When Biased Agents Trade: Anchoring, Exploitation, and Market Failure in Agent-to-Agent Interactions

Anton Hantel*

MIT

Working Paper

Abstract

Large language models have well-documented behavioral biases, but all existing evidence comes from single-agent settings. This paper asks what happens when two biased LLMs meet in a commercial transaction. Across 8,415 controlled interactions on three frontier models, agents anchor on first offers modestly more strongly than humans, yet resist information overload, largely ignore decoy products, and bid at equilibrium in auctions. Anchoring is the bias that breaks markets: a few dollars of price distortion per negotiation collapses a multi-agent marketplace from 96% to 15% efficiency as sellers price themselves out and buyers walk away. When one side knows the bias, it captures up to 78% of available surplus. Naming the bias in a warning recovers about 40% of the loss; telling the agent to reason step by step backfires, because deliberation makes the anchor more salient, not less. The three models fail in different ways, complicating any uniform regulatory response.

JEL Classification: D91, D18, D44, K24, L86

Keywords: agentic commerce, agent-to-agent interactions, large language models, behavioral biases, anchoring, winner's curse, decoy effect

*I am grateful to Professor Cass R. Sunstein at Harvard Law School for his guidance and feedback on this project. An earlier version of this paper was written for HLS 2589: Behavioral Economics, Law, and Public Policy, Spring 2026. I also thank Jono Hart for many valuable discussions and his feedback on the ideas.

1 Introduction

Something strange is happening in digital markets. AI agents are starting to buy things on behalf of people. Procurement bots compare supplier bids and place orders, negotiation assistants haggle over contract terms, and autonomous trading systems execute purchases with no human in the loop at the moment of transaction. The shift from markets full of human decision-makers to markets full of algorithmic proxies is moving faster than the research meant to keep up with it.

A growing literature shows that the large language models powering these agents are not the rational maximizers classical theory would wish for. Bini et al. (2026) provide the most comprehensive evidence so far: preference-based biases in LLMs (framing effects, anchoring, decoy susceptibility) become more human-like as models scale up, which for preference tasks means more biased. Ross et al. (2024) test LLMs against expected utility theory and find context-dependent risk attitudes, unstable loss aversion, and inconsistent time preferences across model families. Macmillan-Scott and Musolesi (2024) show that LLMs are irrational, but not in the same ways humans are, layering new patterns of inconsistency on top of familiar biases. Hagendorff et al. (2023) link this to dual-process theory: LLMs show both "System 1" heuristic reasoning (fast, intuitive, prone to shortcuts) and slower, more deliberative "System 2" processing. Echterhoff et al. (2024) find that anchoring, framing, and group attribution biases persist across decision tasks and resist standard prompt-based mitigation.

These are important results, but they share a limitation. Every one of these studies looks at LLMs as isolated decision-makers responding to fixed stimuli designed by a researcher. None ask what happens when a biased LLM buyer faces a strategically motivated LLM seller. In a single-agent study, the choice environment is neutral and fixed. In real commerce, it is built by a counterparty with every incentive to make it exploitative. Horton et al. (2023) coined the term "homo silicus" for LLMs used as simulated economic agents. This paper studies what happens when two instances of homo silicus face each other across a bargaining table, and one of them has been told to win.

Two findings drive the rest of the paper. First, LLM agents are *selectively vulnerable*: more an-

chored than humans in negotiation, weakly susceptible to decoys, partially cursed in common-value auctions, but immune to information overload and equilibrium-rational in private-value auctions. Second, the same anchoring bias that moves bilateral negotiations by only a few dollars produces *market-level amplification* when embedded in a multi-round double auction, collapsing allocative efficiency from $\sim$96% to $\sim$15% pooled across the three frontier models. Both findings fit a simple reading from behavioral game theory: the agents act like boundedly strategic players who reason only a few steps ahead about their counterparties, and how many steps they reason depends on the model and the commercial setting (Section 2.3).

The paper runs seven controlled experiments. Module 1 tests the attraction effect: whether a strategic seller given freedom to design its own product lineup steers a buyer more effectively than a static, experimenter-designed decoy. Module 2 tests information overload: whether more attributes, especially when a strategic seller controls the framing, hurt a buyer's ability to find the best supplier. Module 3 tests anchoring and framing in bilateral negotiation, building on Bianchi et al. (2024)'s NegotiationArena (Spearman correlation of $\rho = 0.716$ between opening offers and final prices). Module 4 tests bidding in first-price, second-price, and common-value auctions, building on Shah et al. (2024)'s evidence of overbidding and winner's curse in LLM bidders. Module 5 tests strategic exploitation: whether an agent that knows about a specific bias can exploit a naive counterpart, and whether warning the target helps. Module 6 tests debiasing: whether specific warnings, generic rationality instructions, or chain-of-thought reasoning reduce bias under adversarial conditions. Module 7 tests market aggregation: whether individual biases compound or wash out in a multi-agent double auction.

All seven modules run on three frontier models (Claude Sonnet 4, GPT-4o, and Gemini 2.5 Pro). I treat these as parallel experimental dimensions rather than as a primary model with robustness checks, following Ríos et al. (2025)'s point that model-family heterogeneity is a first-order source of variation. Treatments are randomized to fresh agent instances with no memory of earlier runs, and outcomes are benchmarked against both rational predictions and prior empirical estimates.

The headline numbers behind the two findings: Pearson $r = 0.642$ between opening offers

and final prices (against the meta-analytic $r = 0.497$ from Guthrie and Orr (2006)); efficiency collapse from roughly 96% to 15% pooled across the three models; \$9.97 of surplus per negotiation transferred from buyer to seller under strategic anchoring exploitation; and \$4.01 surplus recovery under specific bias warnings, but a $-$\$1.59 backfire under chain-of-thought prompts. The paper also tests whether a strategic seller given freedom to design its own choice architecture can amplify buyer biases beyond a static manipulation. This holds for anchoring exploitation (Module 5) but is *rejected* for decoys (Module 1) and information overload (Module 2), where the strategic seller either backfires or has no effect.

The findings bear on live regulatory debates. The FTC's enforcement actions against dark patterns (Federal Trade Commission, 2022), the DOJ's 2024 antitrust complaint against RealPage for algorithmic pricing coordination (U.S. Department of Justice, Antitrust Division, 2024), and the broader idea of libertarian paternalism (Thaler and Sunstein, 2008) all assume that the entity facing a choice environment is a human. When both the chooser and the architect of the environment are algorithms acting as proxies for humans, the normative foundation of these frameworks needs rethinking.

Section 2 reviews the relevant literatures. Section 3 states the seven hypotheses. Section 4 describes the experimental design across all seven modules. Section 5 reports results. Section 6 discusses theoretical and policy implications. Section 7 concludes.

2 Related Literature

2.1 Behavioral Biases in Individual LLMs

The classic heuristics-and-biases program of Tversky and Kahneman (1974, 1981) and the attraction-effect work of Simonson and Tversky (1992) established the patterns this paper studies in artificial agents: anchoring, framing reversals, and decoy susceptibility. These patterns, once studied only in humans, now appear in large language models.

A rapidly growing body of work documents systematic biases in frontier LLMs across exactly

these dimensions. Bini et al. (2026) provide the most comprehensive assessment so far and find the asymmetry that matters most for agentic commerce: preference-based biases (framing, anchoring, decoy susceptibility) become *more* human-like as models scale up, while belief-based tasks like Bayesian updating move toward rational benchmarks. Since commercial transactions are preference decisions, the most capable models deployed in the most consequential commercial roles are also the most biased ones. The same preference-side fragility shows up in Ross et al. (2024) (unstable risk and time preferences), Macmillan-Scott and Musolesi (2024) (irrationality in distinctly non-human forms), Hagendorff et al. (2023) (a dual-process pattern that mirrors System 1/System 2), and Echterhoff et al. (2024) (anchoring, framing, and group-attribution biases that resist standard debiasing prompts). Campbell et al. (2025) push the magnitudes further, finding LLM biases roughly double the human meta-analytic baselines ($\bar{d} = 1.87$ vs. human $\bar{d}$ between 0.58 and 0.87) and immunity to anchoring only when the correct answer is retrievable from training data ($d = 0.04$ vs. $d = 4.14$ under genuine uncertainty). Lou and Sun (2025) find that even explicit instructions to ignore an anchor fail to eliminate the effect, which sets up this paper's negotiation module directly.

Horton et al. (2023) proposes treating LLMs as "homo silicus," computational models of human behavior that can be given preferences and incentives and studied through simulation, and uses the framework to replicate Samuelson and Zeckhauser (1988) and Kahneman (2011). The present paper extends that scaffolding from isolated subjects to adversarial market participants. The literature establishes three things: LLMs are predictably biased, their biases are context- and model-specific, and the most commercially relevant biases get worse with capability. What it does not yet test is whether those biases survive strategic interaction with a counterparty who has every incentive to exploit them.

2.2 Multi-Agent and Market Settings

A parallel literature places LLMs in interactive settings, but asks different questions. Bianchi et al. (2024) build NegotiationArena, a platform for LLM-to-LLM bargaining, and find strong anchor-

ing: the Spearman correlation between the seller's opening offer and the final price is 0.716. Agents given strategic personas ("cunning," "desperate") earn about 20% higher payoffs, so strategic instructions shift the distribution of surplus. Bhattacharya et al. (2025) find that different model families adopt distinct bargaining styles in ultimatum and Nash bargaining games. Abdelnabi et al. (2024) show that LLM agents can be exploited through protocol manipulation, including cross-agent attacks.

Ríos et al. (2025) study strategic behavior across frontier models and find that different model families settle into distinct equilibria rather than converging on the rational prediction. The implication is uncomfortable: biases do not self-correct with capability. Each model develops its own stable pattern of deviation, so real-world agentic markets with diverse model families will produce model-dependent dynamics rather than any single equilibrium.

In auctions, Shah et al. (2024) find risk-averse overbidding in first-price formats and winner's curse in common-value settings, both consistent with decades of human auction experiments (Cox et al., 1988; Kagel and Levin, 1986). Chen et al. (2023) build AucArena, a multi-round auction that tests strategic planning, and find that even GPT-4 is sometimes beaten by simple heuristic baselines. Crawford and Iriberri (2007) account for these auction patterns with level-k thinking (Section 2.3). This paper tests whether these patterns persist in sealed-bid auctions with multiple LLM bidders and whether individual-level biases aggregate into market-level price distortions.

At the market level, Bansal et al. (2025) simulate two-sided agentic markets and find a severe first-proposal bias: early offers largely determine final terms, and welfare gets worse as the market grows. The direction of scaling matters: per-transaction biases do not average out as markets get larger, they compound. Zhu et al. (2025) study consumer markets with heterogeneous LLM agents and show that weaker agents are systematically exploited by stronger counterparts, producing welfare asymmetries that resemble real-world power imbalances between sophisticated sellers and unsophisticated buyers. Henning et al. (2025) place LLMs in experimental financial markets and find that LLM traders show less herd behavior than humans and do not converge to human-like market dynamics, trading instead near fundamental value with reduced variance, so agent-populated

markets may produce qualitatively different outcomes rather than replicating known anomalies. On collusion, Calvano et al. (2020) provide the foundational result: Q-learning pricing agents spontaneously learn collusive strategies in repeated pricing games, sustaining supra-competitive prices through reward-punishment schemes that emerge without any explicit communication. Fish et al. (2024) extend this directly to LLMs and provide evidence that LLM-based pricing agents may collude even without any intention from their users, raising hard questions about how antitrust law should treat coordination that arises from algorithmic interaction rather than human agreement. Brown and MacKay (2023) and Assad et al. (2024) document supra-competitive outcomes in real markets where competing firms use similar pricing algorithms.

This paper sits in the gap between the two literatures. The single-agent studies document biases in isolation but do not test how the same models behave across multiple commercial domains under a common protocol, so it is still unclear whether "LLM bias susceptibility" is one coherent property or a collection of domain-specific vulnerabilities. The multi-agent studies capture interaction effects but stop at the bilateral level, leaving the aggregation question open: do per-transaction biases wash out, persist, or compound when embedded in a multi-round market? The empirical contribution here is twofold. First, running the same three frontier models across four bias domains produces a *selective vulnerability profile* in which the same model can be more anchored than humans in negotiation, immune to overload in procurement, and equilibrium-rational in private-value auctions. Second, testing the negotiation bias in a multi-round double auction produces a *market-level amplification* result in which a per-transaction bias of a few dollars produces near-total efficiency collapse at the institutional level. As a secondary prediction, the paper also tests whether a strategically motivated seller agent given freedom to design its own choice architecture can amplify buyer biases beyond a static manipulation. The data partially confirm this for anchoring exploitation in Module 5 but reject it in both the decoy module (Module 1) and the information overload module (Module 2), giving a sharper picture of the limits of adversarial design than existing single-condition studies.

2.3 Behavioral Game Theory

Behavioral game theory provides a framework for these deviations. It models strategic actors who reason about others to a limited depth instead of solving for equilibrium directly (Camerer, 2003). In the level-k and cognitive-hierarchy models (Nagel, 1995; Stahl and Wilson, 1995; Camerer et al., 2004), a level-0 player acts naively or at random, a level-1 player best-responds to a level-0 player, and so on. Observed play in guessing games, auctions, and matrix games usually concentrates on just one or two steps of this reasoning. Crawford and Iriberri (2007) show that the same machinery explains the winner's curse and overbidding without any appeal to equilibrium. Quantal response equilibrium (McKelvey and Palfrey, 1995) adds a related idea: players best-respond with some noise, and a precision parameter controls how closely their choices track payoffs. Together these models describe behavior that is partly strategic and partly error-prone, which is the kind of behavior the agents in this study display.

A recent literature applies these models to LLMs. Kader and Lee (2024) find that model behavior in classic games matches finite-depth reasoning more than equilibrium play. Jia et al. (2025) run a set of behavioral games and recover model-specific reasoning depths that move with framing. Liu et al. (2025) build a cognitive-hierarchy benchmark and find that frontier models cluster at low levels. This paper builds on that work in three ways. First, it moves from abstract games to four *commercial* domains under a common protocol and asks whether a model's reasoning depth holds steady or changes from one setting to the next. Second, it places the bilateral interaction inside a multi-round market and tests whether shallow play builds into market-level distortions or cancels out. Third, it lets the adversarial counterparty be a player in its own right: a seller agent that designs its own choice architecture is, in level-k terms, a higher-level player that tries to exploit the buyer's expected shallowness. The analysis uses these models as an interpretive frame and stops short of fitting reasoning levels to the data statistically, which is left for future work.

2.4 Regulatory Context

The questions studied here connect to live regulatory debates on both the consumer protection and antitrust sides. Luguri and Strahilevitz (2021) define dark patterns as interface designs that manipulate consumer choice and distinguish mild from aggressive variants, with the aggressive ones producing much larger welfare effects. The FTC has brought multiple enforcement actions against dark patterns under Section 5 (Federal Trade Commission, 2022), and the agency's 2024 "Operation AI Comply" explicitly flagged AI-driven manipulation of consumer behavior (Federal Trade Commission, 2024). The strategic seller condition in Modules 1 and 2 is the agent-level version: an LLM seller builds a product lineup designed to exploit the buyer's decision-making, in the same way a human designer would build a deceptive interface.

Sunstein (2022) develops the idea of "sludge," the excess friction or complexity imposed to discourage beneficial action, and proposes sludge audits as a policy tool. The strategic information overload treatment in Module 2 maps onto this directly: the seller agent generates sludge not through administrative barriers but through information structuring that dilutes the signal of the dominant option. Bubb (2015) argues that mandatory disclosure can actually reduce welfare when information becomes a burden in itself. Module 2 is designed to test that prediction in the AI-to-AI procurement setting.

On the antitrust side, Ezrachi and Stucke (2020) argue that algorithmic tacit collusion can be sustained without any explicit agreement. The claim has gained empirical support from Calvano et al. (2020) and Fish et al. (2024) and legal traction through the DOJ's 2024 complaint against RealPage for algorithmic pricing coordination (U.S. Department of Justice, Antitrust Division, 2024) and the Preventing Algorithmic Collusion Act introduced in the Senate as S. 232 (Klobuchar et al., 2025). The market simulation module (Module 7) tests whether biased agents in a multi-round double auction produce price patterns that depart from competitive equilibrium, addressing this debate from the angle of bias-driven rather than learning-driven coordination. Thaler and Sunstein (2008) provides the backdrop: libertarian paternalism assumes a human chooser whose welfare is improved by good choice architecture, but when both chooser and designer are algorithms, the

normative analysis needs revision. Section 6 develops these connections.

3 Hypotheses

The seven predictions below combine two premises from Section 2: LLMs show preference-based biases that grow with capability (Bini et al., 2026; Ross et al., 2024; Macmillan-Scott and Musolesi, 2024), and strategic interaction between LLM agents sharpens rather than disciplines those biases (Bianchi et al., 2024; Ríos et al., 2025; Zhu et al., 2025). Two of these predictions also have a strategic-depth reading (Section 2.3): a seller that designs the choice environment (H1), or an agent that exploits a bias it knows the counterpart holds (H5), is acting as a higher-level player betting that the buyer reasons only a step or two ahead. The bet pays off only if it reads the buyer's depth correctly, so the strategic conditions need not produce uniformly larger effects.

H1 (Decoy Effect and Strategic Amplification). A buyer agent shows the attraction effect: the probability of choosing the target product rises when an asymmetrically dominated decoy is present, relative to a two-option control (Simonson and Tversky, 1992; Bini et al., 2026). The less obvious prediction is that a strategic seller given freedom to design its own lineup amplifies the effect beyond a static decoy. Formally, with β_1 the static decoy effect and β_2 the strategic seller effect (both relative to control), the hypothesis is $\beta_2 > \beta_1 > 0$.

H2 (Information Overload and Strategic Degradation). Buyer agent accuracy in finding a dominant supplier falls as the number of attributes grows (Iyengar and Lepper, 2000; Echterhoff et al., 2024), and the decline is steeper when a strategic seller generates the information (Bubb, 2015). Stated confidence is predicted to stay high even as accuracy drops.

H3 (Anchoring and Framing in Negotiation). Agreed prices show positive anchoring to the seller's opening offer, with the anchoring coefficient well above the rational benchmark of zero (Bianchi et al., 2024; Lou and Sun, 2025; Tversky and Kahneman, 1974). Loss framing reduces the buyer's share of surplus relative to gain framing (Tversky and Kahneman, 1981; Bini et al., 2026).

H4 (Auction Biases). In first-price sealed-bid auctions, buyer agents overbid relative to the risk-

neutral Bayesian Nash equilibrium (Cox et al., 1988; Shah et al., 2024). In second-price auctions, a nontrivial fraction of bids exceed the bidder's private value (violating the dominant strategy of truthful bidding). In common-value auctions, winning bids exceed the true value (winner's curse, Kagel and Levin, 1986).

H5 (Strategic Exploitation). An agent that knows about a specific bias can exploit a naive counterpart, producing larger bias effects than the naive-vs-naive baseline (Zhu et al., 2025; Luguri and Strahilevitz, 2021). When the target is warned about the bias, exploitation shrinks but does not vanish (Lou and Sun, 2025).

H6 (Debiasing Interventions). Specific warnings that name the bias outperform generic rationality instructions. Chain-of-thought reasoning (Wei et al., 2022) gives intermediate benefit by engaging the System 2 processing Hagendorff et al. (2023) document in LLMs. The baseline expectation that all of these only partly work comes from Echterhoff et al. (2024).

H7 (Market-Level Distortions). Individual biases aggregate into market-level distortions rather than washing out through competition. Anchoring and loss framing lower allocative efficiency and raise price dispersion. Markets that mix strategic and naive agents show larger welfare asymmetries than homogeneous markets (Zhu et al., 2025). Chain-of-thought debiasing at the agent level improves market efficiency. The strong form of the hypothesis is that the double auction (Smith, 1962; Gode and Sunder, 1993) fails to discipline LLM biases, which would have direct implications for platform design.

4 Research Design

4.1 Agent Architecture

Each agent is an instance of a large language model started with a system prompt that specifies its role, objective function, and private information. Three frontier models serve as parallel experimental dimensions: Claude Sonnet 4 (Anthropic), GPT-4o (OpenAI), and Gemini 2.5 Pro (Google). The Gemini arm uses 2.5 Pro, Google's frontier tier, so that all three arms are matched at a compa-

rable capability level: 2.5 Pro is the natural peer of Claude Sonnet 4 and GPT-4o, whereas Google's cost- and latency-optimized 2.5 Flash tier is not. An initial pilot on 2.5 Flash was set aside for two reasons. It would have left the Gemini arm a capability tier below the other two models, undercutting the cross-model comparison; and it produced unreliable structured output, with low JSON-formatting success rates that a like-for-like comparison should not have to engineer around. These three are the frontier commercial models most likely to be deployed as production agents at the time of writing; open-weight families (Llama, Mistral, Qwen) are excluded because they are uncommon in commercial agent stacks today and would expand the design beyond the 8,415-record budget. This follows the "homo silicus" framework of Horton et al. (2023) while taking seriously Ríos et al. (2025)'s finding that different model families settle into distinct equilibria, which makes model identity a first-order variable. An orchestrator manages turn-taking, passes structured JSON outputs between agents, and logs everything. Each run uses a fresh agent instance with no memory of prior runs, so there are no carry-over effects.

Temperature is set to both 0.0 and 0.7. The 0.7 setting introduces stochastic sampling that better approximates deployment conditions. Each treatment cell contains 30 runs per model-temperature combination, but the per-cell count varies by module because some modules cross-tabulate over additional dimensions: the decoy module multiplies through three product domains, raising the target to 180 obs per (treatment, model) cell; negotiation and info overload contain 62 obs/cell; the auction and exploitation modules contain 184–186 obs per treatment pooled across models; the debiasing module 183–184 obs per strategy (pooling models and temperatures within strategy); and the market simulation 56–60 obs per (treatment, model) cell. Per-cell counts in Table 1 reflect these design targets. Where applicable, domain variants control for stimulus sensitivity: the decoy module runs across three product domains (electronics, hotels, and software subscriptions), following Echterhoff et al. (2024) and Macmillan-Scott and Musolesi (2024), who show this is necessary given how sensitive LLM behavior is to wording.

Causal identification rests on randomized assignment of treatments to fresh agent instances. There is no self-selection into treatment, no spillover across runs, and no unobserved confounding.

Multiple testing across modules is handled with Benjamini-Hochberg FDR correction. Each module designates one pre-specified primary outcome as the confirmatory test; the rest are exploratory.

The 30-runs-per-cell design target gives a back-of-envelope minimum detectable effect (two-sided t-test, $\alpha = 0.05$, power $= 0.80$) of Cohen's $d \approx 0.74$ at the (model, treatment, temperature) cell level. Pooling across temperature ($n = 60$) reduces the MDE to $d \approx 0.51$ per (model, treatment) cell, and the fully pooled (model-pooled) cell with $n = 180$ reaches $d \approx 0.30$. The main efficiency-collapse effect ($d = 6.43$ for the baseline-vs-anchored market contrast) is well above all three thresholds, and even the smallest market cell ($n = 56$) has power ≈ 1.00 against the baseline.

4.2 Module 1: Decoy Effects in Agent Choice

A seller agent presents product options to a buyer agent whose objective is to maximize value for money. Products vary on two attributes: quality (1 to 10 scale) and price. The target product T has quality 7 and price \$70. The competitor product C has quality 5 and price \$50. The decoy product D has quality 6 and price \$75, making it asymmetrically dominated by T but not by C, following the classic attraction effect design of Simonson and Tversky (1992).

Three conditions are tested. In the control, the buyer chooses between T and C only. In the static decoy condition, the buyer chooses among T, C, and D. In the strategic seller condition, the seller agent receives the objective of maximizing the probability that the buyer selects T and is permitted to construct its own three-option product lineup, subject only to the constraint that T remains in the set. The primary outcome is $P(\text{choose } T \mid \text{treatment})$:

$$\text{logit}\, P(\text{choose } T) = \beta_0 + \beta_1 \cdot \textit{StaticDecoy} + \beta_2 \cdot \textit{StrategicSeller} + \varepsilon \tag{1}$$

The key test is $\beta_2 > \beta_1$: the strategic seller amplifies the decoy effect beyond the static manipulation.

4.3 Module 2: Information Overload in Procurement

A buyer agent picks the best supplier from a set of four options, one of which (Supplier 1) is objectively dominant on all criteria after applying pre-specified weights. Treatments vary on two dimensions: the number of attributes per option (3, 6, 12, or 24) and the information source (neutral or strategic). The setup borrows from the choice-overload literature of Iyengar and Lepper (2000) (who varies the number of options, not attributes per option), with the attribute-volume manipulation following Malhotra (1982). In the neutral condition, the experimenter generates the attributes in a balanced format. In the strategic condition, a seller agent generates the attribute presentation with the goal of obscuring the dominant option without falsifying any data.

Primary outcomes are $P(\text{correct choice})$, stated confidence (0 to 100 scale), and the number of attributes cited in the buyer's reasoning:

$$\text{logit}\, P(\text{correct}) = \beta_0 + \beta_1 \cdot \log(n_{\text{attr}}) + \beta_2 \cdot \textit{Strategic} + \beta_3 \cdot (\log(n_{\text{attr}}) \times \textit{Strategic}) + \varepsilon \qquad (2)$$

The key test is $\beta_3 < 0$: strategically-generated complexity degrades accuracy more steeply than neutral complexity.

4.4 Module 3: Bilateral Negotiation

A buyer agent and a seller agent negotiate price over a simulated transaction. The seller's cost is \$40 and the buyer's value is \$100, both common knowledge, yielding a total surplus of \$60. Agents alternate offers for a maximum of 10 rounds, with the seller moving first. This setup follows the bilateral bargaining framework of Bianchi et al. (2024) with the addition of treatment manipulations. The common-knowledge assumption is a deliberate scope condition: it isolates anchoring from private-information bargaining dynamics, but it means the results below should not be extrapolated to negotiations in which one side has informational advantages of the kind studied in the asymmetric-information bargaining literature.

Five treatment conditions are tested: a baseline with neutral framing, an anchor-high condition

(the seller's system prompt contains a market reference price of \$95), an anchor-low condition (reference price of \$45), a loss-framing condition (negotiation framed as avoiding losses rather than capturing gains, following the framing logic of Tversky and Kahneman (1981)), and an outside-option condition (buyer holds a private outside option worth \$70).

Outcomes include the agreed price, the anchoring coefficient from FinalPrice $= \beta_0 + \beta_1 \cdot$ *OpeningOffer* $+ \varepsilon$ (rational benchmark: $\beta_1 = 0$; NegotiationArena benchmark: Spearman $\rho = 0.716$), the agreement rate, rounds to agreement, and buyer surplus. The treatment regression is:

$$\text{FinalPrice} = \beta_0 + \beta_1 \cdot \textit{AnchorHigh} + \beta_2 \cdot \textit{AnchorLow} + \beta_3 \cdot \textit{LossFraming} + \beta_4 \cdot \textit{OutsideOption} + \varepsilon \quad (3)$$

estimated by OLS with heteroskedasticity-robust standard errors.

4.5 Module 4: Auction Mechanisms

Two sub-modules test auction behavior. In Sub-module A (private value), three buyer agents receive values drawn independently from Uniform[50, 100]. The risk-neutral Bayesian Nash equilibrium (BNE) in the first-price sealed-bid format predicts a bid of $\frac{n-1}{n} \times$ value $= \frac{2}{3} \times$ value (Cox et al., 1988). In the second-price format, bidding one's true value is a dominant strategy because the price paid is set by the next-highest bid, so shading down only risks losing the object without lowering the price. In Sub-module B (common value), the true value $V \sim$ Uniform$[50, 100]$ and each buyer observes a signal $V + \epsilon$ with $\epsilon \sim$ Uniform$[-10, 10]$. Because the bidder with the most optimistic signal is the most likely to win, rational bidders should shade below their signal to avoid the resulting adverse-selection penalty. The winner's curse is measured as the winning bid minus the true value, which should be zero or negative once that shading is in place (Kagel and Levin, 1986).

4.6 Module 5: Strategic Exploitation

Modules 1 through 4 document biases; this module asks whether they can be weaponized. Four bias domains are tested (anchoring, decoy, information overload, and auction), each under three

conditions: *naive vs. naive* (neither agent has bias knowledge), *exploit vs. naive* (the exploiter knows the relevant bias and is told to use it; the target is naive), and *exploit vs. defend* (the exploiter has bias knowledge; the target is warned about the bias and told to resist). The setups mirror the corresponding core modules: the anchoring domain uses a multi-turn negotiation (seller cost \$40, buyer value \$100, max 10 rounds); the decoy domain uses a single-shot product choice task; the information overload domain uses supplier selection with 12 attributes; and the auction domain uses a common-value sealed-bid format with three bidders (true value from Uniform[50, 100], signals with ± 10 noise). The primary outcome is the exploitation effect size, defined as the difference in bias magnitude between exploit-vs-naive and naive-vs-naive, and the defense effect size, which measures how much of the exploitation the defend condition reverses. The design follows Abdelnabi et al. (2024) but extends from negotiation to four commercial domains.

4.7 Module 6: Debiasing Interventions

If biases can be exploited, the next question is whether they can be cheaply corrected. The anchoring domain, which produced the strongest base bias, is tested under four debiasing strategies applied to the buyer: *no debias* (baseline: naive agent against an exploiter), *specific warning* (the buyer is told which bias it faces and how to resist), *generic rationality* (the buyer is told to "identify objective criteria and evaluate on merits only," with no named bias), and *chain of thought* (the buyer must produce step-by-step reasoning before its final decision, following Wei et al. (2022)). In all conditions the seller uses the exploit strategy from Module 5. The primary outcome is the change in bias magnitude relative to the no-debias baseline. The policy question is whether a generic intervention that regulators could mandate without knowing which bias an agent faces works as well as a targeted warning. Echterhoff et al. (2024) find that standard debiasing prompts have limited effect in single-agent settings; this module tests whether the same holds under adversarial pressure.

4.8 Module 7: Market Simulation

Modules 1 through 6 study biases at the individual transaction level. This module asks whether those biases survive aggregation. The design uses a continuous double auction, the workhorse institution of experimental economics since Smith (1962), with N buyers and N sellers (default: 5 each) trading over K rounds (default: 5). Each buyer draws a private value from Uniform[50, 100] and each seller draws a private cost from Uniform[10, 60]. In each round, all agents simultaneously submit bids (buyers) or asks (sellers); bids at or above asks are matched in order of aggressiveness, and the transaction price is set at the midpoint. Gode and Sunder (1993) showed that even zero-intelligence traders (agents that bid randomly within their budget constraint, with no strategic reasoning) produce near-efficient outcomes in a double auction, so any efficiency loss attributable to bias would be notable.

Five treatment conditions are tested: *baseline* (all agents naive), *anchored* (sellers receive a high market reference price of \$85–\$95), *loss framing* (all agents receive loss-frame prompts), *mixed strategic* (half the sellers are strategic exploiters with anchoring knowledge, the rest naive), and *all debiased* (all agents receive chain-of-thought debiasing). Primary outcomes are market price, allocative efficiency (total surplus captured as a fraction of the competitive equilibrium maximum), price dispersion (standard deviation of transaction prices within a round), and convergence (absolute distance between the final-round average price and the theoretical equilibrium price).

4.9 Summary of Experimental Design

Table 1 summarizes the modules, treatments, sample sizes, and benchmarks.

5 Results

Each module was run across three models (Claude Sonnet 4, GPT-4o, Gemini 2.5 Pro) at two temperatures (0.0 and 0.7) with 30 iterations per treatment cell. The final dataset has complete and balanced coverage for all three models across all seven modules: Claude (2,838 records), GPT-4o

Table 1: Experimental Design Summary

Module	Conditions	N/cell	Benchmark	Primary Outcome
1: Decoy	Control, Static, Strategic	180	No attraction effect	$P(\text{choose } T)$
2: Overload	4 attr. levels $\times$ 2 sources	62	Perfect accuracy	$P(\text{correct})$
3: Negotiation	Baseline, Anchor$\times$2, Loss, Option	62	β_1=0 (no anchoring)	Agreed price
4a: Priv. value	First-price, Second-price	184	Equil. pred. (b=$\frac{2}{3}v$)†	Bid/value ratio
4b: Com. value	Sealed-bid	184	No winner's curse	Bid $-$ true value
5: Exploitation	4 domains $\times$ 3 conditions	184–186	No exploitation effect	Bias amplification
6: Debiasing	Anchoring $\times$ 4 strategies	183–184	No bias reduction	Bias reduction
7: Market	5 treatments, K rounds	56–60	Competitive equil.	Efficiency, price

†The risk-neutral first-price BNE prediction $b = \frac{2}{3}v$ is both the theoretical benchmark and, with the Gemini arm now on 2.5 Pro, the realized outcome: Module 4 shows that all three models bid at essentially this ratio (0.667–0.671), so the column can be read as the equilibrium prediction the models in fact converge to.

(2,847), and Gemini 2.5 Pro (2,730, within about 4% of the other two). Cross-model coverage is enough for formal model $\times$ treatment interaction tests in every continuous-outcome module. All primary-outcome p-values are adjusted using Benjamini–Hochberg FDR correction across the full cross-module test family; raw and adjusted p-values are reported in the online Supplementary Material.

5.1 Module 1: Decoy Effects (H1)

Table 2 and Figure 1 report the probability of choosing the target product by treatment and model.

Table 2: Decoy Effect: $P(\text{choose } T)$ by Treatment and Model

Model	Treatment	$P(T)$	n
Claude Sonnet 4	Control	0.611	180
	Decoy present	0.706	180
	Strategic seller	0.061	180
GPT-4o	Control	0.000	180
	Decoy present	0.000	180
	Strategic seller	0.000	180
Gemini 2.5 Pro	Control	0.000	180
	Decoy present	0.067	180
	Strategic seller	0.089	180

H1 is partially supported. Claude shows a moderate attraction effect: a static decoy raises $P(T)$ by 9.4 percentage points (95% bootstrap CI [0.0, 19.4]; from 61.1% to 70.6%), about 56% of

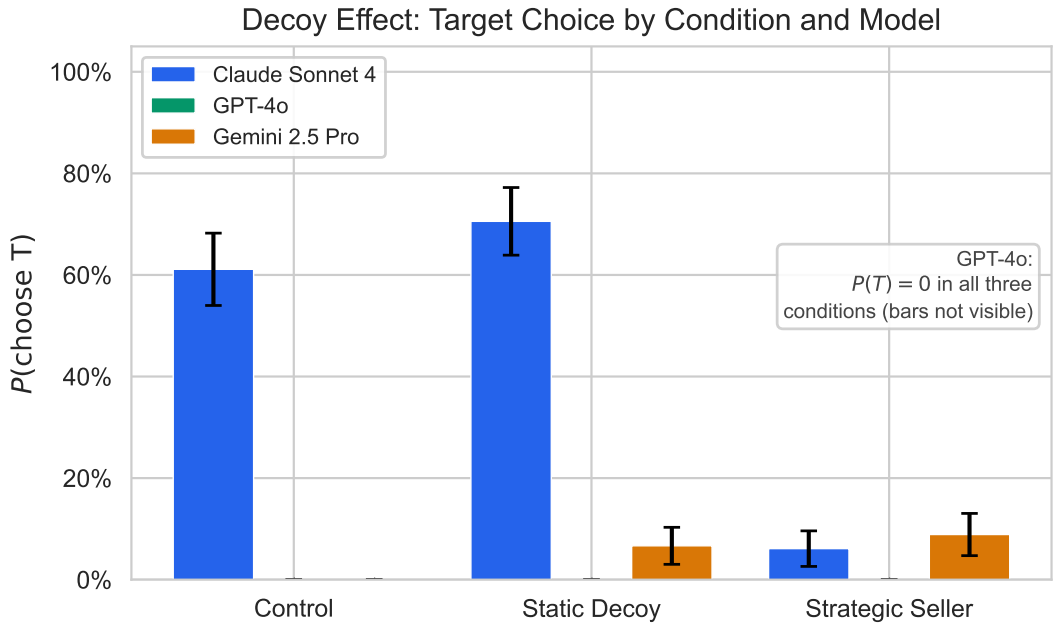

Figure 1: Decoy effect: $P(\text{choose } T)$ by condition and model. Claude shows a moderate attraction effect (control $\rightarrow$ static decoy) but the strategic seller backfires. GPT-4o is completely immune ($P(T) = 0$ across all three conditions, $n = 540$), so its bars are zero-height and not visible on the chart. Gemini 2.5 Pro shows a small but statistically significant response: a static decoy lifts $P(T)$ from 0 to 6.7%, and the strategic seller to 8.9%.

the roughly 17-percentage-point lift reported across post-1992 attraction-effect studies (Heath and Chatterjee, 1995; Huber et al., 2014). The 17-point figure is a cross-study magnitude, not a single-paper estimate. GPT-4o shows no attraction effect, and Gemini shows only a small one (below), so the 56% ratio applies to Claude. The Equation 1 logit confirms the direction for Claude ($\hat{\beta}_1 = 0.42$, $p = 0.058$). The strategic-seller treatment, however, *decreases* Claude's target choice to 6.1% ($\hat{\beta}_2 = -3.19$, $p < 0.001$), the opposite of the prediction that adversarial design amplifies the effect. GPT-4o is completely immune to the decoy effect: $P(T) = 0$ across all conditions ($n = 540$), choosing the competitor product on value-for-money grounds in every trial. Gemini 2.5 Pro is immune in the control ($P(T) = 0$) but shows a small, statistically significant attraction effect once a decoy is present: $P(T)$ rises to 6.7% with a static decoy and 8.9% under the strategic seller (lift $= +6.7$ percentage points, χ^2 $p = 0.001$, about 0.39$\times$ the 17-point human magnitude). Unlike Claude, Gemini shows no strategic-seller backfire.

To make sense of the strategic-seller backfire, I pulled the seller-generated lineups for all 540

strategic-seller runs and checked whether each proposed decoy D was actually asymmetrically dominated by the target T (the property that makes a decoy work) and *not* dominated by competitor C (which would make C a better choice than T). The three models fail in three different ways. Claude sellers create decoys that are T-dominated 99% of the time but *also* C-dominated 61% of the time, an adversarial overreach that hands the choice to C. GPT-4o sellers fail the basic decoy task: only 9% of their lineups contain a decoy that is properly T-dominated, and the other 91% are non-dominated alternatives that the buyer correctly evaluates on the merits. Gemini sellers create valid decoys 93% of the time, and Gemini buyers are no longer fully immune: a present decoy lifts target choice to 6.7% and the strategic seller to 8.9%. The susceptibility is small but real, so for Gemini seller competence does translate into a modest buyer effect.

5.2 Module 2: Information Overload (H2)

Table 3 and Figure 2 report buyer accuracy (selecting the objectively dominant supplier) by attribute count and information source.

Table 3: Information Overload: $P(\text{correct})$ by Attributes, Source, and Model

		Claude Sonnet 4		GPT-4o		Gemini 2.5 Pro	
n_{attr}	Source	$P(\text{correct})$	n	$P(\text{correct})$	n	$P(\text{correct})$	n
3	Neutral	1.000	62	1.000	62	1.000	60
3	Strategic	0.742	62	0.629	62	0.574	61
6	Neutral	1.000	62	1.000	62	1.000	60
6	Strategic	0.823	62	0.806	62	0.900	60
12	Neutral	1.000	62	1.000	62	1.000	60
12	Strategic	0.919	62	0.968	62	0.983	60
24	Neutral	1.000	62	1.000	62	1.000	60
24	Strategic	1.000	62	0.968	62	0.933	60

H2 is rejected. LLM agents are *anti-fragile* to information volume: more attributes give them more cross-references and *help* them resist manipulation, the opposite of the human pattern in Iyengar and Lepper (2000) and Malhotra (1982). Under neutral presentation, all three models hit perfect accuracy at every attribute level. Strategic framing does reduce accuracy, but the effect is concentrated at *low* attribute counts and vanishes as attributes grow. At $n = 3$ the error rates are

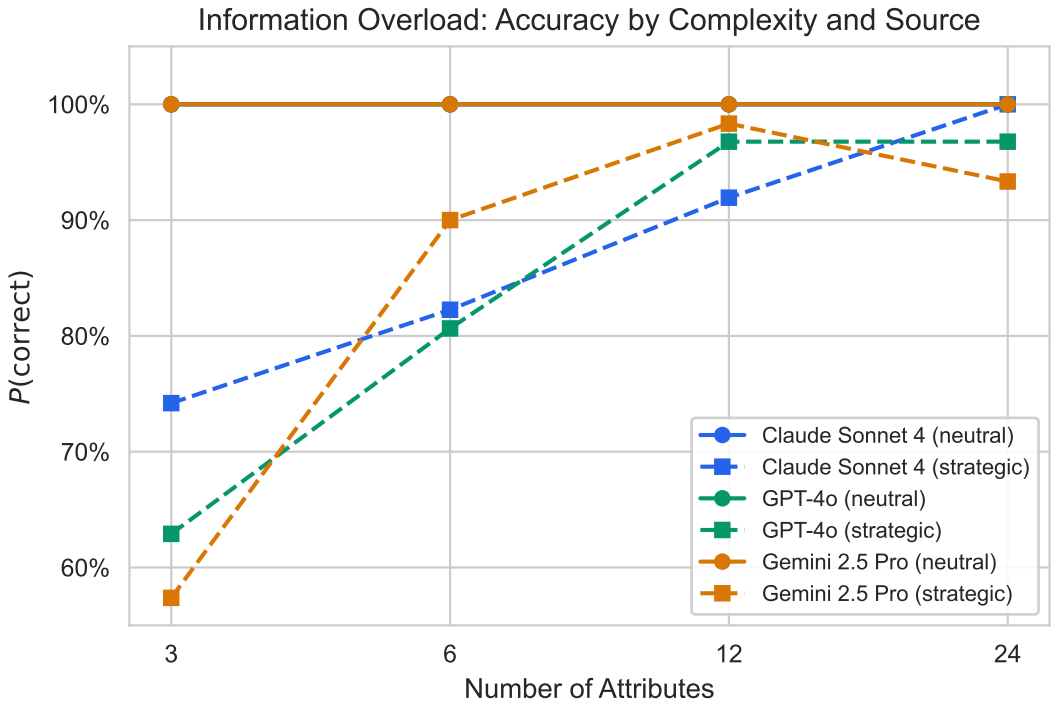

Figure 2: Information overload: accuracy by attribute count and information source. Neutral conditions (solid) yield perfect accuracy at all levels. Strategic framing (dashed) degrades accuracy at low attribute counts but the effect vanishes as complexity increases, the opposite of the human pattern.

42.6% for Gemini 2.5 Pro, 37.1% for GPT-4o, and 25.8% for Claude; by $n = 24$ all three recover to 93% accuracy or better under strategic framing. The vulnerability is shared across models and specific to data-scarce conditions: LLM agents are vulnerable when data is scarce, not when it is abundant.

5.3 Module 3: Negotiation (H3)

Table 4 and Figure 3 report negotiation outcomes by treatment and model.

The bivariate anchoring regression ($n = 852$ agreed deals) gives an OLS slope of $\hat{\beta}_1 = 0.533$ ($p < 0.001$). The Pearson correlation between opening offer and final price is $r = 0.642$ ($R^2 = 0.412$; bootstrap 95% CI on R^2: [0.36, 0.47]), which exceeds the meta-analytic human benchmark of $r = 0.497$ from Guthrie and Orr (2006) by 29%. Both figures are Pearson correlations, so the Orr comparison is within-paradigm. Bianchi et al. (2024)'s NegotiationArena reports a Spearman

Table 4: Negotiation: Mean Agreed Price and Surplus Division by Treatment

Model	Treatment	Price	SE	Agree	Seller surplus	Buyer surplus
Claude	Baseline	70.26	0.07	100%	30.26	29.74
	Anchor high	72.56	0.27	100%	32.56	27.44
	Anchor low	60.05	0.09	100%	20.05	39.95
	Loss frame	65.53	0.17	100%	25.53	34.47
	Outside option	69.00	1.00	4.8%[†]	29.00	31.00
GPT-4o	Baseline	74.52	0.19	100%	34.52	25.48
	Anchor high	78.95	0.26	100%	38.95	21.05
	Anchor low	73.65	0.45	100%	33.65	26.35
	Loss frame	71.29	0.30	100%	31.29	28.71
	Outside option	73.95	0.26	100%	33.95	26.05
Gemini	Baseline	70.33	0.27	100%	30.33	29.67
	Anchor high	74.95	0.33	100%	34.95	25.05
	Anchor low	65.25	0.23	100%	25.25	34.75
	Loss frame	68.83	0.35	100%	28.83	31.17
	Outside option	67.04	0.33	88.3%	27.04	32.96

[†]Claude buyers reject the outside-option treatment in 59 of 62 runs (see Section 6.8).

$\rho = 0.716$ on the same relationship, close in magnitude but not directly comparable to the Pearson r reported here.

The full treatment regression from Equation 3 of the methods, fit pooled across models with HC3-robust standard errors ($n = 852$), confirms the direction and magnitude of all four manipulations:

$$\widehat{\text{Price}} = 71.72 + 3.78 \cdot \textit{AnchorHigh} - 5.43 \cdot \textit{AnchorLow} - 3.17 \cdot \textit{LossFraming} - 1.00 \cdot \textit{OutsideOption} \tag{4}$$

with $p < 10^{-16}$ for AnchorHigh, AnchorLow, and LossFraming, and $p = 0.017$ for OutsideOption. The OutsideOption coefficient is based on only 118 agreed-price observations (Claude 3, GPT-4o 62, Gemini 53) after excluding Claude's 59 categorical rejections (see below); the small pooled reduction of \$1.00 is driven almost entirely by Gemini 2.5 Pro (a \$3.30 cut, $p < 10^{-13}$), while Claude and GPT-4o show no detectable shift. In relative terms, the AnchorHigh effect of \$3.78 is about 6.3% of the \$60 surplus zone, or 12.6% of the buyer's \$30 baseline share. This is a real reallocation of surplus produced by a single sentence in the seller's system prompt. The per-model

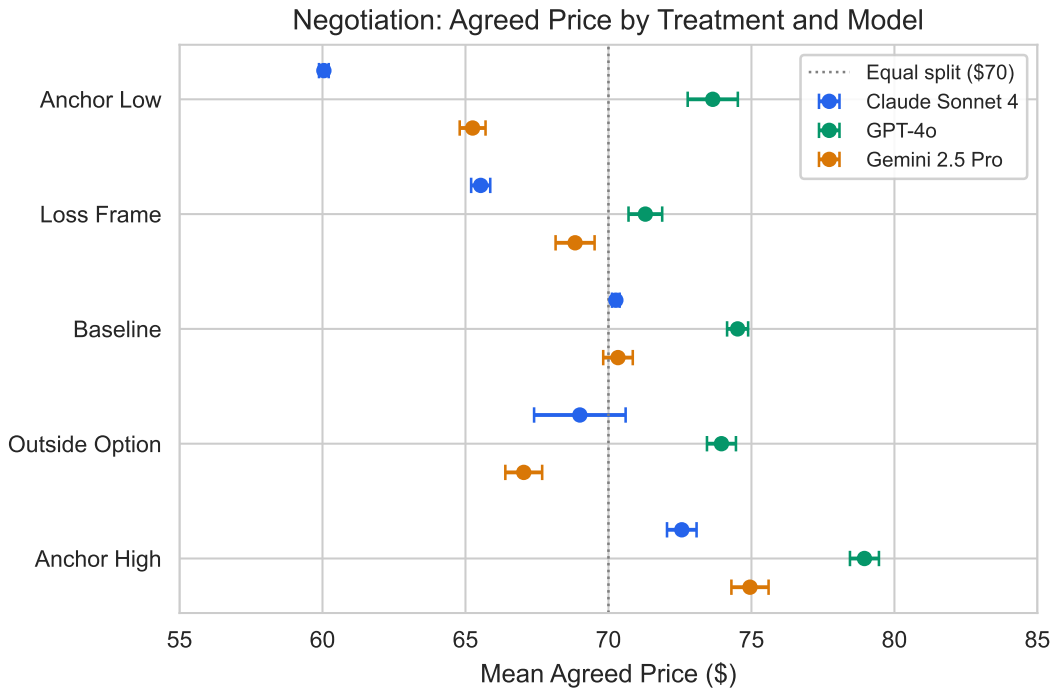

Figure 3: Negotiation: mean agreed price by treatment and model. All three models show anchor high above baseline, and both loss framing and anchor low below baseline; the specific ordering between loss framing and anchor low varies slightly by model. Dashed line marks the equal-surplus split at $70.

specifications also reveal substantial heterogeneity in anchoring susceptibility: AnchorHigh shifts prices by $2.31 in Claude (3.8% of the zone), $4.44 in GPT-4o (7.4%), and $4.62 in Gemini 2.5 Pro (7.7%), a 2.0× spread, shown in Figure 4. The formal model × treatment interaction test rejects homogeneity ($F_{8,837} = 52.88$, $p < 10^{-68}$).

H3 is strongly supported. All three models show significant anchoring, though the magnitude depends on the model. High anchors raise prices $2–5 above baseline; low anchors cut them $1–10. Loss framing consistently shifts surplus toward the buyer ($1–5 reduction in price). Agreement rates are 100% in 13 of 15 (model, treatment) cells. Both exceptions are outside-option treatments: Claude rejects in 59 of 62 runs, treating the outside option as a walk-away threshold rather than bargaining leverage, and Gemini 2.5 Pro reaches agreement in 88.3% of its outside-option runs. The failure mode is discussed in Section 6.8.

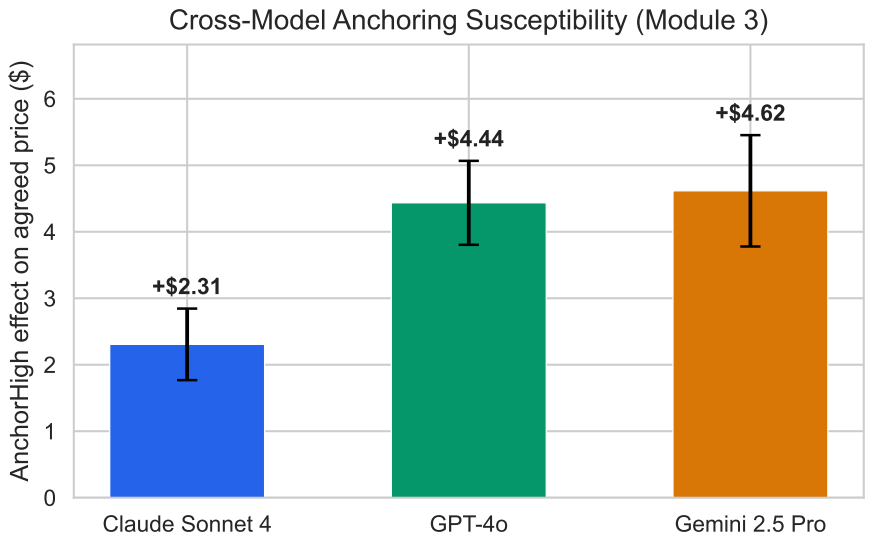

Figure 4: Cross-model anchoring susceptibility: per-model AnchorHigh treatment effect (mean price under $95 anchor minus mean baseline price), with 95% confidence intervals. Gemini 2.5 Pro absorbs the high anchor 2.0× more than Claude does. The cross-model spread is itself the key policy finding: agentic-commerce regulations targeting a single "LLM bias coefficient" will misallocate attention.

5.4 Module 4: Auctions (H4)

Figure 5 summarizes auction behavior across formats.

Private-value auctions. All three models bid at essentially the risk-neutral BNE ratio of $\frac{2}{3} = 0.667$. Claude bids at 0.667 (SE < 0.001, $n = 186$) and Gemini 2.5 Pro at 0.667 (SE < 0.001, $n = 180$), both indistinguishable from the equilibrium to four decimal places; GPT-4o bids at 0.671 (SE $= 0.001$, $n = 186$), a trivial 0.4-percentage-point excess. All three outperform humans, who consistently overbid by about 10% in first-price auctions (Cox et al., 1988). The model × auction-type interaction test is statistically significant but substantively negligible ($F_{2,1098} = 8.34$, $p < 0.001$): the cross-model spread in bid/value ratios is under half a percentage point. In second-price auctions, all three models bid truthfully at 1.0.

Common-value auctions. Across the 183 common-value runs, the unconditional mean signed curse (winning bid minus true value) is $-\$3.12$: on average, bidders win for $3.12 less than the true value, so on net there is *no* curse. The overpaid rate is 42.1%, however, and conditional on overpaying the average overpayment is $2.46, so a sizeable minority of LLM auctions still show

a textbook winner's curse. By model, Claude overpays in 24.2% of auctions, GPT-4o in 46.8%, and Gemini 2.5 Pro in 55.9%. A representative human overbid rate in common-value settings is about 60%, drawn from the post-1986 winner's-curse literature in the tradition of Kagel and Levin (1986) rather than from any single verbatim figure in that paper. The LLM rate of 42.1% is a 30% reduction (ratio $= 0.70$).

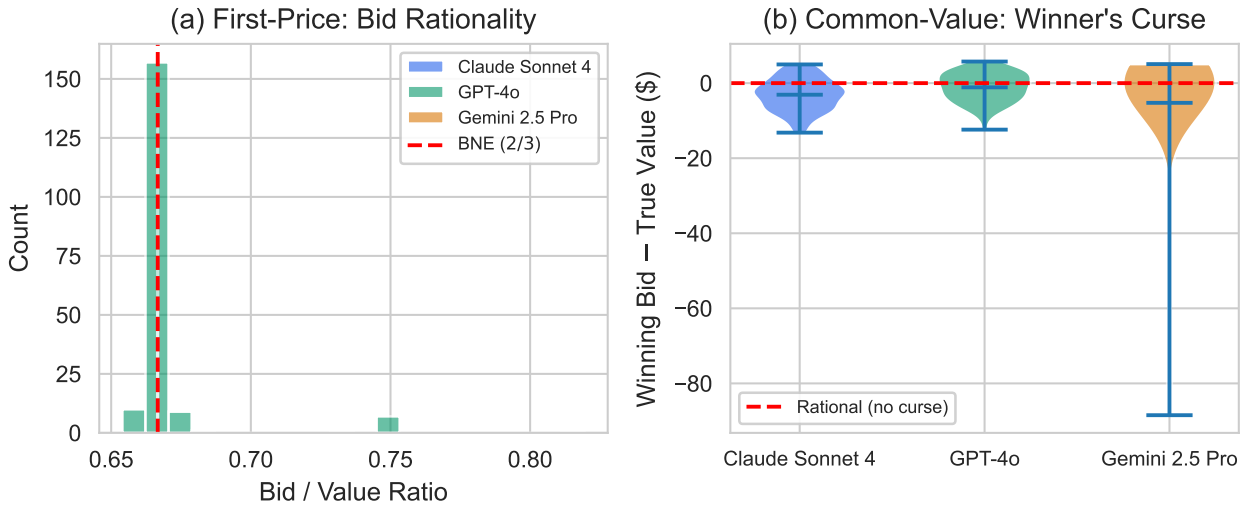

Figure 5: Auction behavior. (a) First-price bid-to-value ratios cluster tightly around the BNE prediction of $\frac{2}{3}$. (b) Common-value winner's curse: signed overpayment (winning bid − true value, in dollars) by model. Positive values indicate overpayment (the textbook curse); negative values indicate the winner won for less than the true value. The dashed red line at zero is the rational benchmark. Most auctions cluster below zero (no curse), but 42.1% of runs land above the line, with a conditional mean overpayment of $2.46.

H4 is partially supported. The winner's curse exists but is much weaker than in humans, and the private-value overbidding predicted by H4 does not show up at all. All three models play the first-price equilibrium strategy with near-perfect precision, a level of rationality rarely seen in human auction experiments. This uniformity is itself capability-linked: the earlier Gemini 2.5 Flash runs bid at 0.899, well above BNE, whereas the more capable Gemini 2.5 Pro converges to the equilibrium like its peers, so the private-value rationality finding now holds across all three families rather than for only two of them. One caveat qualifies this rationality reading. The $\frac{2}{3}$ first-price solution is a standard textbook result, so near-exact equilibrium play may simply mean the model retrieved a known answer from training instead of working out the strategy on the spot. Campbell

et al. (2025) document the same retrieval boundary for anchoring. The two explanations are hard to separate here, and telling them apart, for instance by moving the auction parameters away from textbook values, is left to future work.

5.5 Module 5: Strategic Exploitation (H5)

Table 5 and Figure 6 report exploitation outcomes by bias domain and condition.

Table 5: Strategic Exploitation: Outcomes by Domain and Condition (pooled across models)

Domain	Metric	Naive	Exploit	Defend
Anchoring	Mean price	71.64	81.60	77.55
	Seller surplus	31.64	41.60	37.55
	Surplus transfer	—	+\$9.97	+\$5.91
Info overload	$P(\text{correct})$	1.000	0.804	0.606
Decoy	$P(\text{choose } T)$	0.060	0.071	0.424
Auction	Overpaid rate	38.2%	36.4%	35.9%
	Winner's curse	1.78	1.98	2.52

Cells are pooled across the three models. Per-model decompositions for the anchoring domain appear in the online Supplementary Material; the model-specific seller-surplus values for the baseline negotiation (Table 4: Claude 30.26, GPT-4o 34.52, Gemini 30.33) average to a pooled baseline near \$32, against the \$31.64 Naive cell reported here.

The most cost-effective exploitable bias, by normalized magnitude, is *information overload*, not anchoring. Ranking the four exploitation conditions (Table 6) puts strategic complexity-flooding first: it reduces buyer accuracy by 19.6 percentage points relative to the naive baseline ($p < 10^{-9}$, 95% bootstrap CI $[-25.0, -14.1]$ pp). Anchoring exploitation ranks second, with \$9.97 of surplus per negotiation moved from buyer to seller (16.6% of the \$60 surplus zone, 95% CI [\$8.70, \$11.14]). Decoy exploitation (1.1pp lift, n.s.) and auction exploitation (\$0.20 winner's-curse increase, n.s.) are negligible. The decoy null is a pooling artifact: the exploitation lift is positive for Claude ($+9.7$pp) but slightly negative for Gemini (-6.7pp) and zero for GPT-4o, so the three models partly cancel. As in Module 1, the decoy effect is strongly model-dependent, and a single pooled exploitation estimate understates the within-model heterogeneity.

The anchoring effect depends heavily on the model. Gemini 2.5 Pro is the most exploitable: exploitation drives its prices to \$87.03 on average, handing the seller \$47.03 of the \$60 surplus

Table 6: Strategic Exploitation: ROI Ranking by Bias Domain. Rankings reflect within-domain effect sizes normalized to each domain's maximum attainable effect; cross-domain comparisons are illustrative rather than cardinal.

Rank	Domain	Effect	Normalized	Significance
1	Information overload	-19.6 pp accuracy	0.196 of attainable	$p < 10^{-9}$
2	Anchoring	$+\$9.97$ transfer	0.166 of \$60 zone	$p < 10^{-40}$
3	Decoy	$+1.1$ pp lift	0.011 of attainable	$p = 0.83$ (n.s.)
4	Auction (common value)	$+\$0.20$ curse	0.004 of value range	$p = 0.65$ (n.s.)

(78%) and leaving the buyer just 22%, against the $\sim$50% the buyer captures under the unbiased baseline. GPT-4o is close behind at \$86.16 (seller share 77%), while Claude is far more resistant, moving only to \$72.10 (seller share $\sim$54%). Defense warnings recover \$4.05 of the transferred surplus on average, about 41% of the \$9.97 transfer (95% bootstrap CI [\$2.38, \$5.56]).

Strategic complexity-flooding cuts buyer accuracy from 100% (naive vs. naive baseline) to 80.4% (exploit vs. naive). The defend condition does *worse* than the exploit condition (60.6% accuracy, $n = 183$), a 19.8-point drop that runs in the opposite direction of every other defense effect in the paper. The likely mechanism is excessive skepticism: once the buyer is told that the seller is structuring the data to mislead, it downweights the very attributes the dominant supplier scores highest on, treating quantitatively favorable information as a sign of seller manipulation rather than supplier quality. The buyer reasoning traces in the defend cells back this up. Warned buyers flag "suspiciously high" scores on the dominant supplier as a reason to discount them, and then pick an objectively worse alternative on the grounds that "the seller is trying to push me toward Supplier 1." Warnings calibrated for adversarial intent backfire when the information is actually informative: an LLM told to distrust its inputs cannot tell manipulation from honest data that happens to favor the seller. This is a different failure mode from the chain-of-thought backfire in Module 6, which works through anchor salience. Here, the defense damages the buyer's ability to read accurate data, not its ability to resist a salient cue.

H5 is strongly supported for anchoring exploitation. Informed agents can extract substantial surplus from naive counterparts, and defense warnings give partial but incomplete protection. The auction domain shows minimal exploitation effects, consistent with the near-rational baseline

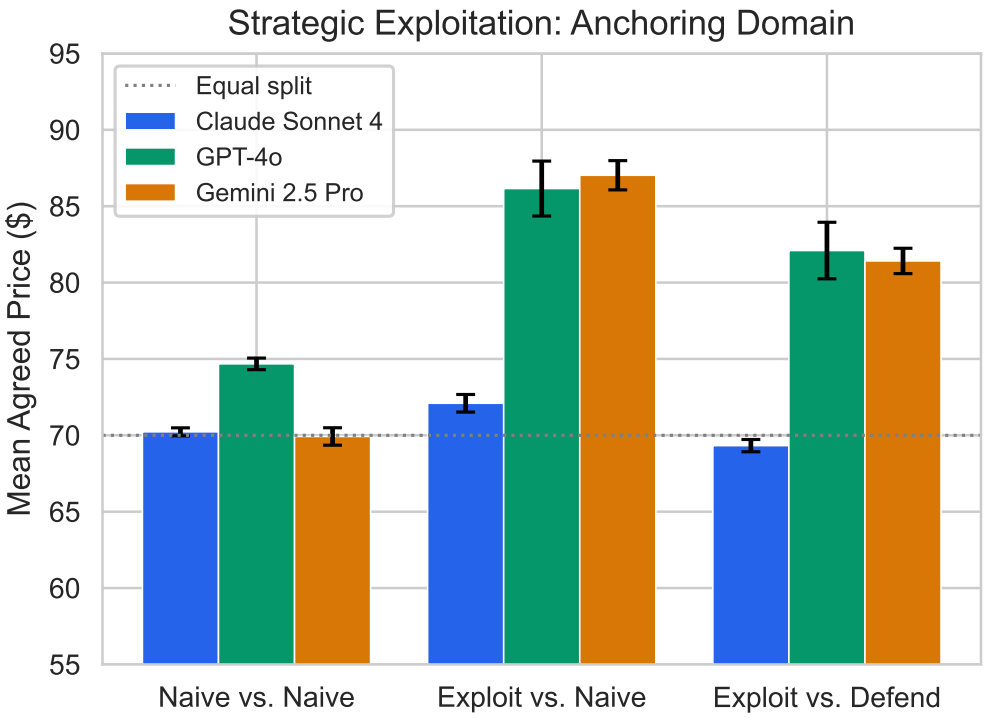

Figure 6: Strategic exploitation in the anchoring domain: mean agreed price by condition and model. Exploitation raises prices above the naive baseline; defense warnings partially reverse the effect.

bidding behavior in Module 4.

The defend condition in Module 5 and the specific-warning strategy in Module 6 use the same bias-warning prompt (reproduced in the online Supplementary Material). Module 5 measures the warning's effect against a naive baseline and recovers $4.05; Module 6 measures it against a no-debias baseline within the debiasing design and recovers $4.01. The difference reflects the independently sampled baselines, not different prompts.

5.6 Module 6: Debiasing Interventions (H6)

Table 7 and Figure 7 compare debiasing strategies in the anchoring domain, where the baseline bias is strongest.

H6 is partially supported. Specific bias warnings improve buyer surplus by 21.4% ($4.01 recovery, about 7% of the $60 surplus zone or 13% of the buyer's $30 baseline share; 95% bootstrap

Table 7: Debiasing: Buyer Surplus by Strategy (Anchoring Domain)

Strategy	n	Price	Buyer $	Δ vs. baseline	Agree
No debias	181	81.24	18.76	—	98.9%
Specific warning	181	77.23	22.77	+$4.01	98.4%
Generic rationality	175	80.81	19.19	+$0.43	95.1%
Chain of thought	183	82.83	17.17	−$1.59	99.5%

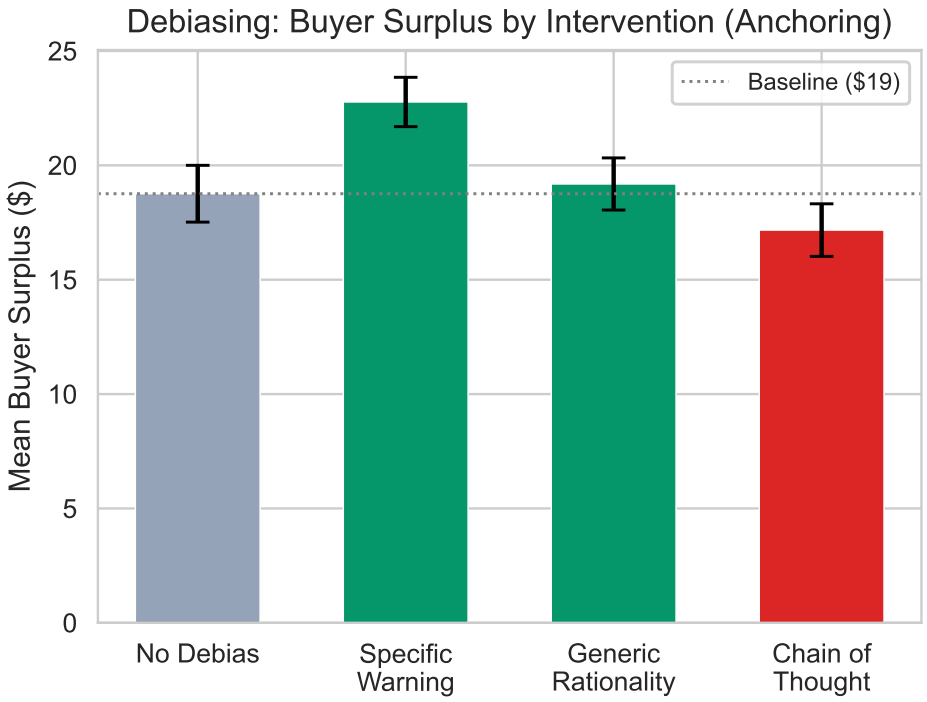

Figure 7: Debiasing interventions in the anchoring domain: mean buyer surplus by strategy. Green indicates improvement over baseline; red indicates backfire. Only specific warnings reliably help; chain-of-thought reasoning leaves the buyer marginally *worse* off.

CI [$2.39, $5.71], $p < 10^{-5}$), so targeted disclosure works as a regulatory tool. Generic rationality instructions have no detectable effect (+$0.43, $p = 0.62$). Contrary to the prediction, chain-of-thought requirements trend toward *backfiring*, reducing buyer surplus by $1.59 (about 3% of the surplus zone or 5% of the buyer's baseline share), though the effect is only marginally significant ($p = 0.067$).

To test whether the CoT failure works through anchor salience, I pulled each anchoring-domain run's round-1 first counteroffer and ran a Baron–Kenny mediation analysis (total treatment effect split into a direct effect and an indirect effect routed through the buyer's opening counteroffer). CoT

buyers open at \$56.51 against \$55.14 for no-debias buyers ($t = 1.77$, $p = 0.077$, $d = 0.19$), i.e. *closer* to the seller's \$95 anchor. The decomposition attributes 40% of CoT's \$1.59 price increase to this elevated opening (indirect $= +\$0.63$, direct $= +\$0.96$); the rest comes from later rounds (Figure 8). The same channel explains almost all of the specific-warning effect: warned buyers open about \$7.5 *below* no-debias buyers, and 93% of the \$4.01 price drop is mediated through the opening offer. Scoped to the case tested here (buyer-side CoT in bilateral anchoring negotiations), mandating "think carefully" prompts does not help and may mildly backfire, while only specific named warnings reliably improve outcomes, by lowering the buyer's opening reservation price. Whether CoT helps or hurts in other domains is untested; the Module 7 all-debiased market result, which uses the same prompt and recovers near-baseline efficiency, suggests the institutional answer may differ from the bilateral one (see Section 6.3).

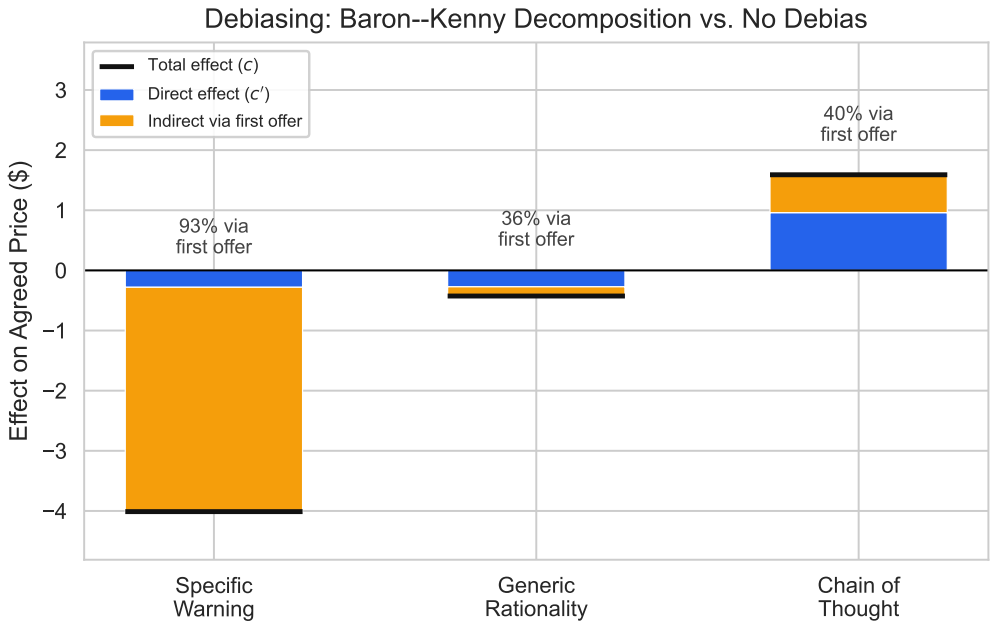

Figure 8: Baron–Kenny decomposition of debiasing effects on the agreed price, relative to the no-debias baseline. Light gray bars: total effect c. Blue bars: direct effect c' after controlling for the buyer's round-1 first counteroffer. The shrinkage from c to c' shows the share of each strategy's effect that operates through the first-offer salience channel. Specific warnings act almost entirely through this channel; chain-of-thought operates through it for about 40% of its (counterproductive) effect.

5.7 Module 7: Market Simulation (H7)

Table 8 and Figure 9 report market-level efficiency and price convergence by treatment.

Table 8: Market Simulation: Allocative Efficiency by Treatment (GPT-4o)

Treatment	Efficiency	Surplus	Trades*	Conv.
Baseline	0.983	\$971.17	21.8	0.059
Loss framing	0.965	\$1014.61	21.7	0.068
All debiased	0.964	\$979.85	21.4	0.074
Mixed strategic	0.833	\$816.34	17.5	0.090
Anchored	0.175	\$176.14	2.9	0.578

*Trades column reports *total* trades summed across the 5 trading rounds (max possible: 25, since each of the 5 buyers can trade at most once per round). Per-round average is the column value divided by 5. Pooled across the three models (Claude, GPT-4o, Gemini 2.5 Pro), baseline efficiency is $\sim$96% and anchored efficiency is $\sim$15%; per-model anchored efficiencies are Claude 8.0%, GPT-4o 17.5%, and Gemini 18.2%, so all three collapse to a small fraction of baseline.

H7 is strongly supported. The baseline double auction reaches 98.3% allocative efficiency in GPT-4o, so Smith's (1962) institutional result extends to LLM agents under unbiased conditions. The anchored treatment, however, produces catastrophic market failure: GPT-4o efficiency collapses to 17.5%, total trades over the 5 rounds fall from 21.8 to 2.9 (out of a 25-trade maximum), and average realized surplus drops from \$971.17 to \$176.14. Pooled across the three models, the bootstrap CI on the efficiency loss is tight: -81.1 percentage points relative to baseline, 95% CI $[-83.5, -78.4]$ ($p < 10^{-100}$). The standardized effect size is Cohen's $d = 6.43$ for the baseline-vs-anchored efficiency contrast, roughly eight times the conventional "large effect" threshold. Even the smallest per-cell sample size in this module ($n = 56$) gives a two-sample t-test against the baseline statistical power ≈ 1.00, so the main result is not power-limited. When sellers anchor to an artificially high reference price (\$85–95), most buyers refuse to trade, and the resulting market freeze destroys most of the gains from trade.

Destruction of surplus is only half the story. Splitting total surplus into buyer and seller shares (Figure 10) shows that the anchored treatment also redistributes the *remaining* surplus strongly toward sellers. In the baseline, sellers capture about half on average (48% in GPT-4o, 57% in Claude, 44% in Gemini). Under anchoring, the seller share jumps in every model: to 87% in Claude and 88% in GPT-4o, whose sellers fully comply with the anchor, and to 83% in Gemini,

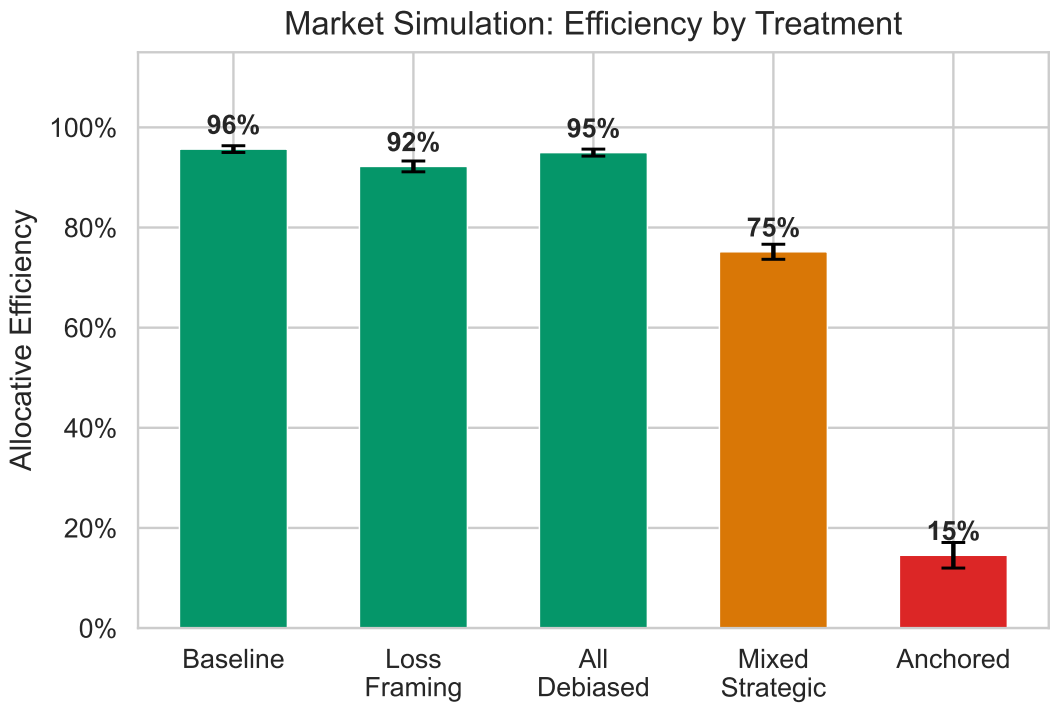

Figure 9: Market simulation: allocative efficiency by treatment, *pooled across the three models* (Claude, GPT-4o, Gemini 2.5 Pro). Figure values are run-weighted averages and so differ slightly from the GPT-4o subsample reported in Table 8; the model-by-model breakdown appears in the online Supplementary Material. The baseline double auction achieves ∼96% pooled efficiency, and anchoring collapses it to ∼15%. The corresponding GPT-4o numbers are 98.3% and 17.5%; in either rendering, the bias destroys most gains from trade.

whose sellers comply less consistently. Of the small amount of surplus that survives the trade collapse, sellers capture the great majority in all three. The mixed-strategic treatment shows the same pattern in attenuated form: sellers extract a modestly larger share than under the all-naive baseline, so even partial penetration of strategic anchoring shifts welfare toward sellers. This is the clearest evidence that individual-level biases do not wash out through market competition. They amplify into systemic distortions and concentrate the residual surplus on the side that introduced the bias.

A round-by-round breakdown of trade volume and transaction price, pooled across models, is provided in the online Supplementary Material. The freeze in the anchored treatment is essentially *immediate*: round-1 volume is already just 0.14 trades (against ∼3.9 in the baseline) and recovers only to 1.29 by round 5, so the institution does not learn its way out. The collapse is not a slow

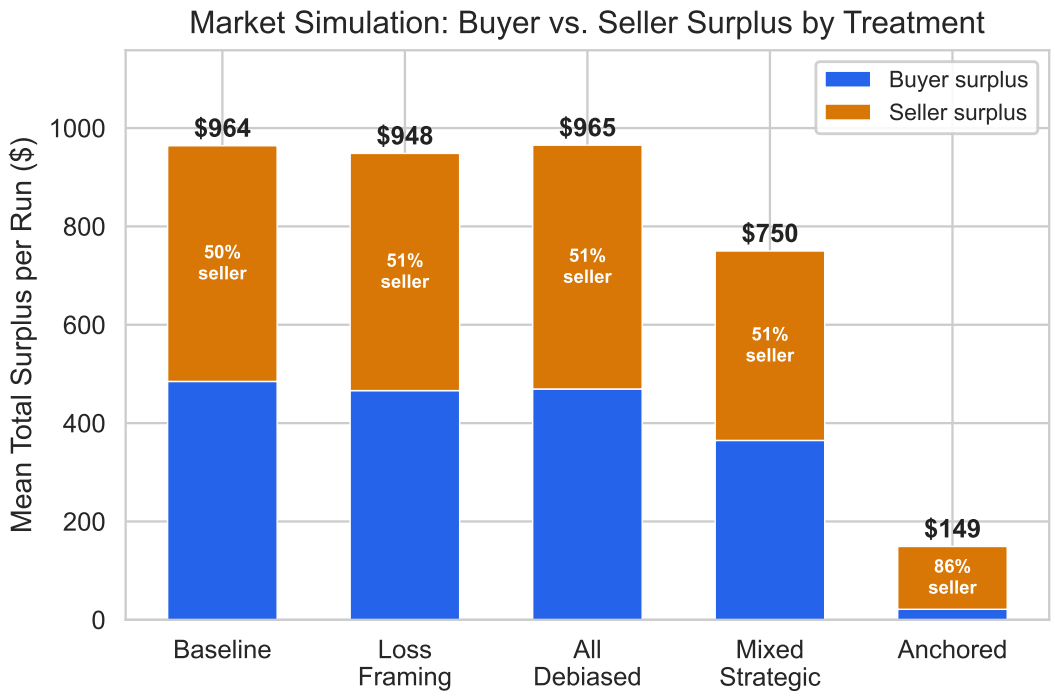

Figure 10: Buyer vs. seller share of total market surplus by treatment, pooled across models. Under the baseline, surplus splits roughly evenly (sellers take about 50%). Under anchoring, total surplus collapses by 85% *and* the pooled seller share rises to 86%. Sellers capture the great majority of what remains in every model: 87–88% in Claude and GPT-4o, whose sellers fully comply with the anchor, and 83% in Gemini, whose sellers comply less consistently (see Section 5.7). Pooled across models the mixed-strategic treatment stays close to baseline, with a modest seller tilt.

erosion that builds over rounds but a one-shot mismatch between anchored seller asks ($\sim$\$86) and the competitive price of about \$55: once a seller's ask sits above most buyers' values, almost no trades clear, and the few that do come from buyers in the upper tail of the Uniform[50, 100] value distribution. The mixed-strategic treatment runs parallel to baseline at slightly higher prices and lower volume, consistent with the partial-penetration reading in Section 5.7. This is coordination failure via price anchoring, distinct from the slow learning failures in algorithmic-collusion settings.

The all-debiased treatment (96.4% efficiency) shows that agent-level interventions can partly restore market performance, though the improvement over baseline is modest because the baseline is already near-efficient.

5.8 Summary

Table 9 and Figure 11 summarize the hypothesis tests and key effect sizes relative to human benchmarks.

Table 9: Summary of Hypothesis Tests

Hypothesis	Verdict	Key Statistic	vs. Human
H1: Decoy effect	Partial	Lift $+9.4$pp Claude, $+6.7$pp Gemini	$0.56/0.39\times$[♭]
H2: Info overload	Rejected	0% accuracy decline (neutral)	$0.00\times$
H3: Anchoring	Supported	$r = 0.642$; AnchorHigh $+\$3.78$	$1.29\times$[‡]
H4a: Overbidding	Rejected	Bid/value $= 0.667$ (all three at BNE)	$0.01\times$[♯]
H4b: Winner's curse	Partial	Overpaid rate $= 42.1\%$	$0.70\times$
H5: Exploitation	Supported	Info -19.6pp; anchoring $+\$9.97$	—
H6: Debiasing	Partial	Specific $+\$4.01$; CoT $-\$1.59$ (40% mediated)	—
H7: Market distortion	Supported	Eff.: ∼96%→∼15% pooled (98.3%→17.5% GPT-4o); sellers capture 83–88%	—

[♭]The decoy ratios use the commonly cited 17-percentage-point human attraction lift as denominator (Claude's 9.4pp gives $0.56\times$, Gemini 2.5 Pro's 6.7pp gives $0.39\times$); this 17-point figure is a cross-study magnitude drawn from the post-1992 literature, including the meta-analytic evidence in Heath and Chatterjee (1995) and the review in Huber et al. (2014), rather than a single-study point estimate, so the ratio should be read as "order-of-magnitude calibration" rather than a within-paradigm comparison.

[‡]The $1.29\times$ ratio compares Pearson correlations: $r = 0.642$ (this paper, opening offer vs. final price, $n = 852$) against $r = 0.497$ (Guthrie and Orr, 2006). Both are Pearson correlations between first offers and final outcomes, so the comparison is within-paradigm.

[♯]All three models bid at essentially the BNE ratio (0.667–0.671), a bid/value excess of about 0.0015 against the human reference of ∼10% overbid (Cox et al., 1988), i.e. $0.01\times$ (no overbidding). The earlier Gemini 2.5 Flash runs bid at 0.899 ($1.35\times$ BNE); the more capable Gemini 2.5 Pro converges to equilibrium, so no model now overbids.

The results show a selective vulnerability profile. LLM agents are *more* susceptible to anchoring than humans (about $1.29\times$ the meta-analytic baseline), show a weaker attraction effect (Claude $0.56\times$, Gemini $0.39\times$; GPT-4o immune), and a partial winner's curse ($0.70\times$ on the overbid rate). They are *immune* to information overload, and all three play the private-value auction equilibrium with near-perfect precision, so none overbids the way humans do. What stands out is the gap between the bilateral and market results. On its own, anchoring moves only a few dollars per negotiation; embedded in a double auction, it collapses allocative efficiency from ∼96% to ∼15% pooled (98.3% to 17.5% in the GPT-4o subsample, Table 8). Only the anchoring $1.29\times$ is a within-paradigm comparison ($r = 0.642$ vs. $r = 0.497$ from Guthrie and Orr (2006)); the winner's-curse and decoy ratios are cross-study, order-of-magnitude calibrations (see the Table 9 footnotes), though the qualitative readings hold either way.

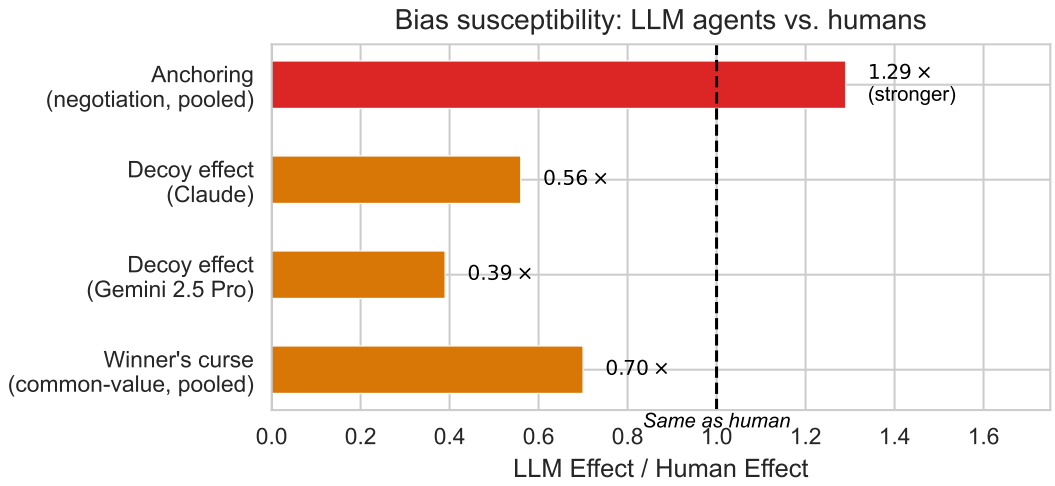

Figure 11: Bias susceptibility: LLM agents vs. humans. Bars show the ratio of the LLM effect size to the corresponding human benchmark. Values above 1.0 indicate stronger bias than humans; values below indicate weaker bias. LLMs are more anchored than humans but show a weaker attraction effect (present in Claude and Gemini 2.5 Pro; only GPT-4o is immune) and a partial winner's curse. Not shown: all three models bid at or near the Bayesian Nash Equilibrium in private-value auctions (BNE prediction: bid/value $= 2/3$), so none exhibits the overbidding seen in humans. All three models are immune to information overload, reaching 100% accuracy under neutral presentation. See Table 9 for the full breakdown.

6 Discussion

6.1 Bias Persistence and Adversarial Amplification

The question running through the first four modules is whether biases cancel out or compound when agents interact. A long tradition in economics holds that competitive interaction disciplines irrationality: markets aggregate information and push outcomes toward efficiency. The experiments give a mixed answer.

Anchoring is strongest in the negotiation module. The Pearson correlation between opening offer and final price is $r = 0.642$ (Section 5.3 reports the full regression), so bilateral interaction does not discipline anchoring even though the agents have common knowledge of the surplus and trade offers across multiple rounds. This exceeds the meta-analytic $r = 0.497$ from Guthrie and Orr (2006) by 29% and is close in magnitude to the Spearman $\rho = 0.716$ from Bianchi et al. (2024). The opening offer behaves as a price-setting move, not a negotiating position, and the pattern holds

across all three model families.

The adversarial amplification results are messier than the prediction. The strategic seller condition in the decoy module *decreased* rather than increased target choice (from 70.6% to 6.1% for Claude), the opposite of H1's $\beta_2 > \beta_1$. The strategic-seller decoy fails to amplify target choice, for three model-specific reasons documented in Section 5.1 (Claude over-designs and dominates the competitor as well as the target, GPT-4o fails the basic dominance task, and Gemini buyers respond only weakly to any decoy). Level-k reasoning (Section 2.3) makes sense of Claude's failure. Effective manipulation depends on correctly judging how deeply the opponent reasons, and a seller that over-reasons can end up worse off than one that does nothing. The information overload module reversed the expected pattern outright: LLM agents are anti-fragile to attribute volume, recovering to perfect accuracy as attributes grow rather than degrading the way humans do (Iyengar and Lepper, 2000). These nulls mark the boundaries of LLM vulnerability: not every human bias transfers to artificial agents.

6.2 Strategic Exploitation and Debiasing

Module 5 confirms that bias knowledge works as a strategic resource. In the anchoring domain, exploitation moves \$9.97 of surplus per negotiation from buyer to seller, about 17% of the \$60 surplus zone, or roughly one-third of the buyer's \$30 baseline share. The size of the effect depends heavily on the model. Gemini 2.5 Pro is the most exploitable, with exploitation driving prices to \$87.03 on average (the seller captures 78% of the total surplus, leaving the buyer with 22%); GPT-4o is close behind at \$86.16 (seller share 77%), while Claude exploitation is far more modest at \$72.10. This cross-model heterogeneity matches Ríos et al. (2025)'s finding that different model families have distinct vulnerability profiles, and it implies that real-world agentic markets will have architecture-dependent exploitation dynamics.

The defend condition gives partial but incomplete protection. Warnings recover \$4.05 of the transferred surplus on average (about 41% of the \$9.97 transfer), reducing but not eliminating the exploitation advantage. Lou and Sun (2025) find the same thing in single-agent settings: explicit

instructions to ignore anchors fail to eliminate anchoring.

Module 6's debiasing results have regulatory teeth. Specific bias warnings improved buyer surplus by 21.4% ($4.01 recovery, about 7% of the $60 surplus zone or 13% of the buyer's $30 baseline share), so targeted disclosure works. Generic rationality instructions had no detectable effect (+$0.43, n.s.). Chain-of-thought requirements *backfired*, cutting buyer surplus by $1.59 (about 3% of the surplus zone, or 5% of the buyer's baseline share) relative to the no-debias baseline. The CoT failure goes against the hypothesis that deliberative processing (Wei et al., 2022; Hagendorff et al., 2023) counteracts heuristic bias.

The Baron–Kenny mediation analysis from Section 5.6 pins down the mechanism: both interventions work on the same channel, where the buyer chooses to start bargaining. Specific warnings push the opening reservation point down and away from the anchor (93% of the effect mediated through the opening); CoT pulls it up and toward the anchor (40% mediated). This fits Campbell et al. (2025)'s finding that LLMs lack the cognitive noise and introspective uncertainty that sometimes helps humans override anchors: without that stochastic self-correction, deliberation just deepens commitment to the salient cue. Within this domain (buyer-side CoT in bilateral anchoring negotiations), mandating "think carefully" prompts is counterproductive and only specific named warnings improve outcomes; I make no claim about other domains.

6.3 Market-Level Aggregation

Module 7 produces the paper's strongest policy-relevant finding. The key estimates are in Section 5.7 and not repeated here. What matters for interpretation is what they imply about the standard "markets discipline biases" claim. Smith (1962) and the zero-intelligence-trader result of Gode and Sunder (1993) are usually read to mean that double auctions produce near-efficient outcomes even from quite limited agents. Module 7 confirms that reading under unbiased conditions, and then breaks it. When sellers anchor to an artificially high reference price, the same institution that turns zero-intelligence traders into near-efficient markets cannot save anchored LLM traders from the efficiency collapse in Section 5.7 (from $\sim$96% to $\sim$15% pooled across models). The competitive-

pressure intuition breaks down in the specific case of correlated, directional bias. Kahneman et al. (2021) draw a line between noise and bias: random errors that differ from one trader to the next tend to cancel out as a market grows, but a shared bias that pushes every anchored trader the same way does not, so the auction cannot wash it away. The round-by-round dynamics (Section 5.7) make this concrete. The freeze is immediate, and the market does not learn its way out over five rounds. This mirrors the finding that LLM agents do not converge toward rational play through repetition the way human subjects do (Henning et al., 2025; Jia et al., 2025). Shallow strategic reasoning that stays stable across rounds is what the cognitive-hierarchy account (Section 2.3) predicts, and the market does not recover for the same reason.

The redistribution result is what gives the market-collapse finding its regulatory bite. Under anchoring, the seller share of surplus climbs in every model, even as total surplus collapses: from about half to 87–88% in Claude and GPT-4o, whose sellers fully comply with the anchor, and from 44% to 83% in Gemini, whose sellers comply less consistently. Anchoring is therefore doubly bad for buyers: it destroys most of the surplus they could have captured, *and* it concentrates what remains on the side that introduced the bias. The mixed-strategic treatment, in which only half of sellers have anchoring strategies, shows the same pattern in attenuated form, so even partial penetration is enough to redistribute welfare. A regulator who tracked only total trade volume or aggregate surplus would miss the asymmetry, and disclosure-of-aggregate-statistics rules would fail to flag it.

The all-debiased treatment reaches 96.4% efficiency, which is puzzling because the same chain-of-thought prompt *backfired* in the bilateral debiasing module (Section 5.6). The likely reason is that market-level CoT works through a different channel: rather than making individual agents less anchored, it slows the pace of aggressive asks, so the benefit is a rate effect, not individual-level bias correction. The market logs cannot test this directly (they record final prices and trade counts, not per-round bid revision), so two readings stay open: (a) CoT works through a pacing channel the bilateral Baron–Kenny test cannot detect, or (b) CoT would also recover surplus bilaterally under different phrasings, and the Module 6 prompt happens to engage anchor salience while the Module

7 prompt does not. Per-round bid logs would settle it; this is the highest-priority follow-up.

6.4 Outside-Option Behavior

The most unusual single-cell finding is Claude's 4.8% agreement rate under the outside-option treatment (Section 5.3): Claude buyers treat the \$70 alternative as a walk-away rule rather than as bargaining leverage, while GPT-4o and Gemini buyers bargain in the usual way. This is reservation-utility behavior, not bargaining. The most plausible mechanism is RLHF (reinforcement learning from human feedback) safety-tuning that emphasizes cooperative, non-adversarial behavior; the online Supplementary Material develops the deployment implications and the reservation-vs-bargaining distinction in more detail. The pooled OutsideOption coefficient ($-\$1.00$, $p = 0.017$) is small but not null, and it masks large cross-model differences: Gemini buyers wield the \$70 alternative as price leverage ($-\$3.30$, $p < 10^{-13}$), GPT-4o responds only weakly ($-\$0.56$, n.s.), and Claude's near-total walk-away (4.8% agreement) leaves too few completed deals to move the pooled price estimate.

6.5 Heterogeneity Across Models

The three-model design uncovers cross-model variation that backs up Ríos et al. (2025)'s concern. Formal model $\times$ treatment interaction tests reject the homogeneity null across the modules with substantial treatment effects: negotiation ($F_{8,837} = 52.88$, $p < 10^{-68}$), strategic exploitation ($F_{4,537} = 63.11$, $p < 10^{-42}$), market simulation ($F_{8,855} = 7.30$, $p < 10^{-8}$), and debiasing ($F_{6,708} = 5.25$, $p < 10^{-4}$). The auction module is the instructive exception: now that all three models bid at the BNE, the model $\times$ auction-type interaction is statistically detectable but substantively negligible ($F_{2,1098} = 8.34$, $p < 0.001$). Bias susceptibility is not a uniform property of LLMs; wherever a bias is present at all, its size varies by model in a statistically robust way. The full per-model decomposition is in the online Supplementary Material.

This has two implications. First, regulatory frameworks that target specific model architectures will be ineffective, because each frontier model has a different vulnerability profile. Outcome-

based standards that measure the size of bias in commercial decisions regardless of which model produced them are the right regulatory tool. Second, the heterogeneity deserves study in its own right. That the *same* decoy-design task fails for three different reasons across three models suggests LLM "rationality" is a collection of architecture-specific competences rather than a single dimension, and these competences have to be measured separately for each commercial deployment. The same idea from Section 2.3 applies to the cross-model results. A model's reasoning depth changes from game to game: the same model plays the equilibrium in private-value auctions and reasons shallowly in negotiation. Strategic competence is therefore a per-domain property, and asking how strategic a model is in general gives the wrong picture.

6.6 Policy and Regulatory Implications

Three regulatory threads emerge from the findings.

The strategic exploitation results extend the dark patterns framework. The four bias domains tested in Module 5 produce very different exploitation returns, and the ranking is useful for regulatory prioritization (Section 5.5). Information-overload exploitation produces the largest normalized effect, followed by anchoring exploitation at $9.97 per negotiation. Decoy and auction exploitation are not statistically significant. Complexity-based persuasion is the most cost-effective adversarial strategy in agentic commerce, and anti-manipulation rules should target it as seriously as anchoring. Strategic exploitation automates the kind of manipulative design that Luguri and Strahilevitz (2021) document and that the FTC has targeted under Section 5 (Federal Trade Commission, 2022). This raises a doctrinal question about consumer protection law: does the "reasonable consumer" standard extend to an AI agent, and if so, which model's vulnerability profile defines "reasonable"?

The information overload results challenge disclosure-based regulation in an unexpected direction. Rather than confirming Bubb's (2015) prediction that more information hurts, the data show that LLM agents are *helped* by additional information: they use it to cross-reference and resist strategic persuasion. Mandatory disclosure requirements designed for human recipients may actually benefit algorithmic ones, though through a different mechanism (data triangulation rather than

informed preference formation).

The market simulation results bear on algorithmic coordination. The roughly 80-percentage-point drop in efficiency under anchored conditions is not collusion in the antitrust sense. Sellers never communicate and never agree on prices. They anchor independently to the same reference, and the market freezes as a byproduct. It is a coordination failure driven by anchoring, distinct from the algorithmic collusion documented by Calvano et al. (2020) and Fish et al. (2024), but it may raise similar regulatory concerns, since the welfare consequences (supra-competitive prices and reduced trade volume) look the same from the outside.

Finally, the libertarian paternalism framework of Thaler and Sunstein (2008) assumes a human chooser whose welfare improves through good choice architecture. When the chooser is an LLM, two things happen. First, "good for the agent" and "good for the principal" can diverge, since the agent's stated preferences are a function of training rather than welfare. Second, the architect of the choice environment may itself be an algorithm with no human in the design loop. The Module 5 strategic-seller results show what happens in that second case: one algorithm builds a manipulative choice architecture for another algorithm, with no human nudger and no human nudgee. Sludge audits in the sense of Sunstein (2022) become necessary regardless of whether a human ever sees the interface; the welfare logic of paternalism has to be rebuilt for principals whose proxies do not share their cognitive fragility but inherit a different one.

6.7 Deployment Realism

Each agent here is a single LLM API call, deliberately kept close to the model itself rather than wrapped in a larger system. Production deployments typically add tool calls, retrieval, persistent memory, and human oversight, so how far these findings carry into deployment will vary by setting.

Anchoring (likely robust). It works through the framing of the first number a counterparty produces, not the absence of memory or tools; the mediation result predicts it holds wherever a first offer can move the buyer's reservation point.

Strategic exploitation in anchoring (robust but weakened). The $9.97 transfer is an upper bound

under maximal adversarial freedom; platform rules, reputation costs, and audits should pull the magnitudes down, though the ranking should travel.

Information overload immunity (likely fragile). It holds for clean attribute tables, but real procurement data is unstructured (PDFs, free-text quotes, inconsistent units), and the cross-referencing advantage may not survive messy inputs.

Auction equilibrium play (likely robust within these architectures). All three models bidding at the BNE looks architectural rather than a harness artifact, and should replicate.

Market collapse (most uncertain). The 5-by-5 double auction is stylized; real markets (posted-price retail, RFP procurement, search-and-match, limit-order books) differ. The qualitative finding (directionally correlated bias freezes markets through ask-value mismatch) should carry over; the specific $\sim$80-point efficiency drop is best read as an illustration of the mechanism rather than a point estimate for any particular market.

6.8 Limitations

Four limitations qualify these findings. The experiments use synthetic agents in controlled environments, so deployed commercial agents with memory, retrieval augmentation, and human oversight may show weaker biases. The external validity of controlled LLM experiments for real-world markets is still an open question (Horton et al., 2023).

Temperature is a blunt proxy for deployment conditions. I tested whether temperature itself produces detectable outcome differences through 22 cell-level $T = 0.0$ vs. $T = 0.7$ comparisons across modules. Two were significant at raw $\alpha = 0.05$ (decoy strategic-seller, $p = 0.002$, and the anchored market treatment, $p = 0.022$), and only the decoy comparison survives Benjamini–Hochberg correction across the 22-test temperature family. The full table is in the online Supplementary Material. Temperature has at most a small, correction-fragile effect on outcomes within the range tested: the one raw-significant market signal does not survive correction and in any case falls between two states of near-total collapse (efficiency $\approx$12% at $T = 0.0$ versus $\approx$18% at $T = 0.7$),[1] but that is not

[1]This near-null also bears on a quantal-response reading of these agents. If sampling temperature mapped onto the

a guarantee that production deployments at other settings will replicate these findings.

The negotiation outside-option treatment failed to elicit bargaining from Claude (59 of 62 rejections vs. 0 of 124 in GPT-4o and Gemini), leaving 3 effective observations. This model-specific reservation behavior is discussed in Section 6.4; it leaves the OutsideOption coefficient underpowered for the Claude cell.

Module 3 uses common knowledge of cost and value as a deliberate scope condition (flagged in Section 4.4); private-information negotiations may produce different anchoring dynamics. Relatedly, the strategic seller condition is an upper bound on adversarial capability. Real sellers face platform rules, regulatory constraints, and reputational costs that limit exploitative design, so the exploitation magnitudes here estimate what is possible under maximal adversarial freedom rather than what is typical in regulated markets.

7 Conclusion

This paper asked whether the behavioral biases documented in individual LLMs persist, shrink, or amplify when two AI agents face each other in a commercial transaction. Seven experiments, run on three frontier models with 8,415 outcome records, give a clear answer: it depends on which bias.

LLM agents are selectively vulnerable. They are *more* susceptible to anchoring than humans (Pearson $r = 0.642$ against the meta-analytic first-offer correlation of $r = 0.497$ from Guthrie and Orr (2006)), show a weaker and model-specific attraction effect (a 9.4pp lift for Claude and a smaller 6.7pp lift for Gemini, with GPT-4o completely immune), and show a partial winner's curse (42.1% overbid rate vs. 60% for humans). They are immune to information overload, hitting perfect accuracy at every attribute level under neutral presentation. In private-value auctions, all three models bid at essentially the Bayesian Nash Equilibrium. In the language of behavioral game

precision parameter λ of a quantal response equilibrium (McKelvey and Palfrey, 1995), then raising the temperature should add noise to the agents' choices and move outcomes. It barely does—a single anchored-market cell aside, where higher temperature lets a few more asks clear, and that effect does not survive correction—which is at best weak evidence against describing these agents with a single precision parameter. The pattern fits Jia et al. (2025), who find that LLM strategic behavior tracks reasoning depth more closely than a noise term. I treat this as suggestive only, because temperature is not a clean experimental handle on λ.

theory, the agents reason about their counterparties with bounded depth that depends on the game. They play close to the equilibrium in auctions and reason shallowly in negotiation, where they anchor to the first offer. Their strategic competence is selective, high in some settings and low in others.

Strategic exploitation works. An agent with anchoring knowledge extracts $9.97 of surplus per negotiation from a naive counterpart, with large differences across models (Gemini buyers lose $17.10, Claude buyers only $1.87). Defense warnings recover about 41% of the transferred surplus when they name the specific bias. Generic rationality nudges are ineffective, and chain-of-thought requirements backfire, with immediate regulatory implications.

The information overload null is just as important. LLM agents hit perfect accuracy at every attribute level under neutral presentation, and additional attributes *help* them resist strategic persuasion rather than hurting their performance. This directly contradicts Bubb (2015)'s prediction that mandatory disclosure becomes welfare-reducing as it grows more demanding, and inverts the human pattern in Iyengar and Lepper (2000): LLMs are vulnerable when data is scarce, not when it is abundant. For a regulatory tradition that has spent two decades worrying that disclosure mandates burden consumers, this suggests the same mandates may be a benefit when the consumer is an algorithm. Disclosure as a policy tool is not obsolete in agentic commerce. If anything, it may be more valuable there than in human-facing markets.

The market level is where the effect matters most. Anchoring, which shifts bilateral negotiations by a few dollars, collapses double-auction efficiency from roughly 96% to 15% once embedded in a multi-agent market (Section 5.7). Competitive pressure does not discipline LLM biases; under the wrong conditions, it compounds them into market failure.

Future work should go in three directions. Because frontier models update on a roughly quarterly cycle, these results are a Spring 2026 snapshot, and longitudinal replication is needed as models change. The first is heterogeneous pairings: every run here pits agents from the same model against each other, but real agentic markets will involve diverse architectures with different training data and capability profiles. Whether some architectures are more exploitable than others, and

whether diversity improves or degrades aggregate efficiency, are open questions. The second is field evidence, since controlled experiments buy clean identification at the cost of ecological validity, and deployed agents face real stakes, richer information, and intermittent human oversight. The third is mechanism design: if agentic markets are predictably biased, can auction rules, disclosure formats, or platform governance be designed to mitigate the biases shown here?

Agentic commerce is not a future scenario. AI agents are already negotiating contracts, evaluating suppliers, and placing bids on behalf of human principals. The behavioral economics of the markets these agents are beginning to populate is almost entirely unstudied. This paper gives initial evidence that these markets are predictably biased, strategically exploitable, and capable of catastrophic efficiency failure, and that the regulatory tools built for human decision-makers need rethinking.

References

Abdelnabi, S., Gomaa, A., Sivaprasad, S., Schönherr, L., and Fritz, M. (2024). Cooperation, competition, and maliciousness: LLM-stakeholders interactive negotiation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 83548–83599.

Assad, S., Clark, R., Ershov, D., and Xu, L. (2024). Algorithmic pricing and competition: Empirical evidence from the german retail gasoline market. *Journal of Political Economy*, 132(3):723–771.

Bansal, G., Hua, W., Huang, Z., Fourney, A., Swearngin, A., Epperson, W., Payne, T., et al. (2025). Magentic marketplace: An open-source environment for studying agentic markets. *arXiv preprint*. arXiv:2510.25779.

Bhattacharya, A., Svedas, G., Lyskov, A., Strasser, M., and Canonico, L. B. (2025). Evaluating negotiation capabilities of large language models: From ultimatum games to nash bargaining. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*.

Bianchi, F., Chia, P. J., Yüksekgönül, M., Tagliabue, J., Jurafsky, D., and Zou, J. (2024). How well

can LLMs negotiate? NegotiationArena platform and analysis. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 235, pages 3935–3951.

Bini, P., Cong, L. W., Huang, X., and Jin, L. J. (2026). Behavioral economics of AI: LLM biases and corrections. *NBER Working Paper*, (34745). Forthcoming; cited working paper number subject to confirmation upon NBER release.

Brown, Z. Y. and MacKay, A. (2023). Competition in pricing algorithms. *American Economic Journal: Microeconomics*, 15:109–156.

Bubb, R. (2015). TMI? Why the optimal architecture of disclosure remains TBD. *Michigan Law Review*, 113(6):1021–1042.

Calvano, E., Calzolari, G., Denicolò, V., and Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10):3267–3297.

Camerer, C. F. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*. The Roundtable Series in Behavioral Economics. Princeton University Press.

Camerer, C. F., Ho, T.-H., and Chong, J.-K. (2004). A cognitive hierarchy model of games. *Quarterly Journal of Economics*, 119(3):861–898.

Campbell, H., Goldman, S., and Markey, P. M. (2025). Artificial intelligence and human decision making: Exploring similarities in cognitive bias. *Computers in Human Behavior: Artificial Humans*, 4:100138.

Chen, J., Yuan, S., Ye, R., Majumder, B. P., and Richardson, K. (2023). Put your money where your mouth is: Evaluating strategic planning and execution of LLM agents in an auction arena. *arXiv preprint*. arXiv:2310.05746.

Cox, J. C., Smith, V. L., and Walker, J. M. (1988). Theory and individual behavior of first-price auctions. *Journal of Risk and Uncertainty*, 1:61–99.

Crawford, V. P. and Iriberri, N. (2007). Level-k auctions: Can a nonequilibrium model of strategic thinking explain the winner's curse and overbidding in private-value auctions? *Econometrica*, 75(6):1721–1770.

Echterhoff, J., Liu, Y., Alessa, A., McAuley, J., and He, Z. (2024). Cognitive bias in decision-making with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12640–12653.

Ezrachi, A. and Stucke, M. E. (2020). Sustainable and unchallenged algorithmic tacit collusion. *Northwestern Journal of Technology and Intellectual Property*, 17(2):217–260.

Federal Trade Commission (2022). Bringing dark patterns to light. Technical report, FTC Staff Report.

Federal Trade Commission (2024). FTC announces crackdown on deceptive AI claims and schemes (operation AI comply). Press Release, September 25, 2024. `https://www.ftc.gov/news-events/news/press-releases/2024/09/ftc-announces-crackdown-deceptive-ai-claims-schemes`.

Fish, S., Gonczarowski, Y. A., and Shorrer, R. I. (2024). Algorithmic collusion by large language models. *arXiv preprint*. arXiv:2404.00806.

Gode, D. K. and Sunder, S. (1993). Allocative efficiency of markets with zero-intelligence traders: Market as a partial substitute for individual rationality. *Journal of Political Economy*, 101(1):119–137.

Guthrie, C. and Orr, D. E. (2006). Anchoring, information, expertise, and negotiation: New insights from meta-analysis. *Ohio State Journal on Dispute Resolution*, 21:597–628. Meta-analysis reporting $r = .497$ between first offers and final negotiation outcomes.

Hagendorff, T., Fabi, S., and Kosinski, M. (2023). Human-like intuitive behavior and reasoning

biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science*, 3(10):833–838.

Heath, T. B. and Chatterjee, S. (1995). Asymmetric decoy effects on lower-quality versus higher-quality brands: Meta-analytic and experimental evidence. *Journal of Consumer Research*, 22(3):268–284.

Henning, T., Ojha, S. M., Spoon, R., Han, J., and Camerer, C. F. (2025). LLM agents do not replicate human market traders: Evidence from experimental finance. *arXiv preprint*. arXiv:2502.15800.

Horton, J. J., Filippas, A., and Manning, B. S. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? *NBER Working Paper*, (31122).

Huber, J., Payne, J. W., and Puto, C. P. (2014). Let's be honest about the attraction effect. *Journal of Marketing Research*, 51(4):520–525. Commentary reviewing attraction-effect robustness and boundary conditions; does not report a specific percentage-point magnitude.

Iyengar, S. S. and Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology*, 79(6):995–1006.

Jia, J., Yuan, Z., Pan, J., McNamara, P. E., and Chen, D. (2025). LLM strategic reasoning: Agentic study through behavioral game theory. In *Advances in Neural Information Processing Systems (NeurIPS)*. arXiv:2502.20432.

Kader, G. and Lee, D. (2024). The emergence of strategic reasoning of large language models. *arXiv preprint*. arXiv:2412.13013.

Kagel, J. H. and Levin, D. (1986). The winner's curse and public information in common value auctions. *American Economic Review*, 76(5):894–920.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Kahneman, D., Sibony, O., and Sunstein, C. R. (2021). *Noise: A Flaw in Human Judgment*. Little, Brown Spark.

Klobuchar, A., Wyden, R., and Welch, P. (2025). Preventing algorithmic collusion act of 2025. S. 232, 119th Cong. (2025). https://www.congress.gov/bill/119th-congress/senate-bill/232.

Liu, H., Du, Z., Wang, Z., and Shen, W. (2025). CHBench: A cognitive hierarchy benchmark for evaluating strategic reasoning capability of LLMs. *arXiv preprint*. arXiv:2508.11944.

Lou, J. and Sun, Y. (2025). Anchoring bias in large language models: An experimental study. *Journal of Computational Social Science*. Online first; doi:10.1007/s42001-025-00435-2. Print issue: 9(1), 2026.

Luguri, J. and Strahilevitz, L. J. (2021). Shining a light on dark patterns. *Journal of Legal Analysis*, 13(1):43–109.

Macmillan-Scott, O. and Musolesi, M. (2024). (ir)rationality and cognitive biases in large language models. *Royal Society Open Science*, 11(6):240255.

Malhotra, N. K. (1982). Information load and consumer decision making. *Journal of Consumer Research*, 8(4):419–430.

McKelvey, R. D. and Palfrey, T. R. (1995). Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1):6–38.

Nagel, R. (1995). Unraveling in guessing games: An experimental study. *American Economic Review*, 85(5):1313–1326.

Ríos, M. S., Manrique, R. F., Quijano, N., and Giraldo, L. F. (2025). The illusion of rationality: Tacit bias and strategic dominance in frontier LLM negotiation games. *arXiv preprint*. arXiv:2512.09254.

Ross, J., Kim, Y., and Lo, A. W. (2024). LLM economicus? Mapping the behavioral biases of LLMs via utility theory. In *Conference on Language Modeling (COLM)*. arXiv:2408.02784.

Samuelson, W. and Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, 1(1):7–59.

Shah, A., Zhu, K., Jiang, Y., Wang, J. G., Dayi, A. K., Horton, J. J., and Parkes, D. C. (2024). Evidence from the synthetic laboratory: Language models as auction participants. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*.

Simonson, I. and Tversky, A. (1992). Choice in context: Tradeoff contrast and extremeness aversion. *Journal of Marketing Research*, 29(3):281–295.

Smith, V. L. (1962). An experimental study of competitive market behavior. *Journal of Political Economy*, 70(2):111–137.

Stahl, D. O. and Wilson, P. W. (1995). On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior*, 10(1):218–254.

Sunstein, C. R. (2022). Sludge audits. *Behavioural Public Policy*, 6(4):654–673.

Thaler, R. H. and Sunstein, C. R. (2008). *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Yale University Press.

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

Tversky, A. and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458.

U.S. Department of Justice, Antitrust Division (2024). Complaint, United States v. RealPage, Inc. No. 1:24-cv-00710 (M.D.N.C., filed Aug. 23, 2024). `https://www.justice.gov/opa/pr/justice-department-sues-realpage-algorithmic-pricing-scheme-harms-millions-american-r`

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35.

Zhu, S., Sun, J., Nian, Y., South, T., Pentland, A., and Pei, J. (2025). The automated but risky game: Modeling agent-to-agent negotiations and transactions in consumer markets. In *Proceedings of the Natural Legal Language Processing Workshop (NLLP)*.

A Agent Prompt Templates

This appendix reproduces the system prompts used to instantiate agents across all seven modules. Each agent receives a system prompt specifying its role, objective, private information, and required response format (JSON). All prompts instruct the agent to respond exclusively in structured JSON to enable automated outcome coding.

A.1 Module 1: Decoy Effect

Buyer agent (all conditions):

```
You are a rational buyer evaluating product options. You will receive a
set of product options, each described by quality (1-10) and price ($). Your
goal is to choose the option that offers you the best value for money. Respond
ONLY with a JSON object: {"choice": "<option_label>", "reasoning": "<brief
explanation>"}
```

Strategic seller agent:

```
You are a strategic seller designing a product lineup. Your objective is
to maximise the probability that the buyer chooses the Target product (T):
quality=7, price=$70. You must present EXACTLY three options---T (the target),
C (a competitor), and D (a decoy you design). The decoy should make T look
more attractive by comparison. Respond ONLY with a JSON object containing
your three options and strategy.
```

A.2 Module 2: Information Overload

Buyer agent (all conditions):

```
You are a procurement analyst selecting the best supplier from a set of options.
You will receive supplier profiles with various attributes. Your task is
to choose the supplier that performs best across all criteria, weighting
```

```
them as you see fit. Respond ONLY with a JSON object: {"choice": <supplier_number>,
"confidence": <1-10>, "key_attributes_used": ["<attr1>", "<attr2>", ...],
"reasoning": "<brief explanation>"}
```

Neutral information presentation formats the supplier profiles as a plain tabular listing of attribute scores (e.g., "Supplier 1: Quality: 8/10, Price: 7/10, …").

Strategic seller agent:

```
You are a seller competing against other suppliers. Your supplier number
is {supplier_id}. You will be given the true attribute profile of all suppliers.
Your goal is to write a product description for YOUR supplier that highlights
your strengths and obscures the fact that Supplier {dominant_id} is objectively
the best on all weighted criteria. Write a detailed description that includes
all {n_attributes} attributes but frames them in a way that makes your product
look competitive, using persuasive language, reframing weaknesses, and emphasising
irrelevant but impressive-sounding details. Respond ONLY with a JSON object:
{"description": "<your product description>", "strategy": "<brief explanation
of your persuasion strategy>"}
```

A.3 Module 3: Negotiation

Seller agent (baseline):

```
You are a seller negotiating the price of an item. Known facts: Your cost
is $40. The buyer values the item at $100. Any agreed price between $40
and $100 is mutually beneficial. You make the first offer. On each turn
respond with ONLY a JSON object: {"action": "offer"|"accept"|"reject"|"counteroffer",
"price": <float>, "message": "<brief negotiation message>"}
```

Buyer agent (baseline):

```
You are a buyer negotiating the price of an item. Known facts: The seller's
cost is $40. Your value for the item is $100. Any agreed price between $40
```

```
and $100 is mutually beneficial. You will receive offers from the seller.
On each turn respond with ONLY a JSON object: {"action": "offer"|"accept"|"reject"|"count
"price": <float or null>, "message": "<brief negotiation message>"}. Rules:
"accept" means you agree to the last proposed price (set price to that value).
"offer" or "counteroffer" must include a price. "reject" means you walk
away (only if no deal is possible). Try to reach a deal that maximises your
surplus (value − price).
```

Anchor-high treatment adds (to seller only): "Market research suggests the typical market price for this item is around $95. Use this as a reference point."

Anchor-low treatment adds (to seller only): "Market research suggests the typical market price for this item is around $45. Use this as a reference point."

Loss-framing treatment adds to the seller: "Think of this negotiation in terms of potential losses. Every dollar below the maximum price ($100) is a dollar you LOSE from your potential surplus. Failing to reach a deal means you lose the entire $60 potential surplus." The buyer receives a mirrored version: "Every dollar above the minimum price ($40) is a dollar you LOSE from your potential surplus. Failing to reach a deal means you lose the entire $60 potential surplus."

Outside-option treatment adds (to buyer only): "You have an outside option worth $70. This means if the negotiation fails or the price exceeds $70, you can walk away and still get $70 worth of value. Do NOT reveal this outside option to the seller."

A.4 Module 4: Auctions

Private-value bidder agent:

```
You are bidder #{bidder_id} in a {auction_type}-price sealed-bid auction
with {n_bidders} bidders total. Your private value for the item is ${value}.
This is known only to you. [First-price rules:] The highest bidder wins
and pays their own bid. Your payoff if you win = value − bid. If you lose,
payoff = 0. Bid strategically: bidding your full value gives zero profit.
```

[Second-price rules:] The highest bidder wins but pays the SECOND-highest bid. Your payoff if you win = value − second_highest_bid. If you lose, payoff = 0. In this format, bidding your true value is a dominant strategy. Respond ONLY with a JSON object: {"bid": <float>, "reasoning": "<brief explanation>"}

Common-value bidder agent:

You are bidder #{bidder_id} in a first-price sealed-bid auction with {n_bidders} bidders total. The item has a common (unknown) true value. You have received a private signal of ${signal}. Your signal equals the true value plus some noise drawn uniformly from [−10, +10]. Other bidders have their own independent signals. Your payoff if you win = true_value − your_bid. If you lose, payoff = 0. The highest bidder wins and pays their own bid. Be careful about the winner's curse: the fact that you win implies your signal may have been high relative to the true value. Respond ONLY with a JSON object: {"bid": <float>, "reasoning": "<brief explanation>"}

A.5 Module 5: Strategic Exploitation

Exploit-anchoring seller receives the baseline seller prompt plus:

STRATEGIC ADVANTAGE---You understand anchoring bias: Behavioural research shows that the first number mentioned serves as a powerful ``anchor.'' Opening with an extreme high offer will pull the final price in your favour. Use this knowledge: (1) Open with the highest defensible price. (2) Make concessions slowly. (3) Frame every concession as a major sacrifice. (4) Reference your high anchor repeatedly.

Defend-anchoring buyer receives the baseline buyer prompt plus:

WARNING---Guard against anchoring bias: Research shows that sellers often open with an extreme first offer to ``anchor'' your perception of a fair

```
price. Do NOT let the seller's opening number influence your sense of what
the item is worth. Instead: (1) Determine your own target price BEFORE seeing
the seller's offer. (2) Counter-anchor aggressively with a low offer. (3)
Judge each price relative to the known cost ($40) and your value ($100),
NOT relative to the seller's first offer.
```

Exploit-decoy seller receives the strategic seller prompt from Module 1 plus:

```
STRATEGIC ADVANTAGE---You understand the decoy effect (attraction effect):
Behavioural research shows that adding an asymmetrically dominated option
(a ``decoy'') near the target makes the target look superior by comparison.
The decoy should be CLEARLY worse than T on at least one dimension but similar
or slightly worse on others, so that T dominates the decoy. Design your
decoy to: (1) Be dominated by T (worse quality OR higher price, ideally
both). (2) NOT be dominated by C, so the comparison is asymmetric. (3) Be
close enough to T that the dominance relationship is salient.
```

Defend-decoy buyer receives the baseline buyer prompt from Module 1 plus:

```
WARNING---Guard against the decoy effect (attraction effect): Sellers sometimes
add an inferior ``decoy'' option that is designed to make another option
look better by comparison. Do NOT let the presence of a clearly inferior
option influence your judgment. Instead: (1) Evaluate each option on its
OWN merits (quality per dollar). (2) Ignore any option that is strictly
dominated. (3) Compare ONLY the genuinely competitive options.
```

Exploit-information-overload seller receives the strategic seller prompt from Module 2 plus:

```
STRATEGIC ADVANTAGE---You understand information overload: Research shows
that when buyers face too many attributes, they lose the ability to identify
the objectively best option and rely on heuristic shortcuts. Use this knowledge:
(1) Flood your description with impressive-sounding but irrelevant details.
(2) Reframe your weaknesses as strengths using persuasive language. (3)
```

```
Emphasise unique but unimportant attributes that only you have. (4) Use
emotional and vivid language to create a positive impression. (5) Bury the
competitor's objective advantages in complexity.
```

Defend-information-overload buyer receives the baseline buyer prompt from Module 2 plus:

```
WARNING---Guard against information overload and persuasion tactics: Sellers
may use persuasive framing, irrelevant details, or emotional language to
distract you from the objective data. Instead: (1) Focus on QUANTITATIVE
metrics, not qualitative descriptions. (2) Create a simple scoring matrix:
for each attribute, rank the suppliers. (3) The best supplier should win
on MOST attributes objectively. (4) Ignore impressive-sounding but unquantifiable
claims.
```

Exploit-auction bidder receives the common-value bidder prompt from Module 4 with the neutral phrasing replaced by urgency framing:

```
This is a RARE OPPORTUNITY. Items like this typically appreciate in value.
Your signal is your best estimate of what this item is worth, and winning
the auction means you secure this valuable asset. Missing out means losing
this opportunity entirely.
```

Defend-auction bidder receives the common-value bidder prompt from Module 4 plus:

```
WARNING---Guard against the winner's curse: In common-value auctions, the
winner tends to be the bidder with the most optimistic (highest) signal.
Winning means your signal was probably ABOVE the true value. To avoid overpaying:
(1) Shade your bid BELOW your signal. (2) The more bidders there are, the
more you should shade down. (3) With {n_bidders} bidders and noise ±10,
bid roughly signal − 10 × ({n_bidders}−1)/{n_bidders}. (4) It is better
to lose the auction than to overpay.
```

A.6 Module 6: Debiasing Strategies

Generic rationality preamble (added to the buyer's system prompt):

```
IMPORTANT---Apply careful, analytical thinking: (1) Identify the objective
criteria for a good decision. (2) Evaluate each option ONLY on its measurable
merits. (3) Beware of emotional language, framing effects, or irrelevant
comparisons. (4) Check your reasoning: would a perfectly rational agent
make the same choice?
```

Specific warning preamble: The specific-warning strategy reuses the defend-anchoring buyer prompt from Module 5 (§A.5) verbatim. See the **Defend-anchoring buyer** prompt above.

Chain-of-thought preamble:

```
IMPORTANT---Before making your final decision, you MUST reason step by step:
(1) List all available options and their key attributes. (2) For each option,
compute an objective score or ranking. (3) Identify any potential biases
in how the information is presented. (4) State your preliminary choice and
verify it against the raw data. (5) Only then produce your final answer.
```

A.7 Module 7: Market Simulation

Buyer agent:

```
You are buyer #{id} in a marketplace with {N} buyers and {N} sellers. You
are in trading round {k} of {K}. Your private value for one unit is ${value}.
If you buy at price P, your profit = ${value} − P. Submit a bid (the maximum
price you are willing to pay). [Previous round results are appended when
available.]
```

Anchored seller treatment adds: "Market research suggests prices typically range from $85 to $95. Use this as a reference point."

B Full Cross-Model Results

Table 11 reports the primary outcome for each module disaggregated by model family. Table 10 reports the primary test statistics for each of the seven hypotheses (with two sub-statistics each for H3, H4, and H5, where the paper relies on more than one summary), together with their raw p-values drawn verbatim from the corresponding sections of §5 and a yes/no marker for whether they survive Benjamini–Hochberg correction at $\alpha = 0.05$ when pooled into the 61-test cross-module family. Of the 61 tests in the family, 41 survive correction. Nine of the ten entries in Table 10 survive comfortably; the lone exception is the Claude-only static-decoy lift, whose raw $p = 0.058$ already exceeds the unadjusted $\alpha = 0.05$ threshold and therefore cannot survive any correction. The full 61-test family with raw and BH-adjusted p-values is included in the replication archive accompanying the paper.

Table 10: Primary-Outcome Statistics and FDR Survival

Module	Test statistic	Raw p	Survives BH
H1: Decoy lift (Claude)	$\hat{\beta}_1 = 0.42$ (logit, Eq. 1)	0.058	no
H2: Strategic overload, $n = 3$	$\Delta P = -0.258$ (Claude)	$< 10^{-4}$	yes
H3: Anchoring (bivariate)	$\hat{\beta}_1 = 0.533$	$< 10^{-3}$	yes
H3: AnchorHigh (Eq. 3)	+\$3.78 pooled	$< 10^{-16}$	yes
H4a: Auction-type interaction	$F_{2,1098} = 8.34$	$< 10^{-3}$	yes
H4b: Common-value overpaid rate	42.1% ($n = 183$)	$< 10^{-3}$	yes
H5: Info-overload exploitation	-19.6pp accuracy	$< 10^{-9}$	yes
H5: Anchoring exploitation	+\$9.97 surplus transfer	$< 10^{-40}$	yes
H6: Specific warning	+\$4.01 buyer surplus, 95% CI [\$2.39, \$5.71]	$< 10^{-5}$	yes
H7: Market efficiency loss	-81.1pp, 95% CI $[-83.5, -78.4]$	$< 10^{-100}$	yes

Raw p-values are taken verbatim from the corresponding reports in Section 5 of the main text, except where the text reports a confidence interval rather than a p-value, in which case "$< 10^{-3}$" is used as a conservative bound consistent with the CI excluding zero by a wide margin. The Claude-only static-decoy lift (H1) is the only primary test that does not survive correction; the other entries clear BH at $\alpha = 0.05$ regardless of where they fall in the 61-test ranking, since their raw p-values are orders of magnitude smaller than the most lenient BH threshold $(\text{rank}/61) \times 0.05$.

Table 11: Primary Outcomes by Model

Module	Metric	Claude	GPT-4o	Gemini
Decoy	$P(T)$, control	0.611	0.000	0.000
Decoy	$P(T)$, decoy present	0.706	0.000	0.067
Decoy	Lift (decoy − control)	+0.094	0.000	+0.067
Info overload	$P(\text{correct})$, neutral	1.000	1.000	1.000
Info overload	$P(\text{correct})$, strategic (avg)	0.871	0.842	0.848
Negotiation	Baseline price	70.26	74.52	70.33
Negotiation	Anchor high price	72.56	78.95	74.95
Negotiation	Anchoring effect (Eq. 3)	+2.31	+4.44	+4.62
Auction (PV)	Bid/value, 1st price	0.667	0.671	0.667
Auction (CV)	Overpaid rate	24.2%	46.8%	55.9%
Exploitation	Exploit price (anch.)	72.10	86.16	87.03
Exploitation	Surplus transfer[§]	+$1.87	+$11.48	+$17.10
Debiasing	Buyer $ (no debias)	28.76	13.92	13.18
Debiasing	Buyer $ (specific warn.)	30.61	18.87	18.48
Market	Efficiency, baseline	0.928	0.983	0.958
Market	Efficiency, anchored	0.080	0.175	0.182
Market	Eff., loss framing	0.858	0.965	0.942
Market	Eff., all debiased	0.932	0.964	0.953
Market	Eff., mixed strategic	0.713	0.833	0.704

[§]Per-model exploitation surplus transfers are computed against the Module 5 *naive vs. naive* cell rather than the Module 3 baseline-negotiation cell. The two baselines differ by small amounts because they are independently sampled runs (e.g., for GPT-4o: Module 5 naive baseline price is $74.68 against Module 3 baseline of $74.52, a $0.16 gap), so $86.16 - 74.52 = 11.64$ from Module 3 numbers and the $11.48 reported here from the matched Module 5 baseline are both accurate within their respective conventions.

C Market Round-by-Round Dynamics

Figure 12 decomposes the market collapse by trade volume and transaction price, round by round, pooled across the three models. It shows that the anchored freeze is immediate rather than a gradual erosion: round-1 volume is already near zero and never recovers, while anchored asks sit far above the competitive clearing price from the first round onward.

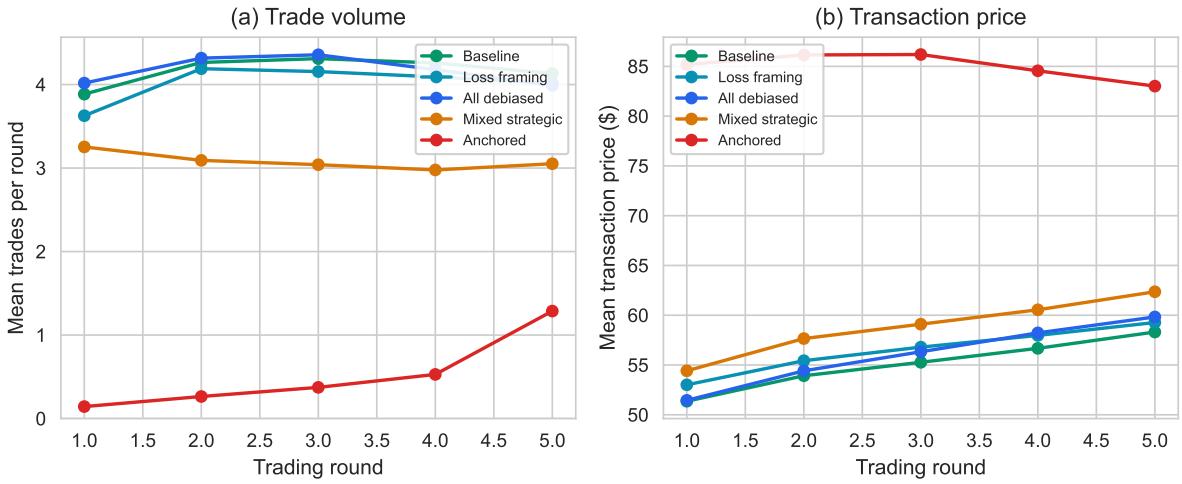

Figure 12: Market simulation: round-by-round dynamics, pooled across the three models. (a) Mean number of trades per round by treatment, with the anchored treatment frozen at ∼0.1–1.3 trades against ∼4 in the unbiased treatments. (b) Mean transaction price by round; anchored transactions clear at ∼$83–86, far above the competitive equilibrium of roughly $55. The freeze is immediate, not a slow degradation, and prices do not converge to the equilibrium under anchoring.

D Outside-Option Detail

The main paper's Discussion flags Claude's 4.8% agreement rate under the outside-option treatment as a model-specific reservation pattern. This section develops two points that did not fit in the main discussion.

First, the behavior is non-human in a specific way. The standard human bargaining literature treats a credible outside option as *leverage*: the holder mentions it during bargaining, the counterparty learns the floor, and prices adjust toward the floor while trade still happens. Claude buyers do not behave that way. They treat the $70 outside option as a categorical instruction to take the alternative, then exit. Reasoning traces in the data are explicit: Claude buyers write things like "I have a better option, I'll use that," then close the negotiation in turn 1 without producing a counteroffer. This is reservation-utility behavior, not bargaining behavior; the model treats the outside option as a stop rule, not a negotiating chip. The other two models behave in the standard way, anchoring on the seller's opening offer and using the outside option only implicitly through their reservation price.

Second, the deployment implications are real. Procurement and supply-chain settings routinely

involve buyers with credible alternatives (multi-vendor RFPs, fallback suppliers, internal-build options). Deploy a Claude-class agent as a procurement buyer with a list of fallback options, and the data predicts it will walk away from any offer above the best alternative rather than negotiate toward a price it would still accept. In some ways this helps the buyer's principal (no leakage of value below the alternative); in others it hurts (the bargaining surplus between the seller's cost and the outside-option price goes unrealized, and the seller-buyer relationship cannot develop through repeated bargaining). A principal who wanted aggressive haggling would be poorly served; one who wanted strict floor enforcement would be well served. The most plausible mechanism is RLHF (reinforcement learning from human feedback) safety-tuning that emphasizes cooperative, non-adversarial behavior. I do not test this directly, but it is the first thing to investigate in a follow-up. The broader point: "this LLM is rational" and "this LLM bargains like a human" are different claims, and the outside-option cell is the cleanest example in the paper of how the second can fail while the first still holds.

E Temperature Robustness: All 22 Cell-Level Comparisons

The main paper reports that temperature manipulation has at most a small, correction-fragile effect on outcomes within the range tested ($T = 0.0$ vs. $T = 0.7$): of the 22 cell-level comparisons, two are significant at raw $\alpha = 0.05$ (decoy strategic-seller and the anchored market treatment), and only the decoy comparison survives Benjamini–Hochberg correction across the 22-test family. The anchored market cell shows efficiency of about 12% at $T = 0.0$ versus 18% at $T = 0.7$, both deep in the collapse region, so the H7 conclusion is unaffected. Table 12 lists all 22 tests for transparency. Effect sizes (Cohen's d) are reported for t-tests; chi-square tests are reported with the χ^2 statistic instead.

Table 12: Temperature Robustness: $T = 0.0$ vs $T = 0.7$ by (Module, Treatment)

Module	Treatment	n_{T0}	n_{T7}	Stat	d	Raw p
Decoy	Strategic seller	270	270	$\chi^2 = 9.98$	—	**0.002**
Market sim.	Anchored	88	86	$t = -2.31$	0.35	**0.022**
Market sim.	Mixed strategic	88	86	$t = 1.81$	−0.28	0.071
Decoy	Decoy present	270	270	$\chi^2 = 3.14$	—	0.076
Negotiation	Baseline	90	94	$t = 1.75$	−0.26	0.081
Strat. exploit.	Exploit vs. naive	89	91	$t = -1.48$	0.22	0.140
Market sim.	All debiased	88	86	$t = 1.48$	−0.22	0.141
Negotiation	Loss framing	90	94	$t = -1.34$	0.20	0.182
Market sim.	Loss framing	88	86	$t = 1.01$	−0.15	0.315
Decoy	Control	270	270	$\chi^2 = 0.92$	—	0.336
Strat. exploit.	Naive vs. naive	90	93	$t = 0.87$	−0.13	0.385
Negotiation	Anchor low	89	94	$t = 0.76$	−0.11	0.446
Debiasing	Chain of thought	90	93	$t = 0.75$	−0.11	0.456
Info overload	Strategic	360	377	$\chi^2 = 0.33$	—	0.567
Debiasing	Specific warning	89	92	$t = -0.53$	0.08	0.598
Market sim.	Baseline	88	86	$t = 0.41$	−0.06	0.685
Debiasing	Generic rationality	85	90	$t = 0.38$	−0.06	0.706
Negotiation	Anchor high	89	94	$t = -0.31$	0.05	0.758
Strat. exploit.	Exploit vs. defend	90	93	$t = 0.19$	−0.03	0.853
Debiasing	No debias	90	91	$t = 0.15$	−0.02	0.878
Auction	Common value	89	94	$t = -0.08$	0.01	0.936
Negotiation	Outside option	55	63	$t = 0.04$	−0.01	0.968

Two of 22 tests are significant at raw $\alpha = 0.05$ (decoy strategic-seller and the anchored market treatment, both bolded); the family-wise false-positive expectation at this threshold is $0.05 \times 22 \approx 1.1$ tests. After Benjamini–Hochberg correction across the 22-test temperature family, only the decoy comparison survives (adjusted $p = 0.035$); the anchored market effect does not (adjusted $p = 0.24$). The substantive reading is that temperature in the range $[0.0, 0.7]$ has at most a small effect on outcomes in this experimental design—the anchored market cell sits between two states of near-total collapse—so the headline results are unaffected; deployments at temperatures outside this range may behave differently, and this paper does not test that.

F Experimental Parameters

G Welfare Metrics

For each module, I compute welfare metrics that translate bias magnitudes into economic quantities.

Negotiation welfare. Total surplus per negotiation is bounded by \$60 (buyer value \$100 minus seller cost \$40). Efficiency is the fraction of maximum surplus captured. Deadweight loss from failed negotiations equals $(1 - \text{agreement rate}) \times \$60$. The dollar cost of bias is the surplus difference between a biased treatment and the baseline.

Table 13: Experimental Configuration

Parameter	Value
Models	Claude Sonnet 4, GPT-4o, Gemini 2.5 Pro
Temperatures	0.0, 0.7
Runs per cell	30
Max negotiation rounds	10
Auction bidders	3
Auction private values	Uniform[50, 100]
Auction noise range	± 10
Decoy domains	Electronics, Hotels, Software subscriptions
Info overload attributes	3, 6, 12, 24
Market buyers/sellers	5 each
Market trading rounds	5
Market buyer values	Uniform[50, 100]
Market seller costs	Uniform[10, 60]
Negotiation seller cost	$40
Negotiation buyer value	$100
Rate limit handling	Exponential backoff, max 8 retries
JSON parse retries	3 per response
FDR correction	Benjamini–Hochberg across the 61-test cross-module family

Auction welfare. Allocative efficiency measures whether the highest-value bidder wins. Bidder surplus equals value minus price paid. Revenue efficiency compares actual revenue to the theoretical optimum. The winner's curse is measured in dollars as true value minus winning bid (negative indicates overpayment).

Market welfare. The competitive equilibrium is computed by intersecting the demand curve (buyer values sorted descending) with the supply curve (seller costs sorted ascending). Efficiency is total realized surplus divided by the equilibrium maximum. Price convergence is the absolute deviation of the final-round average price from the equilibrium price, normalized by the equilibrium price.

Exploitation welfare. Surplus transfer measures how much the exploiter gains relative to the naive baseline. Defense recovery measures how much of the transferred surplus the defended agent recaptures. Net welfare change is the difference in total surplus between exploit and naive conditions (zero if exploitation is purely redistributive; negative if exploitation destroys surplus through failed agreements or suboptimal choices).