Artificially Biased Intelligence:

The Limits of Rationality in LLM Financial Advice

Abstract

Financial advice from large language models changes systematically with how an economic problem is described, even when the underlying economics remain the same. In paired portfolio experiments across 54 models that hold expected returns, risks, constraints, and opportunities fixed, describing identical outcomes as value *preserved* rather than value *lost* raises the recommended weight on the safer option by 22.2 percentage points. Random, explicitly uninformative anchors produce an 8.7-point shift. The resulting reallocations change portfolio factor exposures and impose economically meaningful certainty-equivalent costs. The results caution against treating LLMs as rational agents and show that the representation of financial information can itself introduce systematic bias into LLM advice.

JEL Classification: D91, G11, G18, G23, G41.
Keywords: artificial intelligence, delegated portfolio choice, financial advice, behavioral finance, robo-advising, portfolio recommendations.

1 Introduction

Large language models (LLMs) are becoming inputs into financial decision making. Financial institutions use model-based copilots for research retrieval, client synthesis, and adviser communications (OpenAI, 2024), while robo-advice platforms that already manage retail portfolios (D'Acunto, Prabhala, and Rossi, 2019; Rossi and Utkus, 2024) are adding conversational interfaces. In a January 2026 Schwab survey, 63 percent of registered investment advisers reported using AI tools, more than twice the 2023 share (Schwab Advisor Services, 2026). By early 2026, roughly three in five U.S. retail investors reported doing the same.[1] Academic finance uses LLMs to forecast returns from text (Lopez-Lira and Tang, 2026), analyze financial statements (Kim, Muhn, and Nikolaev, 2024), extract information from disclosures (de Kok, 2025), and increasingly substitute for human subjects in economic experiments (Horton, Filippas, and Manning, 2023; Korinek, 2023).

These applications implicitly assign LLMs a role traditionally reserved for economic agents. Beyond plausible or internally consistent answers, decisions should respond to economically relevant information, not irrelevant variation in its presentation. This matters because all LLM inputs are mediated through language. A client's risk tolerance, horizon, constraints, and financial circumstances reach the model through questionnaires, adviser notes, retrieved documents, or user-entered text. Economically identical states can therefore be described differently; if recommendations change, the variation reflects representation rather than economics.

We study whether LLM financial advice satisfies this requirement. Our benchmark is *description invariance*: logically equivalent descriptions of the same decision problem should generate the same choice (Tversky and Kahneman, 1981, 1986). This implication of rational choice differs from the revealed-preference consistency emphasized in prior work. Chen, Liu,

[1] The adviser figure comes from an online survey of 533 Schwab-custodied advisers conducted October 7 to 26, 2025; most use remains administrative rather than portfolio construction. A March 2026 Investing.com survey of 938 U.S. retail investors finds that 62 percent have used AI to inform investment decisions, and Betterment's 2026 Retail Investor Survey reports that nearly half of Gen Z investors say AI has influenced a financial decision.

Shan, and Zhong (2023) show that GPT's budgetary choices satisfy standard consistency conditions at rates exceeding those of humans, and Mei, Xie, Yuan, and Jackson (2024) find that LLM behavior in strategic games can resemble human behavior. Such evidence establishes choice coherence, not invariance to alternative descriptions of an unchanged problem. Evidence that chatbot behavior varies across strategic framings (Mei, Xie, Yuan, and Jackson, 2024) and theoretically irrelevant features (Ross, Kim, and Lo, 2024) suggests the two notions need not coincide.

The distinction is economically meaningful. Consider a model allocating for a moderately risk-tolerant client with a ten-year horizon across five options with identical stated expected returns and variances. In one prompt, outcomes are framed as value retained: the defensive option is almost certain to *preserve* approximately 85 percent of its value in a bad year, while the higher-risk option has even odds of preserving all of it or approximately 70 percent. In the other, the same outcomes are framed as value surrendered: the defensive option is almost certain to *lose* approximately 15 percent, while the higher-risk option has even odds of losing nothing or approximately 30 percent. Both options are re-expressed, so neither is uniquely favored; the prompts are logically equivalent and leave every stated moment unchanged. Yet the average model assigns 22.2 percentage points more to the defensive option under the preservation frame, exceeding its 20 percent equal-weight allocation. On a $1 million portfolio, one sentence shifts more than $220,000, and every model moves in the same direction.

This distortion in delegated financial advice requires none of the incentives, conflicts, or adviser beliefs emphasized in the traditional literature. It arises before implementation through textual representation. In human choice, framing affects the investor's own decision (Kahneman and Tversky, 1979; Tversky and Kahneman, 1981; Barberis, 2018). With LLMs, it instead affects the technology mapping investor characteristics into advice. At scale, an idiosyncratic adviser response may diversify, while a systematic response embedded in a widely deployed model can recur across clients.

We identify description dependence with paired portfolio-allocation experiments that hold the economic problem fixed while varying one descriptive feature. Each model allocates across five options with identical stated expected returns and variances under a fixed horizon and mandate. Within each model-scenario pair, prompts differ only in one cue sentence; option order is held fixed, and temperature is zero. The paired difference therefore isolates the cue's effect.

We study nine cues differing in whether they can rationally affect recommendations. Six are *information-equivalent*, changing presentation but not economic content. Three, involving peer allocations, recent returns, and a vivid market episode, can convey information and benchmark sensitivity to relevant versus irrelevant inputs. Anchoring and downside wording were pre-registered as confirmatory tests. The main sample contains 15,715 allocation decisions from 54 models across 17 developer families, including open- and closed-weight models. We add an independent confirmation wave, informative-twin experiments varying the information in otherwise similar cues, validation using named U.S. equity ETFs and estimated return moments, and debiasing interventions.

Two information-equivalent cues produce large and robust effects. First, in suggested-allocation anchoring, a random placeholder suggests an allocation for one option and is explicitly described as random and uninformative. Raising the anchor increases the targeted allocation by 8.7 percentage points, with 93 percent of models moving toward it. Telling the model to disregard the cue reduces the effect by 44 percent but does not eliminate it. The response is specific to an allocation-like number: an administrative reference number attached to the same option has essentially no effect, consistent with human anchoring evidence (Northcraft and Neale, 1987). Second, reflection framing replicates independently at 20.4 percentage points, with every model moving as predicted. It survives neutral labels and an explicit statement of decision equivalence, and still shifts allocations by 13.5 percentage points with the complete mean-variance problem. The other pre-registered treatment, downside wording, produces a bounded null, as do defaults and denomination. Description

dependence is therefore concentrated in particular representations rather than reflecting indiscriminate prompt instability.

A central identification problem is separating irrational sensitivity from rational inference about cue information. We use informative twins, pairing each information-equivalent cue with a similar version carrying a legitimate signal. Describing the same suggested allocation as the output of a portfolio optimization for the client's mandate rather than a random placeholder shifts the target weight by 54.8 percentage points, about six times the 8.8-point response to the random anchor in the same 53-model sample; informative peer and recent-return signals show the same asymmetry. LLMs therefore distinguish relevant from irrelevant information, responding much more to economic content; description dependence is the residual once that content is removed. For anchoring, the residual is 16.1 percent as large as the optimizer-attributed effect. Thus, models can use information sensibly while still responding to irrelevant presentation.

These reallocations change the portfolio's economic character. Because options have fixed factor exposures, cue-induced shifts map directly into standard styles: reflection framing raises low-volatility exposure by about ten percentage points and quality exposure with it, while recent-return cues tilt portfolios toward momentum (Jegadeesh and Titman, 1993). Since information-equivalent treatments leave expected returns and variances unchanged, the tilt adds variance without increasing stated expected return. At relative risk aversion $\gamma = 5$, the certainty-equivalent cost of reflection framing is 23.5 basis points per year, roughly an order of magnitude above anchoring's 1 to 2 basis points. The effects survive outside the stylized setting: with five named U.S. equity ETFs, return moments estimated from 2015 to 2024, and a realistic client profile, the preservation frame shifts 9.3 percentage points toward the defensive fund.

We next ask whether prompting can mitigate the problem. Our debiasing ladder strengthens instructions to disregard the cue, culminating in requiring the model to restate each option's identical expected return and variance before allocating. This eliminates anchoring

but reduces reflection framing by only 4 percent. Thus, a model can articulate economic equivalence yet fail to act on it. By contrast, normalizing the description before allocation removes 97 percent of the framing effect. Anchoring is a number the model can be told to disregard; reflection framing is a property of how the outcome is described, which instruction does not reach. Greater capability does not solve it. Reasoning-labeled models show 3.6 percentage points less anchoring but 16.4 points more framing, while moving from the cheapest price quartile to the third increases price nineteenfold and measured intelligence about fourfold with little change in susceptibility. Developer family explains 45 percent of cross-sectional variation, implying that model combinations reduce distortions primarily across developer families.

Our findings contribute to two literatures. First, we sharpen identification of LLM behavioral biases and qualify emerging evidence on LLM rationality. Prior work documents behavioral biases in consumer chatbots and shows that model-generated advice can reinforce investor biases or increase portfolio risk (Winder, Hildebrand, and Hartmann, 2025; Lee, Seo, Park, Lee, Ahn, Choi, Lopez-Lira, and Lee, 2025; Huang, Bie, Na, Ruan, Lei, Yue, and He, 2026; Zhi, Zhang, Wang, Jiang, Ma, Guan, and Shen, 2025; Malberg, Poletukhin, Schuster, and Groh, 2025). A parallel literature finds that LLMs satisfy revealed-preference restrictions (Chen, Liu, Shan, and Zhong, 2023), exhibit human-like decision heuristics (Binz and Schulz, 2023; Chen, Kirshner, Ovchinnikov, Andiappan, and Jenkin, 2025), and perform well on financial questions (Niszczota and Abbas, 2023). Holding the economic problem fixed while varying only its description lets us interpret information-equivalent cue effects as violations of description invariance; informative twins separate them from rational signal responses. None implies description invariance. A model can make consistent choices and use relevant information yet change its decision after a transformation that leaves the economic problem unchanged. This matters when LLMs serve as financial advisers, autonomous decision makers, or experimental substitutes for economic agents.

Second, we identify a distinct distortion in delegated financial advice. Prior work links

poor advice to conflicts, incentives, adviser beliefs, and misconduct (Mullainathan, Noeth, and Schoar, 2012; Gennaioli, Shleifer, and Vishny, 2015; Egan, Matvos, and Seru, 2019; Linnainmaa, Melzer, and Previtero, 2021). Here, the economic environment, preferences, and opportunity set are fixed; distortion comes from how those inputs are encoded in language. As investors increasingly rely on algorithmic recommendations (Dietvorst, Simmons, and Massey, 2015; Logg, Minson, and Moore, 2019), the design of the information pipeline becomes part of the economics of advice.

2 Related Literature

This paper builds on three literatures: behavioral finance evidence on presentation effects in financial choice, research on LLMs as economic agents and advice engines, and the economics of delegated advice and algorithm use.

The heuristics-and-biases literature shows that judgments depend on presentation: people anchor on arbitrary numbers (Tversky and Kahneman, 1974), respond to equivalent outcomes differently across frames (Tversky and Kahneman, 1981), and weight gains and losses asymmetrically, with risk attitudes reversing across domains (Kahneman and Tversky, 1979). These regularities have first-order financial consequences, including the disposition effect (Shefrin and Statman, 1985; Odean, 1998), defaults and inertia in household allocation (Madrian and Shea, 2001; Benartzi and Thaler, 1995), overreaction to news (De Bondt and Thaler, 1985), and belief-based asset-pricing effects (Barberis, Shleifer, and Vishny, 1998; Hirshleifer, 2001); Barberis (2018) surveys this literature. Even equivalent monetary gambles produce different choices when sure outcomes are described as money retained rather than lost (De Martino, Kumaran, Seymour, and Dolan, 2006).

Two studies are especially close to our benchmark because they hold the investment opportunity fixed while varying its presentation. Choi, Laibson, and Madrian (2010) show that allocations across S&P 500 index funds differing only in fees depend on how those fees

are presented. Kronlund, Pool, Sialm, and Stefanescu (2021) find a similar effect outside the laboratory: after a disclosure reform made fees more visible, 401(k) participants shifted toward lower-cost funds although the menu was unchanged. Both document description-invariance failures by human investors in real allocations. We ask whether the technology advising those investors fails the same test, identify cue effects rather than presuming they are biases, and map resulting allocations to factor exposures (Fama and French, 1993). Psychology provides a direct analogue to our informative-twin test: holding a numerical input fixed, Hütter and Fiedler (2019) find that both genuine advice and arbitrary anchors influence judgment, but advice receives greater weight. We apply the same separation to LLM portfolio recommendations and price the residual response to the noninformative cue.

A newer literature studies whether LLM outputs display patterns conventionally labeled as human biases. Since capable LLMs emerged (Brown et al., 2020), they have exhibited human-like heuristics (Binz and Schulz, 2023; Hagendorff, Fabi, and Kosinski, 2023; Jones and Steinhardt, 2022; Chen, Kirshner, Ovchinnikov, Andiappan, and Jenkin, 2025) and been used as simulated economic agents (Horton, Filippas, and Manning, 2023). Demszky et al. (2023) caution that human-like output does not establish human-like cognition, and psychological labels can encourage anthropomorphic inference (Shanahan, McDonell, and Reynolds, 2023). We therefore use bias names for controlled output regularities, not latent processes. Related work on sycophancy shows that models defer to user-stated views and that chain-of-thought rationales can obscure the features driving answers (Perez, Ringer, Lukosiute, Nguyen, Chen, et al., 2023; Sharma, Tong, Korbak, et al., 2024; Turpin, Michael, Perez, and Bowman, 2023). Our anchor instead is explicitly random and expresses no user preference, while the informative twin attributes the same number to portfolio optimization.

A parallel strand asks whether LLMs satisfy rationality criteria. Chen, Liu, Shan, and Zhong (2023) find that GPT's budgetary choices satisfy revealed-preference restrictions more often than human choices, while Mei, Xie, Yuan, and Jackson (2024) find human-like play in classic games that remains sensitive to framing. Ross, Kim, and Lo (2024) find LLM

economic behavior neither reliably human-like nor reliably rational, and Goli and Singh (2024) report discount rates above human benchmarks whose measured patience varies with prompt language. Bini, Cong, Huang, and Jin (2026) find that more advanced models are more human-like on preference tasks and more rational on belief tasks, but do not study paired portfolio advice with a fixed opportunity set and priced allocation consequences. These findings help motivate deployment in advice and the research uses surveyed by Korinek (2023), but revealed-preference consistency does not imply invariance across equivalent descriptions. We test the latter directly.

Closest to our design, Winder, Hildebrand, and Hartmann (2025) find that LLM-generated investment advice reinforces classic biases and raises portfolio risk across three to five consumer chatbots; related studies document confirmation, product, and anchoring biases in panels of up to twenty models (Lee, Seo, Park, Lee, Ahn, Choi, Lopez-Lira, and Lee, 2025; Zhi, Zhang, Wang, Jiang, Ma, Guan, and Shen, 2025; Malberg, Poletukhin, Schuster, and Groh, 2025; Huang, Bie, Na, Ruan, Lei, Yue, and He, 2026). Our design differs in three ways. Paired prompts hold expected returns, risk, horizon, choice set, and constraints fixed, so cue responses identify description-invariance failures rather than changed decision problems. Informative twins separate distortion from rational information use. Factor-exposure and certainty-equivalent calculations price the distortion, while the 54-model panel establishes that it is not confined to a few consumer chatbots.

A related literature examines LLMs in finance and delegated advice. Models forecast returns from text (Lopez-Lira and Tang, 2026), analyze financial statements (Kim, Muhn, and Nikolaev, 2024), serve as textual-analysis instruments (de Kok, 2025), and answer advice questions (Niszczota and Abbas, 2023). Investors already act on these tools: Cheng, Lin, and Zhao (2025) find that retail trading falls during ChatGPT outages. On the supply side, Cao, Jiang, Yang, and Zhang (2023) show that firms rewrite disclosures when they expect machines to process them, which is especially relevant when wording itself can influence recommendations.

The robo-advising literature documents both benefits and costs of delegated allocation (D'Acunto, Prabhala, and Rossi, 2019; Rossi and Utkus, 2024). Automated advice expands portfolio management to smaller and less wealthy accounts (Reher and Sokolinski, 2024), and investors often combine human and algorithmic input (Bianchi and Brière, 2026). Human advice is itself distorted: advisers may reinforce client biases (Mullainathan, Noeth, and Schoar, 2012), trust can sustain fees above cost (Gennaioli, Shleifer, and Vishny, 2015), adviser-specific views can dominate customization (Foerster, Linnainmaa, Melzer, and Previtero, 2017), advisers recommend underperforming strategies they themselves follow (Linnainmaa, Melzer, and Previtero, 2021), and misconduct is persistent (Egan, Matvos, and Seru, 2019).

Our distortion requires none of these commissions, private beliefs, or career concerns: description alone can alter recommendations with the economic problem fixed. It can matter even without mechanical implementation because advice that agrees with recipients can raise confidence without improving accuracy (Soll, Palley, and Rader, 2022), and Steyvers et al. (2025) report a similar effect for longer LLM explanations. Whether such distortions affect outcomes depends on reliance on algorithmic advice, the focus of the algorithm aversion and appreciation literature (Dietvorst, Simmons, and Massey, 2015; Logg, Minson, and Moore, 2019). We contribute evidence that these distortions arise within the recommendation, appear across developers, resist generic instruction, and do not decline monotonically with model scale.

3 Framework and Predictions

3.1 Description invariance as a benchmark

We study whether financial advice changes with the description of an otherwise fixed economic problem. In our experiments, stated economic inputs, the feasible choice set, and constraints fully specify the allocation problem, while surrounding wording adds no decision-

relevant information. An advice rule based only on the economics should therefore be invariant to that wording.

This requirement is *description invariance* (Tversky and Kahneman, 1981): economically equivalent descriptions should generate the same recommendation. The paired prompts in Section 1 provide a canonical example. Client characteristics and the investment opportunity set are identical, while one outcome is described differently; an invariant adviser assigns the same portfolio weights in both. We define a response as biased when an economically irrelevant descriptive change alters the recommended allocation. This definition requires no assumptions about model beliefs, emotions, or preferences, only that the advice rule changes when the underlying economic problem does not.

Description invariance is a natural benchmark for two reasons. First, it is an irrelevance condition: once expected returns, risk, horizon, and the investment opportunity set are fixed, repackaging the same information should not change portfolio weights, analogous to classic irrelevance arguments in finance (Modigliani and Miller, 1958). Second, it is a basic consistency requirement of rational choice (Tversky and Kahneman, 1986). Section 4 implements the benchmark directly: within each scenario, the two arms differ only in the cue sentence, with option labels and ordering fixed. The null hypothesis is therefore a zero arm difference.

3.2 Why description invariance may fail

Description invariance is not a restriction imposed directly by standard LLM training. Large language models are initially trained to predict tokens conditional on preceding token sequences (Brown et al., 2020). Two sentences can describe the same economic state while remaining different model inputs, and the next-token objective does not require them to induce identical internal representations or choices.

Instruction tuning and reinforcement learning from human feedback alter the objective but need not impose invariance. These stages train models to follow instructions and produce responses preferred by human evaluators (Ouyang, Wu, Jiang, Almeida, Wainwright,

et al., 2022), generally treating each prompt-response pair separately rather than restricting decisions across economically equivalent prompts. A model can therefore become more helpful, accurate, or instruction-responsive without becoming invariant to economically irrelevant input transformations. Consistent with this possibility, performance varies across semantically equivalent phrasings and meaning-preserving prompt formats (Sun, Shaib, and Wallace, 2024; Sclar, Choi, Tsvetkov, and Suhr, 2024).

This training structure makes description dependence plausible ex ante. Prompt language can activate learned statistical associations even when they are irrelevant to the stated economics. Loss language, for example, often appears in contexts associated with caution or risk, while a suggested allocation resembles a quantity that normally carries decision-relevant information. Models may therefore respond to such cues even when explicitly told they are uninformative. The issue is not whether LLMs understand the economics: a model may correctly identify expected returns and risks yet still allow their linguistic representation to affect its recommendation.

This reasoning yields three predictions. First, information-equivalent cues can shift allocations despite leaving the economic problem unchanged. Because each treatment makes one option more prominent or attractive along a particular dimension, the alternative to invariance is directional: recommendations should move toward the targeted option. We do not predict effects for every manipulation; dependence should be strongest when the cue resembles information that ordinarily matters for choice.

Second, models should distinguish irrelevant cues from genuine information. If an allocation-shaped number matters partly because such numbers are usually informative, attributing it to a credible optimization procedure should produce a larger response than explicitly labeling it random. The same logic applies to peer choices and recent returns. This comparison separates rational responsiveness to information from residual sensitivity to representation.

Third, interventions should differ according to where description dependence enters the

advice process. If models fail only to recognize equivalence, explicitly stating the equivalence or requiring them to restate relevant economic inputs should reduce the effect. If representation matters even after equivalence is recognized, stronger instructions need not eliminate it, while neutralizing the description before allocation should be more effective. The debiasing experiments distinguish these mechanisms.

We make no monotonic prediction with respect to capability. Scale, reasoning training, and instruction tuning may improve recognition of semantic equivalence, reducing description dependence, but greater capability may also improve extraction of contextual associations embedded in language. Whether more capable models are more description-invariant is therefore empirical.

3.3 Advisory workflows

The advisory setting changes the economic interpretation of these effects. In conventional behavioral experiments, framing operates through the investor's own choice. In an AI-assisted workflow, it can instead enter through the technology mapping client information into recommendations, including questionnaires, prompt templates, retrieved documents, adviser notes, guardrails, and the model. A downside description may originate in a questionnaire, a suggested allocation in an adviser interface, and peer or performance information in retrieved material. We isolate the model-generated allocation while holding investment facts fixed, and Section 5 tests whether confirmed effects survive when synthetic options are replaced with named funds.

3.4 Information-equivalent and information-bearing cues

A central design distinction is between *information-equivalent* and *information-bearing* cues. Information-equivalent cues change representation without changing investment-relevant content, so any resulting recommendation change violates description invariance. Information-bearing cues can contain information a rational adviser may use, so a response

alone does not establish bias.

We study nine cues. Six are information-equivalent: anchoring, reflection framing, global framing, defaults, denomination, and downside wording. Three are potentially information-bearing: herding, recency, and availability. Anchoring introduces an irrelevant number as a suggested allocation; reflection framing describes identical terminal outcomes as value preserved versus lost. Table 1 defines each cue, its behavioral analog, and the wording contrast used to isolate it.

We distinguish design status from empirical support. A *confirmed information-equivalent bias* is an irrelevant cue for which the data reject description invariance in a pre-specified test or independent confirmation wave. An *exploratory information-equivalent bias candidate* is an irrelevant cue that moves advice outside the confirmatory set. An *information-sensitive response* is a response to a potentially informative cue, where movement alone does not imply bias. We use *pre-registered* for design status and *confirmed* for resulting evidence.

The informative-twin design further separates bias from information use. For anchoring, the model sees the same suggested allocation under two source descriptions: one randomly generated and explicitly uninformative, the other attributed to a portfolio optimization for the client's mandate. Holding the number fixed, the response difference measures whether models distinguish a credible signal from an equivalent irrelevant cue. Herding and recency twins similarly compare informative and uninformative versions of the same underlying cue.

3.5 Style footprints and the welfare criterion

Description dependence matters economically when it changes the portfolio. Each experimental option has fixed factor loadings, so weight changes map directly into style exposures. Increasing the defensive option, for example, raises quality and low-volatility exposure, while shifting toward the recent winner raises momentum exposure. We call this mapping the cue's *style footprint*.

Table 1. The Nine Wording Changes and the Biases They Probe

The table lists the nine cue manipulations and groups them by whether they are information-equivalent (Panel A, where any response violates description invariance) or information-bearing (Panel B, where a response may reflect rational signal use). For each cue, it identifies the human behavioral analog, defines the bias, and reports the minimal wording contrast between the control arm (Arm 0) and treatment arm (Arm 1). Every cue is designed to move advice toward the same target option. Figure 1 shows two complete rendered instruments.

Manipulation	Behavioral analog	What the bias is	Cue contrast (Arm 0 → Arm 1)
Panel A: Information-equivalent cues (a response is a description-invariance failure)			
Anchoring	Anchoring heuristic	An irrelevant reference number pulls the recommendation toward itself.	Low → high random suggested allocation, stated to be uninformative.
Framing (reflection)	Reflection effect / framing	Identical terminal outcomes described as a loss versus as preserved value shift risk taking.	Loss framing → preservation framing of the same outcomes.
Framing (global)	Framing	Restating one portfolio-level probability moves advice even though the odds are unchanged.	Loss-probability → keep-probability statement.
Default	Status-quo / default effect	Marking one option as the preselected choice pulls advice toward it.	Broad-market default → target option as default.
Denomination	Money illusion / denomination	The same loss stated in dollars versus percentages moves advice.	Percent loss → dollar loss.
Downside	Loss aversion / framing	Describing the same outcomes as value retained versus value lost moves advice.	Value retained → value lost.
Panel B: Information-bearing cues (a response may reflect rational use of a signal)			
Herding	Herding / social proof	Concentrated peer choices may signal quality; over-response to them is herding.	Peers split evenly → peers choose the target option.
Recency	Recency / return extrapolation	Recent outperformance is extrapolated beyond its informational value.	Similar recent returns → target recently outperformed.
Availability	Availability heuristic	A vivid recent episode makes one option easier to recall and overweight.	No recent episode → vivid recent episode.

The design also provides a natural welfare criterion. Every option has the same stated expected return and variance. A wording-induced allocation change therefore creates a portfolio tilt without a stated expected-return benefit, making any additional variance uncompensated risk (Markowitz, 1952). We measure this distortion by the certainty-equivalent value of the tilt.

Formally, let $d = w_b - w_c$ denote the difference between bias-arm and control-arm al-

locations. At risk aversion γ, the change in certainty equivalent attributable to the tilt is

$$\Delta \mathrm{CE} = -\tfrac{\gamma}{2}\, d'\Sigma d,$$

where Σ is the covariance matrix of portfolio options. Treating d as a self-financing, zero-expected-return overlay isolates the incremental risk created by wording. Comparing the two full portfolios would also include the cross term $\gamma w_c'\Sigma d$ and depend on the control portfolio's level. We instead focus on the tilt to isolate the component attributable to description. In the main specification, Σ is constructed from choice-set factor loadings and each option is rescaled to stated variance v; a robustness specification imposes equal correlations. The expected-return term drops out because all options have the same stated expected return.

4 Experimental Design and Data

The experiment uses a matched-pair design within each model. We ask the model to allocate a portfolio across a fixed menu, then repeat the problem with one sentence rewritten. Expected returns, variances, horizon, mandate, and menu are unchanged. We call the two versions the *control* and *treated arms* and each parameterization a *scenario*. Both arms are run for every model-scenario pair, so differencing removes model- and scenario-specific levels and leaves wording as the identifying variation.

The evidence comes from six collections of model responses, summarized in Table 2. Each uses the same contrast, and we report them separately rather than pooling.

4.1 The allocation task

The model acts as an adviser. Each prompt states a client mandate, common expected return and variance, horizon, five portfolios with brief style descriptions, the experimental cue, and asks for weights summing to one. The mandate is long only with no leverage.

Table 2. The Six Experiments and the Models Used in Each

Every collection holds the economic problem fixed within a model-scenario cell and varies only the cue. Models is the number of checkpoints in that collection. Fifty-two models appear in all six; the exceptions are `meta-llama/llama-3-8b-instruct`, retired during the study, `qwen/qwen3.6-35b-a3b`, which returns reasoning tokens without message content, and `anthropic/claude-opus-4.8`, added only for the named-fund wave. That wave alone applies the pre-specified rule dropping models with invalid-response rates above one half, removing four more models. The recorded pre-analysis plans for the confirmation and twin waves instead retain the realized valid-response panel, including high-invalid models. The rule is prompt-specific: `stepfun/step-3.7-flash`, for example, is unparseable on 84 percent of named-fund prompts but 32 percent of confirmation prompts. Individual tests can have fewer models than the row count when a model has no parseable response for that choice set; table notes identify these cases. No collection drops models based on estimated effects. New scenarios indicates a newly generated scenario bank; explicit equivalence reuses the main-panel scenarios and adds an irrelevance instruction. "Confirmation" denotes a separately designed test with fresh scenarios, not nonoverlapping models, and also fixes decision rules before analysis. Model sets overlap. Overlap, sample sizes, and further design details appear in the Internet Appendix.

Collection	Purpose	Models	New scenarios
Discovery panel	All nine cues: how much each wording change moves the targeted weight	54	Yes
Explicit equivalence	Repeats anchoring and downside after stating that the cue is uninformative	52	No
Confirmation wave	Retests reflection framing on fresh scenarios	53	Yes
Informative twins	Same number, described as uninformative or as a credible signal	53	Yes
Named ETFs	Repeats the confirmed cues on named funds with estimated moments	49	Yes
Debiasing	Three instruction rungs plus a preprocessing rewrite of the prompt	53	Yes

Figure 1 reproduces two complete prompts, and the Internet Appendix gives verbatim text for all nine.

The menu is deliberately generic: a broad market fund, value tilt, momentum tilt, low-volatility and quality tilt, and small higher-risk tilt. None is a real fund or has a name or ticker, preventing prior knowledge of specific funds from driving allocation. Section 5.6 returns to named funds.

Before data collection, we assign each portfolio fixed factor loadings. These never appear in the prompt and do not affect the task; they translate allocation changes into factor exposures and supply the covariance matrix for the welfare calculation in Section 6. They are stylized, not estimated real-fund betas.

The outcome is the weight on the cue-targeted portfolio. Because weights are long only and sum to one, it lies between zero and one, and treatment effects are reported in allocation

percentage points.

Models differ in how they accept structured output. Those supporting a schema are asked for JSON and receive no system message; others receive a short formatting instruction appended to the same user message. Both paths use identical parsing rules. A model's path is fixed across arms and therefore cannot contaminate the contrast.

4.2 What varies, and what does not

Within a scenario, temperature is zero and portfolio order is drawn once and fixed across arms. The prompts therefore agree on model, scenario, option descriptions and order, stated moments, mandate, and response format, differing only in one sentence. This isolates wording from task difficulty, arithmetic load, domain knowledge, and menu composition.

We use two informational regimes. The baseline shows only what an advice system would ordinarily observe. The explicit-equivalence regime adds a sentence stating that the cue is uninformative and should not affect the recommendation, testing whether description dependence survives after irrelevance is made explicit.

The discovery collection uses 12 scenarios per cue and one fixed wording per scenario.[2] Because one wording cannot separate a cue from its particular phrasing, the confirmation wave uses fresh scenarios and tests two paraphrases alongside the core wording (Section 5). The explicit-equivalence, confirmation, and informative-twin collections each have their own design manifest and scenario bank.

Table 3 summarizes the discovery design and realized sample. Whereas Table 1 sorts cues by informational content, Table 3 records pre-registration status, explicit-equivalence coverage, whether the cue names a labeled option, and parseable response counts.

[2]When a model-scenario-arm cell has multiple draws, the collector resends the identical prompt. Repeated draws are not paraphrases; they average residual provider variation even at temperature zero.

Table 3. Pre-Registration Status and Sample Size for Each Wording Change

Each row reports pre-registration class, whether the manipulation was also collected under explicit equivalence, whether the cue names a labeled option, the number of parseable models, and successful observations with refusal or parse-failure rates in parentheses. All manipulations use 12 scenarios with one to three draws per model-scenario cell. The class column records the pre-registration battery (confirmatory or exploratory), not whether the effect is empirically confirmed: reflection framing is exploratory here because it was not preregistered and is independently confirmed in a follow-up wave (Table 5). Table 1 defines each manipulation and behavioral analog.

Manipulation	Class	Explicit-equiv. regime	Labels in cue	Models	N (invalid %)
Anchoring	Confirm.	yes	no	54	1834 (1.6%)
Framing (reflection)	Explor.	no	yes	52	1668 (3.0%)
Herding	Explor.	no	yes	52	1589 (1.7%)
Recency	Explor.	no	yes	52	1712 (1.6%)
Framing (global)	Explor.	no	no	54	1757 (2.3%)
Default	Explor.	no	yes	52	1651 (2.2%)
Denomination	Explor.	no	no	52	1688 (2.3%)
Availability	Explor.	no	yes	52	1710 (1.3%)
Downside	Confirm.	yes	no	54	1798 (1.7%)

Figure 1. The Prompts Used to Test Anchoring and Reflection Framing

Panel A gives the complete suggested-allocation anchoring instrument: the full user message, including the five option descriptions and response-format instruction. The arms differ only in the suggested weight (12 versus 65 percent), and no system message is used. Panel B gives the reflection-framing cue sentences: identical prospects described as losses versus preserved value, replacing the anchoring cue in the same instrument. Both are extracted from the generation code and match what was sent to the models. The Internet Appendix reports the verbatim prompt, both arms, and both response paths for all nine manipulations.

Panel A. Suggested-allocation anchoring (control vs. bias arm).

An adviser allocates a client account. Before allocating, a random, uninformative placeholder was displayed for the value-tilted option P2; it is random and carries no information about the appropriate allocation. You are advising on a portfolio allocation with an expected return of about 0.07 and variance of about 0.04 over a 10 year horizon, under a moderate-risk mandate (long only, no leverage, choose only from the listed portfolios). The random placeholder displayed about **[12 / 65]** percent in the value-tilted option P2. Assign weights summing to 1.0 across the options, using each option's id (for example P1) as the key. Options: P5: Tilted toward small, higher-risk, higher-turnover names; P1: A broad market fund with no style tilt; P4: Tilted toward stable, high-quality, low-volatility companies; P3: Tilted toward recent strong performers (momentum); P2: Tilted toward cheap, out-of-favor value companies.

Response format: a JSON object mapping each option id to its weight, summing to one; the schema path uses no system message.

Panel B. Reflection framing (control vs. bias arm).

Loss frame: The low-volatility option (Portfolio 4) is almost certain to lose about 15 percent of value, while the higher-risk option (Portfolio 5) has a 50 percent chance of losing nothing and a 50 percent chance of losing about 30 percent.

Gain frame: The low-volatility option (Portfolio 4) is almost certain to preserve about 85 percent of value, while the higher-risk option (Portfolio 5) has a 50 percent chance of preserving all of it and a 50 percent chance of preserving about 70 percent.

4.3 The model panel

The panel contains 54 models from 17 developer families, chosen to span the market rather than represent one vendor. Thirty-four are open-weight and 20 closed-weight. Among the 31 reporting parameter counts, sizes range from 4 to 671 billion. Sixteen run in reasoning mode and 38 in standard mode. All models and their characteristics are listed in the Internet Appendix.

A model identifier does not define a permanently fixed object. Reasoning status is recorded for the configuration actually queried; where multiple settings exist, our classification applies only to the observed one. Vendors may also update or retire models under similar identifiers. We therefore study outputs at the recorded collection dates. The replication package preserves raw responses, parser, and collection manifests, while the Internet Appendix records identifiers, dates, temperature, and schema mode for every wave rather than relying on future queries for reproduction.

Only within-model, within-scenario comparisons carry causal weight because the same model faces the same problem under two descriptions. Cross-model comparisons, including heterogeneity analysis in Section 8, are descriptive because size, vendor, and training regime are not randomly assigned.

Attrition is minor. Of 15,715 discovery draws, 308 are invalid, a 1.96 percent rate. Rates by model, cue, and arm appear in the Internet Appendix, which also documents the parser and shows robustness to dropping models with high invalid rates.

4.4 Estimation and inference

Let y_{msar} be the weight on the cue-targeted portfolio for model m, scenario s, arm a, and repeated draw r, where $a = 1$ is treated and $a = 0$ control. We estimate the average cue-induced change, weighting each model equally.

We first average repeated draws within model-scenario-arm cells to obtain $\bar{y}_{msa}$, difference

arms within each pair, average across S scenarios for each model, and then across M models:

$$\hat{\tau}_m = \frac{1}{S}\sum_{s=1}^{S}\big(\bar{y}_{ms1} - \bar{y}_{ms0}\big), \qquad \hat{\tau} = \frac{1}{M}\sum_{m=1}^{M}\hat{\tau}_m. \tag{1}$$

Equal weighting defines the estimand rather than serving as an efficiency choice: $\hat{\tau}$ is the average description effect across panel models, so models with more parseable responses receive no extra weight.

As a draw-level robustness check, we estimate

$$y_{msar} = \alpha_m + \alpha_s + \beta\, T_a + \varepsilon_{msar}, \tag{2}$$

where T_a indicates the bias-consistent arm. Model and scenario fixed effects absorb persistent allocation differences across models and investment problems, so β is identified by within-model-scenario arm contrasts. It differs from $\hat{\tau}$ only in weighting: β weights draws equally, while $\hat{\tau}$ weights models equally. They coincide with balanced cells and differ only when parse failures create imbalance.

For model-level estimates, tables report the standard error of $\hat{\tau}$ across the M model effects, making the model the unit of inference unless noted. For the draw-level specification, we also use a wild cluster bootstrap (Cameron, Gelbach, and Miller, 2008) by developer family, allowing within-developer correlation, and randomization inference (Young, 2019), permuting arm labels within model-scenario blocks. This permutation matches the paired design's exchangeability under the sharp null. The p-value is the share of placebo assignments with effects at least as large as observed. The Internet Appendix reports draw-level fixed-effects estimates and cases where imbalance changes the estimate by at least 0.005.

The design identifies the causal cue effect for the models and scenarios studied because both arms are observed in every model-scenario cell with nothing else varying. The 54 models, however, are a curated panel of commercially and publicly available systems at collection time, not a random sample from a stable LLM population. Cross-model averages

therefore describe this panel rather than all current or future models.

We address multiple testing separately for confirmatory and exploratory inference. Within the pre-specified confirmatory family, we apply Holm (Holm, 1979); across all nine manipulations, we also report Benjamini-Hochberg false-discovery-rate-adjusted q-values (Benjamini and Hochberg, 1995). This provides family-wise error control for the small confirmatory set while evaluating the broader exploratory set jointly.

Anchoring and downside were the two original pre-registered manipulations and were collected under both baseline and explicit equivalence. Anchoring is confirmed and downside is not (Section 5). Reflection framing, herding, and recency emerge from the exploratory panel. We retest reflection framing in the confirmation wave and use informative twins for anchoring, herding, and recency, holding the cue fixed while changing whether it is informative. Pre-specification materials, and all deviations from the recorded designs, are reproduced in the Internet Appendix.

5 Main Results

This section tests whether LLM portfolio advice satisfies description invariance: whether recommendations remain unchanged when economically irrelevant prompt features vary while decision-relevant information is fixed. Unless noted otherwise, effects are model-averaged changes in the target-option weight, so significance reflects systematic responses across models rather than idiosyncratic scenarios.

5.1 Anchoring under information equivalence

Our first confirmatory test examines *suggested-allocation anchoring*. We present a random number as a suggested allocation to one option and explicitly state that it is random and uninformative. The question is therefore whether an arbitrary allocation suggestion pulls advice toward it, not whether any number in the prompt matters. A later confirmation wave

sharpens this distinction by showing that an unrelated administrative number has essentially no effect (Internet Appendix; Figure 1, Panel A).

The high random anchor raises the recommended weight on the anchored option by +0.087 in the 54-model panel, with 93 percent of models moving in the predicted direction (Table 6). Relative to the 0.20 equal-weight benchmark, an 8.7 percentage-point shift moves the option more than two fifths of the benchmark on wording alone. On a $1 million mandate, this corresponds to about $87,000 of recommended holdings moving toward an option for reasons that contain no information about its investment merits. This magnitude is economically meaningful because the cue changes neither the client's mandate nor any stated feature of the investment opportunity set.

Explicitly instructing the model to disregard the cue attenuates but does not eliminate the effect. In the confirmatory-regime panel, anchoring falls from +0.101 under naturalistic wording to +0.057 when the prompt states that the cue is irrelevant, a 44 percent reduction. The remaining effect is highly significant ($p = 0.0001$; Table 4). Thus, explicit equivalence appears to discipline use of the numerical suggestion only partially: recommendations continue to move toward a number the model has been told contains no decision-relevant information. A similar reduction appears in the named-fund setting (Section 5.6), suggesting that instructions can reduce sensitivity to arbitrary numerical suggestions. As shown below, the same is not true for loss-versus-preservation framing. The other pre-registered confirmatory cue, downside wording, is small, negative, and within the bounded-null region.

5.2 Reflection framing: discovery and confirmation

Figure 2 and Table 6 summarize all nine manipulations. The largest information-equivalent effect is *reflection framing*: identical terminal outcomes are described as value lost in one arm and value preserved in the other (Figure 1, Panel B). Expected outcomes, risks, and investment opportunities are unchanged. We use the term reflection framing to distinguish

Table 4. Anchoring and Downside Wording, With and Without an Explicit Irrelevance Instruction

This table splits effects by regime for the two confirmatory manipulations specified in advance, anchoring and downside. The explicit-equivalence regime states directly that the options are equivalent for the decision and that the cue is irrelevant. A manipulation that continues to move advice in this regime therefore violates description invariance under the prompt's own stated terms. Effect SEs are model-level because the table uses the confirmatory-regime panel, and M denotes the number of models contributing a usable pair in each cell. Two of the 54 discovery-panel routes are absent from this wave, leaving 52 models; M falls to 51 for baseline anchoring, where one further model returns no usable pair. No model is dropped here for its invalid-response rate. The seven original non-confirmatory manipulations were collected only under the baseline regime and therefore have no explicit-equivalence counterpart in the original panel (Section 4).

Manipulation	Regime	Effect (SE)	p	M
Anchoring	Baseline	+0.101 (0.011)	<0.001	51
Anchoring	Explicit-equiv.	+0.057 (0.014)	<0.001	52
Downside	Baseline	-0.013 (0.006)	0.035	52
Downside	Explicit-equiv.	-0.003 (0.002)	0.092	52

this contrast from strict prospect-theory reflection, where outcomes cross a reference point.

Preservation wording produces more defensive advice, while loss wording produces more risk-seeking advice. Moving from the loss to preservation description raises the defensive-option weight by +0.222, and every model moves in the same direction. This is not simply favorable wording attracting allocation. Both arms re-express both options, and in the preservation arm the most favorable phrase belongs to the higher-risk option, which has a 50 percent chance of preserving all value, yet that option loses weight. What differs across arms is therefore not which option sounds more favorable, but whether outcomes are expressed as losses or as value retained, and with it the model's willingness to accept dispersion around the same expected outcome. This distinction matters because both options are re-expressed in each arm, leaving no option uniquely targeted by positive wording. Section 6 restates the same result in discrete-choice terms.

Nor is the effect driven by any mention of downside risk. The global-framing cue (Table 1) restates a portfolio-level probability without changing the loss-versus-preservation description and moves allocation by only +0.021. The order-of-magnitude difference points to representation rather than the mere prominence of downside probability.

We conducted a separate confirmation wave before treating reflection framing as a main

result. Using fresh scenarios, 53 models, and decision rules fixed before analysis, the core replication is +0.204, with all 53 models moving in the predicted direction (Table 5). Removing style labels yields +0.232, ruling out the possibility that calling one option "defensive" drives the result. Explicitly stating that the descriptions are equivalent still yields +0.162. This condition directly tells the model that the wording difference should not affect the decision, making the remaining effect especially informative about description dependence. All three tests have Holm-adjusted randomization p-values of 0.001, and two additional paraphrases also move allocations in the predicted direction. We therefore treat reflection framing as the second confirmed information-equivalent bias.

Explicit equivalence reveals a sharp contrast with anchoring. Declaring the frames equivalent reduces reflection framing from +0.204 to +0.162, or 21 percent, versus a 44 percent decline for anchoring (Table 4). Direct instructions therefore appear substantially more effective against an arbitrary number than against linguistic representation, an asymmetry that recurs below.

One alternative is that loss wording conveys information about the client's reference point or concern about downside risk, so the model is inferring an objective rather than violating invariance. Three design features work against this interpretation. First, the frame is scenario text rather than a client statement and is attributed to no particular source. Second, both arms report identical expected return and variance. Third, the effect remains +0.162 when the prompt explicitly says the descriptions are equivalent for the decision. If framing mattered only because it revealed a changed client objective, that inference becomes substantially less defensible once decision-relevant differences are explicitly ruled out.

There is nevertheless a limit to this test. The prompts do not state that the client's objective and reference point are identical across arms. A stricter instrument could fix these directly, for example, "same investor, same objective, same reference point, notation only." The evidence therefore does not eliminate every possible pragmatic interpretation of the wording, but persistence after equivalence is explicitly stated makes such explanations

increasingly difficult to sustain.

Table 5. Independent Replication of the Reflection-Framing Effect

Effects are model-averaged changes in the target-option weight; SEs and family SEs are model-random and family-clustered, respectively. Wild p is a family-clustered wild-cluster-bootstrap p-value, and RI p is a randomization-inference p-value obtained by permuting frame labels within model-by-scenario blocks. Holm RI applies the Holm correction across H1–H3. The three pre-registered tests are the core naturalistic replication (H1), a neutral-label variant that removes style names from the choice set (H2), and a rigorous variant that explicitly states that the two frames are equivalent for the decision (H3). The wave uses the full 53-model panel; $M = 49$ for H2 because four models return no parseable response on the neutral-label choice set.

Test	Estimate	Model SE	Family SE	Wild p	RI p	Holm RI p
H1: Core naturalistic	0.204	0.018	0.014	<0.001	<0.001	0.001
H2: Neutral naturalistic	0.232	0.030	0.038	<0.001	<0.001	0.001
H3: Core rigorous	0.162	0.030	0.034	<0.001	<0.001	0.001

5.3 Bounded nulls and multiple-testing discipline

Not every irrelevant cue moves advice. To distinguish economically small effects from imprecise zeros, we use two one-sided tests of equivalence (TOST). Rather than testing whether an effect equals zero, TOST asks whether it lies within a pre-specified region of economic insignificance, here ± 0.03 of target-option weight, roughly one third of the main anchoring effect.

Denomination, the status-quo default, and downside wording satisfy this bounded-null criterion. Their confidence regions are tight enough to classify the effects as economically small, showing that LLM advice is not mechanically sensitive to every irrelevant presentation change. Global framing and availability are inconclusive: their point estimates are small, but cross-model variation is too large to establish that the true effects lie within the ± 0.03 band.

Multiple-testing adjustments give the same broad picture. Anchoring survives Holm correction within the confirmatory set, while downside does not. Across all nine manipulations, eight estimates differ from zero after Benjamini-Hochberg adjustment at $q < 0.05$, but this does not imply eight economically meaningful biases. Several effects are small, some run op-

posite the corresponding human-bias prediction, and others involve potentially informative cues. We therefore emphasize confirmatory status, economic magnitude, and equivalence bounds rather than significance alone.

Because several models share developers, the Internet Appendix also applies a wild cluster bootstrap at the 17-developer level. The four largest effects are unchanged, while availability loses support. Leave-one-out and distributional analyses likewise show that the headline effects are broadly distributed rather than driven by a few extreme models.

Table 6. How Much Each Wording Change Moves the Recommended Allocation

Effect is the model-averaged change in the recommended target-option weight, with family-clustered SEs in parentheses; p is the model-as-random test and BH q is the Benjamini-Hochberg false-discovery-rate-adjusted value. The % models+ and % families+ columns report the share of models and developer families whose effects run in the predicted direction. The equivalence column uses two one-sided tests against a ±0.03 small-effect band: *moves* indicates that the effect lies outside the band, *bounded null* that the tests place it inside the band, and *inconclusive* that neither conclusion is supported. The class column records discovery-panel status. Reflection framing is exploratory here because it was not preregistered and is independently confirmed in Table 5. $M = 52$ for reflection framing, herding, recency, default, denomination, and availability because two models (`qwen3-32b`, `qwen3.7-plus`) did not return parseable responses. Both arms lose the same two models, so the within-model contrast is unaffected. See the Internet Appendix for the attrition audit.

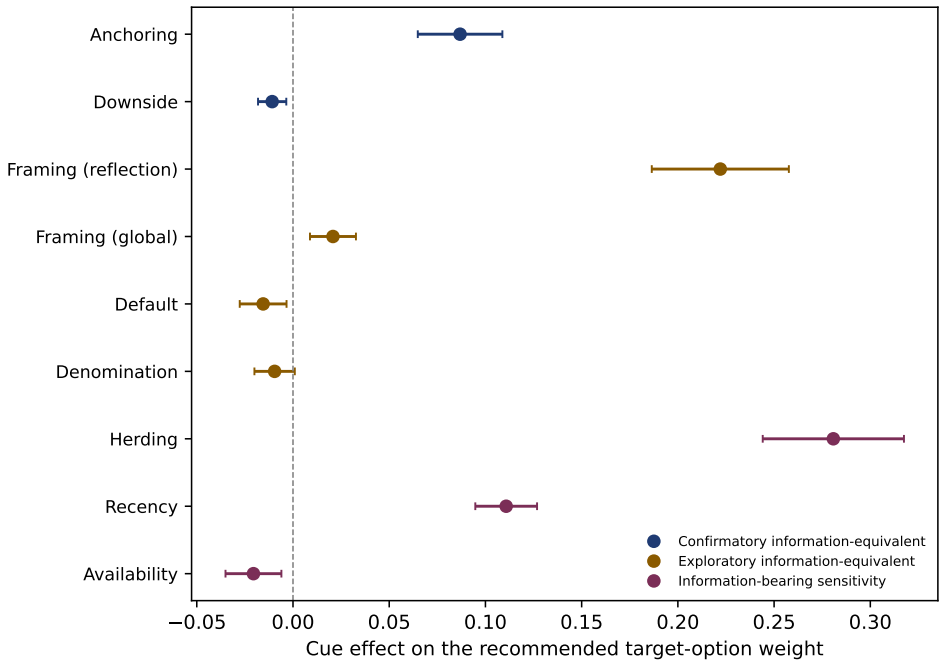
Manipulation	Class	Effect (SE)	p	BH q	% models+	% families+	Equivalence
Panel A: Confirmatory information-equivalent tests							
Anchoring	confirm.	+0.087 (0.014)	<0.001	<0.001	93%	94%	moves
Downside	confirm.	-0.011 (0.003)	0.006	0.008	35%	12%	bounded null
Panel B: Exploratory information-equivalent tests							
Framing (reflection)	explor.	+0.222 (0.021)	<0.001	<0.001	100%	100%	moves
Framing (global)	explor.	+0.021 (0.007)	0.001	0.002	78%	88%	inconclusive
Default	explor.	-0.016 (0.005)	0.016	0.018	25%	12%	bounded null
Denomination	explor.	-0.010 (0.006)	0.080	0.080	38%	24%	bounded null
Panel C: Information-bearing sensitivity tests							
Herding	explor.	+0.281 (0.020)	<0.001	<0.001	100%	100%	moves
Recency	explor.	+0.111 (0.009)	<0.001	<0.001	98%	100%	moves
Availability	explor.	-0.021 (0.010)	0.008	0.010	37%	24%	inconclusive

5.4 Separating bias from information use: the informative twins

The informative-twin tests address whether apparent bias instead reflects attempts to extract information from a cue. As described in Section 3, each test pairs an uninformative cue with

Figure 2. Which Wording Changes Move LLM Advice

Model-averaged effects on target-option weight with 95% intervals across the 54-model panel (52 models for the six cues with two non-parsing models). Colors distinguish confirmatory information-equivalent, exploratory information-equivalent, and information-bearing sensitivity tests.



a counterpart that a rational adviser could reasonably treat as informative.

For anchoring, we compare the random-anchor manipulation with an optimization-attributed twin using matched low and high suggested allocations. In one condition, the number comes from an arbitrary placeholder explicitly labeled uninformative; in the other, the identical number is described as the output of a portfolio optimization for the client's mandate. No optimization is actually performed, so only the stated provenance changes. The contrast measures sensitivity to a source represented as informative, not accuracy in using a verified optimization signal. We use the common 53-model set so the numerical anchors are matched.

Source attribution matters substantially. The optimization-attributed anchor moves target weight by +0.548, versus +0.088 for the random anchor in the same sample (Internet Appendix). Models therefore respond more than six times as strongly when the same number

is represented as optimizer output. This directional response is consistent with models placing greater weight on a cue represented as economically informative. Yet the random-anchor effect remains 16.1 percent as large: almost nine percentage points still follow a number explicitly declared uninformative. Greater responsiveness to a credible source therefore does not imply an ability to ignore the same cue once its information content is removed.

The confirmation wave further clarifies the scope of the finding. An unrelated administrative reference number moves advice by only -0.002 (randomization $p = 0.388$), and a neutral-label anchoring replication is similarly near zero at -0.004 ($p = 0.392$; Internet Appendix). These near-null results narrow rather than overturn the original finding. The suggested-allocation anchor on a labeled choice set survives explicit equivalence (+0.057; Table 4), remains positive in all three tested target positions, and equals +0.088 in the 53-model informative-twin sample. By contrast, a number elsewhere in the prompt, or an allocation suggestion attached to a choice set stripped of style labels, does not reliably anchor recommendations. Thus, the evidence supports a narrower claim than generic numerical anchoring: an arbitrary number presented specifically as an allocation suggestion for a labeled investment option systematically moves advice.

The same logic applies to herding and recency, where peer choices and recent returns can be informative in practice. In the twin-extension wave, informative herding moves target weight by +0.377 and informative recency by +0.270, both robust across inference methods. Their fresh, mechanically generated uninformative twins are much smaller, at +0.020 and +0.064, respectively (Internet Appendix). Because these are distinct instruments from the full-panel herding and recency manipulations in Table 6, their estimates need not coincide. The broader conclusion is asymmetric. Models respond strongly when social or performance cues are represented as economically informative, showing substantial sensitivity to relevant signals. Residual responses to explicitly uninformative versions are much smaller and are not uniformly robust. The strongest description-invariance failures are therefore not explained by a general inability to distinguish signal from noise.

5.5 Robustness of the headline effects

The headline results are robust to measuring movement in the full allocation rather than only the targeted weight. Using half-L_1 distance, one half the sum of absolute changes in all five weights, the estimated movements are 0.165 for anchoring, 0.291 for reflection framing, 0.318 for herding, and 0.151 for recency. Because weights sum to one, half-L_1 has a natural economic interpretation: it measures the share of portfolio weight that must be reallocated to transform one recommendation into the other. Reflection framing, for example, reallocates roughly 29 percent of the portfolio on this measure. A fractional-logit specification on the full allocation gives the same signs for the two confirmed biases. Target-weight and full-portfolio measures therefore identify the same economically important manipulations.

Anchoring is positive across all three tested target positions: P2 (+0.097), P3 (+0.044), and P4 (+0.120), so it is not specific to one menu location. Reflection framing targets P4 by design, so the neutral-label replication in Table 5 instead tests whether the "defensive" label drives the effect. Both headline effects also survive excluding the eight models with elevated invalid-response rates.

Neither effect is driven by a particular developer family or scenario. The full leave-one-out analysis appears in the Internet Appendix. Dropping any of the 17 families leaves anchoring between +0.075 and +0.092 and reflection framing between +0.204 and +0.228; dropping any of the 12 scenarios leaves anchoring between +0.083 and +0.091 and framing between +0.219 and +0.225. No leave-one-out estimate changes sign or approaches zero.

The effects differ, however, with model capability. In the top intelligence-index quartile, anchoring attenuates to +0.033, while reflection framing remains +0.186. This matches the broader heterogeneity results in Section 8: greater capability is associated with less numerical anchoring, but not with a comparable reduction in loss-versus-preservation framing.

5.6 A named-fund validation

The experiments above use synthetic portfolios with equal stated moments to tightly control information, at the cost of a stylized choice environment. We therefore test whether the two confirmed biases survive in more realistic investment advice.

We repeat anchoring and reflection framing using five named, low-cost U.S. equity ETFs representing total market, value, momentum, minimum volatility, and small-cap segments. Expected returns and covariances come from monthly returns over 2015–2024, expense ratios from fund prospectuses, and the client profile is fixed across arms, including age, horizon, taxable-account status, and moderate risk tolerance. We use the full discovery panel rather than selecting models. Of 55 routes attempted, 53 return usable responses and 49 enter the analysis after excluding four with invalid-response rates above one half. The frozen choice set, moments, client information, models, and response rates appear in the Internet Appendix.

Variant S supplies each fund's estimated return and volatility. Variant O additionally supplies the full mean vector, covariance matrix, risk-aversion coefficient, and mean-variance objective, making the same mean-variance solution applicable in both arms.

Both confirmed biases survive, though at smaller magnitudes than in the synthetic setting (Table 7). Under Variant S, anchoring moves the target-fund weight by +0.032 under naturalistic wording and reflection framing moves the defensive-fund weight by +0.093. Every cell in Table 7 has a Holm-adjusted randomization $p \leq 0.001$ and family-clustered wild-bootstrap $p = 0.0001$. The effects are smaller than the synthetic estimates of +0.087 and +0.222. Some attenuation is unsurprising because named funds differ along economically meaningful dimensions and therefore give the model more substantive information on which to base the recommendation. Even so, framing alone shifts 9.3 percentage points under Variant S and 13.5 under Variant O.

Explicit equivalence again separates the biases. For anchoring, stating that the cue is uninformative reduces the effect from +0.032 to +0.018 under Variant S and from +0.057 to

+0.021 under Variant O, reductions of 44 and 63 percent. Reflection framing shows essentially no attenuation: +0.093 becomes +0.098 under Variant S and +0.135 becomes +0.134 under Variant O. This reproduces the asymmetry from the synthetic experiments and the debiasing exercise in Section 7: explicit instruction partially remedies arbitrary numerical anchoring but not sensitivity to how identical outcomes are described. The recurring pattern suggests that numerical anchoring and semantic framing enter the advice process in systematically different ways.

Variant O is especially stringent. With the sufficient economic inputs for the stated portfolio problem supplied directly, the model need not infer returns, risk, or preferences from framing language. The decision problem is therefore fully specified and unchanged across arms. Yet the allocation remains description-dependent, with Holm-adjusted randomization $p = 0.0005$ in both regimes. The implication is stronger than a generic framing effect because the model has no need to infer the sufficient economic inputs from the framing language.

The framing-induced tilt happens to improve measured welfare in this choice set, but this does not make description dependence benign. The five ETFs have an average pairwise return correlation of 0.85 and different estimated expected returns, so a wording-induced reallocation changes both mean and variance, unlike the synthetic experiment where the design isolates uncompensated portfolio movement. Under Variant S, the preservation tilt lowers variance enough to improve certainty equivalent by about 25 basis points per year but lowers expected return enough to reduce it by about 17 basis points. The net change is +7.1 basis points using frozen sample means and +9.1 using Bayes-Stein shrunk means (Jorion, 1986); under Variant O it reaches +44.1 basis points. Anchoring has the opposite sign, with changes between −0.1 and −2.3 basis points.

The positive framing welfare effect is therefore a property of this opportunity set, not evidence of desirable reasoning. The same invariance failure produces losses in the synthetic setting and happens to improve welfare here; nothing in the response indicates that the model distinguishes these cases on economic grounds. The sign instead depends on the return and

covariance structure of the assets toward which the wording happens to tilt the portfolio. In another choice set, the same sensitivity could move advice away from the investor's objective. The named-fund exercise therefore extends the main result beyond a stylized menu: with realistic fund names, estimated moments, explicit client characteristics, and even a fully specified optimization problem, economically identical information still produces materially different recommendations solely because it is described differently.

Table 7. Anchoring and Reflection Framing on a Menu of Named ETFs

Effects are model-averaged changes in the target-fund weight on the named-fund choice set, with moments estimated from monthly returns over 2015–2024, across 49 models. Of 55 routes attempted, two returned no parseable response and four were dropped for invalid-response rates above one half. $M = 48$ rather than 49 in Variant O because one model never returns both arms of the same Variant O scenario. Variant S states each fund's return and volatility; Variant O additionally states the full mean vector, covariance matrix, risk-aversion coefficient, and mean-variance objective. The naturalistic regime uses plain advice wording, while the rigorous regime also states explicitly that the options are decision-equivalent and that the cue is uninformative. Inference is based on randomization rather than model-level standard errors: RI p is obtained by permuting the cue within model-by-scenario blocks. CE change is the signed annual certainty-equivalent change generated by the cue-induced tilt at $\gamma = 5$, calculated within each model-scenario cell and then averaged across models, using both frozen means and Bayes-Stein shrunk means. Positive values indicate that the cue-induced allocation increases the certainty equivalent. The Internet Appendix reports the design, estimated moments, and model-level response rates.

Bias	Variant	Regime	Effect	RI p	CE change (bp)	CE change, shrunk (bp)
Anchoring	S (stated)	nat	+0.032	0.001	−0.1	+0.3
Anchoring	S (stated)	rig	+0.018	0.001	−0.2	+0.2
Anchoring	O (opt.)	nat	+0.057	0.001	−1.4	−1.3
Anchoring	O (opt.)	rig	+0.021	0.001	−2.3	−2.0
Reflection framing	S (stated)	nat	+0.093	0.001	+7.1	+9.1
Reflection framing	S (stated)	rig	+0.098	0.001	+22.1	+23.7
Reflection framing	O (opt.)	nat	+0.135	0.001	+44.1	+45.3
Reflection framing	O (opt.)	rig	+0.134	0.001	+50.0	+50.9

6 Economic Significance for Delegated Advice

6.1 From cue effects to investing styles

Changes in recommended weights also change the economic character of the portfolio. Each option in the synthetic choice set has fixed stylized factor loadings (Section 4), allowing us to map allocation responses into exposures to the market, value, momentum, and a common

quality/low-volatility dimension. We ask how wording changes the investment style of the recommended portfolio.

Table 8 reports these style footprints. The largest again comes from reflection framing. Moving from the loss description to the economically equivalent preservation description makes the portfolio more defensive: its low-volatility loading rises by +0.104 and its shared quality/low-beta loading by +0.062, while market, value, and momentum loadings decline and the small residual size loading rises by +0.010. The loss frame produces the mirror image, with greater exposure to the higher-risk dimensions of the choice set.

Framing therefore does more than redistribute weight across anonymous options. A 0.10 increase in low-volatility exposure represents a substantial change in the client's style mix even though the investment opportunities are unchanged. Other manipulations produce smaller but interpretable footprints. Recency raises momentum exposure by +0.054, herding produces larger quality and low-volatility tilts, and anchoring produces a more modest defensive tilt. Manipulations with bounded or near-null allocation effects also have little style impact.

This mapping matters for delegated advice. If LLM recommendations feed downstream screening, ranking, model-portfolio construction, or allocation workflows, description dependence can propagate into systematic investment exposures. Economically irrelevant language can therefore change the factor characteristics of the portfolio ultimately delivered to the investor, not just an individual recommendation.

6.2 The certainty-equivalent cost of frame dependence

We next translate these allocation changes into investor welfare, first characterizing the portfolios from which treatment effects begin. Equal weighting is not an assumed optimum, but provides a useful benchmark for concentration.

In the control arms, the average recommendation allocates 0.37 to the broad-market

Table 8. How Each Wording Change Shifts Portfolio Style Exposures

The table reports the model-averaged effect of each cue on choice-set factor exposures using the stylized factor loadings assigned to the experimental options. Model-level standard errors, calculated across models following Section 4, appear in parentheses. Quality / low beta reports the common loading assigned to the choice set's quality and low-beta dimensions. Figure 1 reports the frozen choice set and associated loadings.

Manipulation	Market	Size	Value	Momentum	Quality / low beta	Low vol
Anchoring	-0.009 (0.001)	-0.001 (0.001)	+0.009 (0.002)	+0.001 (0.002)	+0.013 (0.002)	+0.022 (0.003)
Framing (reflection)	-0.042 (0.004)	+0.010 (0.004)	-0.030 (0.004)	-0.015 (0.004)	+0.062 (0.005)	+0.104 (0.009)
Herding	-0.024 (0.002)	-0.039 (0.003)	+0.023 (0.002)	-0.025 (0.004)	+0.041 (0.003)	+0.064 (0.005)
Recency	+0.009 (0.001)	-0.002 (0.001)	-0.009 (0.002)	+0.054 (0.004)	-0.013 (0.002)	-0.022 (0.003)
Framing (global)	-0.004 (0.001)	-0.001 (0.001)	+0.000 (0.001)	-0.003 (0.001)	+0.007 (0.002)	+0.011 (0.003)
Default	+0.004 (0.001)	-0.000 (0.001)	-0.002 (0.002)	+0.005 (0.002)	-0.006 (0.002)	-0.009 (0.003)
Denomination	+0.002 (0.001)	+0.002 (0.001)	-0.002 (0.001)	+0.001 (0.001)	-0.004 (0.002)	-0.006 (0.003)
Availability	+0.005 (0.001)	+0.009 (0.001)	-0.002 (0.002)	+0.017 (0.002)	-0.009 (0.002)	-0.013 (0.003)
Downside	+0.003 (0.001)	+0.002 (0.001)	+0.002 (0.002)	+0.002 (0.002)	-0.004 (0.002)	-0.007 (0.002)

option, 0.20 to value, 0.11 to momentum, 0.23 to the defensive quality/low-volatility option, and 0.08 to the small high-risk option, with rounding causing the weights not to sum exactly to one. Baseline advice is therefore already concentrated, particularly in the broad-market and defensive options.

How each manipulation changes distance from equal weight is reported in the Internet Appendix. Herding produces the only large movement away from this benchmark, increasing concentration by 0.254. Reflection framing, recency, availability, default, and downside move portfolios modestly closer to equal weight, while anchoring, global framing, and denomination have little effect.

To value the cue-induced tilt, we use the moments supplied in the prompts rather than realized factor returns. As in Section 3, every synthetic option has the same expected return and variance. A wording-induced reallocation therefore cannot earn a higher expected return under the information available to the model. Any additional variance is uncompensated, so the tilt can be treated as an active portfolio overlay with zero expected-return compensation.[3]

Table 9 reports the resulting certainty-equivalent changes. At risk aversion $\gamma = 5$, reflection framing reduces annual certainty equivalent by 23.5 basis points under the main

[3]We construct the option covariance matrix in two ways. The main calibration uses the choice set's factor-implied correlations, rescaled to the stated variance ($v = 0.045$). As a robustness exercise, we instead impose a common pairwise correlation of $\rho = 0.8$.

covariance calibration and 16.3 basis points under equal correlation. Herding produces an 8.0 basis-point loss and recency 3.3 basis points, while manipulations with small allocation effects generate correspondingly small welfare changes.

The two confirmed biases differ sharply in economic importance. Anchoring is highly systematic statistically but costs only about 1 to 2 basis points per year under this calibration. Reflection framing's certainty-equivalent penalty is roughly an order of magnitude larger. Statistical reliability therefore need not imply economic importance: a reproducible bias can produce a modest portfolio distortion, while another materially changes investor risk.

A 23.5 basis-point annual certainty-equivalent loss is also comparable to, and often larger than, expense ratios on low-cost index products. The framing effect is not literally a fee, but an economically irrelevant prompt feature can create a welfare loss on the same order as costs investors and advisers routinely seek to minimize. This loss is not driven by greater concentration: baseline portfolios are already concentrated, and reflection framing moves them slightly closer to equal weight. It instead arises because equivalent descriptions produce different portfolios and, with equal expected returns, different risk without compensation.

The final column of Table 9 shows why realized returns provide a misleading welfare benchmark. Applying historical factor premia to the same defensive framing tilt implies a 123.6 basis-point annual return increase. Ex post, the frame-dependent portfolio therefore appears profitable. But those premia were not supplied to the adviser and cannot rationalize why preservation wording changed the recommendation. A decision rule can violate description invariance and still perform well ex post when an arbitrary tilt happens to load on a positive realized premium. Our welfare criterion instead evaluates the allocation using the economic information available to the model.

Table 9. The Annual Certainty-Equivalent Cost of Wording-Induced Portfolio Tilts

The CE columns price the cue-induced allocation tilt under the equal expected returns and equal option variance stated in the prompts. The main specification uses factor-implied option correlations rescaled to $v = 0.045$ unless noted otherwise. Entries are annual basis points, with negative values denoting certainty-equivalent losses. The historical-premia column applies realized factor premia to the same tilt as an ex-post benchmark rather than the welfare criterion. The $\gamma = 5$ column reports 2,000 model-bootstrap 95 percent confidence intervals in brackets. Across the variance grid $v \in \{0.030, 0.045, 0.060\}$, the absolute $\gamma = 5$ certainty-equivalent penalty ranges from 0.9 to 1.8 basis points for anchoring and from 15.7 to 31.4 basis points for reflection framing.

Manipulation	$\Delta CE_{\gamma=5}$	$\Delta CE_{\gamma=2}$	$\Delta CE_{\gamma=10}$	ΔCE, $\rho = 0.8$	Hist.-premia return
Anchoring	-1.4 [-2.2, -0.8]	-0.5	-2.7	-1.4	+34.6
Framing (reflection)	-23.5 [-32.4, -16.3]	-9.4	-47.1	-16.3	+123.6
Herding	-8.0 [-10.6, -5.9]	-3.2	-16.1	-3.9	+74.7
Recency	-3.3 [-4.5, -2.4]	-1.3	-6.7	-3.7	+4.7
Framing (global)	-0.2 [-0.4, -0.1]	-0.1	-0.3	-0.1	+12.7
Default	-0.2 [-0.5, -0.0]	-0.1	-0.4	-0.1	-10.4
Denomination	-0.1 [-0.3, -0.0]	-0.0	-0.2	-0.0	-8.3
Availability	-0.4 [-0.7, -0.2]	-0.2	-0.8	-0.4	-5.7
Downside	-0.1 [-0.3, -0.0]	-0.0	-0.2	-0.0	-7.1

6.3 Magnitudes relative to human benchmarks

A final perspective expresses the two confirmed biases in units familiar from the human decision-making literature. Recoding each allocation draw as a sure-versus-risky choice, models place more weight on the sure option in 89.3 percent of draws under preservation wording versus 50.4 percent under economically equivalent loss wording, a 39.0 percentage-point difference. For comparison, the classic risky-choice framing problem produces a reversal of roughly 50 percentage points in human subjects (Tversky and Kahneman, 1981), while meta-analytic evidence across the broader framing literature finds smaller average effects (Kühberger, 1998). For anchoring, dividing the allocation change by the gap between low and high anchors yields an LLM anchoring index of 0.15, compared with approximately 0.50 in the classic human numerical-estimation setting (Jacowitz and Kahneman, 1995).

These comparisons are necessarily approximate and should not be interpreted as showing whether LLMs are more or less susceptible than humans because the instruments, outcomes, and environments differ. The recoding and the comparability caveats are set out in the Internet Appendix. Our primary measure remains the continuous portfolio-weight effect of +0.222, which is directly tied to the financial decision studied here.

Both confirmed effects violate description invariance but differ sharply in magnitude and economic consequence (Table 10). Anchoring produces a smaller portfolio distortion and modest certainty-equivalent cost, while reflection framing changes portfolio weights, investment style, and risk exposure at a scale comparable to canonical human framing effects. For delegated financial advice, the more economically important failure is therefore not that irrelevant information affects recommendations, but that alternative descriptions of identical economic outcomes can materially change the portfolio an investor is advised to hold.

Table 10. LLM Effects Compared with Classic Human Framing and Anchoring Results

The safe-choice share is the fraction of draws in which the model places more weight on the sure option than on the gamble. The anchoring index divides the allocation effect by the difference between high and low anchors. The human framing benchmark is the approximately 50 percentage-point reversal in the classic risky-choice problem of Tversky and Kahneman (1981); the human anchoring benchmark is the approximately 0.50 average anchoring index in the estimation task of Jacowitz and Kahneman (1995). Because the instruments and decision tasks differ, these comparisons establish orders of magnitude rather than direct estimates of relative susceptibility between humans and LLMs.

Manipulation	Measure	LLM panel	Human benchmark
Loss–preservation framing	Safe-choice share, preservation vs. loss framing	89.3% vs. 50.4% (+39.0 pp)	≈50 pp
Suggested-allocation anchoring	Anchoring index (effect / anchor gap)	0.15	≈0.50

7 Debiasing and Supervising AI Advice

7.1 A ladder of prompt interventions

The preceding results show that economically irrelevant prompt features can materially change LLM portfolio advice. We next ask whether simple interventions can restore de-

scription invariance. We evaluate four interventions for the two confirmed biases, suggested-allocation anchoring and reflection framing, using the full 53-model panel (Figure 3).[4]

The interventions form a natural ladder. The first three leave the cue in context but increasingly direct the model to disregard it. The fourth intervenes earlier by neutralizing the prompt before the allocating model sees it. This distinction asks whether description dependence can be corrected by changing how the model reasons about an observed cue or requires changing the representation itself.

The first intervention is a normative system instruction directing the model to use only expected return, variance, horizon, and mandate. The second uses a consider-the-opposite instruction that identifies presentation effects as a potential bias and asks the model to disregard them. The third imposes cognitive forcing: before allocating, the model must restate each option's expected return and variance and recognize that these quantities are equal across the relevant options. The fourth uses prompt preprocessing, where a separate model neutralizes the manipulated description before allocation.

For preprocessing, we generate one common set of neutralized prompts for all allocating models rather than letting each neutralize its own input. Model-specific rewriting would combine the ability to remove the cue with susceptibility to whatever remains. Fixing the rewritten text isolates allocating-model responses to a common intervention and makes the inputs auditable. The neutralizing model sees only one arm at a time, so it is not mechanically instructed to make paired prompts similar. Before collection, we verify that all numerical inputs, options, and response requirements are preserved. The Internet Appendix reports the procedure, validation tests, and textual similarity across neutralized arms.

Each rung is compared with a baseline collected in the same wave, so the reference effects are +0.102 for anchoring and +0.232 for reflection framing rather than the discovery estimates in Section 5. Because the waves use different panels and scenario banks, percentage reductions are measured relative to the within-wave baseline.

[4]The panel consists of the 54 models in Section 4, less one checkpoint retired by its provider before this wave.

The ladder reveals a sharp difference between the two biases. Anchoring responds strongly to increasingly explicit reasoning instructions. The normative system message reduces it by 12 percent, consider-the-opposite by 63 percent, and forced computation eliminates it. Directing attention toward economically relevant quantities can therefore largely restore invariance for anchoring.

Reflection framing behaves differently. Consider-the-opposite reduces the effect by only 24 percent, forced computation by 4 percent, and the normative system message increases it by 7 percent. Forced computation is especially informative: the model restates the equal expected returns and variances, explicitly recognizes that the relevant economic quantities are unchanged, and still recommends different allocations under preservation and loss wording. Recognizing economic equivalence therefore does not make the recommendation invariant. The failure appears downstream from recognizing the relevant numerical information.

Prompt preprocessing produces a different result. Neutralizing the framing language reduces the effect from +0.232 to +0.007, a 97 percent decline, compared with at most 24 percent for the instruction-based interventions. For framing, changing the representation is therefore substantially more effective than instructing the model to disregard it.

The preprocessing result has different interpretations for anchoring and framing. Anchoring enters through a discrete sentence containing the suggested allocation. Removing that sentence makes the neutralized arms textually identical, so the near-zero post-processing effect is mechanical rather than evidence that the allocating model has become less susceptible.[5]

Reflection framing cannot be removed by deleting one sentence because it is embedded in the outcome description. The preprocessing model instead rewrites preservation and loss wording into a symmetric representation while preserving outcomes, probabilities, options, and numerical information. The resulting prompts remain distinct, with mean pairwise similarity of 0.986 versus 1.000 for preprocessed anchoring. The decline from +0.232 to +0.007

[5]Sixteen of the 53 models produce an anchoring effect of exactly zero after preprocessing. Remaining differences reflect provider output variation even at temperature zero.

is therefore substantive: the economic content remains, but the asymmetric representation is removed.

The broader lesson is that mitigation depends on the cue's form. A discrete numerical suggestion can be removed and substantially attenuated through cognitive forcing. A framing cue embedded in semantic representation is harder for the allocating model to disregard even when explicitly instructed. In our setting, mitigating that form of description dependence requires intervening on the input itself.

Figure 4 provides additional evidence on where framing is most consequential. Across models, the framing effect is smaller among those already placing greater baseline weight on the defensive option and larger when baseline portfolios are farther from equal weight. The first relation is inconsistent with preservation wording simply amplifying an existing defensive preference; the second suggests greater scope for framing to move allocations when baseline recommendations are farther from a neutral allocation.

These correlations are descriptive rather than experimentally assigned mechanisms. They nevertheless reinforce a practical implication of the intervention results: susceptibility is not summarized by a single prompting rule, so mitigation should be tested against the particular forms of description dependence relevant to a given workflow.

7.2 The deployment value of debiasing

The value of these interventions depends on the cost of the portfolio distortions they remove, not on measured bias alone. Applying the certainty-equivalent framework of Section 6, the baseline anchoring penalty of 1.9 basis points per year is eliminated by both forced computation and preprocessing. Reflection framing's baseline penalty of 25.7 basis points, more than an order of magnitude larger, is barely reduced by instruction-based interventions and almost fully recovered by preprocessing. The economic ranking therefore reinforces the behavioral results: an intervention may perform well against a statistically reliable bias

Figure 3. How Much Each Debiasing Intervention Reduces the Effect

The figure reports model-averaged cue effects on target-option weight at each intervention rung for the full 53-model panel, with model-level 95% intervals. The first three interventions leave the cue in the prompt and increasingly direct the model to disregard it; preprocessing neutralizes the cue before allocation. Cognitive forcing eliminates suggested-allocation anchoring but leaves reflection framing nearly unchanged, whereas preprocessing nearly eliminates both. For anchoring, the preprocessed arms are textually identical, so the near-zero effect is mechanical. For framing, the neutralized arms remain distinct texts and the reduction is substantive (Section 7).

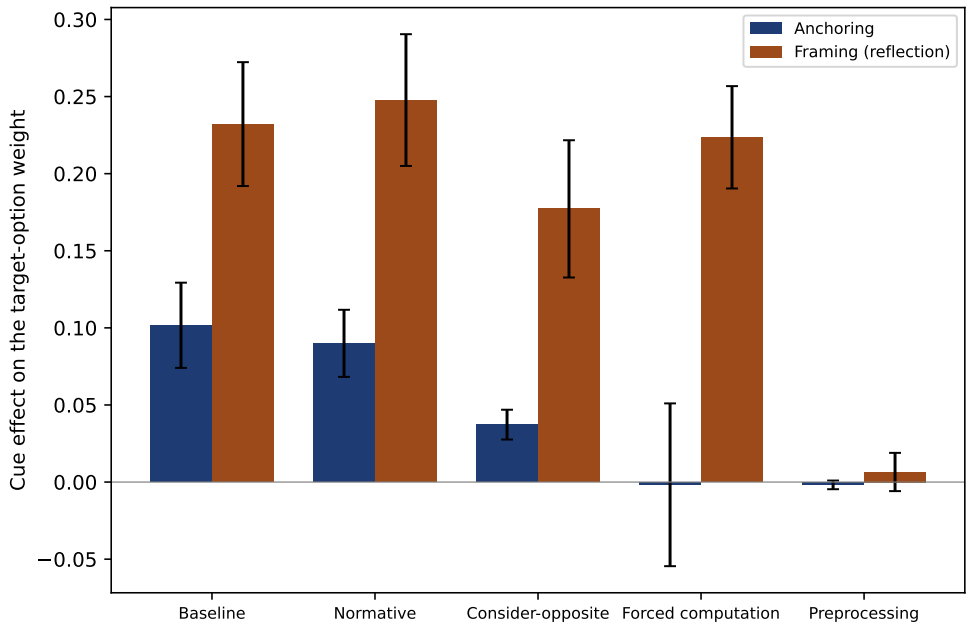
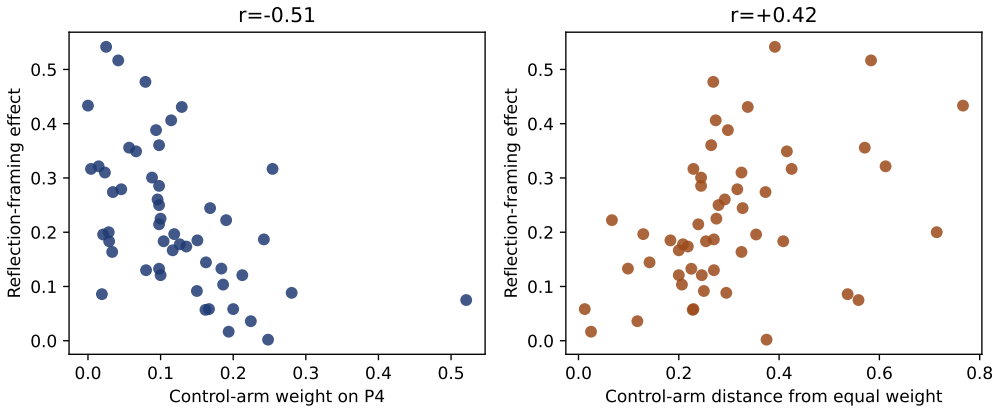


Figure 4. Which Models Show the Largest Reflection-Framing Effect

The left panel relates each model's control-arm weight on the defensive low-volatility option, P4, to its reflection-framing effect. The right panel relates control-arm distance from equal weight to the framing effect. The negative left-panel relation is inconsistent with framing simply amplifying an existing defensive allocation, while the positive right-panel relation indicates larger framing effects for models whose baseline allocations are farther from equal weight. These correlations are descriptive rather than causal.



while doing little against the distortion most important for investor welfare. The Internet Appendix reports the full ladder, historical-premia columns, and the ex-post caveat discussed in Section 6.

7.3 Supervisory implications

The debiasing results shift attention from model-level safeguards toward workflow-level validation. General capability remains relevant for model selection but is insufficient: an advice system should also be tested for whether recommendations remain stable when economically irrelevant aspects of information representation change.

The relevant object of supervision is therefore the full advice pipeline: prompt templates, retrieved context, model selection, preprocessing, guardrails, optimization layers, and the final recommendation. Our experiments isolate behavior primarily at the prompt and model stages. Whether human review, compliance procedures, or formal portfolio optimizers attenuate or amplify these distortions lies outside the design and remains an important direction for future work. The implications here are therefore lessons for workflow design rather than measured properties of deployed systems.

The paired-prompt methodology provides one validation approach. Before deployment, and again after changes to the model, prompts, retrieval system, or preprocessing pipeline, hold the investor's economic problem fixed, vary irrelevant descriptions, and test whether the workflow's recommendation changes. The relevant outcome is the stability of the system's recommendation, not the identity of the underlying model or its score on a general capability benchmark.

Such validation complements rather than replaces fiduciary, suitability, and best-interest review. Supervisory attention to automated investment advice is longstanding (U.S. Securities and Exchange Commission, 2017), and enforcement has begun to address how firms characterize their use of artificial intelligence (U.S. Securities and Exchange Commission, 2024). The risk studied here is different: even a system accurately described and oper-

ating as designed can remain sensitive to economically irrelevant differences in how client information or investment outcomes are expressed.

Several controls follow naturally. Prompt templates can be stress-tested with economically equivalent paraphrases; model selection can incorporate measured description dependence alongside general capability; mitigation instructions and preprocessing can be logged; and stability tests can be repeated after provider or model updates.

Preprocessing has an additional supervisory advantage. Unlike instructing a model to "ignore" a feature, a neutralized prompt creates an observable artifact that can be stored, audited, and compared with the original. A firm can inspect what changed rather than relying solely on model compliance.

That advantage creates a corresponding risk. A preprocessing system that removes a relevant expected return, alters a probability, drops an option, or otherwise changes the economic problem could reduce measured description dependence while degrading the advice. In our experiment, every neutralized prompt was validated before use to ensure preservation of invariant numerical information, investment options, and response format (Section 4). A deployed preprocessing layer would require the same validation as an ongoing control.

The central supervisory implication is therefore not that every system should adopt one debiasing prompt, but that description invariance is measurable. The appropriate control depends on where instability enters the workflow. For some cues, more explicit reasoning may suffice. For others, including reflection framing, stable advice may require controlling the representation presented to the model before the decision is made.

8 Heterogeneity Across Models

8.1 Scale versus reasoning

The average effects above mask substantial heterogeneity across models. We next ask whether observable model characteristics are associated with susceptibility to description

dependence. As noted in Section 4, these regressions are descriptive rather than causal because intelligence, size, reasoning architecture, and developer are jointly determined rather than experimentally assigned.

Table 11 reports model-level regressions for the two confirmed biases. For suggested-allocation anchoring, a one-standard-deviation increase in the intelligence index is associated with a 0.023 smaller effect (SE 0.009), while reasoning models exhibit 0.036 less anchoring (SE 0.016). Open weights contribute little once other characteristics are included. Larger models also exhibit less anchoring among those with disclosed parameter counts, but the multivariate estimate uses only 28 models because many proprietary models lack parameter counts and the specification also requires a non-missing intelligence index. We therefore treat the size result as suggestive.

Reflection framing shows a different pattern. Reasoning models exhibit *more* sensitivity to loss-versus-preservation wording, with a coefficient of +0.164 (SE 0.049). The intelligence coefficient is negative, but intelligence and reasoning capture different model features and are correlated with other characteristics. Neither should be viewed as a general remedy for behavioral instability. The same characteristic associated with greater invariance to an arbitrary numerical suggestion can therefore coincide with greater sensitivity to linguistic representation. Figure 5 illustrates this asymmetry.

Developer family also explains part of the variation. With developer fixed effects, the negative reasoning coefficient for anchoring becomes imprecise, while the positive framing coefficient remains positive but less precise. Leave-one-family-out estimates preserve both signs. The appropriate interpretation is not that reasoning architecture causally lowers anchoring and increases framing, but that reasoning-labeled models in this panel exhibit that pattern, partly reflecting broader developer differences. A model characteristic that appears protective against one failure need not protect against another. Susceptibility therefore requires bias-specific tests rather than a single robustness characteristic or benchmark.

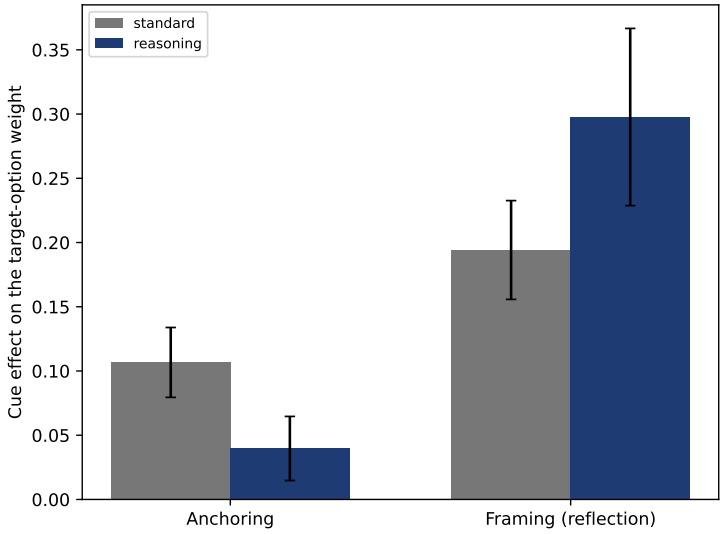
Table 11. Model Characteristics That Predict Anchoring and Reflection Framing

The table reports model-level regressions of each model's cue effect on standardized intelligence, a reasoning-model indicator, an open-weights indicator, and log parameter count. Standard errors clustered by developer family appear in parentheses. The intelligence coefficient is the association with a one-standard-deviation increase in the index; reasoning and open-weights coefficients are level differences, so magnitudes are not directly comparable. Associations are descriptive rather than causal. Specifications using disclosed parameter counts are restricted to models with available parameter data and include log parameters; log parameters are omitted (–) elsewhere. Every model disclosing a parameter count is open-weight, so the open-weights indicator has no variation and is omitted (–) in that subsample. Family-FE specifications include developer fixed effects.

Manipulation	Spec.	Intelligence	Reasoning	Open weights	Log params	N
Anchoring	Main	-0.023 (0.009)	-0.036 (0.016)	-0.001 (0.020)	–	50
Anchoring	Params disclosed	-0.013 (0.014)	-0.044 (0.029)	–	-0.026 (0.011)	28
Anchoring	Family FE	-0.014 (0.029)	-0.039 (0.045)	+0.013 (0.052)	–	50
Framing (reflection)	Main	-0.057 (0.020)	+0.164 (0.049)	-0.002 (0.025)	–	48
Framing (reflection)	Params disclosed	-0.009 (0.031)	+0.126 (0.080)	–	-0.026 (0.030)	27
Framing (reflection)	Family FE	-0.071 (0.031)	+0.146 (0.084)	-0.010 (0.080)	–	48

Figure 5. Anchoring and Reflection Framing in Reasoning and Standard Models

The figure compares reasoning-labeled and standard models for suggested-allocation anchoring and reflection framing. The plotted quantity is the model-averaged cue effect on target-option weight, with 95% intervals. Comparisons are descriptive and do not isolate the causal effect of reasoning architecture from other model and developer differences.



8.2 Bias over the capability frontier

We next ask whether description dependence disappears as models become more capable. We merge each model's Artificial Analysis intelligence index with its release date, reported

in the Internet Appendix, and examine susceptibility along the capability frontier.

Anchoring, recency, denomination, and downside decline with measured intelligence. Reflection framing does not: its relation with intelligence is small and imprecise, and reasoning-labeled models exhibit more framing than standard models. These remain cross-model correlations because capability, release date, architecture, training choices, and developer identity move together. Even so, the pattern shows that some forms of description dependence attenuate with capability while the large framing effect persists among highly capable models. Progress on general benchmarks therefore need not produce monotonic progress toward description-invariant advice.

Figure 6 makes the model-selection problem visible. Composite susceptibility does not decline smoothly with intelligence, and models with similar capability can differ substantially. Several standard models, including Claude Sonnet 4, Mistral Medium 3.5, and Claude Opus 4, combine relatively high capability with comparatively low susceptibility. Reasoning models do not dominate the frontier: three of the five models with the highest composite susceptibility are reasoning-labeled. The implication is not that standard models are generally preferable, but that capability rankings and description invariance capture different dimensions of performance. Selecting a model on a capability benchmark alone can therefore change which behavioral vulnerabilities enter the advice process without making those vulnerabilities visible to the user.

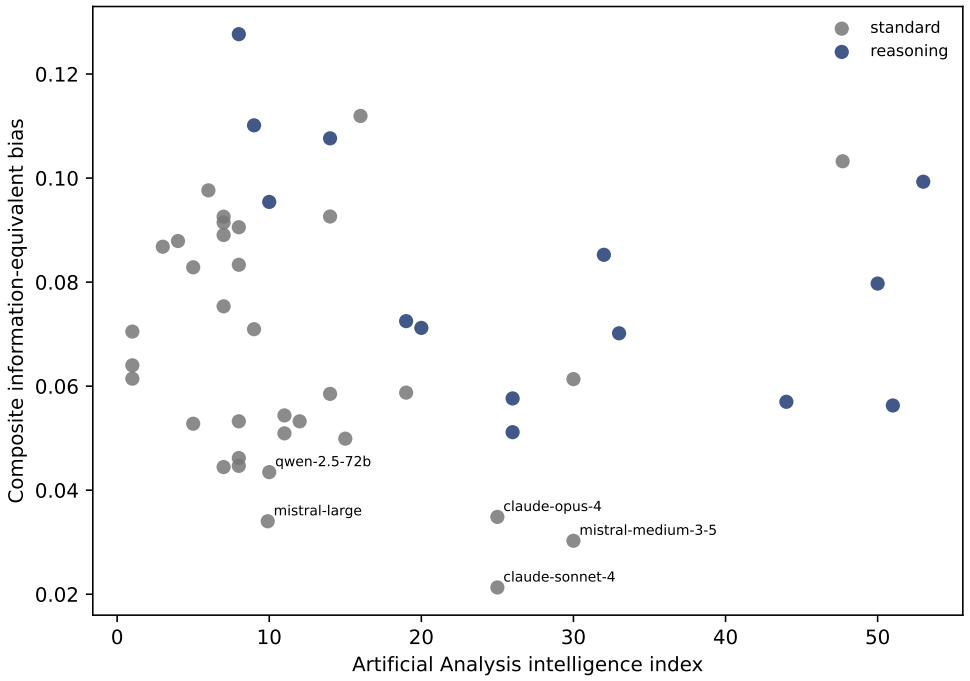
Price provides another selection criterion. If more expensive models produced more stable advice, firms could use price as a rough proxy for behavioral quality. Table 12 provides little evidence of a monotonic relationship. Moving from the cheapest to the third price quartile raises average price by roughly nineteenfold and average measured intelligence by roughly fourfold, yet composite susceptibility is essentially unchanged: the difference is 0.3 percent ($t = 0.03$). Only the most expensive quartile has materially lower composite susceptibility, about one third below the remaining models ($t = -3.48$).

Across models, price and composite susceptibility have a negative rank correlation (Spear-

man -0.36, $p = 0.010$), but the quartile pattern shows that this is driven largely by the top quartile rather than a steady decline with price. Even that decline does not align cleanly with capability: the most expensive quartile has a lower average intelligence index than the third quartile despite substantially less composite bias. Higher price therefore appears to purchase measured capability well before it reliably purchases description invariance.

Figure 6. Overall Susceptibility Against Measured Model Capability

The figure plots composite information-equivalent susceptibility, defined as the mean absolute effect across the six information-equivalent manipulations, against the Artificial Analysis intelligence index. Reasoning-labeled models are marked separately, and the five least-susceptible plotted models are labeled. Because each composite effect is estimated from finitely many draws, point labels indicate locations in capability-susceptibility space rather than precise model rankings.



8.3 Is biased advice diversifiable across models?

Cross-model heterogeneity suggests a potential mitigation strategy: aggregate recommendations across models. If description-dependent errors are mostly idiosyncratic, averaging should reduce them much as imperfectly correlated assets diversify portfolio risk. If errors

Table 12. Susceptibility and Capability by Model Price Quartile

Models are sorted into quartiles of provider list price per million tokens. Composite bias is the model-level mean absolute effect across the six information-equivalent manipulations; 95 percent intervals use 5,000 model-bootstrap draws. The intelligence index is the Artificial Analysis composite measure, and the two rightmost columns report model-averaged effects for the confirmed biases. Prices are OpenRouter provider list prices retrieved after response collection and characterize each model's price tier rather than the exact price when responses were generated. The analysis is descriptive: price reflects capability, serving costs, model vintage, and provider margins and is treated as a buyer-facing model characteristic rather than an assigned treatment. $N = 51$ models have both complete composite susceptibility and a listed price.

		Price	Intelligence	Composite		Confirmed cue effect	
Price quartile	Models	($/M tok.)	index	bias	95% CI	Anchoring	Framing
Q1 (cheapest)	13	0.10	6.1	0.075	[0.064, 0.087]	+0.129	+0.250
Q2	13	0.57	15.6	0.076	[0.064, 0.087]	+0.072	+0.239
Q3	13	1.88	24.1	0.075	[0.063, 0.088]	+0.054	+0.260
Q4 (priciest)	12	6.25	21.0	0.050	[0.039, 0.061]	+0.086	+0.148

share a strong common component, averaging offers less protection.

The evidence lies between these cases. Among 52 models with complete cue-effect vectors, the first principal component explains 26 percent of cross-model variation. Susceptibility therefore has a meaningful common component but is far from one-dimensional; models do not simply load more or less on a single "bias factor."

Developer family provides the clearest residual structure. Models within the same family have a mean pairwise cue-effect correlation of 0.10, versus −0.02 across families, and developer indicators explain 45 percent of in-sample variation in composite susceptibility. Behavioral errors are therefore more similar within developers than across them.

This dependence determines the diversification value of multi-model advice. Averaging closely related models reduces idiosyncratic noise but does little to diversify shared behavioral tendencies. As in portfolio diversification, the covariance of errors matters more than the number of models. Several models from one provider can create the appearance of independent validation while preserving common behavioral exposures, whereas a smaller ensemble spanning developer families may diversify description-dependent errors more effectively.

Averaging is not shown here to eliminate either confirmed bias, and our design does not optimize an ensemble rule. It instead identifies the relevant dependence structure. The Internet Appendix reports the full cross-model correlation matrix, principal-component de-

composition, and developer-family ranking. Multi-model architectures intended to improve robustness should therefore account for dependence across model errors rather than treating each additional model as an independent source of judgment.

9 Conclusion

We ask whether large language models give the same financial advice when the same economic problem is described differently. Across 54 models and 15,715 allocation decisions, they often do not. A random, explicitly uninformative suggested allocation shifts the recommended weight on the targeted option by 8.7 percentage points. Re-expressing identical outcomes as value preserved rather than value lost shifts 22.2 percentage points toward the safer option. The effect replicates at 20.4 points on fresh scenarios, with every model moving in the same direction, and carries a certainty-equivalent cost of roughly 23.5 basis points per year under the moments supplied to the model. Yet denomination and the status-quo default produce bounded nulls, so the result is not generic prompt instability. Nor does it reflect an inability to recognize information. Attributing the same numerical suggestion to a portfolio optimizer rather than a random placeholder produces a response more than six times as large. Models can distinguish signal from noise and still respond systematically to noise.

The distinction matters for deploying AI in finance because capability and description invariance are different properties. Reasoning-labeled models anchor less but respond more strongly to reflection framing, while moving from the cheapest model quartile to the third raises price roughly nineteenfold and measured intelligence fourfold with essentially no change in overall susceptibility. More capable models are therefore not automatically more reliable along this dimension. The appropriate diagnostic is direct: hold the client's economics fixed, vary only their representation, and test the recommendation produced by the full workflow. The appropriate mitigation also depends on the source of instability. Forcing models to restate equal expected returns and variances eliminates anchoring but reduces reflection

framing by only 4 percent; neutralizing the representation before allocation removes 97 percent of the framing effect. Multi-model validation is not automatically sufficient either, because cue effects are more correlated within developer families than across them.

The same issue applies to research that uses LLMs as advisers, evaluators, experimental subjects, or measurement tools. Such applications often treat model output as a response to the economic content of the prompt. Our evidence shows that representation can itself be an economically consequential treatment. When conclusions depend on model judgments, economically equivalent paraphrases therefore provide a natural robustness test, and changes in output should not automatically be interpreted as changes in beliefs, preferences, or reasoning when the underlying information is unchanged. The magnitudes documented here will evolve as models change, but the underlying test is durable. If an AI system is trusted with delegated financial decisions, economically equivalent descriptions should produce economically equivalent advice unless the description itself conveys information. Description invariance is therefore a basic and directly measurable reliability condition for AI-assisted finance.

Data and Code Availability

The study uses responses collected through the OpenRouter API and no proprietary data. The replication package includes the prompt generators, collection and parsing code, analysis scripts, choice set and factor loadings, model characteristics, pricing and intelligence data, and raw responses needed to reproduce all tables and figures. The Internet Appendix documents the instruments, pre-specification, welfare construction, debiasing and preprocessing procedures, informative twins, attrition, and model subsamples. OpenRouter prices are current and may change with provider repricing, while commercial model outputs may not be exactly reproducible if providers update or retire model versions.

Acknowledgements and Disclosure

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The cost of the model queries reported in the paper was borne by the authors. The authors have no financial relationship with any of the model providers or inference platforms studied, hold no position in any entity whose models are evaluated, and have no other potential conflicts of interest to disclose. No provider reviewed the manuscript or had any role in the design, execution, analysis, or reporting of the study.

References

Barberis, Nicholas, 2018, Psychology-based models of asset prices and trading volume, in *Handbook of Behavioral Economics: Applications and Foundations 1* . pp. 79–175 (Elsevier).

——— , Andrei Shleifer, and Robert Vishny, 1998, A model of investor sentiment, *Journal of Financial Economics* 49, 307–343.

Benartzi, Shlomo, and Richard H. Thaler, 1995, Myopic loss aversion and the equity premium puzzle, *The Quarterly Journal of Economics* 110, 73–92.

Benjamini, Yoav, and Yosef Hochberg, 1995, Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society: Series B (Methodological)* 57, 289–300.

Bianchi, Milo, and Marie Brière, 2026, Human-robot interactions in investment decisions, *Management Science* 72, 14–31.

Bini, Pietro, Lin William Cong, Xing Huang, and Lawrence J. Jin, 2026, Behavioral economics of AI: LLM biases and corrections, Working Paper 34745 National Bureau of Economic Research.

Binz, Marcel, and Eric Schulz, 2023, Using cognitive psychology to understand GPT-3, *Proceedings of the National Academy of Sciences* 120, e2218523120.

Brown, Tom B., et al., 2020, Language models are few-shot learners, in *Advances in Neural Information Processing Systems* vol. 33 pp. 1877–1901.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller, 2008, Bootstrap-based improvements for inference with clustered errors, *Review of Economics and Statistics* 90, 414–427.

Cao, Sean, Wei Jiang, Baozhong Yang, and Alan L. Zhang, 2023, How to talk when a machine is listening: Corporate disclosure in the age of AI, *The Review of Financial Studies* 36, 3603–3642.

Chen, Yang, Samuel N. Kirshner, Anton Ovchinnikov, Meena Andiappan, and Tracy Jenkin, 2025, A manager and an AI walk into a bar: Does ChatGPT make biased decisions like we do?, *Manufacturing & Service Operations Management* 27, 354–368.

Chen, Yiting, Tracy Xiao Liu, You Shan, and Songfa Zhong, 2023, The emergence of economic rationality of GPT, *Proceedings of the National Academy of Sciences* 120, e2316205120.

Cheng, Qiang, Pengkai Lin, and Yue Zhao, 2025, Does generative AI facilitate investor trading? Early evidence from ChatGPT outages, *Journal of Accounting and Economics* 80, 101821.

Choi, James J., David Laibson, and Brigitte C. Madrian, 2010, Why does the law of one price fail? An experiment on index mutual funds, *The Review of Financial Studies* 23, 1405–1432.

D'Acunto, Francesco, Nagpurnanand Prabhala, and Alberto G. Rossi, 2019, The promises and pitfalls of robo-advising, *The Review of Financial Studies* 32, 1983–2020.

De Bondt, Werner F. M., and Richard Thaler, 1985, Does the stock market overreact?, *The Journal of Finance* 40, 793–805.

de Kok, Ties, 2025, ChatGPT for textual analysis? How to use generative LLMs in accounting research, *Management Science* 71, 7888–7906.

De Martino, Benedetto, Dharshan Kumaran, Ben Seymour, and Raymond J. Dolan, 2006, Frames, biases, and rational decision-making in the human brain, *Science* 313, 684–687.

Demszky, Dorottya, Diyi Yang, David S. Yeager, Christopher J. Bryan, Margarett Clapper, Susannah Chandhok, Johannes C. Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, Michaela Jones, Danielle Krettek-Cobb, Leslie Lai, Nirel JonesMitchell, Desmond C. Ong, Carol S. Dweck, James J. Gross, and James W. Pennebaker, 2023, Using large language models in psychology, *Nature Reviews Psychology* 2, 688–701.

Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey, 2015, Algorithm aversion: People erroneously avoid algorithms after seeing them err, *Journal of Experimental Psychology: General* 144, 114–126.

Egan, Mark, Gregor Matvos, and Amit Seru, 2019, The market for financial adviser misconduct, *Journal of Political Economy* 127, 233–295.

Fama, Eugene F., and Kenneth R. French, 1993, Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* 33, 3–56.

Foerster, Stephen, Juhani T. Linnainmaa, Brian T. Melzer, and Alessandro Previtero, 2017, Retail financial advice: Does one size fit all?, *The Journal of Finance* 72, 1441–1482.

Gennaioli, Nicola, Andrei Shleifer, and Robert Vishny, 2015, Money doctors, *The Journal of Finance* 70, 91–114.

Goli, Ali, and Amandeep Singh, 2024, Frontiers: Can large language models capture human preferences?, *Marketing Science* 43, 709–722.

Hagendorff, Thilo, Sarah Fabi, and Michal Kosinski, 2023, Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT, *Nature Computational Science* 3, 833–838.

Hirshleifer, David, 2001, Investor psychology and asset pricing, *The Journal of Finance* 56, 1533–1597.

Holm, Sture, 1979, A simple sequentially rejective multiple test procedure, *Scandinavian Journal of Statistics* 6, 65–70.

Horton, John J., Apostolos Filippas, and Benjamin S. Manning, 2023, Large language models as simulated economic agents: What can we learn from Homo Silicus?, Working Paper 31122 National Bureau of Economic Research Revised February 2026.

Huang, Yiming, Biquan Bie, Zuqiu Na, Weilin Ruan, Songxin Lei, Yutao Yue, and Xinlei He, 2026, Understanding the anchoring effect of LLM with synthetic data: Existence, mechanism, and potential mitigations, in *HCAIR Workshop at the Fourteenth International Conference on Learning Representations*. Also available as arXiv:2505.15392.

Hütter, Mandy, and Klaus Fiedler, 2019, Advice taking under uncertainty: The impact of genuine advice versus arbitrary anchors on judgment, *Journal of Experimental Social Psychology* 85, 103829.

Jacowitz, Karen E., and Daniel Kahneman, 1995, Measures of anchoring in estimation tasks, *Personality and Social Psychology Bulletin* 21, 1161–1166.

Jegadeesh, Narasimhan, and Sheridan Titman, 1993, Returns to buying winners and selling losers: Implications for stock market efficiency, *The Journal of Finance* 48, 65–91.

Jones, Erik, and Jacob Steinhardt, 2022, Capturing failures of large language models via human cognitive biases, in *Advances in Neural Information Processing Systems* vol. 35 pp. 11785–11799.

Jorion, Philippe, 1986, Bayes-Stein estimation for portfolio analysis, *Journal of Financial and Quantitative Analysis* 21, 279–292.

Kahneman, Daniel, and Amos Tversky, 1979, Prospect theory: An analysis of decision under risk, *Econometrica* 47, 263–291.

Kim, Alex G., Maximilian Muhn, and Valeri V. Nikolaev, 2024, Financial statement analysis with large language models, Working Paper 2024-65 Becker Friedman Institute for Economics at the University of Chicago Temporarily withdrawn by the authors because of data and analysis inconsistencies; arXiv:2407.17866.

Korinek, Anton, 2023, Generative AI for economic research: Use cases and implications for economists, *Journal of Economic Literature* 61, 1281–1317.

Kronlund, Mathias, Veronika K. Pool, Clemens Sialm, and Irina Stefanescu, 2021, Out of sight no more? The effect of fee disclosures on 401(k) investment allocations, *Journal of Financial Economics* 141, 644–668.

Kühberger, Anton, 1998, The influence of framing on risky decisions: A meta-analysis, *Organizational Behavior and Human Decision Processes* 75, 23–55.

Lee, Hoyoung, Junhyuk Seo, Suhwan Park, Junhyeong Lee, Wonbin Ahn, Chanyeol Choi, Alejandro Lopez-Lira, and Yongjae Lee, 2025, Your AI, not your view: The bias of LLMs in investment analysis, in *Proceedings of the 6th ACM International Conference on AI in Finance* pp. 150–158.

Linnainmaa, Juhani T., Brian T. Melzer, and Alessandro Previtero, 2021, The misguided beliefs of financial advisors, *The Journal of Finance* 76, 587–621.

Logg, Jennifer M., Julia A. Minson, and Don A. Moore, 2019, Algorithm appreciation: People prefer algorithmic to human judgment, *Organizational Behavior and Human Decision Processes* 151, 90–103.

Lopez-Lira, Alejandro, and Yuehua Tang, 2026, Can ChatGPT forecast stock price movements? Return predictability and large language models, *Journal of Financial Economics* 184, 104335.

Madrian, Brigitte C., and Dennis F. Shea, 2001, The power of suggestion: Inertia in 401(k) participation and savings behavior, *The Quarterly Journal of Economics* 116, 1149–1187.

Malberg, Simon, Roman Poletukhin, Carolin M. Schuster, and Georg Groh, 2025, A comprehensive evaluation of cognitive biases in LLMs, in *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities* pp. 578–613. Association for Computational Linguistics.

Markowitz, Harry, 1952, Portfolio selection, *The Journal of Finance* 7, 77–91.

Mei, Qiaozhu, Yutong Xie, Walter Yuan, and Matthew O. Jackson, 2024, A Turing test of whether AI chatbots are behaviorally similar to humans, *Proceedings of the National Academy of Sciences* 121, e2313925121.

Modigliani, Franco, and Merton H. Miller, 1958, The cost of capital, corporation finance and the theory of investment, *American Economic Review* 48, 261–297.

Mullainathan, Sendhil, Markus Noeth, and Antoinette Schoar, 2012, The market for financial advice: An audit study, Working Paper 17929 National Bureau of Economic Research.

Niszczota, Paweł, and Sami Abbas, 2023, GPT has become financially literate: Insights from financial literacy tests of GPT and a preliminary test of how people use it as a source of advice, *Finance Research Letters* 58, 104333 Part A.

Northcraft, Gregory B., and Margaret A. Neale, 1987, Experts, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing decisions, *Organizational Behavior and Human Decision Processes* 39, 84–97.

Odean, Terrance, 1998, Are investors reluctant to realize their losses?, *The Journal of Finance* 53, 1775–1798.

OpenAI, 2024, Morgan Stanley uses AI evals to shape the future of financial services, OpenAI customer story Published December 4, 2024; accessed July 5, 2026.

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, et al., 2022, Training language models to follow instructions with human feedback, in *Advances in Neural Information Processing Systems* vol. 35 pp. 27730–27744.

Perez, Ethan, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, et al., 2023, Discovering language model behaviors with model-written evaluations, in *Findings of the Association for Computational Linguistics: ACL 2023* pp. 13387–13434.

Reher, Michael, and Stanislav Sokolinski, 2024, Robo advisors and access to wealth management, *Journal of Financial Economics* 155, 103829.

Ross, Jillian, Yoon Kim, and Andrew W. Lo, 2024, LLM economicus? Mapping the behavioral biases of LLMs via utility theory, in *First Conference on Language Modeling*.

Rossi, Alberto G., and Stephen Utkus, 2024, The diversification and welfare effects of roboadvising, *Journal of Financial Economics* 157, 103869.

Schwab Advisor Services, 2026, 2026 RIA & AI research study: Advisor AI in action, Charles Schwab, January 22.

Sclar, Melanie, Yejin Choi, Yulia Tsvetkov, and Alane Suhr, 2024, Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting, in *The Twelfth International Conference on Learning Representations*.

Shanahan, Murray, Kyle McDonell, and Laria Reynolds, 2023, Role play with large language models, *Nature* 623, 493–498.

Sharma, Mrinank, Meg Tong, Tomasz Korbak, et al., 2024, Towards understanding sycophancy in language models, in *The Twelfth International Conference on Learning Representations*.

Shefrin, Hersh, and Meir Statman, 1985, The disposition to sell winners too early and ride losers too long: Theory and evidence, *The Journal of Finance* 40, 777–790.

Soll, Jack B., Asa B. Palley, and Christina A. Rader, 2022, The bad thing about good advice: Understanding when and how advice exacerbates overconfidence, *Management Science* 68, 2949–2969.

Steyvers, Mark, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas W. Mayer, and Padhraic Smyth, 2025, What large language models know and what people think they know, *Nature Machine Intelligence* 7, 221–231.

Sun, Jiuding, Chantal Shaib, and Byron C. Wallace, 2024, Evaluating the zero-shot robustness of instruction-tuned language models, in *The Twelfth International Conference on Learning Representations*.

Turpin, Miles, Julian Michael, Ethan Perez, and Samuel R. Bowman, 2023, Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting, in *Advances in Neural Information Processing Systems* vol. 36 pp. 74952–74965.

Tversky, Amos, and Daniel Kahneman, 1974, Judgment under uncertainty: Heuristics and biases, *Science* 185, 1124–1131.

——— , 1981, The framing of decisions and the psychology of choice, *Science* 211, 453–458.

——— , 1986, Rational choice and the framing of decisions, *The Journal of Business* 59, S251–S278.

U.S. Securities and Exchange Commission, 2017, Investor bulletin: Robo-advisers, Investor.gov Investor Bulletin Published February 23, 2017.

——— , 2024, SEC charges two investment advisers with making false and misleading statements about their use of artificial intelligence, Press Release 2024-36 Published March 18, 2024.

Winder, Philipp, Christian Hildebrand, and Jochen Hartmann, 2025, Biased echoes: Large language models reinforce investment biases and increase portfolio risks of private investors, *PLOS ONE* 20, e0325459.

Young, Alwyn, 2019, Channeling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results, *The Quarterly Journal of Economics* 134, 557–598.

Zhi, Yuhan, Xiaoyu Zhang, Longtian Wang, Shumin Jiang, Shiqing Ma, Xiaohong Guan, and Chao Shen, 2025, Exposing product bias in LLM investment recommendation, Working Paper arXiv:2503.08750 arXiv.