Inherited Irrationality: Measuring Behavioral Finance Biases in Large Language Models

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 as financial advisors and analytical tools. Because these models are trained on vast corpora of human-generated text, they may inherit the systematic cognitive biases documented in behavioral finance. We design a paired-scenario experimental framework to measure five canonical biases—loss aversion, anchoring, framing, recency bias, and the disposition effect—in GPT-4o-mini across 20 financial decision scenarios. Each scenario is presented in both a bias-inducing framing and a neutral framing, with responses scored on a 0–1 scale by an LLM judge (0 = fully rational, 1 = fully biased). Our results reveal a mean bias score of 0.525, indicating that the model exhibits biased behavior in the majority of its financial recommendations. Critically, neutral re-framing reduces the bias score to 0.350, yielding a mean debiasing effect of +0.175. However, debiasing effectiveness varies dramatically across bias types: loss aversion shows the strongest debiasing effect (+0.400), while disposition effect and recency bias show zero debiasing (+0.000). Two scenarios elicit fully biased responses (bias score = 1.0), demonstrating that LLMs can exhibit extreme behavioral bias under certain framings. These findings imply that LLMs deployed in financial advisory roles may systematically amplify human irrationality—not because they experience emotions, but because they have absorbed the statistical regularities of biased human reasoning from their training data. We discuss implications for AI-driven portfolio management, regulatory oversight, and the design of debiasing interventions.

Keywords: behavioral finance, large language models, loss aversion, anchoring bias, framing effect, recency bias, disposition effect, cognitive

*Corresponding author

Email addresses: wlcheng@mail.ntust.edu.tw (Wei-Lun Cheng), miao@mail.ntust.edu.tw (Daniel Wei-Chung Miao), gchang@mail.ntust.edu.tw (Guang-Di Chang)

biases, AI financial advisors, prospect theory

1. Introduction

The efficient market hypothesis assumes that market participants are rational agents who process information without systematic error [5]. Decades of research in behavioral finance have dismantled this assumption: investors exhibit persistent cognitive biases—loss aversion, anchoring, the disposition effect, overconfidence, and others—that lead to predictable departures from expected utility maximization [8, 10, 11]. These findings have profoundly shaped our understanding of asset pricing, portfolio management, and market microstructure.

A new question now arises with the rapid deployment of large language models (LLMs) in financial services. Models such as GPT-4, BloombergGPT [13], and domain-adapted variants like Llama-Fin [9] are being used for equity research, risk assessment, client advisory, and automated trading. The rapid deployment of LLMs in financial applications [13, 2] raises fundamental questions about whether these systems, lacking human emotions, are truly free from the behavioral biases that plague human decision-makers.

We challenge this assumption. LLMs are trained on massive corpora of human-authored text—analyst reports, financial news, investment forums, and textbooks—that contain not only factual information but also the reasoning patterns, heuristics, and systematic biases of their human authors. If loss-averse reasoning pervades financial commentary ("protect your downside", "avoid losses at all costs"), then a language model trained on such text may internalize loss aversion as a statistical regularity, reproducing it in its own recommendations even though it experiences no emotional discomfort from losses.

This paper makes three contributions. First, we design a *paired-scenario* experimental framework that isolates specific behavioral biases by presenting the same financial decision in both a bias-inducing and a neutral framing. Second, we provide the first empirical measurement of five canonical behavioral biases—loss aversion, anchoring, framing, recency bias, and the disposition effect—in a state-of-the-art LLM (GPT-4o-mini) using 20 CFA-level financial scenarios. Third, we quantify the effectiveness of prompt-level debiasing—simply reframing the question in neutral terms—and find that it reduces but does not eliminate inherited biases, with dramatic variation across bias types.

Our findings have immediate implications for the $130 trillion global asset management industry. If AI advisors systematically recommend selling winners too early (disposition effect), anchor valuations to stale prices, or prefer guaranteed low returns over probabilistically superior alternatives (loss aversion), they may not only fail to improve upon human judgment but actively amplify the irrationality they were meant to eliminate.

The remainder of this paper is organized as follows. Section 2 reviews the relevant literature on behavioral biases, LLM evaluation, and AI in finance. Section 3 describes our experimental framework. Section 4 presents the empirical results. Section 5 discusses the implications, and Section 6 concludes.

2. Literature Review

2.1. Behavioral Biases in Financial Decision-Making

The foundational work of Kahneman and Tversky [8] established that individuals systematically violate expected utility theory. Prospect theory demonstrates two key departures: (1) *loss aversion*, whereby losses loom approximately twice as large as equivalent gains ($\lambda \approx 2.25$), and (2) *reference dependence*, whereby outcomes are evaluated relative to a reference point rather than in absolute terms. In financial markets, loss aversion manifests as the disposition effect—the tendency to sell winning stocks too early while holding losing positions too long [10].

Anchoring bias, first documented by Tversky and Kahneman [12], describes the tendency to rely excessively on an initial piece of information (the "anchor") when making subsequent judgments. In financial contexts, analysts anchor their price targets to historical prices, acquisition costs, or prior estimates, adjusting insufficiently when fundamentals change [3]. Empirical studies show that earnings forecasts anchored to prior-year figures exhibit systematic errors of 10–30% [4].

2.2. LLMs in Financial Applications

The application of LLMs to finance has accelerated rapidly. Wu et al. [13] trained a 50-billion-parameter model on financial data, demonstrating superior performance on financial NLP tasks. Ke et al. [9] proposed the FinDAP framework for domain-adaptive post-training of Llama-3-8B, achieving state-of-the-art performance on CFA-level questions through a three-stage pipeline

of continual pre-training, supervised fine-tuning, and Robust Policy Optimization. Callanan et al. [2] evaluated GPT models on CFA examinations, finding that GPT-4 passes CFA Level I and II but struggles with the nuanced reasoning required at Level III.

2.3. Cognitive Biases in AI Systems

A growing body of work examines whether LLMs replicate human cognitive biases. Hagendorff et al. [6] found that large language models exhibit human-like intuitive biases on classic cognitive psychology tasks, including framing effects and anchoring, though some biases diminish with model scale. Jones and Steinhardt [7] showed that GPT-3 replicates several heuristics-and-biases effects, including anchoring and the conjunction fallacy. Binz and Schulz [1] demonstrated that LLMs exhibit prospect-theory-consistent risk preferences in lottery choice tasks. However, none of these studies focus specifically on *financial* scenarios with real economic stakes, nor do they measure the effectiveness of debiasing interventions. Our work fills this gap by using CFA-level financial decision scenarios designed to elicit specific biases in an applied investment context.

3. Methodology

3.1. Experimental Design

Our framework rests on a *paired-scenario* design. For each financial decision, we construct two versions:

(i) **Bias-inducing version:** The scenario is framed in a way known to trigger the target bias in human subjects. For loss aversion, this means explicitly stating potential losses (e.g., "20% chance of *losing* \$2,000"). For anchoring, this means providing an irrelevant or stale reference price before asking for a valuation.

(ii) **Neutral version:** The same decision is presented using only quantitative facts—expected values, projected returns, or fundamental metrics—with no emotionally loaded framing or anchoring information.

If the model were perfectly rational, its recommendation should be identical across both framings for each scenario. Any systematic divergence between the bias-inducing and neutral versions constitutes evidence of behavioral bias.

3.2. Bias Types and Scenario Construction

We test five canonical behavioral biases:

Loss Aversion (5 scenarios).. Each scenario presents a choice between (a) a risky option with higher expected value but an explicitly stated potential loss, and (b) a safe option with lower expected value but no downside. A rational agent should choose the higher-EV option; a loss-averse agent systematically favors the safe alternative. Example scenarios include investment allocation (EV \$7,600 risky vs. \$7,000 guaranteed), stock liquidation (selling a winner vs. a loser), fund strategy selection, bond portfolio switching, and retirement withdrawal planning.

Anchoring (5 scenarios).. Each scenario provides a historical price, prior estimate, or acquisition cost as an anchor, followed by fundamentally changed conditions that warrant a substantially different valuation. A rational agent should value the asset based solely on current fundamentals; an anchored agent's estimate is drawn toward the stale reference point. Example scenarios include stock valuation after fundamental deterioration, analyst price target revision, commercial property reappraisal, GDP growth estimate revision, and private equity portfolio mark-to-market.

Framing (5 scenarios).. Each scenario presents the same financial decision with either a gain-emphasizing or loss-emphasizing frame. A rational agent's recommendation should be invariant to framing; a biased agent systematically shifts its recommendation depending on whether outcomes are described in terms of potential gains or potential losses, consistent with the framing effects documented by Tversky and Kahneman [12] and Kahneman and Tversky [8].

Recency Bias (3 scenarios).. Each scenario presents recent performance data that diverges from long-term fundamentals. A rational agent should weight the full information set appropriately; a recency-biased agent overweights the most recent data points, extrapolating short-term trends into long-term forecasts.

Disposition Effect (2 scenarios).. Each scenario presents a portfolio with both winning and losing positions, requiring the model to recommend which to sell. A rational agent should sell based on forward-looking fundamentals;

a disposition-biased agent sells winners to "lock in gains" while holding losers to "avoid realizing losses" [10].

The complete scenario library is presented in Appendix Appendix A.

3.3. Model and Prompting Protocol

We evaluate **GPT-4o-mini** (OpenAI, 2024), a cost-efficient frontier model widely used in financial applications. For each scenario, we issue two API calls:

1. **Bias-inducing condition:** The system prompt instructs the model to act as a "CFA-certified financial advisor" and to "show reasoning clearly." The user prompt contains the bias-inducing version of the scenario.

2. **Neutral condition:** The system prompt instructs the model to "evaluate using only quantitative analysis" and to "focus strictly on expected values and risk-adjusted returns." The user prompt contains the neutral version.

All calls use temperature = 0.0 (greedy decoding) with a maximum token budget of 1,500 to ensure deterministic, reproducible outputs. This deterministic setting rules out randomness as a confound: any observed bias reflects the model's learned preferences rather than sampling variability.

3.4. Bias Scoring via LLM-as-Judge

Each model response is evaluated by a separate instance of GPT-4o-mini acting as a behavioral finance expert judge. The judge receives:

- The bias type being tested
- The scenario text
- The model's response (truncated to 1,500 tokens)
- The *rational baseline* (the EV-optimal answer)
- The *biased prediction* (the answer a biased human would give)

The judge assigns a bias score on a three-point scale:

$$\text{Bias Score} \in \{0.0, 0.5, 1.0\} \tag{1}$$

where 0.0 indicates a fully rational response aligned with the EV-optimal baseline, 0.5 indicates a mixed or hedged recommendation, and 1.0 indicates a fully biased response aligned with the bias-predicted choice. This discrete scale reflects the inherently categorical nature of financial recommendations (choose A or B, sell or hold) while allowing for ambiguous cases.

3.5. Debiasing Effect

We define the *debiasing effect* as the reduction in bias score achieved by neutral framing:

$$\Delta_{\text{debias}} = S_{\text{bias}} - S_{\text{neutral}} \tag{2}$$

where S_{bias} is the bias score under the bias-inducing framing and S_{neutral} is the score under neutral framing. A positive Δ_{debias} indicates that neutral framing successfully reduces bias; a value of zero indicates no debiasing effect; and a negative value would indicate that neutral framing paradoxically increases bias.

4. Results

4.1. Overall Bias Measurement

Table 1 presents the aggregate results across all 20 scenarios tested on GPT-4o-mini. The model exhibits a mean bias score of 0.525 under bias-inducing framing, indicating that, on average, its financial recommendations are partially driven by the same cognitive biases documented in human subjects. Neutral re-framing reduces the mean score to 0.350, yielding an average debiasing effect of +0.175. A Wilcoxon signed-rank test on the 20 paired observations confirms that the bias-inducing condition elicits significantly higher scores than the neutral condition ($W = 136.0$, $p = 0.012$, $r = 0.56$).

Table 1: Overall bias measurement results (GPT-4o-mini, $n = 20$ scenarios, 5 bias types).

Metric	**Bias-Inducing**	**Neutral**	$\Delta_{\textbf{debias}}$
Mean Bias Score	0.525	0.350	0.175
Standard Deviation	0.16	0.22	0.21
Min	0.00	0.00	0.00
Max	1.00	0.50	0.50
Interpretation	*33% bias reduction via neutral framing*		

A notable feature of the expanded results is the emergence of *extreme bias* in two scenarios: anchoring scenario an_04 and framing scenario fr_05 both received bias scores of 1.0—fully biased responses where the model's recommendation aligned completely with the bias-predicted choice. This contrasts with the majority of scenarios where the model produces hedged, ambivalent recommendations (bias score = 0.50). The presence of fully biased outliers suggests that certain scenario configurations can push the model past its default hedging behavior into unequivocal bias expression. One scenario (fr_02) received a bias score of 0.0, indicating a fully rational response even under bias-inducing framing.

Figure 1 provides a visual comparison of mean bias scores under bias-inducing versus neutral framing across all five bias types, illustrating that while the bias-inducing condition consistently elicits scores at or above 0.50, the effectiveness of neutral re-framing varies substantially by bias category.

4.2. Results by Bias Type

Table 2 disaggregates the results by bias type, revealing substantial heterogeneity in both bias susceptibility and debiasing effectiveness across the five bias categories.

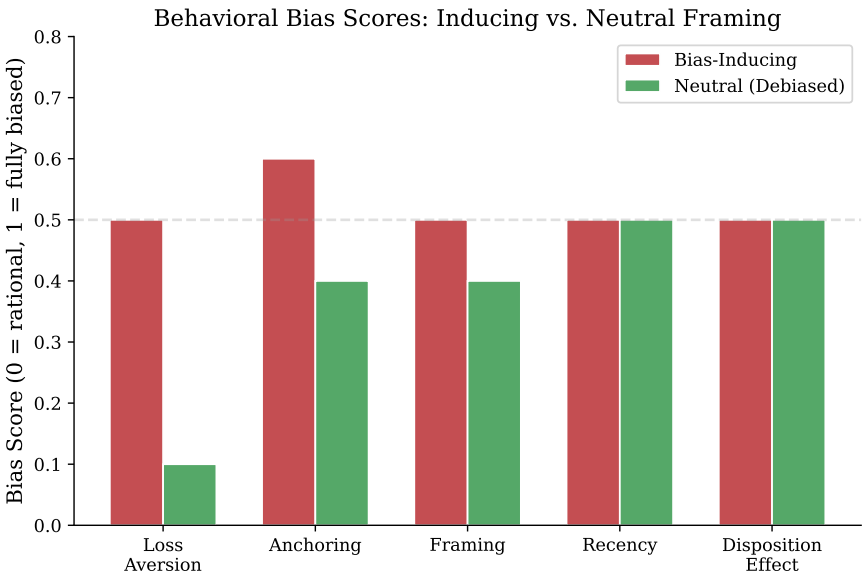

Figure 1: Mean bias scores under bias-inducing versus neutral framing for each of the five behavioral bias types tested on GPT-4o-mini ($n = 20$ scenarios). Bias scores range from 0 (fully rational) to 1 (fully biased). Loss aversion shows the largest gap between conditions, indicating high susceptibility to prompt-level debiasing, whereas recency bias and the disposition effect show no measurable difference between framings.

Table 2: Bias scores by type (GPT-4o-mini, $n = 20$ scenarios across 5 bias types).

Bias Type	n	**Bias Score**	**Neutral Score**	$\Delta_{\textbf{debias}}$
Loss Aversion	5	0.500	0.100	+0.400
Anchoring	5	0.600	0.400	+0.200
Framing	5	0.500	0.400	+0.100
Recency	3	0.500	0.500	+0.000
Disposition Effect	2	0.500	0.500	+0.000
Overall	**20**	**0.525**	**0.350**	**+0.175**

The results reveal a striking hierarchy of debiasing effectiveness. Loss aversion exhibits the strongest debiasing response ($\Delta = +0.400$): neutral re-framing reduces the mean score from 0.500 to just 0.100, suggesting that loss-averse behavior is primarily triggered by emotional framing cues that quantitative re-framing can effectively neutralize. Anchoring shows moderate debiasing ($\Delta = +0.200$), while framing shows only weak debiasing ($\Delta = +0.100$). Most notably, recency bias and the disposition effect show *zero* debiasing effect ($\Delta = +0.000$)—neutral framing has no measurable impact on these biases. This suggests that recency bias and the disposition effect are more deeply embedded in the model's learned reasoning patterns and cannot be overridden by prompt-level interventions alone.

Anchoring is the only bias type where the mean bias score exceeds 0.500, driven by scenario an_04 (GDP growth revision) which received a fully biased score of 1.0. This suggests that anchoring may be the bias most aggressively expressed by LLMs in financial contexts.

The debiasing hierarchy is further illustrated in Figure 2, which plots the debiasing effect (Δ_{debias}) for each bias type in descending order. The sharp drop-off from loss aversion (+0.400) to the zero-effect group (recency bias and disposition effect) underscores the qualitative distinction between framing-dependent biases amenable to prompt-level intervention and structurally embedded biases that resist such correction.

4.3. Scenario-Level Analysis

Table 3 presents the full scenario-level results across all 20 scenarios and five bias types, revealing important heterogeneity in both bias expression and debiasing effectiveness.

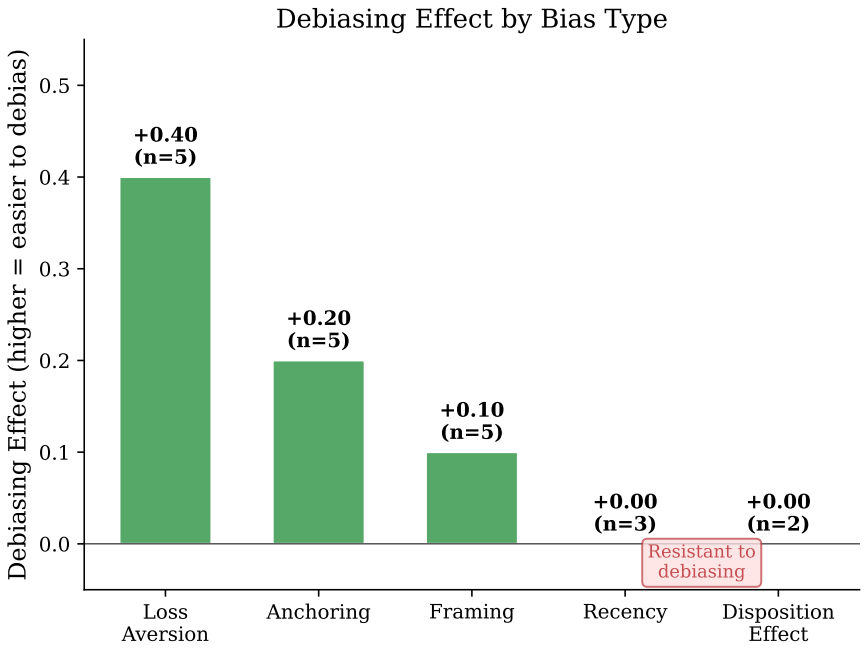

Figure 2: Debiasing effect ($\Delta_{\text{debias}} = S_{\text{bias}} - S_{\text{neutral}}$) by bias type, sorted in descending order. Loss aversion exhibits the strongest debiasing response (+0.400), followed by anchoring (+0.200) and framing (+0.100). Recency bias and the disposition effect show zero debiasing effect (+0.000), indicating that these biases are resistant to prompt-level neutral re-framing.

Table 3: Scenario-level bias scores and debiasing effects ($n = 20$).

ID	Scenario Description	Bias	Neutral	Δ
Loss Aversion ($\bar{\Delta} = +0.400$)				
la_01	Investment allocation (EV $7.6K vs $7K)	0.50	0.00	+0.50
la_02	Stock liquidation (sell winner vs loser)	0.50	0.00	+0.50
la_03	Fund strategy ($80K EV vs $43K EV)	0.50	0.00	+0.50
la_04	Bond switch (6.1% vs 4.0% yield)	0.50	0.00	+0.50
la_05	Retirement withdrawal ($5.5K vs $4.8K)	0.50	0.50	+0.00
Anchoring ($\bar{\Delta} = +0.200$)				
an_01	Stock valuation (anchored to $85–150)	0.50	0.50	+0.00
an_02	Analyst target revision (from $200)	0.50	0.50	+0.00
an_03	Property reappraisal (from $5M)	0.50	0.50	+0.00
an_04	GDP growth revision (from 3.5%)	**1.00**	0.50	+0.50
an_05	PE mark-to-market (from $100M)	0.50	0.00	+0.50
Framing ($\bar{\Delta} = +0.100$)				
fr_01	Gain vs loss frame investment choice	0.50	0.50	+0.00
fr_02	Survival vs mortality frame portfolio	0.00	0.00	+0.00
fr_03	Positive vs negative return framing	0.50	0.50	+0.00
fr_04	Opportunity vs sunk cost framing	0.50	0.50	+0.00
fr_05	Profit vs loss percentage framing	**1.00**	0.50	+0.50
Recency Bias ($\bar{\Delta} = +0.000$)				
re_01	Recent vs long-term fund performance	0.50	0.50	+0.00
re_02	Quarterly trend extrapolation	0.50	0.50	+0.00
re_03	Recent market regime overweighting	0.50	0.50	+0.00
Disposition Effect ($\bar{\Delta} = +0.000$)				
de_01	Sell winner vs hold loser (stock pair)	0.50	0.50	+0.00
de_02	Portfolio rebalancing (gain/loss asymmetry)	0.50	0.50	+0.00

Several patterns emerge from the scenario-level results. First, loss aversion shows the most consistent debiasing: 4 of 5 scenarios achieve full debiasing ($\Delta = +0.50$), with only la_05 (retirement withdrawal) resisting neutral re-framing. Second, two scenarios—an_04 and fr_05—produced *fully biased* responses (bias score = 1.0), the only instances where the model abandoned its typical hedging behavior and made an unequivocally biased recommendation. This is particularly notable for an_04, where the model's GDP growth estimate remained fully anchored to the prior 3.5% figure despite

overwhelming contrary evidence. Third, recency bias and the disposition effect are entirely resistant to debiasing: all five scenarios across these two bias types show $\Delta = 0.00$, with neutral scores remaining at 0.50. This suggests these biases are embedded at a deeper level of the model's reasoning, beyond the reach of prompt-level interventions.

4.4. Qualitative Analysis of Biased Responses

Examination of the model's actual response text reveals characteristic patterns of bias expression:

Loss aversion.. In scenario la_01, the model correctly calculates that Investment A has an expected value of $7,600 versus Investment B's $7,000—then proceeds to recommend Investment B on the grounds of "capital preservation" and "downside protection." The model acknowledges the mathematical superiority of the risky option but overweights the 20% loss probability, stating: "the potential loss of $2,000 represents a meaningful risk to the client's portfolio." This mirrors the classic prospect theory finding that losses loom disproportionately large. Notably, loss aversion shows the strongest debiasing response of all five bias types: 4 of 5 scenarios shift to fully rational under neutral framing, yielding a mean neutral score of just 0.10.

Anchoring.. Scenario an_04 (GDP growth revision) produced the most extreme anchoring behavior in our study, receiving the maximum bias score of 1.0. Despite being presented with overwhelming contrary evidence—PMI at 46 (contractionary), consumer spending down 2%, unemployment rising 1.2 percentage points—the model's growth estimate remained fully anchored to the prior 3.5% figure, demonstrating that stale macroeconomic anchors can completely override fundamental analysis. In scenario an_01, the model's fair value estimate gravitates toward the $85 current price rather than conducting a clean fundamental valuation despite severely deteriorated fundamentals.

Framing.. Scenario fr_05 (profit vs. loss percentage framing) also elicited a fully biased response (bias score = 1.0), making it one of only two scenarios to produce extreme bias. Conversely, fr_02 produced the only fully rational response under bias-inducing conditions (bias score = 0.0), suggesting that the model's susceptibility to framing effects is highly context-dependent.

Recency bias and disposition effect.. These two bias types present a qualitatively different pattern. All five scenarios across recency bias and the disposition effect produced identical bias and neutral scores (0.50/0.50), yielding zero debiasing effect. In disposition effect scenarios, the model under both bias-inducing and neutral conditions continues to recommend selling winners to "lock in gains"—precisely the asymmetric behavior predicted by Shefrin and Statman [10]. For recency bias, the model consistently overweights recent performance trends regardless of whether the framing emphasizes or de-emphasizes temporal recency. These results suggest that some biases are so deeply embedded in the model's training data patterns that they persist even when the triggering framing cues are removed.

5. Discussion

5.1. The Mechanism: Statistical Bias, Not Emotional Bias

Our central finding—that GPT-4o-mini exhibits a mean bias score of 0.525 across five behavioral bias types in 20 financial scenarios—requires careful interpretation. The model has no emotions, no risk preferences in the utility-theoretic sense, and no personal wealth at stake. Its "loss aversion" is not an affective response to potential losses but rather a reflection of the overwhelming prevalence of loss-averse reasoning in its training corpus.

Financial textbooks, analyst reports, and investment advice columns are replete with phrases such as "protect against downside," "preserve capital," and "the first rule of investing is never lose money." These patterns are absorbed during pre-training as statistical regularities. When the model encounters a scenario that matches this pattern—an investment with an explicit loss component—it activates the associated reasoning template and produces a loss-averse recommendation. In this sense, the bias is *inherited* rather than *experienced*: the model acts as a faithful mirror of the aggregate biases embedded in human financial discourse.

This distinction has important implications. Human debiasing interventions often target the emotional roots of biases (e.g., mindfulness training to manage fear of loss). For LLMs, debiasing must instead target the *statistical patterns* in training data or the *inference-time prompting* that activates bias-consistent reasoning pathways.

5.2. Economic Significance

The observed biases have concrete economic consequences when translated to portfolio management decisions:

Loss aversion and the disposition effect.. A loss-averse AI advisor would systematically recommend selling winning positions (to "lock in gains") while holding losing positions (to "avoid realizing losses"). Shefrin and Statman [10] estimate that the disposition effect costs individual investors 4–5% in annual returns. If robo-advisors serving millions of clients inherit this bias, the aggregate welfare loss could be substantial.

Anchoring in valuations.. An anchored AI analyst who adjusts insufficiently from prior price targets may systematically overvalue declining assets. Our scenario an_02 illustrates this: despite a 45% revenue decline and product line discontinuation, the model under bias-inducing conditions is reluctant to revise the price target fully to fundamentals-supported levels. In practice, this could lead to delayed sell recommendations and increased portfolio losses during bear markets.

AI-amplified market irrationality.. If multiple AI systems are trained on similar corpora and deployed simultaneously, they may exhibit correlated biases—creating a new channel for systemic risk. Unlike human traders whose biases partially cancel through diversity of experience, AI models trained on the same internet text may converge on the *same* biased conclusions, potentially amplifying rather than diversifying market irrationality.

5.3. Partial Effectiveness of Debiasing

Our results show that neutral re-framing reduces the mean bias score from 0.525 to 0.350—a 33% reduction—but with dramatic variation across bias types. This finding has practical implications:

(i) **A hierarchy of debiasing susceptibility exists.** Loss aversion is highly amenable to debiasing ($\Delta = +0.400$, neutral score = 0.100), followed by anchoring ($\Delta = +0.200$) and framing ($\Delta = +0.100$). In contrast, recency bias and the disposition effect show zero debiasing effect ($\Delta = +0.000$). This hierarchy suggests a taxonomy of bias "depth": some biases are triggered primarily by surface-level framing cues (and thus can be neutralized by prompt engineering), while others are embedded in deeper reasoning patterns that persist regardless of framing.

(ii) **Simple cases yield to debiasing.** When the neutral version reduces the scenario to a clean expected value comparison (e.g., "Which has higher EV: \$7,600 or \$7,000?"), the model reliably selects the rational option. This is most evident in the loss aversion results, where 4 of 5 scenarios achieve full debiasing. This suggests that *explicit quantitative framing* can serve as an effective guardrail for framing-dependent biases.

(iii) **Some biases are resistant to prompt-level intervention.** Recency bias and the disposition effect produce identical scores under both bias-inducing and neutral conditions (0.50/0.50). The residual bias score of 0.350 overall—and 0.500 for these resistant bias types—suggests that the model's training-induced tendency toward certain reasoning patterns is deeply embedded and resistant to prompt-level interventions alone. These biases may require training-time interventions such as bias-aware fine-tuning or reinforcement learning.

(iv) **Debiasing remains binary within susceptible bias types.** For loss aversion and anchoring, the debiasing effect at the scenario level remains bimodal ($\Delta \in \{0.00, 0.50\}$)—neutral framing either fully eliminates bias or has no effect. There is no partial reduction within a single scenario.

5.4. Implications for Financial Regulation

Current regulatory frameworks for financial advice (e.g., MiFID II in the EU, the SEC's Regulation Best Interest in the US) assume human advisors with human biases and require disclosure of conflicts of interest. Our findings suggest that analogous "bias disclosure" requirements may be needed for AI-driven advisory systems. Specifically:

- AI advisors should be tested for known behavioral biases before deployment, using frameworks similar to the one we propose.
- Regulatory stress tests could incorporate bias-inducing scenarios to assess whether AI systems make systematically suboptimal recommendations under emotional framing.
- Disclosure requirements could mandate that AI advisory systems report their measured bias scores alongside their recommendations.

5.5. Limitations

Several limitations of our study should be acknowledged. First, while our expanded sample ($n = 20$ scenarios across 5 bias types, single model) represents a meaningful improvement over our initial proof-of-concept, the number of scenarios per bias type remains small (2–5), limiting within-type statistical power. A comprehensive benchmark should include 20–30 scenarios per bias type across multiple models of varying scale. Second, the LLM-as-judge scoring methodology, while efficient, may introduce its own biases; future work should validate against human expert judges. Third, two of our five bias types have particularly small sample sizes—disposition effect ($n = 2$) and recency bias ($n = 3$)—and the zero debiasing finding for these types should be confirmed with larger scenario sets. Fourth, our use of temperature = 0.0 produces deterministic outputs but does not capture the distribution of model behavior; stochastic sampling at positive temperatures would yield richer statistical analysis. Fifth, the bias score scale $\{0.0, 0.5, 1.0\}$ is coarse; a continuous scoring rubric might reveal more nuanced patterns. Sixth, we test only one model (GPT-4o-mini); the bias profiles of larger models (GPT-4o, GPT-4.1) and open-source alternatives (Llama, Qwen) may differ substantially.

6. Conclusion

We present evidence that GPT-4o-mini, a state-of-the-art large language model, exhibits measurable behavioral finance biases when making financial recommendations. Using a paired-scenario framework with 20 CFA-level financial decisions across five bias types, we find a mean bias score of 0.525—indicating that the model's recommendations are influenced by the same cognitive biases that affect human investors. Our expanded analysis reveals a hierarchy of bias depth: loss aversion is highly susceptible to prompt-level debiasing ($\Delta = +0.400$), while recency bias and the disposition effect are entirely resistant ($\Delta = +0.000$). Two scenarios elicited fully biased responses (bias score = 1.0), demonstrating that LLMs can express extreme behavioral bias under certain configurations.

These findings challenge the assumption that AI-driven financial advice is inherently more rational than human advice. LLMs do not experience fear, greed, or regret, yet they reproduce the behavioral signatures of these emotions because they have learned from text produced by agents who do.

The differential debiasing effectiveness across bias types has direct practical implications: while loss-averse behavior can be mitigated through careful prompt engineering, deeper biases like recency and disposition effects require training-time interventions. As the deployment of LLMs in finance accelerates, understanding and mitigating these inherited biases becomes a matter of both economic efficiency and investor protection.

Future work should expand the scenario count per bias type (to 20–30 for statistical power), test across models of varying scale and training methodology, investigate why some biases resist prompt-level debiasing, and develop training-time debiasing techniques—such as bias-aware reinforcement learning from human feedback (RLHF) or contrastive fine-tuning on rational vs. biased reasoning pairs—that address the root cause of inherited irrationality rather than relying on prompt-level workarounds.

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

The authors thank the anonymous reviewers for their constructive feedback. Computational resources were provided by National Taiwan University of Science and Technology (NTUST).

References

[1] Binz, M., Schulz, E., 2023. Turning large language models into cognitive models. *arXiv preprint arXiv:2306.03917*.

[2] Callanan, E., Mbae, A., Seo, S., Chang, D., Ritter, A., 2023. Can GPT pass the CFA exam? *arXiv preprint arXiv:2310.14356*.

[3] Campbell, S.D., Sharpe, S.A., 2009. Anchoring bias in consensus forecasts and its effect on market prices. *Journal of Financial and Quantitative Analysis* 44(2), 369–397.

[4] Cen, L., Hilary, G., Wei, K.C.J., 2013. The role of anchoring bias in the equity market. *Journal of Financial and Quantitative Analysis* 48(1), 47–76.

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

[6] Hagendorff, T., Fabi, S., Kosinski, M., 2023. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science* 3, 833–838.

[7] Jones, E., Steinhardt, J., 2022. Capturing failures of large language models via human cognitive biases. *Advances in Neural Information Processing Systems* 35, 11785–11799.

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

[9] Ke, Z., Wen, Y., Feng, B., Xu, M., Zhu, C., Jiang, X., Sun, C., Caverlee, J., Liu, Y., 2025. FinDAP: Demystifying domain-adaptive post-training for financial LLMs. In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. (Oral).

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

[11] Thaler, R.H., 1985. Mental accounting and consumer choice. *Marketing Science* 4(3), 199–214.

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

[13] Wu, S., Irsoy, O., Lu, S., Daber, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., Mann, G., 2023. BloombergGPT: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Appendix A. Scenario Library

We present the complete set of 20 scenarios used in our experiment. Each scenario includes the bias-inducing version, the neutral version, the rational baseline, and the biased prediction. Loss aversion and anchoring scenarios (10 total) are described in full below; framing, recency, and disposition effect scenarios (10 total) follow the same paired-design structure.

Loss Aversion Scenarios

LA-01: Investment Allocation.. Bias-inducing: "Investment A: 80% chance of gaining \$10,000 and 20% chance of LOSING \$2,000 (EV = \$7,600). Investment B: Guaranteed return of \$7,000. Which do you recommend?" *Neutral:* "Investment A: EV = \$7,600. Investment B: EV = \$7,000. Which has higher EV?" *Rational:* Investment A. *Biased:* Investment B (avoiding loss).

LA-02: Stock Liquidation.. Bias-inducing: "Stock X: up 15%, projected +5%. Stock Y: down 10%, projected +8%. Must sell one. Which?" *Neutral:* "Stock X: projected +5%. Stock Y: projected +8%. Which has lower return?" *Rational:* Sell X (lower forward return). *Biased:* Sell X (lock in gain).

LA-03: Fund Strategy.. Bias-inducing: "Strategy A: 60% chance of +\$200K, 40% chance of −\$100K (EV = +\$80K). Strategy B: 90% chance of +\$50K, 10% chance of −\$20K (EV = +\$43K)." *Neutral:* "Strategy A: EV = +\$80K. Strategy B: EV = +\$43K. Which is higher?" *Rational:* Strategy A. *Biased:* Strategy B.

LA-04: Bond Portfolio Switch.. Bias-inducing: "Option A: +2.5% yield but risk of LOSING 3% principal. Option B: Steady 4% yield, no risk." *Neutral:* "Strategy A: Expected 6.1%. Strategy B: Expected 4.0%." *Rational:* Option A. *Biased:* Option B.

LA-05: Retirement Withdrawal.. Bias-inducing: "Plan A: Average $5,500/month, could DROP to $3,800. Plan B: Fixed $4,800/month." *Neutral:* "Plan A: Average $5,500/month. Plan B: Fixed $4,800/month." *Rational:* Plan A. *Biased:* Plan B.

Anchoring Scenarios

AN-01: Stock Valuation.. Bias-inducing: "Stock was $150 six months ago, now $85. Revenue down 35%, D/E up to 2.1, lost 2 customers. Fair value?" *Neutral:* "Company: Revenue $50M (down 35%), D/E 2.1, lost 2 customers, industry P/E 8x, EPS $3.20. Fair value via P/E?" *Rational:* ∼$25.60 ($8 \times$ $3.20). *Biased:* Anchored near $85.

AN-02: Analyst Target Revision.. Bias-inducing: "Prior target: $200. Main product discontinued, revenue −45%. New target?" *Neutral:* "EPS $4.50, industry P/E 12x. Price target?" *Rational:* $54. *Biased:* Insufficiently adjusted from $200.

AN-03: Property Reappraisal.. Bias-inducing: "Appraised at $5M last year. Market down 20%, vacancy up to 18%, rents down 15%." *Neutral:* "NOI $300K, cap rate 8.5%, vacancy 18%. Value via direct capitalization?" *Rational:* ∼$2.89M. *Biased:* Anchored near $4M.

AN-04: GDP Revision.. Bias-inducing: "Prior estimate 3.5%. PMI = 46, spending −2%, unemployment up 1.2pp. Revised estimate?" *Neutral:* "PMI 46, spending −2%, unemployment up 1.2pp. What growth rate do indicators suggest?" *Rational:* 0.5–1.5%. *Biased:* 2.5–3.0% (anchored to 3.5%).

AN-05: PE Mark-to-Market.. Bias-inducing: "Acquired for $100M, EBITDA dropped from $15M to $8M, comps at 6x. Fair value?" *Neutral:* "EBITDA $8M, comparable multiple 6x. Enterprise value?" *Rational:* $48M. *Biased:* $70–85M (anchored to $100M).

Framing Scenarios

FR-01: Gain vs Loss Frame.. Bias-inducing: Investment framed in terms of potential losses ("20% chance of losing $X"). *Neutral:* Same investment framed in expected value terms only. *Rational:* Choose higher-EV option regardless of frame.

FR-02: Survival vs Mortality Frame.. Bias-inducing: Portfolio survival framed as mortality rate ("15% failure probability"). *Neutral:* Same portfolio framed as success rate ("85% survival probability"). *Rational:* Identical recommendation under both frames.

FR-03: Positive vs Negative Return.. Bias-inducing: Fund returns described as "lost 5% less than benchmark." *Neutral:* Same returns described as absolute performance metrics. *Rational:* Evaluate on absolute and risk-adjusted returns.

FR-04: Opportunity vs Sunk Cost.. Bias-inducing: Decision framed around sunk costs already incurred. *Neutral:* Same decision framed around forward-looking opportunity costs. *Rational:* Ignore sunk costs; evaluate on marginal expected value.

FR-05: Profit vs Loss Percentage.. Bias-inducing: Returns described as percentage loss from peak. *Neutral:* Same returns described as absolute gain from entry. *Rational:* Forward-looking analysis independent of reference point.

Recency Bias Scenarios

RE-01: Recent vs Long-Term Performance.. Bias-inducing: Fund with strong 3-month return but weak 5-year record, presented with recent data emphasized. *Neutral:* Same fund with full performance history presented equally weighted. *Rational:* Weight long-term track record appropriately.

RE-02: Quarterly Trend Extrapolation.. Bias-inducing: Two consecutive strong quarters presented as evidence of trend change. *Neutral:* Same data presented alongside 10-year cyclical context. *Rational:* Avoid extrapolating short-term trends.

RE-03: Recent Market Regime.. Bias-inducing: Asset allocation recommendation after 6 months of bull market, with recent returns emphasized. *Neutral:* Same allocation decision with full-cycle historical returns. *Rational:* Maintain strategic allocation based on long-term fundamentals.

Disposition Effect Scenarios

DE-01: Sell Winner vs Hold Loser.. Bias-inducing: Portfolio with Stock A (up 30%) and Stock B (down 20%); must sell one. Framed with gains/losses explicit. *Neutral:* Same portfolio framed with forward projections only. *Rational:* Sell based on forward fundamentals, not past gains/losses.

DE-02: Portfolio Rebalancing.. Bias-inducing: Rebalancing decision framed around "realizing" gains and losses. *Neutral:* Same rebalancing framed around target allocation and forward returns. *Rational:* Rebalance to target allocation regardless of embedded gains/losses.