Beyond Multiple Choice: How Answer Options Inflate LLM Financial Reasoning Scores

Wei-Lun Cheng[a], Daniel Wei-Chung Miao[a,*], Guang-Di Chang[a]

[a] *Graduate Institute of Finance, National Taiwan University of Science and Technology, Taipei, 10607, Taiwan*

Abstract

Current evaluations of Large Language Models (LLMs) on financial benchmarks rely almost exclusively on multiple-choice question (MCQ) formats, yet MCQ options themselves leak information—magnitude clues, sign directions, and elimination opportunities—that inflate perceived reasoning ability. We present two complementary experiments using CFA (Chartered Financial Analyst) examination questions. First, we measure *option bias* by testing GPT-4o-mini on the same 100 questions with and without answer options, finding that MCQ format inflates accuracy by **12.0 percentage points** (85.0% with options vs. 73.0% without); McNemar's test confirms this difference is statistically significant ($p = 0.045$). Second, we apply a *three-tier evaluation framework*—distinguishing exact matches, directionally correct responses, and genuine errors—to open-ended answers, revealing that **28.0% of responses** classified as "incorrect" under binary scoring are actually directionally correct (Level B), employing valid financial reasoning with different assumptions. The strict (exact match) accuracy of 34.0% contrasts sharply with the lenient (including directionally correct) accuracy of 62.0%, exposing a 28-percentage-point gap attributable to the inherent ambiguity of financial calculations. Our findings suggest that current MCQ-based benchmarks systematically overstate LLM financial competence and that the binary correct/incorrect paradigm fails to capture the nuanced nature of financial reasoning.

*Corresponding author

Email addresses: d11018003@mail.ntust.edu.tw (Wei-Lun Cheng), miao@mail.ntust.edu.tw (Daniel Wei-Chung Miao), gchang@mail.ntust.edu.tw (Guang-Di Chang)

Keywords: Large Language Models, Financial Reasoning, Multiple Choice Bias, Open-Ended Evaluation, CFA Examination, Benchmark Design

1. Introduction

The rapid deployment of Large Language Models (LLMs) in financial services has been accompanied by a proliferation of benchmark evaluations. Models are now routinely tested on professional certification exams—CFA, CPA, Bar Exam—with headlines proclaiming that AI can "pass" these tests [1, 4]. These evaluations almost universally employ the multiple-choice question (MCQ) format, reporting a single accuracy number that serves as the basis for deployment decisions.

However, the MCQ format itself introduces a systematic measurement artifact. In classical test theory, this is known as *answer-space restriction bias*: the set of answer options constrains the response space, providing information beyond what the examinee actually knows [5]. For LLMs, this bias manifests in three specific mechanisms:

1. **Magnitude clues**: Options reveal the order of magnitude of the answer (e.g., options of \$1.2M, \$2.4M, \$4.8M, \$9.6M constrain the answer's range).
2. **Sign clues**: Options reveal whether the answer is positive or negative, narrowing computational search.
3. **Elimination opportunities**: LLMs can reject implausible options without computing the exact answer, using heuristics rather than reasoning.

The combined effect is that MCQ accuracy overstates the model's genuine financial reasoning ability. But *by how much?* And when we remove options, what does the model's reasoning actually look like? Is a wrong answer always a "failure," or can a model use correct reasoning with different (but equally valid) financial assumptions?

This paper addresses these questions through two complementary experiments:

1. **Option Bias Quantification (A5)**: We test the same model on the same CFA questions in two formats—MCQ (with options) and open-ended (without options)—and measure the accuracy gap attributable to option-derived information leakage.

2. **Three-Tier Open-Ended Evaluation (A1)**: We evaluate open-ended responses using a three-tier framework that distinguishes exact matches (Level A), directionally correct responses with different assumptions (Level B), and genuinely incorrect answers (Level C).

Our contributions are fourfold:

1. We quantify the MCQ option bias in financial LLM evaluation at +12.0 percentage points, demonstrating that current benchmarks systematically overstate reasoning ability.
2. We introduce a three-tier evaluation framework that accommodates the inherent ambiguity of financial calculations, revealing that 28% of "errors" are actually valid alternative analyses.
3. We decompose errors into structured categories (formula error, calculation error, conceptual error), enabling targeted diagnosis of financial reasoning weaknesses.
4. We argue for a paradigm shift in financial LLM evaluation: from binary MCQ accuracy to nuanced, open-ended assessment that better reflects real-world financial analysis.

2. Related Work

2.1. LLM Evaluation on Professional Examinations

Evaluating LLMs on professional examinations has become standard practice. Callanan et al. [1] tested GPT-4 on CFA Level I, finding pass-rate performance. Ke et al. [4] developed FinDAP, achieving state-of-the-art CFA results. However, all such evaluations use the MCQ format, leaving open the question of how much performance is format-dependent.

2.2. MCQ Format Bias in AI Evaluation

The limitations of MCQ evaluation for AI systems have received growing attention. Gao et al. [3] demonstrate that LLMs exploit MCQ-specific strategies (option anchoring, elimination) that inflate accuracy beyond genuine understanding. Robinson et al. [6] analyze how answer option statistics in training data enable shortcut learning. Our work extends this literature to the financial domain, where the consequences of overestimated AI competence are particularly severe.

2.3. Open-Ended Evaluation of Mathematical Reasoning

Open-ended evaluation removes the "crutch" of answer options, requiring models to generate answers from scratch. Cobbe et al. [2] use this approach for mathematical reasoning, finding substantial accuracy drops compared to multiple-choice equivalents. Our three-tier framework goes further by acknowledging that financial calculations involve legitimate ambiguity (compounding conventions, day-count conventions, rounding policies) that binary scoring fails to capture.

3. Methodology

3.1. Option Bias Measurement

Each CFA question is presented to the same model in two formats:

- **Format A (MCQ)**: Standard format with answer options (A, B, C). The model selects an option letter.
- **Format B (Open-ended)**: Options removed; the model generates a free-form answer.

The *option bias* is defined as:

$$\text{Option Bias} = \text{Acc}_{\text{MCQ}} - \text{Acc}_{\text{open-ended}} \tag{1}$$

Positive values indicate that options inflate accuracy. We use McNemar's test on the paired observations (same question, two formats) to assess statistical significance.

For open-ended answers, evaluation uses a combination of:

- **Numerical tolerance matching**: $|a_{\text{model}} - a_{\text{gold}}|/|a_{\text{gold}}| \leq 0.02$ for exact match
- **Semantic matching**: LLM-as-judge (GPT-4o-mini) for conceptual/textual answers

3.2. Three-Tier Evaluation Framework

We replace binary scoring with a three-tier classification:

Level A — Exact/Acceptable Match. The answer falls within 2% relative tolerance of the gold answer, or the semantic judge confirms equivalence. This is the "strict" correct.

Level B — Directionally Correct. The answer demonstrates correct reasoning approach (correct formula, correct direction, correct order of magnitude) but arrives at a different final value due to alternative assumptions (e.g., different compounding convention, different day-count method, different rounding). Under binary scoring, this would be "incorrect"; under our framework, it is a legitimate alternative analysis.

Level C — Genuinely Incorrect. The answer reflects a fundamental error: wrong formula, wrong concept, logical fallacy, or computational mistake.

We report both *strict accuracy* (Level A only) and *lenient accuracy* (Level A + B), arguing that the gap between them measures the inherent ambiguity of financial evaluation.

3.3. Structured Error Attribution

For Level C responses, we classify errors into categories using LLM-as-judge:

- `formula_error`: Selected the wrong formula or financial model
- `calculation_error`: Correct formula but arithmetic mistake
- `conceptual_error`: Fundamental misunderstanding of the financial concept
- `assumption_mismatch`: Used invalid or inappropriate assumptions
- `extraction_error`: Misread or misinterpreted the question data
- `incomplete_reasoning`: Correct approach but failed to complete all steps

4. Data and Experimental Design

We use 100 questions from the CFA-Easy dataset [4]. The model is GPT-4o-mini (OpenAI) at temperature $\tau = 0.0$. Each question is evaluated in both MCQ and open-ended format, yielding 200 inferences for option bias analysis plus 100 additional inferences for three-tier evaluation.

5. Results

5.1. Option Bias

Table 1 presents the core option bias findings.

Table 1: Option Bias Results (GPT-4o-mini, $n = 100$)

Metric	**Value**	**Interpretation**
Accuracy WITH options (MCQ)	85.0%	Standard benchmark score
Accuracy WITHOUT options	73.0%	True reasoning ability
Option bias	**+12.0 pp**	Format-inflated performance
Biased questions (MCQ ✓, Open ×)	21/100 (21.0%)	Questions where options are a "crutch"
McNemar's p-value	0.045	Significant at $\alpha = 0.05$

Option bias = $\text{Accuracy}_{\text{MCQ}} - \text{Accuracy}_{\text{open-ended}}$. Biased questions are those answered correctly with options but incorrectly without.

MCQ format inflates accuracy by 12 percentage points. In 21% of questions, the model answers correctly *only* when options are provided—suggesting that option-derived information (magnitude clues, elimination strategies) is essential for these responses. McNemar's test confirms that the option bias is statistically significant ($b = 21$, $c = 9$, $\chi^2 = 4.033$, $p = 0.045$), indicating that the 12-percentage-point inflation is not merely a practical observation but a reliable measurement artifact of the MCQ format.

5.2. Three-Tier Evaluation

Table 2 presents the three-tier evaluation of open-ended responses.

The results reveal a striking discrepancy: strict accuracy (34.0%) is less than half of what the MCQ format (85.0%) would suggest, while lenient accuracy (62.0%) is much closer. This means:

Table 2: Three-Tier Evaluation of Open-Ended Responses ($n = 100$)

Level	Count	Percentage	Description
Level A (Exact)	34	34.0%	Correct within 2% tolerance
Level B (Directional)	28	28.0%	Right approach, different assumptions
Level C (Incorrect)	38	38.0%	Genuine error
Strict Accuracy (A only)		**34.0%**	
Lenient Accuracy (A+B)		**62.0%**	

The 28-percentage-point gap between strict and lenient accuracy reflects the inherent ambiguity of financial calculations.

- **34% of responses** are exactly correct—the model demonstrably understands the problem.
- **28% of responses** use valid reasoning but arrive at different answers due to alternative financial conventions—these are not "errors" in a meaningful sense.
- **38% of responses** are genuinely incorrect—reflecting real limitations in financial reasoning.

5.3. Error Attribution

Among the 38 Level C responses, error attribution reveals:

- `conceptual_error`: 21/38 (55.3%)—fundamental misunderstanding of the financial concept
- `unknown`: 7/38 (18.4%)—error type could not be automatically classified
- `arithmetic_error`: 4/38 (10.5%)—correct formula, wrong calculation
- `incomplete_reasoning`: 2/38 (5.3%)—correct approach but stopped too early
- `reading_error`: 2/38 (5.3%)—misread or misinterpreted the question data
- `formula_error`: 1/38 (2.6%)—wrong financial model selected

- `assumption_error`: 1/38 (2.6%)—used invalid or inappropriate assumptions

Conceptual errors dominate, accounting for over half of all genuine errors. This suggests that the primary bottleneck is not arithmetic (which LLMs handle reasonably well) but *selecting the right financial concept or framework* for the given problem. The relatively low formula error rate (2.6%) indicates that when the model identifies the correct conceptual domain, it generally applies the right formula—but it frequently misjudges which domain applies.

6. Discussion

6.1. Economic Significance: The Option Bias Tax

The 12-percentage-point option bias represents a systematic overestimation of AI financial competence. For an institution evaluating AI tools based on MCQ benchmarks:

- A reported accuracy of 85% suggests the AI handles roughly 5 out of 6 financial analyses correctly.
- The true open-ended accuracy of 73% means it handles only roughly 3 out of 4 correctly.
- The additional "correct" responses are artifacts of format-provided shortcuts, not genuine reasoning.

This "option bias tax" is paid when the AI encounters real-world financial analysis—which never comes with multiple-choice options. Institutions deploying AI based on inflated MCQ scores face a reliability gap that manifests as unexpected errors in production.

6.2. The Hidden Competence Problem

The three-tier framework reveals a complementary insight: binary scoring *underestimates* competence on the other end. The 28% Level B rate means that over a quarter of "wrong" answers are actually reasonable alternative analyses using valid financial reasoning with different assumptions. In practical terms:

- An AI that computes a bond yield of 6.32% (using continuous compounding) when the gold answer is 6.45% (using semi-annual compounding) is not "wrong"—it is using a different but legitimate convention.

- A compliance reviewer who flags this as an error wastes review capacity; an AI evaluation that counts it as incorrect underestimates the model's financial competence.

The lenient accuracy (62.0%) is thus a more realistic measure of the model's financial understanding than either the inflated MCQ score (85.0%) or the overly strict open-ended score (34.0%).

6.3. Implications for CFA Exam Design

Our findings have direct implications for the CFA Institute:

1. **AI vulnerability**: The 21% biased question rate identifies questions where AI can "game" the MCQ format. These items should be reviewed for question quality.
2. **Format innovation**: As AI capabilities advance, the CFA Institute should explore partial-credit scoring and open-ended formats for future exam iterations.
3. **Convention sensitivity**: Level B responses highlight the need for clearer specification of computational conventions in exam questions, reducing ambiguity.

6.4. Limitations

Our study uses $n = 100$ questions from the CFA-Easy dataset and a single model (GPT-4o-mini). McNemar's test confirms the option bias is statistically significant ($p = 0.045$). Extension to the full CFA-Easy corpus ($n = 1{,}032$) and multiple models would strengthen generalizability. The LLM-as-judge approach for three-tier classification may itself contain biases, and human validation of a subset of classifications is recommended for future work.

7. Conclusion

This paper demonstrates that MCQ-format evaluations of financial LLMs are simultaneously too generous and too strict. Too generous: answer options inflate accuracy by providing format-derived shortcuts (option bias = +12%). Too strict: binary scoring classifies reasonable alternative analyses as "errors" (28% of "incorrect" answers are directionally correct).

We propose a shift in financial LLM evaluation: from MCQ accuracy to open-ended assessment with three-tier scoring. This approach provides a more realistic picture of AI financial competence—one that acknowledges both the inflation from format shortcuts and the ambiguity inherent in financial calculations.

The question is not whether AI can choose the right option, but whether it can reason to the right answer. Our evidence suggests a meaningful gap between these two abilities.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT Authorship Contribution Statement

Wei-Lun Cheng: Conceptualization, Methodology, Software, Formal Analysis, Data Curation, Writing – Original Draft, Visualization. **Daniel Wei-Chung Miao**: Supervision, Writing – Review & Editing. **Guang-Di Chang**: Supervision, Writing – Review & Editing.

Acknowledgments

The authors thank the anonymous reviewers for their constructive feedback.

Data Availability

The CFA-Easy dataset is available via HuggingFace under the FinEval benchmark [4]. Experiment code is available from the corresponding author upon reasonable request.

References

[1] Callanan, E., Mbae, A., Selle, S., Gupta, V., & Houlihan, R. (2023). Can GPT-4 pass the CFA exam? *arXiv preprint arXiv:2310.09542*.

[2] Cobbe, K., Kosaraju, V., Bavarian, M., et al. (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

[3] Gao, J., Guo, C., Zhang, Y., et al. (2024). Are LLMs good at multiple choice questions? A benchmark for MCQ evaluation. *arXiv preprint*.

[4] Ke, Z., Ming, Y., Nguyen, X. P., Xiong, C., & Joty, S. (2025). Demystifying domain-adaptive post-training for financial LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[5] Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates.

[6] Robinson, J., Sloane, C., Liang, P., & Tenenbaum, J. (2023). Leveraging large language models for multiple choice question answering. *arXiv preprint arXiv:2210.12353*.

[7] Wu, S., Irsoy, O., Lu, S., et al. (2023). BloombergGPT: A large language model for finance. *arXiv preprint arXiv:2303.17564*.