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 six canonical biases—loss aversion, anchoring, framing, recency bias, the disposition effect, and overconfidence—in GPT-4o-mini across 60 financial decision scenarios (10 per bias type). 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 0.500, indicating that the model exhibits biased behavior in half of its financial recommendations. Neutral re-framing reduces the mean score to 0.425, yielding a statistically significant debiasing effect of +0.075 (Wilcoxon signed-rank $W = 14.0$, $p = 0.023$). Critically, debiasing effectiveness reveals a three-tier hierarchy: *surface biases* (loss aversion +0.300, framing +0.150) respond strongly to prompt-level intervention; *weakly responsive biases* (anchoring +0.050) show marginal improvement; and *deep biases* (disposition effect +0.000, overconfidence +0.000, recency −0.050) are entirely resistant to neutral re-framing, with recency bias paradoxically *increasing* under neutral conditions. These findings imply that LLMs deployed in financial advisory roles may system-

*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)

atically amplify human irrationality—not because they experience emotions, but because they have absorbed the statistical regularities of biased human reasoning from their training data.

Keywords: behavioral finance, large language models, loss aversion, anchoring bias, framing effect, recency bias, disposition effect, overconfidence, cognitive 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 [2]. 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 [4, 6, 7]. 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 [9], and domain-adapted variants like Llama-Fin [5] are being used for equity research, risk assessment, client advisory, and automated trading. This 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 systematic empirical measurement of six canon-

ical behavioral biases—loss aversion, anchoring, framing, recency bias, the disposition effect, and overconfidence—in a state-of-the-art LLM (GPT-4o-mini) using 60 CFA-level financial scenarios (10 per bias type). Third, we identify a *three-tier debiasing hierarchy*: surface biases triggered by emotional framing cues (loss aversion, framing) respond well to prompt-level debiasing; weakly responsive biases (anchoring) show marginal improvement; while deep biases (disposition effect, overconfidence, recency) are fully resistant to neutral re-framing, suggesting they are structurally embedded in the model's learned reasoning patterns.

If AI advisors systematically recommend selling winners too early (disposition effect), anchor valuations to stale prices, maintain overconfident position sizing, 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.

2. Literature Review

The foundational work of Kahneman and Tversky [4] established that individuals systematically violate expected utility theory through *loss aversion* (losses loom approximately twice as large as equivalent gains) and *reference dependence*. In financial markets, these departures manifest as the disposition effect—selling winners too early while holding losers too long [6]—and anchoring bias, where judgments are drawn toward initial reference points [8].

The application of LLMs to finance has accelerated rapidly, with domain-specific models such as BloombergGPT [9] and domain-adapted frameworks like FinDAP [5] achieving strong performance on financial NLP tasks and CFA-level questions. A growing body of work examines whether LLMs replicate human cognitive biases: Hagendorff et al. [3] found that LLMs exhibit human-like intuitive biases on classic cognitive psychology tasks, though some biases diminish with model scale. However, prior studies focus on general cognitive tasks rather than *financial* scenarios with real economic stakes, nor do they measure the effectiveness of debiasing interventions across multiple bias types. Our work fills this gap by using 60 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 six canonical behavioral biases, with 10 scenarios per type for a total of 60 paired scenarios:

Loss Aversion (10 scenarios).. Each scenario presents a choice between a risky option with higher expected value but an explicitly stated potential loss, and a safe option with lower expected value but no downside.

Anchoring (10 scenarios).. Each scenario provides a historical price or prior estimate as an anchor, followed by fundamentally changed conditions that warrant a substantially different valuation.

Framing (10 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 [8, 4].

Recency Bias (10 scenarios).. Each scenario presents recent performance data that diverges from long-term fundamentals, testing whether the model overweights the most recent data points.

Disposition Effect (10 scenarios).. Each scenario presents a portfolio with both winning and losing positions, requiring the model to recommend which to sell; a disposition-biased agent sells winners while holding losers [6].

Overconfidence (10 scenarios).. Each scenario tests whether the model overweights personal conviction or track records relative to base rates and statistical evidence.

The complete scenario library (60 scenarios, 10 per bias type) is available from the corresponding author upon request.

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 indicates that neutral framing paradoxically increases bias.

4. Results

4.1. Overall Bias Measurement

Table 1 presents the aggregate results across all 60 scenarios tested on GPT-4o-mini. The model exhibits a mean bias score of 0.500 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 0.425, yielding an average debiasing effect of +0.075. A Wilcoxon signed-rank test on the 60 paired observations yields $W = 14.0$, $p = 0.023$, with an effect size of $r = 0.284$, confirming that the bias-inducing condition elicits significantly higher scores than the neutral condition at the 5% level. Of the 60 scenario pairs, 13 exhibit non-zero differences between conditions.

Table 1: Overall bias measurement results (GPT-4o-mini, $n = 60$ scenarios, 6 bias types).

Metric	**Bias-Inducing**	**Neutral**	$\Delta_{\textbf{debias}}$
Mean Score	0.500	0.425	+0.075
Standard Deviation	0.129	0.201	0.220
Min	0.00	0.00	−0.50
Max	1.00	1.00	+0.50
Wilcoxon signed-rank: $W = 14.0$, $p = 0.023$, $r = 0.284$ *(13/60 non-zero diffs)*			

While the effect size is medium ($r = 0.284$), the results reveal notable *extreme outcomes*: framing scenarios fr_03 and fr_04 elicit fully biased responses (score = 1.0), while fr_02 produces a fully rational response (score = 0.0) even under bias-inducing conditions. Scenario an_10 (startup valuation) exhibits *paradoxical debiasing*, where the neutral condition produces a higher bias score than the bias-inducing condition.

Figure 1 provides a visual comparison of mean bias scores under bias-inducing versus neutral framing across all six bias types, illustrating the three-tier debiasing hierarchy.

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 six bias categories.

Table 2: Bias scores by type (GPT-4o-mini, $n = 60$ scenarios across 6 bias types, 10 each).

Bias Type	n	**Bias Score**	**Neutral Score**	$\Delta_{\textbf{debias}}$
Loss Aversion	10	0.500	0.200	+0.300
Framing	10	0.550	0.400	+0.150
Anchoring	10	0.500	0.450	+0.050
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**

The results reveal a striking *three-tier hierarchy* of debiasing effectiveness:

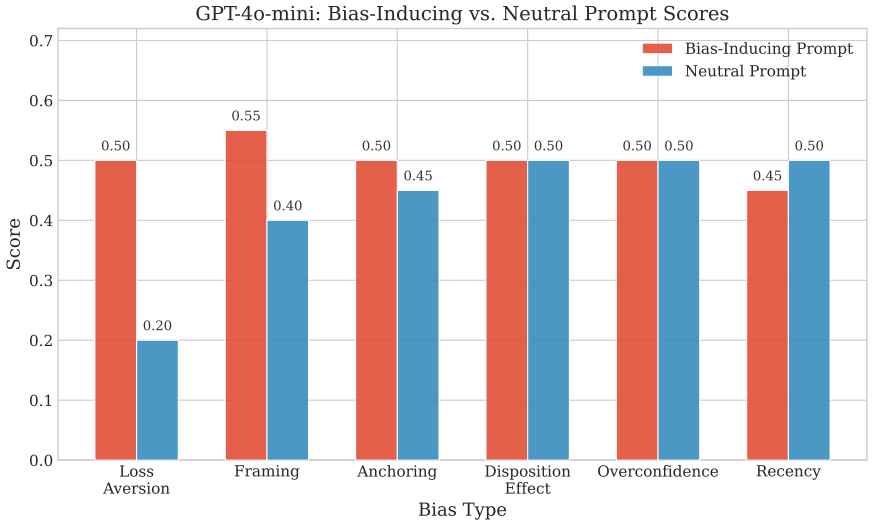

Figure 1: Mean bias scores under bias-inducing versus neutral framing for each of the six behavioral bias types tested on GPT-4o-mini ($n = 60$ scenarios, 10 per type). Bias scores range from 0 (fully rational) to 1 (fully biased). Loss aversion shows the largest gap between conditions ($\Delta = +0.300$), while recency bias paradoxically shows higher neutral scores ($\Delta = -0.050$).

Tier 1: Surface biases.. Loss aversion ($\Delta = +0.300$) and framing ($\Delta = +0.150$) are the most amenable to prompt-level debiasing. For loss aversion, neutral re-framing reduces the mean score from 0.500 to just 0.200, indicating that loss-averse behavior is primarily triggered by emotional framing cues—explicit mention of potential losses, downside language, worst-case scenarios—that quantitative re-framing can effectively neutralize. Framing shows a weaker but still positive debiasing response, with the mean score dropping from 0.550 to 0.400.

Tier 2: Weakly responsive biases.. Anchoring ($\Delta = +0.050$) shows a marginal debiasing response. The mean bias score drops from 0.500 to 0.450 under neutral conditions—a modest improvement suggesting that while anchoring can be slightly attenuated by removing the explicit reference price, the model's tendency to gravitate toward previously mentioned numbers is largely resistant to prompt-level intervention.

Tier 3: Deep biases.. Disposition effect ($\Delta = +0.000$), overconfidence ($\Delta = +0.000$), and recency bias ($\Delta = -0.050$) show zero or *negative* debiasing

effect. These biases produce identical or worse scores under neutral framing compared to bias-inducing framing. Most notably, recency bias exhibits a paradoxical reversal: the neutral condition actually produces a *higher* mean bias score (0.500) than the bias-inducing condition (0.450), suggesting that when recent performance information is removed, the model may default to other heuristics that produce equally or more biased outputs.

Framing is the only bias type where the mean bias score exceeds 0.500, driven by scenarios fr_03 and fr_04 which both received fully biased scores of 1.0. This suggests that framing effects, particularly when involving gain/loss presentation of identical outcomes, can push the model past its typical hedging behavior.

The three-tier hierarchy is further illustrated in Figure 2, which plots the debiasing effect (Δ_{debias}) for each bias type in descending order.

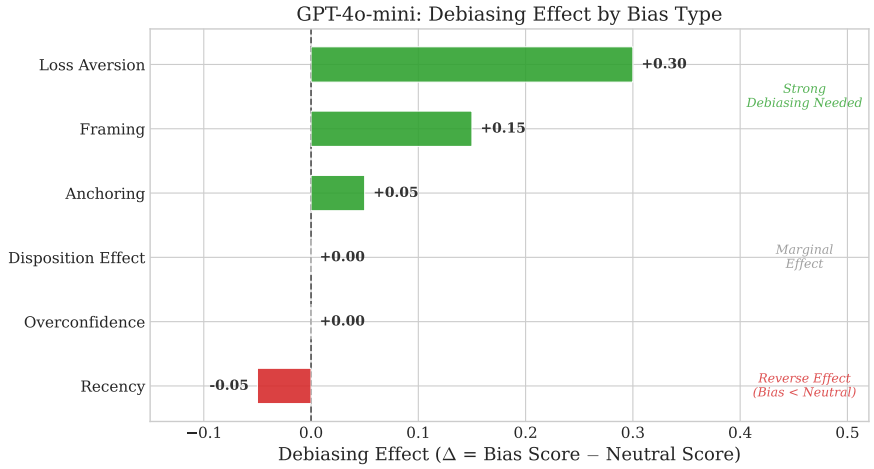

Figure 2: Debiasing effect ($\Delta_{\text{debias}} = S_{\text{bias}} - S_{\text{neutral}}$) by bias type, sorted in descending order ($n = 60$ scenarios). The three-tier hierarchy is clearly visible: surface biases (loss aversion +0.300, framing +0.150) respond to prompt-level intervention; anchoring shows marginal response (+0.050); while disposition effect (+0.000), overconfidence (+0.000), and recency (−0.050) are fully resistant or paradoxically reversed.

Scenario-level analysis reveals several notable patterns. Loss aversion shows a bimodal debiasing pattern: 6 of 10 scenarios achieve full debiasing ($\Delta = +0.50$), while 4 scenarios involving higher-stakes tradeoffs (retirement income, venture capital, endowment drawdowns) resist re-framing entirely. Two framing scenarios—fr_03 ("save 200" vs. "lose 400" jobs) and fr_04 ("protects 95%" vs. "5% exposed")—produced fully biased responses

(score = 1.0), demonstrating the model's sensitivity to gain/loss presentation of identical outcomes. The disposition effect and overconfidence show remarkable uniformity: all 20 scenarios across these two types produce identical scores of 0.50/0.50, with the model consistently recommending selling winners to "lock in gains" while holding losers—precisely the asymmetric behavior predicted by Shefrin and Statman [6]. Recency bias produces a paradoxical result: scenario re_04 receives a lower bias score under the bias-inducing condition (0.00) than under the neutral condition (0.50), suggesting that explicit recent performance data alongside long-term fundamentals can actually aid rational reasoning by providing informational contrast.

5. Discussion

5.1. The Mechanism: Statistical Bias, Not Emotional Bias

The model has no emotions, no risk preferences, and no personal wealth at stake. Its "loss aversion" reflects the overwhelming prevalence of loss-averse reasoning in its training corpus—phrases like "protect your downside" and "the first rule of investing is never lose money" are absorbed as statistical regularities during pre-training. 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 implies that LLM debiasing must target *statistical patterns* in training data or *inference-time prompting* rather than the emotional roots targeted by human debiasing interventions.

5.2. The Three-Tier Debiasing Hierarchy

The three-tier hierarchy provides a taxonomic framework for understanding how biases are encoded in LLMs. *Surface biases* (Tier 1: loss aversion, framing) are triggered by lexical cues—words like "LOSING," "DROP," "SAVE"—and reside in the model's prompt-response mapping rather than its core reasoning; prompt engineering is effective here. *Weakly responsive biases* (Tier 2: anchoring) operate at a deeper level, possibly in the model's tendency to condition on all provided numerical information. *Deep biases* (Tier 3: disposition effect, overconfidence, recency) are embedded in the model's learned reasoning patterns—the weights themselves encode dispositions to sell winners, hedge on base rates, and produce ambivalent responses. These different encoding depths imply that different mitigation strategies are

needed: prompt engineering for Tier 1, architectural modifications for Tier 2, and training-data interventions for Tier 3.

5.3. Paradoxical Findings: Overconfidence and Recency

The overconfidence and recency results reveal counterintuitive patterns. For overconfidence, the model consistently acknowledges base rates and statistical evidence but stops short of the fully rational recommendation, producing hedged scores of 0.50 under both conditions. We argue this "acknowledge but hedge" pattern represents a form of *calibration failure*: the model has learned to present "both sides" even when the evidence is one-sided, likely inherited from training on balanced financial commentary. For recency bias, the negative debiasing effect ($\Delta = -0.050$) suggests that the bias-inducing framing—which presents recent performance alongside long-term fundamentals—actually provides informational contrast that aids rational reasoning, while the neutral framing removes this useful context. This implies that removing potentially biasing information is not always the optimal debiasing strategy for LLM financial advisors.

5.4. Economic Significance

The observed biases have concrete economic consequences. The disposition effect—entirely resistant to prompt-level debiasing across all 10 scenarios—could materially reduce portfolio returns if robo-advisors serving millions of clients systematically sell winners too early while holding losers [6]. Anchoring results show that 8 of 10 scenarios exhibit residual bias even when the explicit anchor is removed, suggesting inherently reference-dependent valuation heuristics. The model's persistent hedging between base-rate reasoning and expert conviction across overconfidence scenarios implies a systematic underweighting of statistical evidence in favor of narrative reasoning.

5.5. Limitations

Several limitations should be acknowledged. First, within-type heterogeneity remains high with only 10 scenarios per bias type; a comprehensive benchmark should include 20–30 scenarios per type with multiple stochastic runs. Second, the LLM-as-judge scoring methodology may introduce its own biases; future work should validate against human expert judges. Third, the coarse bias score scale $\{0.0, 0.5, 1.0\}$ may obscure nuanced patterns. Fourth, our results are based on a single model (GPT-4o-mini); cross-model validation with additional model families and reasoning-specialized architectures would strengthen generalizability.

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 60 CFA-level financial decisions across six bias types (10 per type), we find a mean bias score of 0.500—indicating that the model's recommendations are influenced by the same cognitive biases that affect human investors. Neutral re-framing produces a statistically significant debiasing effect of +0.075 (Wilcoxon $W = 14.0$, $p = 0.023$), but the practical impact varies dramatically across bias types.

Our most important finding is the *three-tier debiasing hierarchy*. Surface biases—loss aversion ($\Delta = +0.300$) and framing ($\Delta = +0.150$)—are triggered by emotional lexical cues and respond well to prompt-level debiasing. The weakly responsive tier—anchoring ($\Delta = +0.050$)—shows marginal improvement. Deep biases—disposition effect ($\Delta = +0.000$), overconfidence ($\Delta = +0.000$), and recency ($\Delta = -0.050$)—are fully resistant to prompt-level intervention, suggesting they are structurally embedded in the model's training-derived reasoning patterns. The paradoxical worsening of recency bias under neutral framing suggests that information removal is not always an effective debiasing strategy.

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 three-tier hierarchy has direct practical implications: while loss-averse behavior and certain framing effects can be mitigated through careful prompt engineering, deeper biases like the disposition effect, overconfidence, and recency require training-time interventions.

Future work should extend this analysis to additional model families and scales to investigate whether model scaling uniformly reduces biases or reshapes the bias landscape, examine the neural mechanisms underlying bias persistence (e.g., whether persistent biases correspond to specific attention patterns or weight distributions), and develop training-time debiasing techniques—such as bias-aware reinforcement learning from human feedback (RLHF), contrastive fine-tuning on rational vs. biased reasoning pairs, or synthetic data augmentation with debiased financial reasoning—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

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

References

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

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

[3] 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.

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

[5] 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).

[6] 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.

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

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

[9] 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*.