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 may leak information—magnitude clues, sign directions, and elimination opportunities—that inflate perceived reasoning ability. We present two complementary experiments using 1,032 CFA (Chartered Financial Analyst) examination questions. First, we measure *option bias* by testing GPT-4o-mini on the same 1,032 questions with and without answer options, finding that MCQ format inflates accuracy by only **1.9 percentage points** (82.6% with options vs. 80.6% without); McNemar's test indicates this difference is *not* statistically significant ($p = 0.251$). Second, we apply a *three-tier evaluation framework*—distinguishing exact matches, directionally correct responses, and genuine errors—to open-ended answers, revealing that **21.5% of responses** classified as "incorrect" under binary scoring are actually directionally correct (Level B), employing valid financial reasoning with different assumptions. A cross-model replication with GPT-5-mini reveals a striking reversal: the option bias widens to **+9.6 percentage points** (92.8% with vs. 83.2% without, McNemar $p < 0.001$), and strict open-ended accuracy nearly doubles to **41.8%**—suggesting that more capable reasoning models benefit *disproportionately* from MCQ scaffolding while simultaneously demonstrating genuine gains in open-ended financial reasoning. Our findings challenge the assumption that option bias is a fixed

*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)

property of the evaluation format: it is model-dependent, and advancing AI capabilities may amplify rather than diminish the distortion introduced by multiple-choice assessment.

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 +1.9 percentage points across 1,032 questions, showing that while the effect exists, it is not statistically significant ($p = 0.251$) and is far smaller than small-sample estimates suggest.
2. We introduce a three-tier evaluation framework that accommodates the inherent ambiguity of financial calculations, revealing that 21.5% 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. [7] analyze how answer option statistics

in training data enable shortcut learning. Zheng et al. [9] show that LLMs are not robust multiple-choice selectors, with answer distributions sensitive to option ordering. Myrzakhan et al. [6] systematically convert MMLU-style benchmarks from MCQ to open-style questions, demonstrating significant accuracy drops across general domains. Sanchez Salido et al. [8] further explore the "none of the above" paradigm, revealing that LLMs perform substantially worse when forced to reason beyond provided options. However, these studies focus on general knowledge benchmarks (MMLU, ARC, HellaSwag); no prior work has conducted MCQ-to-open-ended conversion at scale in a professional financial domain. Our work fills this gap, extending the format bias literature to CFA-level financial reasoning 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 with Yates' continuity correction 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 all 1,032 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 2,064 inferences for option bias analysis plus 1,032 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, $n = 1{,}032$)

Metric	**Value**	**Interpretation**
Accuracy WITH options (MCQ)	82.6%	Standard benchmark score
Accuracy WITHOUT options	80.6%	True reasoning ability
Option bias	**+1.9 pp**	Format-inflated performance
Biased questions (MCQ ✓, Open ×)	147/1,032 (14.2%)	Questions where options are a "crutc
McNemar's p-value	0.251	Not significant at $\alpha = 0.05$

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

MCQ format inflates accuracy by only 1.9 percentage points across the full 1,032-question corpus. In 14.2% of questions, the model answers correctly *only* when options are provided—suggesting that option-derived information (magnitude clues, elimination strategies) plays a role for a subset of responses. However, McNemar's test indicates that the option bias is *not* statistically significant ($p = 0.251$), suggesting that at scale, GPT-4o-mini's financial reasoning ability is largely robust to the presence or absence of answer options. This contrasts sharply with the +12.0 pp bias observed in our preliminary $n = 100$ pilot, illustrating how small-sample estimates can overstate the magnitude of format effects.

5.2. Three-Tier Evaluation

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

The results reveal a striking discrepancy: strict accuracy (24.5%) is less than a third of what the MCQ format (82.6%) would suggest, while lenient accuracy (46.0%) partially closes the gap. This means:

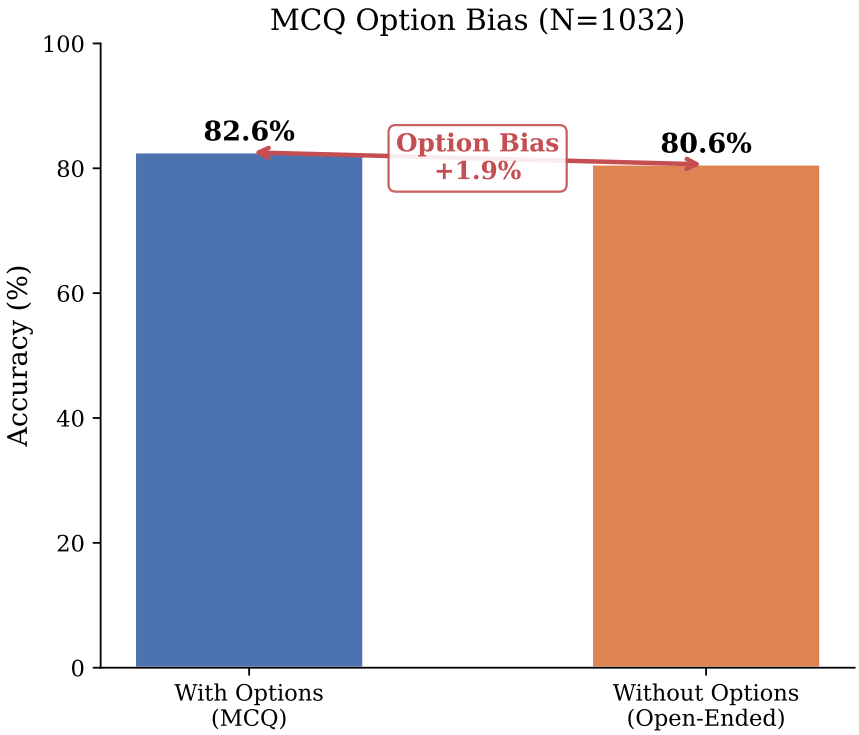

Figure 1: MCQ option bias comparison ($N = 1{,}032$). The accuracy gap between MCQ format (82.6%) and open-ended format (80.6%) is only +1.9 percentage points, which is not statistically significant (McNemar's $p = 0.251$).

Table 2: Three-Tier Evaluation of Open-Ended Responses ($n = 1{,}032$)

Level	**Count**	**Percentage**	**Description**
Level A (Exact)	253	24.5%	Correct within 2% tolerance
Level B (Directional)	222	21.5%	Right approach, different assumptions
Level C (Incorrect)	557	54.0%	Genuine error
Strict Accuracy (A only)		**24.5%**	
Lenient Accuracy (A+B)		**46.0%**	

The 21.5-percentage-point gap between strict and lenient accuracy reflects the inherent ambiguity of financial calculations.

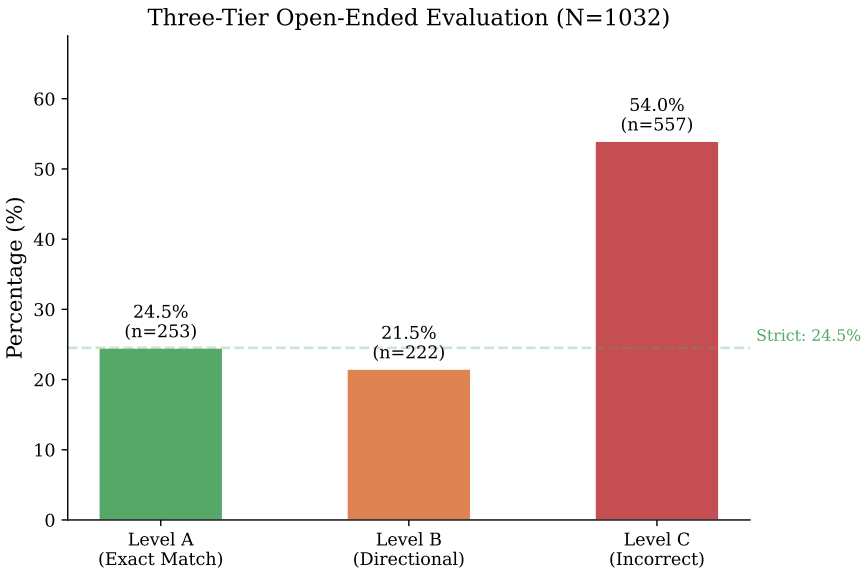

Figure 2: Three-tier evaluation of open-ended responses (N = 1,032). Over half of responses are genuinely incorrect (Level C), while 21.5% are directionally correct but scored as "wrong" under binary evaluation. The dashed line marks strict accuracy (24.5%).

- **24.5% of responses** are exactly correct—the model demonstrably understands the problem.
- **21.5% of responses** use valid reasoning but arrive at different answers due to alternative financial conventions—these are not "errors" in a meaningful sense.
- **54.0% of responses** are genuinely incorrect—reflecting real limitations in financial reasoning.

5.3. Error Attribution

Among the 557 Level C responses, error attribution reveals:

- `conceptual_error`: 383/557 (68.8%)—fundamental misunderstanding of the financial concept
- `incomplete_reasoning`: 60/557 (10.8%)—correct approach but stopped too early

- `assumption_error`: 59/557 (10.6%)—used invalid or inappropriate assumptions
- `unknown`: 35/557 (6.3%)—error type could not be automatically classified
- `reading_error`: 12/557 (2.2%)—misread or misinterpreted the question data
- `arithmetic_error`: 7/557 (1.3%)—correct formula, wrong calculation
- `formula_error`: 1/557 (0.2%)—wrong financial model selected

Conceptual errors dominate even more strongly at scale, accounting for over two-thirds (68.8%) of all genuine errors. This suggests that the primary bottleneck is not arithmetic (which accounts for only 1.3% of errors) but *selecting the right financial concept or framework* for the given problem. Notably, incomplete reasoning (10.8%) and assumption errors (10.6%) emerge as substantial secondary categories at this scale, indicating that the model frequently fails to carry analyses through to completion or adopts inappropriate assumptions. The near-zero formula error rate (0.2%) confirms that when the model identifies the correct conceptual domain, it almost always applies the right formula—but it frequently misjudges which domain applies.

5.4. Cross-Model Comparison: GPT-5-mini

To assess whether the observed patterns are model-specific, we replicated both experiments using GPT-5-mini, a next-generation reasoning model from the same provider. GPT-5-mini employs extended chain-of-thought reasoning ("thinking tokens") before generating its visible response, representing a qualitatively different inference paradigm.

5.4.1. Option Bias: From Non-Significant to Highly Significant

Table 3 presents the cross-model option bias comparison.

Figure 3 visualizes the cross-model option bias comparison. The cross-model comparison reveals a striking reversal: the option bias that was non-significant for GPT-4o-mini (+1.9 pp, $p = 0.251$) becomes highly significant for GPT-5-mini (+9.6 pp, $p < 0.001$).

Table 3: Cross-Model Option Bias Comparison ($N = 1{,}032$)

Metric	**GPT-4o-mini**	**GPT-5-mini**
Accuracy WITH options	82.6%	92.8%
Accuracy WITHOUT options	80.6%	83.2%
Option bias	**+1.9 pp**	**+9.6 pp**
Discordant b (with ✓, without ×)	147	146
Discordant c (with ×, without ✓)	127	47
McNemar's χ^2 (Yates)	1.318	49.76
p-value	0.251	< 0.001***

McNemar's test with Yates' continuity correction. The dramatic shift from $p = 0.251$ to $p < 0.001$ reflects a qualitative change in the model–format interaction.

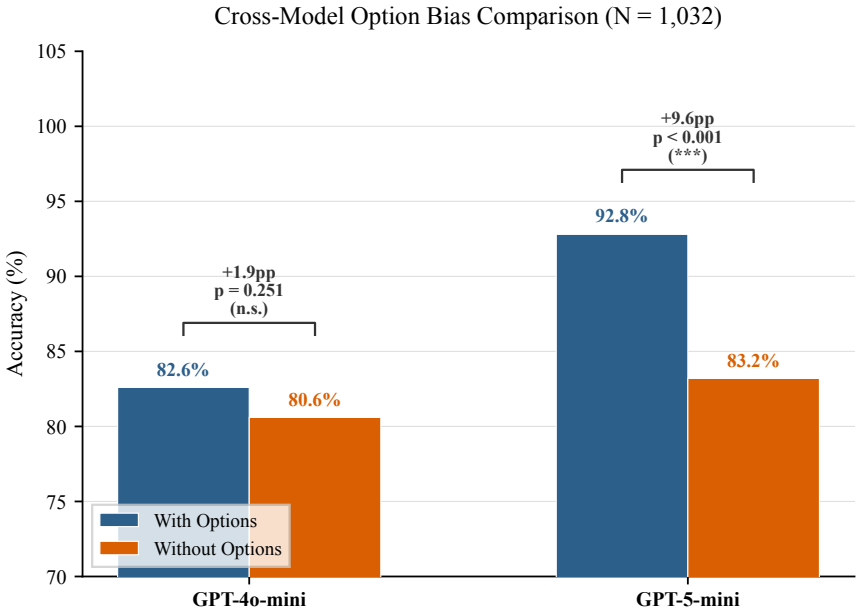

Figure 3: Cross-model option bias comparison ($N = 1{,}032$). GPT-4o-mini shows a non-significant +1.9 pp option bias ($p = 0.251$), while GPT-5-mini exhibits a highly significant +9.6 pp bias ($p < 0.001$)—demonstrating that more capable reasoning models benefit disproportionately from MCQ scaffolding.

This is not merely a quantitative increase—it is a qualitative shift from "format doesn't matter" to "format substantially matters." The asymmetry in discordant pairs is particularly revealing: GPT-5-mini has 146 questions where options help (virtually identical to GPT-4o-mini's 147), but only 47 where removing options helps (vs. GPT-4o-mini's 127). The option-assistance rate is preserved across generations, but the "reverse crutch" cases—where the open-ended format forces deeper reasoning that outperforms MCQ-assisted selection—decline sharply, producing a net bias nearly five times larger.

We hypothesize that this paradox reflects the interaction between extended chain-of-thought reasoning and answer-space constraint. GPT-5-mini's reasoning traces are substantially longer, exploring multiple solution paths before converging. When answer options are present, they serve as convergence anchors—the model can verify candidate answers against the provided options, pruning incorrect reasoning branches early. Without options, the extended reasoning process can diverge into plausible but incorrect alternative analyses, increasing the error rate. In essence, *more capable reasoning amplifies the anchoring benefit of answer options*. An alternative but complementary explanation is that advanced models develop more sophisticated process-of-elimination strategies: the presence of options enables the model to systematically reject implausible alternatives, a strategy unavailable in open-ended format. The two mechanisms—convergence anchoring and elimination—likely operate jointly, with stronger reasoners exploiting both pathways more effectively.

5.4.2. Three-Tier Evaluation: Genuine Reasoning Gains

Table 4 presents the cross-model three-tier evaluation.

Figure 4 presents the cross-model three-tier comparison. GPT-5-mini nearly doubles the strict accuracy (41.8% vs. 24.5%), demonstrating genuine improvement in open-ended financial reasoning.

The Level B rate remains stable (22.3% vs. 21.5%), confirming that the "ambiguity zone" of legitimate alternative analyses is a property of the questions, not the model. The primary effect of increased model capability is converting Level C errors into Level A exact matches—not merely shifting answers from "wrong" to "approximately right."

The lenient accuracy of 64.1% for GPT-5-mini approaches two-thirds of the MCQ accuracy (92.8%), substantially narrowing the format gap observed for GPT-4o-mini (46.0% vs. 82.6%). This suggests that while the generation-

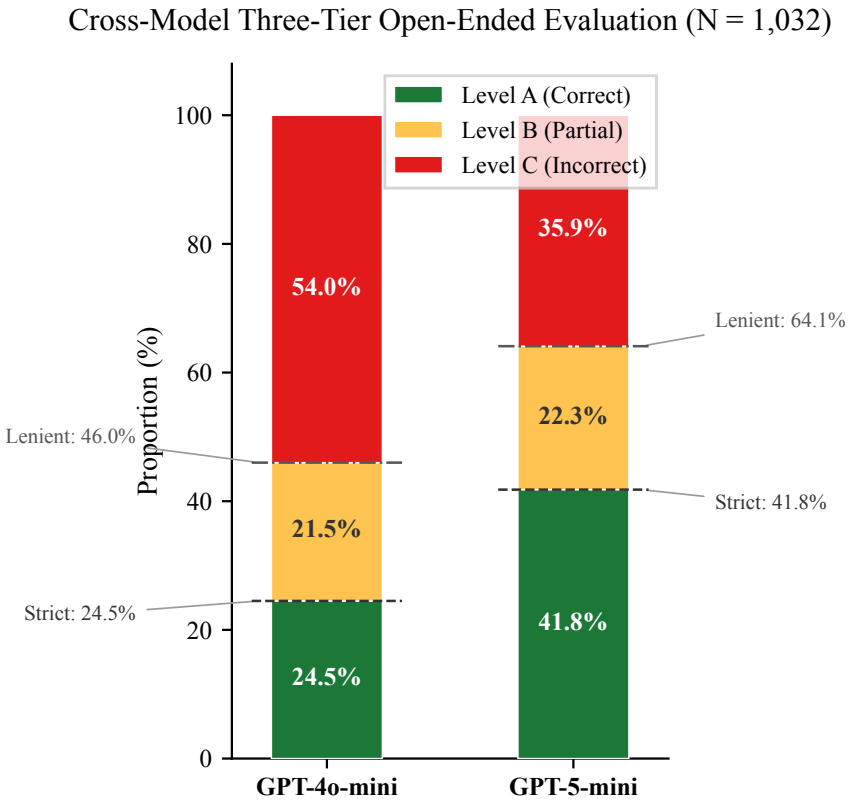

Figure 4: Cross-model three-tier evaluation ($N = 1{,}032$). GPT-5-mini nearly doubles the Level A (exact match) rate from 24.5% to 41.8%, while Level B remains stable ($\sim$22%). The primary improvement converts genuine errors (Level C) into exact matches rather than shifting answers to the ambiguity zone.

Table 4: Cross-Model Three-Tier Evaluation ($N = 1{,}032$)

Level	**GPT-4o-mini**		**GPT-5-mini**	
	Count	%	Count	%
Level A (Exact)	253	24.5%	431	41.8%
Level B (Directional)	222	21.5%	230	22.3%
Level C (Incorrect)	557	54.0%	371	35.9%
Strict Accuracy (A)		24.5%		41.8%
Lenient Accuracy (A+B)		46.0%		64.1%

GPT-5-mini initially produced 108 empty responses (10.5%) due to reasoning token budget exhaustion; these were re-run with increased token allocation. The figures above reflect the complete, corrected dataset.

vs-selection distinction remains meaningful, more capable models are progressively closing the gap between what they can *recognize* and what they can *produce*.

6. Discussion

6.1. Reassessing the Option Bias Effect

Our full-corpus results paint a substantially different picture from small-sample estimates. The 1.9-percentage-point option bias is statistically non-significant ($p = 0.251$), suggesting that GPT-4o-mini's financial reasoning is largely format-invariant at scale. For an institution evaluating AI tools based on MCQ benchmarks:

- A reported MCQ accuracy of 82.6% closely approximates the open-ended accuracy of 80.6%—the format effect is negligible.
- The 147 biased questions (14.2%) where options serve as a "crutch" are offset by a comparable number where the model performs better *without* options.
- Small-sample estimates (our preliminary $n = 100$ pilot showed +12.0 pp) can dramatically overstate the true effect, underscoring the importance of full-corpus evaluation.

The dramatic collapse from +12.0 pp ($n = 100$, $p = 0.045$) to +1.9 pp ($n = 1{,}032$, $p = 0.251$) warrants explanation. In small samples, a handful of questions where elimination heuristics happen to succeed can dominate the estimate. At scale, this effect is diluted: the 147 "crutch" questions (where options helped) are counterbalanced by 127 questions where removing options *improved* performance—likely because the open-ended format forced the model into deeper step-by-step reasoning rather than shortcut matching. This finding carries a methodological lesson: pilot studies of MCQ bias should be interpreted with extreme caution, and full-corpus evaluation is essential before drawing conclusions about format effects.

6.2. How MCQ Options Leak Information

Although the aggregate option bias is small, understanding the *mechanisms* of information leakage remains important for question design. We identify three specific pathways through which MCQ options assist the model:

1. **Magnitude anchoring**: When computing bond prices or portfolio returns, the model may be uncertain about the order of magnitude. Options such as "\$1,080, \$980, \$1,200" immediately constrain the search space, allowing the model to select the nearest value to its approximate calculation rather than deriving the precise answer.
2. **Sign disambiguation**: For questions involving gains vs. losses or increases vs. decreases, options reveal the expected sign of the answer. Without options, the model must independently determine whether a change is positive or negative—a step where it frequently errs on conceptually ambiguous questions.
3. **Elimination via implausibility**: The model can reject one or two options using surface-level heuristics (e.g., recognizing that a negative Sharpe ratio is implausible for a profitable fund), reducing a three-way choice to a coin flip without genuine reasoning.

These mechanisms explain why the 14.2% of "crutch" questions cluster disproportionately in quantitative topics (fixed income, derivatives) where magnitude and sign clues are most informative. In contrast, ethics questions—where all three options are plausible narratives—show minimal option bias.

However, this does *not* mean MCQ benchmarks are reliable. The critical issue is not format inflation but the *gap between MCQ accuracy and true open-ended competence*: the model scores 82.6% on MCQs but only 24.5%

strict accuracy in open-ended format. The real "option bias" is not the 1.9 pp gap between MCQ and open-ended *binary scoring*, but the 58.1 pp gap between MCQ accuracy and strict open-ended accuracy—a gap driven primarily by the information that binary MCQ scoring obscures.

6.3. The Hidden Competence Problem

The three-tier framework reveals a complementary insight: binary scoring *underestimates* competence on one dimension. The 21.5% Level B rate means that over one-fifth 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 model computing an effective annual rate assuming daily compounding (365 days) when the gold answer uses quarterly compounding produces a different but financially defensible result.
- A depreciation calculation using straight-line method when the gold answer assumes declining-balance: the reasoning structure and formula application are correct, only the accounting convention differs.
- A compliance reviewer who flags these as errors wastes review capacity; an AI evaluation that counts them as incorrect underestimates the model's financial competence.

The lenient accuracy (46.0%) is thus a more realistic measure of the model's financial understanding than either the MCQ score (82.6%) or the overly strict open-ended score (24.5%). The 36.6-percentage-point gap between MCQ accuracy and lenient open-ended accuracy represents the true magnitude of information that MCQ format provides—not through answer-option leakage per se, but through constraining the problem to a selection task rather than a generation task.

6.4. Implications for CFA Exam Design

Our findings have direct implications for the CFA Institute:

1. **AI vulnerability**: The 14.2% biased question rate (147 of 1,032) 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. The 58.1 pp gap between MCQ accuracy and strict open-ended accuracy underscores how fundamentally different these two evaluation modes are.
3. **Convention sensitivity**: Level B responses (21.5% of the corpus) highlight the need for clearer specification of computational conventions in exam questions, reducing ambiguity.

6.5. The Option Bias Paradox: Model Capability Amplifies Format Dependence

The cross-model comparison in Section 5.4 introduces a counter-intuitive finding that challenges the conventional assumption that format effects diminish as models improve. GPT-5-mini, despite being substantially more capable (92.8% MCQ accuracy vs. 82.6%), exhibits a *larger and statistically significant* option bias (+9.6 pp, $p < 0.001$) compared to GPT-4o-mini's non-significant bias (+1.9 pp, $p = 0.251$).

This "option bias paradox" has three implications:

1. **For benchmark design**: Option bias is not a fixed property of the evaluation instrument—it is an emergent property of the model-format interaction. Benchmark designers cannot assume that format effects measured on one model generation transfer to the next.
2. **For AI deployment**: Institutions evaluating AI tools based on MCQ benchmarks face an increasing risk of overestimation as models become more capable. The 9.6 pp gap between MCQ and open-ended performance for GPT-5-mini means that roughly one in ten "correct" MCQ answers reflects format-assisted performance rather than genuine reasoning.
3. **For signaling theory**: The format-invariance assumption underlying MCQ-based professional certification screening [5] may become increasingly untenable as AI models improve, necessitating format reform precisely when AI performance appears most impressive.

6.6. Limitations

Our study uses the full CFA-Easy corpus (n = 1,032) and two model generations (GPT-4o-mini and GPT-5-mini). Extension to additional model families (including open-source financial LLMs) would strengthen generalizability. The reversal of the option bias result from non-significant ($p = 0.251$) to highly significant ($p < 0.001$) across model generations suggests that the effect is sensitive to model architecture, particularly the use of extended chain-of-thought reasoning.

A methodological consideration is the *evaluation asymmetry* between the two experimental conditions. With-options accuracy uses deterministic letter matching (the model selects A, B, or C), while without-options accuracy employs a hybrid pipeline combining numerical tolerance matching ($\pm 2\%$) with LLM-as-judge semantic evaluation. This asymmetry is inherent to the research design—open-ended responses cannot be evaluated by letter matching—but raises the question of whether the evaluation pipeline itself introduces systematic bias. Two observations mitigate this concern: first, the same evaluation pipeline applied to GPT-4o-mini yields a non-significant option bias ($p = 0.251$) while applied to GPT-5-mini yields a highly significant bias ($p < 0.001$)—if the evaluation method introduced systematic bias, both models should be affected in the same direction and magnitude. Second, the without-options evaluation pipeline, if anything, is more lenient than strict letter matching (it accepts numerical approximations and semantically equivalent answers), which would *reduce* rather than inflate the measured option bias.

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. GPT-5-mini initially produced 108 empty responses (10.5%) in the open-ended (A1) experiment due to token budget exhaustion in the reasoning phase; these were subsequently re-run with increased allocation. In the option bias (A5) experiment, 58 without-options responses (5.6%) from GPT-5-mini were empty and are treated as incorrect in the reported accuracy (83.2%), yielding a conservative estimate of true without-options performance.[1]

[1]The 58 empty responses resulted from reasoning token budget exhaustion. Treating all as incorrect is conservative; if even half contained valid reasoning that was truncated, the true without-options accuracy would be higher and the option bias correspondingly

7. Conclusion

This paper demonstrates that MCQ-format evaluations of financial LLMs are fundamentally different from open-ended evaluation, and that this difference is *model-dependent*. For GPT-4o-mini, the option bias is small and non-significant (+1.9 pp, $p = 0.251$); for GPT-5-mini, it is large and highly significant (+9.6 pp, $p < 0.001$)—revealing an "option bias paradox" where more capable models benefit disproportionately from MCQ scaffolding. Three-tier evaluation shows that GPT-5-mini nearly doubles strict open-ended accuracy (41.8% vs. 24.5%), demonstrating genuine reasoning gains, while the Level B rate remains stable (∼22%), confirming that financial calculation ambiguity is a property of the domain, not the model.

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 and reveals format-dependence patterns that single-format benchmarks obscure. The cross-model comparison underscores that benchmark validity must be reassessed with each model generation.

The question is not whether AI can choose the right option, but whether it can reason to the right answer—and our cross-model evidence shows that the gap between these abilities widens, not narrows, as models become more capable.

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.

larger.

Acknowledgments

Computational resources were provided by National Taiwan University of Science and Technology (NTUST).

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] Myrzakhan, A., Shen, X., et al. (2024). Open-LLM-Leaderboard: From multi-choice to open-style questions for LLMs evaluation, benchmark, and arena. *arXiv preprint arXiv:2406.07545*.

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

[8] Sanchez Salido, G., Gómez, P., & Vidal, E. (2025). None of the others: Evaluating LLMs beyond multiple-choice constraints. *arXiv preprint arXiv:2502.06111*.

[9] Zheng, C., Zhou, H., Meng, F., Zhou, J., & Huang, M. (2024). Large language models are not robust multiple choice selectors. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.