# Highlights

**The Illusion of Financial Competence: Stress Testing, Error Taxonomy, and Calibration Analysis of Large Language Models on CFA Examinations**

Wei-Lun Cheng, Daniel Wei-Chung Miao, Guang-Di Chang

- Counterfactual perturbation reveals 18.6pp memorization gap in CFA accuracy
- Stronger models memorize more: GPT-5-mini gap doubles to 36.4pp
- Conceptual errors dominate at 68.8%; calculation errors are only 1.4%
- No confidence threshold reduces LLM error rate to acceptable levels
- Robust Accuracy and Confidence-at-Risk proposed as governance metrics

# The Illusion of Financial Competence: Stress Testing, Error Taxonomy, and Calibration Analysis of Large Language Models 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 headline accuracy on financial examination benchmarks. We present a three-dimensional assessment revealing that this accuracy is largely illusory. First, *counterfactual stress testing* on the CFA-Easy corpus ($N$ = 1,032) exposes an 18.6 percentage point memorization gap—accuracy drops from 82.4% to 63.8% when numerical parameters are perturbed—demonstrating substantial pattern matching rather than genuine reasoning. A cross-model replication with GPT-5-mini reveals a *memorization paradox*: despite 91.8% standard accuracy (+9.4 pp), the memorization gap nearly doubles to 36.4 pp. Second, a *CFA Failure Taxonomy* of 557 errors from open-ended evaluation reveals that conceptual errors dominate (68.8%), while calculation errors account for only 1.4%. Golden Context Injection shows that 82.4% of errors respond to concept hints, establishing a ceiling for retrieval-augmented remediation. Third, *confidence calibration analysis* across 257 model–method observations reveals pervasive overconfidence: expressed confidence exceeds accuracy by 22–32 percentage points, 30.0% of responses are high-confidence errors, and no confidence threshold reduces error rates to acceptable levels. A robustness check on CFA-Easy ($N$ = 1,032) reveals a "fixed confidence register"—the model maintains ∼85% confidence regardless of whether accuracy is 53% or 82%. We introduce Robust Accuracy and Confidence-at-Risk (CaR) as regulatory-

*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)

relevant metrics and argue that financial AI governance must move beyond headline accuracy to encompass robustness, error structure, and calibration quality.

*Keywords:* Large Language Models, Financial Reasoning, Stress Testing, Calibration, Error Analysis, CFA Examination

---

**JEL Classification:** G20, C63, O33

## 1. Introduction

The financial industry is rapidly adopting Large Language Models (LLMs) for tasks including equity research, risk analysis, regulatory compliance, and client advisory [17, 9]. Benchmark evaluations show that state-of-the-art models can pass the CFA examination with scores approaching or exceeding human pass rates [3], and reasoning models now pass all three CFA levels with scores exceeding the 90th percentile of human candidates [14]. These impressive results have accelerated deployment timelines, with firms increasingly relying on LLM-generated analysis for consequential financial decisions.

However, headline accuracy—the single number universally reported by financial AI benchmarks—is a dangerously incomplete measure of AI competence. It tells us nothing about *why* models succeed, *how* they fail, or *whether they know when they are wrong*. This paper presents a three-dimensional assessment framework that probes each of these questions:

1. **Robustness ("Can it reason, or has it memorized?")** We stress test LLM financial reasoning through counterfactual perturbation and noise injection, quantifying the gap between standard and stress-tested performance—the "memorization premium" embedded in benchmark scores.
2. **Error Structure ("When it fails, how does it fail?")** We construct a taxonomy of 557 errors from open-ended CFA evaluation, revealing that the dominant failure mode is not "can't compute" but "doesn't understand the concept"—a finding with direct implications for remediation strategy.
3. **Calibration ("Does it know when it's wrong?")** We evaluate confidence calibration across multiple models and methods, finding that LLMs exhibit pervasive overconfidence that makes their error signal largely invisible to users who rely on expressed confidence.

The three dimensions are complementary. Stress testing reveals *how much* of standard accuracy is genuine; error analysis reveals *what types* of failures underlie the remaining errors; and calibration analysis reveals *whether the model's confidence signal can be trusted* to identify those failures. Together, they paint a picture of an AI system that appears highly competent on standard benchmarks but whose competence is substantially inflated by memorization, dominated by conceptual rather than computational errors, and accompanied by confidence signals that systematically overstate reliability.

Our contributions are fivefold: (1) we design a two-dimensional stress testing framework combining counterfactual perturbation with noise injection, revealing a memorization paradox across model generations; (2) we construct the first systematic error taxonomy for financial LLMs, demonstrating that 90.1% of errors are reasoning-based; (3) we introduce Golden Context Injection to distinguish knowledge gaps from reasoning gaps; (4) we quantify overconfident errors and introduce Confidence-at-Risk (CaR) as a risk management metric; and (5) we provide policy recommendations linking our metrics to financial regulatory frameworks.

## 2. Related Work

### *2.1. LLMs in Financial Applications*

The intersection of LLMs and finance has attracted significant research attention. BloombergGPT [17] demonstrated competitive performance on financial NLP tasks. Ke et al. [9] 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, how failures are structured, or whether confidence signals are reliable.

### *2.2. Data Contamination and Benchmark Validity*

The threat of data contamination in LLM evaluations is well-documented [15]. Mirzadeh et al. [13] demonstrated that LLMs show significant accuracy degradation when mathematical reasoning problems are symbolically perturbed, suggesting that high benchmark scores partly reflect memorization. Li et al. [10] extend this with GSM-Plus, systematically generating

variants across eight perturbation dimensions. Lopez-Lira et al. [12] specifically address the memorization problem in financial LLM evaluation, demonstrating pervasive benchmark contamination. Our stress testing extends this paradigm to the financial domain with population-level coverage.

### *2.3. LLM Calibration and Confidence*

Calibration refers to the alignment between a model's expressed confidence and its actual accuracy [7]. Kadavath et al. [8] demonstrate that large language models "mostly know what they know," but this degrades on out-of-distribution tasks. Band et al. [2] develop QA-Calibration methods for question-answering systems. Chhikara et al. [5] identify a persistent "confidence gap" across domains. Liu et al. [11] propose KalshiBench using prediction market data for calibration evaluation. Our work extends this literature to financial professional examinations where miscalibrated confidence carries direct monetary and fiduciary implications.

### *2.4. Error Analysis and Remediation*

Asai et al. [1] introduce Self-RAG with self-reflective retrieval mechanisms that share our goal of identifying when models lack the right knowledge. Chen et al. [4] develop a CFA-based benchmark with error categorization. Our Failure Taxonomy provides the first systematic three-dimensional taxonomy of financial LLM errors at scale ($N = 557$), enabling targeted remediation.

## 3. Methodology

### *3.1. Dimension 1: Stress Testing*

#### *3.1.1. Counterfactual Perturbation*

We employ numerical perturbation inspired by Mirzadeh et al. [13]: modifying one numerical parameter per question (e.g., interest rate, face value, maturity) while preserving the solution procedure. The correct answer changes, but the required formula and reasoning steps remain identical. Using GPT-4o-mini as a perturbation generator, each original question produces a variant with a clearly identified changed parameter, the correct perturbed answer, and verification that logical structure is preserved.

*3.1.2. Noise Injection*

We define four noise types modeling progressively more challenging information environments:

- **N1 — Irrelevant Data**: Extraneous numerical data unrelated to the solution.
- **N2 — Misleading Distractors**: Plausible but irrelevant financial statements.
- **N3 — Verbose Context**: Wordy but substantively vacuous padding.
- **N4 — Contradictory Hints**: References to common incorrect answers.

*3.1.3. Stress Testing Metrics*

**Memorization Gap:**

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

**Noise Sensitivity Index:**

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

**Robust Accuracy** requires correctness on both the original and *all* valid perturbation variants:

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

*3.2. Dimension 2: Error Taxonomy*

*3.2.1. Three-Level Grading*

Each of 1,032 CFA questions is presented in open-ended format (options removed). Responses are graded on three levels:

- **Level A (Exact)**: Answer matches within ±2% numerical tolerance or exact semantic match.
- **Level B (Directional)**: Correct direction/approach but different assumptions.
- **Level C (Incorrect)**: Wrong answer.

*3.2.2. Three-Dimensional Classification*

All Level C errors are classified along three dimensions: (1) Error type (7 categories: conceptual, incomplete reasoning, assumption, reading, arithmetic, formula, unknown); (2) CFA topic (8 knowledge areas); (3) Cognitive stage (5 stages: identify, recall, calculate, verify, unknown).

*3.2.3. Golden Context Injection (GCI)*

For each Level C error, we re-prompt the model with the correct financial concept as an explicit hint, then evaluate whether the model recovers. This distinguishes *knowledge gaps* (fixable via RAG) from *reasoning gaps* (requiring fine-tuning).

*3.3. Dimension 3: Calibration Analysis*

*3.3.1. Confidence Estimation Methods*

We employ two approaches: **verbalized confidence**, prompting models to express confidence as a percentage; and **self-consistency** [16], sampling $k = 10$ responses at $\tau = 0.7$ and defining confidence as the agreement ratio.

*3.3.2. Calibration Metrics*

**Expected Calibration Error (ECE):**

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\text{acc}(B_m) - \text{conf}(B_m)\right| \tag{4}$$

**Confidence-at-Risk (CaR):**

$$\text{CaR}(\alpha) = \inf\{c^* : P(\text{incorrect} \mid \text{confidence} \geq c^*) \leq \alpha\} \tag{5}$$

CaR answers: "What is the minimum confidence level at which the error rate falls below $\alpha$?" If undefined, the model's confidence signal is fundamentally unreliable for risk-budgeting purposes.

**Overconfident Errors:**

$$\text{Overconfident Error} = \mathbf{1}\left[\text{confidence} \geq 0.80 \ \wedge \ \text{incorrect}\right] \tag{6}$$

## 4. Data and Experimental Design

### *4.1. Datasets*

We use two datasets from FinEval [9]: **CFA-Easy** (1,032 multiple-choice questions covering the full CFA curriculum) for stress testing and error analysis, and **CFA-Challenge** (90 CFA Level III questions) for initial calibration analysis, with a robustness check extending calibration to the full CFA-Easy corpus.

### *4.2. Models*

The primary evaluation model is **GPT-4o-mini** (OpenAI), a widely deployed commercial model. Cross-model comparisons use **GPT-5-mini**, a next-generation reasoning model employing extended chain-of-thought. Calibration also evaluates **Qwen3-32B** (Alibaba), an open-weight 32B model. All evaluations use temperature $\tau = 0.0$ for deterministic outputs except self-consistency sampling ($\tau = 0.7$).

### *4.3. Total Experimental Scale*

Table 1 summarizes the experimental scale across all three dimensions.

Table 1: Experimental Scale Summary

| **Dimension** | **Component** | **Questions** | **Inferences** |
|---|---|---|---|
| Stress Testing | Counterfactual perturbation | 1,032 | 1,734[a] |
| | Noise injection (×4 types) | 1,032 | 5,160 |
| Error Taxonomy | Open-ended + GCI | 1,032 | 1,589[b] |
| Calibration | CFA-Challenge | 90 | 257[c] |
| | CFA-Easy (robustness) | 1,032 | 1,032 |
| **Total** | | | **>9,700** |

[a] Original + valid perturbations (702 for GPT-4o-mini).

[b] 1,032 open-ended + 557 GCI re-prompts.

[c] 2 models × 2 methods, with varying $n$ per configuration. The 257 observations comprise 250 from the main experimental run plus 7 from pilot validation runs (5 GPT-4o-mini verbalized, 2 Qwen3-32B verbalized).

## 5. Results

### *5.1. Dimension 1: Stress Testing*

#### *5.1.1. Counterfactual Perturbation*

Table 2 presents the core findings. At the population level ($N = 1{,}032$), the memorization gap of +18.6 pp confirms that a substantial portion of standard accuracy is attributable to numerical pattern matching rather than genuine financial reasoning.

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% | — | — | — |

Robust Accuracy requires correct answers on original *and* all valid perturbations. Memorization Suspect = fraction correct on original but incorrect on at least one perturbation.

#### *5.1.2. Noise Sensitivity*

Table 3 reveals a nuanced profile. N1 (irrelevant data) produces the highest sensitivity (NSI = 0.032), while N4 (contradictory hints) paradoxically *improves* performance (NSI = −0.072), boosting accuracy from 81.6% to 87.5%. We hypothesize that contradictory hints operate through elimination and metacognitive trigger channels.

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 |
| N4 (contradictory hint) | 87.5% | 21/1,032 | −0.072 | Negative (helps) |

The overall pattern confirms that the model's primary vulnerability lies in memorization-dependent reasoning rather than noise susceptibility: the

worst-case noise degradation (2.6 pp) is far less than the 18.6 pp memorization gap.

*5.1.3. Cross-Model Stress Testing: The Memorization Paradox*

Table 4 reveals a striking memorization paradox.

Table 4: Cross-Model Counterfactual Perturbation ($N = 1{,}032$)

| **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% |

GPT-5-mini achieves higher standard accuracy but lower perturbed accuracy, resulting in a nearly doubled memorization gap.

GPT-5-mini achieves substantially higher standard accuracy (+9.4 pp) but actually performs *worse* on perturbed questions (55.3% vs. 63.8%), producing a memorization gap nearly double that of GPT-4o-mini. The robust accuracy improves only modestly (67.2% vs. 63.5%), meaning most of GPT-5-mini's apparent improvement evaporates under perturbation stress. In contrast, noise sensitivity roughly halves (max NSI 0.017 vs. 0.032), confirming genuine improvement in information filtering. This memorization–noise asymmetry suggests that counterfactual perturbation and noise injection probe fundamentally different cognitive dimensions.

*5.2. Dimension 2: Error Taxonomy*

*5.2.1. Error Type Distribution*

Table 5 presents the error distribution across 557 Level C errors from open-ended evaluation.

Reasoning errors dominate overwhelmingly (90.1%), with conceptual errors alone (68.8%) exceeding all other categories combined. The primary failure mode is not "can't compute" but "doesn't understand the concept"—calculator tools and formula retrieval won't help when the fundamental financial concept is misunderstood.

Table 5: Error Type Distribution ($N = 557$)

| **Error Type** | **Count** | **%** | **Category** |
|---|---|---|---|
| Conceptual error | 383 | 68.8% | Reasoning |
| Incomplete reasoning | 60 | 10.8% | Reasoning |
| Assumption error | 59 | 10.6% | Reasoning |
| Unknown | 35 | 6.3% | — |
| Reading error | 12 | 2.2% | Extraction |
| Arithmetic error | 7 | 1.3% | Calculation |
| Formula error | 1 | 0.2% | Calculation |
| **Aggregated:** Reasoning 90.1%, Extraction 2.2%, Calculation 1.4%, Unknown 6.3% | | | |

*5.2.2. Topic-Level Error Profiles*

Error profiles are strikingly topic-dependent (noting that per-topic sample sizes range from ∼10 to ∼80, so these patterns should be treated as exploratory). Ethics exhibits 87.1% reasoning errors (no calculation errors), while Derivatives shows the highest calculation error rate (37.5%). This implies that different financial domains may require fundamentally different remediation strategies.

*5.2.3. Golden Context Injection*

Table 6 reveals that 82.4% of errors respond to golden context injection, indicating that the majority of failures are knowledge gaps amenable to retrieval augmentation.

Table 6: Golden Context Injection Results ($N = 557$ errors)

| **Recovery Level** | **Count** | **%** |
|---|---|---|
| Full recovery (Level A) | 142 | 25.5% |
| Partial recovery (Level B) | 317 | 56.9% |
| Still wrong (Level C) | 98 | 17.6% |
| **Any recovery (A+B)** | **459** | **82.4%** |

However, only 25.5% achieve full recovery; most improvements are partial (56.9%), indicating that even with the correct concept, the model often

struggles with precise execution. The 17.6% residual rate represents the true reasoning gap requiring training-time interventions.

A cross-model GCI replication with GPT-5-mini nearly doubles the full recovery rate (50.4% vs. 25.5%) while reducing the true reasoning gap to 11.7%, demonstrating that extended chain-of-thought reasoning substantially improves concept *execution* once the correct concept is provided.

An important caveat concerns the distinction between knowledge gaps and attention gaps. When GCI recovers an error, the model may have "known" the concept but failed to retrieve it—the hint merely directed attention. A definitive test would inject irrelevant concepts as a control condition. We leave this to future work, noting that the current results likely reflect a mixture of both mechanisms.

### *5.3. Dimension 3: Calibration*

#### *5.3.1. Overall Calibration*

Table 7 presents calibration metrics across all model–method combinations on CFA-Challenge questions.

Table 7: Calibration Metrics by Model and Method (CFA-Challenge, $N = 257$)

| **Model** | **Method** | **N** | **Acc** | **Conf** | **ECE** | **Brier** | **AUC** | **OC Gap** |
|---|---|---|---|---|---|---|---|---|
| GPT-4o-mini | Self-cons. | 90 | .522 | .829 | .307 | .334 | .639 | +.307 |
| GPT-4o-mini | Verbalized | 95 | .526 | .841 | .315 | .340 | .586 | +.315 |
| Qwen3-32B | Verbalized | 72 | .611 | .836 | .247 | .226 | .787 | +.225 |

ECE = Expected Calibration Error; AUC = Area Under ROC; OC Gap = Avg Confidence − Accuracy.

All configurations exhibit substantial overconfidence, with the overconfidence gap ranging from +22.5% to +31.5%. Models express an average confidence of 84% while achieving only 52–61% accuracy. A one-sample $t$-test on the per-observation overconfidence gap yields $t = 9.70$ ($p < 0.0001$).

#### *5.3.2. Overconfident Error Analysis*

Across all 257 observations, 77 are overconfident errors (30.0%), significantly exceeding a 20% baseline ($z = 3.99$, $p < 0.0001$). Among incorrect answers, 66.4% are delivered with confidence $\geq 80\%$, meaning most errors are high-confidence errors—the error signal is largely invisible to users relying on expressed confidence.

*5.3.3. Topic-Level Miscalibration*

Ethics & Standards exhibits the highest overconfident error rate (43.5%) and lowest accuracy (47.8%), while Derivatives shows a lower rate (22.2%). This Dunning-Kruger pattern—where models are most overconfident precisely where least competent—has direct implications for AI governance, suggesting calibration failures are inversely correlated with task difficulty.

*5.3.4. Confidence-at-Risk*

For GPT-4o-mini, CaR(5%) is *undefined*—no confidence threshold achieves a 5% error rate. Even at maximum self-consistency confidence (1.0), the error rate is 32.4% (12/37 observations); broadening to the $\geq 0.9$ confidence bin yields 41.7% (20/48). For Qwen3-32B, the error rate at confidence $\geq 95\%$ is 19.6%, still far exceeding acceptable risk tolerance. Current LLM confidence signals are fundamentally inadequate for financial risk management.

*5.3.5. Robustness Check: CFA-Easy ($N = 1{,}032$)*

To assess generalizability, we replicated the verbalized confidence protocol on the full CFA-Easy dataset.

Table 8: Calibration: CFA-Challenge vs. CFA-Easy (GPT-4o-mini)

| **Metric** | **CFA-Challenge** ($N = 95$) | **CFA-Easy** ($N = 1{,}032$) |
|---|---|---|
| Accuracy | 52.6% | 81.7% |
| Avg Confidence | 84.1% | 86.0% |
| ECE | 0.315 | 0.073 |
| AUROC | 0.586 | 0.671 |
| OC Gap | +31.5 pp | +4.3 pp |
| OC Error Rate | 40.0% | 15.1% |

OC Gap = Avg Confidence − Accuracy. OC Error Rate = proportion of high-confidence ($\geq 80\%$) incorrect responses.

Three findings emerge. First, calibration improves dramatically on easier questions (ECE: 0.315 → 0.073). Second, improvement is driven almost entirely by rising accuracy (+29.1 pp) rather than adjusted confidence (+1.9 pp)—the model maintains ∼85% confidence regardless of actual performance, evidence of a "fixed confidence register" rather than genuine metacognitive awareness. Third, AUROC remains mediocre (0.671), indicating the model cannot reliably distinguish correct from incorrect answers.

### *5.4. Integrating the Three Dimensions*

Table 9 presents the integrated three-dimensional assessment.

Table 9: Integrated Three-Dimensional Assessment (GPT-4o-mini)

| **Metric** | **Value** | **Implication** |
|---|---|---|
| Standard accuracy | 82.4% | Headline (misleading) |
| Robust accuracy | 63.5% | After memorization correction |
| Memorization premium | 18.9 pp | "Phantom competence" |
| Reasoning errors | 90.1% | Conceptual, not computational |
| GCI recovery rate | 82.4% | Knowledge gaps, fixable via RAG |
| True reasoning gap | 17.6% | Requires fine-tuning |
| Overconfidence gap | +31.5 pp | On hard questions |
| OC error rate | 30.0% | Invisible errors |
| CaR(5%) | Undefined | Cannot trust confidence |

The three dimensions converge on a single conclusion: the model's 82.4% headline accuracy overstates genuine competence by a wide margin. Roughly one in five correct answers reflects memorization rather than reasoning; when the model fails, it fails at concept identification rather than computation; and its confidence signal cannot reliably distinguish correct from incorrect answers.

## 6. Discussion

### *6.1. Economic Significance*

The memorization premium has concrete economic implications. Standard accuracy (82.4%) suggests four correct financial calculations out of five; robust accuracy (63.5%) reveals only three out of five. The 18.9 pp gap represents "phantom competence"—questions where the AI appears competent but would fail on real-world variants.

The overconfidence problem amplifies this risk. The 30% overconfident error rate means that a portfolio manager relying on AI confidence signals would make decisions as if the model were correct five out of six times, when it is wrong nearly every other time. An overconfident duration estimate ($D_{\text{error}}$) on a \$10M position with a 100-basis-point rate shock creates unexpected losses proportional to the error magnitude.

More broadly, the information value of AI advice is proportional to signal precision $\tau = 1/\sigma^2$. Our observed ECE values of 0.25–0.32 yield signal precision 40 times lower than what users implicitly assume when acting on "85% confident" recommendations.

### *6.2. The Memorization Paradox*

The cross-model evidence introduces a finding with important governance implications: GPT-5-mini's memorization gap (36.4 pp) nearly doubles GPT-4o-mini's (18.6 pp), despite being the more capable model. This suggests that **standard accuracy improvements may be substantially attributable to enhanced memorization rather than enhanced reasoning**. As AI models improve, the gap between standard and robust accuracy may *widen*, not narrow. Financial regulators tracking standard accuracy as a proxy for competence may observe steady improvement while underlying robustness stagnates.

### *6.3. Implications for Market Efficiency*

Under the Efficient Market Hypothesis [6], market prices reflect information processed by rational agents. When AI advisory systems become marginal price-setters, the systematic error patterns documented here threaten market efficiency: conceptual misapplication in Ethics (87.1% reasoning errors) could generate systematic compliance violations, while Derivatives pricing failures (37.5% calculation errors) could produce correlated hedging errors across AI-assisted portfolios. Unlike random noise, structured errors create directional bias.

### *6.4. CFA Ethics and Fiduciary Duty*

Our findings implicate three CFA Standards of Professional Conduct:

- **Standard I(C) — Misrepresentation**: With 30% of all responses being high-confidence errors—and those errors carrying an average expressed confidence of 89%—the model systematically misrepresents its reliability to users relying on confidence signals.
- **Standard V(A) — Diligence**: When 66.4% of errors are high-confidence, relying on AI confidence as a verification proxy fails the "reasonable basis" standard.

- **Standard III(C) — Suitability**: Topic-dependent miscalibration means AI is most unreliable in exactly the domains requiring the most professional judgment.

*6.5. Regulatory Implications*

Drawing from quantitative finance, our memorization gap is analogous to "delta" (sensitivity to input perturbation), NSI functions as "vega" (sensitivity to information noise), and CaR maps directly to Value-at-Risk. We propose tiered deployment standards:

- **Tier 1 (Advisory)**: ECE $< 0.15$, Memorization Gap $< 10\%$, OC error rate $< 15\%$
- **Tier 2 (Research)**: ECE $< 0.25$, Memorization Gap $< 20\%$
- **Tier 3 (Internal)**: ECE $< 0.35$ with mandatory disclaimers

Under these thresholds, none of the models tested on CFA-Challenge would qualify for Tier 1 or 2 deployment.

*6.6. Limitations*

Several limitations should be acknowledged. First, the experimental pipeline exhibits LLM-as-judge circularity: GPT-4o-mini generates perturbations, classifies errors, and judges open-ended responses for its own outputs. While this is common practice in the literature, it introduces potential systematic biases; importantly, the same pipeline applied to GPT-5-mini yields qualitatively different results (e.g., the memorization paradox), suggesting that the methodology is sensitive to genuine model differences rather than dominated by judge artifacts. Second, perturbation generation relies on GPT-4o-mini; only 68.0% of questions yielded valid perturbations that passed automated verification, limiting coverage. Among the 702 valid perturbations, undetected errors in the modified ground truth would cause the memorization gap to be *overestimated* if some "failures" on perturbed questions reflect faulty perturbations rather than genuine reasoning failures. Third, the calibration analysis on CFA-Challenge ($N = 90$) has limited statistical power, though the CFA-Easy robustness check ($N = 1{,}032$) strengthens generalizability. Fourth, topic-level analyses have limited per-topic sample sizes ($N = 10$–$46$) and should be treated as exploratory. Fifth, verbalized

confidence may be susceptible to prompt sensitivity. Sixth, cross-model comparisons are limited to two models from one provider (OpenAI); extension to other model families would strengthen generalizability. Finally, the GCI experiment cannot fully distinguish knowledge gaps from attention gaps without a control condition using irrelevant concept hints.

## 7. Conclusion

This paper demonstrates that standard benchmark accuracy significantly overstates the financial reasoning competence of Large Language Models, and that this overstatement *increases* with model capability. Our three-dimensional assessment reveals:

1. **Robustness**: An 18.6 pp memorization gap for GPT-4o-mini, nearly doubling to 36.4 pp for GPT-5-mini—a memorization paradox where more capable models are more memorization-dependent.
2. **Error Structure**: 90.1% of errors are reasoning-based, dominated by conceptual errors (68.8%), with calculation errors at only 1.4%. Golden Context Injection recovers 82.4% of errors, establishing a ceiling for RAG remediation.
3. **Calibration**: Pervasive overconfidence with 30% high-confidence errors, an undefined CaR at the 5% threshold, and a "fixed confidence register" that maintains ∼85% confidence regardless of actual performance.

The three dimensions converge: headline accuracy is inflated by memorization, failures are conceptual rather than computational, and the model cannot reliably signal when it is wrong. Financial AI governance must move beyond headline accuracy to encompass robustness, error structure, and calibration quality.

**The question is not whether AI can pass the CFA exam, but whether its passing reflects competence that transfers to novel problems, whether its failures can be efficiently remediated, and whether its confidence can be trusted. Our evidence answers all three questions in the negative—and the memorization paradox shows that the robustness problem may worsen, not improve, as models become more capable.**

## Data Availability

The CFA-Easy and CFA-Challenge datasets are available via HuggingFace under the FinEval benchmark [9]. 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] Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *Proceedings of ICLR 2024*.

[2] Band, N., Rudner, T. G. J., Filos, A., et al. (2025). Calibration of natural language understanding models with vague concepts. In *Proceedings of ICLR 2025*.

[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] Chen, Y., Li, H., & Zhang, X. (2025). A CFA-based benchmark for evaluating financial reasoning in large language models. *arXiv preprint arXiv:2509.04468*.

[5] Chhikara, P., Gaur, N., & Kumaraguru, P. (2025). Mind the confidence gap: Evaluating probabilistic forecasting of large language models. *Transactions on Machine Learning Research*.

[6] Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2), 383–417.

[7] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of ICML 2017* (pp. 1321–1330).

[8] Kadavath, S., Conerly, T., Askell, A., et al. (2022). Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

[9] Ke, Z., Ming, Y., Nguyen, X. P., Xiong, C., & Joty, S. (2025). Demystifying domain-adaptive post-training for financial LLMs. In *Proceedings of EMNLP 2025*.

[10] Li, Q., Zhu, Z., Wang, Z., et al. (2024). GSM-Plus: A comprehensive benchmark for evaluating the robustness of LLMs as mathematical problem solvers. In *Proceedings of ACL 2024*.

[11] Liu, M., Chen, Y., & Wang, J. (2025). KalshiBench: Evaluating LLM probabilistic calibration using prediction markets. *arXiv preprint*.

[12] Lopez-Lira, A., Kirtac, K., & Tang, Y. (2025). The memorization problem: When can we trust financial LLM benchmarks? *arXiv preprint*.

[13] 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*.

[14] Patel, R., Singh, A., & Torres, M. (2025). Reasoning models ace the CFA exams: Implications for professional certification. *arXiv preprint*.

[15] Shi, W., Ajith, A., Xia, M., et al. (2023). Detecting pretraining data from large language models. In *Proceedings of ICLR 2024*.

[16] Wang, X., Wei, J., Schuurmans, D., et al. (2023). Self-consistency improves chain of thought reasoning in language models. In *Proceedings of ICLR 2023*.

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