When AI Is Confidently Wrong: Calibration and Risk Analysis of Large Language Models in Financial Decision-Making

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) are increasingly deployed in financial applications, yet their reliability in high-stakes decision-making remains understudied. We evaluate the confidence calibration of LLMs on 90 CFA (Chartered Financial Analyst) examination questions, generating 257 model–method observations using two models (GPT-4o-mini, Qwen3-32B) and two confidence estimation methods (verbalized, self-consistency) across three model–method configurations. We find pervasive overconfidence: the average expressed confidence exceeds actual accuracy by 22–32 percentage points ($t = 9.70$, $p < 0.0001$). Critically, **30.0% of all responses are high-confidence errors** (confidence $\geq 80\%$), and among incorrect answers, 66.4% are delivered with high confidence. These overconfident errors are not uniformly distributed: Ethics & Standards questions exhibit a 43.5% overconfident error rate, compared to 22.2% for Derivatives ($\chi^2 = 12.37$, $p = 0.030$). We introduce Confidence-at-Risk (CaR), adapting Value-at-Risk methodology to AI confidence assessment, and connect our findings to CFA Institute Ethics Standards, arguing that deployment of miscalibrated AI may violate fiduciary duty requirements. A robustness check on the larger CFA-Easy corpus ($N = 1{,}032$) reveals that calibration improves dramatically (ECE = 0.073), but this improvement is driven entirely by higher accuracy (+29.1 pp) rather than adjusted confidence (+1.9 pp)—the model maintains a near-fixed con-

*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)

fidence register of ∼85% regardless of actual performance, indicating the absence of genuine metacognitive awareness. Our results suggest that financial regulators should establish minimum calibration standards—specifically, Expected Calibration Error (ECE) below 0.15—before permitting AI deployment in advisory roles.

Keywords: Large Language Models, Calibration, Financial AI, Risk Management, CFA Examination, Overconfidence

---

1. Introduction

The deployment of Large Language Models (LLMs) in financial services has accelerated rapidly, with applications spanning automated financial advice, risk assessment, equity research, and regulatory compliance [15, 11]. A growing body of work evaluates LLM *accuracy* on financial benchmarks—whether models can pass the CFA exam, correctly price derivatives, or interpret financial statements. However, accuracy alone is a poor guide to deployment safety.

Consider two AI systems: Model A achieves 70% accuracy with well-calibrated confidence (it says "85% confident" when it is correct 85% of the time), while Model B achieves 75% accuracy but systematically overstates confidence. Model B is more dangerous despite being more accurate, because its confidence signal cannot be trusted to identify errors. A financial advisor who says "I'm 95% certain this bond has a duration of 4.2 years" but is wrong poses a far greater risk than one who acknowledges uncertainty. This phenomenon—*overconfident error*—represents the most dangerous failure mode for AI systems in finance.

In behavioral finance, this failure mode has a well-established name: *overconfidence bias*—the systematic tendency to overestimate the precision of one's knowledge [3]. Decades of research document overconfidence bias among human investors, linking it to excessive trading, under-diversification, and poor risk assessment. However, human overconfidence is heterogeneous: it varies across individuals, is modulated by experience, and can be partially corrected through feedback. LLM overconfidence is qualitatively more dangerous because it is *systematic*—every user of the same model receives the same miscalibrated confidence signal, and no amount of individual experience corrects it. When an AI system deployed to thousands of financial advisors simultaneously expresses 90% confidence on an answer that is wrong 40% of

the time, the resulting correlated errors can produce systemic risk at a scale that idiosyncratic human overconfidence never could.

Yet confidence calibration of LLMs in financial contexts remains largely unexplored. Prior calibration studies focus on general knowledge [8], medical diagnosis [13], or scientific reasoning [12], leaving a critical gap in understanding how LLMs behave in domains where miscalibrated confidence carries direct monetary consequences.

This paper makes four contributions:

1. We systematically evaluate LLM calibration on CFA examination questions—a standardized benchmark for financial professional competency—using 257 observations across two models and two confidence estimation methods.
2. We identify and quantify *overconfident errors*, finding that 30.0% of all responses are high-confidence errors (confidence $\geq 80\%$), significantly exceeding a 20% baseline ($z = 3.99$, $p < 0.0001$).
3. We introduce *Confidence-at-Risk* (CaR), adapting Value-at-Risk methodology to quantify the reliability of AI confidence signals for risk management.
4. We connect calibration findings to the CFA Institute's Code of Ethics, arguing that poorly-calibrated AI deployment may violate fiduciary duty standards, and propose minimum calibration thresholds for financial AI regulation.

2. Related Work

2.1. LLM Calibration

Calibration refers to the alignment between a model's expressed confidence and its actual accuracy [7]. A well-calibrated model expressing 80% confidence should be correct approximately 80% of the time. Kadavath et al. [8] demonstrate that large language models "mostly know what they know," but this self-knowledge degrades on out-of-distribution tasks. Lin et al. [12] show that models can be prompted to verbalize uncertainty, though the resulting confidence estimates often exhibit systematic biases. Xiong et al. [16] survey confidence estimation methods for LLMs, identifying verbalized confidence, consistency-based, and logit-based approaches as the three primary paradigms. More recently, Band et al. [1] develop QA-Calibration methods specifically for question-answering systems, finding that calibration techniques effective in classification tasks often fail in open-ended QA. Chhikara

et al. [2] identify a persistent "confidence gap" between expressed and actual model performance across diverse domains, while Liu et al. [9] propose KalshiBench, using real-world prediction market data to evaluate probabilistic calibration in economic forecasting. Our work extends this literature to financial professional examinations, where the consequences of miscalibrated confidence carry direct monetary and fiduciary implications.

2.2. AI in Financial Applications

Domain-specific financial LLMs have emerged rapidly. BloombergGPT [15] demonstrated competitive performance on financial NLP tasks. Ke et al. [11] introduced FinDAP, a three-stage training pipeline that adapts Llama-3 to finance via continual pre-training, supervised fine-tuning, and preference alignment, achieving state-of-the-art results on CFA examination benchmarks. Callanan et al. [4] evaluated GPT-4 on CFA Level I, finding pass-rate performance but without examining confidence calibration.

2.3. Risk and Trust in AI-Assisted Financial Decision-Making

The literature on trust in algorithmic advice reveals a paradox: users both over-rely on and under-rely on algorithmic recommendations depending on context [5]. Green & Chen [6] argue that AI transparency alone is insufficient—what matters is whether users can accurately assess when AI is reliable. Our work contributes to this debate by showing that LLM confidence signals, which serve as the primary transparency mechanism, are themselves unreliable in financial domains.

3. Methodology

3.1. Confidence Estimation Methods

We employ two complementary confidence estimation approaches:

Verbalized Confidence. Following Lin et al. [12], we prompt models to express confidence as a percentage alongside their answer:

> *"Answer the following CFA question. After your answer, state your confidence level as a percentage (0–100%). Be honest about your confidence—if you are uncertain, say so with a lower percentage."*

Self-Consistency. Following Wang et al. [14], we sample $k = 10$ responses at temperature $\tau = 0.7$ for each question. Confidence is defined as the agreement ratio: $c = n_{\text{majority}}/k$, where n_{majority} is the count of the most frequent answer.[1]

3.2. Calibration Metrics

Our primary metric is the *Expected Calibration Error* (ECE) [7]:

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

where B_m denotes the set of predictions in confidence bin m, $\text{acc}(B_m)$ is the accuracy within that bin, and $\text{conf}(B_m)$ is the average confidence. We use $M = 10$ equal-width bins.

We also report the *Brier Score* $= \frac{1}{n}\sum_{i=1}^{n}(c_i - y_i)^2$, where c_i is the expressed confidence and $y_i \in \{0, 1\}$ is the correctness indicator; AUROC measuring whether confidence scores can discriminate correct from incorrect answers; and the *Overconfidence Gap* $= \overline{c} - \overline{y}$, which is positive when the model is systematically overconfident.

3.3. Overconfident Error Identification

We define *overconfident errors* as cases satisfying:

$$\text{Overconfident Error} = \mathbf{1}\left[\text{confidence} \geq \theta \ \wedge \ \text{answer incorrect}\right] \tag{2}$$

with threshold $\theta = 0.80$. This threshold is motivated by the decision-making context: a financial professional receiving a signal with $\geq 80\%$ confidence would typically act on it without extensive verification.

[1] The self-consistency confidence measure is bounded below by 1/3 (for three-option MCQ) and above by 1.0, with discrete steps of 0.1. This discretization means that the confidence distribution is inherently granular, and the comparison with verbalized confidence (which produces continuous percentages) should be interpreted with this methodological difference in mind. Despite this limitation, self-consistency has been shown to produce better-calibrated estimates than verbalized confidence in prior work [16].

3.4. Confidence-at-Risk (CaR)

Drawing from Value-at-Risk (VaR) methodology, we introduce *Confidence-at-Risk*:

$$\text{CaR}(\alpha) = \inf\{c^* : P(\text{incorrect} \mid \text{confidence} \geq c^*) \leq \alpha\} \tag{3}$$

CaR answers the question: "What is the minimum confidence level at which the error rate falls below α?" If CaR is undefined (no threshold achieves the target error rate), the model's confidence signal is fundamentally unreliable for risk-budgeting purposes.

4. Data and Experimental Design

4.1. Dataset

Table 1 summarizes our dataset: the CFA-Challenge corpus from FinEval [11], comprising 90 questions drawn from CFA Level III curriculum materials (SchweserNotes). All questions are multiple-choice with three options (A, B, C), spanning the full CFA curriculum. CFA Level III questions emphasize application and analysis, representing the most cognitively demanding tier of the CFA Program.

Table 1: Dataset Summary

Dataset	**Questions**	**Options**	**Source**
CFA-Challenge	90	3 (A/B/C)	SchweserNotes Level III

4.2. Models

We evaluate two LLMs representing different architectures and scales:

- **GPT-4o-mini** (OpenAI): A proprietary, cloud-hosted model optimized for efficient inference. Evaluated using both verbalized confidence ($n = 95$) and self-consistency with $k = 10$ samples ($n = 90$).[2]

[2]The verbalized run includes 5 additional questions from an extended SchweserNotes set. Self-consistency and Qwen3-32B results are on the standard 90-question CFA-Challenge corpus.

- **Qwen3-32B** (Alibaba): An open-weight, 32-billion parameter model run locally via Ollama. Evaluated using verbalized confidence ($n = 72$; 18 responses failed confidence extraction).

The total dataset comprises 257 model–question–method observations. Both models are evaluated at temperature $\tau = 0.0$ for single-shot methods and $\tau = 0.7$ for self-consistency sampling. Answers and confidence values are extracted using a five-layer regex chain with fallback parsing.

5. Results

5.1. Overall Calibration

Table 2 presents calibration metrics across all model–method combinations. All configurations exhibit substantial overconfidence, with the overconfidence gap ranging from +22.5% (Qwen3-32B) to +31.5% (GPT-4o-mini verbalized).

Table 2: Calibration Metrics by Model and Confidence Estimation Method

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; Brier = Brier Score; AUC = Area Under ROC; OC Gap = Avg Confidence − Accuracy.

To put this in concrete terms: across all configurations, models express an average confidence of 84% while achieving only 52–61% accuracy—a gap of 22–32 percentage points. The AI's confidence signal systematically overstates its reliability by more than half. A financial professional acting on these signals would make decisions as if the model were correct five out of six times, when in reality it is wrong nearly every other time.

A one-sample t-test on the per-observation overconfidence gap (confidence minus correctness indicator) yields $t = 9.70$ ($p < 0.0001$), confirming that LLM overconfidence on CFA questions is highly statistically significant (H1).

Figure 1 presents reliability diagrams. All models exhibit a consistent pattern: the calibration curve lies well below the diagonal (perfect calibration), indicating that models are more confident than they are accurate across virtually all confidence levels.

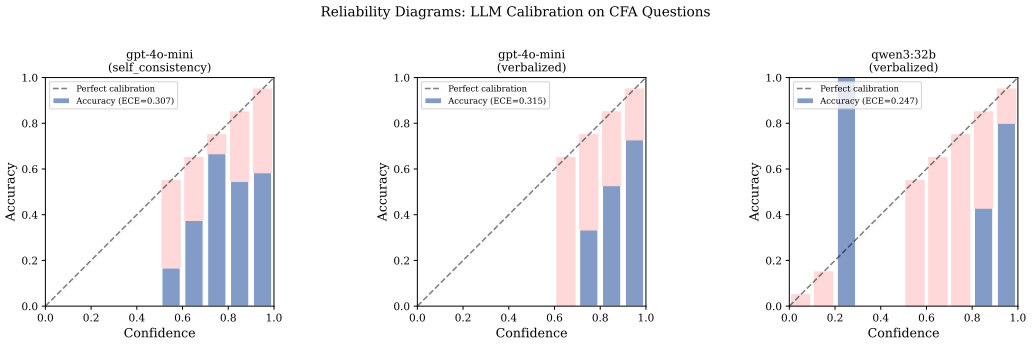

Figure 1: Reliability diagrams for three model–method configurations on CFA-Challenge questions. The dashed diagonal represents perfect calibration. Red-shaded regions indicate overconfidence: the gap between expressed confidence and actual accuracy. All configurations exhibit systematic overconfidence, particularly in the 0.8–1.0 confidence range where most predictions concentrate.

Qwen3-32B achieves the best calibration (ECE = 0.247) and highest discriminative ability (AUROC = 0.787), while GPT-4o-mini shows the worst calibration regardless of estimation method. Self-consistency marginally improves ECE over verbalized confidence for GPT-4o-mini (0.307 vs. 0.315), but substantially improves AUROC (0.639 vs. 0.586), suggesting it provides more discriminative confidence estimates.

Figure 2 visualizes the overconfidence gap across all configurations. In every case, expressed confidence substantially exceeds actual accuracy, confirming that overconfidence is a pervasive, not idiosyncratic, phenomenon.

5.2. Overconfident Error Analysis

Table 3 details the overconfident error profile. Across all 257 observations, 77 are overconfident errors (30.0%), significantly exceeding a 20% baseline (binomial test, $z = 3.99$, $p < 0.0001$; H2). Among the 116 incorrect answers, 66.4% are delivered with confidence $\geq 80\%$ ($z = 3.53$, $p = 0.0002$), meaning that *most errors are high-confidence errors*. The average confidence of overconfident errors is 89.0%.

5.3. Topic-Level Miscalibration

Table 4 reveals significant variation in overconfident error rates across CFA knowledge domains ($\chi^2 = 12.37$, $p = 0.030$; H3). Given the limited per-topic sample sizes (ranging from $N = 10$ for Economics to $N = 46$ for Ethics

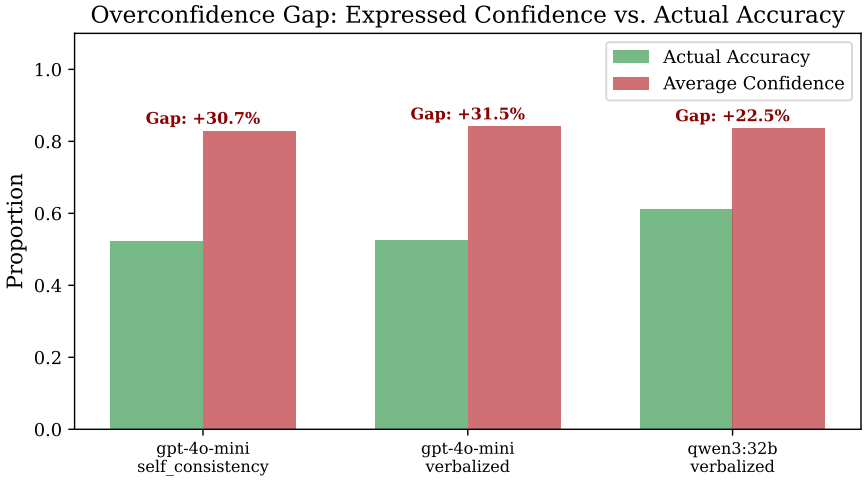

Figure 2: Overconfidence gap across models and methods. Blue bars represent actual accuracy; red bars represent average expressed confidence. The gap between them—ranging from +22.5% to +31.5%—quantifies the degree of systematic overconfidence.

Table 3: Overconfident Error Analysis (Confidence $\geq$ 80%)

Model / Method	**Total**	**Errors**	**OC Errors**	**OC Rate**
GPT-4o-mini / Self-cons.	90	43	25	27.8%
GPT-4o-mini / Verbalized	95	45	38	40.0%
Qwen3-32B / Verbalized	72	28	14	19.4%
Overall	**257**	**116**	**77**	**30.0%**

& Standards), the topic-level analysis should be treated as exploratory; the statistical significance of the overall χ^2 test should not be interpreted as evidence that each individual topic-level estimate is precise.

Table 4: Calibration Metrics by CFA Topic (Topics with $N \geq 10$)

CFA Topic	**N**	**Acc.**	**ECE**	**OC Errors**	**OC Rate**
Ethics & Standards	46	.478	.360	20	43.5%
Portfolio Management	14	.500	.357	6	42.9%
Economics	10	.600	.340	4	40.0%
Fixed Income	20	.650	.223	6	30.0%
Derivatives	27	.593	.291	6	22.2%

Ethics & Standards—the CFA curriculum's foundational domain—exhibits the highest overconfident error rate (43.5%) and lowest accuracy (47.8%). This is particularly concerning: ethics questions require nuanced professional judgment, exactly the domain where overconfident AI poses the greatest fiduciary risk. Derivatives, despite involving complex quantitative reasoning, shows a lower overconfident error rate (22.2%), suggesting that models may be better calibrated on computation-heavy topics where confidence can be partially grounded in mathematical consistency.

5.4. Coverage-Accuracy Tradeoff

Figure 3 presents the selective prediction analysis. If a financial institution restricts AI to answer only questions where confidence exceeds a threshold, what accuracy can be achieved?

For Qwen3-32B, restricting to predictions with confidence $\geq 90\%$ yields 81.3% accuracy at 69% coverage—exceeding the CFA passing threshold. However, GPT-4o-mini cannot achieve 70% accuracy at *any* confidence threshold, reaching only 67.6% even when restricted to its highest-confidence predictions (41% coverage). This demonstrates that selective prediction viability is highly model-dependent.

5.5. Confidence-at-Risk

Applying our CaR framework to the data: for GPT-4o-mini, CaR(5%) is *undefined*—no confidence threshold achieves a 5% error rate. Even at maximum confidence (self-consistency = 1.0), the error rate remains 41.7%. For

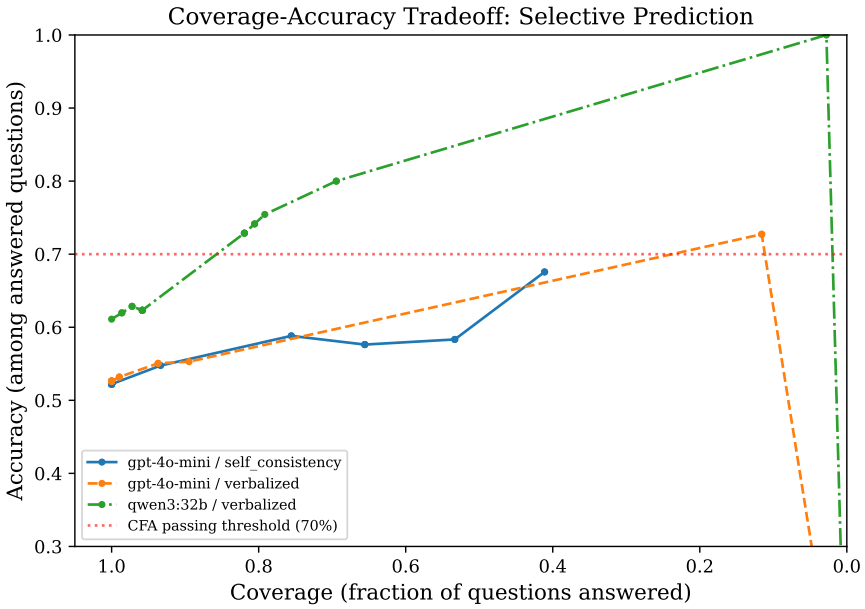

Figure 3: Coverage-accuracy tradeoff under selective prediction. As the confidence threshold increases (moving right to left), fewer questions are answered (lower coverage) but with higher accuracy. The horizontal dashed line marks the CFA passing threshold (70%). Qwen3-32B reaches 81.3% accuracy at 69% coverage, while GPT-4o-mini cannot reach 70% accuracy at any coverage level.

Qwen3-32B, the error rate at confidence $\geq 95\%$ is 19.6%, still far exceeding a 5% risk tolerance. These results demonstrate that current LLM confidence signals are fundamentally inadequate for financial risk management purposes.

5.6. Statistical Summary

Table 5 summarizes all hypothesis tests:

Table 5: Statistical Tests for Research Hypotheses

Hypothesis	**Test**	**Statistic**	***p*-value**
H1: Systematic overconfidence	One-sample t-test on gap	$t = 9.70$	$< 0.0001^{***}$
H2: OC error > 20%	Binomial, $\hat{p} = 0.300$	$z = 3.99$	$< 0.0001^{***}$
H2b: Most errors are OC	Binomial, $\hat{p} = 0.664$	$z = 3.53$	0.0002^{***}
H3: Topic-dependent	Chi-squared test	$\chi^2 = 12.37$	0.030^{*}

$^{***}p < 0.001$; $^{*}p < 0.05$. OC = overconfident.

6. Discussion

6.1. Economic Significance

The finding that 30% of AI responses are high-confidence errors has concrete economic implications. In portfolio management, an overconfident duration estimate can directly impact portfolio value:

$$\Delta V \approx -D_{\text{error}} \times \Delta y \times V \tag{4}$$

If an AI system reports "duration = 4.2 years, confidence = 95%" but the true duration is 6.8 years, a 100-basis-point rate shock on a $10 million portfolio position creates an unexpected loss of approximately $260,000—a 2.6% portfolio-level loss that was "invisible" to the risk model.

More broadly, our Confidence-at-Risk analysis reveals that no confidence threshold can reduce the error rate below 19.6% for the best model (Qwen3-32B) and 41.7% for GPT-4o-mini. In traditional risk management, a VaR model with 41.7% exceedance rate would be immediately rejected. Our CaR metric formalizes this analogy, providing risk managers with a framework to

evaluate AI confidence signals using the same rigor applied to financial risk models.

The overconfidence gap of 22–32 percentage points has implications for information economics. In rational expectations models, information value is proportional to signal precision $\tau = 1/\sigma^2$. A well-calibrated model with ECE = 0.05 provides signal precision $\tau \approx 400$, while our observed ECE values of 0.25–0.32 yield $\tau \approx 10$–16—*40 times less informative* than what users implicitly assume when acting on "85% confident" recommendations.

6.2. CFA Ethics Framework

We map overconfident errors to three CFA Institute Standards of Professional Conduct:

Standard I(C) — Misrepresentation. The CFA Standards prohibit misstatement of performance or analysis. An AI system that expresses 89% average confidence on answers that are wrong 30% of the time systematically misrepresents its analytical reliability. Whether this constitutes misrepresentation depends on whether the firm presents the AI's confidence as a calibrated probability—if so, our evidence suggests such presentation is materially misleading.

Standard V(A) — Diligence and Reasonable Basis. CFA charterholders must have a "reasonable and adequate basis" for investment recommendations. When an AI system expresses high confidence, the natural human response is reduced verification effort. Our finding that 66.4% of errors are high-confidence errors means that relying on AI confidence signals as a proxy for verification fails the "reasonable basis" standard—the very cases where verification is most needed are the ones where the AI most discourages it.

Standard III(C) — Suitability. Our topic-level analysis reveals that Ethics & Standards questions—which directly test professional judgment—exhibit the highest overconfident error rate (43.5%). An AI system confidently recommending actions that violate fiduciary standards undermines the suitability requirement.

6.3. The Dunning-Kruger Pattern in LLM Confidence

Our topic-level results reveal a striking pattern that parallels the Dunning-Kruger effect [10]: models are most overconfident precisely where they are least competent. Ethics & Standards—the domain where models perform worst (47.8% accuracy)—simultaneously exhibits the highest overconfident

error rate (43.5%). This is the hallmark of the Dunning-Kruger effect: individuals who lack competence in a domain also lack the metacognitive ability to recognize their incompetence, leading to inflated self-assessments.

In human cognition, the Dunning-Kruger effect arises because the skills needed to produce correct answers are the same skills needed to recognize what a correct answer looks like. An analogous mechanism may operate in LLMs: ethics questions require subtle professional judgment about competing obligations and contextual nuance—precisely the kind of reasoning where models lack reliable internal signals to gauge their own uncertainty. By contrast, Derivatives questions (22.2% overconfident error rate, 59.3% accuracy) involve more structured, computation-grounded reasoning where models can partially verify their own logic, producing better-calibrated confidence.

This Dunning-Kruger parallel has practical implications for financial AI governance. It suggests that calibration failures are not uniform but are *inversely correlated with task difficulty*—the hardest questions receive the most dangerously overconfident answers. Regulatory calibration standards should therefore be evaluated on a per-domain basis, not merely in aggregate, as overall ECE can mask catastrophic miscalibration in the domains that matter most.

6.4. Regulatory Implications and Policy Recommendations

Our findings have direct implications for emerging AI financial regulation. Table 6 maps our findings to specific regulatory frameworks:

We propose a tiered minimum calibration standard for financial AI deployment:

- **Tier 1 (Advisory/Execution)**: ECE < 0.15, overconfident error rate $< 15\%$
- **Tier 2 (Screening/Research)**: ECE < 0.25, overconfident error rate $< 25\%$
- **Tier 3 (Internal tool)**: ECE < 0.35 with mandatory confidence disclaimers

Under this framework, none of the models tested would qualify for Tier 1 or Tier 2 deployment. Qwen3-32B (ECE $= 0.247$) would marginally qualify for Tier 2, while GPT-4o-mini (ECE > 0.30) would be restricted to Tier 3.

Table 6: Mapping Findings to Regulatory Frameworks

Framework	**Relevant Provision**	**Our Finding**
EU AI Act (2024)	High-risk AI in financial services requires accuracy assessment	ECE > 0.30 fails accuracy transparency
SEC AI Guidance	AI-driven recommendations need "reasonable basis"	30% OC error rate undermines this standard
MAS FEAT Principles	AI must be fair, ethical, accountable, transparent	Miscalibrated confidence violates transparency
CFA Institute	AI should augment, not replace, judgment	Overconfident AI suppresses professional skepticism

6.5. Robustness Check: CFA-Easy (N = 1,032)

To assess whether our calibration findings generalize beyond the 90-question CFA-Challenge corpus, we replicated the verbalized confidence protocol on the full CFA-Easy dataset ($N = 1{,}032$) using GPT-4o-mini. Table 7 presents the comparison.

Three findings emerge. First, calibration improves dramatically on easier questions: ECE drops from 0.315 to 0.073 and the overconfidence gap narrows from +31.5 pp to +4.3 pp. Second, and crucially, the improvement is driven almost entirely by rising accuracy (+29.1 pp) rather than adjusted confidence (+1.9 pp). The model expresses nearly identical confidence (84–86%) regardless of whether it achieves 53% or 82% accuracy—evidence of a "fixed confidence register" rather than genuine metacognitive awareness. Third, AUROC remains mediocre (0.671), indicating that even on easier questions the model cannot reliably distinguish when it is right from when it is wrong.

The CFA-Easy ECE of 0.073 would meet our proposed Tier 1 threshold (ECE < 0.15) for advisory deployment, yet this apparent calibration is an artifact of difficulty: the model happens to be correct at roughly the rate it claims confidence, not because it has calibrated uncertainty. The fixed confidence behavior means that on any new question set where accuracy deviates from $\sim$85%, the calibration will deteriorate proportionally. This finding reinforces our recommendation that calibration standards should be

Table 7: Calibration comparison: CFA-Challenge vs. CFA-Easy (GPT-4o-mini, verbalized confidence)

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
MCE	0.600	0.303
Brier Score	0.340	0.143
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 all responses that are high-confidence ($\geq 80\%$) and incorrect.

evaluated per-domain and per-difficulty tier rather than in aggregate.

6.6. Limitations

Several limitations should be acknowledged. First, our primary dataset comprises 90 unique CFA-Challenge questions, yielding 257 observations, supplemented by a 1,032-question robustness check on CFA-Easy. While these provide consistent evidence of overconfidence, larger-scale validation across additional models would strengthen generalizability. Second, topic classification is based on keyword matching from question text, which may introduce noise. Third, our CFA question benchmark, while standardized, may not fully represent the range of financial reasoning tasks encountered in practice. Fourth, verbalized confidence may be susceptible to prompt sensitivity; we use a single prompt template and acknowledge that alternative prompts could yield different calibration profiles. Fifth, next-generation reasoning models (e.g., GPT-5-mini) present methodological challenges for calibration analysis: their use of internal "thinking tokens" alters the confidence generation mechanism, and their API restrictions on logprobs and temperature control limit the applicability of self-consistency methods. Cross-generational calibration comparison is an important direction for future work, particularly given that our companion studies observe substantial behavioral differences between standard and reasoning models on financial tasks.

7. Conclusion

This paper demonstrates that Large Language Models exhibit significant calibration failures on financial reasoning tasks. Our key finding is stark: 30% of all AI responses to CFA questions are high-confidence errors—cases where the model is confident but wrong. Among all incorrect answers, two-thirds are delivered with high confidence, meaning that *the error signal is largely invisible* to users who rely on expressed confidence.

We introduce Confidence-at-Risk (CaR), adapting Value-at-Risk methodology to evaluate the reliability of AI confidence signals. Applying CaR reveals that no confidence threshold can reduce the error rate to acceptable levels for financial risk management, even for the best-performing model.

Our analysis connects technical calibration metrics to the CFA Institute's ethical framework, demonstrating that three Standards of Professional Conduct—Misrepresentation, Diligence, and Suitability—are implicated by overconfident AI deployment. We propose tiered minimum calibration standards for financial AI, ranging from ECE < 0.15 for advisory roles to ECE < 0.35 for internal tools.

The question is not whether AI can pass the CFA exam, but whether it knows when it cannot. Our evidence suggests it does not.

Data Availability

The CFA-Challenge dataset is available via HuggingFace under the FinEval benchmark [11]. 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] Band, N., Rudner, T. G. J., Filos, A., et al. (2025). Calibration of natural language understanding models with vague concepts. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*.

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

[3] De Bondt, W. F. M. and Thaler, R. H. (1995). Financial decision-making in markets and firms: A behavioral perspective. In *Handbooks in Operations Research and Management Science*, Vol. 9 (pp. 385–410). Elsevier.

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

[5] Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 114–126.

[6] Green, B., & Chen, Y. (2019). The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 1–24.

[7] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning* (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] Liu, M., Chen, Y., & Wang, J. (2025). KalshiBench: Evaluating LLM probabilistic calibration using prediction markets. *arXiv preprint*.

[10] Kruger, J. and Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6), 1121–1134.

[11] 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)*.

[12] Lin, S., Hilton, J., & Evans, O. (2022). Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.

[13] Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). Capabilities of GPT-4 on medical competence examinations. *arXiv preprint arXiv:2303.13375*.

[14] Wang, X., Wei, J., Schuurmans, D., et al. (2023). Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*.

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

[16] Xiong, M., Hu, Z., Lu, X., et al. (2024). Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.