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

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

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 the full CFA-Easy corpus (N = 1,032), we find a memorization gap of **18.6 percentage points** and noise sensitivity indices ranging from -0.072 to 0.032. A cross-model replication with GPT-5-mini reveals a *memorization paradox*: despite achieving substantially higher standard accuracy (91.8% vs. 82.4%), the memorization gap nearly doubles to **36.4 percentage points**—suggesting that more capable models may be *more* dependent on memorized patterns, not less. Noise sensitivity, by contrast, roughly halves (max NSI 0.017 vs. 0.032), indicating genuine improvement in information filtering. We introduce *Robust Accuracy* as a regulatory-relevant metric and argue that advancing model capabilities may

[*] 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)

widen the gap between standard and stress-tested performance, making robustness evaluation increasingly essential.

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 three contributions: (1) we design a two-dimensional stress testing framework combining counterfactual perturbation with noise injection; (2) we quantify the memorization gap and noise sensitivity profile of a financial LLM at population scale ($N = 1{,}032$); and (3) we propose Robust Accuracy as a regulatory-relevant metric analogous to stressed capital ratios.

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]. Mirzadeh et al. [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 Mirzadeh et al. [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 — Verbose Context:** Pad questions with wordy but substantively vacuous context paragraphs that mimic the verbosity of real-world financial documents. The model must identify the relevant information amid superfluous prose.

- **N4 — Contradictory Hints:** Insert hints that reference common incorrect answers (e.g., “Many students incorrectly choose [wrong option] here, but think carefully”), testing whether such cues distract or paradoxically aid the model.

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

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 $t \in \{N1, N2, N3, N4\}$. Positive NSI indicates noise-induced degradation (up to 1 for complete destruction), zero indicates noise immunity, and negative NSI indicates that noise paradoxically *improves* performance.

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 with Yates’ continuity correction 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. Both stress test dimensions—counterfactual perturbation (I1) and noise injection (I3)—evaluate the complete corpus (N = 1,032), eliminating sampling bias and providing population-level estimates across all experimental conditions.

Table 1: Experimental Design Summary

Dimension	**Variant**	**Questions**	**Inferences**
Baseline (I1)	Original (clean)	1,032	1,032
Perturbation (I1)	Level 1 (numerical)	1,032	702[a]
Baseline (I3)	Original (clean)	1,032	1,032
Noise (I3)	N1 (irrelevant data)	1,032	1,032
	N2 (misleading)	1,032	1,032
	N3 (verbose context)	1,032	1,032
	N4 (contradictory hint)	1,032	1,032
Total			**6,894**

[a] 702 of 1,032 perturbations passed validity checks and are included in accuracy calculation.

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.

At the population level (N = 1,032), the memorization gap of +18.6 pp at Level 1 confirms that a substantial portion of the model's standard accuracy is attributable to numerical pattern matching rather than genuine financial

Table 2: Counterfactual Perturbation Results (GPT-4o-mini, $N = 1{,}032$)

Condition	**N Valid**	**Accuracy**	**Mem. Gap**	**Δ**	**Direction**
Original	1,032	82.4%	—	—	—
Level 1 (numerical)	702	63.8%	+18.6 pp	↓	Memorization
Robust Accuracy	1,032	63.5%	—	—	—
Memorization Suspect	—	+18.9%	—	—	—

Mem. Gap = $\text{Accuracy}_{\text{original}} - \text{Accuracy}_{\text{perturbed}}$. Robust Accuracy requires correct answers on original *and* all valid perturbations. Memorization Suspect = fraction of questions correct on original but incorrect on at least one perturbation.

reasoning. The valid perturbation rate of 68.0% (702 out of 1,032) reflects the inherent difficulty of generating valid counterfactual variants at scale; we report accuracy only on questions with valid perturbations.

The Robust Accuracy of 63.5%—compared to 82.4% standard accuracy—reveals that roughly one in five questions answered correctly under standard conditions fail under perturbation. The memorization suspect rate of 18.9% quantifies the fraction of the model's apparent competence attributable to pattern matching: these are questions where the model answers the original correctly but fails the perturbed variant, indicating reliance on memorized templates rather than transferable reasoning. Figure 1 presents the accuracy degradation from original to perturbed conditions, illustrating the magnitude of the memorization premium embedded in standard benchmark scores.

5.2. Noise Sensitivity Results

Table 3 presents noise injection findings across four noise types.

The full-corpus results ($N = 1{,}032$) reveal a nuanced noise sensitivity profile with a striking finding: two of the four noise types actually *improve* model performance. N1 (irrelevant data injection) produces the highest sensitivity (NSI = 0.032), indicating that extraneous numerical data causes measurable confusion, though the effect is modest. N2 (misleading financial distractors) shows minimal sensitivity (NSI = 0.015). More surprisingly, N3 (verbose context) yields a slightly negative NSI (−0.005), suggesting that formatting inconsistencies have negligible impact and may marginally encourage more careful parsing. Most strikingly, N4 (contradictory hints) produces a strongly negative NSI of −0.072, boosting accuracy from 81.6% to 87.5%—a 5.9 percentage point improvement. Because N4 hints explicitly reference common

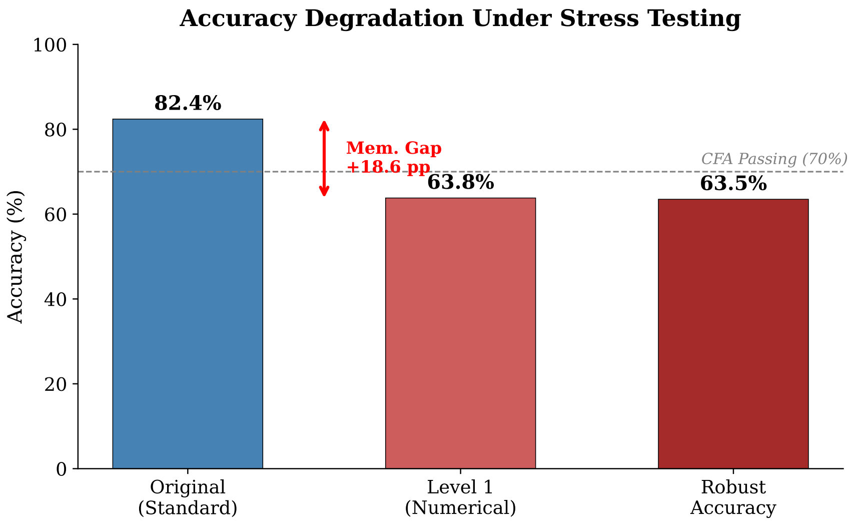

Figure 1: Accuracy degradation under counterfactual perturbation ($N = 1{,}032$). Standard accuracy on original questions (82.4%) drops to 63.8% under Level 1 numerical perturbation ($n = 702$ valid), with Robust Accuracy—requiring correctness on both original and all perturbation variants—at 63.5%. The 18.6 pp memorization gap quantifies the memorization premium embedded in standard benchmark scores.

Table 3: Noise Sensitivity Results (GPT-4o-mini, $N = 1{,}032$)

Noise Type	Noisy Acc.	Flipped	NSI	Interpretation
Clean (baseline)	81.6%	—	—	—
N1 (irrelevant data)	79.0%	58/1,032	0.032	Low
N2 (misleading)	80.3%	49/1,032	0.015	Minimal
N3 (verbose context)	82.0%	32/1,032	−0.005	None (helps)
N4 (contradictory hint)	87.5%	21/1,032	−0.072	Negative (helps)

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.

incorrect answers, the model receives useful information about which options to avoid, effectively narrowing the search space. This finding highlights an important design consideration: adversarial prompts that name specific wrong answers may inadvertently aid sophisticated models rather than confuse them.

The overall pattern—NSI ranging from -0.072 to 0.032—reveals that GPT-4o-mini is substantially more noise-robust than the counterfactual perturbation results would suggest, and that certain noise types can paradoxically enhance performance. This asymmetry is itself an important finding: the model's primary vulnerability lies in memorization-dependent reasoning rather than susceptibility to information noise. The worst-case noise degradation (N1, $\mathrm{NSI} = 0.032$, corresponding to a 2.6 pp absolute accuracy drop) is far less than the 18.6 pp memorization gap from counterfactual perturbation. As shown in Figure 2, the NSI values span both positive and negative territory, with N1 (irrelevant data) as the only substantial degradation source and N4 (contradictory hints) producing a large beneficial effect.

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**[a]
Original	82.4% (Standard)	79.0% (Noise-degraded)
Perturbed	63.5% (Robust)	— (Worst-case)

[a] All values are computed on the full corpus (N = 1,032). The I1 clean baseline (82.4%) and I3 clean baseline (81.6%) differ slightly due to separate experimental runs. Noise-degraded accuracy reflects worst-case N1 (irrelevant data) injection from the I3 experiment.

Standard accuracy (82.4%) is the metric currently reported by all financial LLM benchmarks. Robust accuracy (63.5%) accounts for memorization effects. Noise-degraded accuracy (79.0% for worst-case N1 noise) accounts for real-world information noise. Since both dimensions are now evaluated on the full corpus (N = 1,032), the comparison is direct: the standard-to-robust gap (18.9 pp) dwarfs the noise degradation (2.6 pp from the I3 clean baseline of 81.6%), confirming that counterfactual perturbation is the

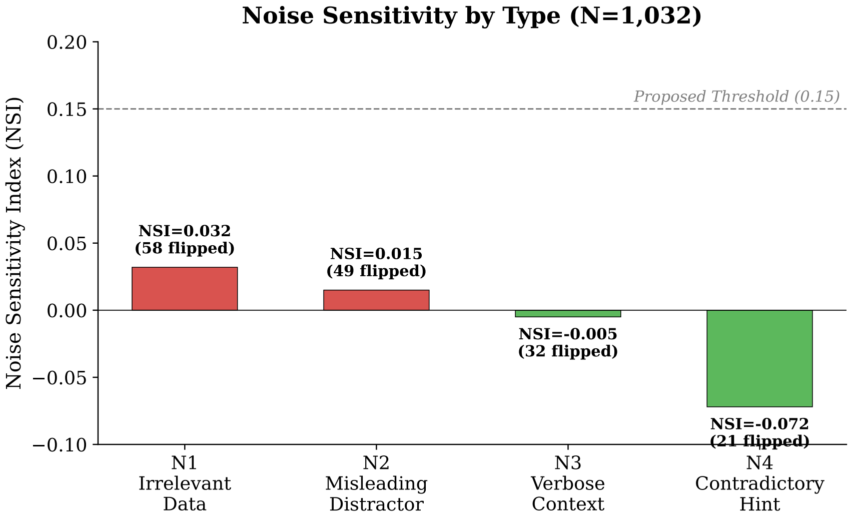

Figure 2: Noise Sensitivity Index (NSI) by noise type (N = 1,032). N1 (irrelevant data injection) causes the greatest performance degradation (NSI = 0.032). N2 (misleading distractors) shows minimal sensitivity (NSI = 0.015). N3 (verbose context) and N4 (contradictory hint) yield negative NSI values (−0.005 and −0.072, respectively), indicating that these noise types actually *improve* performance—particularly N4, which boosts accuracy by 5.9 pp. Overall, noise sensitivity is substantially lower than the memorization gap observed under counterfactual perturbation, and contradictory hints appear to trigger more deliberate reasoning.

more discriminating stress test.[1] Figure 3 visualizes the combined stress test results, presenting the evaluation conditions side by side to highlight the relative magnitude of each degradation pathway.

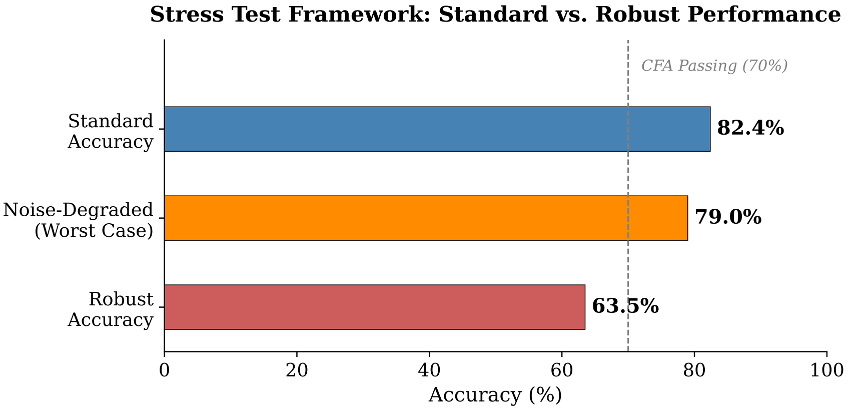

Figure 3: Combined stress test framework results. The bar chart compares standard accuracy (82.4%), worst-case noise-degraded accuracy (79.0%, N1 irrelevant data), and robust accuracy under perturbation (63.5%)—all evaluated on the full corpus ($N = 1{,}032$). The dominant source of performance loss is memorization-dependent reasoning (18.9 pp standard-to-robust gap) rather than noise susceptibility (2.6 pp from the I3 clean baseline), confirming that counterfactual perturbation is the more discriminating stress test dimension.

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 ($n = 702$)	McNemar's	$\chi^2 = 53.33$	< 0.001***
Clean vs. N1 (worst noise, $n = 1{,}032$)	McNemar's	$\chi^2 = 8.19$	0.004**

[1] The I1 and I3 experiments were run independently, yielding slightly different clean baselines (82.4% and 81.6%, respectively). The 18.9 pp figure is standard accuracy minus robust accuracy; the 2.6 pp figure is the I3 clean baseline minus worst-case noisy accuracy.

5.5. Cross-Model Comparison: The Memorization Paradox

To assess whether the stress test patterns are model-specific, we replicated the full framework on GPT-5-mini, a next-generation reasoning model that employs extended chain-of-thought ("thinking tokens") before producing its answer.

5.5.1. Counterfactual Perturbation

Table 6 presents the cross-model perturbation comparison.

Table 6: Cross-Model Counterfactual Perturbation Comparison ($N = 1{,}032$)

Metric	**GPT-4o-mini**	**GPT-5-mini**
Standard accuracy	82.4%	91.8%
Level 1 accuracy (n valid)	63.8% ($n = 702$)	55.3% ($n = 638$)
Memorization gap	**18.6 pp**	**36.4 pp**
Robust accuracy	63.5%	67.2%
Memorization suspect	18.9%	24.5%

GPT-5-mini's lower valid perturbation count (638 vs. 702) reflects stricter answer extraction requirements for the reasoning model's longer outputs.

Figure 4 visualizes the memorization paradox. The cross-model comparison reveals a *memorization paradox*: GPT-5-mini achieves substantially higher standard accuracy (+9.4 pp) but exhibits a nearly doubled memorization gap (36.4 pp vs. 18.6 pp).

In absolute terms, GPT-5-mini actually performs *worse* on perturbed questions (55.3% vs. 63.8%), despite dominating on originals. The robust accuracy improves only modestly (67.2% vs. 63.5%), meaning most of GPT-5-mini's apparent improvement evaporates under perturbation stress.

This result has a natural interpretation: GPT-5-mini, trained on a larger corpus with deeper reasoning capabilities, has likely encountered more CFA question templates during training and developed stronger template-matching associations. When original questions match these templates, the model exploits both memorization and reasoning, yielding exceptional accuracy. When perturbation breaks the template match, the model must fall back on pure reasoning—and the 55.3% perturbed accuracy reveals that this reasoning baseline, while adequate, is far below the 91.8% headline figure.

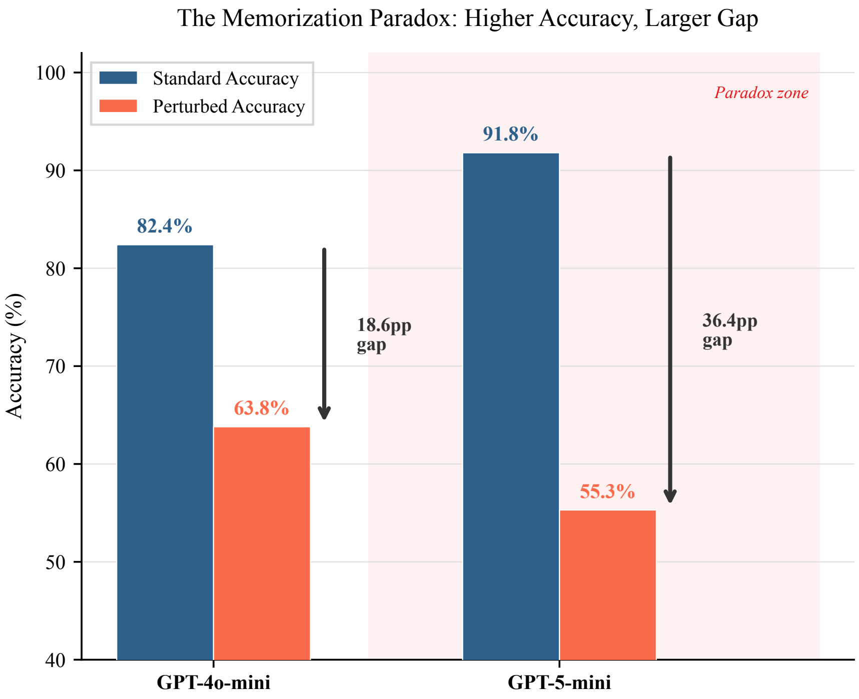

Figure 4: The Memorization Paradox ($N = 1{,}032$). GPT-5-mini achieves higher standard accuracy (91.8% vs. 82.4%) but lower perturbed accuracy (55.3% vs. 63.8%), resulting in a memorization gap (36.4 pp) nearly double that of GPT-4o-mini (18.6 pp). More capable models may be more memorization-dependent.

5.5.2. Noise Sensitivity

Table 7 presents the cross-model noise sensitivity comparison.

Table 7: Cross-Model Noise Sensitivity Comparison ($N = 1{,}032$)

	GPT-4o-mini		**GPT-5-mini**	
Condition	**Acc.**	**NSI**	**Acc.**	**NSI**
Clean (baseline)	81.6%	—	92.3%	—
N1 (irrelevant data)	79.0%	0.032	90.8%	0.017
N2 (misleading)	80.3%	0.015	91.0%	0.015
N3 (verbose context)	82.0%	−0.005	92.4%	−0.001
N4 (contradictory hint)	87.5%	−0.072	96.1%	−0.041

In contrast to the memorization paradox, the noise sensitivity results show clear improvement: GPT-5-mini's worst-case NSI (0.017 for N1) is roughly half of GPT-4o-mini's worst case (0.032 for N1). The model is substantially better at filtering irrelevant and misleading information, likely because extended chain-of-thought reasoning allows it to explicitly identify and discard noise. The N4 (contradictory hint) benefit is also smaller (−0.041 vs. −0.072), consistent with a model that relies less on elimination heuristics.

The memorization–noise asymmetry is thus amplified across generations: GPT-5-mini is simultaneously more memorization-dependent *and* more noise-robust than its predecessor. This dissociation confirms that counterfactual perturbation and noise injection probe fundamentally different cognitive dimensions (Section 6).

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** (82.4%) suggests the AI correctly handles roughly four out of five financial calculations.

- **Robust accuracy** (63.5%) reveals it reliably handles only about three out of five.
- The 18.9 pp gap between standard and robust accuracy 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. Dual Perturbation Taxonomy: Two Cognitive Dimensions

A central contribution of this paper is the recognition that our two stress test dimensions—counterfactual perturbation and noise injection—probe fundamentally different cognitive vulnerabilities. They should not be treated as variations of the same test but as assessments of orthogonal competence dimensions.

Numerical perturbation (I1) tests *memorization versus calculation.* By changing a single numerical parameter (e.g., a coupon rate from 5% to 5.13%, or a face value from $1,000 to $1,047), we hold the financial logic constant while invalidating any answer retrieved from memory. If the model has genuinely learned to compute bond duration, it should produce the correct answer regardless of the specific numbers. A performance drop under perturbation is diagnostic evidence that the model is retrieving a memorized answer rather than executing the underlying mathematical procedure.

Semantic perturbation (I3) tests *understanding versus attention.* By injecting irrelevant data, misleading context, or contradictory signals, we hold the numerical content constant while increasing the cognitive load on information filtering. The model must identify which pieces of information are decision-relevant and which are noise—a skill that depends on comprehension of the problem's causal structure rather than arithmetic ability. A performance drop under noise injection indicates that the model lacks a robust internal representation of what information matters and why.

This distinction maps onto a 2 × 2 competence framework:

	Passes I1 (perturbation)	**Fails I1**
Passes I3 (noise)	Genuine reasoning	Memorized but attentive
Fails I3	Calculates but distractible	Neither

Our results place GPT-4o-mini predominantly in the “memorized but attentive” quadrant: it handles noise well (NSI range −0.072 to 0.032) but struggles significantly with numerical perturbation (memorization gap of 18.6 pp). This pattern is consistent with a model that has developed strong attention mechanisms for filtering irrelevant information but relies substantially on pattern matching for numerical computation.

6.3. Why Numerical Perturbation Breaks LLMs

CFA examination questions are extensively distributed across the internet and almost certainly well-represented in LLM training corpora. A bond pricing question with canonical parameters (5% coupon, $1,000 face value, 10-year maturity) has appeared thousands of times in study guides and forums. The model learns strong associations between these specific patterns and their answers, achieving high accuracy through sophisticated retrieval rather than computation. When we perturb the coupon rate to 5.13% or face value to $1,047—combinations unlikely to appear in training data—the model must execute the underlying mathematical logic rather than retrieve a memorized answer. The 18.6 pp accuracy drop measures precisely this retrieval-to-computation transition, echoing Mirzadeh et al. [1] who found analogous fragility in mathematical reasoning benchmarks.

6.4. Prompt Fairness and Reproducibility

A potential concern with adversarial testing is that prompts may be unfairly constructed to elicit failure. We address this concern explicitly. All questions in our evaluation use the standard CFA multiple-choice format exactly as distributed in the CFA-Easy dataset [5]; we do not alter question structure, option format, or phrasing conventions. Perturbations are applied only to numerical parameters or through clearly defined noise injection templates—never to the question’s logical structure or option layout. The model receives the same system prompt and response format across all conditions (original, perturbed, and noisy), ensuring that any performance difference is attributable to the perturbation itself rather than to prompt engineering artifacts. All perturbation templates and evaluation code are available from the corresponding author, enabling full reproducibility of our stress test results.

6.5. Stress Testing as Regulatory Due Diligence

Drawing from quantitative finance, our memorization gap is analogous to a "delta" (first-order sensitivity to input perturbation), while NSI functions as "vega" (sensitivity to information noise). This mapping positions AI robustness within a framework familiar to financial risk managers. CFA Standard V(A)—Diligence and Reasonable Basis—requires a "reasonable and adequate basis" for recommendations. Deploying AI based solely on standard accuracy, without stress testing its reasoning, arguably fails this standard. Institutions should compute Robust Accuracy and NSI profiles before deployment.

6.6. 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.

6.7. The Memorization Paradox: Why Better Models May Be Less Robust

The cross-model comparison in Section 5.5 introduces a finding with important implications for AI governance: GPT-5-mini's memorization gap (36.4 pp) is nearly double GPT-4o-mini's (18.6 pp), despite GPT-5-mini being the more capable model. This "memorization paradox" suggests that **standard accuracy improvements may be substantially attributable to enhanced memorization rather than enhanced reasoning**.

Three mechanisms may contribute: (1) larger training corpora increase the probability that specific question templates—including CFA study materials—are well-represented; (2) reasoning models' extended thinking may facilitate more effective template retrieval through elaborative search; and (3) the higher standard accuracy itself creates a higher ceiling for memorization-dependent performance to occupy.

The regulatory implication is direct: as AI models improve, the gap between standard and robust accuracy may *widen*, not narrow. Financial regulators who track standard accuracy as a proxy for AI competence may observe steady improvement while the underlying robustness stagnates or even deteriorates. Mandatory stress testing—analogous to bank capital stress tests—becomes more, not less, important as headline accuracy rises.

6.8. Limitations

Several limitations should be acknowledged. First, perturbation generation relies on GPT-4o-mini, which may introduce its own errors in generating valid perturbations; the valid perturbation rate of 68.0% for GPT-4o-mini (702/1,032) and 61.8% for GPT-5-mini (638/1,032) reflects this challenge. Second, the current study evaluates only Level 1 (numerical) perturbation; Level 2 (conditional inversion) perturbation would provide additional evidence on deeper reasoning failures. Third, our noise types, while domain-informed, are synthetic; real-world financial noise may be more subtle or more severe. Fourth, the memorization paradox could partly reflect differences in the reasoning model's answer extraction rate rather than pure memorization effects.

7. Conclusion

This paper demonstrates that standard benchmark accuracy significantly overstates the financial reasoning ability of Large Language Models, and that this overstatement *increases* with model capability. Our two-dimensional stress testing framework—combining counterfactual perturbation with noise injection—reveals a memorization paradox: GPT-5-mini achieves 91.8% standard accuracy (+9.4 pp over GPT-4o-mini) but exhibits a 36.4 pp memorization gap (nearly double the 18.6 pp for GPT-4o-mini), while its noise sensitivity roughly halves. More capable models are simultaneously more memorization-dependent and more noise-robust, confirming that these stress tests probe orthogonal cognitive dimensions.

We introduce Robust Accuracy as a regulatory-relevant metric. The memorization paradox implies that robust accuracy should be mandatory in AI evaluation: headline accuracy improvements may mask stagnant or declining robustness. **The question is not whether AI can pass the CFA exam, but whether it can reason through problems it hasn't memorized—and our cross-model evidence suggests this question becomes more urgent, not less, as models improve.**

Data Availability

The CFA-Easy dataset is available via HuggingFace under the FinEval benchmark [5]. Experiment code and raw results are available from the corresponding author upon reasonable request.

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 Author Contributions

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

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

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*.