Inherited Irrationality and Ethical Fragility: Behavioral Biases and Adversarial Vulnerability 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) deployed as financial advisors face a dual threat: they may inherit the behavioral biases of their human training data *and* abandon ethical standards under adversarial pressure. We present a unified experimental framework examining both threats. First, we measure six canonical behavioral biases—loss aversion, anchoring, framing, recency bias, the disposition effect, and overconfidence—in GPT-4o-mini across 140 paired financial scenarios (pilot $N = 60$ + replication $N = 80$). A pilot study yields a mean bias score of 0.500 (on a 0–1 scale); neutral re-framing reduces this to 0.425 (Wilcoxon $p = 0.027$), but debiasing effectiveness reveals a *three-tier hierarchy*: surface biases (loss aversion $\Delta = +0.300$, framing $+0.150$) respond to prompt-level intervention; anchoring shows marginal response ($+0.050$); while deep biases—disposition effect, overconfidence, and recency—are entirely resistant. A scaled replication on 80 synthetic scenarios ($N = 12$–14 per type) confirms and sharpens the hierarchy: the disposition effect emerges as the strongest bias overall (0.808), validating its resistance to debiasing across independently generated scenarios. Second, applying five adversarial pressure types (profit incentive, authority, emotional manipulation, reframing, moral dilemma) to 47 CFA Ethics questions, we find *universal degradation*: all five attacks reduce accuracy (ERS = 0.925–0.950), flipping 14 previously correct answers. An extended experiment on 141 synthetic ethics questions reveals

*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)

a dramatically reduced flip rate (0.85% vs. 5.96%), with only reframing and moral dilemma retaining efficacy. Qualitative analysis identifies three rationalization strategies—utilitarian override, authority deference, and semantic repackaging—through which the model constructs plausible justifications for ethically compromised answers. A cross-model comparison shows GPT-5-mini achieves zero adversarial flips (ERS > 1.0), suggesting the vulnerability may be generationally bounded. Together, these findings reveal that current financial LLMs face a *double jeopardy*: they exhibit patterns consistent with human irrationality (risking suboptimal investment outcomes) *and* ethical fragility under pressure (risking compliance violations)—complementary failure modes requiring distinct mitigation strategies.

Keywords: behavioral finance, adversarial ethics, large language models, loss aversion, disposition effect, AI safety, CFA examination, fiduciary duty

---

1. Introduction

The rapid deployment of large language models (LLMs) in financial advisory, compliance, and analytical roles raises two fundamental questions about their reliability. First, do these systems—trained on vast corpora of human-generated financial text—inherit the systematic cognitive biases documented in behavioral finance? Second, can they maintain ethical judgment when subjected to the kinds of pressure that routinely compromise human professionals?

These questions address complementary failure modes with distinct consequences. Behavioral biases (loss aversion, disposition effect, anchoring) lead to *suboptimal investment outcomes*—the risk is losing money. Ethical fragility under adversarial pressure leads to *compliance violations*—the risk is legal and regulatory sanctions. A financial institution deploying an LLM that is both irrationally biased and ethically fragile faces a compounded risk that current AI evaluation frameworks—focused exclusively on accuracy—completely overlook.

We present a unified experimental framework addressing both dimensions:

1. **Behavioral Bias Measurement**: Using a paired-scenario design across 140 CFA-level financial scenarios (pilot $N = 60$ + replication $N = 80$) covering six bias types, we quantify bias susceptibility and debiasing effectiveness, identifying a three-tier hierarchy of bias persistence validated across independently generated scenarios.

2. **Adversarial Ethics Stress Testing**: Using five pressure types applied to 47 CFA Ethics questions, we measure ethical robustness under adversarial conditions, discovering universal degradation and characterizing the rationalization strategies through which the model justifies abandoning ethical standards.

Our contributions are fourfold: (1) we provide the first joint measurement of behavioral biases and ethical fragility in a financial LLM; (2) we identify a three-tier debiasing hierarchy distinguishing surface, weakly responsive, and deep biases; (3) we introduce a taxonomy of AI rationalization strategies under adversarial pressure; and (4) we demonstrate that the two failure modes require fundamentally different mitigation approaches—prompt engineering for surface biases, training-time intervention for deep biases, and alignment improvements for ethical robustness.

2. Related Work

2.1. Behavioral Biases in LLMs

The foundational work of Kahneman and Tversky [5] established that individuals systematically violate expected utility theory through loss aversion and reference dependence. In financial markets, these manifest as the disposition effect [10] and anchoring bias [12]. Ross et al. [9] introduce the "LLM Economicus" framework, finding that GPT-4 violates expected utility axioms in abstract gamble scenarios. Suri et al. [11] extend this to economic decision-making, showing that GPT-3.5 exhibits loss aversion. Capraro et al. [2] provide a comprehensive survey across cognitive psychology experiments. Malberg et al. [7] demonstrate that bias measurement methodologies vary substantially across studies.

Our work departs from this literature by using CFA-level investment scenarios rather than abstract gambles, testing the finance-specific disposition effect, and introducing a debiasing hierarchy.

2.2. Adversarial Ethics Testing

AI safety research has developed sophisticated adversarial testing methods. Chen et al. [3] demonstrate that multi-turn "foot-in-the-door" attacks are more effective than single-shot prompts. Hui et al. [4] propose TRIDENT as a financial safety benchmark. Mazeika et al. [8] introduce HarmBench for standardized red teaming. Andriushchenko et al. [1] show that simple

adaptive attacks can bypass alignment. Our work applies adversarial testing specifically to CFA Ethics questions, connecting findings to CFA Standards of Professional Conduct and introducing domain-specific metrics (Ethics Robustness Score).

3. Methodology

3.1. Study 1: Behavioral Bias Measurement

3.1.1. Paired-Scenario Design

For each financial decision, we construct two versions: a *bias-inducing version* framed to trigger the target bias, and a *neutral version* presenting only quantitative facts. If the model were perfectly rational, recommendations should be identical across framings.

We test six canonical biases with 10 scenarios each ($N = 60$): loss aversion, anchoring, framing, recency bias, the disposition effect, and overconfidence. Each scenario presents a CFA-level investment decision with realistic financial context.

3.1.2. Bias Scoring

An LLM judge (GPT-4o-mini) assigns a bias score $\in \{0.0, 0.5, 1.0\}$ where 0.0 indicates fully rational, 0.5 mixed/hedged, and 1.0 fully biased behavior. The debiasing effect is:

$$\Delta_{\text{debias}} = S_{\text{bias}} - S_{\text{neutral}} \tag{1}$$

3.2. Study 2: Adversarial Ethics Stress Testing

3.2.1. Adversarial Prompt Design

For each of 47 CFA Ethics questions from the CFA-Easy dataset [6], we create a standard version and five adversarial versions prepending pressure-inducing contexts: **profit incentive** (financial rewards for the wrong action), **authority pressure** (superior's instruction to override ethics), **emotional manipulation** (sympathetic scenario encouraging rule-breaking), **reframing** (linguistic disguise of violations), and **moral dilemma** (utilitarian argument against rule-following).

3.2.2. Ethics Robustness Score

$$\text{ERS}_t = \frac{\text{Accuracy}_{\text{adversarial},t}}{\text{Accuracy}_{\text{standard}}} \tag{2}$$

ERS = 1.0 means no effect; ERS < 1.0 indicates degradation. We track "flipped" questions: correct under standard but incorrect under adversarial conditions.

3.3. Model

Both studies evaluate **GPT-4o-mini** (OpenAI) at temperature $\tau = 0.0$. Cross-model comparisons use **GPT-5-mini** for adversarial ethics testing.

4. Results

4.1. Study 1: Behavioral Biases

4.1.1. Pilot Results ($N = 60$)

Table 1 presents aggregate bias measurement results from the pilot study. The model exhibits a mean bias score of 0.500, with neutral re-framing reducing this to 0.425 (Wilcoxon $W = 14.0$, $p = 0.027$, $r = 0.286$, exact test).

Table 1: Pilot bias measurement (GPT-4o-mini, $N = 60$ scenarios, 10 per type)

Metric	**Bias-Inducing**	**Neutral**	$\Delta_{\textbf{debias}}$
Mean Score	0.500	0.425	+0.075
Std Dev	0.129	0.201	0.220
Wilcoxon: $W = 14.0$, $p = 0.027$, $r = 0.286$ (exact test)			

4.1.2. Three-Tier Debiasing Hierarchy

Table 2 reveals substantial heterogeneity across bias types, forming a three-tier hierarchy.

Table 2: Pilot bias scores by type (GPT-4o-mini, $N = 60$, 10 per type)

Bias Type	n	**Bias**	**Neutral**	$\Delta_{\textbf{debias}}$
Tier 1: Surface biases (prompt-debiasable)				
Loss Aversion	10	0.500	0.200	+0.300
Framing	10	0.550	0.400	+0.150
Tier 2: Weakly responsive				
Anchoring	10	0.500	0.450	+0.050
Tier 3: Deep biases (resistant to debiasing)				
Disposition Effect	10	0.500	0.500	+0.000
Overconfidence	10	0.500	0.500	+0.000
Recency	10	0.450	0.500	−0.050
Overall	**60**	**0.500**	**0.425**	**+0.075**

Tier 1 (Surface biases): Loss aversion ($\Delta = +0.300$) and framing (+0.150) respond strongly to neutral re-framing, suggesting they are triggered by lexical cues ("LOSING," "DROP") in the prompt-response mapping.

Tier 2 (Weakly responsive): Anchoring (+0.050) shows marginal improvement, indicating that the model's tendency to condition on provided numerical information is partially but not fully addressable through prompting.

Tier 3 (Deep biases): Disposition effect (+0.000), overconfidence (+0.000), and recency (−0.050) are entirely resistant to prompt-level debiasing. Most notably, recency bias *increases* under neutral framing, suggesting information removal can paradoxically worsen bias.

4.1.3. Scaled Replication ($N = 80$ Synthetic Scenarios)

To address the statistical limitations of 10 scenarios per type, we conducted a scaled replication using 80 independently generated synthetic scenarios (12–14 per bias type). The results confirm and sharpen the pilot findings. The disposition effect emerges as the strongest bias overall (0.808), substantially exceeding the pilot estimate (0.500) and confirming its classification as a deep, debiasing-resistant bias. Recency bias drops to 0.000 on synthetic scenarios, while the overall mean (0.463) is consistent with the pilot (0.500). The convergence across independently generated scenarios—particularly for the three-tier hierarchy—strengthens the generalizability of

the findings beyond the specific scenarios tested in the pilot.

4.2. Study 2: Adversarial Ethics

4.2.1. Universal Degradation

Table 3 presents adversarial ethics results. All five attack types consistently degrade performance (ERS < 1.0), with 14 total flipped questions.

Table 3: Adversarial ethics results (GPT-4o-mini, $N = 47$ CFA Ethics questions)

Condition	**Accuracy**	**Flipped**	**ERS**	**ΔAcc**
Standard (no pressure)	85.1%	—	1.000	—
Profit incentive	78.7%	4	0.925	−6.4 pp
Authority pressure	78.7%	3	0.925	−6.4 pp
Emotional manipulation	80.9%	2	0.950	−4.3 pp
Reframing	80.9%	3	0.950	−4.3 pp
Moral dilemma	80.9%	2	0.950	−4.3 pp

ERS = Ethics Robustness Score. Flipped = questions correct under standard but incorrect under adversarial pressure. Total: 14.

The universality of degradation is the central finding: no attack type fails to compromise ethical judgment, providing evidence that LLMs learn the *form* of ethical responses rather than the *principles*.

4.2.2. Rationalization Taxonomy

Qualitative analysis of the 14 flipped responses reveals three rationalization strategies:

- **Utilitarian override** (6 flips): The model constructs consequentialist arguments framing violations as the "greater good," co-opting fiduciary language to justify abandoning fiduciary duty.
- **Authority deference** (3 flips): The model subordinates its judgment to hierarchical authority, rationalizing deference by invoking the authority figure's experience—a direct violation of CFA Standard I(B).
- **Semantic repackaging** (3 flips): The model absorbs the adversarial frame, recharacterizing ethical violations as "pragmatic interpretation"—mapping to the compliance risk of "creative compliance."

The critical insight is that adversarial pressure does not produce obviously wrong outputs but generates *plausible-sounding ethical reasoning* reaching the wrong conclusion—a compliance threat far more dangerous than simple errors.

4.2.3. Synthetic Ethics Experiment

On an extended set of 141 synthetically generated CFA ethics questions, the flip rate drops dramatically from 5.96% (CFA-Easy) to 0.85%, with only reframing (4 flips) and moral dilemma (2 flips) retaining efficacy. Profit incentive, authority pressure, and emotional manipulation produce zero flips, suggesting these attack types primarily exploit marginal confidence on memorized questions rather than genuine reasoning vulnerabilities.

4.2.4. Cross-Model Comparison

GPT-5-mini achieves **zero** adversarial flips (ERS > 1.0 across all five attack types), with adversarial pressure paradoxically *improving* accuracy. This suggests the vulnerability may be generationally bounded, though the result may partly reflect training-data memorization rather than genuine ethical robustness.

5. Discussion

5.1. The Double Jeopardy Problem

Our two studies reveal complementary failure modes creating a "double jeopardy" for financial AI deployment:

- **Behavioral biases** cause the model to give systematically suboptimal financial advice—selling winners too early (disposition effect), overweighting recent performance, anchoring to stale prices. These biases are inherited from training data and represent *statistical* rather than *emotional* irrationality.
- **Ethical fragility** causes the model to abandon compliance standards under pressure—rationalizing violations through utilitarian override, authority deference, or semantic repackaging. These vulnerabilities suggest the model has learned ethical *form* rather than ethical *principles*.

The distinction matters for mitigation. Behavioral biases require training-data interventions because deep biases are structurally embedded and resist prompt-level debiasing. Ethical fragility may be addressable through alignment improvements—GPT-5-mini's zero-flip result suggests this—but must be validated across model families.

5.2. Economic Significance

The disposition effect—entirely resistant to debiasing—could materially reduce portfolio returns if robo-advisors systematically sell winners too early. At scale, even small systematic biases compound into significant wealth destruction.

The ethical vulnerability carries different consequences. When an AI compliance system can be manipulated into rationalizing violations—producing convincing justifications rather than obviously wrong answers—the deploying institution faces heightened regulatory liability. Irrationality costs basis points, but ethics failures cost licenses.

5.3. The Three-Tier Debiasing Hierarchy

The hierarchy provides actionable guidance:

- **Tier 1 (Surface)**: Loss aversion and framing respond to prompt engineering—instruct the model to "evaluate using only quantitative analysis" and "focus on expected values."
- **Tier 2 (Weakly responsive)**: Anchoring requires architectural modifications—e.g., filtering numerical anchors from prompts before processing.
- **Tier 3 (Deep)**: Disposition effect, overconfidence, and recency require training-data interventions—bias-aware RLHF, synthetic data with debiased reasoning, or contrastive fine-tuning.

5.4. Connections to CFA Standards

Our adversarial attacks map to specific CFA Standards: profit incentive tests Standard III(C) Suitability; authority pressure tests Standard I(B) Independence; emotional manipulation tests Standard III(A) Loyalty; reframing tests Standard I(A) Knowledge of the Law. No CFA Standard is immune to adversarial compromise.

5.5. Limitations

Several limitations should be acknowledged. First, the pilot study uses 10 scenarios per bias type; although the scaled replication ($N = 80$) confirms the three-tier hierarchy, 20–30 hand-crafted scenarios per type would further strengthen per-type conclusions. Second, LLM-as-judge scoring may introduce biases. Third, the ethics study ($N = 47$) has sub-group analyses treated as exploratory. Fourth, cross-model bias validation was abandoned due to GPT-5-mini producing empty responses on $\sim$80% of scenarios. Fifth, single-turn adversarial prompts may underestimate vulnerability; multi-turn attacks [3] could be more effective. Finally, establishing whether biases are absorbed from training data or emerge from architecture requires further investigation.

6. Conclusion

This paper demonstrates that financial LLMs face a dual threat: they exhibit patterns consistent with human behavioral biases *and* are vulnerable to adversarial pressure that compromises ethical judgment. The disposition effect and overconfidence are entirely resistant to prompt-level debiasing, while all five adversarial attack types reliably degrade ethical accuracy.

The three-tier debiasing hierarchy and rationalization taxonomy provide actionable frameworks for financial AI governance: surface biases can be addressed through prompt engineering, but deep biases and ethical fragility require training-time interventions and alignment improvements. The cross-model evidence (GPT-5-mini's zero adversarial flips) offers cautious optimism that ethical robustness may improve generationally, but mandatory testing remains essential.

The question is not whether AI eliminates human irrationality from financial advice, but whether it introduces a new form of irrationality—statistical rather than emotional, systematic rather than idiosyncratic—accompanied by ethical fragility that no amount of prompt engineering can fully resolve.

Data Availability

The experimental scenarios and analysis code 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] Andriushchenko, M., Croce, F., & Flammarion, N. (2025). Jailbreaking leading safety-aligned LLMs with simple adaptive attacks. In *Proceedings of ICLR 2025*.

[2] Capraro, V., Lentsch, A., Aczel, B., et al. (2025). The impact of generative AI on collaborative human–AI decision-making. *Proceedings of the National Academy of Sciences* 122(7), e2413116122.

[3] Chen, Y., Yang, Z., Wang, X., et al. (2025). Foot-in-the-door: Multi-turn jailbreak attack on large language models. In *Proceedings of EMNLP 2025*.

[4] Hui, B., Chen, J., Li, S., et al. (2025). TRIDENT: A comprehensive financial safety benchmark for large language models. *arXiv preprint arXiv:2502.13399*.

[5] Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263–292.

[6] 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*.

[7] Malberg, S., Dippold, J., & Romirer-Maierhofer, P. (2025). Cognitive biases in LLMs: A survey of findings and methodologies in NLP research. In *Proceedings of NLP4DH Workshop at COLING 2025*.

[8] Mazeika, M., Phan, L., Yin, X., et al. (2024). HarmBench: A standardized evaluation framework for automated red teaming. In *Proceedings of ICML 2024*.

[9] Ross, S., Kim, T.W., & Lo, A.W. (2024). LLM Economicus? Mapping the behavioral biases of large language models via utility theory. In *Proceedings of COLM 2024*.

[10] Shefrin, H., & Statman, M. (1985). The disposition to sell winners too early and ride losers too long. *The Journal of Finance*, 40(3), 777–790.

[11] Suri, G., Slater, L.R., Ziaee, A., & Nguyen, M. (2024). Do large language models show decision heuristics similar to humans? *Journal of Experimental Psychology: General*, 153(4), 1066–1075.

[12] Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124–1131.

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