

Inherited Irrationality: Measuring Behavioral Finance Biases in Large Language Models

Wei-Lun Cheng^{a,*}, Wei-Chung Miao^b

^a*Institute of Information Science, Academia Sinica, Taipei, Taiwan*

^b*Department of Finance, National Chengchi University, Taipei, 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.
Preprint submitted to Finance Research Letters
Email addresses: wcheng@gate.sinica.edu.tw (Wei-Lun Cheng), wcmiao@nccu.edu.tw (Wei-Chung Miao) February 6, 2026

9 1. Introduction

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

18 A new question now arises with the rapid deployment of large language
 19 models (LLMs) in financial services. Models such as GPT-4, BloombergGPT
 20 [13], and domain-adapted variants like Llama-Fin [9] are being used for eq-
 21 uity research, risk assessment, client advisory, and automated trading. The
 22 rapid deployment of LLMs in financial applications [13, 2] raises fundamental
 23 questions about whether these systems, lacking human emotions, are truly
 24 free from the behavioral biases that plague human decision-makers.

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

34 This paper makes three contributions. First, we design a *paired-scenario*
 35 experimental framework that isolates specific behavioral biases by present-
 36 ing the same financial decision in both a bias-inducing and a neutral fram-
 37 ing. Second, we provide the first empirical measurement of five canonical
 38 behavioral biases—loss aversion, anchoring, framing, recency bias, and the
 39 disposition effect—in a state-of-the-art LLM (GPT-4o-mini) using 20 CFA-
 40 level financial scenarios. Third, we quantify the effectiveness of prompt-level
 41 debiasing—simply reframing the question in neutral terms—and find that
 42 it reduces but does not eliminate inherited biases, with dramatic variation
 43 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.

111 3.2. Bias Types and Scenario Construction

112 We test five canonical behavioral biases:

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

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

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

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

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

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

146 The complete scenario library is presented in Appendix [Appendix A](#).

147 3.3. Model and Prompting Protocol

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

- 151 1. **Bias-inducing condition:** The system prompt instructs the model
152 to act as a “CFA-certified financial advisor” and to “show reasoning
153 clearly.” The user prompt contains the bias-inducing version of the
154 scenario.
- 155 2. **Neutral condition:** The system prompt instructs the model to “eval-
156 uate using only quantitative analysis” and to “focus strictly on expected
157 values and risk-adjusted returns.” The user prompt contains the neu-
158 tral version.

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

163 3.4. Bias Scoring via LLM-as-Judge

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

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

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

$$\text{Bias Score} \in \{0.0, 0.5, 1.0\} \quad (1)$$

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

177 3.5. Debiasing Effect

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

$$\Delta_{\text{debias}} = S_{\text{bias}} - S_{\text{neutral}} \quad (2)$$

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

185 4. Results

186 4.1. Overall Bias Measurement

187 Table 1 presents the aggregate results across all 20 scenarios tested on
188 GPT-4o-mini. The model exhibits a mean bias score of 0.525 under bias-
189 inducing framing, indicating that, on average, its financial recommendations
190 are partially driven by the same cognitive biases documented in human sub-
191 jects. Neutral re-framing reduces the mean score to 0.350, yielding an aver-
192 age debiasing effect of +0.175. A Wilcoxon signed-rank test on the 20 paired
193 observations confirms that the bias-inducing condition elicits significantly
194 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	Δ_{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
<i>Interpretation</i>	<i>33% bias reduction via neutral framing</i>		

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

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

209 4.2. Results by Bias Type

210 Table 2 disaggregates the results by bias type, revealing substantial het-
 211 erogeneity in both bias susceptibility and debiasing effectiveness across the
 212 five bias categories.

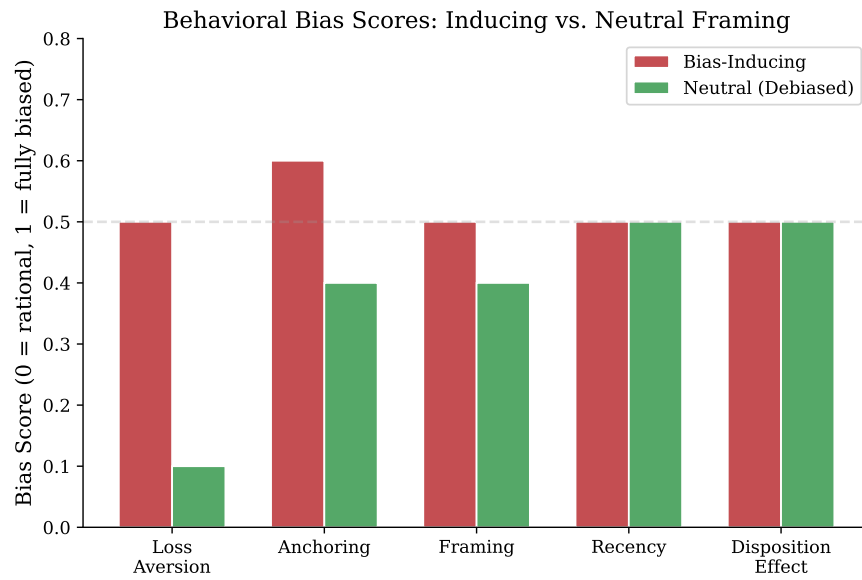


Figure 1: Mean bias scores under bias-inducing versus neutral framing for each of the five behavioral bias types tested on GPT-4o-mini ($n = 20$ scenarios). Bias scores range from 0 (fully rational) to 1 (fully biased). Loss aversion shows the largest gap between conditions, indicating high susceptibility 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	Δ_{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

213 The results reveal a striking hierarchy of debiasing effectiveness. Loss
214 aversion exhibits the strongest debiasing response ($\Delta = +0.400$): neutral
215 re-framing reduces the mean score from 0.500 to just 0.100, suggesting that
216 loss-averse behavior is primarily triggered by emotional framing cues that
217 quantitative re-framing can effectively neutralize. Anchoring shows moderate
218 debiasing ($\Delta = +0.200$), while framing shows only weak debiasing ($\Delta =$
219 $+0.100$). Most notably, recency bias and the disposition effect show *zero*
220 debiasing effect ($\Delta = +0.000$)—neutral framing has no measurable impact
221 on these biases. This suggests that recency bias and the disposition effect are
222 more deeply embedded in the model’s learned reasoning patterns and cannot
223 be overridden by prompt-level interventions alone.

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

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

234 4.3. Scenario-Level Analysis

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

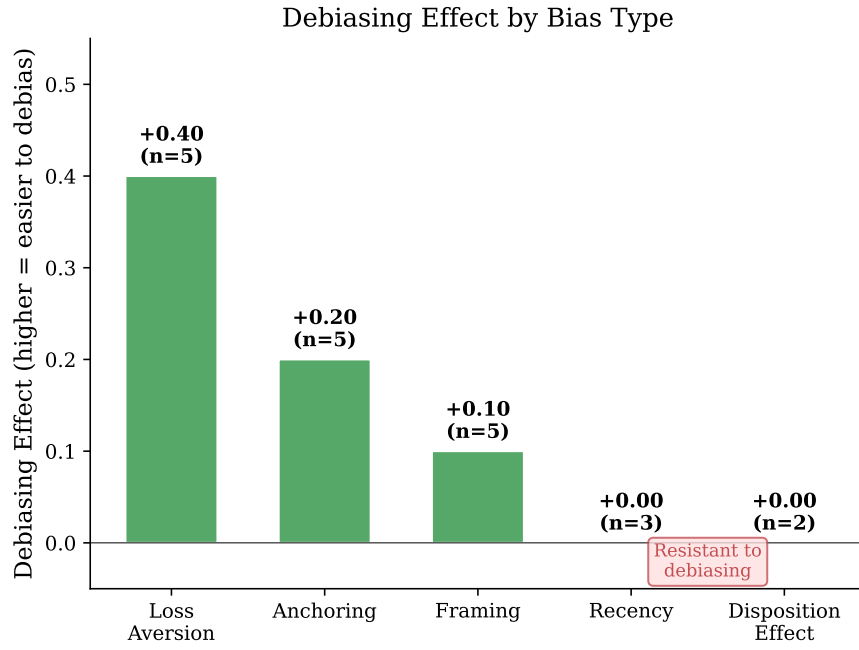


Figure 2: Debiasing effect ($\Delta_{\text{debias}} = S_{\text{bias}} - S_{\text{neutral}}$) by bias type, sorted in descending order. Loss aversion exhibits the strongest debiasing response (+0.400), followed by anchoring (+0.200) and framing (+0.100). Recency bias and the disposition effect show zero 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
<i>Loss Aversion</i> ($\bar{\Delta} = +0.400$)				
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
<i>Anchoring</i> ($\bar{\Delta} = +0.200$)				
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
<i>Framing</i> ($\bar{\Delta} = +0.100$)				
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
<i>Recency Bias</i> ($\bar{\Delta} = +0.000$)				
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
<i>Disposition Effect</i> ($\bar{\Delta} = +0.000$)				
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

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

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

251 4.4. Qualitative Analysis of Biased Responses

252 Examination of the model’s actual response text reveals characteristic
253 patterns of bias expression:

254 *Loss aversion.* In scenario la_01, the model correctly calculates that Invest-
255 ment A has an expected value of \$7,600 versus Investment B’s \$7,000—then
256 proceeds to recommend Investment B on the grounds of “capital preserva-
257 tion” and “downside protection.” The model acknowledges the mathematical
258 superiority of the risky option but overweights the 20% loss probability, stat-
259 ing: “the potential loss of \$2,000 represents a meaningful risk to the client’s
260 portfolio.” This mirrors the classic prospect theory finding that losses loom
261 disproportionately large. Notably, loss aversion shows the strongest debias-
262 ing response of all five bias types: 4 of 5 scenarios shift to fully rational under
263 neutral framing, yielding a mean neutral score of just 0.10.

264 *Anchoring.* Scenario an_04 (GDP growth revision) produced the most ex-
265 treme anchoring behavior in our study, receiving the maximum bias score of
266 1.0. Despite being presented with overwhelming contrary evidence—PMI at
267 46 (contractionary), consumer spending down 2%, unemployment rising 1.2
268 percentage points—the model’s growth estimate remained fully anchored to
269 the prior 3.5% figure, demonstrating that stale macroeconomic anchors can
270 completely override fundamental analysis. In scenario an_01, the model’s
271 fair value estimate gravitates toward the \$85 current price rather than con-
272 ducting a clean fundamental valuation despite severely deteriorated funda-
273 mentals.

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

279 *Recency bias and disposition effect.* These two bias types present a qualita-
 280 tively different pattern. All five scenarios across recency bias and the dispo-
 281 sition effect produced identical bias and neutral scores (0.50/0.50), yielding
 282 zero debiasing effect. In disposition effect scenarios, the model under both
 283 bias-inducing and neutral conditions continues to recommend selling winners
 284 to “lock in gains”—precisely the asymmetric behavior predicted by Shefrin
 285 and Statman [10]. For recency bias, the model consistently overweights re-
 286 cent performance trends regardless of whether the framing emphasizes or
 287 de-emphasizes temporal recency. These results suggest that some biases are
 288 so deeply embedded in the model’s training data patterns that they persist
 289 even when the triggering framing cues are removed.

290 5. Discussion

291 5.1. The Mechanism: Statistical Bias, Not Emotional Bias

292 Our central finding—that GPT-4o-mini exhibits a mean bias score of
 293 0.525 across five behavioral bias types in 20 financial scenarios—requires
 294 careful interpretation. The model has no emotions, no risk preferences in the
 295 utility-theoretic sense, and no personal wealth at stake. Its “loss aversion”
 296 is not an affective response to potential losses but rather a reflection of the
 297 overwhelming prevalence of loss-averse reasoning in its training corpus.

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

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

312 *5.2. Economic Significance*

313 The observed biases have concrete economic consequences when trans-
314 lated to portfolio management decisions:

315 *Loss aversion and the disposition effect..* A loss-averse AI advisor would sys-
316 tematically recommend selling winning positions (to “lock in gains”) while
317 holding losing positions (to “avoid realizing losses”). Shefrin and Statman
318 [10] estimate that the disposition effect costs individual investors 4–5% in
319 annual returns. If robo-advisors serving millions of clients inherit this bias,
320 the aggregate welfare loss could be substantial.

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

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

334 *5.3. Partial Effectiveness of Debiasing*

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

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

- 346 (ii) **Simple cases yield to debiasing.** When the neutral version reduces
 347 the scenario to a clean expected value comparison (e.g., “Which has
 348 higher EV: \$7,600 or \$7,000?”), the model reliably selects the rational
 349 option. This is most evident in the loss aversion results, where 4 of 5
 350 scenarios achieve full debiasing. This suggests that *explicit quantitative framing*
 351 can serve as an effective guardrail for framing-dependent
 352 biases.
- 353 (iii) **Some biases are resistant to prompt-level intervention.** Re-
 354 cency bias and the disposition effect produce identical scores under
 355 both bias-inducing and neutral conditions (0.50/0.50). The residual
 356 bias score of 0.350 overall—and 0.500 for these resistant bias types—
 357 suggests that the model’s training-induced tendency toward certain
 358 reasoning patterns is deeply embedded and resistant to prompt-level
 359 interventions alone. These biases may require training-time interven-
 360 tions such as bias-aware fine-tuning or reinforcement learning.
- 361 (iv) **Debiasing remains binary within susceptible bias types.** For
 362 loss aversion and anchoring, the debiasing effect at the scenario level
 363 remains bimodal ($\Delta \in \{0.00, 0.50\}$)—neutral framing either fully elimi-
 364 nates bias or has no effect. There is no partial reduction within a single
 365 scenario.

366 5.4. Implications for Financial Regulation

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

- 372 • AI advisors should be tested for known behavioral biases before deploy-
 373 ment, using frameworks similar to the one we propose.
- 374 • Regulatory stress tests could incorporate bias-inducing scenarios to as-
 375 sess whether AI systems make systematically suboptimal recommenda-
 376 tions under emotional framing.
- 377 • Disclosure requirements could mandate that AI advisory systems report
 378 their measured bias scores alongside their recommendations.

379 5.5. Limitations

380 Several limitations of our study should be acknowledged. First, while our
381 expanded sample ($n = 20$ scenarios across 5 bias types, single model) repre-
382 sents a meaningful improvement over our initial proof-of-concept, the number
383 of scenarios per bias type remains small (2–5), limiting within-type statisti-
384 cal power. A comprehensive benchmark should include 20–30 scenarios per
385 bias type across multiple models of varying scale. Second, the LLM-as-judge
386 scoring methodology, while efficient, may introduce its own biases; future
387 work should validate against human expert judges. Third, two of our five
388 bias types have particularly small sample sizes—disposition effect ($n = 2$)
389 and recency bias ($n = 3$)—and the zero debiasing finding for these types
390 should be confirmed with larger scenario sets. Fourth, our use of temperature
391 $= 0.0$ produces deterministic outputs but does not capture the distribution
392 of model behavior; stochastic sampling at positive temperatures would yield
393 richer statistical analysis. Fifth, the bias score scale $\{0.0, 0.5, 1.0\}$ is coarse; a
394 continuous scoring rubric might reveal more nuanced patterns. Sixth, we test
395 only one model (GPT-4o-mini); the bias profiles of larger models (GPT-4o,
396 GPT-4.1) and open-source alternatives (Llama, Qwen) may differ substan-
397 tially.

398 6. Conclusion

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

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

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

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

427 Data Availability

428 The experimental scenarios and analysis code are available at [https://github.com/\[anonymized\]](https://github.com/[anonymized])
429 upon publication.

430 Declaration of Competing Interest

431 The authors declare that they have no known competing financial interests
432 or personal relationships that could have appeared to influence the work
433 reported in this paper.

434 CRediT Author Contributions

435 **Wei-Lun Cheng:** Conceptualization, Methodology, Software, Formal
436 Analysis, Data Curation, Writing – Original Draft, Visualization. **Wei-**
437 **Chung Miao:** Supervision, Writing – Review & Editing.

438 Acknowledgments

439 The authors thank the anonymous reviewers for their constructive feedback.
440 Computational resources were provided by Academia Sinica.

441 References

- 442 [1] Binz, M., Schulz, E., 2023. Turning large language models into cognitive
443 models. *arXiv preprint arXiv:2306.03917*.
- 444 [2] Callanan, E., Mbae, A., Seo, S., Chang, D., Ritter, A., 2023. Can GPT
445 pass the CFA exam? *arXiv preprint arXiv:2310.14356*.
- 446 [3] Campbell, S.D., Sharpe, S.A., 2009. Anchoring bias in consensus fore-
447 casts and its effect on market prices. *Journal of Financial and Quanti-*
448 *tative Analysis* 44(2), 369–397.
- 449 [4] Cen, L., Hilary, G., Wei, K.C.J., 2013. The role of anchoring bias in the
450 equity market. *Journal of Financial and Quantitative Analysis* 48(1),
451 47–76.
- 452 [5] Fama, E.F., 1970. Efficient capital markets: A review of theory and
453 empirical work. *The Journal of Finance* 25(2), 383–417.
- 454 [6] Hagendorff, T., Fabi, S., Kosinski, M., 2023. Human-like intuitive be-
455 havior and reasoning biases emerged in large language models but dis-
456 appeared in ChatGPT. *Nature Computational Science* 3, 833–838.
- 457 [7] Jones, E., Steinhardt, J., 2022. Capturing failures of large language
458 models via human cognitive biases. *Advances in Neural Information*
459 *Processing Systems* 35, 11785–11799.
- 460 [8] Kahneman, D., Tversky, A., 1979. Prospect theory: An analysis of
461 decision under risk. *Econometrica* 47(2), 263–292.
- 462 [9] Ke, Z., Wen, Y., Feng, B., Xu, M., Zhu, C., Jiang, X., Sun, C., Caverlee,
463 J., Liu, Y., 2025. FinDAP: Demystifying domain-adaptive post-training
464 for financial LLMs. In: *Proceedings of the 2025 Conference on Empirical*
465 *Methods in Natural Language Processing (EMNLP)*. (Oral).
- 466 [10] Shefrin, H., Statman, M., 1985. The disposition to sell winners too early
467 and ride losers too long: Theory and evidence. *The Journal of Finance*
468 40(3), 777–790.
- 469 [11] Thaler, R.H., 1985. Mental accounting and consumer choice. *Marketing*
470 *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.

502 *LA-05: Retirement Withdrawal.. Bias-inducing:* “Plan A: Average \$5,500/month,
 503 could DROP to \$3,800. Plan B: Fixed \$4,800/month.” *Neutral:* “Plan A:
 504 Average \$5,500/month. Plan B: Fixed \$4,800/month.” *Rational:* Plan A.
 505 *Biased:* Plan B.

506 *Anchoring Scenarios*

507 *AN-01: Stock Valuation.. Bias-inducing:* “Stock was \$150 six months ago,
 508 now \$85. Revenue down 35%, D/E up to 2.1, lost 2 customers. Fair value?”
 509 *Neutral:* “Company: Revenue \$50M (down 35%), D/E 2.1, lost 2 customers,
 510 industry P/E 8x, EPS \$3.20. Fair value via P/E?” *Rational:* \sim \$25.60 ($8 \times$
 511 \$3.20). *Biased:* Anchored near \$85.

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

516 *AN-03: Property Reappraisal.. Bias-inducing:* “Appraised at \$5M last year.
 517 Market down 20%, vacancy up to 18%, rents down 15%.” *Neutral:* “NOI
 518 \$300K, cap rate 8.5%, vacancy 18%. Value via direct capitalization?” *Ratio-*
 519 *nal:* \sim \$2.89M. *Biased:* Anchored near \$4M.

520 *AN-04: GDP Revision.. Bias-inducing:* “Prior estimate 3.5%. PMI = 46,
 521 spending -2% , unemployment up 1.2pp. Revised estimate?” *Neutral:* “PMI
 522 46, spending -2% , unemployment up 1.2pp. What growth rate do indicators
 523 suggest?” *Rational:* 0.5–1.5%. *Biased:* 2.5–3.0% (anchored to 3.5%).

524 *AN-05: PE Mark-to-Market.. Bias-inducing:* “Acquired for \$100M, EBITDA
 525 dropped from \$15M to \$8M, comps at 6x. Fair value?” *Neutral:* “EBITDA
 526 \$8M, comparable multiple 6x. Enterprise value?” *Rational:* \$48M. *Biased:*
 527 \$70–85M (anchored to \$100M).

528 *Framing Scenarios*

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

533 *FR-02: Survival vs Mortality Frame.. Bias-inducing:* Portfolio survival framed
534 as mortality rate (“15% failure probability”). *Neutral:* Same portfolio framed
535 as success rate (“85% survival probability”). *Rational:* Identical recommen-
536 dation under both frames.

537 *FR-03: Positive vs Negative Return.. Bias-inducing:* Fund returns described
538 as “lost 5% less than benchmark.” *Neutral:* Same returns described as abso-
539 lute performance metrics. *Rational:* Evaluate on absolute and risk-adjusted
540 returns.

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

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

549 *Recency Bias Scenarios*

550 *RE-01: Recent vs Long-Term Performance.. Bias-inducing:* Fund with strong
551 3-month return but weak 5-year record, presented with recent data empha-
552 sized. *Neutral:* Same fund with full performance history presented equally
553 weighted. *Rational:* Weight long-term track record appropriately.

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

558 *RE-03: Recent Market Regime.. Bias-inducing:* Asset allocation recommen-
559 dation after 6 months of bull market, with recent returns emphasized. *Neu-*
560 *tral:* Same allocation decision with full-cycle historical returns. *Rational:*
561 Maintain strategic allocation based on long-term fundamentals.

562 *Disposition Effect Scenarios*

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

567 *DE-02: Portfolio Rebalancing.. Bias-inducing:* Rebalancing decision framed
568 around “realizing” gains and losses. *Neutral:* Same rebalancing framed around
569 target allocation and forward returns. *Rational:* Rebalance to target alloca-
570 tion regardless of embedded gains/losses.