Stress Testing Financial LLMs: Counterfactual Perturbation and Noise Sensitivity Analysis on CFA Examinations

Wei-Lun Cheng[a], Wei-Chung Miao[b,*]

[a] *Institute of Information Science, Academia Sinica, Taipei, Taiwan*
[b] *Department of Finance, National Chengchi University, Taipei, Taiwan*

Abstract

Large Language Models (LLMs) achieve impressive accuracy on financial examination benchmarks, but these scores may be inflated by memorization of training data rather than genuine financial reasoning. We introduce two complementary stress tests for financial LLMs using CFA (Chartered Financial Analyst) examination questions. First, *counterfactual perturbation* modifies numerical parameters and conditions while preserving the underlying financial logic, measuring whether models can consistently reason through novel variants. Second, *noise injection* introduces irrelevant data, misleading statements, format noise, and contradictory information, measuring whether models can filter signal from noise—a critical skill in real-world financial analysis. Testing GPT-4o-mini on 100 CFA questions with two perturbation levels and four noise types, we find: (1) a memorization gap of **13.1–23.5 percentage points** between original and perturbed accuracy, suggesting substantial reliance on pattern matching; (2) noise sensitivity indices ranging from **−0.012 to 0.046**, with irrelevant data injection (N1) causing the greatest degradation; and (3) robust accuracy—requiring correctness on both original and all variants—of only **58.0%**, substantially below the **86.0%** standard accuracy. We introduce *Robust Accuracy* as a more realistic measure of AI financial competence and argue that financial regulators should require stress-tested performance metrics alongside standard accuracy when evaluating AI deploy-

[*] Corresponding author

Email addresses: wlcheng@iis.sinica.edu.tw (Wei-Lun Cheng), wcmiao@nccu.edu.tw (Wei-Chung Miao)

ment readiness.

Keywords: Large Language Models, Financial Reasoning, Stress Testing, Robustness, CFA Examination, Memorization

1. Introduction

The financial industry is rapidly adopting Large Language Models (LLMs) for tasks including equity research, risk analysis, regulatory compliance, and client advisory [8, 5]. Benchmark evaluations show that state-of-the-art models can pass the CFA examination with scores approaching or exceeding human pass rates [3]. These impressive results have accelerated deployment timelines, with firms increasingly relying on LLM-generated analysis for consequential financial decisions.

However, a fundamental question remains largely unexamined: *do these models genuinely understand financial logic, or have they memorized patterns from extensively available exam preparation materials?* CFA exam questions—sourced primarily from SchweserNotes, Kaplan, and AnalystPrep—are widely distributed across the internet and likely well-represented in LLM training corpora. The CFA question space is structurally narrow: a limited set of financial concepts, stereotypical numerical patterns (e.g., 5% coupon rate, $1,000 face value, 10-year maturity), and fixed problem templates. This creates ideal conditions for rote memorization to masquerade as genuine reasoning.

The distinction between memorization and reasoning has profound implications for financial practice. Consider two scenarios:

- **Reasoning AI**: Correctly computes bond duration for any combination of coupon rate, maturity, and yield—including combinations never seen in training data.
- **Memorizing AI**: Achieves high accuracy on standard questions but fails when numerical parameters or problem conditions are changed, because its “understanding” is pattern matching against memorized templates.

A portfolio manager using the Memorizing AI faces a dangerous illusion of competence: the system performs well on familiar calculations but fails unpredictably on novel ones—precisely the situations where AI assistance is most valuable.

This paper proposes a *stress testing framework* for financial LLMs, drawing directly from established risk management methodology. Just as banks stress test capital adequacy under adverse scenarios (Basel III, CCAR/DFAST), we stress test AI cognitive adequacy under adversarial perturbations. Our framework comprises two complementary dimensions:

1. **Counterfactual Perturbation (I1)**: We modify numerical parameters and problem conditions in CFA questions while preserving the underlying financial logic. If a model truly reasons, its accuracy should be preserved under perturbation. The *memorization gap*—the difference between original and perturbed accuracy—quantifies the extent of pattern-matching reliance.
2. **Noise Injection (I3)**: We inject irrelevant data, misleading statements, format inconsistencies, and contradictory information into questions. Real-world financial analysis requires filtering signal from noise—a skill untested by clean benchmark questions. The *Noise Sensitivity Index* (NSI) measures how much noise degrades performance.

We introduce *Robust Accuracy*—requiring correctness on both the original question and all stress-tested variants—as a more realistic measure of AI financial competence. Our key insight is that standard benchmark accuracy overstates the practical reliability of financial LLMs, and that robustness metrics should be reported alongside accuracy for any AI system deployed in financial contexts.

This paper makes four contributions:

1. We design and validate a two-dimensional stress testing framework for financial LLMs combining counterfactual perturbation and noise injection.
2. We quantify the memorization gap in financial LLMs, providing evidence that a meaningful portion of benchmark performance derives from pattern matching rather than reasoning.
3. We measure noise sensitivity across four distinct noise types, identifying which types of real-world information noise most degrade AI financial analysis.
4. We propose Robust Accuracy as a regulatory-relevant metric and argue that financial AI deployment should require stress-tested performance reporting analogous to bank stress testing requirements.

2. Related Work

2.1. LLMs in Financial Applications

The intersection of LLMs and finance has attracted significant research attention. BloombergGPT [8] demonstrated competitive performance on financial NLP tasks. Ke et al. [5] introduced FinDAP, achieving state-of-the-art results on CFA benchmarks through domain-adaptive post-training. Callanan et al. [3] evaluated GPT-4 on CFA Level I, finding pass-rate performance. However, these evaluations assess accuracy on standard questions without examining whether performance reflects genuine understanding.

2.2. Data Contamination and Benchmark Validity

The threat of data contamination in LLM evaluations is well-documented [6]. Apple [1] demonstrated that LLMs show significant accuracy degradation when mathematical reasoning problems are symbolically perturbed—changing variable names and numerical values while preserving logical structure—suggesting that high benchmark scores partly reflect memorization. Our work extends this methodology to the financial domain, where the contamination risk is arguably higher due to the narrow and widely-distributed nature of CFA exam materials.

2.3. Robustness and Adversarial Testing

Jia & Liang [4] pioneered adversarial examples for reading comprehension, demonstrating that adding irrelevant sentences to passages dramatically reduces model accuracy. Subsequent work has developed comprehensive robustness benchmarks for NLP systems [7]. In finance, Black [2] established the theoretical importance of distinguishing signal from noise. Our noise injection framework operationalizes this distinction for AI evaluation, creating a financial-domain-specific robustness test.

2.4. Stress Testing in Financial Regulation

Stress testing is a cornerstone of financial regulation. The Basel III framework requires banks to demonstrate capital adequacy under adverse macroeconomic scenarios. The Federal Reserve's Comprehensive Capital Analysis and Review (CCAR) evaluates whether banks can continue lending during severe recessions. We draw a direct analogy: if financial institutions must stress test their capital models, they should also stress test their AI models. Our framework provides the methodology.

3. Methodology

3.1. Counterfactual Perturbation Design

We employ a multi-level perturbation scheme inspired by Apple [1]:

Level 1 — Numerical Perturbation. We modify one numerical parameter in the question (e.g., interest rate, face value, maturity period) while preserving the solution procedure. The correct answer changes accordingly, but the required formula and reasoning steps remain identical. This tests whether the model can re-compute answers with new inputs or is anchored to memorized values.

Using GPT-4o-mini as a perturbation generator, each original question produces a perturbed variant with:

- A clearly identified changed parameter and its new value
- The correct answer for the perturbed version
- Verification that the perturbation preserves the question's logical structure

Level 2 — Conditional Inversion. We change the problem's structural conditions (e.g., annual → continuous compounding, call → put option, long → short position). This requires the model to select a different formula or adjust its reasoning direction—a more demanding test of genuine understanding.

3.2. Noise Injection Design

We define four noise types that model progressively more challenging real-world information environments:

- **N1 — Irrelevant Data Injection:** Insert numerical data points unrelated to the solution (e.g., company founding year, employee count, ESG score in a bond pricing question). The model must identify and ignore this information.
- **N2 — Misleading Financial Distractors:** Insert plausible-sounding but irrelevant financial statements (e.g., "According to consensus estimates, sector growth is expected to be 15%" in a historical portfolio return calculation). These compete with relevant information for the model's attention.

- **N3 — Format Noise:** Introduce formatting inconsistencies mimicking real-world documents: mixed number formats ($1,000 vs. 1000.00), redundant text, and structural variations.

- **N4 — Contradictory Information:** Introduce data that contradicts other parts of the question, requiring the model to identify and resolve the inconsistency.

Noise is generated programmatically using financial domain-specific templates, with intensity controlled by the number of noise elements injected.

3.3. Evaluation Metrics

Memorization Gap. For counterfactual perturbation:

$$\text{Memorization Gap}_\ell = \text{Acc}_{\text{original}} - \text{Acc}_{\text{Level } \ell} \tag{1}$$

Positive values indicate reliance on memorized patterns at perturbation level ℓ.

Noise Sensitivity Index. For noise injection:

$$\text{NSI}_t = \frac{\text{Acc}_{\text{clean}} - \text{Acc}_{\text{noisy},t}}{\text{Acc}_{\text{clean}}} \tag{2}$$

where $t \in \{N1, N2, N3, N4\}$. NSI ranges from 0 (noise-immune) to 1 (completely noise-destroyed).

Robust Accuracy. The most conservative metric:

$$\text{Robust Acc} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1} \left[\text{correct}_i^{\text{orig}} \wedge \bigwedge_{\ell} \text{correct}_i^{\text{Level } \ell} \right] \tag{3}$$

A question contributes to Robust Accuracy only if the model answers both the original and *all* perturbation variants correctly. This metric is analogous to “stressed” capital ratios in banking regulation.

Statistical Testing. We use McNemar’s test to assess whether accuracy differences between original and perturbed/noisy conditions are statistically significant, treating each question as a paired observation.

4. Data and Experimental Design

4.1. Dataset

We draw questions from the CFA-Easy dataset from FinEval [5], comprising 1,032 multiple-choice questions covering the full CFA curriculum. We use a random sample of $N = 100$ questions, providing sufficient statistical power to detect meaningful accuracy differences across conditions.

Table 1: Experimental Design Summary

Dimension	**Variant**	**Questions**	**Inferences**
Baseline	Original (clean)	100	100
Perturbation (I1)	Level 1 (numerical)	100	100
	Level 2 (conditional)	100	100
Noise (I3)	N1 (irrelevant data)	100	100
	N2 (misleading)	100	100
	N3 (format noise)	100	100
	N4 (contradictory)	100	100
Total			**700**

4.2. Model

We evaluate GPT-4o-mini (OpenAI), a widely-used commercial model representative of the class of LLMs increasingly deployed in financial applications. All evaluations use temperature $\tau = 0.0$ for deterministic outputs. Answers are extracted using a five-layer regex chain with fallback parsing.

5. Results

5.1. Counterfactual Perturbation Results

Table 2 presents the core counterfactual perturbation findings.

The memorization gap of +23.5% at Level 1 indicates that approximately one-quarter of the model's standard accuracy is attributable to numerical pattern matching rather than genuine financial reasoning. Notably, Level 2 (conditional perturbation) shows a smaller gap of +13.1%, though this level had a higher valid perturbation rate (85 vs. 64 out of 100), suggesting

Table 2: Counterfactual Perturbation Results (GPT-4o-mini, $N = 100$)

Condition	N Valid	Accuracy	Mem. Gap	Δ	Direction
Original	100	86.0%	—	—	—
Level 1 (numerical)	64	62.5%	+23.5%	$\downarrow$	Memorization
Level 2 (conditional)	85	72.9%	+13.1%	$\downarrow$	Memorization
Robust Accuracy	100	58.0%	—	—	—

Mem. Gap = $\text{Accuracy}_{\text{original}} - \text{Accuracy}_{\text{perturbed}}$. Robust Accuracy requires correct answers on original *and* all valid perturbations.

that questions amenable to valid conditional perturbation may be inherently easier for the model to reason through.

The Robust Accuracy of 58.0%—compared to 86.0% standard accuracy—reveals that only about two-thirds of the model's apparent performance is genuinely robust. The remaining 28.0 percentage points represent brittle performance: correct on the memorized version but failing under perturbation. This "memorization suspect" rate of 28.0% suggests that more than a quarter of the model's benchmark success may be attributable to pattern matching.

5.2. Noise Sensitivity Results

Table 3 presents noise injection findings across four noise types.

Table 3: Noise Sensitivity Results (GPT-4o-mini, $N = 100$)

Noise Type	Noisy Acc.	Flipped	NSI	Interpretation
Clean (baseline)	86.0%	—	—	—
N1 (irrelevant data)	82.0%	7/100	0.046	Low
N2 (misleading)	85.0%	5/100	0.012	Minimal
N3 (format noise)	85.0%	3/100	0.012	Minimal
N4 (contradictory)	87.0%	3/100	−0.012	None

NSI = Noise Sensitivity Index = $(\text{Acc}_{\text{clean}} - \text{Acc}_{\text{noisy}}) / \text{Acc}_{\text{clean}}$. Flipped = questions correct when clean but incorrect with noise.

The results reveal that noise sensitivity at scale is considerably lower than initial small-sample estimates suggested. N1 (irrelevant data injection) produces the highest sensitivity (NSI = 0.046), indicating that even clearly irrelevant data causes some confusion, though the effect is modest. N2 (misleading financial distractors) and N3 (format noise) both show minimal sensitivity (NSI = 0.012), while N4 (contradictory information) actually yields a slightly negative NSI (-0.012), suggesting that contradictory hints may occasionally prompt more careful reasoning.

The overall pattern—NSI ranging from -0.012 to 0.046—indicates that GPT-4o-mini is substantially more noise-robust than the counterfactual perturbation results would suggest. This asymmetry is itself an important finding: the model's primary vulnerability lies in memorization-dependent reasoning rather than susceptibility to information noise. Real-world noise degrades performance by at most 4.6%, far less than the 13–24% degradation from counterfactual perturbation.

5.3. Combined Stress Test Framework

Table 4 presents the combined 2×2 analysis integrating both dimensions.

Table 4: Combined Stress Test Results

	Clean	**Worst-Case Noise**
Original	86.0% (Standard)	82.0% (Noise-degraded)
Perturbed	58.0% (Robust)	— (Worst-case)

Standard accuracy (86.0%) is the metric currently reported by all financial LLM benchmarks. Robust accuracy (58.0%) accounts for memorization effects. Noise-degraded accuracy (82.0% for worst-case N1 noise) accounts for real-world information noise. The gap between standard and robust accuracy—28.0 percentage points—represents the "memorization premium" in current benchmarks: inflated performance that evaporates when the AI encounters genuinely novel problems.

5.4. Statistical Significance

We apply McNemar's test to the paired observations (original vs. perturbed for each question):

Table 5: Statistical Tests

Comparison	**Test**	**Statistic**	***p*-value**
Original vs. Level 1	McNemar's	$\chi^2 = 10.56$	0.001**
Original vs. Level 2	McNemar's	$\chi^2 = 6.75$	0.009**
Clean vs. N1 (worst noise)	McNemar's	$\chi^2 = 2.78$	0.096

6. Discussion

6.1. Economic Significance: The Memorization Premium

Our findings reveal a "memorization premium" in financial LLM benchmarks: the gap between standard accuracy and robust accuracy represents performance that is artificially inflated by pattern matching against known question templates. For a financial institution evaluating AI tools:

- **Standard accuracy** (86.0%) suggests the AI correctly handles roughly 6 out of 7 financial calculations.
- **Robust accuracy** (58.0%) reveals it reliably handles only about three out of five.
- The 28.0-percentage-point difference represents a "phantom competence" zone—questions where the AI appears competent but would fail on real-world variants.

In capital allocation terms: if AI-assisted analysis informs investment decisions, the memorization premium means that a fraction of the AI's "correct" outputs arise from memorization artifacts. These will fail unpredictably on novel financial scenarios—precisely when AI assistance is most valuable.

6.2. Sensitivity Analysis Interpretation

Drawing from quantitative finance methodology, our framework can be interpreted through the lens of sensitivity analysis:

- **Delta** (first-order sensitivity): The Level 1 memorization gap measures how accuracy changes with numerical input perturbation, analogous to a financial instrument's delta.

- **Gamma** (convexity): The acceleration of degradation from Level 1 to Level 2 measures the non-linearity of the model's sensitivity, analogous to gamma risk.
- **Vega** (noise sensitivity): NSI measures how accuracy changes with information noise, analogous to an instrument's sensitivity to volatility changes.

This analogy is not merely illustrative—it positions AI robustness assessment within a framework familiar to financial risk managers, facilitating adoption of stress testing practices for AI systems.

6.3. Stress Testing as Due Diligence

CFA Standard V(A)—Diligence and Reasonable Basis—requires that investment professionals have a "reasonable and adequate basis" for their recommendations. We argue that deploying an AI system based solely on standard benchmark accuracy, without stress testing its reasoning, fails this standard. The memorization gap reveals that standard accuracy is an unreliable basis for assessing AI competence.

Our framework provides a practical due diligence tool: before deploying an AI system for financial analysis, institutions should compute its Robust Accuracy and NSI profile. Systems with high memorization gaps or noise sensitivity should receive additional human oversight, particularly for non-routine financial analysis.

6.4. Regulatory Implications

The EU AI Act classifies AI in financial services as "high-risk," requiring providers to demonstrate accuracy and robustness. Our metrics provide concrete, quantifiable criteria:

1. **Memorization Gap Threshold**: Financial AI systems should demonstrate Memorization Gap $< 10\%$, ensuring that performance is not substantially inflated by rote memorization.
2. **Noise Sensitivity Threshold**: NSI < 0.15 across all noise types, ensuring the system can tolerate real-world information noise without significant performance loss.
3. **Robust Accuracy Reporting**: Regulators should require Robust Accuracy alongside standard accuracy, analogous to how banks report both unstressed and stressed capital ratios.

6.5. Limitations

Several limitations should be acknowledged. First, while $N = 100$ provides meaningful statistical power, validation across the complete CFA-Easy corpus (1,032 questions) and multiple models would strengthen generalizability claims. Second, perturbation generation relies on GPT-4o-mini, which may introduce its own errors in generating valid perturbations; the relatively low valid perturbation rate at Level 1 (64 out of 100) reflects this challenge, and we report accuracy only on valid perturbations. Third, our noise types, while domain-informed, are synthetic; real-world financial noise may be more subtle or more severe. Fourth, we test a single model; cross-model comparison is needed to determine whether the observed memorization-noise asymmetry generalizes across model families.

7. Conclusion

This paper demonstrates that standard benchmark accuracy significantly overstates the financial reasoning ability of Large Language Models. Our two-dimensional stress testing framework—combining counterfactual perturbation with noise injection—reveals a meaningful memorization gap between standard and robust accuracy, and significant sensitivity to information noise that is absent from clean benchmark evaluations.

We introduce Robust Accuracy as a more realistic metric for evaluating AI systems intended for financial deployment. Just as banks must demonstrate capital adequacy under stressed conditions, AI systems should demonstrate reasoning adequacy under cognitive stress tests. **The question is not whether AI can pass the CFA exam, but whether it can reason through problems it hasn't memorized.**

Our stress testing framework provides financial institutions and regulators with practical tools for assessing AI deployment readiness. We recommend that any AI system used for financial analysis be evaluated using both standard accuracy and stress-tested metrics before deployment, with minimum robustness thresholds analogous to capital adequacy requirements.

Data Availability

The CFA-Easy dataset is available via HuggingFace under the FinEval benchmark [5]. All experiment code, perturbation generation scripts, noise injection templates, and raw results are available at `https://github.com/[anonymized]`.

References

References

[1] Mirzadeh, I., Alizadeh, K., Shahrokhi, H., et al. (2024). GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229.*

[2] Black, F. (1986). Noise. *The Journal of Finance*, 41(3), 528–543.

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

[4] Jia, R., & Liang, P. (2017). Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2021–2031).

[5] 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).*

[6] Shi, W., Ajith, A., Xia, M., et al. (2023). Detecting pretraining data from large language models. In *Proceedings of the 12th International Conference on Learning Representations (ICLR).*

[7] Wang, B., Xu, C., Wang, S., et al. (2022). Adversarial GLUE: A multitask benchmark for robustness evaluation of language models. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track.*

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