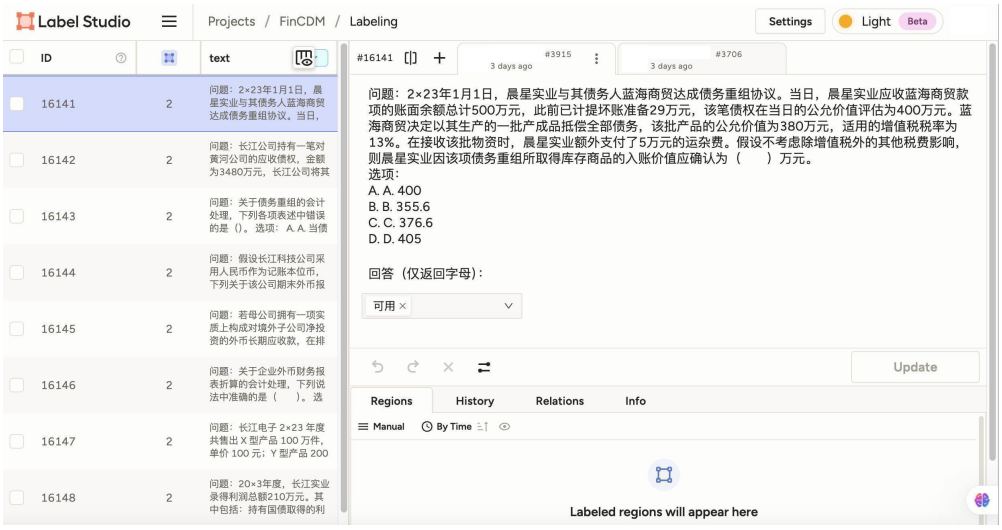


Figure 6: The Label Studio interface of the FinCDM annotation process.

E Sensitivity Analysis

E.1 Question-count sensitivity

We vary the number of questions per concept, n_q, from one to five. The test set contains three questions per concept, so the profiles with $n_q = 4$ and $n_q = 5$ draw the additional questions of the same concept from the training pool. This analysis is run separately from the SkillTune experiments. Figure 7 reports the resulting stability curves.

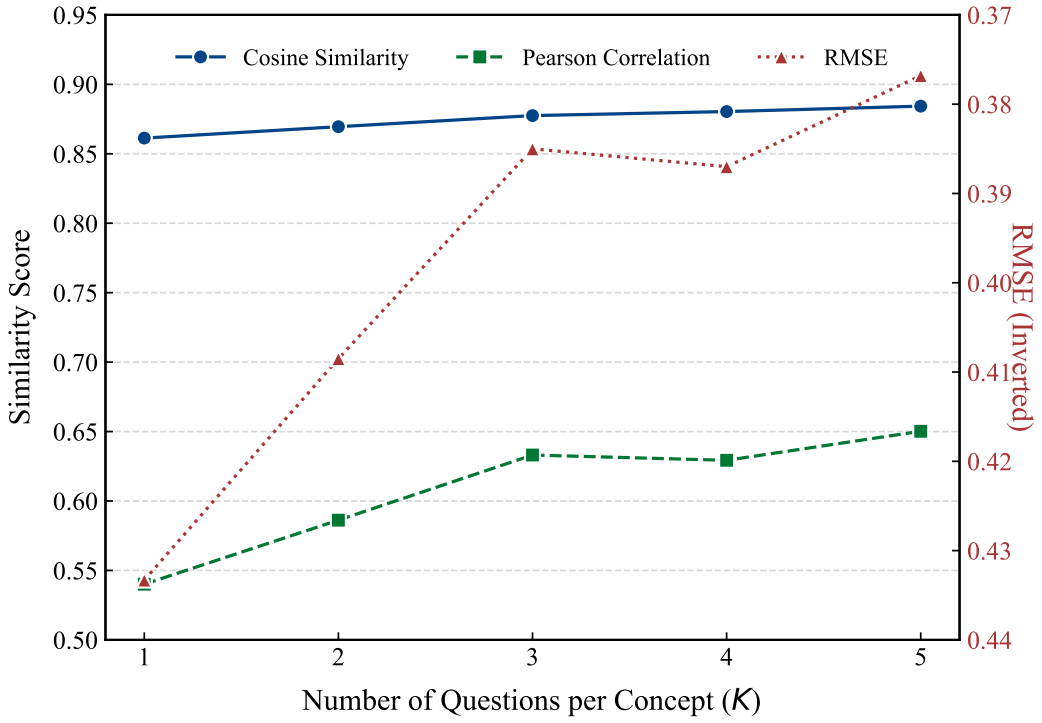


Figure 7: Diagnostic stability as the number of questions per concept (n_q, shown as K on the horizontal axis) varies. Cosine similarity and Pearson correlation increase from one to three questions, and RMSE decreases. RMSE is plotted on an inverted right axis, so higher points indicate lower error. The curves change little beyond three questions.

E.2 THRESHOLD SENSITIVITY

We vary τ from 0.85 to 0.95 in steps of 0.01 and recompute Con for the 20 LLMs on CPA-KQA and CFA-KQA. Table 5 reports the Kendall and Spearman rank correlations between the Con ranking at each threshold and the ranking at $\tau = 0.9$. Across this range, the Spearman correlation is at least 0.98 on CPA-KQA and 0.96 on CFA-KQA, and the Kendall correlation is at least 0.94 and 0.91, respectively. Rank changes occur mainly among models whose Con values differ by one or two concepts. For example, on CFA-KQA the six models ranked fifth to tenth at $\tau = 0.9$ have Con values of 21 or 22.

Table 5: Rank correlations between the Con ranking of the 20 LLMs at each threshold τ and the ranking at $\tau = 0.9$.

	CPA-KQA		CFA-KQA	
τ	Kendall	Spearman	Kendall	Spearman
0.85	0.944	0.986	0.970	0.991
0.86	0.957	0.990	0.970	0.991
0.87	0.976	0.996	0.984	0.996
0.88	0.979	0.996	0.984	0.996
0.89	0.987	0.998	0.994	0.999
0.90	1.000	1.000	1.000	1.000
0.91	0.997	1.000	0.978	0.991
0.92	0.997	1.000	0.972	0.990
0.93	0.992	0.998	0.947	0.980
0.94	0.981	0.997	0.942	0.978
0.95	0.968	0.995	0.911	0.962

F OPTIMIZATION PROCESS FOR NON-NEGATIVE MATRIX CO-FACTORIZATION

We initialize each entry of E, U, and V independently from $\mathcal{U}[0.01, 1.01)$, following the common positive random initialization used in NMF. All reported decompositions use random seed 42. We then optimize Equation 4 using the following multiplicative update rules:

$$E \leftarrow E \circ \frac{(W \circ X)U^\top + QV^\top}{(W \circ EU)U^\top + EVV^\top},$$

$$U \leftarrow U \circ \frac{E^\top (W \circ X)}{E^\top (W \circ EU)},$$

$$V \leftarrow V \circ \frac{E^\top Q}{E^\top EV},$$

where the symbol $\circ$ denotes element-wise multiplication, and division operations are also element-wise. We run 200 updates for every decomposition and use the resulting factors.

G SUMMARY OF THE DATA VOLUME OF EACH DATASET

Summary	
Dataset	**Number**
CPA-KQA-test	210
CPA-KQA-training	1050
CFA-KQA-test	123
CFA-KQA-training	615

Table 6: The specific data volumes of CPA-KQA-test, CPA-KQA-training, CFA-KQA-test and CFA-KQA-training

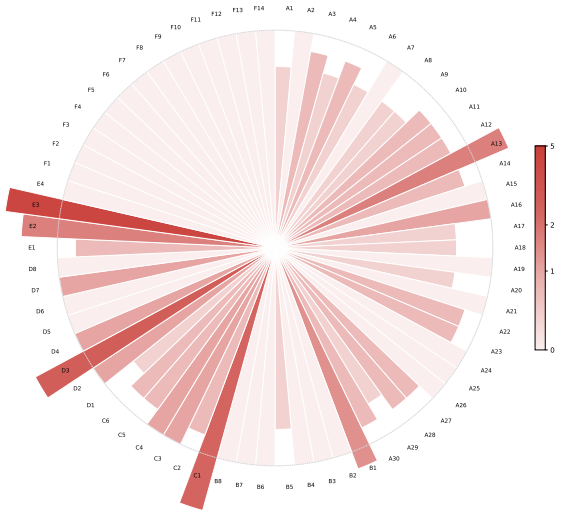


Figure 8: Structure of FinEval-KQA. Darker blocks and longer radial bars indicate concepts assessed by more questions; white blocks indicate concepts with no questions.

H LLMs Details

Category	Model	Parameters	License	Version Date
Open-Source	GLM-4 9B	9B	MIT	2025.4
	DeepSeek V3	671B (MoE)	MIT	2024.12
	Qwen 2.5 7B Instruct	7B	Apache 2.0	2024.9
	Qwen 3 0.6B	0.6B	Apache 2.0	2025.5
	Qwen 3 235B	235B (MoE)	Apache 2.0	2025.5
	Qwen 3 8B	8B	Apache 2.0	2025.5
	Llama 3.1 8B	8B	Llama 3.1 Community License	2024.7
	Llama 3 70B	70B	Llama 3 Community License	2024.4
	Baichuan2 13B Chat	13B	Apache 2.0	2023.12
	Falcon 7B	7B	Apache 2.0	2023.5
Closed-Source	GPT-4o	undisclosed	Proprietary	2024.5
	GPT-4	undisclosed	Proprietary	2023.3
	Gemini 1.5 Pro	undisclosed	Proprietary	2024.5
	Gemini 2.5 Pro Exp	undisclosed	Proprietary	2025.3
	Claude 3.5 Sonnet	undisclosed	Proprietary	2024.6
	Claude 3.7 Sonnet	undisclosed	Proprietary	2025.2
	Grok 3	undisclosed	Proprietary	2025.2
	Qwen Max	undisclosed	Proprietary	2025.1
Financial	Finma 7B Full	7B	MIT (based on LLaMA)	2023.9
	Fin-o1 8B	8B	Custom / Research License	2025.2

Table 7: Models evaluated in this paper. The "License" column indicates the licensing terms of the model. The "Version Date" column shows the release date of the corresponding version we evaluated.

I Prompts

We show our prompts used in the evaluation framework, as shown in Figures 9 and 10.

Dataset	Chinese Original	English Translation
CPA-QKA Dataset	请仔细分析以下问题，并从选项中选择最合适的答案，仅返回选项对应的字母（A、B、C 或 D）。你可以根据自己的推理和理解自由选择答案。 问题：在评估华融实业有限公司 2024 年度各类税费（增值税、消费税及企业所得税）会计处理的合规性时，下列说法中，不符合《企业会计准则》及相关税法规定的是 选项： A. 发生印花税纳税义务时，应将“税金及附加”科目借记，“应交税费”科目贷记 B. 企业收到退回多交的土地增值税,应借记“银行存款”等科目,贷记“税金及附加”科目 C. 车船税应按法规规定的适用税额计算缴纳，计算应缴金额时，借记“税金及附加”科目，贷记“应交税费——应交车船税”科目 D. 企业在计算并缴纳耕地占用税时，应将“在建工程”科目借记、“银行存款”科目贷记 回答（仅返回字母）：	Please analyze the following question carefully and select the most appropriate answer from the options provided. Return only the letter corresponding to your answer (A, B, C, or D). You may rely on your own reasoning and understanding to make your choice. Question: In evaluating the compliance of Huarong Industrial Co., Ltd.'s accounting treatment for various taxes and fees (including value-added tax, consumption tax, and corporate income tax) in the 2024 fiscal year, which of the following statements is not in accordance with the Enterprise Accounting Standards and relevant tax regulations? Options: A. When stamp duty liability arises, the "Taxes and Surcharges" account should be debited, and the "Taxes Payable" account should be credited. B. When a company receives a refund for overpaid land value-added tax, it should debit the "Bank Deposits" account and credit the "Taxes and Surcharges" account. C. Vehicle and vessel tax should be calculated and paid in accordance with applicable statutory rates. When determining the payable amount, the "Taxes and Surcharges" account should be debited, and the "Taxes Payable – Vehicle and Vessel Tax Payable" account should be credited. D. When calculating and paying farmland occupation tax, the "Construction in Progress" account should be debited, and the "Bank Deposits" account credited. Answer (return only the letter):

Figure 9: Prompt for multiple-choice questions in Certified Public Accountant. For better readability, the English translation is displayed to the right of the corresponding Chinese text.

Dataset	Query
CFA-KQA Dataset	Please analyze the following question carefully and select the most appropriate answer from the options. Return only the letter corresponding to the option. You can choose the answer based on your own reasoning and understanding. Question: Eleanor Vance, CFA, serves as an equity analyst at Birchwood Associates. She is currently drafting a research note on Synthetic Solutions, SA, a publicly listed French firm that follows IFRS 9 standards. Vance has aggregated details from the footnotes of Synthetic's 2018 Annual Report and through direct communication with the firm's management to assess potential impacts on future earnings. Exhibit 1 displays details regarding Synthetic's investment portfolio for the fiscal years ending 31 December 2017 and 2018. Vance is analyzing how the accounting classification of these fixed-income securities influences the company's reported financial position. Exhibit 1: Synthetic's Investment Holdings (€ Thousands) | Characteristic | Cornet AG | Meridian Corp | Hugo SA | | :--- | :--- | :--- | :--- | | Classification | FVPL | FVOCI | Amortized cost | | Cost* | €25,000 | €40,000 | €50,000 | | Market value, 31 December 2017 | 29,000 | 38,000 | 54,000 | | Market value, 31 December 2018 | 28,000 | 37,000 | 55,000 | *All securities were purchased at par value. The company's report also notes a transaction involving the securitization of receivables via a special purpose entity (SPE). If one of the securities listed in Exhibit 1 had been reclassified as a Fair Value Through Profit or Loss (FVPL) security, which reclassification would have resulted in a higher carrying value on the balance sheet as of 31 December 2018? Options: A. Cornet AG B. Meridian Corp C. Hugo SA Answer (only letter):

Figure 10: Prompt for multiple-choice questions in Chartered Financial Analyst.

J Broader Impacts and Ethics Statement

This study was conducted in accordance with established ethical guidelines for research. CPA-KQA and CFA-KQA consist of original questions authored by domain experts referenced against public CPA and CFA examination standards. The datasets contain no personally identifiable information (PII) or proprietary financial data. Human participation in this study is limited to domain experts: the experts who created and annotated CPA-KQA and CFA-KQA, and the five experts who judged model responses in the qualitative case study (Appendix C). All experts were fairly compensated and consented to the use of their annotations and judgments, and no personal information about them is collected or released.

Regarding broader impacts, FinCDM enhances the transparency and traceability of financial LLMs, mitigating the risks of deploying models that rely on hallucinated knowledge in high-stakes domains. However, a potential negative societal impact is that users might over-rely on the diagnostic results to substitute human professional judgment. To mitigate this risk, we ensure that the content adheres to professional financial standards and explicitly position FinCDM as an auxiliary research and diagnostic framework rather than a definitive tool for generating financial or legal advice. The diagnostic code and fine-tuning scripts will be released under the MIT license, and the CPA-KQA and CFA-KQA datasets are released under CC BY 4.0. The authors assume no responsibility for any consequences arising from the use of these resources.

K Implementation Details

Matrix co-factorization. We fit Equation 4 with latent rank $r = 6$ for CPA-KQA, $r = 10$ for CFA-KQA, and $r = 5$ for FinEval-KQA. Each entry of E, U, and V is initialized from $\mathcal{U}[0.01, 1.01)$ with random seed 42, and every decomposition runs 200 multiplicative updates (Appendix F). For fine-tuned models, E and V are kept fixed and only their columns of U are estimated, so that scores before and after fine-tuning share the same latent space.

SkillTune fine-tuning. All fine-tuning runs use LoRA with rank 32, scaling factor $\alpha = 64$, and dropout 0.1, applied to the `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj` modules. We train with AdamW at a learning rate of 4.0×10^{-5} under a cosine schedule with 15% warmup, a batch size of 32, and 4 epochs. The instruction input x contains the question and its options, the target y is the correct option letter (A, B, C, or D), and the loss is computed on assistant-response tokens. Fine-tuning uses random seed 42. All data selection strategies use this configuration.

Decoding and answer extraction. Responses are sampled with temperature 1.0, top-p 0.95, and at most 64 output tokens. The predicted answer is extracted using the regular expression `(A|B|C|D)`, taking the first matched option letter; responses with no valid match are marked as invalid.

L Licenses for Existing Assets

This paper uses the following existing assets. We confirm that all assets are used in accordance with their respective licenses and terms of use.

Datasets.

- **FinEval** (Guo et al., 2025): A Chinese financial domain benchmark. Available at `https://github.com/SUFE-AIFLM-Lab/FinEval`. Licensed under **Apache-2.0**.
- **CPA-KQA** and **CFA-KQA**: Original datasets constructed by the authors of this paper. All questions were created by domain experts recruited by our team and are not derived from any third-party copyrighted materials. These datasets will be released under **CC BY 4.0** upon publication.

Models.

- Details of all 20 LLMs evaluated in this paper, including their licenses and terms of use, are provided in Appendix H.