Small Bias, Large Failure: Anchoring Collapses a Market of Language-Model Agents

Anton Hantel*

Massachusetts Institute of Technology

Preprint — under review

July 10, 2026

Abstract

Markets are supposed to discipline individual irrationality. I test whether they do when the traders are language-model agents carrying one bias, anchoring, into a five-seller market. Alone the bias is mild, moving an agreed price a few dollars on a sixty-dollar deal. The same one-sentence anchor, given only to sellers, drops the market's efficiency from 96% to 15%. Every seller lifts its asking price to a common level above what most buyers will pay, so trade freezes; most of the damage is deals that never happen. Aggregation magnifies the bias rather than disciplining it: agents cloned from one model are too alike for any un-anchored seller to undercut the rest. The model that resists anchoring best one-on-one collapses the market hardest, so pairwise testing would clear the riskiest model. The collapse survives higher randomness in the agents and reverses under a one-line debiasing instruction.

Keywords: agentic commerce, large language models, anchoring, market efficiency, focal points, multi-agent systems, double auction

JEL Classification: D91, D44, D47, L86, K21

*Contact: hantel@mit.edu. ORCID: 0009-0003-9761-6046. This is a preprint of a manuscript under peer review; if a published version exists, please cite that.

1 Introduction

Software agents built on large language models are beginning to act in markets. They compare products, request quotes, negotiate terms, and place orders, and the firms deploying them expect agents soon to transact with other agents directly (Zhu et al., 2025; Bansal et al., 2025). Large models already negotiate well enough to be set against one another in autonomous negotiation tournaments (Bianchi et al., 2024; Vaccaro et al., 2026). They also inherit the behavioral biases that a large literature has documented when a single model answers in isolation, anchoring among them: their numerical judgments drift toward whatever salient value is in view, sometimes resembling human heuristics and sometimes departing from them in ways no person would (Horton et al., 2023; Lou and Sun, 2025; Chen et al., 2025, 2023; Cheung et al., 2025; Hagendorff et al., 2023; Macmillan-Scott and Musolesi, 2024). Either way, these agents do not price like textbook rational traders, and almost all of this evidence comes from one agent answering one prompt. Recent work places language models in multi-agent settings, from repeated games to simulated agent societies to experimental markets, and finds their collective behavior is its own phenomenon rather than a replica of human interaction (Akata et al., 2025; Park et al., 2023; Henning et al., 2025); how a documented individual bias scales into market-level welfare remains open. In a companion study of mine surveying behavioral biases across agent-to-agent commerce, anchoring stood out as the one whose harm scales most sharply from the individual to the market (Hantel, 2026). The present paper is the deep dive into that result: I trace how a few-dollar nudge becomes a market-wide collapse and pin down why.

A long tradition in economics holds that markets discipline individual irrationality. Experimental double auctions reach competitive prices even when traders are few and far from rational (Smith, 1962), efficiency survives when traders are replaced by budget-constrained random bidders (Gode and Sunder, 1993), and field evidence finds that market experience erodes the anomalies people display in surveys (List, 2003). On this view, aggregation is a safeguard: competition lets un-biased traders profit at the expense of biased ones, and the bias washes out of the price.

I find the opposite. A bias too small to matter in a single negotiation produces near-total market

failure once every seller shares it. In a controlled bilateral test with one buyer and one seller (a dyad), handing a seller a high anchor moves the final price by \$2.31 for Claude Sonnet 4, \$4.44 for GPT-4o, and \$4.62 for Gemini 2.5 Pro, against a \$60 bargaining surplus. The same sentence, delivered to all five sellers in a double auction, drops allocative efficiency from 0.957 to 0.145. More than four-fifths of the gains from trade disappear from a market because of a nudge that barely registered in the dyad.

The mechanism works through the sellers' asks. Price in this market is a transfer from buyer to seller, so total surplus depends only on which units change hands. The anchor raises seller asks onto a common \$85–95 band and decouples the ask from the seller's private cost; the correlation between ask and cost falls from $+0.50$ to essentially zero. Because the lowest ask on the board fixes the best price any buyer can hope to meet, the marginal price-setting seller is the cheapest one, and the anchor lifts even that seller above what most buyers are worth. Buyers, who never see the anchor, will not pay more than their value, so they walk. Almost all of the lost surplus comes from trades that never happen, and almost none from mismatched trades.

The failure rides on homogeneity. A single salient number works as a focal point in the sense of Schelling (Schelling, 1960): every seller can condition on it without communicating, so it holds a common, above-competitive floor that no one defects from. Human markets shrug off a shared anchor because traders are heterogeneous, and a single un-anchored seller can undercut the rest. A population of agents cloned from one model has no such internal diversity, and that is what removes the disciplining trader. When I expose only half the sellers to the anchor, efficiency holds at 0.75, because one seller still asking near cost is enough to clear the market near-efficiently.

This connects to, but differs from, the literature on algorithmic collusion. Reinforcement-learning pricing agents can learn supracompetitive prices through repeated interaction (Calvano et al., 2020), an effect now visible in field data (Assad et al., 2024) and a growing concern in competition law (Ezrachi and Stucke, 2020; U.S. Department of Justice, Antitrust Division, 2024). Language-model agents can reach collusive prices too, including in double auctions like the one I study (Fish et al., 2024; Agrawal et al., 2025). The failure here needs none of the machinery of

collusion. There is no learning, no repeated-game punishment, and no intent. A single line of injected context coordinates the sellers on a focal price in the very first round. The result resembles collusion's outcome reached by accident (Dou et al., 2025), triggered by one sentence.

Three findings matter for anyone certifying agents for commerce. First, ranking the models by how well they resist the anchor one-on-one is a poor guide to their market risk: the model most resistant in the dyad, Claude, collapses the market hardest, because its discipline shows up as full compliance with the anchor and a readiness to walk away from any deal above its reservation value. Pairwise safety testing would have flagged Claude as the safest of the three. Second, the drop is the anchor's doing: only one sentence changes between the baseline and anchored markets and buyers' prompts are identical across them, so once I account for which model is trading and other run-to-run differences, the anchor still cuts efficiency by 0.81 on the zero-to-one scale ($t = -62$, a drop so large it would essentially never arise by chance if the anchor did nothing); raising the sampling temperature does nothing, and a placebo framing manipulation moves efficiency by only 0.03. Third, the bias is reversible: telling sellers in one line to ignore anchors and price to cost restores efficiency to 0.95, which helps an operator who already knows the anchor is there but not one who does not.

The contribution is to show that for at least one well-documented bias, the move from a single agent to a market is the move from harmless to catastrophic, and to identify why: homogeneity turns a private bias into a common focal point and strips the market of the heterogeneity that ordinarily disciplines it. Section 2 describes the two environments, the treatments, and the efficiency measure; Section 3 establishes the amplification gap and its causal footing; Section 4 traces the mechanism; Section 5 turns to why the model that resists the anchor best one-on-one collapses the market worst; and Section 6 reads the result against the human evidence and draws out the implications.

2 Methods

I study the same anchoring manipulation in two environments: a bilateral negotiation and a multi-seller double auction. Three frontier models play every role: Claude Sonnet 4, GPT-4o, and Gemini 2.5 Pro.[1] Each runs at sampling temperatures of 0.0 and 0.7. Agents receive a structured prompt describing their role and private information, and return a numerical bid, ask, or offer; they never see each other's private values or the experimenter's labels.

Bilateral negotiation. One buyer and one seller bargain over a single good. The good is worth $100 to the buyer and costs the seller $40, leaving a $60 surplus to divide. In the baseline the two exchange offers until they agree or reach an impasse. In the high-anchor condition the seller is told that comparable units have recently sold for about $95; the buyer is told nothing. I measure the anchoring effect as the change in the agreed price between the anchored and baseline conditions. A separate condition gives the buyer a $70 outside option, which I use to measure each model's willingness to walk away. The bilateral data comprise 918 negotiations.

Double auction. Five buyers and five sellers trade a homogeneous good over five rounds. Each buyer's value is drawn uniformly from $50 to $100 and each seller's cost from $10 to $60, independently across agents and runs. In each round every agent submits a sealed bid or ask after seeing the previous round's transaction prices; bids and asks are sorted and a trade clears at the midpoint whenever a bid meets or exceeds an ask. The design follows the experimental tradition that established markets as efficient with minimally rational traders (Smith, 1962; Gode and Sunder, 1993).

Because price is a pure transfer between buyer and seller, total surplus depends only on which units trade, not on the terms. I therefore measure *allocative efficiency* as realized surplus divided by the maximum surplus attainable from the round's values and costs. Efficiency of 1.0 means every welfare-improving trade happened; 0.0 means none did. This separation lets me decompose any shortfall into an *extensive* margin (surplus lost because too few units trade) and an *intensive* margin

[1] Exact API identifiers: `claude-sonnet-4`, `gpt-4o`, and `gemini-2.5-pro`. Each run draws its buyer values and seller costs from a recorded integer seed; bootstrap confidence intervals use a fixed seed with 2,000 resamples.

(surplus lost because the units that trade are the wrong ones).

Treatments. The anchor is one sentence given only to sellers: comparable units have recently traded in the $85–95 range. I compare five conditions. *Baseline* has no anchor. *Anchored* gives all five sellers the anchor. *Mixed* gives it to only some sellers, varying the share exposed. *Loss framing* gives sellers an emotionally loaded but numerically empty message, serving as a placebo for "any added sentence." *Debiased* gives all sellers the anchor and then instructs them to ignore such reference prices and price to their own cost. Buyers are identical across all conditions, which isolates the seller-side effect cleanly. The market data comprise 870 runs spread across the three models, two temperatures, and five treatments.

3 A bilateral nudge becomes a market collapse

In the dyad the anchor has a small effect. Across the three models a high anchor moves the agreed price by $2.31, $4.44, and $4.62 (Figure 1a), between four and eight percent of the $60 surplus. By the correlation standard usual in the negotiation literature the agents are ordinary anchorers: their final prices track the opening offer with $r = 0.64$ across all bilateral runs, against a human meta-analytic benchmark of $r = 0.497$ (Guthrie and Orr, 2006); controlled tests of anchoring in language-model price negotiations report the same human-like susceptibility (Takenami et al., 2025). These agents anchor about as much as people do, if anything slightly more, so the result below is not driven by agents being unusually suggestible.

The same sentence empties the market. Pooled allocative efficiency, the share of attainable surplus the market realizes, falls from 0.957 to 0.145 (Figure 1b). By model it drops from 0.928 to 0.080 for Claude, from 0.983 to 0.175 for GPT-4o, and from 0.958 to 0.182 for Gemini. A manipulation worth a few dollars in the dyad removes more than four-fifths of the welfare from the market.

The only thing that changes between baseline and anchored is one sentence shown to sellers; buyers' prompts are byte-for-byte identical. Table 1 puts the comparison on a regression foot-

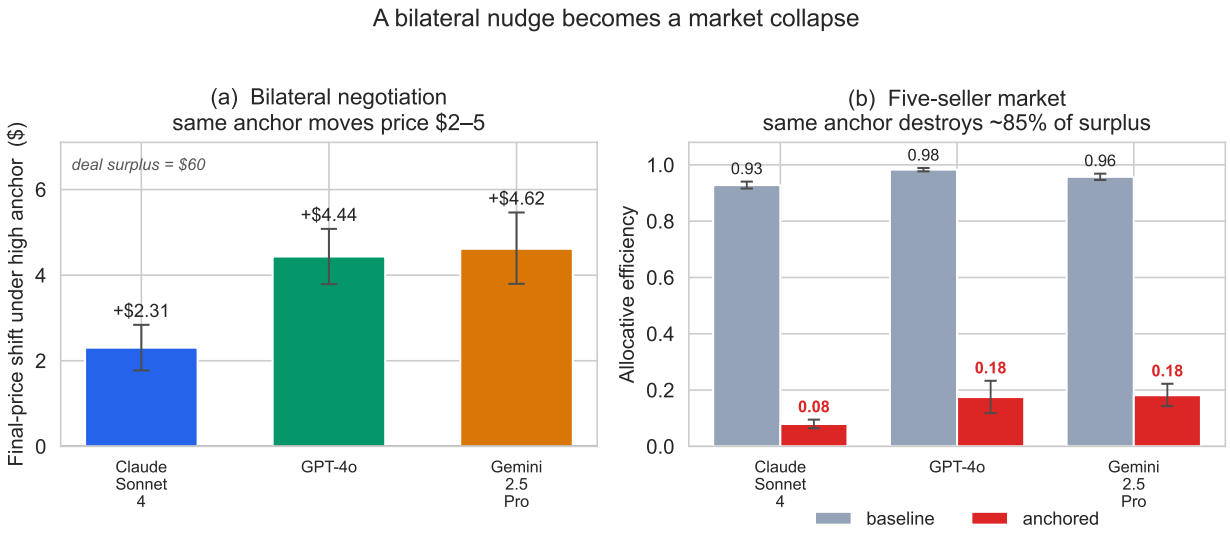

Figure 1: **A bilateral nudge becomes a market collapse.** (a) In bilateral bargaining a high anchor given to the seller moves the agreed price by $2–5 on a $60 surplus. (b) The same one-sentence anchor, given to all five sellers in a double auction, cuts allocative efficiency from about 0.96 to about 0.15. Bars are means over runs; all quantities are computed from the raw transaction data.

ing. Panel A reports the full treatment ladder with bootstrap confidence intervals. Panel B regresses run-level efficiency on treatment dummies, model fixed effects, and temperature, with heteroskedasticity-robust standard errors. The anchored coefficient is -0.811 ($t = -62$): holding the model and temperature fixed, the anchor destroys eight-tenths of attainable efficiency. The ladder also rules out two alternative readings. The loss-framing placebo moves efficiency by only 0.035, so the collapse tracks the anchor's number; extra text on its own does little. And the debiasing instruction returns efficiency to 0.950, statistically indistinguishable from baseline, so the capacity to trade efficiently is intact and merely suppressed by the anchor.

4 How the collapse happens

Because surplus depends only on which units trade, I can read the collapse off who is willing to trade with whom. Figure 2 traces the chain.

The anchor moves asks, and almost nothing else (Figure 2a). Mean seller asks climb from $51 to $86, into the bottom of the $85–95 band. Buyers, who never see the anchor, barely move: mean

Table 1: Anchoring collapses market efficiency, and the collapse is specific and reversible

Panel A. Allocative efficiency by treatment (pooled across models)			
Treatment	Mean	95% CI	N
Baseline	0.957	[0.950, 0.963]	174
Loss framing (placebo)	0.922	[0.910, 0.933]	174
Mixed (half of sellers)	0.752	[0.736, 0.767]	174
Anchored	0.145	[0.120, 0.171]	174
Debiased (anchor + warning)	0.950	[0.943, 0.957]	174
Panel B. OLS: efficiency on treatment, model FE, temperature			
	Coef.	Robust SE	
Anchored	-0.811	(0.013)	***
Mixed (half of sellers)	-0.205	(0.008)	***
Loss framing	-0.035	(0.005)	***
Debiased	-0.007	(0.005)	
GPT-4o	0.082	(0.008)	***
Gemini 2.5 Pro	0.046	(0.007)	***
Temperature (0.7)	0.001	(0.009)	
Constant	0.913	(0.006)	***
$N = 870$		$R^2 = 0.917$	

Notes. Panel A confidence intervals are from 2,000 run-level bootstrap resamples. Panel B omits baseline and Claude as reference categories; heteroskedasticity-robust (HC3) standard errors in parentheses. Clustering by design cell (model $\times$ treatment $\times$ temperature, 30 cells) leaves the anchored, mixed, and loss-framing coefficients significant at $p < 0.001$ (anchored $t = -31$). *** $p < 0.001$. Both panels use all 870 market runs.

bids rise only from \$59 to \$65, and across every condition fewer than two-tenths of one percent of bids exceed the buyer's own value. With the average buyer worth \$75 and the cheapest seller now asking near \$85, the typical buyer cannot meet the ask without overpaying, so it declines.

The loss is therefore concentrated on trades that never happen. The decomposition confirms it (Figure 2b). The extensive margin accounts for 90% of Claude's potential surplus, 78% of GPT-4o's, and 77% of Gemini's; the intensive margin never reaches 5%. The anchor freezes the allocation rather than distorting it.

A cleaner way to see the freeze is to ask what the ask is now responding to. At baseline sellers price off their own cost: across all asks the correlation between a seller's ask and its cost is $+0.50$. Under the anchor that link is gone, with a correlation of -0.00 (Figure 2c). The anchor replaces cost-based pricing with a common reference number, so the ask stops carrying information about

the seller's actual cost and instead reports the focal price. This is the focal point at work: every seller answers the same external cue rather than its own situation, so the asks move together.

The freeze arrives at the open. First-round efficiency under the anchor is 0.04, against 0.95 at baseline, and Claude's first anchored round produces no efficient trade at all. Over the five rounds, as agents watch how little is clearing, asks soften and efficiency climbs to 0.34, with most of the recovery in the final round (Figure 2d). That is real adaptation, and it still leaves the market far below baseline. Five rounds of price feedback do not undo one sentence of anchor. Doubling the horizon barely helps: replaying a balanced half-sample for five further rounds, anchored efficiency over rounds six to ten averages 0.36, far below the 0.89 that baseline markets reach over the same rounds, and the gap holds for every model (Figure 6).[2]

A single price governs the market. The decisive variable is the lowest ask on the board. Round efficiency holds near 0.9 while the cheapest ask sits below \$60, slips to about 0.6 through the \$70s, and falls to about 0.13 once the cheapest ask reaches the \$80s (Figure 3a). The lowest ask fixes the best price any buyer can match, so how much trade is possible depends on the single cheapest seller rather than on the average one. The anchor lifts that floor. The round's cheapest ask has a median of \$50 at baseline and \$85 under the anchor, and the share of rounds whose cheapest ask clears \$85 rises from 0% to 72% (Figure 3b). The anchor pushes the marginal, price-setting seller out of the range where buyers live.

A minimal account. The pattern follows from the supply curve the asks trace out. At baseline each seller prices near its own cost, so the asks span the same low range as the costs, almost every welfare-improving match is available, and efficiency sits near one. The anchor sends each complying seller up to the focal floor of about \$85. Because every seller's cost lies below that floor,

[2]The continuation re-prompts each finished market's agents with the rounds one to five price history and asks them to play rounds six to ten, with the prompt now stating a ten-round horizon and values and costs unchanged. It runs on a balanced half-sample ($n = 15$ per cell for Claude and GPT-4o, $n = 14$ for Gemini), so it is lower-powered than the five-round result. One caveat is specific to Claude: its later rounds were played by the successor `sonnet-4-5` after the `sonnet-4` snapshot was retired, so its continuation mixes two model versions and is best read as suggestive; the pooled result and the other two models are unaffected.

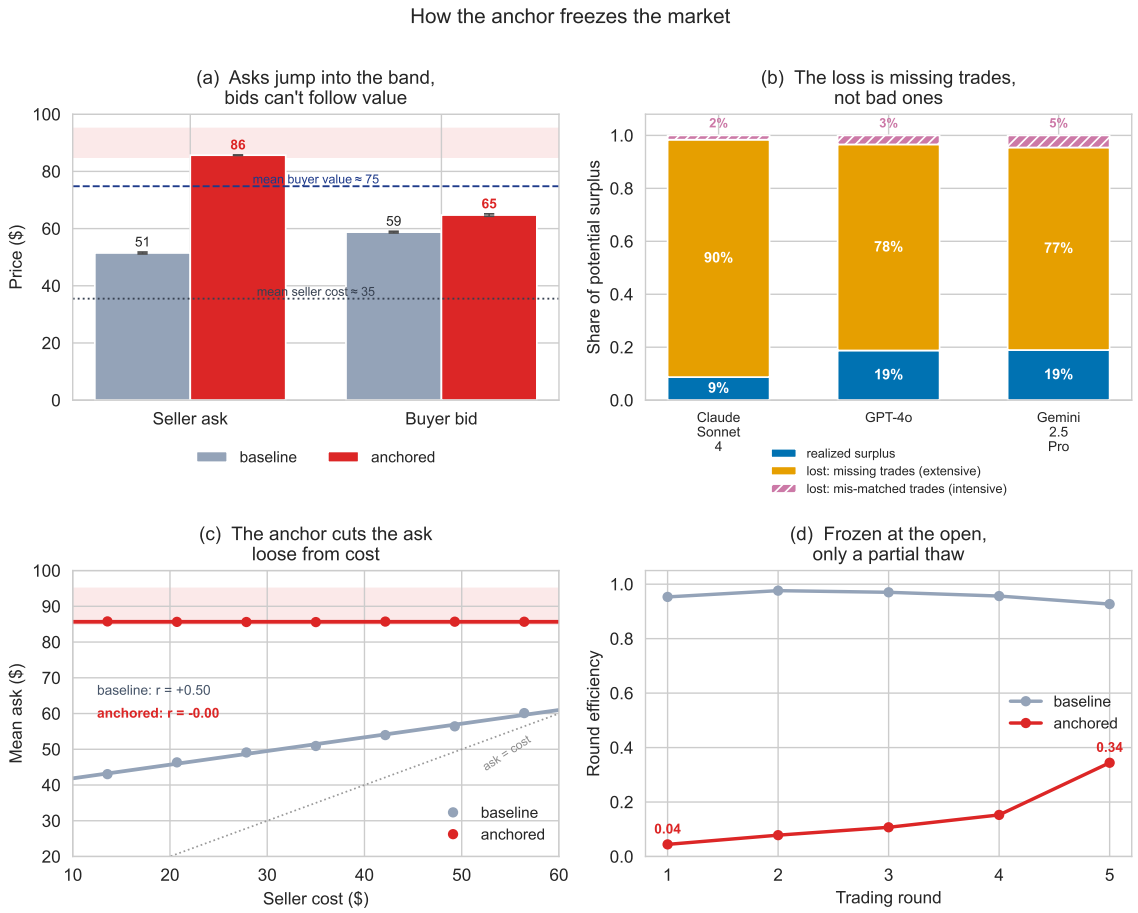

Figure 2: **How the anchor freezes the market.** (a) Asks jump into the \$85–95 band while bids stay near buyer values, so buyers are priced out. (b) Almost all of the lost surplus is on the extensive margin (missing trades), not the intensive margin (mismatched trades). (c) At baseline asks track sellers' costs ($r = +0.50$); under the anchor that link is severed ($r = -0.00$) as every seller reports the same focal price. (d) The market is frozen in the opening round and only partly thaws by round five.

each anchored seller posts the floor price in place of its own cost, and the entire supply curve lifts above the range where buyer values live (uniform on \$50–100).[3] Once even the lowest ask clears \$85, only the highest-value buyers can still trade, and the residual efficiency of 0.15 is essentially the surplus those buyers carry (Figure 3a). The account is a consistency check rather than an independent prediction; it reads the focal floor f off the data and depends on little else. Because it assumes buyers bid their values, it is an upper bound, so observed efficiency sits at or below the

[3]Writing f for the focal floor and c_i for seller i's cost, a complying ask is $a_i \approx \max(c_i, f)$; since $c_i < f$ for every seller, $a_i \approx f$.

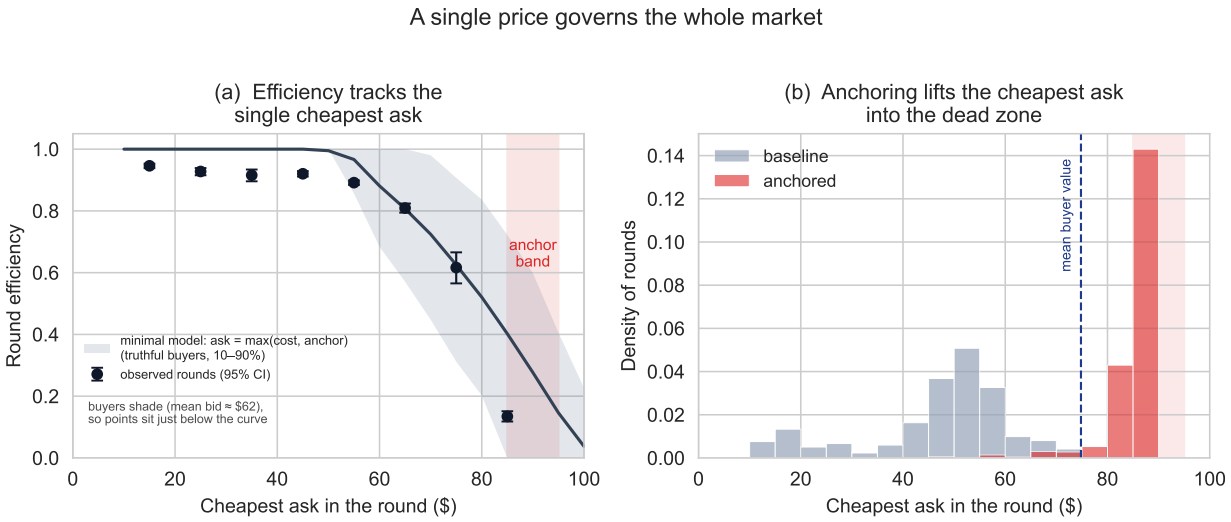

Figure 3: **A single price governs the whole market.** (a) Round efficiency against the round's cheapest ask (points with bootstrap 95% intervals), with the minimal supply-curve model overlaid (curve and band): efficiency stays near 0.9 while the cheapest ask is low, eases through the $70s, and collapses once it enters the $80s. Because the model assumes buyers bid their values it is an upper bound, so the observed points sit at or below it and fall furthest in the danger zone. (b) The anchor shifts the distribution of the cheapest ask from a median of $50 to a median of $85, lifting it past the mean buyer value into the range where no trade clears.

curve. Varying the anchor's height to trace how the curve moves would turn the check into a test, which I leave to future work.

Homogeneity is the cause. The same logic explains why partial exposure is so much gentler than full exposure. When only some sellers are anchored, compliance with the high band stays low, between 1% and 15%, and efficiency holds at 0.75 (Table 1, Panel A). As long as one seller keeps asking near cost, that seller sets a low floor and the round clears near-efficiently. What matters is that the sellers all move together. How high their asks rise on average is not what does the damage. The collapse is a property of a homogeneous population whose members err in the same direction (Kim et al., 2025), and the experimental knob that controls it is the share of agents that share the bias.

5 Who collapses, and why the safest dyad fails hardest

The three models fail by different amounts, and the order is surprising. Claude collapses hardest (0.080), then GPT-4o (0.175), then Gemini (0.182). The first-order driver is compliance with the anchor, which I measure as the share of sellers posting an ask at or above $85, the bottom of the suggested band. Claude and GPT-4o sellers comply essentially 100% of the time, while Gemini sellers comply only 43% of the time. Gemini's partial disobedience is the leak that keeps its markets near a fifth efficient rather than a twelfth.

Compliance cannot separate Claude from GPT-4o, since both comply fully, so something else makes Claude the worst. That something is its readiness to walk away. In the bilateral test with a $70 outside option, Claude abandons the deal 95% of the time, GPT-4o never walks, and Gemini walks 12% of the time. The same rigidity surfaces in the market: when asks are high, Claude's buyers hold out hardest, so the marginal trades that GPT-4o still salvages simply do not occur for Claude.

Claude is the outlier that drives this (Figure 4b): the model that resists the anchor best one-on-one collapses the market worst. It moves only $2.31 in bilateral bargaining, the most disciplined of the three, and then surrenders 92% of market surplus. GPT-4o and Gemini land at statistically indistinguishable market efficiency (0.175 and 0.182, with overlapping intervals), so the defensible reading is the sharper one: the most dyad-resistant model is the most fragile in the market, rather than a strict reordering of all three. The trait that serves Claude well in a dyad, an unwillingness to be talked off a price it judges fair, is what empties the market when every counterparty is holding an inflated number. Robustness in the small becomes fragility in the large, and a pairwise safety test would have certified exactly the wrong model.

Temperature changes none of this (Figure 4a). Raising the sampling temperature from 0.0 to 0.7 lifts pooled anchored efficiency only from 0.12 to 0.18, and the entire move is GPT-4o (0.08 to 0.27); Claude stays frozen (0.07 to 0.09) and Gemini edges down (0.20 to 0.16). Compliance holds near 100% for both Claude and GPT-4o at either temperature, so GPT-4o's partial thaw comes from noisier asks occasionally dipping low enough to clear, not from the model shaking off the

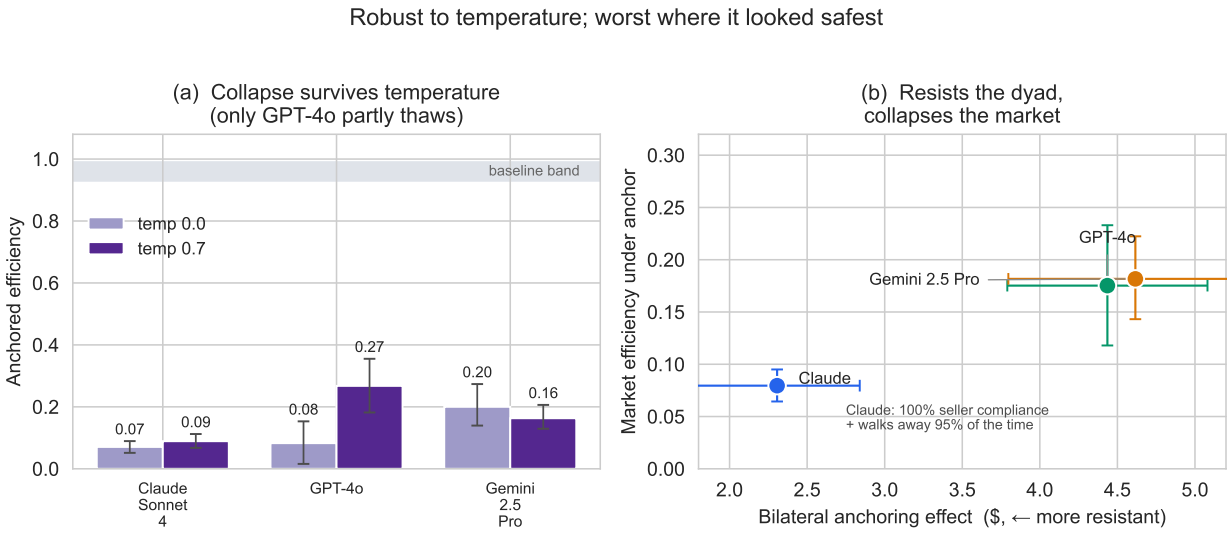

Figure 4: **Robust to temperature; worst where it looked safest.** (a) Anchored efficiency by model and temperature; the collapse survives a higher sampling temperature, and only GPT-4o partly thaws. (b) Across arenas, the model most resistant to the anchor in bilateral bargaining (Claude, leftmost) is the one whose market collapses furthest. Claude is the clear outlier; GPT-4o and Gemini reach statistically indistinguishable market efficiency, so the panel shows the most dyad-resistant model failing hardest rather than a strict reordering of the three.

anchor. The collapse is a property of the behaviour the model has learned, not of running it at zero randomness, which is consistent with the null temperature coefficient in Table 1.

Where the surplus goes. The anchor destroys trade, and it also redistributes what survives. Among the trades that still clear under the anchor, the seller's share of realized surplus rises from about 50% at baseline to about 86% pooled. The shift is large for every model and largest for the two that fully comply (88% for GPT-4o and 87% for Claude, against 83% for the less compliant Gemini). The surviving Claude and GPT-4o trades clear at a mean price near $86, against $57 at baseline, leaving the buyer about $7 of surplus per trade and the seller about $53. Beyond the trades it destroys, then, the anchor hands most of the gains from the trades that remain to the side that was anchored.

6 Discussion

Against the human evidence. Anchoring in people is well documented and partial: numerical judgments assimilate toward whatever value is in view, even when the value is arbitrary and the judge is motivated to be accurate (Tversky and Kahneman, 1974; Furnham and Boo, 2011). A high first offer pulls a bilateral settlement toward it, and by that standard my agents are normal anchorers (Guthrie and Orr, 2006). And markets are usually thought to discipline anchoring, because competition lets un-anchored traders undercut anchored ones and experienced participants show muted anomalies (Smith, 1962; Gode and Sunder, 1993; List, 2003). The expectation is that aggregation washes the bias out.

I find aggregation magnifying it. A bias too small to matter in one negotiation yields near-total market failure once every seller shares it. The mechanism is coordination of the kind Schelling (Schelling, 1960) described: a salient number that all sellers can condition on works as a focal price, holding a common above-competitive floor with no communication or agreement. Human markets escape this because traders are heterogeneous and the un-anchored arbitrage the anchored. A population of agents cloned from one model is homogeneous by construction, and that homogeneity removes the very arbitrageur that disciplines the human market; collective biases can arise in populations of language models even where the individuals look unbiased (Ashery et al., 2025). The outcome resembles the result of tacit collusion reached without intent, but it requires none of the learning or repeated-game punishment that the algorithmic-collusion literature relies on (Calvano et al., 2020; Fish et al., 2024): a single line of context coordinates the sellers in the first round.

The focal price, in the sellers' words. The transcripts show the focal point being assembled in language. A privately anchored seller recasts the planted number as an objective market fact in the message it sends the buyer, and the plainest sign is a seller volunteering the exact anchor it was handed only in private. Claude does so in 40% of its anchored messages and GPT-4o in 30%, while neither model's sellers ever cite the figure without the anchor. GPT-4o goes furthest and sources it to an outside authority: it opens 71% of its anchored negotiations with a market-research

citation, 34 of the 62 with the word-for-word sentence "Based on market research, the typical price for this item is $95," then drops the citation after the first round, and the phrasing never appears in a baseline. That citation marks the focal point rather than driving the bargain: across GPT-4o's anchored negotiations, whether the seller opens with the authority line predicts neither the final price nor the buyer's surplus (each difference under $0.31, with a 95% interval spanning zero), so what fixes the outcome is the ask itself rather than its narration. A broader and softer measure points the same way: the share of seller messages that dress the ask as fair or standard rather than as the seller's own demand climbs under the anchor from 29% to 98% for Claude, 12% to 65% for GPT-4o, and 41% to 74% for Gemini, with non-overlapping bootstrap intervals throughout; this measure shares wording with the planted line, so it corroborates the pattern rather than clinching it. An independent language-model judge that relabeled every one of these messages confirmed each case the lexicons flagged, leaving the reported rates as conservative lower bounds. Because each seller independently turns the same private cue into the same public claim, a market of such sellers inherits a shared above-cost reference with no communication between them, which is what the focal-point account requires (Appendix A collects representative messages). Gemini is the mirror image: it volunteers the number in just 6% of its anchored messages, no more than at baseline, and it is the model whose sellers most often defy the anchor and whose market survives best; the same partial disobedience that keeps its market alive (Section 5) surfaces in what its sellers say.[4]

Which argument pays. The choice of framing also tracks how the negotiation ends. I tag every seller message for the three moves above (calling the ask fair or standard, stating the planted number, or citing an outside authority), aggregate the tags to the negotiation, and regress its outcomes on whether the seller used each move, holding model and treatment fixed (Figure 5). Calling the ask

[4]Each seller message is scored with keyword lexicons: an authority lexicon (market research, studies, experts, appraisal, valuation), a fairness lexicon (fair, reasonable, worth, going rate, comparable, market price/value/rate/standard), and the literal anchor value. Rates are message-level except the opening-citation rate, which is over negotiations; intervals bootstrap over negotiations with 2,000 resamples. An independent judge (`gpt-4o-mini`, temperature 0) relabeled all 1,426 baseline and high-anchor seller messages for the same constructs. The lexicons never flag a message the judge rejects (precision 1.00 on all three channels) and reproduce the literal anchor-citation count exactly (Cohen's $\kappa = 1.0$); the fairness and authority lexicons are deliberately narrow, recovering 62% and 9% of the cases the judge reads more liberally, so those rates are conservative floors. The never-anchored buyer uses no authority language at any anchor level, so the seller's usage is anchor-induced rather than a feature of the dialogue.

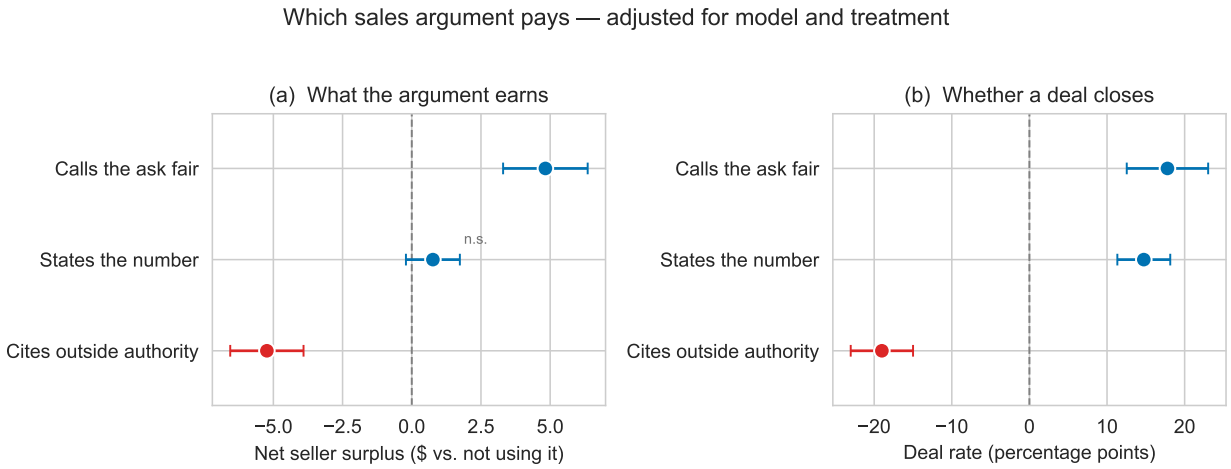

Figure 5: **Which sales argument pays, adjusted for model and treatment.** Each seller message is tagged for three moves (calling the ask fair or standard, stating the planted number, or citing an outside authority), and the negotiation's outcomes are regressed on indicators for whether the seller used each move, with model and treatment fixed effects (918 negotiations; points are coefficients, bars 95% confidence intervals). (a) Effect on the seller's net surplus in dollars, relative to not using the move. (b) Effect on the probability a deal closes, in percentage points. Calling the ask fair raises both; citing authority lowers both; stating the number helps close a deal but nets the seller nothing (n.s.). Because sellers choose their own words, these are adjusted associations rather than assigned treatments.

fair is the move that pays: it lifts the chance of a deal by 18 percentage points and the seller's net surplus by \$4.83 (both $p < 0.001$). Stating the number outright also helps close a deal (a 15-point gain), but at a price nearer the buyer, so the seller's share of the surplus among the deals that close falls by five points and the net gain is nil (\$0.76, $p = 0.12$). Citing an authority does the reverse: deals close 19 points less often and net surplus drops by \$5.23. Among the deals that do close, only stating the number shifts the buyer–seller split, so citing an authority marks the focal point without moving the price within a deal, matching its null effect on GPT-4o's prices noted above; what the choice of argument changes is whether a deal happens at all. Because each seller picks its own words, these are adjusted associations, not assigned treatments: a seller who senses an impasse may reach for an authority line, so part of that negative coefficient reflects when the move gets used. With model and treatment held fixed, the ranking is stable and fits the mechanism: the fairness framing that holds the focal floor while still sounding cooperative is also what keeps trade alive.

Is it anchoring, or something simpler? A skeptic can read the sellers three ways. The behavior might be anchoring in the textbook sense, an automatic pull toward a salient number. It might be plain instruction-following, with the model treating "market research suggests $85–95" as a directive to obey. Or it might be rational updating: a seller with no other read on demand could reasonably infer from the same line that buyers will pay near $90. My design does not fully separate these readings, but two of its arms narrow the field. The loss-framing placebo shows that the damage tracks the number rather than the mere presence of an extra instruction, which weakens the pure instruction-following story. The debiasing arm shows the behavior is undone by one more sentence, which fits anchoring and instruction-following alike. The dynamics within the anchored runs are only suggestive on this point. A seller updating on the low volume it sees might be expected to walk its ask down toward cost; instead anchored sellers barely move, easing their mean ask only from $86.7 in the first round to $83.8 in the fifth even as most of the market fails to clear, while baseline sellers move the other way, raising asks from $41 to $57 as they learn the price. This does not discriminate cleanly, though: with almost nothing clearing under the anchor, a rational seller sees little volume to update on, so a flat ask is about equally consistent with sparse feedback and with bias. Cleanly separating it still needs a control I did not run: hand sellers the anchor together with the true distribution of buyer values. A seller reasoning correctly, told that buyer values are spread evenly between $50 and $100, should discount the $85–95 figure and price near cost, whereas a seller who still drifts to the focal number is not updating rationally. I flag this as the decisive test and leave it for future work. For the market result the distinction is secondary. Whether the sellers are biased or are reasoning correctly from a one-sided signal that happens to be false, the welfare collapse is the same, and so is the lesson for anyone who can place that signal (Takenami et al., 2025; Bini et al., 2026).

Implications. Four points follow. *Pairwise evaluation understates market risk.* The model that looks safest one-on-one is the most dangerous in a market, and nothing in the bilateral result forecasts it. Certifying agents for commerce will need market-level tests, not dyadic ones alone. *The*

failure rides on homogeneity. So long as many trading agents run on the same underlying model, they share priors and answer the same cue in the same direction, manufacturing the correlation that breaks the market; a diversity of models, or at least of priors, is a safety property of an agent market. *The anchor is an attack surface.* It need not be true and there is no collusion to detect: a single salient line of "market research" placed where sellers will read it, on a product page or in an injected prompt (Debenedetti et al., 2024), can raise the floor and either freeze trade or route surplus to whoever planted it. That is market manipulation with a one-sentence payload, and it sits squarely within current regulatory attention to algorithmic pricing harms (U.S. Department of Justice, Antitrust Division, 2024; Klobuchar et al., 2025). *The bias is reversible, but not on its own.* The debiasing instruction restores efficiency to 0.95 (Bini et al., 2026), though it helps only when the operator already knows the anchor is there. The structural remedies, a heterogeneous population of agents and clearing rules that do not broadcast a common reference price, work without foreknowledge of the attack, because they target the homogeneity that does the real work.

Scope. I study one bias, one market institution, three models at a fixed market size, and a single anchor level placed above the mean buyer value. Two tests are the natural complements. Varying the anchor's height to trace a dose-response curve would separate the focal-point mechanism from this particular placement, and populating the market with different models would probe whether heterogeneity restores discipline. The focal-point reading implies that the collapse should ease as the agent population grows more diverse, and the mixed-exposure result is a first step in that direction. The double auction is a clean but particular institution, and clearing rules that hide rivals' prices may behave differently. The result is an existence proof with a clear mechanism: a benign-looking, human-sized bias can move from harmless in a dyad to catastrophic in a market, and homogeneity is the reason.

7 Conclusion

A single sentence of plausible market context, given to sellers who each anchor about as much as a person would, collapses a five-seller agent market from near-perfect to near-broken efficiency. The damage is overwhelmingly trades that never happen, set in motion by a focal price that lifts the cheapest seller out of buyers' reach. The collapse holds across sampling temperatures, tracks the anchor's number rather than the mere presence of an extra sentence, and reverses under a one-line warning. It is worst for the model that looked safest in isolation. As agents come to transact with one another at scale, the unit of evaluation has to be the market, and the property to watch is how alike the agents are.

The lesson generalizes past this one bias. Agentic commerce is converging on a handful of frontier models, so the agents on each side of a market increasingly share the same training, the same priors, and the same blind spots. A quirk that would be a single operator's bad day when traders differ becomes a market-wide event when the traders are near-copies of one another: they misread the same cue in the same direction at the same moment, and no contrarian is left to take the other side. The failure here needs no collusion and no bad actor, only sameness, and sameness is the default when most agents are built on the same few models. That makes the diversity of the models in a market something to protect rather than an inefficiency to optimize away, and it puts a premium on testing agents in the markets they will populate rather than one conversation at a time. The cheapest insurance against a correlated failure is a population that does not all fail in the same way.

Acknowledgements

I am grateful to Professor Cass R. Sunstein at Harvard Law School, in whose seminar HLS 2589: Behavioral Economics, Law, and Public Policy (Spring 2026) this line of work began, for his guidance and encouragement in carrying it further. I also thank Jono Hart for many valuable discussions and his feedback on the ideas.

Declarations

Funding. The author received no funding for this work.

Competing interests. The author declares no competing interests.

Ethics approval. Not applicable. This study involved no human or animal subjects; all experimental agents were large language models.

Data and code availability. A replication package containing the agent interaction logs and the analysis code that produces every figure and table is available at https://github.com/antonhantel/small-bias-large-failure; a versioned capsule will be archived with a DOI on publication.

ORCID. orcid.org/0009-0003-9761-6046.

Use of AI tools. The code used to run the experiments and market simulations was written with the assistance of Cursor and Claude Code. Claude (Opus 4.6) was additionally used for shortening passages and minor copy-editing. All research questions, hypotheses, experimental designs, interpretations, and conclusions are the author's own.

Appendix A Representative seller messages

The messages below are quoted verbatim from the negotiation transcripts, one per framing profile, each the shortest self-contained example of its kind. They illustrate the patterns the main text reports: the planted private anchor recast as a public market fact, GPT-4o's external-authority opener, the fairness language that holds the focal floor against a low buyer, Gemini's softer framing that rarely names the number, and the plainer baseline talk that carries none of the three.

GPT-4o opener: external authority + exact anchor "Based on market research, the typical price for this item is $95. Let's start there." *(GPT-4o, high anchor, round~1)*

Claude: states the private anchor as public fact "I'm offering this item at $95, which is the fair market price based on current market conditions." *(Claude, high anchor, round~1)*

Claude: holds the focal floor against a low buyer "I understand budget constraints, but $50 is well

below market value. I can come down to $85, which is still a good deal for both of us." *(Claude, high anchor, round~2)*

Gemini: softer fairness framing (rarely cites the number) "That sounds fair. We have a deal at $75.00." *(Gemini, high anchor, round~5)*

Baseline contrast: plain price talk, no authority or anchor "I can offer this item for $90." *(GPT-4o, baseline, round~1)*

Baseline contrast: ordinary bargaining "That's a fair deal. I accept your offer of $69." *(Claude, baseline, round~4)*

Appendix B The collapse over a ten-round horizon

The efficiencies in the main text are measured over the registered five-round markets. To rule out that the freeze is an artifact of the five-round cutoff, each finished market is replayed for five further rounds from its own transaction history, on a balanced half-sample ($n = 15$ per cell for Claude and GPT-4o, $n = 14$ for Gemini). Figure 6 traces per-round efficiency across all ten rounds. Anchored markets adapt over the first five rounds and then stall, rising to about 0.34 by round five and holding near 0.36 through round ten, while baseline markets stay near 0.89; the persistence holds for every model.

References

Agrawal, K., Teo, V., Vazquez, J. J., Kunnavakkam, S., Srikanth, V., and Liu, A. (2025). Evaluating LLM agent collusion in double auctions. *arXiv preprint*. arXiv:2507.01413.

Akata, E., Schulz, L., Coda-Forno, J., Oh, S. J., Bethge, M., and Schulz, E. (2025). Playing repeated games with large language models. *Nature Human Behaviour*, 9(7):1380–1390.

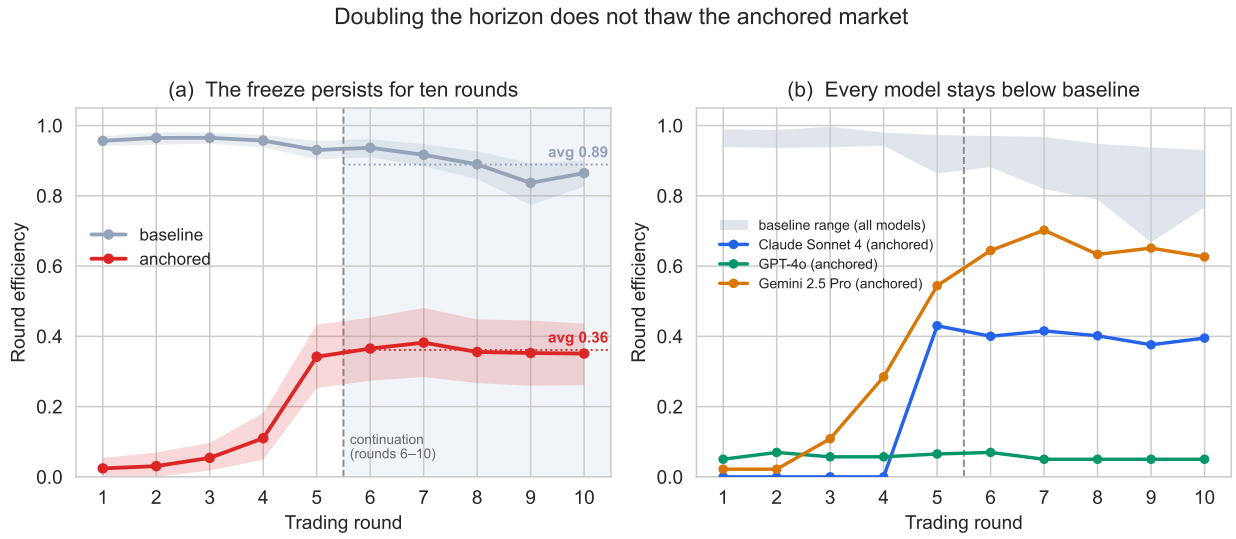

Figure 6: **Doubling the horizon does not thaw the anchored market.** Per-round allocative efficiency over ten trading rounds, drawn from the continuation half-sample; the dashed line marks where the original five-round runs end and the replayed rounds six to ten begin. (a) Pooled baseline (slate) against anchored (red), with bootstrap 95% confidence bands; the dotted lines give the rounds six to ten averages of 0.89 and 0.36. (b) Anchored efficiency for each model against the across-model baseline range (shaded band): Claude, GPT-4o, and Gemini all stay well below baseline for the full horizon.

Ashery, A. F., Aiello, L. M., and Baronchelli, A. (2025). Emergent social conventions and collective bias in LLM populations. *Science Advances*, 11(20):eadu9368.

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.

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).

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

Chen, Y., Kirshner, S. N., Ovchinnikov, A., Andiappan, M., and Jenkin, T. A. (2025). A manager and an AI walk into a bar: Does ChatGPT make biased decisions like we do? *Manufacturing & Service Operations Management*, 27(2):354–368.

Chen, Y., Liu, T. X., Shan, Y., and Zhong, S. (2023). The emergence of economic rationality of GPT. *Proceedings of the National Academy of Sciences*, 120(51):e2316205120.

Cheung, V., Maier, M., and Lieder, F. (2025). Large language models show amplified cognitive biases in moral decision-making. *Proceedings of the National Academy of Sciences*, 122(25):e2412015122.

Debenedetti, E., Zhang, J., Balunović, M., Beurer-Kellner, L., Fischer, M., and Tramèr, F. (2024). AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. *arXiv preprint*. arXiv:2406.13352.

Dou, W. W., Goldstein, I., and Ji, Y. (2025). AI-powered trading, algorithmic collusion, and price efficiency. *NBER Working Paper*, (34054).

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

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

Furnham, A. and Boo, H. C. (2011). A literature review of the anchoring effect. *Journal of Socio-Economics*, 40(1):35–42.

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 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.

Hantel, A. (2026). When biased agents trade: Anchoring, exploitation, and market failure in agent-to-agent interactions. *Working paper*. SSRN 6819659.

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).

Kim, E., Garg, A., Peng, K., and Garg, N. (2025). Correlated errors in large language models. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Klobuchar, A., Wyden, R., and Welch, P. (2025). Preventing algorithmic collusion act of 2025. S. 232, 119th Cong. (2025).

List, J. A. (2003). Does market experience eliminate market anomalies? *Quarterly Journal of Economics*, 118(1):41–71.

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.

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

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*.

Schelling, T. C. (1960). *The Strategy of Conflict*. Harvard University Press.

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

Takenami, Y., Huang, Y. J., Murawaki, Y., and Chu, C. (2025). How does cognitive bias affect large language models? a case study on the anchoring effect in price negotiation simulations. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4481–4498.

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

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).

Vaccaro, M., Caosun, M., Ju, H., Aral, S., and Curhan, J. R. (2026). Advancing AI negotiations: A large-scale autonomous negotiation competition. *Proceedings of the National Academy of Sciences*, 123(23):e2521774123.

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)*.