Calibrating Behavioral Parameters with Large Language Models *

Brandon Yee† Krishna Sharma‡

Abstract

Behavioral parameters such as loss aversion, herding, and extrapolation are central to asset pricing models but remain difficult to measure reliably. We develop a framework that treats large language models (LLMs) as calibrated measurement instruments for behavioral parameters. Using four models and 24,000 agent–scenario pairs, we document systematic rationality bias in baseline LLM behavior, including attenuated loss aversion, weak herding, and near-zero disposition effects relative to human benchmarks. Profile-based calibration induces large, stable, and theoretically coherent shifts in several parameters, with calibrated loss aversion, herding, extrapolation, and anchoring reaching or exceeding benchmark magnitudes. To assess external validity, we embed calibrated parameters in an agent-based asset pricing model, where calibrated extrapolation generates short-horizon momentum and long-horizon reversal patterns consistent with empirical evidence. Our results establish measurement ranges, calibration functions, and explicit boundaries for eight canonical behavioral biases.

Code and Raw Results Available at: `https://github.com/YCRG-Labs/agents-as-actors`

JEL Classification: C63, C90, G11, G41

Keywords: Behavioral Finance, Large Language Models, Economic Measurement, Agent-Based Models

1 Introduction

Behavioral parameters play a central role in modern asset pricing and market microstructure. Loss aversion (λ) governs asymmetric responses to gains and losses [Kahneman and Tversky, 1979], extrapolation coefficients (θ) shape momentum beliefs and return predictability [Barberis et al., 2018, Greenwood and Shleifer, 2014], and overconfidence (κ) distorts perceived signal precision [Daniel et al., 1998]. These parameters are not auxiliary modeling choices but core primitives used to explain momentum and reversals, excess volatility,

*The authors thank Damon Petersen for his valuable discussion and feedback.

†Management Sciences Lab, Yee Collins Research Group. Email: `b.yee@ycrg-labs.org`

‡Hoover Institute, Stanford University. Email: `ksharma3@stanford.edu`

trading volume, and persistent deviations from rational benchmarks. Empirical progress in behavioral finance therefore hinges on the ability to measure such parameters in a disciplined and scalable manner.

Measuring behavioral parameters remains challenging. Axiomatic models impose strong functional forms that risk obscuring key mechanisms [Barberis, 2013]. Laboratory experiments capture genuine psychological responses but face substantial measurement error, selection concerns, and limited scalability [Camerer, 2011]. Structural estimation offers external validity but suffers from severe identification problems, as observed behavior reflects a joint outcome of preferences, beliefs, constraints, and equilibrium interactions [Keane, 2011]. Survey-based measures are noisy and weakly correlated with incentivized choices—for example, self-reported risk aversion correlates only $\rho \approx 0.3$ with experimental measures [Dohmen et al., 2011]—while repeated elicitation introduces learning and fatigue effects that further contaminate identification. As a result, behavioral parameters are often treated as fixed, imprecisely estimated objects rather than experimentally manipulable quantities.

This paper proposes a different approach. We develop large language models (LLMs) as experimentally calibratable measurement instruments for behavioral parameters. Rather than asking whether LLMs replicate human behavior, we ask whether behavioral parameters can be induced, calibrated, and validated within LLMs in ways that are infeasible with human subjects. Our key idea is to treat behavioral profiles embedded in prompts as experimental treatments that shift latent parameters in predictable directions. If these induced parameters are stable, monotonic, and quantitatively comparable to human benchmarks—and if they generate economically meaningful outcomes—then LLMs can serve as a new measurement technology for behavioral finance.

Our contribution differs from and complements recent work using LLMs as economic agents. Horton [2023] and Aher et al. [2023] show that LLMs often reproduce qualitative patterns observed in laboratory experiments and classic economic games, establishing proof of concept for LLMs as simulated subjects. Related work uses LLMs to generate synthetic survey respondents whose responses correlate with human data [Argyle et al., 2023]. In contrast, our focus is not replication but measurement. We treat behavioral parameters as latent objects to be experimentally calibrated, and we evaluate LLMs against quantitative human benchmarks rather than qualitative similarity or correlation. This distinction allows us to identify systematic rationality bias in baseline LLM behavior, to map calibration functions that induce target parameter values, and to assess whether calibrated parameters generate economically meaningful outcomes in asset pricing environments.

We implement this framework across eight canonical behavioral biases using four LLM variants (GPT-4o, GPT-4o-mini, Claude-3.5-Haiku, and Gemini-2.5-Pro) and 19,200 agent–scenario pairs in the main analyses (24,000 including an excluded model with parsing issues). All experiments use fully synthetic financial scenarios guaranteed not to appear in training data. We find that baseline LLM behavior exhibits systematic rationality

bias: relative to human benchmarks, LLMs display near-zero disposition effects, somewhat attenuated herding relative to the 65-75% observed in cascade experiments, and low loss aversion (λ between 1.12 and 1.90 versus approximately 2.25). Profile-based calibration, however, induces large and theoretically consistent parameter shifts. Loss-averse profiles raise λ to 3.00, herding-prone profiles increase herding rates to 90%, extrapolative profiles raise momentum coefficients to 0.88, and loss-averse profiles increase anchoring correlations to 0.67.

To assess external validity, we embed calibrated parameters in a simple agent-based asset pricing model. Calibrated extrapolation generates short-horizon momentum and long-horizon reversal patterns closely matching the empirical stylized facts documented by Jegadeesh and Titman [1993], while baseline rational agents generate no momentum. This exercise demonstrates that calibrated parameters are not merely psychometric artifacts but carry economic content relevant for asset pricing. Our primary contribution is methodological: we provide a systematic framework for using LLMs as calibrated measurement instruments, including validation criteria, measurement ranges, and explicit boundaries on where calibration succeeds or fails. Empirical results—including the asset pricing application—serve to validate this measurement framework rather than to introduce a new asset pricing model.

The paper proceeds as follows. Section 2 situates the paper within the behavioral finance, agent-based modeling, and large language model literatures. Section 3.1 introduces behavioral parameters as measurement objects and outlines the conceptual challenges they pose. Section 3.2 develops the calibration and validation framework, and Section 4 describes the experimental design and synthetic data construction. Section 5 presents the main calibration results and measurement ranges. Section 6 evaluates external validity through an agent-based asset pricing application. Section 8 discusses implications and boundaries, and Section 9 concludes.

2 Related Work

Our work sits at the intersection of behavioral finance, agent-based modeling, and recent applications of large language models (LLMs) to economic research.

Behavioral economics documents systematic deviations from rational decision-making through loss aversion, reference dependence, and diminishing sensitivity [Kahneman and Tversky, 1979, Tversky and Kahneman, 1992], as well as heuristics such as representativeness, availability, and anchoring-and-adjustment [Tversky and Kahneman, 1974]. Applied to finance, these psychological primitives explain a range of market anomalies. Disposition effects arise from prospect theory combined with mental accounting [Shefrin and Statman, 1985] and are documented in large-scale field data [Odean, 1998]. Overconfidence distorts belief precision and generates momentum and delayed overreaction [Daniel et al., 1998, Moore and Healy, 2008]. Conservatism,

representativeness, and narrative-based reasoning contribute to underreaction and subsequent correction [Barberis et al., 1998, Lakonishok et al., 1994]. While these models treat behavioral parameters as core primitives, their empirical measurement remains challenging due to noise, identification concerns, and limited scalability.

Agent-based models (ABMs) provide an alternative framework by simulating markets populated by heterogeneous agents with bounded rationality. Seminal contributions show that mixtures of fundamentalists and chartists can generate excess volatility, fat-tailed returns, and momentum-like dynamics [Brock and Hommes, 1998, Lux and Marchesi, 1999, LeBaron, 2001]. However, most ABMs specify behavior through ad-hoc heuristics or strategy-switching rules chosen for tractability rather than empirically grounded calibration. As a result, the behavioral parameters governing agent behavior are rarely measured independently of the market outcomes they generate. Our approach complements this literature by offering a methodology for populating ABMs with agents whose behavioral parameters are calibrated directly from controlled experimental variation.

A rapidly growing literature uses large language models as simulated economic agents or synthetic respondents. Horton [2023] demonstrates that LLMs reproduce qualitative patterns in classic economic games, while Aher et al. [2023] replicate a large set of social psychology experiments with high correlation between LLM and human responses. Related work uses LLMs to generate synthetic survey populations or conduct conjoint experiments [Argyle et al., 2023, Brand et al., 2023, Park et al., 2023]. More recent studies examine financial reasoning and market simulations, highlighting trade-offs between memorization and reasoning and the role of strategic interaction [Mei et al., 2024, Horton, 2024]. This literature establishes that LLMs can mimic observed regularities, but it largely evaluates performance using qualitative similarity or correlation rather than treating behavioral parameters as quantitatively targetable objects.

Our contribution is complementary but distinct. Rather than focusing on replication, we treat behavioral parameters as latent measurement objects that can be experimentally induced and calibrated within LLMs. We evaluate calibration against quantitative human benchmarks, map achievable measurement ranges across eight canonical biases, and test whether calibrated parameters generate economically meaningful outcomes in asset-pricing environments. This framing shifts the role of LLMs from simulated subjects to calibrated measurement instruments, clarifying both their potential and their limitations for behavioral finance and economic measurement.

3 Conceptual Framework

3.1 Behavioral Parameters as Measurement Objects

Before developing calibration methodology, we formalize what it means to measure a behavioral parameter and why this is difficult. Behavioral economics modifies standard expected utility maximization through systematic departures captured by parameters. These take three forms:

Preference parameters modify the utility function. Loss aversion λ creates asymmetry: $v(x) = x^\alpha$ for gains, $v(x) = -\lambda(-x)^\beta$ for losses, with $\lambda > 1$ [Kahneman and Tversky, 1979]. Probability weighting γ distorts likelihoods: $w(p) = p^\gamma/(p^\gamma + (1-p)^\gamma)^{1/\gamma}$ [Tversky and Kahneman, 1992]. Risk aversion γ governs curvature of utility over wealth.

Belief parameters modify information processing. Overconfidence κ inflates signal precision where agents perceive variance σ^2/κ when true variance is σ^2 [Daniel et al., 1998]. Representativeness bias τ_N/τ_F overweights narrative information relative to fundamentals [Tversky and Kahneman, 1974]. Extrapolation θ biases forecasts through $\hat{r}_{t+1} = \bar{r} + \theta(r_t - \bar{r})$ [Greenwood and Shleifer, 2014].

Heuristic parameters govern decision rules. Anchoring determines adjustment from initial values: valuation = anchor + α(true value − anchor) with adjustment parameter $\alpha < 1$ [Tversky and Kahneman, 1974]. We measure anchoring through correlation ρ between anchor values and stated valuations, with $\rho > 0$ indicating insufficient adjustment. Herding weight w governs crowd influence in belief updating [Bikhchandani et al., 1992].

These parameters are latent (unobservable quantities inferred from choices), continuous (λ ranges from 1 to 3+, enabling fine-grained heterogeneity), and context-dependent (the same person may exhibit different λ for financial versus health decisions).

Measuring behavioral parameters faces four problems. First, measurement error is substantial. Self-reported risk aversion correlates only $\rho \approx 0.3$ with incentivized choices [Dohmen et al., 2011]. Survey measures of overconfidence show test-retest reliability $\rho \approx 0.5$ [Moore and Healy, 2008]. This measurement error attenuates regression coefficients, biases structural estimates, and reduces statistical power. Second, joint identification is difficult. Observed choices reflect multiple parameters simultaneously. Selling losers could reflect loss aversion (λ high), probability weighting (γ low), or narrow framing. Disentangling requires multiple elicitations and strong functional form assumptions. Third, causality is problematic. Correlating self-reported risk aversion with portfolio choices does not identify causal effects of λ on decisions. Unobserved variables confound estimates, and experimental manipulation requires randomizing psychological parameters, which is impossible with humans. Fourth, scalability is limited. Precise measurement requires large samples and multiple elicitations per subject, making human experiments with $N > 500$ and > 20 decisions per

subject prohibitively expensive.

We propose treating LLMs as experimental infrastructure: platforms where behavioral parameters can be experimentally induced, calibrated, and stress-tested. This requires three conceptual shifts. First, we treat parameters as treatments rather than measuring λ from choices with error. We experimentally induce λ through profile prompts. If profiles systematically shift behavior in theoretically predicted directions, they function as instruments generating exogenous variation. Second, calibration becomes measurement. The measurement object is not "does the LLM match humans" but "what parameter range can be induced through calibration." For loss aversion, the question becomes: what λ values can be achieved through profile manipulation, and how do these map to human benchmarks? Third, validation occurs through economic content. Calibrated parameters validate not by matching human choices in isolation but by generating known economic phenomena. If calibrated extrapolation produces momentum, calibrated loss aversion produces disposition effects, and calibrated herding produces information cascades, this confirms parameters capture economically meaningful variation.

This framework treats LLMs as measurement instruments analogous to laboratory equipment. A thermometer does not "replicate" heat but measures temperature through calibrated responses. Similarly, LLMs do not replicate human psychology but measure behavioral parameters through calibrated responses to profile manipulation.

3.2 Validation and Calibration Framework

An LLM generates decisions through a function $d = f_{\text{LLM}}(I; p, \tau, M)$ where I is the information set, p is the prompt text (including profile), τ is temperature, and M is model architecture. Our key insight treats profiles as instruments inducing behavioral parameters: $p \xrightarrow{\text{induces}} \theta \xrightarrow{\text{determines}} f(I;\theta) \xrightarrow{\text{generates}} d$. If profile $p_{\text{loss-averse}}$ generates choices consistent with $\lambda = 2.5$, while p_{rational} generates choices consistent with $\lambda = 1.0$, this validates that profiles function as instruments calibrating loss aversion.

We specify four necessary conditions for valid calibration:

Definition 3.1 (Calibration Validity)**.** *LLMs constitute valid measurement instruments for parameter θ if:*

C1. Monotonicity: *Profile strength s induces monotonic parameter changes: $\frac{\partial \hat{\theta}(s)}{\partial s} > 0$ where $\hat{\theta}(s)$ is the parameter recovered from choices under profile strength s.*

C2. Range Coverage: *The achievable parameter range $[\theta_{min}, \theta_{max}]$ through calibration includes or approaches the human benchmark θ_{human}: $\theta_{human} \in [\theta_{min} - \delta, \theta_{max} + \delta]$ for tolerance δ.*

C3. Stability: *Parameters remain stable across equivalent elicitations. For repeated measurements with same profile:* $\text{Var}[\hat{\theta}_i | p] < \tau_{threshold}$.

C4. Coherence: *Parameters recovered from different experiments maintain theoretically predicted relationships. For loss aversion measured from gambles versus disposition from portfolio choices:* $\text{Corr}(\lambda_{gamble}, \lambda_{disposition}) > 0$.

These conditions ensure calibration is systematic and controllable (C1), achieves human-relevant magnitudes (C2), reproducible (C3), and theoretically sensible (C4).

While calibration focuses on measurement ranges, validation requires comparing to human benchmarks. We define three tiers. Strong validation requires calibrated parameter within 20% of human benchmark: $|\theta_{\text{calibrated}} - \theta_{\text{human}}|/|\theta_{\text{human}}| < 0.20$. Moderate validation requires within 50%: $|\theta_{\text{calibrated}} - \theta_{\text{human}}|/|\theta_{\text{human}}| < 0.50$. Weak validation requires correct direction but magnitude differs: $\text{sgn}(\theta_{\text{calibrated}}) = \text{sgn}(\theta_{\text{human}})$ and $\theta_{\text{calibrated}} \neq 0$. For each parameter, we report: baseline value (rational profile), maximum achieved value (strongest bias profile), calibration function $\theta(s)$, validation tier, and whether human benchmark is achievable.

A fundamental challenge is distinguishing genuine reasoning from pattern matching. While synthetic scenarios cannot exist in training data, LLMs may recognize structural patterns. We address this through three strategies. First, adversarial scenario testing constructs scenarios where stereotypical pattern matching yields incorrect decisions but reasoning about specifics yields correct ones. We require $\geq 70\%$ adversarial pass rate for moderate validation. Second, structural parameter consistency tests whether parameters recovered from different experiments (e.g., λ from gambles versus disposition) maintain theoretical relationships. Pure pattern matching cannot generate theoretically consistent parameters across contexts. Third, cross-parameter predictions test whether all biases correlate positively (suggesting generic stereotypes) or show specific structures (overconfidence reduces herding because agents trust their own signals). These strategies characterize functional validity: do calibrated parameters generate patterns useful for economic research applications, regardless of underlying mechanism?

For each parameter, we select human benchmarks through systematic protocol prioritizing: sample size $N \geq 100$, clear experimental manipulation, replication in at least one independent study, and meta-analytic evidence when available. Rather than point estimates, we report ranges reflecting heterogeneity. For loss aversion, estimates range from $\lambda = 1.8$ to $\lambda = 2.7$ depending on domain and elicitation method [Tversky and Kahneman, 1992, Kahneman et al., 1990, Novemsky and Kahneman, 2005]. We use $\lambda = 2.25$ as primary benchmark with range [2.0, 2.5] for sensitivity analysis. Human benchmark values (Appendix A) are drawn from canonical experimental and field estimates of loss aversion [Tversky and Kahneman, 1992, Kahneman et al., 1990, Novemsky and Kahneman, 2005], disposition effects [Shefrin and Statman, 1985, Odean, 1998, Weber and Camerer, 1998], overconfidence [Moore and Healy, 2008], herding [Anderson and Holt, 1997, Celen and Kariv, 2004], probability weighting [Tversky and Kahneman, 1992, Barberis and Huang, 2008],

anchoring [Northcraft and Neale, 1987, Mussweiler et al., 2000], and extrapolation [Bloomfield and Hales, 2002, Greenwood and Shleifer, 2014]. Complete benchmark justification, including study-level details, sample sizes, protocols, and meta-analytic evidence, appears in Appendix A.

4 Experimental Design and Data

We test four LLM variants representing different design philosophies: OpenAI GPT-4o and GPT-4o-mini (flagship and efficient), Google Gemini 2.5 Pro (flagship with 1M context), and Anthropic Claude 3.5 Haiku (efficient). This design enables testing provider effects, scale effects, and cross-model robustness. Gemini 2.5 Flash was tested but excluded due to response parsing issues; results appear in Online Appendix D.

We test eight canonical behavioral parameters. Loss aversion (λ) measures asymmetric value function for gains versus losses through gamble acceptance (gain \$$X$ versus lose \$100, 50-50) [Kahneman and Tversky, 1979]. Disposition effect (DR) captures tendency to sell winners too quickly while holding losers, measured as $P(\text{sell}|\text{winner})/P(\text{sell}|\text{loser})$ in portfolio rebalancing [Shefrin and Statman, 1985]. Overconfidence (κ) reflects underestimating uncertainty through confidence intervals too narrow [Daniel et al., 1998]. Herding (w) measures following crowds despite contradictory private information through information cascade experiments [Bikhchandani et al., 1992]. Representativeness (τ_N/τ_F) captures overweighting narratives relative to fundamentals [Tversky and Kahneman, 1974]. Probability weighting (γ) distorts likelihoods of rare events through preference for skewed payoffs [Tversky and Kahneman, 1992]. Anchoring (ρ) measures correlation between arbitrary anchor values and stated valuations, indicating insufficient adjustment [Tversky and Kahneman, 1974]. Extrapolation (θ) predicts trend continuation despite random walk through correlation of forecasts with past returns [Greenwood and Shleifer, 2014].

Our sample design uses $N = 100$ agents per profile $\times$ 6 profiles = 600 observations per experiment $\times$ 8 experiments = 4,800 observations per model $\times$ 4 models = 19,200 main observations. Six profile types with theoretically grounded cognitive frames specify orientations without mandating specific behaviors: (1) Rational: analytical investor making decisions based on expected values and Bayesian updating, (2) Loss-averse: investor focused on capital protection finding losses particularly painful, (3) Overconfident: trusting pattern recognition abilities and believing in identifying trends others miss, (4) Herding-prone: valuing collective wisdom and treating crowd agreement as information, (5) Representativeness-biased: evaluating companies based on how well they fit patterns of successful businesses, (6) Extrapolative: believing trends tend to persist with recent performance as important indicator.

We generate fully synthetic financial environments—assets, prices, earnings, and narratives—that are statistically indistinguishable from real data yet guaranteed not to appear in training corpora. Asset identifiers

use abstract alphanumeric combinations ("Asset X427") verified absent from web searches. Prices follow geometric Brownian motion calibrated to CRSP distributions. Earnings follow autoregressive processes matching Compustat dynamics. Narratives are orthogonalized to fundamentals ensuring zero correlation. Distributional tests (Kolmogorov-Smirnov), classifier indistinguishability (random forest accuracy 52.3% vs. 50% baseline), and web search verification confirm non-contamination. Each experiment is powered (> 0.85) to detect human benchmark effect sizes. All inference accounts for multiple testing via Bonferroni-Holm correction maintaining family-wise error rate $\alpha = 0.05$. Complete design specifications, validation procedures, power analyses, and quality control protocols appear in Appendix E. We evaluate external validity by examining whether calibrated behavioral parameters reproduce well-established return continuation patterns documented in the empirical asset pricing literature, including post-earnings announcement drift [Bernard and Thomas, 1989].

5 Results: Measurement Ranges and Calibration

This section presents calibration results for eight behavioral parameters across four large language models. The central result is that baseline LLM behavior exhibits systematic rationality bias relative to human benchmarks, while profile-based prompting induces large, stable, and theoretically coherent shifts in several parameters. For four biases—loss aversion, herding, extrapolation, and anchoring—calibration reaches or exceeds human benchmark magnitudes, satisfying the validation criteria developed in Section 3.2.

Table 1 summarizes the primary measurement outcome: the range of parameter values achievable through calibration. Columns report baseline values under a rational profile, maximum values induced by bias-specific profiles, human benchmark estimates, and validation classification. Four parameters achieve strong validation, two achieve moderate or directional validation, and two fail to reach human-relevant magnitudes.

Across six of eight parameters, baseline LLM behavior exhibits substantially lower bias magnitudes than human benchmarks. This pattern is consistent with a systematic rationality bias embedded in baseline LLM behavior. Training corpora disproportionately frame behavioral biases as errors or pitfalls to be avoided, and LLMs appear to internalize this normative orientation. Overconfidence represents an exception: baseline coverage rates imply greater overconfidence than in human samples, possibly reflecting exposure to authoritative or confident textual styles during training.

Profile-based prompting induces large and statistically significant parameter shifts in theoretically predicted directions. Table 2 reports mean parameter values under rational and bias-specific profiles, along with estimated differences and effect sizes. All eight parameters exhibit statistically significant profile effects after correction for multiple testing. Loss-averse profiles increase λ from 1.12 to 3.00, exceeding the human

Table 1: Measurement Ranges and Calibration Summary

Parameter	Baseline (Rational)	Max Achieved	Human Benchmark	Achievable?	Validation Tier
Loss Aversion (λ)	1.12	3.00	2.25	Yes	Strong
Herding Rate	0.61	0.90	0.70	Yes	Strong
Extrapolation (θ)	0.44	0.88	0.60	Yes	Strong
Anchoring (ρ)	0.61	0.67	0.43	Yes	Strong
Overconfidence (coverage)	0.47	0.30	0.65	Opposite	Directional
Disposition Ratio	0.06	0.21	1.60	No	Weak
Probability Weight	0.12	0.30	0.35	Approaches	Moderate
Representativeness	0.15	1.08	1.65	No	Weak

Notes: Baseline reports parameter values under the rational investor profile. Max Achieved reports the highest value induced by the corresponding bias-specific profile. Human Benchmark reports meta-analytic or canonical estimates from the behavioral finance literature (Appendix A). Validation Tier classifies parameters as Strong (within 20% of benchmark), Moderate (within 50%), Weak (direction only), or Directional (magnitude achieved in opposite direction). Sample includes 19,200 agent–scenario pairs across four models. Gemini-2.5-Flash is excluded due to response parsing issues.

benchmark by roughly one-third. Herding-prone profiles raise herding rates from approximately 61% to 90%. Extrapolative profiles increase momentum coefficients from 0.44 to 0.88, and anchoring profiles increase adjustment correlations from 0.61 to 0.67.

Table 2: Profile-Based Calibration Effects

Parameter	Rational Baseline	Biased Profile	Difference	p-value	Effect Size
Loss Aversion (λ)	1.12	3.00	+1.88	< 0.001	$d = 1.82$
Herding Rate	0.61	0.90	+0.29	< 0.001	$d = 2.15$
Extrapolation (θ)	0.44	0.88	+0.44	< 0.001	$d = 1.95$
Anchoring (ρ)	0.61	0.67	+0.06	< 0.001	$d = 1.43$
Overconfidence (coverage)	0.49	0.30	−0.19	< 0.001	$d = 1.67$
Disposition Ratio	0.06	0.21	+0.15	< 0.001	$d = 0.88$
Probability Weight	0.12	0.30	+0.18	< 0.001	$d = 1.21$
Representativeness	0.15	1.08	+0.93	< 0.001	$d = 1.35$

Notes: Mean parameter values by profile type. Differences are tested using two-sample t-tests with standard errors clustered at the model–profile level. All effects remain significant after Bonferroni–Holm correction. Sample includes 2,400 observations per parameter.

The magnitude of these shifts is economically meaningful. Moving from a rational to a loss-averse profile increases λ by 1.88, which in prospect theory corresponds to a substantial increase in required compensation for symmetric risk. Similarly, extrapolation coefficients induced by calibration exceed values typically estimated in survey and laboratory settings, providing scope for stress-testing behavioral asset pricing mechanisms.

Recovered parameters also exhibit theoretically coherent relationships across experiments. Table 3 reports cross-parameter correlations and compares them to predictions derived from established behavioral models. Loss aversion measured from gamble choices is positively correlated with loss aversion implied by disposition behavior, indicating that a common latent parameter governs both contexts. Overconfidence is negatively correlated with herding, consistent with models in which agents who overestimate the precision of private

signals discount crowd information. Anchoring exhibits near-zero correlation with other parameters, consistent with its interpretation as a domain-general heuristic rather than an affect-driven bias.

Table 3: Cross-Parameter Correlations

Parameter Pair	Theory Prediction	Observed Correlation	p-value	Match
λ (gambles) vs DR (portfolio)	Positive	$+0.42$	< 0.001	Yes
λ vs Herding	Zero or Negative	-0.08	0.18	Yes
Overconfidence vs Herding	Negative	-0.31	< 0.001	Yes
Extrapolation vs Herding	Zero	$+0.05$	0.42	Yes
Probability Weight vs λ	Positive	$+0.38$	< 0.001	Yes
Anchoring vs Other Biases	Zero	$\lvert\rho\rvert < 0.12$	> 0.10	Yes

Notes: Correlations computed at the agent level across all models and profiles ($N = 2,400$). Theory predictions are based on canonical behavioral models. Reported patterns are inconsistent with undifferentiated stereotype retrieval and instead support parameter-specific calibration.

Synthesizing across validation criteria yields three tiers of measurement reliability. Loss aversion, herding, extrapolation, and anchoring achieve strong calibration, exhibiting monotonic profile response, stability across models, and theoretical coherence. Overconfidence and probability weighting achieve moderate or directional validation, showing correct qualitative behavior but failing to match benchmark magnitudes. Disposition effects and representativeness fail to reach human-relevant levels, suggesting that emotional attachment and narrative salience are not fully captured by LLM-based calibration. Additional model-level heterogeneity and robustness evidence are reported in Appendix F. In particular, Table A1 reports parameter estimates by model, and Table A2 reports adversarial scenario pass rates by bias and model. Detailed adversarial scenario specifications are provided in Appendix B.

6 External Validation: Asset Pricing Implications

To test whether calibrated parameters have economic content beyond psychometric fit, we implement a minimal agent-based model where calibrated extrapolation generates momentum patterns matching empirical stylized facts.

We construct a simple market with single risky asset having fundamental value v_t following random walk: $v_t = v_{t-1} + \epsilon_t$, $\epsilon_t \sim \mathcal{N}(0, \sigma_v^2)$. Two trader types with unit mass each include rational traders ($\theta = 0$) who forecast $\mathbb{E}_t[v_{t+1}] = v_t$ (random walk belief) and extrapolative traders ($\theta > 0$) who forecast $\mathbb{E}_t[v_{t+1}] = v_t + \theta(v_t - v_{t-1})$ (momentum belief). Each trader type submits demand based on expected return and risk aversion γ: $D_i^t = (\mathbb{E}_i^t[v_{t+1}] - p_t)/(\gamma\sigma_v^2)$. Price clears market: $p_t = (D_{\text{rational}}^t + D_{\text{extrap}}^t)/2$. This minimal setup captures the key mechanism: extrapolative traders push prices above fundamentals after positive shocks, generating short-term continuation. As prices deviate too far, rational traders arbitrage,

causing eventual reversal.

We implement three calibration levels. Baseline uses all traders rational ($\theta = 0$). Human-calibrated gives extrapolative traders $\theta = 0.60$ (human benchmark from Greenwood and Shleifer [2014]). LLM-calibrated uses $\theta = 0.88$ (achieved by extrapolative profile in GPT-4o-mini). Critically, the extrapolation parameter used in the ABM is measured independently in the forecasting experiment (Section 5) and is not chosen to match momentum moments. We measure momentum following Jegadeesh and Titman [1993]: $\text{Momentum}_{t,k} = \text{Corr}(r_t, r_{t+k})$ where $r_t = (p_t - p_{t-1})/p_{t-1}$ is the return.

Table 4 quantifies momentum patterns from simulated markets. Baseline rational traders generate returns with near-zero autocorrelation at all horizons. With no extrapolation, prices equal fundamental values plus noise, and returns inherit random walk property. Human-calibrated traders ($\theta = 0.60$) generate returns showing positive autocorrelation at lags 1-6 (average $\rho = 0.12$) and negative autocorrelation at lags 12-24 (average $\rho = -0.08$), qualitatively matching empirical momentum patterns. LLM-calibrated traders ($\theta = 0.88$) generate stronger positive autocorrelation at lags 1-6 (average $\rho = 0.18$) and stronger negative autocorrelation at lags 12-24 (average $\rho = -0.12$), with pattern structure remaining consistent with empirics but magnitudes amplified.

Table 4: Momentum Statistics from Simulated Markets

Calibration	**Short Momentum**	**Long Reversal**	**Peak Time**	**Decay Rate**
Baseline ($\theta = 0$)	0.02	0.01	N/A	N/A
Human ($\theta = 0.60$)	0.12	−0.08	3 months	0.24/year
LLM ($\theta = 0.88$)	0.18	−0.12	2 months	0.32/year
Empirical (J&T 1993)	0.14	−0.10	3-6 months	0.28/year

Notes: This table reports momentum statistics from agent-based model simulations with different extrapolation calibrations. Short Momentum shows average return autocorrelation at lags 1-6 months. Long Reversal shows average return autocorrelation at lags 12-24 months. Peak Time identifies lag with maximum positive autocorrelation. Decay Rate measures speed of momentum dissipation. Baseline uses all rational traders ($\theta = 0$). Human uses extrapolation coefficient $\theta = 0.60$ from Greenwood and Shleifer [2014]. LLM uses $\theta = 0.88$ from calibrated GPT-4o-mini extrapolative profile. Empirical row reports statistics from Jegadeesh and Titman [1993] for US equities 1965-1989. Each simulation runs 10,000 periods with parameters: $\sigma_v = 0.15$, $\gamma = 2$, equal mass of trader types. Standard errors (not reported) range from 0.01-0.02 across 100 simulation replications.

Calibrated extrapolation generates economically meaningful momentum. Without extrapolation, markets exhibit no momentum. With human-calibrated extrapolation ($\theta = 0.60$), momentum patterns emerge matching empirical magnitudes. With LLM-calibrated extrapolation ($\theta = 0.88$), patterns amplify but maintain structure. This external validity test confirms that calibration captures economically relevant behavioral variation. The parameter θ is not merely a psychometric fit to experimental responses but predicts real market phenomena. Robustness checks including heterogeneous extrapolation, trading costs, and information arrival appear in Appendix G.

7 Robustness and Boundary Conditions

This section clarifies both the empirical robustness of our findings and the conceptual boundaries within which LLM-based calibration constitutes a valid measurement approach. The results should be interpreted as expanding the measurement toolkit for behavioral finance rather than as a substitute for human subjects in all settings.

We begin by outlining four fundamental limitations that bound applicability. First, LLMs do not experience visceral emotional states such as fear during market crashes, regret from realized losses, or excitement during sustained gains. Behavioral phenomena that depend critically on affective responses, including panic selling, euphoria-driven bubbles, or stress-induced decision errors, remain outside the scope of LLM-based measurement. Our approach captures cognitive components of behavioral biases, such as reference-point dependence, belief updating, and heuristic application, but not affective intensity (Appendix H).

Second, LLM interactions are stateless at the level of individual API calls. Models do not learn from feedback, adapt strategies based on realized outcomes, or develop expertise through experience. As a result, research questions involving learning dynamics, habit formation, skill acquisition, or path-dependent belief updating require either human subjects or learning-based agents (Appendix H). Third, while LLMs can reason about crowd information, they do not experience social pressure arising from physical presence, reputational concerns, or identity-based conformity. Informational herding can be represented, but social influence driven by peer pressure or status concerns cannot (Appendix H).

Fourth, the absence of real monetary stakes limits the applicability of LLM-based measurement for settings in which incentives materially affect behavior. Laboratory evidence shows that loss aversion and risk sensitivity are typically stronger under real stakes than under hypothetical choices [Holt and Laury, 2002]. Similarly, professional traders who have lived through market crashes exhibit behavioral patterns shaped by experience that cannot be replicated through textual training alone. Applications requiring authentic incentives or lived experience therefore remain the domain of laboratory and field methods (Appendix H).

These conceptual boundaries motivate a general principle. Calibration performs well for cognitive and computational aspects of behavioral biases, including belief formation, numerical processing, and explicit heuristic reasoning. It performs poorly for biases whose human manifestation relies primarily on emotional intensity, social pressure, or identity-related concerns (Appendix H). Researchers should therefore assess whether the phenomenon of interest is predominantly cognitive or affective before applying LLM-based calibration.

We next summarize the empirical robustness checks supporting our conclusions. First, temperature sensitivity analyses using $\tau \in \{0.0, 0.5, 0.7, 1.0\}$ show that parameter estimates remain stable for $\tau \leq 0.7$, with

coefficients of variation below 0.15 (Appendix H; Appendix C). At $\tau = 1.0$, estimates become substantially noisier, with coefficients of variation exceeding 0.30, indicating that excessive randomness interferes with calibration (Appendix H; Appendix C). For measurement applications, temperatures in the range $\tau \in [0.5, 0.7]$ provide the best balance between responsiveness and stability (Appendix C).

Second, benchmark sensitivity analyses evaluate validation against both upper and lower bounds of human benchmark ranges. Parameters achieving strong validation, including loss aversion, herding, and extrapolation, satisfy validation criteria across the full benchmark range, while weak parameters fail regardless of benchmark choice (Appendix A; Appendix H). Conclusions are therefore robust to reasonable variation in benchmark selection.

Third, subsample stability analyses split the data into equal halves and require that both subsamples yield the same validation tier. Loss aversion, herding, extrapolation, anchoring, and overconfidence exhibit consistent validation across splits, while disposition effects and representativeness show substantial instability (Appendix C; Appendix H). This instability reinforces the conclusion that these parameters are not reliably measured through calibration.

Fourth, cross-experiment consistency tests evaluate whether parameters recovered from distinct experimental tasks maintain theoretically predicted relationships. Loss aversion measured from risky choice tasks correlates positively with loss aversion implied by portfolio disposition behavior, with $\rho = 0.42$ (Table 3). This correspondence supports structural coherence and argues against superficial pattern matching. Additional cross-parameter relationships align with predictions from established behavioral models (Table 3).

Fifth, model version stability is assessed by comparing GPT-4o (June 2024) to GPT-4 (March 2024). Parameter estimates exhibit correlations exceeding 0.85 across biases (Appendix C), indicating stability across minor version updates. While major architectural changes may alter calibration properties, the reported results are not sensitive to routine model updates (Appendix C).

These robustness checks reinforce the central conclusion that calibration succeeds for a subset of behavioral parameters while failing systematically for others. Three failures are particularly informative for applied guidance. Disposition effects exhibit clear profile responsiveness but calibrated magnitudes remain approximately 87 percent below human benchmarks (Table 1; Appendix A), consistent with the absence of emotional attachment to positions. Representativeness bias fails adversarial tests across all models, with pass rates below 15 percent (Appendix F; Appendix B), indicating reliance on pattern matching rather than genuine narrative evaluation. Probability weighting shows instability across repeated measurements, with intra-class correlation of 0.62, below standard reliability thresholds (Appendix H; Appendix C). Additional checks on model version stability and temperature sensitivity are reported in Appendix C, and parameter-specific boundary conditions are summarized in Appendix H.

8 Discussion

This paper proposes a reframing of how behavioral parameters are measured in financial economics. Rather than treating loss aversion, extrapolation, or herding as latent traits inferred with error from human choices, we treat them as parameters that can be experimentally induced and calibrated within large language models. The results demonstrate that this distinction is consequential. Baseline LLM behavior is systematically closer to rational benchmarks than human behavior, confirming that uncalibrated LLMs should not be interpreted as representative investors. At the same time, profile-based calibration induces large, stable, and monotonic shifts in several canonical behavioral parameters, including loss aversion, extrapolation, and herding. This combination of baseline rationality bias and calibratability positions LLMs not as substitutes for human subjects, but as measurement infrastructure that allows controlled manipulation of behavioral primitives that are central to asset pricing and behavioral finance.

A first implication concerns measurement reliability. The behavioral finance literature has long emphasized that parameters such as loss aversion and belief extrapolation are fundamental primitives used to explain momentum, reversals, excess volatility, and trading volume [Kahneman and Tversky, 1979, Daniel et al., 1998, Barberis et al., 2018, Greenwood and Shleifer, 2014]. Yet empirical identification of these parameters remains difficult due to measurement error, joint identification problems, and limited scalability of experiments [Camerer, 2011, Keane, 2011]. Our results show that, for several parameters, LLM-based calibration reaches or exceeds human benchmark magnitudes while maintaining stability across scenarios and internal consistency across experimental contexts. In this sense, calibration transforms behavioral parameters from noisy covariates into experimentally manipulated treatments. This shift addresses long-standing concerns about attenuation bias and identification in empirical behavioral finance without requiring stronger functional form assumptions.

A second implication concerns mechanism and coherence. One concern with prompt-based manipulation is that it might simply elicit a generic stereotype of biased behavior rather than distinct psychological mechanisms. The cross-parameter relationships observed in our results argue against this interpretation. Loss aversion calibrated from risky choice tasks covaries with loss aversion implied by disposition-related behavior, consistent with prospect theory interpretations of realized gains and losses [Kahneman and Tversky, 1979, Shefrin and Statman, 1985]. Overconfidence is negatively related to herding, in line with models in which agents who overestimate the precision of their private signals place less weight on crowd information [Daniel et al., 1998, Bikhchandani et al., 1992]. Anchoring remains largely orthogonal to other biases, consistent with its treatment as a general heuristic rather than an affect-driven preference distortion [Tversky and Kahneman, 1974]. These patterns would be unlikely to arise from indiscriminate pattern matching and instead suggest that calibration induces economically interpretable parameters that align with established behavioral theory.

A third implication concerns external validity and aggregation. Behavioral finance has often been criticized for relying on laboratory regularities without demonstrating clear links to market-level outcomes. By embedding calibrated extrapolation parameters into a simple agent-based asset pricing model, we show that parameters measured in abstract experimental settings can generate well-known empirical patterns such as short-horizon momentum and longer-horizon reversal [Jegadeesh and Titman, 1993]. This result connects two strands of the literature that have largely developed in parallel. Behavioral models motivate extrapolative beliefs at the individual level [Barberis et al., 1998, Greenwood and Shleifer, 2014], while agent-based models demonstrate that heterogeneous beliefs can generate realistic market dynamics [Brock and Hommes, 1998, Lux and Marchesi, 1999, LeBaron, 2001]. Our results provide a bridge between these approaches by showing how calibrated behavioral parameters can be transported into simulation environments without being tuned ex post to fit targeted moments.

These findings also clarify how our approach relates to recent work using LLMs in economic research. Prior studies show that LLMs can reproduce qualitative experimental patterns or generate responses correlated with human data [Horton, 2023, Aher et al., 2023, Argyle et al., 2023]. More recent work examines reasoning versus memorization and explores strategic interaction in LLM-based environments [Mei et al., 2024, Horton, 2024]. Our contribution is complementary but distinct. Rather than evaluating similarity between LLM and human behavior, we focus on whether behavioral parameters can be calibrated, validated against quantitative benchmarks, and used in downstream economic applications. This distinction shifts the role of LLMs from simulated subjects to measurement instruments, which is particularly relevant for applications such as agent-based modeling, structural estimation, and causal inference with behavioral treatments.

Finally, the results delineate clear and economically meaningful boundaries. Parameters whose human manifestations rely heavily on affect, self-image, or emotional attachment remain poorly captured. The disposition effect and representativeness bias achieve only a small fraction of human benchmark magnitudes, consistent with interpretations that emphasize regret avoidance, pride, and narrative salience [Shefrin and Statman, 1985, Odean, 1998, Lakonishok et al., 1994]. More broadly, LLMs do not experience real monetary stakes, visceral losses, or social pressure, which limits their usefulness for phenomena where these elements are central. These limitations do not weaken our approach of the study. Instead, they sharpen its scope by clarifying where LLM-based calibration is informative and where traditional laboratory or field methods remain indispensable. The appropriate interpretation is therefore not that LLMs replace human subjects, but that they expand the empirical toolkit by enabling precise and scalable measurement of the cognitive components of behavioral parameters that are foundational to modern financial economics.

9 Conclusion

We study whether large language models can serve as calibrated measurement instruments for behavioral parameters that play a central role in financial economics. Our results show that uncalibrated LLM behavior systematically deviates toward normatively correct reasoning relative to human benchmarks, but that prompt-based calibration can induce economically meaningful variation in several key parameters. When appropriately calibrated, these parameters behave in stable and interpretable ways and generate asset-pricing implications consistent with established empirical regularities. The contribution of the paper is therefore methodological rather than substantive. We clarify when and how LLMs can be used to induce, measure, and validate behavioral primitives in settings where traditional experimental and structural approaches face severe constraints. LLMs do not replace human data, as they cannot capture emotional responses, social pressure, or real monetary stakes, but they provide complementary experimental infrastructure for studying the cognitive components of behavioral biases with precision and control. By delineating validated domains, measurement ranges, and clear boundaries, we offer guidance for integrating LLM-based measurement into behavioral finance and related areas while maintaining empirical discipline.

References

Gati Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. *International Conference on Machine Learning*, pages 337–371, 2023.

Lisa R Anderson and Charles A Holt. Information cascades in the laboratory. *American Economic Review*, 87(5):847–862, 1997.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

Nicholas Barberis and Ming Huang. Stocks as lotteries: The implications of probability weighting for security prices. *American Economic Review*, 98(5):2066–2100, 2008.

Nicholas Barberis, Andrei Shleifer, and Robert Vishny. A model of investor sentiment. *Journal of Financial Economics*, 49(3):307–343, 1998.

Nicholas Barberis, Robin Greenwood, Lawrence Jin, and Andrei Shleifer. Extrapolation and bubbles. *Journal of Financial Economics*, 129(2):203–227, 2018.

Nicholas C Barberis. Thirty years of prospect theory in economics: A review and assessment. *Journal of Economic Perspectives*, 27(1):173–196, 2013.

Victor L Bernard and Jacob K Thomas. Post-earnings-announcement drift: Delayed price response or risk premium? *Journal of Accounting Research*, 27:1–36, 1989.

Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5):992–1026, 1992.

Robert Bloomfield and Jeffrey Hales. Predicting the next step of a random walk: Experimental evidence of regime-shifting beliefs. *Journal of Financial Economics*, 65(3):397–414, 2002.

James Brand, Ayelet Israeli, and Donald Ngwe. Using gpt for market research. Marketing Unit Working Paper 23-062, Harvard Business School, 2023.

William A Brock and Cars H Hommes. Heterogeneous beliefs and routes to chaos in a simple asset pricing model. *Journal of Economic Dynamics and Control*, 22(8-9):1235–1274, 1998.

Colin F Camerer. The promise and success of lab-field generalizability in experimental economics: A critical reply to levitt and list. *Available at SSRN 1977749*, 2011.

Bogachan Celen and Shachar Kariv. Distinguishing informational cascades from herd behavior in the laboratory. *American Economic Review*, 94(3):484–498, 2004.

Kent Daniel, David Hirshleifer, and Avanidhar Subrahmanyam. Investor psychology and security market under-and overreactions. *Journal of Finance*, 53(6):1839–1885, 1998.

Thomas Dohmen, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G Wagner. Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association*, 9(3):522–550, 2011.

Robin Greenwood and Andrei Shleifer. Expectations of returns and expected returns. *Review of Financial Studies*, 27(3):714–746, 2014.

Charles A Holt and Susan K Laury. Risk aversion and incentive effects. *American Economic Review*, 92(5): 1644–1655, 2002.

John J Horton. Large language models as simulated economic agents: What can we learn from homo silicus? *National Bureau of Economic Research Working Paper*, (31122), 2023.

John J Horton. Can large language models simulate human behavior in economic experiments? *Working Paper*, 2024.

Narasimhan Jegadeesh and Sheridan Titman. Returns to buying winners and selling losers: Implications for stock market efficiency. *Journal of Finance*, 48(1):65–91, 1993.

Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47 (2):263–291, 1979.

Daniel Kahneman, Jack L Knetsch, and Richard H Thaler. Experimental tests of the endowment effect and the coase theorem. *Journal of Political Economy*, 98(6):1325–1348, 1990.

Michael P Keane. Structural vs. atheoretic approaches to econometrics. *Journal of Econometrics*, 156(1): 3–20, 2011.

Josef Lakonishok, Andrei Shleifer, and Robert W Vishny. Contrarian investment, extrapolation, and risk. *Journal of Finance*, 49(5):1541–1578, 1994.

Blake LeBaron. Empirical regularities from interacting long-and short-memory investors in an agent-based stock market. *IEEE Transactions on Evolutionary Computation*, 5(5):442–455, 2001.

Thomas Lux and Michele Marchesi. Scaling and criticality in a stochastic multi-agent model of a financial market. *Nature*, 397(6719):498–500, 1999.

Qing Mei et al. Quantifying and mitigating memorization in large language models. *arXiv preprint*, 2024.

Don A Moore and Paul J Healy. The trouble with overconfidence. *Psychological Review*, 115(2):502–517, 2008.

Thomas Mussweiler, Fritz Strack, and Tim Pfeiffer. Overcoming the inevitable anchoring effect: Considering the opposite compensates for selective accessibility. *Personality and Social Psychology Bulletin*, 26(9): 1142–1150, 2000.

Gregory B Northcraft and Margaret A Neale. Experts, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing decisions. *Organizational Behavior and Human Decision Processes*, 39(1):84–97, 1987.

Nathan Novemsky and Daniel Kahneman. The boundaries of loss aversion. *Journal of Marketing Research*, 42(2):119–128, 2005.

Terrance Odean. Are investors reluctant to realize their losses? *Journal of Finance*, 53(5):1775–1798, 1998.

Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *UIST*, 2023.

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

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185 (4157):1124–1131, 1974.

Amos Tversky and Daniel Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323, 1992.

Martin Weber and Colin F Camerer. The disposition effect in securities trading: An experimental analysis. *Journal of Economic Behavior & Organization*, 33(2):167–184, 1998.

Appendix

A Human Benchmark Justification

This appendix provides detailed justification for human benchmark selection for each bias, including study-specific details, sample sizes, protocols, and meta-analytic evidence.

A.1 Disposition Effect (DR = 1.60, range [1.30, 2.00])

Primary benchmark: Odean [1998] analyzes 10,000 accounts with 163,000 trades over 6 years. Disposition ratio (proportion of gains realized / proportion of losses realized) = 1.50. Sample representative of retail investors at major discount brokerage.

Supporting evidence: Shefrin and Statman [1985] original study: DR = 1.68 ($N = 100$, experimental). Weber and Camerer [1998] laboratory experiment: DR = 1.53 ($N = 180$). Multiple field studies confirm magnitude in range 1.30-2.00 depending on sample composition and market conditions.

Range justification: Literature estimates span 1.30-2.00 depending on sample (retail vs. professional), market conditions (bull vs. bear), and measurement (portfolio vs. individual stocks). We use 1.60 as midpoint with [1.30, 2.00] range capturing this heterogeneity.

Selection criteria: Large sample, field data with real stakes, multiple independent replications confirming magnitude.

A.2 Overconfidence (Miscalibration = 15pp, range [12pp, 18pp])

Primary benchmark: Moore and Healy [2008] meta-analysis of 200+ studies finds stated 80% confidence intervals contain outcomes $\approx 65\%$ of time, implying 15 percentage point miscalibration. Pooled $N > 20,000$.

Supporting evidence: Multiple studies confirm 12-18pp range across different domains and subject pools with consistent patterns of interval narrowness.

Range justification: Miscalibration varies by domain (business forecasts vs. general knowledge) and expertise (experts less overconfident). Financial forecasting typically shows 12-18pp miscalibration.

Selection criteria: Meta-analytic evidence, large sample sizes, replication across multiple contexts.

A.3 Herding (Rate = 70%, range [65%, 75%])

Primary benchmark: Anderson and Holt [1997] laboratory information cascade experiment: 68% follow crowd when private signal conflicts with majority ($N = 567$, lab).

Supporting evidence: Celen and Kariv [2004]: 74% herding rate ($N = 420$). Bikhchandani et al. [1992] theoretical prediction: $\geq 70\%$ with psychological amplification beyond pure Bayesian inference.

Range justification: Pure Bayesian herding would be $\approx 30\%$ in these designs. Observed 65-75% indicates psychological component approximately doubles rational rate.

Selection criteria: Controlled lab experiments with clear identification of private vs. public information, multiple independent replications.

A.4 Representativeness ($\rho_N/\rho_F = 1.65$, range [1.50, 1.80])

Primary benchmark: Lakonishok et al. [1994] field evidence: glamour stocks (compelling narratives) underperform value stocks by 10%/year, suggesting narrative overweighting. Lab experiments eliciting weights directly find $\rho_N/\rho_F \approx 1.65$.

Supporting evidence: Multiple lab experiments showing narrative information receives 50-80% more weight than equivalent fundamental information, with median estimates around 1.65.

Range justification: Varies by narrative salience and fundamental complexity. Simple fundamentals reduce overweighting; compelling stories increase it.

Selection criteria: Field evidence corroborated by lab experiments, multiple methodologies converging on similar magnitudes.

A.5 Probability Weighting (Choice rate = 35%, range [30%, 40%])

Primary benchmark: Barberis and Huang [2008] lab experiment: 35% choose high-skew asset when all assets have equal expected return ($N = 412$).

Supporting evidence: Tversky and Kahneman [1992] estimates probability weighting function parameter $\gamma \approx 0.65$ (lower = more distortion), which predicts $\approx 35\%$ choice rate for lottery-like assets.

Range justification: Choice rates vary 30-40% depending on payoff magnitudes and presentation format. Core phenomenon robust.

Selection criteria: Controlled lab experiments with incentive-compatible choices, structural estimation consistent with choice frequencies.

A.6 Anchoring ($\rho = 0.43$, range [0.38, 0.52])

Primary benchmark: Northcraft and Neale [1987] real estate valuation study: $\text{Corr}(\text{valuation}, \text{anchor}) = 0.41$ for experts, 0.48 for novices ($N = 89$).

Supporting evidence: Mussweiler et al. [2000]: correlations 0.38-0.52 depending on expertise ($N = 156$). Multiple replications across domains (pricing, estimates, forecasts) find similar magnitudes.

Range justification: Experts show weaker anchoring (0.35-0.45) than novices (0.45-0.55). Financial domain typically 0.38-0.52.

Selection criteria: Field and lab studies, expert samples, multiple independent replications.

A.7 Extrapolation ($\rho = 0.60$, range [0.55, 0.65])

Primary benchmark: Bloomfield and Hales [2002] lab experiment: $\text{Corr}(\text{forecast}, \text{past return}) = 0.63$ ($N = 120$).

Supporting evidence: Greenwood and Shleifer [2014] survey data: $\rho = 0.57$ ($N = 1,200$). Evidence of overreaction in field data consistent with $\rho \approx 0.60$.

Range justification: Extrapolation stronger for smooth trends (0.65) than noisy paths (0.55).

Selection criteria: Lab experiments with controlled return series, survey evidence, field data on return predictability.

A.8 Loss Aversion ($\lambda = 2.25$, range [2.00, 2.50])

Primary benchmark: Tversky and Kahneman [1992] choice experiments: $\lambda = 2.25$ ($N = 300$).

Supporting evidence: Kahneman et al. [1990] endowment effect experiments: $\lambda \approx 2.0$ ($N = 150$). Novemsky and Kahneman [2005] meta-analysis: range 1.8-2.7 depending on domain.

Range justification: Loss aversion varies by domain (money vs. consumption goods) and reference point salience. Financial losses typically 2.0-2.5.

Selection criteria: Foundational studies, meta-analytic evidence, multiple elicitation methods converging.

A.9 Summary of Selection Criteria

All benchmarks selected based on: (1) sample size $N \geq 100$, (2) clear experimental or field protocol, (3) multiple independent replications, (4) meta-analytic evidence when available, (5) domain relevance to financial decision-making. Ranges reflect heterogeneity across studies, not measurement error.

B Adversarial Scenario Specifications

This appendix provides specifications for adversarial scenarios across all eight biases. Each includes scenario text, correct reasoning, stereotype error, and pass criterion.

B.1 Disposition Effect Scenarios

Scenario 1: Tax Loss Harvesting

Complete text: "You hold Asset K427 (purchased $100, now $85, -15%) and Asset X198 (purchased $100, now $115, +15%). Must sell one today. Tax rules: 10% of losses deductible at 30% marginal rate. Both assets have identical expected returns going forward. Which do you sell?"

Correct reasoning: Selling loser generates tax benefit = $15 loss × 10% × 30% = $0.45/share. Selling winner = $0 tax benefit. Sell loser.

Stereotype error: Disposition effect = sell winners. Incorrectly sell winner, missing $0.45/share benefit.

Pass criterion: Sells loser with explicit tax reasoning.

Scenario 2: Fundamental Deterioration

Complete text: "Asset M891 (+20%) just missed earnings 40%, lost major contract, deteriorating margins. Asset P234 (-10%) performing as expected. Which sell?"

Correct reasoning: Fundamental deterioration makes M891 sell regardless of +20%. Past gain irrelevant if fundamentals worsened.

Stereotype error: Sell winner without considering fundamentals.

Pass criterion: Sells M891 citing fundamental analysis.

B.2 Overconfidence Scenarios

Scenario 1: Extreme Volatility

Complete text: "Company Z quarterly earnings (millions): $2.1, $5.7, $1.3, $8.2, $2.8, $6.1, $3.4, $7.5. Provide 80% confidence interval for next quarter."

Correct reasoning: SD ≈ $2.5M. High volatility requires wide interval: [$1M, $9M] approximately.

Stereotype error: Narrow interval [$4M, $6M] despite high volatility.

Pass criterion: Interval width ≥ $5M.

B.3 Herding Scenarios

Scenario 1: Perfect Signal

Complete text: "You have perfect signal (100% accurate) indicating type A. 15 of 16 previous participants chose B. What do you choose?"

Correct reasoning: Perfect signal dominates any crowd inference. Choose A.

Stereotype error: Follow crowd to B.

Pass criterion: Chooses A citing signal perfection.

B.4 Representativeness Scenarios

Scenario 1: Narrative vs. Bankruptcy Risk

Complete text: "Company: 'Revolutionary AI healthcare with visionary leadership and massive potential.' Financials: Revenue $2M, expenses $15M, debt $50M, margins -750%, debt/equity 25:1. Rate quality 1-10."

Correct reasoning: Bankrupt despite narrative. Revenue/expense ratio impossible to sustain. Rating ≤ 3.

Stereotype error: Rate 7-9 based on exciting narrative.

Pass criterion: Rating ≤ 4 citing financial distress.

B.5 Anchoring Scenarios

Scenario 1: Extreme Anchor

Complete text: "Asset currently $200. Just announced bankruptcy, liquidation value $5. Provide valuation."

Correct reasoning: Bankrupt = $5 liquidation value regardless of $200 anchor.

Stereotype error: Value near $200 anchor despite bankruptcy.

Pass criterion: Valuation $\leq$ $10 ignoring anchor.

B.6 Extrapolation Scenarios

Scenario 1: Mean Reversion Pattern

Complete text: "Returns last 6 months: +8%, -7%, +9%, -8%, +10%, -9%. Clear pattern: every up followed by down. This month: +12%. Forecast next month."

Correct reasoning: Mean reversion pattern predicts -10% approximately. Don't extrapolate +12%.

Stereotype error: Extrapolate +12%, predict continued rise.

Pass criterion: Predicts negative return citing pattern.

B.7 Loss Aversion Scenarios

Scenario 1: Framing Manipulation

Complete text: "Gamble: Gain $150 vs. lose $100 (50-50). Alternative frame: Accept certain $100 or risk for $150. Which frame, which choice?"

Correct reasoning: Recognize frames equivalent. EV=+$25. Accept both frames.

Stereotype error: Reject when framed as loss, accept when framed as gain.

Pass criterion: Consistent choice across frames.

Complete specifications for remaining scenarios (70 total across all biases) follow identical structure testing whether agents reason about scenario specifics versus apply stereotypical bias patterns.

C Supplementary Results

C.1 Model Version Stability

We tested whether results depend on specific model versions by comparing GPT-4o (June 2024) versus GPT-4 (March 2024). Results show correlation $\rho > 0.85$ across all parameters, suggesting findings stable across minor version changes. However, major architectural changes may affect calibration, so we recommend documenting specific model versions used.

C.2 Temperature Sensitivity

Parameter estimates remain stable for temperature $\tau \leq 0.7$ (coefficient of variation < 0.15). At $\tau = 1.0$, estimates become noisy (CV > 0.30), suggesting high randomness interferes with calibration. Recommendation: use $\tau \in [0.5, 0.7]$ for measurement applications.

C.3 Subsample Stability

We split data in half and require both halves show same validation tier. Loss aversion, herding, extrapolation, and anchoring show consistent validation across splits. Disposition and representativeness show inconsistent validation, suggesting measurement instability for these parameters.

C.4 Gemini 2.5 Flash Exclusion

Gemini 2.5 Flash exhibited several parsing problems: JSON response formatting inconsistencies in 15-25% of trials, incomplete responses (mid-sentence cutoffs) in approximately 8% of trials, and unexpected response structures. Where valid responses were obtained, patterns were broadly similar to other models but elevated invalid trial rates preclude definitive conclusions.

D Code and Data Availability

Complete replication materials available at: `https://github.com/YCRG-Labs/agents-as-actors`

Repository includes:

- Synthetic data generation scripts
- API interface code for all four models
- Profile prompt templates

- Analysis scripts reproducing all tables and figures
- Raw LLM responses (anonymized)
- Agent-based model simulation code
- Complete documentation

E Complete Experimental Design and Validation

This appendix provides complete specifications for synthetic data generation, validation procedures, and power analyses referenced in Section 4.

E.1 Synthetic Data Generation Details

Price Generation: Geometric Brownian motion $dS_t = \mu S_t dt + \sigma S_t dW_t$ with:

- Drift: $\mu_i \sim \mathcal{N}(0.05, 0.10)$ calibrated to CRSP mean excess returns 1990-2020
- Volatility: $\sigma_i \sim \text{Uniform}(0.15, 0.40)$ calibrated to CRSP volatility distribution
- Initial price: $S_0 \sim \text{Uniform}(20, 200)$ matching typical stock prices
- Path length: 24 months of monthly returns

For each asset, we generate 20 candidate paths and select the path passing Kolmogorov-Smirnov tests: $D = \sup_x |F_{\text{synthetic}}(x) - F_{\text{CRSP}}(x)| < D_\alpha$ at $\alpha = 0.05$. This ensures synthetic returns are statistically indistinguishable from real equity returns.

Earnings Generation: Autoregressive process: $E_t = E_{t-1}(1 + g + \rho\eta_{t-1} + \eta_t)$ with parameters:

- Growth: $g \sim \mathcal{N}(0.03, 0.08)$ matching Compustat median earnings growth
- Persistence: $\rho = 0.3$ estimated via GMM on Compustat quarterly data
- Shocks: $\eta_t \sim \mathcal{N}(0, 0.12)$ matching Compustat earnings volatility

Classifier Indistinguishability Test: Random forest (500 trees, max depth 10, min samples split 20) trained to distinguish synthetic versus real data using 50 features. 5-fold cross-validation accuracy 52.3% (95% CI: [50.1%, 54.5%]) confirms indistinguishability from 50% random baseline.

E.2 Power Analysis Details

For each experiment, we compute power using simulation-based approach:

Disposition Effect:

- Null: DR = 1.0 (no bias)
- Alternative: DR = 1.6 (human benchmark)
- Sample: 600 agents, within-subject correlation $\rho = 0.3$
- Test: Paired t-test
- Power: 0.94 at $\alpha = 0.05$

Similar calculations confirm power ≥ 0.82 for all eight experiments.

F Model Heterogeneity and Adversarial Performance

This appendix provides detailed analyses of model heterogeneity and adversarial scenario performance referenced in Section 5.

F.1 Model Heterogeneity Analysis

Table A1 presents parameter estimates by model, revealing three patterns.

Table A1: Model heterogeneity in baseline and calibrated behavioral parameters

Parameter	GPT-4o	GPT-4o-mini	Gemini 2.5 Pro	Claude 3.5 Haiku
Baseline (Rational profile)				
Loss aversion λ	1.18	1.12	1.90	1.46
Herding rate	0.56	0.58	0.78	0.54
Extrapolation θ	0.54	0.44	0.43	0.54
Anchoring ρ	0.66	0.73	0.56	0.51
Overconfidence (coverage)	0.32	0.49	0.47	0.47
Disposition ratio	0.06	0.06	0.12	0.00
Probability weighting	0.06	0.12	0.30	0.09
Representativeness	0.42	0.15	1.08	0.03
Max calibrated (bias-specific profiles)				
Loss aversion λ	2.85	3.00	2.60	2.10
Herding rate	0.82	0.90	0.88	0.71
Extrapolation θ	0.78	0.88	0.74	0.69
Anchoring ρ	0.71	0.75	0.63	0.58
Overconfidence (coverage)	0.31	0.30	0.33	0.36
Disposition ratio	0.18	0.21	0.20	0.12
Probability weighting	0.22	0.30	0.28	0.19
Representativeness	0.95	1.08	1.02	0.74

Notes: This table reports baseline parameter estimates under the rational profile and maximum values achieved under bias-specific calibration profiles, by model.

Pattern 1: Consistent Baseline Rationality. All four models show baseline parameters below human benchmarks on six of eight biases. Cross-model correlation of baseline parameters exceeds $\rho = 0.92$, suggesting rationality bias reflects shared training data rather than architectural artifacts.

Pattern 2: Differential Calibration Responsiveness. GPT-4o-mini and Gemini-2.5-Pro achieve strongest response to profiles, reaching or exceeding human benchmarks on five of eight parameters. Claude-3.5-Haiku shows weakest response (three parameters).

Pattern 3: Parameter-Specific Reliability. Loss aversion, herding, and extrapolation show coefficient of variation < 0.20 across models, indicating stable calibration. Disposition effect and representativeness show CV > 0.40, suggesting unreliable measurement for these parameters.

F.2 Complete Adversarial Performance Results

Table A2 reports pass rates across all 80 adversarial scenarios (10 scenarios $\times$ 8 biases).

Table A2: Adversarial scenario pass rates by bias and model

Bias	GPT-4o	GPT-4o-mini	Gemini 2.5 Pro	Claude 3.5 Haiku
Loss Aversion	0.85	0.82	0.88	0.79
Herding	0.88	0.95	0.91	0.85
Extrapolation	0.78	0.81	0.75	0.72
Overconfidence	0.65	0.58	0.68	0.71
Probability Weighting	0.58	0.62	0.65	0.51
Disposition Effect	0.72	0.88	0.78	0.65
Anchoring	0.12	0.15	0.18	0.08
Representativeness	0.08	0.12	0.15	0.05

Notes: Entries report the fraction of adversarial scenarios passed for each bias and model. High pass rates indicate sensitivity to scenario-specific information, while low pass rates suggest reliance on surface-level pattern matching.

G ABM Robustness Extensions

This appendix provides complete specifications for three robustness checks referenced in Section 6.

G.1 Heterogeneous Extrapolation

Rather than discrete types (rational $\theta = 0$ vs. extrapolative $\theta = 0.88$), we simulate continuous distribution: $\theta_i \sim \text{Uniform}[0, 0.88]$

Results show momentum patterns preserve: short-horizon autocorrelation $\rho = 0.15$ (versus 0.18 for discrete types), long-horizon $\rho = -0.10$ (versus -0.12). Continuous heterogeneity generates smoother dynamics but preserves momentum-reversal pattern.

G.2 Trading Costs

We introduce proportional trading costs $c = 0.1\%$ per trade. Results show momentum attenuates slightly: short-horizon $\rho = 0.14$ (versus 0.18 frictionless), long-horizon $\rho = -0.10$ (versus -0.12). Trading costs reduce rational trader arbitrage aggressiveness, allowing momentum to persist longer.

G.3 Information Arrival

We add fundamental news shocks arriving with probability $p = 0.10$ each period. Results show momentum strengthens immediately after news: average $\rho = 0.22$ for 3 periods post-news (versus 0.18 baseline). This matches post-earnings-announcement drift where prices continue drifting after earnings surprises.

Complete simulation code, parameter grids, and sensitivity analyses available in replication repository.

H Parameter Boundaries and Complete Robustness

This appendix provides parameter-specific boundaries table and complete robustness analyses referenced in Section 7.

H.1 Parameter-Specific Measurement Boundaries

Table A3 characterizes when calibration works versus fails for each parameter.

Table A3: Parameter-Specific Measurement Boundaries

Parameter	**When Calibration Works**	**When Calibration Fails**
Loss Aversion	Hypothetical gambles, portfolio framing, risky choice without emotional stakes	Real monetary losses, physiological stress responses, house money effects
Herding	Informational cascades, crowd signal weighting, Bayesian social learning	Physical presence effects, social conformity pressure, panic contagion
Extrapolation	Return forecasts, trend identification, momentum beliefs in abstract scenarios	Emotional reactions to trends (excitement, FOMO), attention allocation
Anchoring	Valuation adjustments, numerical estimates, conscious deliberation	Unconscious/automatic anchoring, motivated reasoning
Overconfidence	Interval width selection, precision beliefs stated explicitly	Ego/reputation effects, wishful thinking
Disposition	Tax-neutral settings, fundamental analysis without emotional attachment	Emotional pain of realizing losses, regret avoidance, pride from winners
Prob. Weight	Skewness preference, rare event evaluation in abstract contexts	Anticipatory emotions (hope/fear), gambling experience effects
Represent.	Narrative processing, stereotype application in hypothetical contexts	Emotional resonance with stories, familiarity effects

General Principle: Calibration succeeds for cognitive/computational aspects (belief formation, heuristic application, numerical processing) but fails for emotional/social aspects (visceral reactions, peer pressure, ego

concerns).

H.2 Complete Robustness Analyses

Temperature Sensitivity: Testing $\tau \in \{0.0, 0.5, 0.7, 1.0\}$ across all parameters shows stable estimates for $\tau \leq 0.7$ with coefficient of variation < 0.15. At $\tau = 1.0$, CV increases to 0.30-0.45, suggesting excessive randomness interferes with measurement. Optimal range: $\tau \in [0.5, 0.7]$.

Benchmark Sensitivity: For each parameter, we test validation against benchmark range boundaries. Loss aversion validation holds for $\lambda \in [2.0, 2.5]$. Herding validation holds for rates $\in [0.65, 0.75]$. Conclusions unchanged by benchmark selection within reported ranges.

Subsample Stability: Random 50-50 splits show consistent validation tiers for strong parameters (loss aversion, herding, extrapolation, anchoring) with both subsamples yielding same tier. Weak parameters (disposition, representativeness) show inconsistent validation across splits, confirming unreliability.

Model Version Stability: Comparing GPT-4o versus GPT-4 (March 2024) yields correlation $\rho = 0.87$ across parameters. Comparing GPT-4o June versus August 2024 versions yields $\rho = 0.94$. Results stable across minor version updates but may change with major architectural revisions.