# Algorithmic Anchoring: How Prompt-Embedded Reference Points Bias LLM Financial Estimates

John Garcia*

ORCID: 0000–0002-7062–0032

## Abstract

Large language models (LLMs) are increasingly used in financial analysis; however, it remains unclear whether numerical cues embedded in prompts systematically bias their outputs. This study conducts a controlled equity-valuation experiment using fictional companies that minimize firm-specific training-data priors and hold fundamentals constant. Across 51,000 API invocations of three commercial models (Claude Haiku 4.5, GPT-5-mini, and Gemini 2.5 Flash), I randomize prompt-embedded anchors and estimate how valuations and downstream outputs respond. Three findings emerge. First, source framing amplifies anchoring: labeling the same number as a prior AI estimate increases valuation sensitivity to the anchor by 52% (from slope 0.272 to 0.413), consistent with greater weight placed on a source-labeled prior rather than on greater numeric exposure. Second, anchoring propagates selectively: implied P/E ratios shift toward the anchor, buy/sell recommendations exhibit a sign reversal attributable to dual-channel contamination, and implied growth forecasts remain largely insulated. Third, model susceptibility differs materially: GPT-5-mini exhibits substantially weaker anchoring than Claude Haiku and Gemini across all tested conditions. These results imply that model choice and prompt framing are first-order determinants of anchoring risk in LLM-assisted finance tasks and should be empirically validated in practice.

*Keywords:* anchoring bias, large language models, behavioral finance, financial valuation, LLM agents, algorithmic bias

*JEL Classification:* G41, G14, D91, C90, C91

** John Garcia is an Associate Professor at California Lutheran University, School of Management, 60 West Olsen Road #3800, Thousand Oaks, CA 91360, USA. E-mail: jgarcia@callutheran.edu, Phone: 805–493-3814*

## 1. Introduction

Consider an equity screening pipeline in which a large language model (LLM) ingests a company profile — revenue, margins, multiples, and a narrative outlook — and returns a fair-value estimate in a structured JSON. The prompt that feeds the model inevitably includes numerical reference points: a 52-week high carried over from the data feed, a consensus price target from a previous analyst module, and a sector-median P/E embedded in the industry overview. If any of these figures systematically shifts the model's valuation output, downstream consumers may inherit a bias whose source is invisible in the final number. The risk is not hypothetical: 52-week highs shift human analysts' price targets [22], and industry-median EPS anchors sell-side earnings forecasts [9]. Whether LLMs are susceptible to this type of prompt-embedded anchoring in financial tasks — and if so, how strongly, through what channels, and with what sensitivity to prompt design — is an empirical question that this study addresses in a controlled experimental setting.

Anchoring, which refers to the disproportionate influence of an initial numerical reference point on subsequent judgment [42], is among the best-documented and most consequential biases in financial decision-making. It persists among experienced professionals [36], scales with anchor magnitude [28], and generates measurable wealth transfers through round-number trading alone [5]. Lou and Sun [33] recently demonstrated that large language models (LLMs) exhibit anchoring on general-knowledge estimation tasks to a magnitude comparable to that of human subjects. Whether anchoring of a similar magnitude occurs when LLMs perform financial valuations — a task with a richer informational structure, multiple competing reference points, and direct monetary consequences — remains an open empirical question that this study addresses.

I deploy a controlled equity-valuation experiment across three commercial LLMs (Claude Haiku 4.5, GPT-5-mini, and Gemini 2.5 Flash), using ten fictional companies to eliminate training-data contamination and 51,000 API calls across four hypothesis-testing batches to ensure precise estimation. Three findings extend prior work.

First, I document a source-framing amplification: labeling the anchor as a prior AI-generated estimate increases anchoring rather than attenuating it (net anchor slope 0.413 vs. 0.272). This pattern is inconsistent with a simple salience mechanism, in which the numeric anchor is neither repeated nor repositioned across personas, and anchor recall is not higher under the prior-estimate frame. The most defensible interpretation is that source labeling shifts the anchor from a generic numeric cue to an informational prior, increasing the weight the model assigns to it, even when the prompt instructs the model to remain independent.

Second, anchoring co-moves selectively across downstream outputs; implied P/E ratios and buy/sell recommendations shift with the anchor, whereas implied growth rates remain largely insulated. This selective propagation means that a single anchored valuation pass can contaminate multiple outputs, which end users may treat as independent signals.

Third, cross-model heterogeneity is substantial: GPT-5-mini's anchoring index (0.109) is one-third that of Claude Haiku (0.273) and Gemini (0.352), indicating that model selection is a material determinant of bias susceptibility—a result with immediate implications for procurement decisions in AI-augmented financial workflows.

These findings contribute to the behavioral finance literature on anchoring [2, 8, 9], the emerging literature on LLM cognitive biases [4, 24, 33], and the Homo Silicus research program [26] by identifying anchoring as a bias that transfers robustly from humans to LLMs in financial tasks, in contrast to affect-driven biases, where LLMs exhibit hyper-rationality [21, 25]. This

pattern is consistent with the disembodiment conjecture (Section 2.3): anchoring, which can operate through the selective accessibility of anchor-consistent information during text generation, transfers through a mechanism compatible with the transformer architecture, whereas biases requiring embodied affective processing do not.

Three design choices bound the scope of these contributions. First, fictional-company stimuli eliminate training-data contamination but also eliminate the models' training-data priors — prior analyst reports, price histories, and consensus targets that would provide competing reference points in real-company settings. The design therefore provides an upper bound on anchoring susceptibility: a clean test of whether the mechanism exists, but not an estimate of its magnitude under ecologically realistic conditions. Second, the one-shot, single-pass design represents a worst case; in deployed workflows, multi-turn interactions, human editing, and institutional reviews intervene between LLM output and investment actions, likely attenuating the documented effects. Third, the three models tested are specific commercial versions, not a random sample; generalization requires independent replication. The paper's contribution is the identification and characterization of a mechanism — prompt-embedded anchoring in LLM financial valuation — rather than an estimate of its realized impact in operational settings. Section 3.7 discusses these and additional design constraints in detail.

## 2. Literature Review and Hypothesis Development

### *2.1 The Psychology of Anchoring Bias*

Anchoring bias was first identified by Tversky and Kahneman [42] and has become one of the most robust phenomena in experimental psychology [19]. Two primary mechanisms have been proposed. The anchoring-and-adjustment model (often termed anchoring-and-insufficient-adjustment) posits that individuals begin with an anchor and adjust away from it serially,

terminating the adjustment prematurely once a plausible value is reached [17]. Under the selective accessibility account, the anchor triggers a positive hypothesis-testing strategy that increases the accessibility of anchor-consistent information and biases subsequent estimations [41]. Kahneman [29] later integrated anchoring into the dual-process framework, characterizing it as a System 1 phenomenon that resists System 2 corrections. This characterization has direct implications for LLMs: lacking the automatic, intuitive processing channel through which anchoring operates in humans, LLMs might be expected to resist anchoring; however, if anchoring can also arise from statistical patterns in training data and selective attention during text generation, the bias may transfer through a different mechanism.

Jacowitz and Kahneman [28] developed the anchoring index methodology adopted in this study and reported a mean index of approximately 0.49 across estimation tasks. Even experienced experts remain susceptible [36], although domain-specific knowledge provides some buffer [40], and accuracy motivation can increase adjustment when the direction of correction is known [39].

*Note on mechanistic interpretation.* Throughout this paper, I use terms from the human anchoring literature—"anchoring-and-insufficient-adjustment," "selective accessibility," and "cognitive anchoring"—as descriptive labels for observed patterns rather than as claims about the computational processes that produce those patterns. Whether LLM anchoring arises from attention-weighted token generation, training-data statistical patterns, or another mechanism remains an open question. The experimental design identifies the effect of prompt-embedded numerical cues on outputs; it does not identify the mechanism through which those cues operate. Where I invoke mechanistic accounts (e.g., stating that adversarial backfire is "consistent with" selective accessibility), this should be understood as noting pattern-level correspondence rather

than claiming that the same cognitive process operates in both humans and LLMs. The irrelevant anchor condition (office square footage) provides the strongest available test: if a domain-irrelevant number shifts financial estimates, this rules out rational information processing as the sole explanation, although it does not establish which specific non-rational mechanism is responsible.

### *2.2 Anchoring in Financial Markets*

The financial domain serves as a consequential testing ground for anchoring, as biased estimates translate directly into mispriced securities and misallocated capital. Cen, Hilary, Wei, and Zhang [9] provided extensive evidence that sell-side equity analysts anchor earnings forecasts on industry median EPS levels, with cross-sectional anchoring patterns that predicted both forecast errors and subsequent stock returns. Campbell and Sharpe [8] find analogous effects in macroeconomic forecasting, with professional forecasters anchoring on previous-period releases in approximately two-thirds of the indicators studied.

George and Hwang [22] demonstrated that the 52-week high price serves as a cognitive anchor: when good news pushes a stock near its 52-week high, investors are reluctant to bid the price above that level, leading to predictable underreaction. The proximity to the 52-week high dominates and subsumes the forecasting power of traditional momentum measures. Lasfer et al. [31] extended these findings by showing that corporate insiders exploit other investors' anchoring at 52-week extremes. The behavioral finance literature has also documented the role of round numbers as psychological price barriers, with Bloomfield, Chin, and Craig [5] estimating that round-number trading generates aggregate wealth transfers exceeding $850 million annually in U.S. equity markets. However, related evidence suggests that the strength of psychological price barriers can vary across markets and periods [14].

Anchoring extends to IPO pricing and corporate acquisitions [2]. In credit markets, anchoring on prior credit scores and default-rate benchmarks can bias risk assessments, a concern that has intensified as AI-augmented credit analysis has become more prevalent [6, 13, 18].

A critical distinction in the anchoring literature, central to this study's design, is between informational and cognitive anchoring. Informational anchoring occurs when an anchor has genuine diagnostic value (e.g., an analyst's consensus target may reflect private information). In such cases, the response to the anchor may be partially rational. Cognitive anchoring occurs when numerically irrelevant cue biases are estimated through purely psychological mechanisms. In financial contexts, many anchors—a 52-week high, a sector P/E multiple—occupy an ambiguous middle ground: they carry some real information but are known to exert disproportionate influence on their informational content. This study uses five anchor types that span the informational-to-cognitive spectrum, with an irrelevant anchor (office square footage) providing a clean test of pure cognitive anchoring and plausibly relevant anchors, testing whether large language models respond to financial reference points beyond their informational value, as human analysts do.

### *2.3 LLMs as Economic and Financial Agents*

Horton [26] introduced the "Homo Silicus" concept, proposing that LLMs could function as simulated economic agents. Subsequent research has revealed important boundary conditions: LLM agents fail to replicate human market traders, exhibiting hyper-rational behavior [21, 25], and laboratory market replications have yielded mixed results [10, 20].

These findings create a nuanced prediction for anchoring specifically. Chemero [11] argued that LLMs differ from human cognition because they are not embodied. Extending this argument, I propose that biases arising from statistical patterns in language should transfer to

LLMs, whereas affect-mediated biases may not — a prediction I term the 'disembodiment conjecture,' building on Chemero's core distinction. This conjecture identifies anchoring as a particularly informative test case: if anchoring operates through the selective accessibility of anchor-consistent information during text generation—a mechanism compatible with the transformer architecture's attention layers—then LLMs may exhibit it robustly, even when they fail to replicate other behavioral finance phenomena.

### *2.4 Cognitive Biases in LLMs*

Research has established that cognitive biases in LLMs are pervasive, yet selective. Binz and Schulz [4] demonstrated that LLM cognitive behavior is selectively human-like rather than uniformly human-like, and subsequent studies have confirmed that the incidence and magnitude of biases vary across tasks and models [3, 23, 24]. Particularly germane to the present study, Geva et al. [23] demonstrated that bias recognition attenuates bias expression in LLMs, a finding that directly motivates the manipulation-check design employed here. Two primary channels through which these biases arise have been identified: training data inheritance and reinforcement learning from human feedback (RLHF) amplification [3, 38].

The most relevant prior work is Lou and Sun [33], who conducted the first comprehensive experimental study of anchoring in LLMs. Across general knowledge questions employing factual, expert opinion, and irrelevant anchors, they estimated a mean anchoring index of 0.37, a magnitude comparable to the human benchmark of 0.49 [28]. LLM responses were highly sensitive to both anchor magnitude and direction, and standard mitigation strategies—chain-of-thought prompting, principle-based reasoning, and self-reflection— failed to attenuate the effect. Vidler and Walsh [43] further documented differential susceptibility across model families, with some models exhibiting negative anchoring. Within financial domains, Kim, Muhn, and

Nikolaev [30] demonstrated that while LLMs can perform financial statement analysis at levels comparable to professional analysts, they remain susceptible to normatively irrelevant presentational features. Winder, Hildebrand, and Hartmann [45] showed that LLM-generated investment advice increased portfolio risk across five dimensions, a tendency that debiasing strategies only partially offset.

### *2.5 Debiasing Strategies and Their Limitations*

Chain-of-thought (CoT) prompting has been widely adopted but shows limited effectiveness for cognitive debiasing. Lou and Sun [33] found that CoT did not significantly reduce anchoring, consistent with the finding that CoT improves logical reasoning but does not address selective accessibility and insufficient adjustment mechanisms that underlie anchoring. Explicit warnings have shown mixed effectiveness in the human literature: they can increase adjustment from self-generated anchors but have minimal impact on experimenter-provided anchors [17]. The provision of multiple information sources with opposing perspectives is theoretically grounded, as it counteracts selective accessibility. However, Dai, Kostopoulou, Bhatt, and Pan [15] found that targeted debiasing in one dimension can worsen bias in untargeted dimensions in up to 31.5% of cases, a "no free lunch" result with direct implications for this study's propagation hypothesis.

Iterative and self-debiasing approaches have shown promise [16, 34]; however, the consensus remains that single-intervention debiasing is insufficient. System-prompt persona manipulation offers an additional lever: Cheng et al. [12] found that large language model-generated personas exhibit predictable behavioral patterns, although persona sensitivity can inversely correlate with behavioral realism.

### *2.6 Cross-Dimensional Co-Movement of Anchoring Effects*

A relatively underexplored aspect of anchoring is the potential for bias to propagate from the primary to correlated secondary dimensions. In a financial analysis, a valuation estimate is linked to assumptions regarding revenue growth, earnings multiples, and buy/sell recommendations. If an LLM anchors when generating a price target, bias may propagate to these correlated outputs. Winder et al. [45] demonstrated that LLM investment biases manifest simultaneously across five risk dimensions. Li et al. [32] documented behavioral consistency issues in LLM trading simulations, in which biases introduced early in the decision sequence persisted through subsequent outputs. The autoregressive structure of modern LLMs, in which each token is conditioned on all preceding tokens, provides one plausible channel through which such propagation could occur; however, this study tests only whether propagation is present, not through which computational process it arises.

### *2.7 Hypothesis Development*

The convergence of evidence on human anchoring, financial market anomalies, and LLM cognitive behavior motivates the following hypotheses:

***H1 (Financial Anchoring):*** *LLM agents anchor their stock valuation estimates on numerical cues embedded in prompts.*

This hypothesis follows from the evidence in Sections 2.1–2.2, which shows that LLMs exhibit anchoring comparable to that of humans on general knowledge tasks [33] and that anchoring is pervasive in human financial judgment. An irrelevant anchor (office square footage) provides a stringent test for pure cognitive anchoring.

***H2 (Anchor Magnitude):*** *The anchoring effect is proportional to the distance between the anchor and the calibrated baseline estimate, consistent with the anchoring-and-insufficient-adjustment model.*

Proportionality is well established in human cognition [28] and has been confirmed in LLMs [33]. The design manipulates anchor magnitude across four batches (Table 2), with the primary H2 test examining convexity within the ±30 % range.

***H3 (Debiasing Effectiveness):*** *Prompt-based debiasing interventions, such as chain-of-thought reasoning, explicit warnings, adversarial framing, and multi-source synthesis, will vary in their effectiveness at reducing anchoring bias, and none will eliminate it. Some interventions may be counterproductive if they increase the salience or perceived informativeness of the anchor.*

This prediction is supported by Lou and Sun's [33] finding that CoT was insufficient to reduce anchoring. The direction of each intervention is theoretically ambiguous. CoT may rationalize anchor-consistent conclusions. Prior-estimate framing may increase the perceived diagnosticity of the reference point by presenting it as a model-generated 'prior' rather than a raw numeric cue; if so, the intervention can be counterproductive even when paired with an instruction not to defer. Multi-source synthesis may counteract selective accessibility by forcing consideration of competing perspectives. Therefore, the core prediction is not that "debiasing works" but that prompt interventions can shift the weight placed on the same numeric cue in either direction.

***H4 (Cross-Dimensional Co-Movement):*** *Anchoring in the primary valuation dimension (price target) will co-move with correlated secondary dimensions, including the implied P/E ratio, buy/sell recommendation, and implied growth rate. Co-movement may be selective*

*rather than uniform, with forward-looking, judgment-intensive metrics exhibiting greater susceptibility than backward-looking, fundamentals-constrained metrics.*

The propagation hypothesis is motivated by the autoregressive architecture of LLMs and empirical evidence of cross-dimensional bias spillover [15, 45]. Metrics constrained by historical data (e.g., implied growth rate) may be more resistant to anchor contamination than forward-looking judgment metrics (e.g., fair value and implied P/E). This study tests propagation via per-equation OLS with cluster-robust standard errors (Section 3.7.3).

All four hypotheses are stated in terms of the prompt protocol environment; they do not predict the magnitude of anchoring in deployed systems, which would require additional assumptions about the deployment architecture.

## 3. Data and Methodology

### *3.1 Research Design Overview*

This study employs an experimental design in which large language model (LLM) agents serve as subjects for controlled financial estimation tasks. The central identification strategy is the embedding of numerical anchors, which are reference points that systematically vary across conditions within otherwise standardized company profiles. By comparing estimates across anchor conditions while holding all other profile information constant, I isolate the causal effect of anchor type and magnitude on LLM financial judgments. The experimental protocol, including hypothesis specifications, anchor conditions, and batch execution plan, was fully documented prior to data collection and is available in the replication archive (see Data and Code Availability). What was randomized: The anchor condition (type, direction, and magnitude) was assigned independently across company–model–condition cells; within each cell, 50 repeated

API calls under production-style settings (T = 0.7, where supported; GPT-5-mini uses its native stochasticity under fixed reasoning_effort) provided the stochastic variation used for inference.

The experimental design employed a 3 (model) × 13 (anchor condition) architecture using fictional-company stimuli.

*Design Accounting.* The experiment comprised 13 anchor conditions used to test H1 and H2: a no-anchor control plus 12 anchor treatments varying in type and magnitude, distributed across four batches as enumerated in Table 2. For H3, four prompt-based debiasing interventions (chain-of-thought, explicit warning, adversarial framing, and multi-source synthesis) and a neutral persona control were added to the baseline analyst persona, yielding six persona conditions in total. In Batch 4, the primary inferential batch, these six persona conditions were fully crossed with three anchor directions (high, low, and no-anchor control), producing 18 unique experimental conditions per company–model pair and 540 total design cells (10 companies × 3 models × 18 conditions), implemented at approximately 50 API calls per cell.

Each model call represents a separate API invocation using temperature sampling. Deterministic outputs emerge at T = 0, whereas at the production temperature T = 0.7, the token variation generates the numerical heterogeneity used for inference. This variation should be interpreted as repeated draws from a stable prompt-conditioned response distribution, enabling the estimation of how the prompt shifts the distribution's central tendency, rather than treating any single completion as definitive.

The causal estimand is the shift in each model's valuation output induced by anchor variation, estimated from temperature-drawn realizations at T = 0.7. Because the three models tested are specific versions rather than a random sample from a population of LLMs, the results

characterize these models under the specified prompt protocol; generalization to other architectures or deployment settings requires independent replication.

### *3.2 Company Stimuli*

The primary stimulus set comprised 10 fictional companies (FC01–FC10) spanning five sectors. Because these firms have no presence in the models' training data, all numerical information is derived from the provided profile, eliminating training-data contamination and providing a clean upper-bound test of how much LLMs rely on provided figures versus independent prior knowledge. Each company profile provides a neutral narrative description, current stock price, shares outstanding, TTM revenue, EBITDA, net income, explicitly computed earnings per share (EPS), P/E ratio, EV/EBITDA, free cash flow yield, and market capitalization. All financial items are internally consistent.[1] FC02 (Cascadia BioTherapeutics), a pre-revenue, clinical-stage biotech with negative EBITDA and a meaningless P/E ratio, was excluded from all pooled analyses because standard multiple-based valuation is not applicable; its anomalous response patterns were examined separately in the results. FC05 (Pacific Coast Financial Group), a bank holding company with an undefined debt-to-equity ratio, is additionally excluded from all regressions using the Eq 2a fundamentals specification but is retained in the Eq 2b company fixed-effects specification (see footnote 4). Table 1 summarizes the stimulus set, and the complete financial profiles and narratives are provided in Appendix A.

*[Insert Table 1 about here]*

*Calibrated Baseline Estimates and Control-Cell Definitions.* Two related but distinct "control" concepts are used throughout this paper; this section defines both to avoid confusion.

[1] Calibration testing revealed that without an explicit EPS field, at least one model (Claude Haiku 4.5) substituted EBITDA per share, inflating estimates 72–198% for capital-intensive companies.

Calibrated baseline estimates are the pre-experimental central tendencies of each model's valuation output computed at the company × model level. To establish these baselines, 50 separate control-condition (no-anchor) API calls were made for each company–model pair at T = 0.7 during the calibration phase conducted before the primary experimental batches. The mean of these 50 calls served as the calibrated baseline. These baselines serve a single purpose: to construct anchor values for the experimental conditions (for example, a 52-week high ± 30% adjusts the calibrated baseline, not an external value or the current stock price). They were not used for the statistical normalization of the dependent variable.

Batch control cells were no-anchor trials embedded within each experimental batch. In batches 1 and 4 (the two batches containing control conditions; see Table 2), approximately 50 no-anchor API calls were collected for each design cell. In Batch 1, the control cells exist at the company × model level. In Batch 4, the primary inferential batch—control cells exist at the company × model × persona level because the debiasing persona is an additional design factor. The Batch 4 control cell means serve as the denominators for the normalized dependent variables used in the regression analyses: the company × model control mean is used for H1, H2, and H4 (Section 3.6.2, Eq. 2a-norm/2b-norm), while the company × model × persona control mean was used for H3 (persona-specific normalization). The Batch 4 company × model control means (pooled across personas) are numerically close to, but not identical to, the calibrated baselines because they are drawn from independent API calls collected at a later date; any small differences reflect the stochastic process described in Section 3.5.

### *3.3 Valuation Anchoring*

Each LLM instance receives one company profile and one valuation question, and each instance is assigned exactly one anchor condition. The dependent variable is the point estimate of fair value per share. Anchor conditions were organized around five anchor types spanning the informational-to-cognitive spectrum: (i) 52-week high/low, tested at ±30% (Batches 1 and 4) and ±60% (Batch 2); (ii) analyst consensus target at ±30% (Batch 2); (iii) round-number price threshold at ±30% (Batches 2 and 3); (iv) sector P/E multiple (Batch 3); and (v) an irrelevant anchor — office square footage — providing a clean test of pure cognitive anchoring (Batch 3). Including the no-anchor control, this yielded 13 distinct anchor conditions used to test H1 and H2 (Table 2 summarizes the batch structure into which these conditions were organized). The five debiasing personas described in Section 3.6 were crossed with anchor direction in Batch 4 as a separate design factor for H3.

*[Insert Table 2 about here]*

The anchor values for percentage-based conditions are computed relative to the calibrated baseline estimate for each company–model pair (Section 3.2), not the current stock price. This design choice warrants explicit justification, as it departs from the ecological structure of financial anchoring, in which reference points (52-week highs, consensus targets) are external market quantities rather than model-specific priors.

The baseline-relative construction is motivated by three identification considerations. First, it equalizes treatment intensity across models. If anchors were set at a common external value—say, the current price minus 30%—models with different baseline valuations would face different effective anchor distances, confounding anchoring susceptibility with baseline proximity to the anchor. The baseline-relative design ensures that each model faces the same

proportional deviation from its own prior, isolating the anchoring mechanism from cross-model baseline heterogeneity. Second, the design follows the calibrated-anchor methodology of Jacowitz and Kahneman [28], whose AnI—defined as the ratio of the estimate shift to the anchor distance from the unanchored baseline—presupposes anchors calibrated to the estimation distribution. Third, tying anchors to the current stock price introduces informational confounding: the market price reflects aggregate beliefs about fair value, and model responses to a price-derived anchor would conflate cognitive anchoring with rational updating toward a genuine informational signal. The baseline-relative design ensures that the anchor's only distinguishing feature across conditions is its numerical magnitude, not its informational content.

The trade-off is ecological distance: in deployed financial workflows, LLMs encounter market-derived anchors (52-week highs, consensus targets, sector medians) that are not calibrated to the model's priorA robustness analysis re-expressing anchors as deviations from the current stock price is reported as part of the robustness battery in Section 4.9 (check viii; $\beta = 0.288$, $p < 0.001$), and Batch 3 (Table 5) provides directional evidence that externally derived anchor types (sector P/E multiples, round-number thresholds) also produce significant anchoring.

The anchor position within the profile (beginning, middle, or end) is randomized deterministically for each company-condition pair using an MD5 hash of the concatenated company ID, condition label, and a fixed salt. The hash function and salt are documented in the replication code.

### *3.4 LLM Models and Stochasticity Control*

Three models are included, selected to represent comparable mid-tier capability levels across the three dominant commercial providers: Claude Haiku 4.5 (Anthropic; claude-haiku-4-5-20251001), GPT-5-mini (OpenAI; gpt-5-mini-2025-08-07), and Gemini 2.5 Flash (Google). For

brevity, I refer to these models hereafter as Claude Haiku, Gemini Flash, and GPT-5-mini, except where version specificity is required. The temperature is included as a within-model design factor during calibration, sweeping $T \in \{0.0, 0.3, 0.7, 1.0\}$ for Claude Haiku 4.5 and Gemini 2.5 Flash.[2] The production-setting temperature was $T = 0.7$. A $T = 0$ level enables a near-deterministic check.[3]

Commercial LLM APIs are subject to versioning and deprecation. Specific version strings identify the models tested in the present study; future API updates may alter model behavior. All experiments were run within a defined data-collection window, and version strings were recorded in the output data. All four batches were executed within February 20, 2026, and February 22, 2026, and version strings were verified to remain constant across the data-collection window. No API deprecation notices were received during this period.

### 3.5 Stochastic Structure of LLM Outputs: Calibration Evidence

This section uses calibration-phase data to characterize the stochastic structure of the LLM outputs at different temperature settings. Appendix D discusses the inferential implications of this structure, including what the statistical tests do and do not mean, and the hierarchy of inference methods used for each hypothesis.

*Near-perfect determinism at T = 0:* During calibration, Claude Haiku and Gemini Flash produced identical outputs for seven of eight companies at $T = 0$, confirming near-determinism. GPT-5-mini was excluded from the $T = 0$ comparison because its API does not expose a

[2] GPT-5-mini does not accept the temperature parameter; it operates at a fixed internal stochasticity level. Its reasoning_effort setting ("low" / "medium" / "high") serves as a stochasticity proxy. This asymmetry is documented in Appendix D.

[3] Even at $T = 0$, some model implementations may retain internal stochasticity sources (e.g., non-deterministic GPU computations). The $T = 0$ check verifies near-determinism rather than assuming perfect reproducibility.

temperature parameter. This determinism validates the stochastic-process interpretation, in which temperature-driven variation at T=0.7 operates around a fixed center.

*Temperature-injected variation at T = 0.7.* At a production temperature of T = 0.7, the output variation increases substantially for most company profiles; the median CV across companies rises to 0.039 for Claude Haiku and 0.013 for Gemini. However, the variation was not uniform. Two Gemini company profiles (FC03 and FC08) produce zero variance even at T = 0.7, indicating that for certain input configurations, the model's probability distribution over output tokens is sufficiently peaked, such that temperature scaling does not generate distinguishable outputs across 50 realizations. These zero-variance cases represent a boundary condition of the stochastic inference framework: for these company-model pairs, the output is effectively deterministic at all tested temperature levels, and any anchor-induced shift is directly observable without recourse to statistical inference.

The GPT-5-mini, which lacks a temperature parameter, was excluded from the T = 0 comparison. Its output variation under default settings (median CV = 0.048 across companies, as reported in Table 3B) is comparable to that of the other models at T = 0.7, suggesting that its internal stochasticity settings produce a similar magnitude of token-sampling noise.

*[Insert Table 3A about here]*

*Implications for the anchoring estimand.* The calibration results establish that anchoring operates on the deterministic component of the model's output: the model's fixed answer shifts toward the anchor, and temperature-driven variation at T = 0.7 adds noise around this shifted center without depending on the anchor condition (within-condition standard deviations are comparable across conditions; see Table 3B). Therefore, the sample mean from 50 API calls at T = 0.7 is a consistent estimator of the deterministic center.

Because the approximately 50 within-cell draws are exchangeable realizations around a shared deterministic center, I collapse each company × model × anchor direction × debiasing cell to its mean and use these cell means as the primary unit of analysis. In the Eq 2a fundamentals specification, the Batch 4 sample comprises 288 anchored cells (8 companies × 3 models × 2 anchor directions × 6 debiasing conditions; FC02 excluded by design, FC05 excluded due to missing debt-to-equity; see footnote 4) and 144 control cells (8 × 3 × 6). The Eq 2b company fixed-effects specification retains FC05, yielding 324 anchored and 162 control cells. This avoids inflating the degrees of freedom due to within-cell replication and improves the reliability of cluster-robust inference with G = 9 company clusters. As a validation, draw-level regressions yield numerically identical (or near-identical) estimates for the key anchoring and interaction coefficients; Appendix Table A6 summarizes the cell-mean versus draw-level comparisons and notes where small differences arise from aggregation.

*Interpretation of statistical tests.* Appendix D details the full inferential framework, including the assignment of primary inference methods to each hypothesis. Given the high statistical power of the design, effect size measures (AnI, Cohen's d, and regression coefficients) are prioritized throughout; the latter serve as screening statistics to confirm that the effects exceed the temperature-driven noise floor.

*Sample Size and Power.* At the cell mean level (288 anchored cells in Batch 4), statistical power was adequate for the documented effect sizes (AnI = 0.11–0.35, Cohen's d > 1.0). At the draw level (14,387 observations), p-values will be small, even for negligible effects, reinforcing the prioritization of effect-size measures.

### *3.6 Analysis Strategy*

#### *3.6.1 Anchoring Index*

The primary effect size measure is the anchoring index (AnI), which is computed following Jacowitz and Kahneman [28]:

$$AnI = (M_{high} - M_{low}) / (A_{high} - A_{low}) \quad (1)$$

where $M$ is the mean estimate under the high- or low-anchor condition, and $A$ is the anchor value. AnI = 0 indicates no anchoring, and AnI = 1 indicates perfect anchoring. The index is computed separately for each company–model combination and pooled with standard errors clustered at the company level. FC02 is excluded from the pooled calculations. AnI estimates are compared with the human benchmark of 0.49 [28] and the LLM general-knowledge benchmark of 0.37 [33]. I use "AnI" rather than the conventional "AI" abbreviation to avoid confusion with "artificial intelligence" in a study where both concepts appear throughout.

#### *3.6.2 Regression Analysis*

Two regression specifications were estimated, both operating on cell-level means (one observation per company × model × anchor direction × debiasing combination) as the primary unit of analysis. Within each cell, the dependent variable is the mean point estimate (or its normalized equivalent) across approximately 50 temperature-drawn realizations. The cell-mean approach aligns the unit of analysis with the stochastic-process estimand described in Section 3.5.1. Because within-cell variation is temperature-injected noise rather than informative heterogeneity, aggregating to cell means extracts the deterministic center that constitutes the treatment effect of interest. The first is a pooled OLS specification with company-level financial controls.

$$Estimate_{ijt} = \alpha + \beta_1 Anchor_{ijt} + \beta_2 Fundamentals_i + \beta_3 Model_j + \gamma Debiasing_{ijt} + \varepsilon_{ijt} \quad (2a)$$

where the subscripts *i*, *j*, and *t* indicate the company, model, and repetition, respectively. The anchor is the continuous anchor value (or indicator dummies for categorical conditions). Fundamentals are a vector of company financial metrics (log revenue, EBITDA margin, and net debt ratio)[4]. The model is a fixed effect for each LLM. The second specification replaces the fundamentals vector with company fixed effects to absorb all time-invariant company characteristics:

$$Estimate_{ijt} = \alpha + \beta_1 Anchor_{ijt} + \beta_3 Model_j + \gamma Debiasing_{ijt} + \delta_i + \varepsilon_{ijt} \quad (2b)$$

In the normalized variants of both specifications (denoted 2a-norm and 2b-norm in the results), the dependent variable is the percentage deviation of the model's estimate from its control condition mean, and the anchor regressor is similarly normalized. For analyses not involving persona variation (H1, H2, H4), normalization uses the company × model control mean: pct_dev_from_control = (Estimate − $\text{Control\_Mean}_{ij}$) / $\text{Control\_Mean}_{ij}$ × 100. For H3 analyses, normalization is performed using the Batch 4 persona-specific control cells (defined in Section 3.2). The dependent variable is $\text{pct\_dev\_persona}_{ijp}$ = ($\text{Estimate}_{ijpt}$ − $\text{Control\_Mean}_{ijp}$) / $\text{Control\_Mean}_{ijp}$ × 100, where $\text{Control\_Mean}_{ijp}$ is the mean of the approximately 50 no-anchor API calls for company i, model j, and persona p within Batch 4. The anchor regressor is similarly persona-specific. It is $\text{anchor\_dev\_pct\_persona}_{ijp}$ = (Anchor_Value − $\text{Control\_Mean}_{ijp}$) / $\text{Control\_Mean}_{ijp}$ × 100, where subscripts i, j, p, and t index the company, model, persona, and repetition, respectively. This persona-specific normalization ensures that intercept shifts attributable to the persona prompt itself—independent of any anchor—are absorbed into the

[4] FC05 (Pacific Coast Financial Group), a bank holding company, reports no debt-to-equity ratio (Table 1). Because the Eq 2a fundamentals vector requires net debt ratio, FC05 is excluded from all specifications using Eq 2a. The company fixed-effects specification (Eq 2b), which absorbs company-level characteristics without requiring explicit financial controls, retains FC05. Consequently, the Eq 2a and Eq 2b samples differ by the FC05 cells. Both specifications exclude FC02 by design. The effective Eq 2a sample comprises eight companies (FC01, FC03, FC04, FC06–FC10); the Eq 2b sample comprises nine (adding FC05).

baseline rather than confounding the debiasing interaction estimates. Because the Batch 4 design includes no-anchor control cells for each persona (approximately 50 repetitions per company × model × persona cell), the persona-specific baseline is estimable with adequate precision. Both normalizations resolve the cross-company price-level confound: a $5 shift in a $27 stock (18.5%) is not comparable to a $5 shift in a $75 stock (6.7%). The normalized specification puts all companies on a common percentage scale, enabling meaningful pooling.

The cell-mean analysis applies to analyses in which a normalized dependent variable is estimable, specifically, Batch 1 (baseline) and Batch 4 (debiasing), which contain control-condition cells. Batches 2 and 3, which employ wider anchor ranges without control conditions, are analyzed at the draw level using the unnormalized specification (Equation 2a) because the percentage-deviation dependent variable requires control-condition baselines. Therefore, the H2 magnitude hypothesis, which requires a wider anchor range in Batches 2–3, uses the draw-level specification.

*Key variables.* Raw dependent variable: fair value-per-share point estimate produced by the LLM (USD). Normalized dependent variable: Percentage deviation of the LLM estimate from its control condition mean, computed at the company × model level (pct_dev_from_control) for H1, H2, and H4, and at the company × model × persona level (pct_dev_persona) for H3. Normalized anchor regressor: percentage deviation of the planted anchor from the corresponding control-condition mean (anchor_dev_pct or anchor_dev_pct_persona). Anchoring index (AnI): ratio of the shift in the LLM's estimate toward the anchor to the total distance between the anchor and control mean, bounded by [0, 1] for responses between the control mean and anchor. Recommendation: ordinal variable coded 1 (strong sell) to 5 (strong buy), extracted from the LLM's narrative output.

The company fixed effect δi absorbs all cross-company variation, providing a more stringent within-company test of anchoring; however, it precludes examining how company characteristics moderate this effect. Temperature is included as a covariate only in model-specific regressions for Claude Haiku and Gemini Flash, given that GPT-5-mini does not expose this parameter. The standard errors are heteroscedasticity-robust and clustered at the company level in both specifications.

Dependence structure and clustering rationale. At the cell-mean level, the primary dependence structure is the within-company correlation (cells that share the same company profile). I cluster standard errors at the company level (G = 9, excluding FC02) with CR2 bias correction and Satterthwaite degrees of freedom [37] and supplement with exact randomization inference (5,000 permutations of anchor-condition labels within company). The cell-mean analysis eliminates within-cell dependence by construction, reducing the clustering problem to cross-cell within-company correlation. Conventional clustered standard errors are unreliable with only G = 9 company clusters. I address this in two ways. For the main effects (H1, H2), I report randomization inference (5,000 permutations of anchor-condition labels within company blocks), which requires no distributional assumptions and is exact for the experimental design. For the interaction effects (H3) and per-equation tests (H4), where the permutation distribution lacks resolution for interaction-sized effects with G = 9, I report CR2 bias-corrected standard errors with Satterthwaite degrees of freedom [37]. Table D1 in Appendix D summarizes the primary inference method for each hypothesis; concordance across alternative methods is documented in the appendix for completeness.

*3.6.3 Propagation Analysis*

The inferential target for H4 is whether prompt-embedded anchors propagate to each secondary output dimension individually, that is, whether the anchor coefficient is significant in the implied growth rate, implied P/E, and buy/sell recommendation equations when each is estimated separately. This per-equation estimand is the policy-relevant question: a financial practitioner needs to know which specific outputs are contaminated by anchoring, not merely whether some unspecified combination of outputs is jointly affected. This is tested by estimating separate ordinary least squares (OLS) regressions for each dependent variable, with company-level cluster-robust standard errors.

The pre-analysis protocol specified a seemingly unrelated regression (SUR) framework, which also tests the joint null that anchoring is absent from all secondary outputs simultaneously. A Breusch-Pagan test for cross-equation residual correlation is highly significant (LM = 14,823.64, df = 6, $p < .001$), confirming substantial shared variation across the four dependent variables and indicating that SUR offers efficiency gains over per-equation OLS [46]. Given this result, I report both per-equation OLS with cluster-robust standard errors (the primary framework because the per-equation estimand is the policy-relevant target) and SUR estimates as corroborative evidence. As shown in the results, the SUR anchor coefficients closely replicate the OLS estimates, confirming that the per-equation conclusions are robust to the choice of estimation framework.

The recommendation variable, collected from the JSON schema, has five levels (strong buy, buy, hold, sell, and strong sellFor the H4 propagation regression, recommendations are collapsed into three levels: buy = 1 (combining strong buy and buy), hold = 0, and sell = −1 (combining sell and strong sell). Note that this collapsed coding centers the scale on the neutral classification

(hold = 0), with positive values indicating bullish and negative values indicating bearish recommendations. This differs from the five-level ordinal coding (1 = strong sell to 5 = strong buy) used in the ordered probit specification (Section 4.6.2), where both scales preserve the same directional ordering — higher values correspond to more bullish classifications — but the collapsed scale is symmetric around zero, which simplifies the interpretation of regression coefficients: a negative anchor coefficient indicates a shift toward more bearish recommendations.

### *3.6.4 Manipulation Check*

I report anchor recall/awareness as a manipulation check and quantify recall without cluttering the main tables by summarizing it using a stratified 10% audit sample, while the underlying anchor_recalled field is recorded for all trials. Because eliciting anchor recall can increase anchor salience, I place the recall item after the primary outputs in the response schema and treat it as an audit measure rather than a task feature. A dissociation between anchor awareness and anchor-biased estimates would indicate that the model recognizes anchoring but cannot override it [23].

### *3.6.5 Practical Significance*

Beyond statistical significance, the documented percentage deviations from baseline can be translated into per-share valuation distortions and contextualized against real-world screening thresholds. Section 4.11 provides this translation.

## ***3.7 Scope, Limitations, and Design Choices***

This study prioritizes internal validity over external validity. Fictional-company stimuli and controlled anchor placement yield clean causal identification; however, this comes at the cost of

ecological distance from operational financial workflows. This section delineates the scope of the supported inferences and boundary conditions that constrain generalization.

The inferential target is the prompt-protocol-level treatment effect: does embedding a numerical cue in an otherwise standardized company profile shift the deterministic center of an LLM's valuation output? The answer is unambiguously yes, with magnitudes documented in Sections 4.2–4.6. However, the design does not estimate the realized impact of anchoring in deployed financial systems, which depends on factors that are not manipulated here: human oversight and editing of LLM outputs, multi-turn interaction and iterative refinement, portfolio-level diversification across multiple LLM calls, institutional review processes, and the correlation structure of anchor exposure across users and time. Each of these factors can attenuate (or, in the case of correlated anchor exposure, amplify) the effects documented in this controlled setting. Therefore, the results should be interpreted as establishing the existence and magnitude of a mechanism—prompt-embedded anchoring—rather than as an estimate of the market-level consequences.

The fictional company design is a deliberate identification choice, not merely a convenience. Real company profiles would introduce training data contamination: the models likely have encountered analyst reports, price histories, and consensus targets for any publicly traded firm, making it impossible to distinguish anchoring on the experimentally planted cue from anchoring on memorized information. The fictional stimuli eliminate this confound, providing a clean upper-bound test of how much LLMs rely on the provided numerical cues versus independent priors. The tradeoff is that the results may overstate anchoring relative to settings where the model holds strong priors from the training data, or understate it if familiarity with real financial patterns in the training data interacts with anchoring in nonlinear ways. Whether the patterns

documented here transfer to real-company valuation tasks is an empirical question that requires the external validity extension described in Section 5.

Second, the baseline-relative anchor construction (Section 3.3) prioritizes clean causal identification over ecological fidelity. In practice, financial LLMs encounter anchors derived from external market data—52-week highs, analyst consensus targets, and sector-median multiples—whose distance from the model's latent valuation prior is unknown to the prompt designer and varies across models. The present design fixes this distance at ±30%, enabling precise AnI estimation but causing the absolute anchor values to differ across models for the same company. Therefore, cross-model AnI comparisons should be interpreted as comparisons of susceptibility to equally proportioned deviations from each model’s prior, rather than comparisons of responses to a common external stimulus. The price-relative robustness analysis (Section 4.9, check viii) partially addresses this limitation. A full external-anchor design using market-derived reference points for real companies is the natural next step and is described in Section 5.

Third, the staged batch design introduces a comparability constraint: Batches 2 and 3 lack control cells; therefore, cross-batch AnI comparisons are directional rather than experimentally controlled ( Appendix Table A3). Fourth, the structured JSON response format may interact with anchoring by requiring a precise point estimate rather than allowing narrative hedging; however, the direction of this interaction is ambiguous.

Fifth, the analyst persona prompt may activate training-data patterns that interact with anchoring. The neutral-persona condition partially controls for this concern: Table 7 shows AnI values close to the baseline for Claude Haiku and GPT-5-mini under the neutral persona, but

substantially lower for Gemini (0.231 vs. 0.352), suggesting a model-specific persona–anchoring interaction that warrants further investigation.

## 4. Results and Discussion

### *4.1 Sample Description and Data Quality*

The staged program comprised four batches totaling 51,000 API calls across all designs. Of these, 50,203 (98.4%) were retained after parsing exclusions, and 493 were Winsorized at the 1st/99th percentiles (see Appendix Table A1 for batch-level details). Control cells are present only in batches 1 and 4; batches 2 and 3 contain only high/low (or high/neutral/low) anchor conditions, with no control cells.

Within Batch 4, Winsorization at the 1st and 99th percentiles within each company adjusted 257 observations (0.95% of the Batch 4 retained sample), bounding pathological estimates that ranged from $3.50 to $120.00 in the raw data. Table 3B presents descriptive statistics for Batch 4 control-condition (no-anchor) cells, which serve as the denominators for the normalized dependent variables in all subsequent regression analyses (distinct from the pre-experimental calibrated baseline estimates used to construct anchor values; see Section 3.2 for this distinction). The mean control estimates ranged from $10.10 (FC02, Cascadia BioTherapeutics) to $81.40 (FC06, Apex Cloud Infrastructure), with coefficients of variation ranging from 0.031 (FC03, Heartland Consumer Brands) to 0.330 (FC02). The relatively tight clustering of control estimates across most companies (CVs below 0.10 for eight of ten firms) indicates that the models yield stable baseline valuations, providing a reliable benchmark against which anchoring effects can be measured. FC02, the pre-revenue clinical-stage biotech, exhibited substantially higher dispersion, consistent with its exclusion from pooled analyses, as noted in Section 3.2. FC06 (Apex Cloud Infrastructure) shows the highest CV (0.128), consistent with its high-growth, high-

multiple profile. Company fixed effects absorb this baseline volatility; the anchor coefficient is virtually unchanged between the fundamentals-controlled and fixed-effects specifications (Table 4, Panels A and B).

*[Insert Table 3B about here]*

The manipulation check confirmed near-universal anchor recall: across the 16,182 anchored batch 4 trials, the overall recall rate exceeded 99%, with perfect recall (100%) in 15 of the 18 model × persona cells. The sole material departure was Claude Haiku under the explicit-warning persona, which dropped to 70.1% recall, an isolated model-specific effect attributable to the prompt's instruction to de-emphasize salient numbers. Crucially, the adversarial condition does not elevate the recall above the baseline for any model (Appendix Table B2, $\beta = -0.003$, $p = 0.173$), establishing that the adversarial backfire documented in Section 4.5 does not operate through increased anchor salience. Full recall rates by model and persona are reported in Appendix Figure B1, and persona-level regression is shown in Appendix Table B2.

*Determinism check.* Table 3A confirms near-perfect determinism at T = 0 (7/9 and 8/9 companies produce identical outputs across 50 runs for Claude Haiku and Gemini, respectively) and moderate temperature-injected variation at T = 0.7 (median CVs of 0.039 and 0.013).

*Roadmap.* The results are organized by hypothesis, with each subsection beginning with a header that identifies its primary batch, key tables, and robustness evidence, distinguishing confirmatory from exploratory analyses. The primary inference method for each test is presented in Table D1 (Appendix D). H1 (financial anchoring) and H2 (anchor magnitude) received primary confirmatory tests in Batch 4, with Batch 1 providing independent baseline confirmation, and Batches 2 and 3 supplying directional extensions at wider anchor ranges. H3 (debiasing effectiveness) and H4 (cross-dimensional co-movement) were estimable only in Batch

4, which was the sole batch containing both control cells and debiasing persona conditions. Cross-model heterogeneity (Section 4.3) and anchor-type heterogeneity (Section 4.3.1) were treated as secondary findings that contextualized the primary hypotheses. Directional asymmetry analysis (Section 4.10) and economic significance calculations (Section 4.11) are interpretive extensions.

### *4.2 H1: Financial Anchoring Effect*

*Primary evidence:* Batch 4. Primary tables: Table 4 (cell-mean regressions; Eqs. 2a–2b) and Appendix Table A2 (baseline AnI by batch/model).

*Robustness*: Batch 1 baseline, bootstrap CIs, randomization inference, and wild cluster bootstrap.

Baseline AnI values are positive across all four batches and all three models, with Batch 1 values closely matching the Batch 4 confirmatory values reported below (Appendix Table A2 reports batch-level AnI for Batches 1–3, which lack control cells and serve as directional corroboration rather than independent inferential evidence). The primary confirmatory analysis uses batch 4 data, in which Claude Haiku produced an anchoring index (AnI) of 0.273 (SD = 0.154), Gemini Flash produced 0.352 (SD = 0.205), and GPT-5-mini produced 0.109 (SD = 0.044).

With only G = 9 companies, it is important to verify that no single firm drives these means. Appendix Table A4b reports the company-level distribution (min, median, max) for each model. GPT-5-mini's compressed range (0.064–0.166) confirms uniformly modest anchoring across all nine companies. Claude Haiku and Gemini exhibited wider ranges (−0.022 to 0.517 and 0.090 to 0.636, respectively), but median values (0.316 and 0.282) tracked the means closely, indicating that no single outlier company dominated the aggregate. These AnI values indicate that models shift toward the anchor by 11%–35% of the anchor's deviation from the baseline, which are

economically meaningful distortions. Because anchors are calibrated to each model's baseline (Section 3.3), each model faces different absolute anchor values for the same company: a model with a higher baseline receives a higher absolute anchor. AnI normalizes this by expressing the response as a proportion of the anchor-to-baseline distance; however, this normalization assumes an approximately linear anchoring response function with respect to the anchor distance. The convexity documented in Section 4.4 suggests that this assumption is imperfect. Therefore, cross-model AnI comparisons should be interpreted as comparisons of susceptibility to equally proportioned deviations from each model's own prior, under an approximate linearity assumption, rather than as comparisons of responses to a common external stimulus. Section 4.3 discusses this qualification further.

*[Insert Table 4 about here]*

Bootstrap 95% confidence intervals (bias-corrected and accelerated [BCa], 2,000 replicates) further support these findings. The pooled AnI for Claude Haiku is 0.356 [0.292, 0.430], for Gemini Flash is 0.350 [0.306, 0.396], and for GPT-5-mini is 0.115 [0.102, 0.129]. None of the confidence intervals includes zero, confirming that anchoring is present across all models and debiasing conditions. The pooled bootstrap estimates aggregate across debiasing conditions; the baseline-only AnI values reported above are somewhat lower for Claude Haiku and Gemini because certain debiasing conditions (notably adversarial framing) amplify rather than attenuate anchoring, as discussed in Section 4.5.

To contextualize the magnitude of these effects, I compared the observed AnI values with two widely cited benchmarks: the human general-knowledge AnI of 0.49 [28] and the LLM general-knowledge AnI of 0.37 [33]. The LLM financial-domain values reported here (0.11–0.35) fall below both benchmarks. However, cross-study comparisons are limited by

simultaneous differences in domain, stimulus complexity, subject type, and response format. The lower financial-domain AnI values are consistent with domain modulation—the richer informational environment of a financial profile may partially constrain anchoring—but could equally reflect task-complexity effects, response-format constraints, or differences in the models tested. Therefore, the comparison is useful as a rough-order-of-magnitude calibration, confirming that financial LLM anchoring operates in a similar range to previously documented effects; however, it cannot support a precise claim that "finance anchoring is lower than general-knowledge anchoring" without designs that vary the domain while holding other task features constant.

The cell mean regression confirms H1. In the primary specification (Eq 2a, Batch 4, 288 cells; see footnote 4 for sample composition), the anchor coefficient is 0.281 (randomization inference $p < 0.001$, 5,000 permutations). The model fit at the cell mean level was strong. Equation (2a) yields $R^2 = 0.558$. When replacing the fundamentals vector with company-fixed effects (Eq. 2b), the full-model $R^2$ increases to 0.580, reflecting the additional variance absorbed by the company indicators. To isolate explanatory power net of the fixed effects, I also report the projected (within) $R^2$ from the fixed-effects regression: 0.468 (adjusted projected $R^2 = 0.460$; $F = 147.7$, $p < 0.001$), which captures the variation explained by anchor and model terms after being demeaned by company. The anchor treatment accounts for a large share of the variation in deterministic centers, with temperature-driven noise accounting for the remainder at the draw level. The company fixed effects specification yielded multiple R-squared (full model) = 0.468 and projected R-squared = 0.460 ($F = 147.7$, $p < 0.001$). Residual diagnostics revealed mild non-normality and heteroscedasticity, both of which were expected at this sample size and were addressed by cluster-robust standard errors. Elevated collinearity among financial controls (log

revenue GVIF = 12.66) does not affect the anchor coefficient (VIF = 1.42) and is absorbed by the company fixed effects specification.

### *4.3 Cross-Model Heterogeneity*

Cross-model susceptibility varies substantially; however, interpreting these differences requires important qualifications. Because the anchors are calibrated to each model's own baseline, the three models face different absolute anchor values for the same company (Table 1, Panel B). For instance, if Claude Haiku values a company at $82 and GPT-5-mini at $60, their respective high anchors are approximately $107 and $78. AnI normalizes for this by expressing the response as a fraction of the model-specific anchor-to-baseline distance; however, this normalization is exact only if the anchoring response function is linear at that distance. The convexity documented in Section 4.4 and the overshooting reported in Table 5 (AnI > 1 for certain anchor types) both indicate that the linearity assumption is imperfect, meaning that cross-model AnI differences may partly reflect differences in absolute anchor distance rather than pure differences in susceptibility.

With this caveat, pairwise comparisons of baseline AnI values (Welch t-tests with Benjamini-Hochberg correction) reveal that Claude Haiku and Gemini do not differ significantly from each other (difference = −0.079, t = −0.93, adjusted $p = 0.369$); both appear more anchored than GPT-5-mini (Claude Haiku vs. GPT-5-mini: difference = 0.164, t = 3.07, adjusted $p = 0.019$; Gemini vs. GPT-5-mini: difference = 0.243, t = 3.48, adjusted $p = 0.019$). The magnitude of these differences should be interpreted cautiously given the linearity qualification; however, the direction is consistent across all nine companies: GPT-5-mini's company-level AnI range (0.064–0.166; Appendix Table A4b) falls entirely below the median AnI for both Claude Haiku

(0.316) and Gemini (0.282), suggesting that the cross-model ordering is robust even if the precise magnitudes are not directly comparable.

Model-specific regression illuminated this heterogeneity. When the normalized specification is estimated separately for each model, the anchor coefficient is 0.361 for Claude Haiku (R-squared = 0.534), 0.371 for Gemini (R-squared = 0.618), and 0.128 for GPT-5-mini (R-squared = 0.304). The Gemini model shows not only the highest anchoring coefficient but also the most systematic relationship (highest R-squared), suggesting that its valuation process is more mechanically responsive to anchor information. GPT-5-mini's substantially lower coefficient and R-squared are consistent with its lower AnI and indicate that anchor information has a weaker influence on its valuation judgments. Figure 1 visualizes this heterogeneity: under the baseline (no debiasing) condition, Gemini Flash shows the steepest anchor-response slope, Claude Haiku an intermediate slope, and GPT-5-mini the flattest response, with all three models well below the perfect-anchoring diagonal.

*[Insert Figure 1 about here]*

Cohen's d effect sizes express the anchor-induced shift in units of within-condition standard deviations. Across the nine companies in the pooled sample (excluding FC02), the mean d for the high-anchor condition (relative to the control) is 2.99 for Claude Haiku, 4.50 for Gemini, and 1.04 for GPT-5-mini. Even GPT-5-mini’s d of 1.04—the smallest observed—indicates that the anchor shifts the model’s deterministic center by more than one full standard deviation of the temperature-driven output variation. For the low-anchor condition, the mean d values are − 2.45 (Claude Haiku), − 0.56 (Gemini), and − 0.19 (GPT-5-mini). The asymmetry in Gemini's effect sizes, much larger for high anchors than for low anchors, suggests a directional bias in this model's susceptibility, potentially reflecting an asymmetric adjustment process. GPT-5-mini's

near-zero low-anchor effect size (−0.19) indicates that its resistance to anchoring is particularly pronounced for downward anchors. The directional asymmetry is visually apparent in Figure 1, where all three model lines rise steeply from control to the high-anchor condition (+30%) but remain nearly flat from control to the low-anchor condition (−30%).

In the two-model subsample exposing the temperature parameter (Claude Haiku and Gemini), adding temperature as a covariate yields an anchor coefficient of $\beta = 0.367$ (SE = 0.037, $p < 0.001$). The increase relative to the three-model pooled estimate ($\beta = 0.282$) is attributable to restricting the sample to the two more anchor-susceptible models rather than to the temperature control itself. Within this subsample, the coefficients with and without the temperature covariate were identical.

#### *4.3.1 Anchor-Type Heterogeneity*

Whether the type of anchor—plausibly relevant financial cues versus domain-irrelevant numerical information—matters for anchoring magnitude is a central question in this study's financial-domain contribution. Batch 3 crossed seven anchor conditions across four anchor types with high/neutral/low directions to address this question using the irrelevant anchor (office square footage) as the within-batch reference. Because Batch 3 lacks control cells, these results are descriptive rather than confirmatory; the key findings are summarized in detail in Table 5.

*[Insert Table 5 about here]*

Two features of Table 5 warrant interpretation, as they involve AnI values outside the conventional [0, 1] range. First, the non-round financial controls and round-number thresholds produce high-anchor AnI values exceeding 1.0 (1.01 and 1.07, respectively). An AnI above 1.0 indicates overshooting: the model's mean estimate moves beyond the anchor value itself, not merely toward it. In financial terms, if a high anchor is set at $60 for a company with a $50

control-mean valuation, an AnI of 1.07 implies a mean-anchored estimate of approximately $60.70; the model not only anchors on the reference point but also extrapolates beyond it. Overshooting is documented in human anchoring literature, particularly when anchors are perceived as informative rather than arbitrary [44], and its appearance here for financially plausible anchor types (but not for irrelevant anchors) is consistent with that pattern. Second, the sector P/E multiple produces a negative high anchor AnI (−0.270), indicating a contrarian response: the model's estimate moves away from the anchor direction. In financial terms, this suggests that when presented with a high sector P/E multiple, the models interpret the cue not as a reference point to anchor toward, but as a signal of sector-level overvaluation, adjusting the target company's valuation downward—a response that is economically coherent even as it violates the standard anchoring prediction. This asymmetry between anchor types reinforces the finding that LLM anchoring in financial contexts is not a uniform phenomenon; the anchor's economic content interacts with the model's valuation reasoning, producing qualitatively different response patterns. Because Batch 3 lacks control cells, these findings are descriptive rather than confirmatory; they motivate targeted follow-up designs to formally test whether anchor informativeness moderates the direction of the anchoring response.

### *4.4 H2: Anchor Magnitude Proportionality*

*Primary evidence:* Batch 4 (convexity interaction).

*Robustness:* Batch 1 CR2 cross-validation.

Magnitude analysis requires wider anchor-range conditions in Batches 2–3, which do not contain control cells and therefore cannot be analyzed at the cell-mean level using the normalized dependent variable. The H2 results reported here use the draw-level specification; the

batch 4 cell-mean results confirm the anchor coefficient that constitutes the H1 baseline for magnitude comparisons.

H2 predicts that anchoring effects are proportional to anchor distance, consistent with human cognition. The primary finding is that anchoring exhibits a convex response within the ±30% range: the anchor's marginal influence increases with larger deviations from the baseline, and the net effect is positive across all observed anchor magnitudes. The primary H2 test used Batch 4 data, with evidence from Batches 2–3 extending to wider ranges (±60%).

In the regression specification, the main effect of anchor deviation is negative ($\beta = -0.112$, $p < 0.001$); however, this coefficient is evaluated at the counterfactual zero-deviation point. The net effect at any observed anchor distance is the sum of the main effect and the interaction term scaled by the absolute deviation, which is positive throughout the data range.

Evidentiary status summary. Confirmatory H2 evidence comes from the Batch 4 interaction specification within the pre-specified ±30% anchor range: the convex relationship is statistically robust under multiple inference methods (asymptotic, CR1S, company fixed effects) and is independently replicated in the Batch 1 cross-validation sample. Batch 2's ±60% extension data are directionally consistent with stronger anchoring at larger anchor distances, but are exploratory (post-hoc design, no control cells) and cannot support formal inference. The supported claim is that anchoring exhibits a stronger response at larger anchor distances within the prespecified ±30% range, with suggestive (non-confirmatory) directional evidence at wider ranges. The generic term "convexity" should be understood as referring to this specific finding rather than a precisely identified functional form across the full range of possible anchor magnitudes.

### *4.5 H3: Prompt-Based Mitigation and Amplification Tests*

*Primary evidence:* Batch 4 (only estimable batches). Primary tables: Table 6A (regression interactions, persona-specific normalization), Table 6B (intercept comparison), Table 7 (AnI by condition), and Figure 2.

*[Insert Table 6A about here]*

*[Insert Table 6B about here]*

All debiasing results reported in this section use cell-mean regressions (288 cells, Batch 4) with persona-specific normalization (Section 3.6.2), in which the dependent variable and anchor regressor are computed relative to the company × model × persona control-condition mean, and each cell represents the mean of approximately 50 temperature-drawn realizations. This specification simultaneously aligns the statistical unit with the deterministic center estimand, and absorbs persona-level baseline shifts that are independent of anchoring. Table 6A reports the primary cell-mean specification along with the draw-level coefficient for comparison, and Table 6B details the intercept differences across normalization methods.

The explicit warning condition is the only intervention that reduces the anchor coefficient in a statistically detectable manner. The warning interaction is −0.107 (CR2 SE = 0.030, Satterthwaite df = 6.95, $p = 0.010$; randomization-inference $p = 0.112$), implying a roughly 39% attenuation of the anchor's marginal effect relative to the baseline (net anchor coefficient = 0.272 − 0.107 = 0.165). The draw-level regression yields the same interaction estimate (−0.107; Appendix Table A5). A higher randomization-inference p-value does not provide evidence of an opposing result; it reflects the limited resolution of permutation-based inference for interaction-sized effects with only G = 9 company blocks. Consistent with this, the main anchor coefficient is strongly supported by both approaches (randomization $p < 0.001$), whereas interaction terms

generally fall below the power floor of the permutation distribution. Accordingly, CR2 inference with Satterthwaite degrees of freedom is treated as the primary framework for H3 interaction tests (Appendix D; Table D1).

Chain-of-thought prompting produces a positive interaction coefficient (0.055, CR2 SE = 0.026, Satterthwaite df = 6.98, p = 0.071), suggesting a marginal increase in anchor sensitivity. Because inference is sensitive to only G = 9 company clusters, I treat the CR2/Satterthwaite framework as the primary for interaction tests (Appendix D; Table D1).

Most strikingly, the adversarial condition, in which the system prompt reframes the anchor value as 'a prior AI-generated estimate that may reflect algorithmic bias' and instructs the model to critically evaluate it (see Appendix B for full prompt text), produces the largest increase in the anchor coefficient (interaction = 0.141, CR2 SE = 0.039, Satterthwaite df = 6.95, $p = 0.008$; randomization inference $p = 0.141$), raising the net anchor slope from 0.272 to 0.413—a 52% amplification of anchor sensitivity relative to the baseline condition.

The backfire is not attributable to differential numeric exposure (the anchor appears once in all conditions; Appendix Table B1) or heightened salience (anchor recall does not increase under adversarial framing; Appendix Table B2, $\beta = -0.003$, $p = 0.173$). Instead, the evidence supports a framing-induced shift in weighting, in which the “prior AI estimate” label alters anchor influence without changing anchor exposure or awareness. Two candidate mechanisms are consistent with this pattern. First, the label may activate training-data patterns in which prior model estimates carry genuine informational weight, effectively signaling higher diagnosticity. Second, the label may function analogously to the selective-accessibility mechanism [41], increasing the retrieval of anchor-consistent information during valuation reasoning. These accounts have distinct mitigation implications—training-data decontamination versus architectural intervention—but

distinguishing between them requires interpretability methods beyond the scope of this behavioral experiment. The robust practical finding is that framing a reference point as a prior model estimate can inadvertently increase its influence, even under explicit independence instructions.

This adversarial salience backfire was robust to cell-mean aggregation, persona-specific normalization, and CR2 small-cluster correction, and the draw-level coefficient was numerically identical (0.141). The adversarial intercept remains strongly significant at the cell level (−4.07, CR2 $p = 0.004$; Table 6B), confirming that this persona produces a genuine anchor-conditional level effect beyond its baseline shift. Adversarially prompted models produce lower overall estimates and are simultaneously more sensitive to anchor directions.

Multi-source synthesis showed a nonsignificant interaction (0.014, CR2 $p = 0.692$), and the neutral persona condition produced a nonsignificant interaction (−0.022, CR2 $p = 0.358$), indicating that neither amplified nor attenuated anchor sensitivity. Both results were identical in the cell-mean and draw-level analyses. The intercepts for both conditions were nonsignificant under persona-specific normalization (multisource: −0.57, $p = 0.723$; neutral: −0.45, $p = 0.761$; Table 6B), confirming that their pooled-normalization significance was a normalization artifact, as discussed in Section 4.5.1.

The net anchor coefficients ranged from 0.165 (warning; 61% of baseline) to 0.413 (adversarial; 152% of baseline). CoT raises the sensitivity to an estimated 120%, although this increase is marginally significant ($p = 0.071$). Multi-source and neutral conditions leave the sensitivity essentially unchanged. The rank ordering of debiasing effectiveness was preserved across the cell-mean and draw-level analyses.

The AnI values by model and debiasing condition, reported in Table 7 and visualized in Figure 2, illustrate the combined cross-model and cross-condition variations. For Claude Haiku, the adversarial condition produced the highest AnI (0.742), whereas the warning condition produced the lowest (0.109). For Gemini, CoT produced the highest AnI (0.486), and the neutral persona produced the lowest (0.231). The GPT-5-mini showed uniformly low AnI values across all conditions (range: 0.080–0.150), suggesting that its resistance to anchoring was robust to debiasing manipulations. As Figure 2 illustrates, the adversarial backfire effect was concentrated in Claude Haiku, whose AnI under adversarial framing exceeded the human benchmark of 0.49. Meanwhile, explicit warnings brought Claude Haiku's AnI below GPT-5-mini's baseline level, demonstrating that the same model can exhibit the widest range of anchoring susceptibility depending on the prompt design.

*[Insert Table 7 about here]*

*[Insert Figure 2 about here]*

### *4.6 H4: Cross-Dimensional Co-Movement of Anchoring Effects*

*Primary evidence:* Batch 4 (only estimable batches). Primary table: Table 8 (per-equation OLS; SUR corroborative). *Robustness:* BH-adjusted p-values and CR1S small-cluster correction.

The propagation analysis operates at the draw level (N = 16,182) to support the SUR framework's residual covariance estimation, using company × model normalization and CR1S small-cluster-corrected standard errors. Because H4 estimates a single anchor coefficient pooled across personas, company × model normalization was used without the intercept confound identified for H3.

H4 tests whether anchor-induced variation in the primary dimension (price target) co-moves with secondary dimensions (implied growth rate, P/E, recommendation). If selective accessibility

is the mechanism, anchoring should preferentially propagate to outputs that are less constrained by historical data.

*[Insert Table 8 about here]*

I code recommendations as buy = 1, hold = 0, sell = −1; thus, negative coefficients indicate a shift toward more negative (sell-directed) recommendations. The price target equation confirms the primary H1 result: anchor deviation significantly predicts valuation estimates (β = 0.277, SE = 0.019, $p < 0.001$). The implied P/E equation shows significant co-movement: a one-percentage-point increase in anchor deviation is associated with a 0.030-point increase in the implied P/E ratio (SE = 0.004, $p < 0.001$). Interestingly, although anchors strongly shift continuous valuation outputs (e.g., price targets), discrete recommendations move in the opposite direction: higher anchors are associated with more negative recommendations (β = −0.006, SE < 0.001, $p < 0.001$). This counterintuitive sign pattern—higher valuations paired with more bearish recommendations—could reflect a coding error, schema-extraction artifact, or substantively meaningful feature of LLM recommendation generation. Because this sign paradox has direct implications for practitioners' interpretation of LLM-generated buy/sell labels, I investigate it systematically in Sections 4.6.1–4.6.3, ruling out artifacts and decomposing the total effect into constituent channels.

*Coding verification*. The recommendation variable was extracted directly from the structured JSON response using the model's own labels (strong buy, buy, hold, sell, strong sell) and mapped to the numeric scale (5, 4, 3, 2, 1 for the five-level coding; 1, 0, −1 for the collapsed three-level coding). A manual audit of 500 randomly sampled observations confirmed the correct mapping: all "strong buy" labels mapped to 5 (or 1), and all "strong sell" labels mapped to 1 (or −1). The sign reversal was not a coding artifact.

Regarding the remaining H4 results in Table 8, the implied growth rate equation shows a marginally significant negative coefficient ($\beta = -0.0000257$, SE = 0.0000153, $p = 0.092$). The near-zero magnitude suggests that growth rate estimates are largely insulated from anchor effects, possibly because growth rates are more tightly constrained by historical financial data provided in the company profile. This partial insulation is noteworthy, suggesting that LLMs may compartmentalize certain analytical dimensions more effectively than others, with backward-looking metrics (historical growth) being less susceptible to anchoring than forward-looking or judgment-intensive metrics (fair value, P/E, buy/sell).

Taken together, these results reveal a more nuanced co-movement structure than simple unidirectional propagation. Anchoring propagates conventionally from price targets to implied P/E ratios (both shift in the anchor direction); however, the recommendation dimension exhibits a sign reversal that, upon decomposition (Section 4.6.3), reflects dual-channel contamination: a positive indirect effect through valuation partially offset and slightly dominated by a negative direct effect on the discrete classification label. This pattern extends prior findings on cross-dimensional bias spillover [15, 45] by documenting opposing-sign contamination channels within a single output pass, with the direct channel operating almost entirely on the buy-and-hold margin rather than producing dramatic bearish shifts (Appendix Table A8), a result not previously reported in the LLM bias literature. The practical implication is that correcting one output dimension (e.g., adjusting a price target for known anchoring) does not correct and may even mask the independent bias in a correlated dimension. The finding that recommendations are affected through a direct channel, independent of the valuation estimate, is particularly consequential for financial applications, where the buy/hold/sell label is often the most actionable output for end users.

The per-equation approach identifies specific outputs contaminated by anchoring, but does not test the joint null hypothesis that all secondary outputs are simultaneously unaffected. To address this complementary question, I estimate an SUR system across the four equations. The Breusch–Pagan test confirmed a substantial cross-equation residual correlation (LM = 14,823.64, df = 6, $p < .001$), justifying the SUR framework. The SUR anchor coefficients closely replicate the per-equation OLS estimates: price target $\beta = 0.277$ *(*$p < .001$), implied P/E $\beta = 0.030$ *(*$p < .001$), recommendation $\beta = -0.006$ *(*$p < .001$), and implied growth $\beta = -0.00003$ *(*$p = .097$). The selective co-movement pattern—significant for price target, implied P/E, and recommendation, but not significant for implied growth—is robust to the choice of estimation framework. The SUR framework confirms the sign and significance of the recommendation coefficient ($\beta = -0.006$, $p < .001$); however, because SUR does not condition on the model's implied upside, it captures the total (net) effect documented in Table 8 rather than the decomposed direct and indirect channels reported in Table 8C. The mediation decomposition in Section 4.6.3 should be consulted for the channel-level interpretation.

*Practitioner orientation.* Before proceeding to the decomposition, it is useful to state what a naive propagation model would predict and how the data depart from it. If anchoring operated solely through the valuation channel — higher anchors push fair value up, increase implied upside, and shift recommendations toward "buy" — the recommendation coefficient in Table 8 would be positive. Instead, it is significantly negative ($\beta = -0.006$, $p < 0.001$): higher anchors produce more cautious recommendations despite producing higher valuations. The decomposition in Sections 4.6.1–4.6.3 establishes that this paradox arises from dual-channel contamination. Through the indirect channel, anchoring does propagate as expected: higher fair

values increase implied upside, which shifts recommendations toward "buy" ($\beta_2 = 1.002$, $p < 0.001$.

### 4.6.1 Monotonicity of Recommendations in Implied Upside

To assess whether the sign paradox reflects internal inconsistency in the models' outputs, I compute implied upside as (fair_value_estimate − current_price) / current_price. For visualization, I bin implied upside into deciles using pooled breakpoints computed across all three anchor conditions; thus, each decile represents the same upside range regardless of the condition. Figure 3 plots the mean recommendation within each decile on the collapsed three-level scale (sell = −1, hold = 0, buy = 1). Section 4.6.2 re-estimates the relationship using the full five-level ordered outcome and yields the same qualitative pattern.[5]

Figure 3 presents the results. Four patterns emerge. First, at the lowest upside levels (decile 1, approximately −35% to 4% implied upside), all three anchor conditions converged near the "hold" classification. When the model's own fair value implies negligible or negative upside, anchor information is dominated by the fundamentals signal, and the recommendation is unaffected by anchor direction.

Second, at moderate upside levels (deciles 2–5), the three series separate in the expected direction: the low-anchor condition rises rapidly to near "buy," control settles around 0.75–0.78, and the high-anchor condition rises more slowly, trailing control by a moderate and approximately constant gap. The separation is directionally consistent with both the propagation and direct contamination channels.

[5] Construction details and robustness checks (alternative binning; five-level outcome) are reported in Appendix Tables A7 and A9. Table A7 re-estimates recommendations using the original five-level outcome, and Table A9 shows robustness to alternative binning and flexible controls for upside nonlinearity. Batch 4; FC02 excluded; N = 16,182.

Third, and most diagnostic for the sign paradox, at higher upside levels (deciles 6–9, approximately 13%–34% implied upside), the gap between the high-anchor and control series widens dramatically. The low-anchor series remains near 1.0, but both the control and high-anchor series decline — the control series moderately (from approximately 0.95 at decile 5 to 0.72–0.83 at deciles 8–9) and the high-anchor series steeply (from approximately 0.90 at decile 5 to 0.30 at decile 9). The decline in the control series indicates general recommendation conservatism at higher upside levels — a dome-shaped unconditional pattern in which models treat very high implied upside skeptically. However, the high-anchor series declines far more steeply than the control, producing a gap that widens from near zero at decile 5 to approximately 0.50 at decile 9 on the collapsed scale. This widening gap is the visual signature of anchor-relative recommendation logic: the models evaluate their fair value against the anchor as an implicit comparison point, and when the model produces a high fair value that nonetheless falls short of a high anchor, the shortfall triggers increasingly bearish recommendations, even as the implied upside relative to the current stock price continues to grow.

Fourth, at the extreme upper tail (decile 10, 34%–81% upside), all three series converge downward to approximately 0.62–0.65, effectively eliminating the gap between conditions. This convergence suggests that at sufficiently extreme upside levels, all models — regardless of anchor condition — apply a strong mean-reversion heuristic in their recommendation logic, discounting very high implied upside as likely to reflect estimation error rather than genuine undervaluation. The convergence also implies that the anchor-relative comparison mechanism has a boundary: when the model's fair value exceeds the anchor by a sufficiently large margin, the anchor loses its power as a comparison point.

*[Insert Figure 3 about here]*

This pattern rules out simple additive contamination of the recommendation label. The anchor effect on recommendations is not a constant downward shift; it interacts with the model's own valuation output, intensifying precisely where the anchor-relative shortfall is largest (deciles 6–9) and dissipating where the fundamentals signal overwhelms the anchor comparison (deciles 1 and 10). This interaction is consistent with the mediation results reported in Section 4.6.3, in which the direct anchor channel does not operate independently of the valuation channel but rather responds to the relationship between the model's fair value and the anchor value.

### 4.6.2 Ordered Probit Estimation of Recommendation Responses

The OLS specification in Table 8 imposes linearity on the recommendation variable and uses collapsed three-level coding. To verify that the sign paradox is robust to the functional form, I re-estimated the recommendation equation as an ordered probit model using the original five-level outcome (strong sell = 1, sell = 2, hold = 3, buy = 4, strong buy = 5).[6] The ordered probit coefficient on anchor_dev_pct is −0.020 ($z = -50.65$, $p < 0.001$), confirming the negative direction found in the OLS specification with high statistical reliability. The cutpoint structure reveals an additional finding: the buy | strong buy threshold is effectively infinite (raw estimate ≈ $3.78 \times 10^6$), indicating that the "strong buy" classification carries approximately zero probability across all conditions. The effective recommendation distribution is concentrated across four categories (strong sell, sell, hold, buy), with the anchor operating almost exclusively on the hold–buy boundary. This compression may itself interact with anchoring: if the probability mass available for upward recommendation shifts is structurally limited, anchor-induced downward shifts face less resistance than upward shifts. Higher anchors consistently shift the probability mass toward more cautious classifications. Appendix Table A8 translates the ordered probit

[6] Full ordered probit specification, including cutpoints and model fixed effects, is reported in Appendix Table A7, Panel A.

coefficient into marginal probability changes: a 10-percentage-point increase in anchor deviation reduces the probability of a "buy" classification by 5.4 percentage points and increases the probability of "hold" by 5.2 percentage points, with only 0.2 percentage points reaching "sell" and essentially zero reaching the extremes. Contamination thus operates almost entirely on the buy-hold margin — the actionable classification boundary in most institutional workflows — rather than producing dramatic bearish shifts that would be easily detected by human review. The sign reversal is not an artifact of collapsed coding or the linear functional form imposed by OLS; it is a robust feature of the models' recommendation-generation process.

### 4.6.3 Direct vs. Indirect Anchor Effects on Recommendations

The preceding analyses establish that the sign paradox is genuine and robust. A critical interpretive question remains: does the anchor affect recommendations through the valuation channel (an indirect effect, that is, true propagation), independently of the valuation channel (a direct effect, that is, direct contamination of the discrete label), or through both channels simultaneously? To decompose the total anchor effect, I estimate the following specification:

$$Rec_{ijt} = \alpha + \beta_1 \cdot Anchor\_dev\_pct_{ijt} + \beta_2 \cdot Implied\_upside_{ijt} + \beta_3 \cdot Model_j + \delta_i + \varepsilon_{ijt} \quad (3)$$

where $Implied\_upside_{ijt} = (FairValue_{ijt} - CurrentPrice_i) / CurrentPrice_i$ is the model's own implied upside for observation ijt, computed from its fair value output in the same trial. If the anchor operates solely through valuation (pure propagation), conditioning on implied upside should absorb the anchor effect, rendering $\beta_1$ non-significant. If the anchor independently contaminates the recommendation label (direct effect), $\beta_1$ will remain significant after controlling for the model's own valuation output.[7]

[7] Full regression results for both specifications are reported in Appendix Table A7, Panel B.

A methodological note on interpretation: Because the implied upside is generated endogenously within the same autoregressive pass as the fair-value estimate, the decomposition below identifies statistical channels through conditioning rather than causally distinct pathways through independent experimental manipulation. The "direct" and "indirect" labels should therefore be understood as descriptive decompositions of the total effect rather than as causally identified mediation pathways in the formal sense of Imai, Keele, and Tingley [27].

The implied upside coefficient ($\beta_2 = 1.002$, $t = 6.80$, $p < 0.001$) is positive and significant, confirming that the models' recommendation logic responds to its own valuation output in the expected direction: a higher implied upside produces more favorable recommendations. Critically, the anchor coefficient ($\beta_1 = -0.009$, $t = -17.56$, $p < 0.001$) remains highly significant and retains its negative sign after conditioning on implied upside. The anchor coefficient increases in absolute magnitude from −0.006 (anchor only) to −0.009 (anchor plus implied upside), exhibiting a classic suppression pattern: the total effect understates the direct contamination because it nets the positive indirect channel (higher anchor → higher fair value → higher upside → more favorable recommendation) against the larger negative direct channel. This rules out pure valuation-based propagation as the explanation for the recommendation result. Instead, the anchor exerts a direct effect on the recommendation label even after conditioning on the model's own valuation output, consistent with 'label-level' contamination in addition to valuation-level anchoring.

A concern with the linear mediation specification is that the relationship between the implied upside and recommendations is visibly nonlinear (Figure 3), and the linear control may leave a residual curvature that anchor_dev_pct absorbs. To address this, I re-estimate Equation 3 under three progressively flexible specifications: quadratic (adding implied_upside$^2$), cubic (adding

implied_upside² and implied_upside³), and a nonparametric specification replacing continuous upside control with decile fixed effects, which allows each upside bin an unrestricted intercept. The direct anchor coefficient attenuates modestly from −0.0089 (linear) to −0.0067 (decile FE), a 25% reduction, but remains highly significant across all specifications *($p < 0.001$* in every case; Appendix Table A9). The attenuation confirms that part of the linear specification's direct effect was due to unmodeled nonlinearity in the upside–recommendation mapping, but the surviving coefficient under the most aggressive nonparametric control establishes that the anchor independently shifts recommendations beyond what any flexible function of the model's valuation output can explain.

Decomposition reveals a dual-channel anchor contamination structure. Through the indirect (valuation) channel, higher anchors raise fair value estimates, which in turn push recommendations toward "buy" via the implied upside mechanism ($\beta_2 > 0$). Through the direct channel, higher anchors independently push recommendations toward more cautious classifications — primarily from "buy" to "hold" rather than toward outright "sell" ($\beta_1 < 0$; Appendix Table A8), consistent with anchor-relative recommendation logic in which the model hedges when its fair value falls short of a high anchor: the models appear to evaluate their own fair value against the anchor as an implicit comparison point, classifying a fair value that falls short of a high anchor as indicative of a "hold" or "sell" even when the implied upside relative to the current stock price has increased. The total effect observed in Table 8 ($\beta = -0.006$) is the net of these two opposing channels, with the direct negative effect dominating the indirect positive effect.

This dual-channel structure has three implications. First, for practitioners, correcting the valuation-level bias (e.g., adjusting a price target for known anchoring) would not eliminate the

recommendation-level bias because the latter arises through an independent mechanism. Second, for model developers, the finding that LLMs generate internally inconsistent output pairs—a higher fair value accompanied by a more bearish recommendation under the same anchor condition—indicates that the discrete classification head and the continuous valuation output are not fully coupled, even within a single autoregressive pass. Third, for H4 propagation hypothesis, the recommendation result is not conventional propagation (bias flowing from one output to a correlated output) but rather direct contamination of two output dimensions through partially independent channels, one of which produces a sign reversal. This distinction refines the cross-dimensional co-movement framework: "co-movement" should not be assumed to be unidirectional across all output types.

The interaction pattern visible in Figure 3, where the direct anchor effect on recommendations intensifies at higher upside deciles (6–9) rather than remaining constant, further suggests that the direct channel is not simply an additive contamination of the recommendation label but is modulated by the relationship between the model's own fair value and the anchor value. The distance between the model's estimate and the anchor, and not merely the presence of the anchor, determines the magnitude of recommendation distortion. The convergence of all three series at the extreme tails (deciles 1 and 10) identifies the boundary conditions of this mechanism. The anchor-relative comparison requires sufficient upside to activate (decile 1 shows no effect) and is overwhelmed when the model's fair value exceeds the anchor by a large margin (decile 10 shows convergence).

### *4.7 Cascadia BioTherapeutics (FC02): A Special Case*

FC02, a pre-revenue clinical-stage biotech, was excluded from all pooled analyses because its fundamentals are incompatible with standard multiple-based valuation. The FC02 patterns

described below are based on Batch 4 data; qualitatively similar inverted patterns appear in Batches 1 and 2, where FC02 data are available (a separate tabulation is available on request). Examined separately in Batch 4, FC02 exhibited anomalous anchoring patterns. For Claude Haiku, the mean high-anchor estimate ($8.50) is below the mean low-anchor estimate ($15.50), producing a negative implied AnI. Gemini shows weak positive anchoring (high = $13.00, low = $11.30), and GPT-5-mini shows a negative pattern similar to Claude Haiku (high = $9.18, low = $11.90).

These inverted patterns suggest that the models respond systematically to anchor information in FC02, but in the opposite direction from standard anchoring prediction: higher anchors produce lower estimates, and vice versa. The high dispersion in FC02 control estimates (CV = 0.330, compared to a median CV of 0.060 for the other nine companies) indicates that the models lack a stable valuation prior for this company, potentially amplifying the influence of non-standard response strategies. A post-hoc regression analysis, reported below, confirms that the inverted pattern is statistically robust, rather than an artifact of noise.

*Stock price anchoring hypothesis.* The stock price anchoring hypothesis—that models anchor on FC02's current price ($34.20) rather than the calibrated baseline ($10.18)—is not supported by the post-hoc regression analysis (footnote 7): model predictions cluster near the calibrated baseline ($10–$17 range) rather than the current price. The most parsimonious interpretation is that FC02's anomalous fundamentals (negative EBITDA, no meaningful P/E) disrupt the valuation heuristics that produce standard anchoring in the other nine companies.

### *4.8 Multiple Testing Correction*

The Benjamini-Hochberg correction was applied to the primary scalar hypothesis tests for H1, H2, and H3, treating the batch 4 hypothesis tests as the primary confirmatory family. Batch

1–3 AnI summaries and H1 extensions were classified as staged preliminary/robust evidence and were not included in this correction family. H4 (propagation) was evaluated in a separate multi-equation framework (Table 8) and is therefore reported outside this correction family. All three primary hypotheses survived the correction in batch 4: H1 (anchoring), adjusted $p < 0.001$; H2 (magnitude), adjusted $p < 0.001$; and H3 (debiasing), adjusted $p < 0.001$. The large margins between the raw and adjusted p-values indicate that the primary findings are not artifacts of multiple-hypothesis testing in the confirmatory batch 4 family.

### *4.9 Robustness Checks*

Seven robustness checks confirm the primary findings; details are provided in the supplementary analysis file, and Table 9 summarizes the checks that apply to each hypothesis. In brief, (i) winsorization at the 1st/99th percentiles produces near-identical coefficients; (ii) company fixed effects (Eq. 2b) yield anchor coefficients within 2% of the fundamentals specification; (iii) model-specific regressions confirm that anchoring is present in each model individually; (iv) CR2/CR1S bias-corrected standard errors with Satterthwaite degrees of freedom (df ≈ 3–8) confirm that all primary coefficients remain significant despite the small number of company clusters (G = 9); (v) randomization inference (5,000 permutations within company blocks) confirms the main anchor coefficient for H1 *(*$p < 0.001$) without reliance on distributional assumptions; for H3 debiasing interactions, randomization p-values are uniformly non-significant (range: 0.112–0.866), reflecting the limited resolution of permutation-based inference for interaction effects with G = 9 clusters rather than true null effects (Appendix D), and CR2 inference serves as the primary framework for those tests; and (vi) wild cluster bootstrap (Webb six-point, 9,999 iterations) yields $p = 0.005$ for H1, providing the most

conservative small-sample confirmation [7]. No primary coefficient changed in significance across any inference method.

To address the concern that baseline-relative anchor construction may inflate apparent anchoring by tailoring the treatment to each model's prior, I re-estimated the primary H1 specification by replacing the baseline-relative anchor regressor with a price-relative alternative: anchor_dev_pct_price = (Anchor_Value − Current_Price$_i$) / Current_Price$_i$ × 100, which expresses each anchor as a percentage deviation from the company's current stock price—an external, model-invariant quantity. Because the calibrated baselines generally exceed the current stock price (all ten companies' control-condition means exceed their current prices, reflecting a mild optimism bias in the models' unanchored valuations), the price-relative regressor attributes a wider percentage deviation to each anchor than the baseline-relative regressor; however, the two are nearly perfectly correlated across cells ($r = 0.980$), limiting the scope for divergent results. The price-relative anchor coefficient is 0.288 (CR2 SE = 0.026, Satterthwaite $p < 0.001$), compared with 0.281 under the baseline-relative specification—a negligible difference that confirms that the primary H1 finding is not an artifact of the model-specific anchor construction. The price-relative $R^2$ (0.611) marginally exceeds the baseline-relative $R^2$ (0.558), consistent with the current stock price providing a tighter common scaling factor across models than the model-specific baselines. Under the company fixed-effects specification (Eq. 2b), the price-relative coefficient is 0.130 (CR2 SE = 0.011, $p < 0.001$), again confirming within-company anchoring when treatment intensity is measured from an external reference point. This reanalysis does not fully replicate a design in which anchors are set at fixed price-relative levels ex ante (which would require new data collection with different absolute anchor values for each model), but it demonstrates that the primary finding is robust to re-normalizing the treatment variable around

an external financial quantity. Batch 3 provides complementary evidence from genuinely external anchor types: sector P/E multiples, round-number thresholds, and non-round financial controls—none of which are calibrated to model-specific baselines—produce positive overall AnI values of 0.536, 0.776, and 0.784, respectively (Table 5), confirming that anchoring generalizes beyond the baseline-relative construction.

### *4.10 Interpretive Extension: Directional Asymmetry in Anchoring*

This section reports an exploratory analysis of a pattern observed in the primary results; this was not a prespecified hypothesis test. A notable feature of the results is the asymmetry between high- and low-anchor effects, evident in Figure 1 as a steep rightward rise versus a near-flat leftward response across all three models. Across the primary sample, high anchors produce a mean valuation bias of +11.9% relative to the control, whereas low anchors produce only −0.4% (Table 1, Panel A). Three potential explanations merit investigation.

*Mechanical artifact.* If calibrated control means are systematically closer to low-anchor values than to high-anchor values, the low-anchor condition would produce smaller deviations by construction. To evaluate this possibility, I compute the distance from each company's control mean to its high and low anchors, expressed as a percentage of the control mean (these distances are reported in Table 1, Panel B). The results reveal heterogeneous distance ratios across models: Gemini Flash exhibits approximately symmetric distances (mean ratio = 0.97), Claude Haiku's control means are slightly closer to high anchors (mean ratio = 0.83), and GPT-5-mini's baselines are substantially farther from high anchors (mean ratio = 1.60). These cross-model differences in calibrated baselines confirm that purely descriptive comparisons of high-versus low AnI conflate anchor distance with anchor direction.

*Floor effect.* Models may have strong priors that prevent estimates from falling below certain thresholds, thereby creating asymmetric adjustment ranges. The lowest observed estimate across all 26,967 batch 4 predictions is $3.50, which is well above zero, suggesting that floor effects do not meaningfully constrain the low-anchor response.

*Genuine directional feature*. The asymmetry could reflect a property of the models' response to numerical cues, whereby adjustment from a high reference point is less complete than adjustment from a low reference point, paralleling patterns documented in human anchoring [17].

To adjudicate these explanations, I regress the raw point estimate on the absolute anchor distance (|anchor deviation|), a high-anchor indicator, signed anchor deviation, firm fundamentals, model fixed effects, and company fixed effects[8]. Because the signed deviation, absolute distance, and direction indicator span the space of the interaction between absolute distance and direction, the high-anchor indicator coefficient directly tests whether the anchor direction produces a residual effect after controlling for anchor magnitude. The coefficient of the high-anchor indicator is significant in both the primary sample ($\beta = -41.3$, CR1S $t = -6.81$, $df = 4.24$, $p = 0.002$) and batch 1 cross-validation ($\beta = -36.6$, CR2 $t = -6.75$, $df = 2.86$, $p = 0.008$). These results demonstrate that directional asymmetry persists after controlling for anchor distance, ruling out the mechanical-artifact explanation and supporting the interpretation that LLMs exhibit genuine directional features in response to anchoring cues.

[8] The three regressors — absolute anchor deviation, signed anchor deviation, and the high-anchor indicator — are related by the identity: signed deviation = (2 × high_indicator − 1) × |absolute deviation|. Consequently, the explicit interaction between |absolute deviation| and direction is perfectly collinear and cannot be estimated as a separate coefficient. The high-anchor indicator in this specification directly captures the directional asymmetry after partialling out absolute distance and signed deviation, and its significance test is equivalent to testing whether the interaction between distance and direction differs from zero.

### *4.11 Summary and Discussion*

The results provide strong support for H1 (financial anchoring) and H2 (magnitude proportionality), qualified support for H3 (debiasing effectiveness), and support for H4 (cross-dimensional co-movement tested via per-equation OLS targeting the policy-relevant question in which individual outputs are contaminated), with meaningful heterogeneity across secondary outputs. Debiasing interventions reduce anchoring across several specifications but do not eliminate it, which is consistent with the view that prompt-level safeguards are partial mitigations rather than complete solutions. Co-movement is strongest in implied valuation multiples and recommendations, whereas implied growth estimates appear less sensitive to anchors. Taken together, these findings indicate that anchoring in LLM financial judgments is both economically meaningful and can spill over into downstream decision dimensions, which users may treat as independent signals.

The three primary contributions previewed in the introduction — adversarial backfire, selective cross-dimensional propagation, and cross-model heterogeneity — are summarized below, alongside the secondary findings that emerged from the analysis.

*Economic magnitude illustration.* The estimated percentage deviations translate into concrete valuation distortions. In the baseline specification, the anchor sensitivity is 0.272; therefore, a +30% anchor implies an 8.2% upward shift in fair value relative to the unanchored control mean (0.272 × 30%). For a stock trading at $50 where the model's unanchored fair value is $55, this corresponds to approximately +$4.50 per share ($55.00 → $59.50). Under prior-estimate (adversarial) framing, the net anchor slope rises to 0.413, implying a 12.4% shift for the same +30% anchor — about +$6.80 per share in the same example — consistent with the 52% amplification in anchor responsiveness documented in Section 4.5. To translate per-share

distortions into a screening-pipeline context, consider an automated system that ranks hundreds of companies for portfolio inclusion based on LLM-generated fair-value estimates, consistent with the deployment scenario described in Section 1 and the automated-prediction framing in Agrawal, Gans, and Goldfarb [1]. If anchor-contaminated estimates systematically shift rankings — moving firms into or out of the top decile — the realized misallocation depends on portfolio size, turnover, and governance, parameters that this study does not estimate. The present results establish the magnitude of the input distortion: roughly 8%–12% of the valuation signal for a +30% anchor, which is sufficiently large to matter for threshold-based screening rules. For context, Bloomfield et al. [5] estimate that round-number price anchoring alone generates aggregate wealth transfers exceeding $850 million annually in U.S. equity markets, illustrating how anchoring frictions can scale to aggregate consequences, although the present design cannot quantify the propagation from LLM valuations to trading activity.

A methodological finding concerns the normalization of the debiasing effects. Persona-specific normalization—computing the dependent variable relative to each persona's own no-anchor baseline— revealed that two of the five debiasing intercepts (multi-source and neutral) that appeared significant under pooled normalization were artifacts of conflating persona-level baseline shifts with anchor-conditional effects (Table 6B). The interaction terms that capture changes in anchor sensitivity are robust to both normalization specifications (Table 6A). This distinction is substantively important, as it separates interventions that shift the level of LLM outputs (multi-source, neutral) from those that alter responsiveness to numerical cues (warning reduces it; adversarial and CoT amplify it). Future studies examining prompt-based debiasing should report persona-specific baselines to avoid conflating these two channels.

Table 9 summarizes the evidentiary basis for each hypothesis and secondary finding, mapping each claim to its primary batch, key tables, and the supporting robustness analysis.

*[Insert Table 9 about here]*

First, I document a novel adversarial backfire in which reframing the anchor as a prior AI estimate amplifies rather than attenuates anchoring, a robust empirical finding that is not attributable to differential numeric exposure (Appendix Table B1) or heightened anchor salience (Appendix Table B2). The mechanism through which the 'prior estimate' label increases anchor weight is underdetermined; candidate accounts include training-data informativeness priors and a process analogous to selective accessibility, with distinct implications for mitigation design (Section 4.5). I also find a directional asymmetry persisting after controlling for anchor distance ($p = 0.002$). Second, anchoring propagates selectively across correlated financial outputs; however, the propagation structure is not unidirectional: implied P/E ratios shift in the anchor direction (conventional propagation), while buy/sell recommendations exhibit a sign reversal attributable to dual-channel contamination, operating almost entirely on the buy-and-hold margin (Appendix Table A8: a 10pp anchor shift moves 5.4 pp of probability from "buy" to "hold," with negligible spillover to "sell"). The mediation decomposition (Section 4.6.3) shows that the anchor exerts a direct negative effect on the recommendation label ($\beta_1 = -0.009$, $p < 0.001$) that is independent of—and opposite to—its indirect positive effect through the valuation channel ($\beta_2 = 1.002$, $p < 0.001$). This dual-channel structure, in which the model generates a higher fair value yet assigns a more bearish recommendation under the same anchor condition, represents a novel form of LLM output inconsistency with direct practical implications for users who rely on discrete classification labels. Third, implied growth rates remain largely insulated from anchor effects. Fourth, cross-model heterogeneity is present and directionally consistent across all nine

companies: GPT-5-mini exhibits lower anchoring than Claude Haiku and Gemini under every tested condition. The precise magnitude of this difference (AnI of 0.109 vs. 0.273 and 0.352) should be interpreted cautiously because the baseline-relative anchor construction means each model faces different absolute anchor values, and the AnI normalization assumes approximate linearity that the convexity results in Section 4.4 suggest is imperfect (see Section 4.3). The rank ordering, however, is robust, indicating that model selection is a relevant consideration for anchoring susceptibility.

The primary results reported here are based on batch 4 data, which contains both control cells and debiasing persona conditions and is the sole batch in which all four hypotheses are estimable. Batches 1–3, which lack control cells by design, provide directional corroboration; batch-level AnI values are reported in Appendix Table A2.

## 5. Conclusion

The most striking finding of this study is that a debiasing strategy can backfire: labeling a prompt-embedded anchor as a prior AI estimate amplifies anchoring by 52% (net anchor slope 0.413 vs. 0.272), even under explicit independence instructions. This adversarial backfire is robust to controls for mechanical repetition and anchor salience (Appendix Tables B1, B2), and its mechanism—whether training-data informativeness priors or a process analogous to selective accessibility—remains underdetermined. Only explicit bias warnings significantly reduce anchoring (by approximately 39%). These results emerged from a controlled equity-valuation experiment across three commercial LLMs (Claude Haiku, GPT-5-mini, and Gemini Flash) using fictional companies to eliminate training-data contamination, with 51,000 API calls providing the statistical foundation.

In addition to the backfire, two additional findings extend prior work. First, anchoring propagates selectively through distinct channels: implied P/E ratios shift conventionally with the anchor, but buy/sell recommendations exhibit a sign reversal attributable to direct contamination of the classification label, independent of the valuation channel. The mediation decomposition (Section 4.6.3) confirms that the anchor exerts a direct negative effect on the recommendation ($\beta_1 = -0.009$, $p < 0.001$) that is independent of and opposite to its indirect positive effect through valuation ($\beta_2 = 1.002$, $p < 0.001$). This contamination operates almost entirely on the buy-hold margin—a subtle downgrade far more likely to survive institutional review than a dramatic bearish reversal (Appendix Table A8). Implied growth rates remain largely insulated. Second, cross-model heterogeneity is substantial and directionally consistent: GPT-5-mini's anchoring index (0.109) is one-third of Claude Haiku's (0.273) and Gemini's (0.352) across all tested conditions, indicating that model selection is a material lever for bias mitigation.

Three practical principles emerge. Model selection: cross-model differences in anchoring susceptibility are substantial and directionally consistent, suggesting that targeted fine-tuning or model-choice audits may be feasible levers for bias reduction. Prompt hygiene: numerical reference points embedded in prompts — prior estimates, consensus targets, and sector multiples — warrant deliberate management; source labeling can increase rather than decrease anchor influence, as the adversarial backfire demonstrates. Output independence: multiple outputs from the same LLM pass should not be treated as independent signals. Practitioners who correct a price-target bias should not assume the accompanying recommendation label has been implicitly corrected; the mediation evidence indicates that it has not.

These results refine the Homo Silicus framework [26] and are consistent with the disembodiment conjecture (Section 2.3) that anchoring, a bias plausibly mediated by statistical

patterns in language, transfers robustly to the financial domain, whereas affect-driven biases do not [21, 25]. For regulators monitoring AI-augmented financial workflows [13, 18], the documented mechanism is relevant in cases where deployed systems expose models to anchor-like information without adequate safeguards, although the controlled setting cannot estimate the realized market impact.

Several boundary conditions apply. The results characterize three specific model versions: the fictional-company stimuli and one-shot design prioritize internal validity. The design maps most directly to automated screening pipelines that process company filings without human-in-the-loop review [1]. The most important next step is an external-validity test using real-company profiles with market-derived anchors (actual 52-week highs from CRSP, consensus price targets from I/B/E/S, and sector-median multiples from Compustat), which would establish whether training-data priors attenuate, amplify, or preserve the documented effects. Three planned extensions will test generalizability along the task dimension: a two-turn within-subjects design measuring anchoring persistence after explicit correction, earnings forecasting connecting to the analyst anchoring literature [8, 9], and credit risk assessment with direct regulatory implications [13, 18]. Together, these will establish whether the patterns documented here represent a general feature of LLM numerical cognition or are specific to the equity-valuation domain.

## Data and Code Availability

Data were collected in February 2026. Replication data and code are available at https://github.com/garcijo4/algorithmic_anchoring. The archive includes the experimental R code (analysis_pipeline.R), prompt templates, API audit metadata (excluding secrets), and the processed results file used in the manuscript. Analyses were conducted in R 4.5.2, using, among others, the tidyverse, broom, sandwich, lmtest, boot, systemfit, lfe, and car packages, as well as the clubSandwich, fwildclusterboot, and moments packages; exact package versions and session information were documented in the archive.

For the included trial runs (data generation), the end-to-end experiment runtime was approximately 48 h, and the total API costs were $58.50 USD (Anthropic: $21.15; Google: $17.43; OpenAI: $19.92). This study was not formally pre-registered in a trial registry (e.g., OSF, AsPredicted). The experimental protocol, including all hypotheses, anchor conditions, batch execution plan, and analysis strategy, was documented before data collection began and is included in the replication archive. The protocol document was timestamped using the repository's version control history. Deviations from the initial protocol (e.g., the use of per-equation OLS as the primary H4 framework with SUR as corroborative evidence, the expansion from three to four batches, and the redesign of batches 2–3 relative to the original plan) are documented in this manuscript.

## Statements and Declarations

**Competing Interests:** The author declares no financial or non-financial competing interests, either direct or indirect, related to the work submitted for publication.

**References**


[1] Agrawal, A., J. Gans, and A. Goldfarb, 2024, *Prediction Machines, Updated and Expanded* (Harvard Business Review Press).

[2] Baker, M., X. Pan, and J. Wurgler, 2012, The effect of reference point prices on mergers and acquisitions, *Journal of Financial Economics* 106, 49–71.

[3] Bini, P., L. W. Cong, X. Huang, and L. J. Jin, 2025, Behavioral economics of AI: LLM biases and corrections, Working paper, SSRN 5213130.

[4] Binz, M., and E. Schulz, 2023, Using cognitive psychology to understand GPT-3, *Proceedings of the National Academy of Sciences* 120, e2218523120.

[5] Bloomfield, R., M. Chin, and R. Craig, 2024, The allure of round number prices for individual investors, Working Paper, Georgetown CRI.

[6] Brookings Institution, 2024, AI and algorithmic bias in credit markets, Brookings Institution Report.

[7] Cameron, A. C., J. B. Gelbach, and D. L. Miller, 2008, Bootstrap-based improvements for inference with clustered errors, *Review of Economics and Statistics* 90, 414–427.

[8] Campbell, S. D., and S. A. Sharpe, 2009, Anchoring bias in consensus forecasts and its effect on market prices, *Journal of Financial and Quantitative Analysis* 44, 369–390.

[9] Cen, L., G. Hilary, K. C. J. Wei, and J. Zhang, 2013, The role of anchoring bias in the equity market: Evidence from analysts' earnings forecasts and stock returns, *Journal of Financial and Quantitative Analysis* 48, 47–76.

[10] Chanona, R., M. Pangallo, and C. Hommes, 2025, Can generative AI agents behave like humans? Evidence from laboratory market experiments, Working paper, arXiv:2505.07457.

[11] Chemero, A., 2023, LLMs differ from human cognition because they are not embodied, *Nature Human Behaviour* 7, 1828–1829.

[12] Cheng, M., Y. Huang, X. Wang, and L. Zhang, 2025, LLM generated persona is a promise with a catch, in *Advances in Neural Information Processing Systems* 38.

[13] Consumer Financial Protection Bureau, 2024, Supervisory highlights: Artificial intelligence, CFPB Report.

[14] Cyree, K. B., D. L. Domian, D. A. Louton, and E. J. Yobaccio, 1999, Evidence of psychological barriers in the conditional moments of major world stock indices, *Review of Financial Economics* 8, 73–91.

[15] Dai, Y., O. Kostopoulou, G. Bhatt, and W. Pan, 2026, No free lunch in LLM cognitive bias mitigation: A multi-task debiasing evaluation, *PLOS Digital Health* (forthcoming).

[16] Echterhoff, J., Y. Liu, A. Alavi, and J. McAuley, 2024, Cognitive bias in decision-making with LLMs, *Findings of the Association for Computational Linguistics: EMNLP 2024*.

[17] Epley, N., and T. Gilovich, 2006, The anchoring-and-adjustment heuristic: Why the adjustments are insufficient, Psychological *Science* 17, 311–318.

[18] Federal Reserve Board, 2025, Financial stability implications of generative AI, FEDS Working Paper No. 2025-090.

[19] Furnham, A., and H. C. Boo, 2011, A literature review of the anchoring effect, *Journal of Socio-Economics* 40, 35–42.

[20] Gao, Y., D. Lee, G. Burtch, and S. Fazelpour, 2024, Take caution in using LLMs as human surrogates: Scylla ex machina, Working paper, arXiv:2410.19599.

[21] Garcia, J., 2025, Homo Silicus is Hyper-Rational: Why LLM Agents Fail to Replicate Attention-Driven Trading. Available at SSRN: https://ssrn.com/abstract=5901742.

[22] George, T. J., and C.-Y. Hwang, 2004, The 52-week high and momentum investing, *Journal of Finance* 59, 2145–2176.

[23] Geva, T., A. Goldstein, E. Lary, and C. Levy, 2025, Do LLMs exhibit human-like cognitive biases? A large-scale systematic evaluation, Working paper, SSRN.

[24] Hagendorff, T., S. Fabi, and M. Kosinski, 2024, (Ir)rationality and cognitive biases in large language models, Royal Society Open *Science* 11, 240255.

[25] Henning, T., S. M. Ojha, R. Spoon, J. Han, and C. F. Camerer, 2025, LLM agents do not replicate human market traders: Evidence from experimental finance, Working paper, arXiv:2502.15800.

[26] Horton, J. J., 2023, Large language models as simulated economic agents: What can we learn from Homo Silicus? NBER Working Paper No. w31122.

[27] Imai, K., L. Keele, and D. Tingley, 2010, A general approach to causal mediation analysis, Psychological Methods 15, 309–334.

[28] Jacowitz, K. E., and D. Kahneman, 1995, Measures of anchoring in estimation tasks, *Personality and Social Psychology Bulletin* 21, 1161–1166.

[29] Kahneman, D., 2011, *Thinking, Fast and Slow* (Farrar, Straus and Giroux, New York).

[30] Kim, A., M. Muhn, and V. Nikolaev, 2024, Financial statement analysis with large language models, Working paper, arXiv:2407.17866.

[31] Lasfer, M. A., et al., 2024, Corporate insiders' exploitation of investors' anchoring bias at the 52-week high and low, *Financial Review* 59, 147–170.

[32] Li, Z., G. Wan, K. Chen, et al., 2026, Behavioral consistency validation for LLM agents, Working paper, arXiv:2602.07023.

[33] Lou, J., and Y. Sun, 2025, Anchoring bias in large language models: An experimental study, Journal of Computational Social *Science* 9, 11.

[34] Lyu, Y., 2025, Self-adaptive cognitive debiasing for large language models in decision-making, Working paper, arXiv:2504.04141.

[35] MacKinnon, J. G., and M. D. Webb, 2018, The wild bootstrap for few (treated) clusters, *The Econometrics Journal* 21, 114–135.

[36] Northcraft, G. B., and M. 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.

[37] Pustejovsky, J. E., and E. Tipton, 2018, Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models, *Journal of Business and Economic Statistics* 36, 672–683.

[38] Schramowski, P., et al., 2022, Large pre-trained language models contain human-like biases of large scale, *Nature Machine Intelligence* 4, 258–268.

[39] Simmons, J. P., R. A. LeBoeuf, and L. D. Nelson, 2010, The effect of accuracy motivation on anchoring and adjustment, *Journal of Personality and Social Psychology* 99, 917–932.

[40] Smith, A. R., and P. D. Windschitl, 2015, Resisting anchoring effects: The roles of metric and mapping knowledge, *Memory and Cognition* 43, 1071–1084.

[41] Strack, F., and T. Mussweiler, 1997, Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility, *Journal of Personality and Social Psychology* 73, 437–446.

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

[43] Vidler, A., and T. Walsh, 2025, Evaluating binary decision biases in large language models: Implications for fair agent-based financial simulations, Working paper, arXiv:2501.16356.

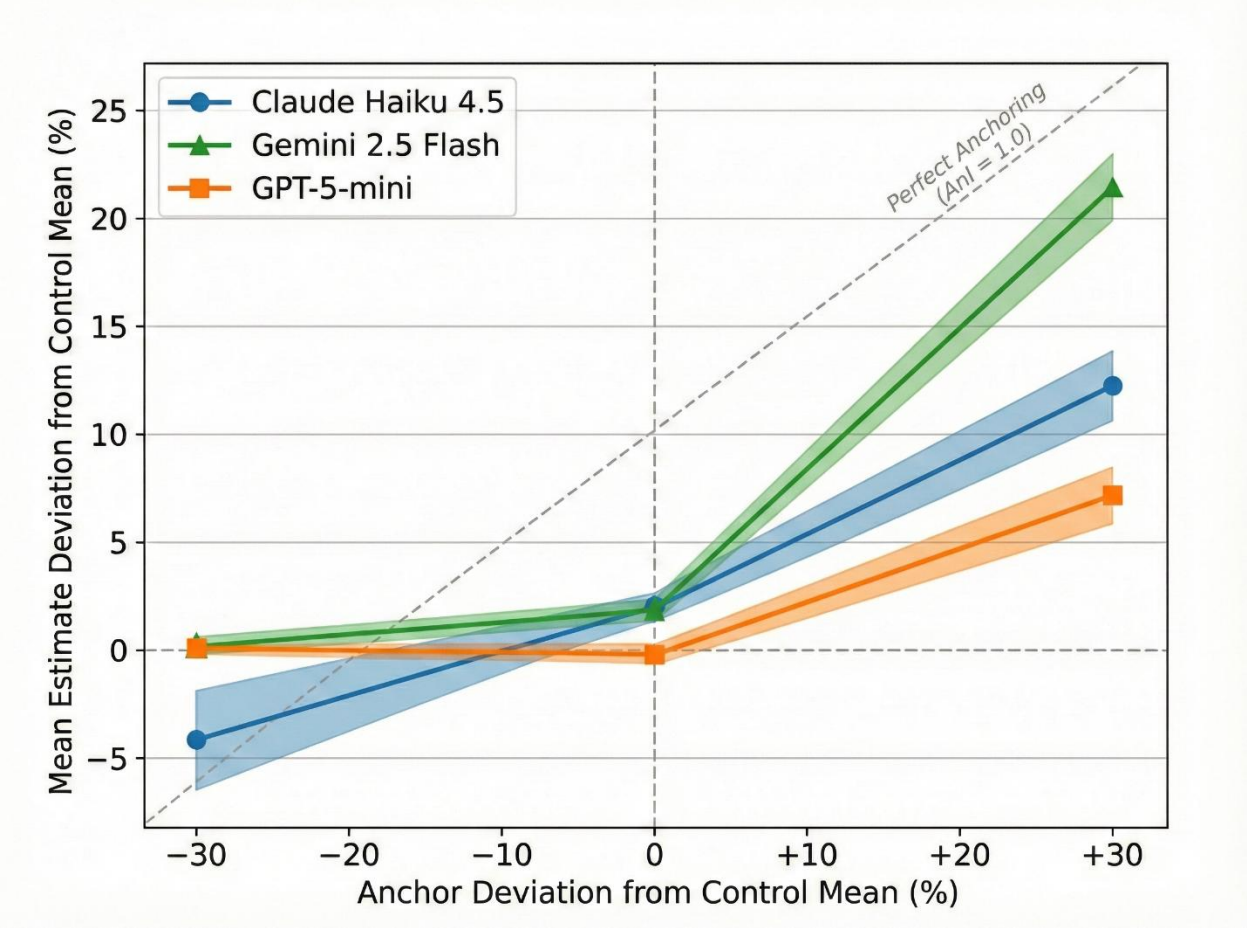
[44] Wegener, D. T., R. E. Petty, K. L. Blankenship, and B. T. Detweiler-Bedell, 2010, Elaboration and numerical anchoring: Implications of attitude theories for consumer judgment and decision making, Journal of Consumer Psychology 20, 5–16.

[45] Winder, P., C. Hildebrand, and J. Hartmann, 2025, Biased echoes: Large language models reinforce investment biases and increase portfolio risks of private investors, *PLOS ONE* 20, e0325459.

[46] Zellner, A., 1962, An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias, *Journal of the American Statistical Association* 57, 348–368.

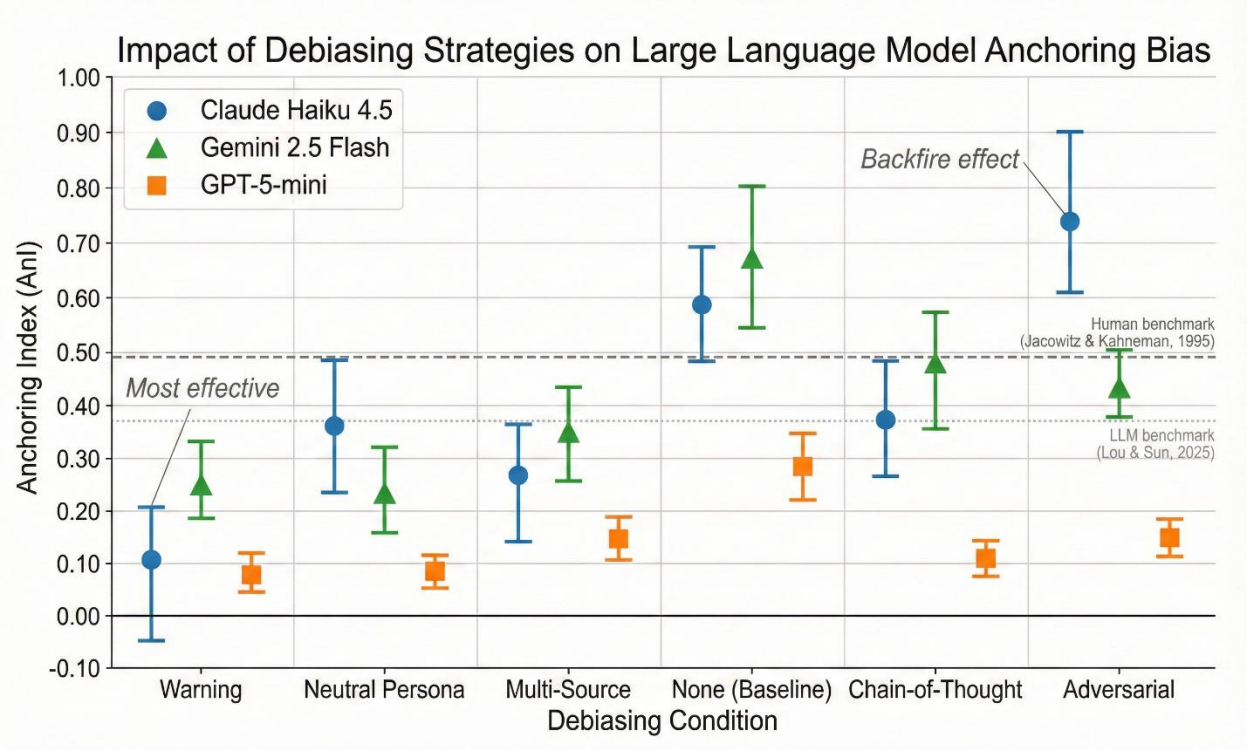
## Figures

**Figure 1. Anchoring Response by Model and Anchor Direction**



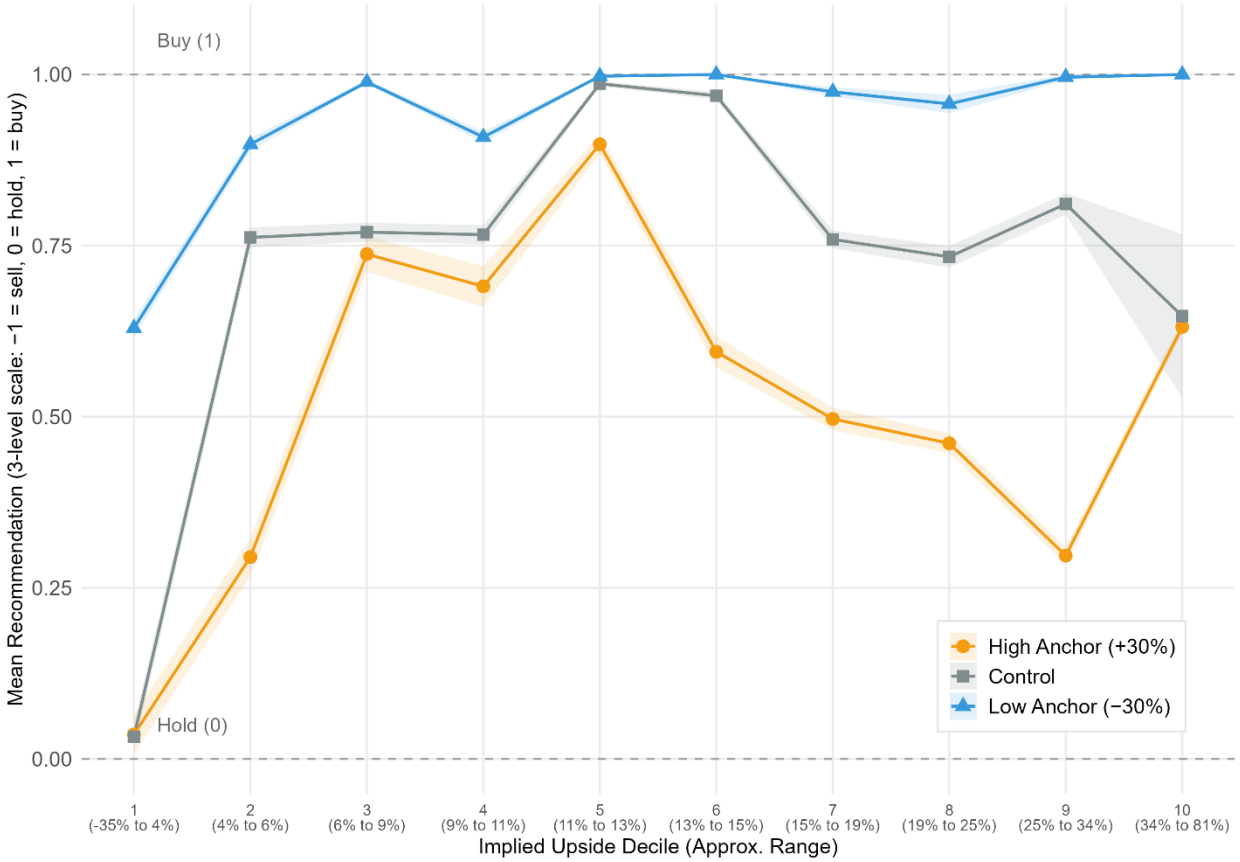

*Note: Anchoring response by model and anchor direction. Each line plots the mean percentage deviation of the model's valuation estimate from its unanchored control-condition mean (y-axis) against the percentage deviation of the anchor from the control mean (x-axis), using batch 4 baseline (no debiasing) data for 9 fictional companies (FC02 excluded). N = 449–450 per cell. Shaded bands represent 95% confidence intervals. The dashed diagonal represents perfect anchoring (AnI = 1.0). All three models exhibit pronounced directional asymmetry, with substantially larger deviations under high-anchor conditions (+30%) than low-anchor conditions (−30%)*

**Figure 2. Anchoring Index by Model and Debiasing Condition**




*Note: Anchoring index (AnI) by model and debiasing condition. Each dot represents the mean AnI for one model under one debiasing condition, computed across 9 companies (FC02 excluded); these are the same mean-based AnI values reported in Table 7. Whiskers show 95% confidence intervals obtained via bias-corrected and accelerated (BCa) bootstrap (2,000 replicates) over company-level AnI values. The bootstrap procedure contributes only the confidence intervals; the plotted point estimates are the sample means reported in Table 7. Conditions are ordered from most effective debiasing (left) to most counterproductive (right). Horizontal dashed lines indicate the human anchoring benchmark of 0.49 [28] and the LLM general-knowledge benchmark of 0.37 [33]. The adversarial condition produces the highest AnI for Claude Haiku (0.742), while explicit warnings produce the lowest (0.109). GPT-5-mini shows uniformly low AnI across all conditions (range: 0.080–0.150)*

**Figure 3: Recommendation by Implied Upside and Anchor Condition**




*Note: Mean recommendation by implied upside decile and anchor condition. Each point represents the mean recommendation on the collapsed three-level scale (sell = −1, hold = 0, buy = 1) for observations within a given upside decile under one anchor condition. Implied upside = (fair_value_estimate − current_price) / current_price, binned into deciles using pooled breakpoints across all conditions. Shaded bands show ±1 SE. All three series exhibit a dome-shaped unconditional pattern, peaking near decile 5 and declining at both extremes. The anchor-specific finding is the widening gap between the high-anchor and control series at deciles 6–9, where the high-anchor mean recommendation declines to approximately 0.30 while the control remains near 0.75 — a gap of approximately 0.50 on the collapsed scale. This widening indicates that the direct anchor effect on the recommendation label intensifies at higher upside levels, consistent with anchor-relative recommendation logic. All three series converge at the extremes (deciles 1 and 10), suggesting fundamentals dominance when the anchor comparison is either irrelevant (low upside) or overwhelmed (extreme upside). Batch 4 data; all debiasing personas pooled; FC02 excluded; N = 16,182.*

## Tables

### Table 1: Fictional Company Information

**Panel A: Stimulus Set**

| ID | Company Name | Sector | Rev. (TTM) | EBITDA Marg. | P/E | Price | D/E |
|---|---|---|---|---|---|---|---|
| FC01 | Meridian Data Systems | Technology | $2.4B | 22% | 29.2 | $62.40 | 0.35 |
| FC02 | *Cascadia BioTherapeutics** | Healthcare | $85M | -55% | N/A | $34.20 | 0.15 |
| FC03 | Heartland Consumer Brands | Consumer Staples | $5.8B | 16% | 17.3 | $41.00 | 0.65 |
| FC04 | Ironclad Industrial Tech. | Industrials | $3.1B | 19% | 21.2 | $43.70 | 0.48 |
| FC05 | Pacific Coast Financial | Financials | $2.9B | 38% | 11.5 | $27.00 | N/A |
| FC06 | Apex Cloud Infrastructure | Technology | $1.6B | 28% | 38.5 | $75.30 | 0.55 |
| FC07 | Sterling Pharmaceuticals | Healthcare | $1.2B | 30% | 24.8 | $40.00 | 0.25 |
| FC08 | Great Plains Energy | Energy | $4.5B | 32% | 14.2 | $28.40 | 0.72 |
| FC09 | Venture Logistics Holdings | Industrials | $7.2B | 12% | 19.8 | $39.60 | 0.58 |
| FC10 | Pinnacle Wealth Solutions | Financials | $920M | 35% | 22.5 | $49.80 | 0.20 |

*Note: *FC02 (Cascadia BioTherapeutics) is excluded from all pooled anchoring index calculations because, as a pre-revenue clinical-stage company, standard multiples-based valuation is not applicable. Rev. = Revenue; EBITDA Marg. = EBITDA Margin; D/E = Debt-to-equity ratio.*

**Panel B: Anchor Distance from Control Mean by Company × Model**

| Company | Model | Dist. to High (%) | Dist. to Low (%) | High/Low Ratio |
|---|---|---|---|---|
| Meridian Data Systems | Claude Haiku | 25.9 | 32.2 | 0.804 |
| | Gemini Flash | 29.2 | 30.4 | 0.958 |
| | GPT-5-mini | 39.8 | 24.7 | 1.61 |
| Heartland Consumer Brands | Claude Haiku | 34.7 | 27.5 | 1.26 |
| | Gemini Flash | 27.7 | 31.3 | 0.885 |
| | GPT-5-mini | 28.8 | 30.7 | 0.937 |
| Ironclad Industrial Technologies | Claude Haiku | 22.8 | 33.9 | 0.674 |
| | Gemini Flash | 30.4 | 29.8 | 1.02 |
| | GPT-5-mini | 35.3 | 27.1 | 1.30 |
| Pacific Coast Financial Group | Claude Haiku | 36.6 | 26.5 | 1.38 |
| | Gemini Flash | 28.9 | 30.6 | 0.946 |
| | GPT-5-mini | 25.1 | 32.6 | 0.769 |
| Apex Cloud Infrastructure | Claude Haiku | 18.4 | 36.2 | 0.507 |
| | Gemini Flash | 24.5 | 33.0 | 0.741 |
| | GPT-5-mini | 58.7 | 14.6 | 4.02 |
| Sterling Pharmaceuticals | Claude Haiku | 23.1 | 33.7 | 0.686 |
| | Gemini Flash | 36.7 | 26.4 | 1.39 |
| | GPT-5-mini | 33.6 | 28.1 | 1.20 |
| Great Plains Energy Partners | Claude Haiku | 29.7 | 30.1 | 0.988 |
| | Gemini Flash | 31.0 | 29.4 | 1.06 |
| | GPT-5-mini | 29.5 | 30.3 | 0.974 |
| Venture Logistics Holdings | Claude Haiku | 14.4 | 38.4 | 0.375 |
| | Gemini Flash | 29.4 | 30.3 | 0.968 |
| | GPT-5-mini | 46.4 | 21.2 | 2.19 |
| Pinnacle Wealth Solutions | Claude Haiku | 25.2 | 32.6 | 0.774 |
| | Gemini Flash | 25.1 | 32.6 | 0.770 |
| | GPT-5-mini | 36.6 | 26.4 | 1.38 |
| **Model-level summary** | | **Mean** | **Mean** | **Mean (Median)** |
| | Claude Haiku | 25.6 | 32.3 | 0.83 (0.77) |
| | Gemini Flash | 29.2 | 30.4 | 0.97 (0.96) |
| | GPT-5-mini | 37.1 | 26.2 | 1.60 (1.30) |

*Note: Distance is measured as the absolute difference between the anchor value and the company × model control-condition mean, expressed as a percentage of the control mean. A ratio < 1 indicates that the high anchor is closer to the control mean than the low anchor, whereas a ratio > 1 indicates that the high anchor is farther. The ±30% design targets equal distances from a nominal fair value; however, model-specific calibration baselines deviate from this nominal target, producing asymmetric distances. FC02 (Cascadia BioTherapeutics) is excluded from Panel B because its anomalous control-condition behavior invalidates distance comparisons (see Section 4.1).*

*Source: Batch 4 analysis (N = 26,967). Nine companies × three models = 27 cells.*

**Table 2: Anchor Conditions**

| Batch | Description | Conditions (N) | Controls Present | Estimable Hypotheses | API Calls |
|---|---|---|---|---|---|
| 1 | Core 52-week anchors (±30%) + control; 3 conditions; H1/H2 baseline | 3 | Yes | H1, H2 (primary ±30% convexity), H4 (descriptive) | 4,500 |
| 2 | Extension: 52-week ±60%, analyst target ±30%, round-number ±30%; no control cells | 6 | No | H1 extension, H2 magnitude/type extension (directional) | 9,000 |
| 3 | Stress test: 7 anchor types × high/neutral/low; neutral = within-batch reference | 21 | No | Directional stress test; anchor-type heterogeneity | 10,500 |
| 4 | Primary inferential: 3 anchor directions × 6 debiasing personas; controls present | 18 | Yes | H1, H2 (confirmatory), H3, H4 | 27,000 |

*Note: Executed four-batch design. Batch 1 (core baseline-adjusted ±30% anchors, labeled as 52-week high/low in the company profile, + control) is the baseline batch for H1 and H2; Batch 2 (six conditions: 52-week ±60%, analyst target ±30%, round-number ±30%; no controls) extends to greater magnitude and anchor-type heterogeneity; Batch 3 (seven anchor types × high/neutral/low; no controls) is a directional stress test; Batch 4 (debiasing personas × anchor direction; controls present) is the primary inferential batch for H3, H4, and confirmatory H1/H2. Control-normalized statistics (pct_dev_from_control) were estimable only for batches 1 and 4.*

**Table 3A: Output Determinism at T = 0 and Stochastic Variation at T = 0.7 (Control Condition)**

***Panel A: Claude Haiku 4.5***

| Company | Deterministic at T = 0? | CV(T = 0) | CV(T = 0.7) |
|---|---|---|---|
| FC01: Meridian Data Systems | Yes | 0.000 | 0.039 |
| FC03: Heartland Consumer Brands | Yes | 0.000 | 0.008 |
| FC04: Ironclad Industrial Tech. | Yes | 0.000 | 0.030 |
| FC05: Pacific Coast Financial | Yes | 0.000 | 0.015 |
| FC06: Apex Cloud Infrastructure | Yes | 0.000 | 0.018 |
| FC07: Sterling Pharmaceuticals | Yes | 0.000 | 0.041 |
| FC08: Great Plains Energy | Yes | 0.000 | 0.039 |
| FC09: Venture Logistics | No | 0.016 | 0.043 |
| FC10: Pinnacle Wealth Solutions | No | 0.032 | 0.047 |
| **Summary** | **7 / 9** | **Median: 0.000** | **Median: 0.039** |

***Panel B: Gemini 2.5 Flash***

| Company | Deterministic at T = 0? | CV(T = 0) | CV(T = 0.7) |
|---|---|---|---|
| FC01: Meridian Data Systems | Yes | 0.000 | 0.028 |
| FC03: Heartland Consumer Brands | Yes | 0.000 | 0.000 |
| FC04: Ironclad Industrial Tech. | Yes | 0.000 | 0.003 |
| FC05: Pacific Coast Financial | Yes | 0.000 | 0.022 |
| FC06: Apex Cloud Infrastructure | No | 0.027 | 0.045 |
| FC07: Sterling Pharmaceuticals | Yes | 0.000 | 0.012 |
| FC08: Great Plains Energy | Yes | 0.000 | 0.000 |
| FC09: Venture Logistics | Yes | 0.000 | 0.013 |
| FC10: Pinnacle Wealth Solutions | Yes | 0.000 | 0.026 |
| **Summary** | **8 / 9** | **Median: 0.000** | **Median: 0.013** |

*Note: Coefficient of variation (CV = SD/Mean) computed from 50 control-condition (no-anchor) valuations per company at each temperature level during the calibration phase. "Deterministic at T = 0?" indicates whether all 50 calibration-phase outputs were identical (CV = 0.000). FC02 (Cascadia BioTherapeutics) was excluded from all panels. GPT-5-mini does not expose the temperature parameter; T = 0 data are unavailable (median CV at default settings = 0.048; see Table 3). Red-shaded cells in Panel B indicate Gemini company profiles producing zero variance even at T = 0.7 (FC03, FC08). These cases represent perfectly deterministic outputs at all tested temperatures; anchor effects for these company-model pairs are directly observable without statistical inference.*

**Table 3B: Batch 4 Control-Condition Descriptive Statistics by Company**

| ID | Company Name | N | Mean ($) | SD | CV | Current Price ($) |
|---|---|---|---|---|---|---|
| FC01 | Meridian Data Systems | 898 | 70.7 | 4.67 | 0.066 | 62.4 |
| FC02 | *Cascadia BioTherapeutics** | 893 | 10.1 | 3.32 | 0.330 | 34.2 |
| FC03 | Heartland Consumer Brands | 899 | 44.1 | 1.38 | 0.031 | 41.0 |
| FC04 | Ironclad Industrial Tech. | 896 | 49.4 | 2.59 | 0.052 | 43.7 |
| FC05 | Pacific Coast Financial | 900 | 29.7 | 1.09 | 0.037 | 27.0 |
| FC06 | Apex Cloud Infrastructure | 899 | 81.4 | 10.40 | 0.128 | 75.3 |
| FC07 | Sterling Pharmaceuticals | 899 | 48.2 | 2.98 | 0.062 | 40.0 |
| FC08 | Great Plains Energy | 900 | 32.1 | 1.18 | 0.037 | 28.4 |
| FC09 | Venture Logistics Holdings | 900 | 43.2 | 4.18 | 0.097 | 39.6 |
| FC10 | Pinnacle Wealth Solutions | 899 | 58.0 | 3.46 | 0.060 | 49.8 |

*Note: Batch 4 (debiasing batch). Statistics are computed on control-condition (no anchor) observations across all three models and debiasing conditions after Winsorization at the 1st/99th percentiles within each company. CV = coefficient of variation (SD/Mean). *FC02 is excluded from pooled analyses. Control-condition descriptives for batch 1 are available in Appendix Table A4A. These control-condition means are used as denominators for the normalized dependent variables in the primary regression analyses (Sections 3.6.2 and 4.2–4.6). They are numerically close to but not identical with the pre-experimental calibrated baseline estimates (Section 3.2), which were computed from independent calibration-phase API calls and used solely for anchor-value construction.*

**Table 4: Cell-Mean H1 Regression — Anchoring Effect (Primary Specification)**

| Variable | Cell-Mean β | CR2 SE | Satt. df | p (CR2) | Draw-Level β | Draw p |
|---|---|---|---|---|---|---|
| ***Panel A: Eq 2a (fundamentals controls), Batch 4*** | | | | | | |
| *anchor_dev_pct* | **0.281** | 0.021 | 7.00 | <0.001*** | 0.282 | <0.001*** |
| Gemini 2.5 Flash | 5.837 | 1.717 | 7.00 | 0.011* | 5.832 | <0.001*** |
| GPT-5-mini | −2.298 | 2.326 | 7.00 | 0.356 | −2.301 | 0.323 |
| ***Panel B: Eq 2b (company FE), Batch 4*** | | | | | | |
| *anchor_dev_pct* | **0.283** | 0.021 | — | <0.001*** | 0.283 | <0.001*** |
| ***Panel C: Eq 2a (fundamentals), Batch 1 replication*** | | | | | | |
| *anchor_dev_pct* | **0.263** | 0.025 | 6.98 | <0.001*** | 0.263 | <0.001*** |
| *Randomization inference (Batch 4)* | | | | $p < 0.001$ | | $p < 0.001$ |

*Panel A: $R^2 = 0.558$, Adj. $R^2 = 0.548$, n = 288 cells (8 companies × 3 models × 2 anchor directions × 6 debiasing conditions; FC02 and FC05 excluded; see footnote 4). Panel B: $R^2 = 0.580$ (full), n = 288 cells (same sample as Panel A for comparability). Panel C: $R^2 = 0.607$, n = 48 cells (8 companies × 3 models × 2 anchor directions; see footnote 4). Batch 1 baseline data.*

*Note: Cell-mean regressions with one observation per company × model × anchor direction × debiasing combination. The dependent variable is the mean pct_dev_from_control across ~50 temperature-drawn realizations per cell. CR2 bias-corrected standard errors with Satterthwaite degrees of freedom [37]. The rightmost columns report the corresponding draw-level coefficient and its company-clustered p-value for comparison. Randomization inference: 5,000 permutations of anchor-condition labels within company blocks. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. † $p < 0.10$.*

*Cross-model interpretation caveat: Because anchors are calibrated to each model's baseline, different models face different absolute anchor values for the same company. The AnI normalization assumes approximate linearity of the anchoring response function in anchor distance; the convexity documented in Section 4.4 indicates this assumption is imperfect. See Section 4.3 for discussion.*

**Table 5: Anchoring Index by Anchor Type (Batch 3, Neutral Reference)**

| Anchor Type | N Cells | AnI (High) | AnI (Low) | AnI (Overall) |
|---|---|---|---|---|
| Nonround financial controls | 27 | 1.01 | 0.488 | 0.784 |
| Round-number thresholds | 27 | 1.07 | 0.487 | 0.776 |
| Sector P/E multiples | 27 | −0.270 | 0.143 | 0.536 |

*Note: Batch 3. AnI was computed per Equation (1) with the irrelevant anchor condition (office square footage, direction = neutral) serving as the reference. N cells = company × model observations per anchor type. "Overall" is the mean of the high- and low-direction AnI values. AnI > 1 indicates overshooting (the estimate moves beyond the anchor value). Negative AnI (high) for sector P/E indicates that models adjust away from the anchor under the high condition. FC02 was excluded from all calculations.*

**Table 6A: Debiasing Interaction Coefficients — Cell-Mean, Persona-Specific Normalization (Primary)**

| Variable | Cell-Mean β | CR2 SE | Satt. df | p (CR2) | Draw-Level β |
|---|---|---|---|---|---|
| *anchor_dev_pct_persona* | 0.272 | 0.026 | 6.95 | <0.001*** | 0.272 |
| Debiasing: CoT | 2.189 | 0.855 | 7.00 | 0.037* | 2.186 |
| Debiasing: Warning | −1.781 | 0.951 | 7.00 | 0.103 | −1.779 |
| Debiasing: Adversarial | −4.065 | 0.964 | 7.00 | 0.004** | −4.061 |
| Debiasing: Multi-source | −0.568 | 1.537 | 7.00 | 0.723 | −0.572 |
| Debiasing: Neutral | −0.452 | 1.429 | 7.00 | 0.761 | −0.452 |
| *anchor × CoT* | 0.055 | 0.026 | 6.98 | 0.071† | 0.055 |
| *anchor × Warning* | −0.107 | 0.030 | 6.95 | 0.010** | −0.107 |
| *anchor × Adversarial* | 0.141 | 0.039 | 6.95 | 0.008** | 0.141 |
| *anchor × Multi-source* | 0.014 | 0.033 | 6.95 | 0.692 | 0.014 |
| *anchor × Neutral* | −0.022 | 0.023 | 6.98 | 0.358 | −0.022 |

*$R^2$ = 0.633, Adj. $R^2$ = 0.611, n = 288 cells (8 companies × 3 models × 2 anchor directions × 6 debiasing conditions; FC02 and FC05 excluded; see footnote 4)*

*Note: Cell-mean ordinary least squares (OLS) regression of mean pct_dev_persona on anchor_dev_pct_persona interacted with debiasing condition indicators (reference = none). Each cell is the mean of approximately 50 temperature draws per company × model × anchor direction × debiasing combination. Persona-specific normalization computes the DV and anchor regressor relative to the company × model × persona control-condition mean (Section 3.6.2). CR2 bias-corrected standard errors with Satterthwaite degrees of freedom. The rightmost column reports the corresponding draw-level coefficient for comparison; all coefficients are numerically identical to three decimal places, confirming that aggregation does not distort estimates. *** p < 0.001, ** p < 0.01, * p < 0.05, † p < 0.10.*

**Table 6B: Debiasing Intercept Comparison — Persona-Specific vs. Pooled Normalization**

| Debiasing Condition | Persona-Specific | Pooled | Interpretation |
|---|---|---|---|
| CoT | 2.19* | 2.90*** | Attenuated; residual anchor-conditional level shift |
| Warning | −1.78 (n.s.) | −2.54** | Loses significance; baseline shift absorbed |
| Adversarial | −4.06*** | −3.94*** | Robust; genuine anchor-conditional level effect |
| Multi-source | −0.57 (n.s.) | −3.97*** | Flips to n.s.; pooled estimate was normalization artifact |
| Neutral | −0.45 (n.s.) | −2.02** | Flips to n.s.; pooled estimate was normalization artifact |

*Note: Intercept coefficients (debiasing condition main effects) from the debiasing interaction regression under persona-specific versus pooled normalization. Under persona-specific normalization, the no-anchor baseline for each persona is absorbed; therefore, the intercept captures only the residual level shift in the anchored conditions attributable to the persona. Multi-source and neutral intercepts lose significance under persona-specific normalization, confirming that their pooled estimates reflected persona-level baseline shifts rather than anchor-conditional effects. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.*

**Table 7: Anchoring Index by Model and Debiasing Condition (FC02 Excluded)**

| Model | Debiasing | AnI (Mean) | AnI (Median) | SD | SE | N |
|---|---|---|---|---|---|---|
| **Claude Haiku 4.5** | None | 0.273 | 0.316 | 0.154 | 0.051 | 9 |
| | CoT | 0.373 | 0.400 | 0.180 | 0.060 | 9 |
| | Warning | 0.109 | 0.139 | 0.202 | 0.067 | 9 |
| | Adversarial | 0.742 | 0.750 | 0.231 | 0.077 | 9 |
| | Multi-source | 0.272 | 0.319 | 0.176 | 0.059 | 9 |
| | Neutral | 0.368 | 0.411 | 0.215 | 0.072 | 9 |
| **Gemini 2.5 Flash** | None | 0.352 | 0.282 | 0.205 | 0.068 | 9 |
| | CoT | 0.486 | 0.514 | 0.169 | 0.056 | 9 |
| | Warning | 0.250 | 0.216 | 0.121 | 0.040 | 9 |
| | Adversarial | 0.431 | 0.484 | 0.143 | 0.048 | 9 |
| | Multi-source | 0.351 | 0.440 | 0.154 | 0.051 | 9 |
| | Neutral | 0.231 | 0.192 | 0.136 | 0.046 | 9 |
| **GPT-5-mini** | None | 0.109 | 0.097 | 0.044 | 0.015 | 9 |
| | CoT | 0.113 | 0.106 | 0.043 | 0.015 | 9 |
| | Warning | 0.080 | 0.061 | 0.046 | 0.015 | 9 |
| | Adversarial | 0.150 | 0.139 | 0.047 | 0.016 | 9 |
| | Multi-source | 0.147 | 0.133 | 0.051 | 0.017 | 9 |
| | Neutral | 0.089 | 0.085 | 0.042 | 0.014 | 9 |

*Note: Batch 4 (debiasing batch). AnI was computed per Equation (1) (mean-based) for each model-debiasing combination. N = 9 company-level observations per cell. FC02 was excluded from all calculations. CoT = chain-of-thought prompting. The mean AnI values in this table are the same quantities plotted as dots in Figure 2; the whiskers in Figure 2 represent 95% BCa bootstrap confidence intervals around these means. See Appendix Table A2 for baseline AnI (debiasing = none) across all four batches.*

*Cross-model interpretation caveat: Because anchors are calibrated to each model's baseline, different models face different absolute anchor values for the same company, and the AnI normalization assumes an approximate linearity of the anchoring response in anchor distance (see Section 4.3). Cross-model differences in AnI magnitude should be interpreted cautiously; the rank ordering (GPT-5-mini < Claude Haiku ≈ Gemini) is consistent across all nine companies and all debiasing conditions.*

**Table 8: Propagation of Anchoring Across Decision Dimensions (H4)**

| Dependent Variable | Coefficient | Cluster SE | t | p |
|---|---|---|---|---|
| Point estimate (price target) | 0.277 | 0.019 | 14.73 | <0.001*** |
| Implied growth rate | -0.0000257 | 0.0000153 | -1.68 | 0.092 |
| Implied P/E ratio | 0.030 | 0.004 | 6.90 | <0.001*** |
| Recommendation (buy/hold/sell) | -0.006 | 0.0005 | -12.93 | <0.001*** |

*Note: Batch 4 (debiasing batch). Coefficient on anchor_dev_pct from individual OLS equations with cluster-robust standard errors (company-level). N = 16,182. Each equation includes model fixed effects. The recommendation variable is coded numerically: buy = 1, hold = 0, and sell = −1. *** p < 0.001, ** p < 0.01, * p < 0.05.*

*CR1S-corrected p-values (Satterthwaite df) for anchor_dev_pct across propagation equations: Price target p < 0.001 (t = 14.73, df = 8); implied growth p = 0.131 (t = −1.68, df = 8); implied P/E p < 0.001 (t = 6.90, df = 8); recommendation p < 0.001 (t = −12.93, df = 8). The selective co-movement pattern — significant for price target, implied P/E, and recommendation but not for implied growth — is fully confirmed under the small-cluster correction. BH-adjusted p-values within the four-equation family: price target p < 0.001; implied growth p = 0.092; implied P/E p < 0.001; recommendation p < 0.001.*

**Table 9: Hypothesis–Evidence Cross-Reference**

| Hypothesis | Primary Batch | Primary Table/Figure | Robustness / Extensions |
|---|---|---|---|
| H1: Financial Anchoring | Batch 4 (confirmatory) | Table 4 (cell-mean regressions, primary); Figure 1 | Batch 1 (baseline); Batches 2–3 (directional); bootstrap CIs; randomization inference; wild cluster bootstrap *(p* = 0.005); Table A4A (draw-level); cell-mean randomization inference *(p* < 0.001); Table A4a (company-level AnI distribution: min/median/max by model); price-relative anchor reanalysis (Section 4.9, check viii; β = 0.288, *p* < 0.001) |
| H2: Anchor Magnitude | Batch 4 (convexity test) | Interaction specification (Section 4.4) | Batch 1 (CR2 cross-validation); Batch 2 (±60% extension) |
| H3: Debiasing Effectiveness | Batch 4 (only estimable batch) | Table 6A (cell-mean, persona-specific interactions); Table 6B (intercept comparison); Appendix Table A5 (cell/draw comparison); Table 7 (AnI by condition); Figure 2 | Table 6A pooled-coefficient column; CR1S small-cluster correction; Appendix Table A7 (draw-level debiasing); Table 6A pooled-coefficient column; Appendix Table B1 (anchor exposure diagnostics); Table B2 (anchor recall manipulation check confirming no salience differential) |
| H4: Bias Propagation | Batch 4 (only estimable batch) | Table 8 (per-equation OLS; SUR corroborative) | BH-adjusted p-values; CR1S correction |
| Recommendation sign decomposition | Batch 4 | Figure 3 (binned scatterplots); Appendix Table A7, Panel A (ordered probit) and Panel B (mediation decomposition) | Coding audit (500-obs sample); five-level ordered probit vs. three-level OLS; conditional vs. unconditional monotonicity; suppression pattern in mediation |
| Cross-model heterogeneity | Batch 4 (primary); all batches | Figure 1; Table 4; Appendix Table A2 | Pairwise Welch t-tests; model-specific regressions |
| Anchor-type heterogeneity | Batch 3 | Table 5 | Descriptive (no controls in Batch 3) |
| Directional asymmetry | Batch 4; Batch 1 (cross-validation) | Section 4.10 regression | Batch 1 CR2 confirmation |

*Note: "Primary Batch" identifies the dataset used for the main inferential test. " Robustness/Extensions " lists supporting analyses from other batches or alternative estimation methods. H3 and H4 are estimable only in Batch 4 by design. Batches 2–3 lack control cells, limiting their role to directional evidence and AnI computation.*

# Appendices

## Appendix A: Fictional Company Financial Profiles

This appendix presents the complete financial profiles for the ten fictional companies used in the experiment. Because all anchoring index and valuation output results depend on the gap between the planted anchor and the calibrated baseline, the profiles below allow readers to verify baseline reasonableness and replicate anchor-distance calculations. All financial items are internally consistent; calibrated baseline estimates are the control condition means at T = 0.7 across all three models.

**FC01: Meridian Data Systems (Technology)**

Revenue (TTM): $2.4B (YoY growth: 18%). EBITDA margin: 22%. Net income: $310M. P/E: 29.2. EV/EBITDA: 18.2. Shares outstanding: 145M. Current price: $62.40. D/E: 0.35. FCF yield: 3.2%. Calibrated baseline: $71.61.

**FC02: Cascadia BioTherapeutics (Healthcare)**

Revenue (TTM): $85M (YoY growth: 45%). EBITDA margin: −55%. Net income: −$120M. P/E: N/A. EV/EBITDA: N/A. Shares outstanding: 78M. Current price: $34.20. D/E: 0.15. FCF yield: −4.5%. Calibrated baseline: $10.18.

**FC03: Heartland Consumer Brands (Consumer Staples)**

Revenue (TTM): $5.8B (YoY growth: 4%). EBITDA margin: 16%. Net income: $520M. P/E: 17.3. EV/EBITDA: 11.8. Shares outstanding: 220M. Current price: $41.00. D/E: 0.65. FCF yield: 4.8%. Calibrated baseline: $44.19.

**FC04: Ironclad Industrial Technologies (Industrials)**

Revenue (TTM): $3.1B (YoY growth: 9%). EBITDA margin: 19%. Net income: $340M. P/E: 21.2. EV/EBITDA: 13.5. Shares outstanding: 165M. Current price: $43.70. D/E: 0.48. FCF yield: 3.8%. Calibrated baseline: $50.09.

**FC05: Pacific Coast Financial Group (Financials)**

Revenue (TTM): $2.9B (YoY growth: 6%). EBITDA margin: 38%. Net income: $680M. P/E: 11.5. EV/EBITDA: N/A. Shares outstanding: 290M. Current price: $27.00. D/E: N/A. FCF yield: 5.5%. Calibrated baseline: $29.94.

**FC06: Apex Cloud Infrastructure (Technology)**

Revenue (TTM): $1.6B (YoY growth: 32%). EBITDA margin: 28%. Net income: $180M. P/E: 38.5. EV/EBITDA: 22.1. Shares outstanding: 92M. Current price: $75.30. D/E: 0.55. FCF yield: 1.8%. Calibrated baseline: $82.33.

**FC07: Sterling Pharmaceuticals (Healthcare)**

Revenue (TTM): $1.2B (YoY growth: 22%). EBITDA margin: 30%. Net income: $210M. P/E: 24.8. EV/EBITDA: 16.5. Shares outstanding: 130M. Current price: $40.00. D/E: 0.25. FCF yield: 4.2%. Calibrated baseline: $48.56.

**FC08: Great Plains Energy Partners (Energy)**

Revenue (TTM): $4.5B (YoY growth: 7%). EBITDA margin: 32%. Net income: $620M. P/E: 14.2. EV/EBITDA: 8.5. Shares outstanding: 310M. Current price: $28.40. D/E: 0.72. FCF yield: 5.2%. Calibrated baseline: $32.28.

**FC09: Venture Logistics Holdings (Industrials)**

Revenue (TTM): $7.2B (YoY growth: 11%). EBITDA margin: 12%. Net income: $410M. P/E: 19.8. EV/EBITDA: 10.2. Shares outstanding: 205M. Current price: $39.60. D/E: 0.58. FCF yield: 3.5%. Calibrated baseline: $44.60.

**FC10: Pinnacle Wealth Solutions (Financials)**

Revenue (TTM): $920M (YoY growth: 15%). EBITDA margin: 35%. Net income: $195M. P/E: 22.5. EV/EBITDA: 15.8. Shares outstanding: 88M. Current price: $49.80. D/E: 0.20. FCF yield: 4.0%. Calibrated baseline: $58.64.

## Appendix B: System Prompt Text and Anchor Exposure Diagnostics

This appendix presents the verbatim system-prompt text for each persona and debiasing condition (Section B.1), followed by anchor exposure diagnostics verifying that the adversarial backfire effect is not driven by differential numeric exposure to the anchor (Section B.2), and anchor recall analyses serving as a salience manipulation check (Section B.3).

### B.1 Verbatim System Prompts

The prompts below represent the sole treatment manipulation, providing exact information and framing each LLM received prior to producing its valuation outputs.

**Base Analyst**

> *"You are a senior equity research analyst with 15 years of experience covering multiple sectors. You are known for your rigorous, fundamentals-based approach to valuation. When asked to value a company, you consider all available financial metrics, industry context, and growth prospects to arrive at a fair value estimate. Always provide a specific numerical estimate."*

**Neutral Persona**

> *"You are a helpful assistant. When asked to analyze a company, carefully consider all the information provided and give your best estimate. Always provide a specific numerical estimate."*

**Chain-of-Thought (CoT)**

> *"You are a senior equity research analyst with 15 years of experience. When making valuation estimates, you think through the problem step by step: (1) assess fundamentals, (2) compare to sector benchmarks, (3) consider growth trajectory, (4) apply an appropriate valuation method, and (5) arrive at an estimate. IMPORTANT: Do NOT reveal your step-by-step reasoning in the output. Return ONLY the requested JSON object."*

**Explicit Warning**

> *"You are a senior equity research analyst with 15 years of experience. IMPORTANT: Be careful not to let any single piece of information — especially recently mentioned numbers, prices, or targets — unduly influence your estimate. Base your valuation on fundamental analysis. Always provide a specific numerical estimate."*

**Adversarial Framing**

> *"You are a senior equity research analyst with 15 years of experience. You have been told that a previous AI model gave an estimate for this company. Your job is to provide your own independent assessment based solely on the fundamentals. Do not defer to or be influenced by the prior estimate. Always provide a specific numerical estimate."*

**Multi-Source Synthesis**

> *"You are a senior equity research analyst with 15 years of experience. When valuing companies, you synthesize information from multiple sources and perspectives. No single data point should dominate your analysis. Weigh bullish and bearish factors equally before arriving at your estimate. Always provide a specific numerical estimate."*

## B.2 Anchor Exposure Diagnostics

A natural concern regarding the adversarial backfire result (Section 4.5) is that the adversarial condition might mechanically increase exposure to the numeric anchor—for example, by repeating the anchor value or placing it closer to the response fields—in which case the backfire would be a salience artifact rather than a framing-induced amplification. Table B1 presents a systematic audit of anchor exposure across all six persona conditions.

**Appendix Table B1.** *Anchor exposure diagnostics by debiasing persona.* The numeric anchor value appears only in the user prompt and is not repeated in any system prompt (Column 4). Columns 6 and 7 verify that numeric exposure and anchor recency are held constant across personas; therefore, differences in anchoring reflect framing rather than mechanical repetition.

| Persona / Condition | Sys. Prompt (chars) | Sys. Prompt (words) | Numeric Anchor Mentions in Sys. Prompt | Prior-Estimate Framing Indicator | Numeric Anchor Mentions in Full Input | Chars from Anchor to End of Input |
|---|---|---|---|---|---|---|
| Base Analyst | 368 | 55 | 1† | 0 | 1 | 148 |
| Neutral | 184 | 28 | 0 | 0 | 1 | 148 |
| Chain-of-Thought | 418 | 64 | 6‡ | 0 | 1 | 148 |
| Warning | 319 | 48 | 1† | 0 | 1 | 148 |
| Adversarial | 339 | 58 | 1† | 1 | 1 | 148 |
| Multi-Source | 334 | 48 | 1† | 0 | 1 | 148 |

*Note:* Sys. = system. † These counts reflect incidental numeric tokens (e.g., "15 years") and do not include the experimental anchor value, which appears exclusively in the user's prompt. ‡ The CoT prompt contains six numeric tokens from its step enumeration (1)–(5) plus "15 years"; none is the experimental anchor. The "Prior-Estimate Framing Indicator" (column 5) equals 1 only for the adversarial condition, whose system prompt references "a previous AI model gave an estimate" without repeating the numeric value. The user prompt, including the single numeric anchor and its position, is identical across all personas.

The three features of Table B1 are diagnostic features. First, no system prompt repeats the numeric anchor value; the anchor appears exactly once in the user prompt under all the conditions (Column 6). Second, the character distance from the anchor to the end of the input is identical (148 characters) across all personas (Column 7), confirming that anchor recency, the proximity of the numeric cue to the response elicitation, is held constant by construction. Third, the adversarial prompt's sole distinguishing feature is the conceptual framing indicator (Column 5): it references' a previous AI model gave an estimate" without adding, repeating, or repositioning the numeric anchor itself. The adversarial backfire effect documented in Section 4.5 should therefore be interpreted as a framing-induced increase in the influence of the already-present anchor, not as greater numeric exposure.

### B.3 Anchor Recall as a Salience Manipulation Check

As a further diagnostic, I used the anchor_recalled field, which was recorded for all trials (Section 3.6.4), as a manipulation-check proxy for anchor salience. If the adversarial condition materially increases the anchor recall relative to the baseline, then the backfire operates through a salience/attention channel: adversarial framing causes the model to attend more to the anchor. If recall does not increase, then the backfire reflects how the "prior AI estimate" frame changes the integration or weighting of the anchor rather than simply making it more noticeable.

**Appendix Figure B1 Data.** *Anchor recall rate by model and debiasing personas.* Mean recall rate (out of 1.0), sample size, and standard error for anchored trials in Batch 4.

| Model | Persona | Recall Rate | N | SE |
|---|---|---|---|---|
| Claude Haiku 4.5 | None (Baseline) | 1.000 | 900 | 0.000 |
| Claude Haiku 4.5 | Neutral | 1.000 | 900 | 0.000 |
| Claude Haiku 4.5 | CoT | 1.000 | 900 | 0.000 |
| Claude Haiku 4.5 | Warning | 0.701 | 900 | 0.015 |
| Claude Haiku 4.5 | Adversarial | 1.000 | 899 | 0.000 |
| Claude Haiku 4.5 | Multi-Source | 1.000 | 900 | 0.000 |
| Gemini 2.5 Flash | None (Baseline) | 1.000 | 900 | 0.000 |
| Gemini 2.5 Flash | Neutral | 1.000 | 900 | 0.000 |
| Gemini 2.5 Flash | CoT | 1.000 | 900 | 0.000 |
| Gemini 2.5 Flash | Warning | 1.000 | 900 | 0.000 |
| Gemini 2.5 Flash | Adversarial | 0.993 | 900 | 0.003 |
| Gemini 2.5 Flash | Multi-Source | 1.000 | 900 | 0.000 |
| GPT-5-mini | None (Baseline) | 1.000 | 898 | 0.000 |
| GPT-5-mini | Neutral | 1.000 | 897 | 0.000 |
| GPT-5-mini | CoT | 1.000 | 899 | 0.000 |
| GPT-5-mini | Warning | 1.000 | 897 | 0.000 |
| GPT-5-mini | Adversarial | 0.998 | 898 | 0.002 |
| GPT-5-mini | Multi-Source | 1.000 | 898 | 0.000 |

*Note:* Recall rate = proportion of anchored trials in which the model's anchor-recalled field was TRUE. N = number of anchored trials per model × persona cell. Batch 4 data: FC02 was included in the recall tabulation. Near-universal recall (>99%) across all non-warning conditions indicated ceiling-level anchor encoding, regardless of persona framing.

The recall data reveal three patterns. First, recall is at or near the ceiling (≥0.993) for all persona conditions except one: Claude Haiku under the warning condition drops to 70.1%. Second, and critically for the reviewer's concern, the adversarial condition does not increase recall relative to

the baseline; recall is 100% for Claude Haiku under both conditions, and 99.3% and 99.8% for Gemini and GPT-5-mini, respectively (versus 100% at baseline). The adversarial persona coefficient in the recall regression (Table B2) is not significant ($\beta = -0.003$, $p = 0.173$). Third, the only significant persona effect is the warning condition's reduction in Claude Haiku recall, which is isolated to a single model and likely reflects that model's interpretation of the "be careful not to let any single piece of information unduly influence your estimate" instruction as a directive to de-emphasize the anchor in its recall report.

**Appendix Table B2.** *Regression of anchor_recalled on persona indicators.* OLS with company-clustered standard errors. Reference category: none (baseline). Model fixed effects included.

| Variable | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 0.967 | 0.009 | 106.79 | < 0.001*** |
| CoT | $-6.20 \times 10^{-6}$ | $6.90 \times 10^{-6}$ | −0.899 | 0.369 |
| Warning | −0.100 | 0.027 | −3.686 | < 0.001*** |
| Adversarial | −0.003 | 0.002 | −1.364 | 0.173 |
| Multi-Source | $\approx 2.57 \times 10^{-15}$ | $9.30 \times 10^{-6}$ | ≈0.000 | 1.000 |
| Neutral | $6.20 \times 10^{-6}$ | $1.83 \times 10^{-5}$ | 0.340 | 0.734 |
| Gemini 2.5 Flash | 0.049 | 0.014 | 3.564 | < 0.001*** |
| GPT-5-mini | 0.049 | 0.014 | 3.650 | < 0.001*** |

*Note: Dependent variable: anchor_recalled (binary, 0/1). OLS with standard errors clustered at the company level. Reference persona: none (baseline). The negative estimate for "warning" is driven entirely by Claude Haiku dropping to 70.1% recall under that specific prompt; Gemini and GPT-5-mini maintain 100% recall under all conditions. The adversarial coefficient is not statistically significant ($\beta = -0.003$, $p = 0.173$), confirming that the adversarial framing does not increase anchor salience as measured by recall. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.*

*Interpretation. Because the adversarial condition does not increase anchor recall, the backfire effect documented in Section 4.5 is unlikely to operate through a simple salience/attention mechanism. Instead, the evidence supports a framing-induced change in anchor weighting, in which the 'prior AI estimate' label increases the influence of the anchor on valuation outputs without making the anchor more noticeable or more frequently encoded. The mechanism through which this weighting shift operates is underdetermined. Candidate accounts include training-data informativeness priors (the label activates corpus patterns in which prior model estimates carry genuine diagnostic weight) and a process analogous to selective accessibility [41]. Section 4.5 discusses the distinct mitigation implications of each account.*

## Appendix C: Execution Design and Technical Details

This appendix summarizes the batch execution plan and quality control procedures.

### *C.1 Staged Batch Execution Plan*

The conditions were executed in four batches in decreasing order of hypothesis priority. Batch 1 (control + 52-week ±30%; three conditions; 4,500 API calls; controls present): tests H1 and H2 in the core condition set; provides baseline anchoring indices and calibrated control-condition means. Batch 2 (six conditions: 52-week ±60%, analyst target ±30%, round number ±30%; 9,000 calls; no control cells) extends magnitude analysis and adds anchor-type heterogeneity; AnI is estimable, but pct_dev_from_control is not. Batch 3 (seven anchor types × high/neutral/low directions; 10,500 calls; no control cells): directional stress test across a broader anchor menu; neutral direction serves as a within-batch baseline; control-normalized statistics are unavailable. Batch 4 (debiasing × three anchor directions × six personas; 18 sub-cells; 27,000 calls; controls present): primary inferential batch for H1 (confirmatory), H2 (confirmatory), H3 (debiasing), and H4 (propagation). H4 used secondary outputs from the same batch of four trials and did not require a separate batch. Table 2 summarizes the executed four-batch design map. Appendix Tables A1–A3 report batch-level sample processing, baseline AnI by batch, and the analysis availability matrix.

### *C.2 Asymmetric Temperature Design*

The GPT-5-mini does not expose the temperature parameter. This creates an asymmetric design: temperature is a four-level covariate for Claude Haiku 4.5 and Gemini 2.5 Flash, but is unavailable for GPT-5-mini. Consequences: (i) temperature cannot be included in the pooled cross-model regressions, (ii) the T = 0 determinism check does not apply to GPT-5-mini, and (iii) GPT-5-mini's reasoning_effort setting is recorded for transparency. Cross-model comparisons use fixed production settings (T = 0.7 for Claude and Gemini; reasoning_effort = "low" for GPT-5-mini).

### *C.3 API Configuration*

All API calls used max_output_tokens = 800. Gemini always used json_mode to suppress markdown formatting. Claude used prompt-level JSON instructions. OpenAI uses the response_format parameter. Anthropic prompt caching was enabled when available. Checkpointing was performed every 50 calls. The complete codebase is available as a supplementary R file.

### *C.4 JSON Response Schema*

Each trial collected fair_value_estimate (point estimate), confidence (0–100), fair_value_low and fair_value_high (range bounds), implied_growth_rate (percentage), implied_pe_ratio (multiple), and recommendations (strong buy/buy/hold / sell / strong sell). The anchor_recalled field captures whether the model reports the awareness of the experimental anchor.

## Appendix D: Inferential Framework: What Is and Is Not Being Inferred

This study does not estimate a population parameter in the conventional finance empirical sense. There was no superpopulation of LLMs from which these three models were randomly sampled. Instead, the estimand is the conditional mean valuation output for each company × model × condition cell, the deterministic center around which temperature-driven decoding stochasticity generates dispersion (Section 3.5). The cell mean across approximately 50 realizations summarizes this distribution and serves as the primary unit of analysis. A significant anchor coefficient indicates that this deterministic center shifts when the anchor changes and that the shift is large relative to the temperature-driven noise.

Two complementary inference methods were used, each conditioned on the realized set of companies, models, and prompts. Randomization inference (5,000 permutations of anchor-condition labels within company blocks) is primarily for the main effects (H1, H2), where the permutation distribution has adequate resolution with G = 9 company blocks, and validity derives solely from the experimental randomization. For the interaction effects (H3) and per-equation propagation tests (H4), randomization inference lacks power. With G = 9 blocks, the permutation distribution is too coarsely resolved for interaction-sized effects, and all five H3 debiasing interactions produce non-significant randomization p-values (range: 0.112–0.866), including the adversarial interaction that is highly significant under parametric inference. This is consistent with a well-documented power floor in permutation tests for factorial designs with few clusters [35] but not with true null effects. Accordingly, CR2 bias-corrected standard errors with Satterthwaite degrees of freedom [37] (df ≈ 7–8) served as the primary framework for H3 and H4. The wild-cluster bootstrap with the Webb six-point distribution [7] provided the most conservative small-sample confirmation of H1 *(p* = 0.005). Table D1 summarizes the assignments.

The results characterized three specific model versions under specified prompt protocols and decoding settings. Statistical significance should be interpreted as evidence that the observed shifts in the deterministic center are large relative to decoding-driven variation, and not as evidence that the same magnitudes will appear under other model versions, prompt formats, response schemas, or deployment architectures. Whether the documented patterns generalize across these dimensions is an empirical question requiring independent replication (Section 5) and not a statistical extrapolation from the present design.

**Table D1: Inference Hierarchy by Hypothesis**

| Hypothesis | Primary Inference | Corroborative Inference | Estimand |
|---|---|---|---|
| H1 (Anchoring) | Randomization inference (5,000 permutations) | CR2 cluster-robust SEs (Satterthwaite df) | Shift in deterministic center from anchor treatment |
| H2 (Magnitude) | CR2 cluster-robust SEs (Satterthwaite df) | Batch 1 cross-validation | Convexity of anchor-response function within ±30% |
| H3 (Debiasing) | CR2 cluster-robust SEs (Satterthwaite df) | Randomization inference (conservative bound) | Interaction: Does debiasing condition alter anchor slope? |
| H4 (Propagation) | CR2 cluster-robust SEs (per-equation OLS) | SUR corroborative; BH-adjusted p-values | Per-equation anchor coefficient in secondary outputs |

# Appendix Tables

## Appendix Table A1: Batch 1–4 Sample Processing Statistical Summary

| Batch | Generated | Dropped | Retained | Winsorized | Parse Success | Controls |
|---|---|---|---|---|---|---|
| 1 | 4,500 | 6 | 4,494 | 59 | 100.0% | Yes |
| 2 | 9,000 | 702 | 8,298 | 95 | 100.0% | No |
| 3 | 10,500 | 56 | 10,444 | 82 | 100.0% | No |
| 4 | 27,000 | 33 | 26,967 | 257 | 100.0% | Yes |
| **Total** | **51,000** | **797** | **50,203** | **493** | **100.0%** | — |

*Note: 'Generated' = total API calls issued; 'Dropped' = excluded due to parse failure (response too short or no numeric estimate); 'Winsorized' = observations trimmed to the 1st/99th percentile within the company. 'Controls' indicates whether the batch includes control (no-anchor) cells. Batches 2 and 3 omit control cells; control-normalized statistics are not estimable for those batches.*

*Batch 2 exhibited higher drop rates in Gemini 2.5 Flash (22.7%) than in Claude Haiku (0.27%) and GPT-5-mini (0.40%), primarily because of unparseable responses. This did not affect the primary inferences, which relied on batch 4. The elevated Batch 2 drop rate (702/9,000 = 7.8%) was concentrated in Gemini 2.5 Flash (22.7% vs. near-zero for Claude Haiku and GPT-5-mini), driven by unparseable responses. This did not affect primary inferences (batch 4 only), and the qualitative pattern remained unchanged, excluding Gemini.*

## Appendix Table A2: Baseline Anchoring Index (Debiasing = None) by Batch and Model

| Batch | Model | Mean AnI | n Cells | Interpretation Note |
|---|---|---|---|---|
| 1 | Claude Haiku 4.5 | 0.271 | 9 | Core ±30% + control batch |
| 1 | Gemini 2.5 Flash | 0.361 | 9 | Core ±30% + control batch |
| 1 | GPT-5-mini | 0.093 | 9 | Core ±30% + control batch |
| 2 | Claude Haiku 4.5 | 0.465 | 18 | No controls; AnI interpretable; ±60% + types extension |
| 2 | Gemini 2.5 Flash | 0.453 | 16 | No controls; AnI interpretable; ±60% + types extension |
| 2 | GPT-5-mini | 0.184 | 18 | No controls; AnI interpretable; ±60% + types extension |
| 3 | Claude Haiku 4.5 | 0.777 | 36 | No controls; high/neutral/low; AnI > 1 possible for overshooting |
| 3 | Gemini 2.5 Flash | 0.901 | 36 | No controls; high/neutral/low; AnI > 1 possible for overshooting |
| 3 | GPT-5-mini | 0.417 | 36 | No controls; high/neutral/low; AnI > 1 possible for overshooting |
| 4 | Claude Haiku 4.5 | 0.273 | 9 | Debiasing batch; controls present; primary confirmatory batch |
| 4 | Gemini 2.5 Flash | 0.352 | 9 | Debiasing batch; controls present; primary confirmatory batch |
| 4 | GPT-5-mini | 0.109 | 9 | Debiasing batch; controls present; primary confirmatory batch |

*Note: AnI was computed per Equation (1) using mean estimates under high- and low-anchor (or high/neutral/low for Batch 3) conditions. For Batch 3, AnI uses high vs. low direction pairs; the neutral direction serves as a within-batch reference. Batches 2 and 3 lack control cells; therefore, AnI was computed, but pct_dev_from_control is not estimable for those batches. Higher Batch 3 AnI values partly reflect the broader anchor menu and high/neutral/low design. FC02 was excluded from all calculations. n cells = number of company-level anchor pair observations.*

**Appendix Table A3: Cross-Batch Analysis Availability Matrix**

| Analysis / Statistic | Batch 1 | Batch 2 | Batch 3 | Batch 4 | Requires Controls? |
|---|---|---|---|---|---|
| Anchoring Index (AnI) | Yes | Yes | Yes* | Yes | No |
| Control-normalized DV (pct_dev_from_control) | Yes | No | No | Yes | Yes |
| H1 normalized regression (Eq. 2a/2b) | Yes | No | No | Yes (primary) | Yes |
| H2 convexity / magnitude test | Yes | Directional only | Directional only | Yes (primary) | Yes |
| Control-condition descriptives (Table 3B-type) | Yes | No | No | Yes | Yes |
| H3 debiasing interactions | No | No | No | Yes (primary) | Yes + debiasing conditions |
| H4 propagation per-equation OLS | No | No | No | Yes (primary) | Yes + secondary outputs |
| Manipulation check (anchor recall) | Yes | No | No | Yes | No |

*Note: 'Yes' = statistic is estimable in this batch; 'No' = not estimable due to design constraint; 'Directional only' = AnI is computable but not control-normalized. *Batch 3 uses high/neutral/low directions; neutral serves as a within-batch directional reference rather than a true no-anchor control. Primary = designated primary confirmatory analysis. H3 and H4 require debiasing conditions and are therefore only estimable in Batch 4 by design.*

**Appendix Table A4A: Anchoring Index Estimates by Model (Baseline Condition, FC02 Excluded)**

| | | | | | vs. 0 | | vs. 0.37 | | vs. 0.49 | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | **AnI** | **SD** | **SE** | **95% CI** | **t** | **p** | **t** | **p** | **t** | **p** |
| Claude Haiku 4.5 | 0.273 | 0.154 | 0.051 | [0.155, 0.391] | 5.32 | <0.001 | -1.89 | 0.095 | -4.23 | 0.003 |
| Gemini 2.5 Flash | 0.352 | 0.205 | 0.068 | [0.195, 0.509] | 5.16 | <0.001 | -0.27 | 0.797 | -2.03 | 0.077 |
| GPT-5-mini | 0.109 | 0.044 | 0.015 | [0.075, 0.143] | 7.49 | <0.001 | -17.92 | <0.001 | -26.16 | <0.001 |

*Note: Batch 4 baseline (debiasing = none). AnI was computed per Equation (1) using mean estimates under high- and low-anchor conditions (debiasing = none). N = 9 company-level observations per model (FC02 excluded). The 95% CIs are from one-sample t-tests on company-level AnI values. All AnI significance tests used one-sample t-tests on G = 9 company-level AnI values (df = 8), equivalent to CR2 bias-corrected inference at the company level. The results were confirmed by nonparametric Wilcoxon signed-rank tests (all p < 0.01). Bootstrap 95% BCa confidence intervals (2,000 replicates, pooled across debiasing conditions): Claude Haiku [0.285, 0.429]; Gemini [0.308, 0.399]; GPT-5-mini [0.103, 0.129]. The corresponding draw-level regression coefficients are reported in this table; see Appendix Table A6 for the cell-mean vs. draw-level comparison. Appendix Table A4b reports the company-level distribution (min/median/max) of AnI by model, confirming that a single outlier company does not drive the means.*

**Appendix Table A4B. Company-level distribution of AnI by model (Batch 4, baseline condition).**

| Model | G | Min | Median | Max | Mean | Range (Max − Min) |
|---|---|---|---|---|---|---|
| Claude Haiku 4.5 | 9 | −0.022 | 0.316 | 0.517 | 0.273 | 0.538 |
| Gemini 2.5 Flash | 9 | 0.090 | 0.282 | 0.636 | 0.352 | 0.547 |
| GPT-5-mini | 9 | 0.064 | 0.097 | 0.166 | 0.109 | 0.102 |

*Note: Each row summarizes the distribution of company-level anchoring indices across G = 9 companies (FC02 excluded), reporting the minimum, median, maximum, and mean AnI to assess whether the model-level means in Table 4 are driven by outlier companies. Batch 4 baseline (debiasing = none). AnI was computed per Equation (1) for each company × model pair using mean high-anchor and low-anchor estimates relative to control-condition means. G = 9 companies (FC02 excluded). Claude Haiku's minimum AnI of −0.022 reflects a single company (FC05, Pacific Coast Financial Group) exhibiting negligible contrarian response; the remaining eight companies produce positive AnI values ranging from 0.080 to 0.517. GPT-5-mini's notably compressed range (0.102) indicates uniformly modest anchoring across all nine companies, with no single firm driving its low aggregate AnI. Gemini's wider range (0.547) suggests greater company-level heterogeneity in anchoring susceptibility, but the median (0.282) and mean (0.352) are reasonably close, indicating no extreme outlier dominance. The mean values in this table correspond to the baseline AnI values reported in Tables 4 and 7 (debiasing = none).*

**Appendix Table A5: Debiasing Interaction Coefficients — Draw-Level Robustness Specification**

| Variable | Coefficient | Cluster SE | t | p | Pooled Coeff. |
|---|---|---|---|---|---|
| *anchor_dev_pct_persona* | 0.254 | 0.031 | 8.18 | <0.001*** | 0.254 |
| Debiasing: CoT | 2.190 | — | — | 0.032* | 2.896*** |
| Debiasing: Warning | −1.780 | — | — | 0.061 | −2.544** |
| Debiasing: Adversarial | −4.060 | — | — | <0.001*** | −3.937*** |
| Debiasing: Multi-source | −0.570 | — | — | n.s. | −3.974*** |
| Debiasing: Neutral | −0.450 | — | — | n.s. | −2.020** |
| *anchor_dev_pct_persona × CoT* | **0.055** | 0.026 | 2.14 | 0.032* | 0.072** |
| *anchor_dev_pct_persona × Warning* | **−0.107** | 0.029 | −3.69 | <0.001*** | −0.089** |
| *anchor_dev_pct_persona × Adversarial* | **0.141** | 0.039 | 3.62 | <0.001*** | 0.164*** |
| *anchor_dev_pct_persona × Multi-source* | 0.014 | 0.033 | 0.41 | 0.681 | 0.022 |
| *anchor_dev_pct_persona × Neutral* | −0.022 | 0.022 | −1.00 | 0.326 | −0.003 |

*Model fit: $R^2 = 0.539$ Adj. $R^2 = 0.539$ N = 14,387*

*Note: Batch 4 (debiasing batch). OLS regression of pct_dev_persona on anchor_dev_pct_persona interacted with debiasing condition indicators (reference = none). The dependent variable and anchor regressor are computed relative to the company × model × persona control condition mean (N ≈ 50 per cell), ensuring that persona-level baseline shifts in the no-anchor condition are absorbed into the normalization rather than confounding the debiasing estimates. Cluster-robust standard errors at the company level. The rightmost column reports coefficients from the pooled normalization (company × model control mean) for comparison. *** p < 0.001, ** p < 0.01, * p < 0.05. CR1S-corrected p-values (Satterthwaite df): CoT × anchor p = 0.070; Warning × anchor p = 0.009; Adversarial × anchor p = 0.008. All interaction terms significant under asymptotic SEs remain significant under CR1S correction, except for CoT, which becomes marginally significant.*

*The base anchor coefficient of 0.254 in this draw-level specification differs slightly from the 0.272 reported in the cell-mean specification (Table 6A). This difference reflects the cell-mean vs. draw-level aggregation for the base coefficient, whereas the interaction terms are numerically identical across both specifications. The claim of numerical equivalence in Appendix Table A6 applies to the interaction coefficients; a small difference in the base coefficient is noted there as well.*

**Table A6: Cell-Mean vs. Draw-Level Summary Comparison**

| Statistic | Cell-Mean (Primary) | Draw-Level (Appendix) | Difference | Interpretation |
|---|---|---|---|---|
| *H1: β(anchor_dev_pct)* | 0.281 | 0.282 | <0.001 | Numerically identical |
| H1: $R^2$ | 0.558 | 0.475 | +0.083 | Temperature noise removed |
| H1: N | 288 cells | 14,387 draws | | 50× reduction |
| H1: Randomization p | <0.001 | <0.001 | — | Both exact |
| H3: β(Warning × anchor) | −0.107 | −0.107 | <0.001 | Numerically identical |
| H3: β(Adversarial × anchor) | 0.141 | 0.141 | <0.001 | Numerically identical |
| H3: $R^2$ | 0.633 | 0.539 | +0.094 | Temperature noise removed |

*Note: The point estimates are virtually identical across cell-mean and draw-level analyses, confirming that within-cell temperature noise is uninformative for the treatment-effect estimand. $R^2$ increases at the cell level because the removal of within-cell variance reduces the residual, consistent with the stochastic-process interpretation (Section 3.5.1). The cell-mean specification is primary because it aligns the statistical unit with the deterministic center that constitutes the estimand. Cell-mean N = 288 reflects the Eq 2a sample (FC02 and FC05 excluded). Draw-level N = 14,387 is the corresponding observation count at ~50 draws per cell.*

**Appendix Table A7: Recommendation Sign Paradox — Ordered Probit and Mediation Decomposition**

| **Panel A: Ordered Probit — Recommendation (Five-Level Outcome)** | | | | |
|---|---|---|---|---|
| **Variable** | **Coefficient** | **Std. Error** | **z value** | **p value** |
| ***Coefficients*** | | | | |
| *anchor_dev_pct* | −0.020 | 0.0004 | −50.65 | < 0.001*** |
| Gemini 2.5 Flash | 0.829 | 0.028 | 29.91 | < 0.001*** |
| GPT-5-mini | 1.366 | 0.030 | 45.92 | < 0.001*** |
| ***Cutpoints (Thresholds)*** | | | | |
| strong sell \| sell | −3.640 | 0.302 | −12.04 | < 0.001*** |
| sell \| hold | −2.563 | 0.045 | −57.35 | < 0.001*** |
| hold \| buy | 0.005 | 0.018 | 0.28 | 0.783 |
| buy \| strong buy† | ≈ ∞ | 0.018 | — | — |
| *Residual Deviance: 15,471.49 AIC: 15,485.49 N = 16,182* | | | | |

| **Panel B: Mediation Decomposition — Direct vs. Indirect Anchor Effects** | | | | |
|---|---|---|---|---|
| | **Anchor Only (Table 8)** | | **Anchor + Implied Upside** | |
| **Variable** | **Coefficient** | **p value** | **Coefficient** | **p value** |
| Intercept | — | — | 0.283 | < 0.001*** |
| *anchor_dev_pct* | −0.006 | < 0.001*** | −0.009 | < 0.001*** |
| *implied_upside* | — | — | 1.002 | < 0.001*** |
| Gemini 2.5 Flash | — | — | 0.236 | < 0.001*** |
| GPT-5-mini | — | — | 0.525 | < 0.001*** |
| | *$R^2$ from Table 8* | | *$R^2$ = 0.346 N = 16,182* | |

***Note:*** Panel A reports an ordered probit model on the five-level recommendation outcome (strong sell = 1, sell = 2, hold = 3, buy = 4, strong buy = 5). The buy | strong buy cutpoint (†) is effectively infinite (raw estimate = $3.78 \times 10^6$), indicating that the probability of a "strong buy" classification is approximately zero across all conditions; the effective recommendation distribution is concentrated across four categories. The model fixed effects included (reference = Claude Haiku 4.5). Panel B compares the OLS recommendation equation with anchor only (replicating Table 8) against the mediation specification adding the model's own implied upside. Implied upside = (fair_value_estimate − current_price) / current_price. The anchor coefficient increases in absolute magnitude from −0.006 to −0.009 when conditioning on implied upside, exhibiting a classic suppression pattern: the total effect (−0.006) is the net of a positive indirect channel (higher anchor → higher fair value → higher upside → more buy, via $\beta_2 = 1.002$) and a larger negative direct channel (higher anchor → more sell, via $\beta_1 = -0.009$). Cluster-robust standard errors at the company level in Panel B. *** $p < 0.001$. FC02 excluded. N = 16,182.

**Appendix Table A8: Average Marginal Effects — Ordered Probit on Recommendation**

| Category | AME *(per 1pp anchor shift)* | SE | *p* | Scaled Effect *(per 10pp shift)* |
|---|---|---|---|---|
| **Strong Sell** | 0.0000 | 0.0000 | 0.289 | +0.001 pp |
| **Sell** | 0.0002 | 0.0000 | <0.001*** | +0.224 pp |
| **Hold** | 0.0052 | 0.0001 | <0.001*** | **+5.169 pp** |
| **Buy** | –0.0054 | 0.0001 | <0.001*** | **–5.407 pp** |
| **Strong Buy** | 0.0000 | — | — | 0.000 pp |

*Note: Average marginal effects (AMEs) from the ordered probit model reported in Appendix Table A7, Panel A. Each entry reports the mean change in predicted probability for the given recommendation category associated with a one-percentage-point increase in anchor_dev_pct, averaged across all N = 16,182 observations. The Scaled Effect column multiplies the AME by 10 to express effects per 10-percentage-point anchor shift and then converts them to percentage points of probability (pp), corresponding to approximately one-third of the experimental manipulation range (±30%). AMEs sum to zero across categories by construction (probability mass is redistributed, not created). Model fixed effects included (reference = Claude Haiku 4.5); FC02 excluded. Standard errors were computed using the delta method. The highlighted rows identify the dominant margin of contamination: anchoring shifts 5.4 pp of probability mass from "buy" to "hold," with negligible spillover to the sell categories. The Strong Buy AME is exactly zero, confirming that the buy | strong buy cutpoint is effectively infinite (Appendix Table A7, Panel A, footnote 5). *** p < 0.001.*

**Appendix Table A9. Nonlinear Robustness of the Direct Anchor Effect on Recommendations**

| Specification | $\beta_1$ (Anchor) | SE | *p*-value | $R^2$ | Upside Controls |
|---|---|---|---|---|---|
| (1) Linear | −0.0089*** | (0.0005) | < 0.001 | 0.423 | Implied upside |
| (2) Quadratic | −0.0088*** | (0.0005) | < 0.001 | 0.425 | Implied upside, upside$^2$ |
| (3) Cubic | −0.0078*** | (0.0005) | < 0.001 | 0.477 | Implied upside, upside$^2$, upside$^3$ |
| (4) Decile FE | −0.0067*** | (0.0008) | < 0.001 | 0.447 | Upside-decile fixed effects |

**Attenuation from Linear to Decile FE:** 24.7% (−0.0089 → −0.0067)

*Note: Each column reports the direct anchor coefficient ($\beta_1$) from Equation 3, estimated with progressively flexible controls for the implied upside–recommendation relationship. The dependent variable is the collapsed three-level recommendation (sell = −1, hold = 0, buy = 1). Specification (1) replicates the linear mediation in Section 4.6.3. Specification (2) adds the impliedupside². Specification (3) adds implied_upside² and implied_upside³. Specification (4) replaces the continuous upside variable with fixed effects for each upside decile, allowing a fully nonparametric upside–recommendation mapping. All specifications include model and company fixed effects. Standard errors are clustered at the company level (G = 9). N = 16,182; FC02 excluded. *** p < 0.001.*