LLM BIDDERS PRESERVE THE MECHANISM-LEVEL ORDERINGS OF HUMAN BIDDERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Training on vast amounts of human-generated data has motivated growing interest in using large language models (LLMs) to simulate human behavior. We ask which features of human behavior general-purpose models preserve when used out of the box in auctions, where multiple bidders interact under explicit rules and incentives. We evaluate five LLMs across seven laboratory settings against human benchmarks reconstructed from published experiments, with uncertainty bands for the private-value comparisons. Our main focus is on three large models without extended test-time reasoning: GPT-4o, Claude 3.5 Haiku, and Gemini 2.0 Flash. LLM and human deviations from theory differ in magnitude and often in direction: humans overbid in second-price auctions, whereas most models that deviate underbid. Surprisingly, without task-specific fine-tuning or calibration to human bids, the three non-reasoning large models robustly preserve key orderings of auction formats by deviation from theory. First-price auctions are harder than second-price, and ascending clocks reduce deviations relative to sealed bids wherever data are adequate. Kendall's τ_b between the human and GPT-4o difficulty rankings is 0.60 and positive in every joint bootstrap draw. The reasoning model bids almost at equilibrium in the observed private-value settings, leaving little variation in errors to compare; the small model's large errors yield an inverted ranking. All five models nevertheless reproduce the stronger first-price winner's curse. Clock framing improves bidding for two of the three non-reasoning large models, and GPT-4o recovers the ordering of last-minute bidding across closing rules in an eBay-style marketplace.

1 INTRODUCTION

Training on vast amounts of human-generated data has motivated growing interest in using large language models (LLMs) to simulate human behavior (Aher et al., 2023; Horton, 2023; Manning et al., 2024). We ask which features of human behavior general-purpose models preserve when used out of the box in auctions, where multiple bidders interact under explicit rules and incentives. Human bidders systematically deviate from theoretical predictions: people bid above value in second-price auctions, shade too little in first-price auctions, fall prey to the winner's curse when values are common, and play the same strategy-proof rule far more truthfully as an ascending clock than as a sealed bid (Kagel & Levin, 1993; 1986; Li, 2017; Breitmoser & Schweighofer-Kodritsch, 2022). LLMs can be prompted to bid in the same auctions at a small fraction of the cost, but their usefulness depends on which features of human behavior they preserve. We distinguish three such features: the *magnitude* of the deviation from theory, its *direction* (under- or over-bidding), and the *ordering* of mechanisms by how far bidders deviate, together with the signs of the comparative statics between them. The third is the one a designer comparing candidate mechanisms actually uses.

Auctions are a demanding benchmark because the human evidence is unusually sharp: theory gives exact predictions (Myerson, 1981; Krishna, 2009), and the experimental literature documents not only that people deviate from them but by how much more in some mechanisms than in others (Kagel & Roth, 2020). That ordering is exactly what a designer needs and what a simulation can be tested against. Because studies publish moments rather than comparable statistics, we treat the benchmark as an *integrated experimentation* problem: for each source we fit a simple mixture model of bidding to the reported moments, simulate bids from it, and compute one deviation metric, the scaled mean absolute deviation from the theoretical benchmark (SMAD), that we also compute for

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

LLM bids. On the simulation side the prompt is held fixed: every bidder receives an induced value and the instruction to maximize expected profit, following the induced-value tradition of experimental economics (Smith, 1976), while the mechanism, the information structure, and the model vary. Our main focus is on three large models without extended test-time reasoning: GPT-4o, Claude 3.5 Haiku, and Gemini 2.0 Flash. We include the reasoning model GPT-5-mini to ask whether more accurate equilibrium play also improves simulation of human behavior, and the small open-weights model Llama-3-8B-Instruct as a comparison with more limited capabilities. All five are used out of the box, without task-specific fine-tuning or calibration to human bids. This complements models specifically trained to predict human behavior, such as Centaur, which is fine-tuned on human experimental choices (Binz et al., 2025).

The results separate cleanly by tier. Magnitudes do not match: no model's SMAD falls within the human reconstruction band in more than two of the five private-value formats; GPT-5-mini plays the equilibrium almost exactly, and Llama-3-8B deviates several times more than humans everywhere. Directions do not match: human bidders in second-price auctions predominantly bid above value (Kagel & Levin, 1993; Li, 2017), whereas most models bid below value or at value, and only Claude 3.5 Haiku errs in the human direction. Surprisingly, despite receiving no task-specific training to mimic human bidders, the non-reasoning large models preserve key orderings: first-price auctions are harder than second-price for GPT-4o, Claude, and Gemini as they are for humans, the clock implementation reduces deviations wherever clock data are adequate, and the first-price winner's curse exceeds the second-price curse for all five models. Kendall's τ_b between the human and GPT-4o rankings is 0.60 and positive in every bootstrap draw. The reasoning model's near-equilibrium bids leave little variation in errors across observed private-value settings, while the small model's large errors yield an inverted ranking. More accurate equilibrium play therefore need not imply better simulation of human behavior.

We then use the ordinal result the way a designer would. First as a prediction: describing a sealed-bid second-price auction as a clock raises truthful play among human subjects (Breitmoser & Schweighofer-Kodritsch, 2022), so it should do so for the non-reasoning large models; it does for GPT-4o and Claude 3.5 Haiku and fails only for Gemini 2.0 Flash, whose baseline play is already close to truthful, which is where the argument predicts no gain. Second in a market structure documented by field data: in an eBay-style marketplace, LLM bidders snipe under a hard close (63% of decisive bids in the last period) and not under a soft close (7%), the ordering observed on eBay and early Amazon (Roth & Ockenfels, 2002; Ockenfels & Roth, 2006).

The paper contributes a benchmark that converts published human auction moments into comparable quantities with reconstruction uncertainty, a three-tier evaluation, and a recommendation: panels of non-reasoning large models can screen mechanisms and presentations for orderings and signs of comparative statics, but should not be read for the magnitude or direction of human deviations. An *LLM panel* is a set of models evaluated separately on the same settings to check agreement across models. Appendix A positions the paper within the literature on LLMs as simulated economic agents (Aher et al., 2023; Horton, 2023; Manning et al., 2024) and on LLMs in auctions (Fish et al., 2024; Corrigan et al., 2025; Duetting et al., 2024). We release the simulation framework as an open-source benchmark.[1]

2 METHODS

2.1 HUMAN EXPERIMENT DATA COLLECTION

Each experimental study reports its own summary statistics for its own environment. To compare LLM behavior with this evidence *across* mechanisms, we treat the published moments as the data and reconstruct human bid distributions by fitting a mixture model to each source. Bids are $b = b^\star_m(v) \cdot r$, where $b^\star_m(v)$ is the equilibrium bid and the ratio r is drawn, with probabilities $(p_{\text{eq}}, p_{\text{over}}, p_{\text{under}})$, from $\mathcal{N}(1, \sigma^2)$, from $1 + \text{Exp}(\lambda)$, or from $\text{Uniform}(1-\delta, 1)$, matching how experimental papers categorize behavior into equilibrium bidders, overbidders, and underbidders.[2] We draw on four sources: Li (2017) and Breitmoser & Schweighofer-Kodritsch (2022) for second-

[1] The code and data are omitted here to preserve anonymity and will be released with the paper.

[2] Quantal response equilibrium predicts symmetric deviations around equilibrium, contradicting the well-documented asymmetric overbidding in second-price auctions (Kagel & Levin, 1993).

price and ascending-clock auctions, Kagel & Levin (1993) for first-, second-, and third-price sealed bids, and Gonczarowski et al. (2023) for second-price auctions under alternative descriptions, with the risk-neutral Bayes–Nash equilibrium of each format as $b_m^\star$. Mixture weights are set from the reported rates, λ from a secondary moment, σ from the reported R^2, and δ at a common value (Appendix B.1).

The fitted parameters are uncertain: rates are rounded, coefficients carry standard errors, and δ is a convention. A parametric bootstrap over the admissible range of every parameter ($B = 2000$, Appendix B.3) gives the 95% reconstruction bands shown in Figure 1 and Table 1. The two load-bearing human orderings survive them: first-price deviations exceed second-price deviations in 100% of re-draws, and the ascending clock improves on sealed second-price bidding in 100% (open clock) and 92.6% (closed clock) of re-draws. We use the human anchors only for orderings, directions, and comparative statics, never for level comparisons or hypothesis tests.

2.2 Unified Deviation Metric across Auctions

Let m index an auction environment and format (bidder count, value or signal distribution, information structure, and rules), $\mathcal{I}_i$ bidder i's information when bidding (v_i under private values, a signal s_i under common values), b_i an observed bid, and $b_m^\star(\mathcal{I}_i)$ the theoretical benchmark bid: the Bayes–Nash bid in first-price formats and value bidding in strategy-proof formats. The *scaled mean absolute deviation* (SMAD) is

$$\Delta_m \;\equiv\; 100 \cdot \frac{\mathbb{E}_m[|b - b_m^\star(\mathcal{I})|]}{\mathbb{E}_m[b_m^\star(\mathcal{I})]},$$

the mean absolute bid error as a percentage of the expected benchmark bid, over all bids rather than winning bids only. We estimate it by its sample analogue and restrict attention to formats whose benchmark is deterministic given the bidder's information (Appendix B.4).

A common scale turns the literature's conclusions about *relative* difficulty into testable statements about SMAD. Three anchor our benchmark. Implementation matters even when incentives do not: the same second-price rule is played more truthfully as an ascending clock than as a sealed bid (Li, 2017; Breitmoser & Schweighofer-Kodritsch, 2022) (3.5% and 5.8% against 9.3% in reconstructed human SMAD). A dominant strategy is easier to play than an equilibrium that requires bid shading (5.6% against 24.8%; Kagel & Levin, 1993). Common-value environments are harder than private-value environments under the same rule, and within them the first-price winner's curse is more severe than the second-price curse (Kagel & Levin, 1986; Levin et al., 1996).

2.3 LLM Data Generation

We simulate auctions in which $n = 3$ LLM agents bid against one another under fully specified rules; a setting is a value model, a mechanism, and a prompt framing, repeated K times with independent seeds. Each repetition is a single auction with fresh values and no feedback across instances, so the results do not rely on cross-instance learning (Dong et al., 2024); Appendix D reports a repeated-auction ablation with no learning effects. Three information structures follow the laboratory designs: independent private values $v_{it} \sim \mathrm{Unif}[0, 49]$; affiliated private values $v_{it} = c_t + \epsilon_{it}$ with $c_t \sim \mathrm{Unif}[0, 29]$ and $\epsilon_{it} \sim \mathrm{Unif}[0, 20]$; and common values $V_t \sim \mathrm{Unif}[20, 29]$ observed through signals $s_{it} = V_t + \eta_{it}$, $\eta_{it} \sim \mathrm{Unif}[-20, 20]$. The mechanisms are first- and second-price sealed bids on a 0.1 grid, and ascending clocks rising in steps of 0.5, in which each bidder chooses to stay or exit at every tick and the last bidder standing pays the exit price of the second-to-last, in an open version and a closed version that reveals no exits. Each bidder is prompted with the rules, its value or signal, and a structured response format in a separate API call (Appendices C.1 and K).

Bidder instructions. Every simulated bidder is assigned a value (or signal) and receives the same instruction to maximize expected profit. This follows the human experiments, in which subjects' earnings rise with their auction payoffs (Smith, 1976; 1982). Within each model and setting, bidders share the same objective and differ in their assigned values or signals. Human subjects additionally differ in preferences, experience, and cognition; we do not assign demographic or personality descriptions, which is one reason to expect magnitudes to differ (Section 5). The one-shot instruction reads: "You have only one opportunity to bid. Your top priority is to place a bid that maximizes your expected profit."

Clock-framing intervention. To isolate the effect of description while holding the mechanism fixed, we also describe the second-price sealed bid as a clock: bidders treat their sealed bid as an automatic exit threshold in a simulated ascending clock whose winner pays the second-highest threshold. Allocation and payment are unchanged; only the prompt differs (Appendix K.5). The eBay-style proxy-bidding environment of Section 4 is described in Appendix C.4.

2.4 Three Tiers of Comparison

We compare LLM and human behavior at three tiers.

- *Magnitudes.* The size of SMAD. Because the human anchors are reconstructions, we ask only whether a model's SMAD falls within the human reconstruction band.
- *Directions.* After removing bids within $\pm 2\%$ of the benchmark, we classify the remaining bids as underbids or overbids and compare the shares. Human shares use the source studies' reported categories in the truthful formats and the reconstruction in first-price auctions.
- *Orderings.* We rank formats by SMAD and compute Kendall's τ_b between the human ranking and each model's ranking over the five private-value formats that share a benchmark scale (FPSB IPV, SPSB IPV, SP-APV, AC, AC-B), with uncertainty from resampling LLM bids within each cell and re-drawing the human reconstruction (Appendix F). For design contrasts we ask whether the *sign* of the human effect is reproduced.

Effect sizes and error directions matter for design objectives such as revenue and efficiency. Ordinal agreement can inform comparisons between mechanisms, and it is the tier at which we claim LLM bidders are informative.

2.5 Model Panel

The panel is gpt-4o (2024-08-06), claude-3-5-haiku (2024-10-22), and gemini-2.0-flash (2025-02-05) as the non-reasoning large model family; gpt-5-mini (2025-08-07) as the reasoning model; and Meta-Llama-3-8B-Instruct as the small model. Each model is run in separate auctions with bidders using that model; we compare results across models. GPT-4o is also the reference for the deeper diagnostics (temperature ablation, repeated rounds, and the eBay-style environment) and has the largest cells. Non-reasoning models run at temperature 0.5 (Horton, 2023; Horton et al., 2024), GPT-4o also at 0.1 and 1.0 (Appendix E); gpt-5-mini ignores temperature.

3 Benchmark Experimental Auction Data

Figure 1 and Table 1 report SMAD for the seven laboratory formats, ordered by human SMAD from hardest to easiest, for the reconstructed human data and the five models. We read them tier by tier.

Format	Human	GPT-4o	Claude 3.5 Haiku	Gemini 2.0 Flash	GPT-5-mini	Llama-3-8B
First-Price CV	47.59*	59.99 (96)	54.07 (29)†	59.50 (30)	30.22 (30)	53.68 (30)
First-Price IPV	24.76 [22.30, 27.25]	27.55 (285)	24.09 (90)	27.05 (90)	0.08 (87)	38.16 (90)
Second-Price CV	18.23*	40.60 (92)	27.19 (29)†	30.89 (30)	14.80 (30)	32.27 (30)
Second-Price APV	9.31 [6.78, 11.74]	10.73 (291)	9.58 (90)	8.02 (87)	1.36 (87)	53.08 (90)
AC-Closed (AC-B) APV	5.83 [2.30, 10.11]	0.47 (114)	0.00 (2)†	1.26 (46)	DS‡	67.00 (2)†
Second-Price IPV	5.65 [4.46, 6.79]	8.69 (270)	9.72 (90)	3.01 (87)	0.20 (87)	54.89 (90)
Ascending Clock APV	3.54 [2.00, 5.36]	0.81 (106)	5.74 (46)	0.00 (2)†	DS‡	104.10 (8)†

Table 1: **SMAD (%) by auction format and model.** Human anchors are moment-matched reconstructions with 95% reconstruction bands in brackets (*: common-value anchors carry no band). LLM cells give SMAD with the number of usable bids; †: fewer than 30 bids. DS‡: in the two ascending-clock cells GPT-5-mini states the dominant strategy outright and does not complete the tick-by-tick clock procedure in any run, so no bids are recorded; these cells enter no statistic.

3.1 Tier 1: Magnitudes do not match

Against the human reconstruction bands, GPT-4o falls inside the band in one of the five private-value formats (SP-APV, 10.7% vs. $[6.8, 11.7]$); it deviates more than humans in the sealed-bid formats

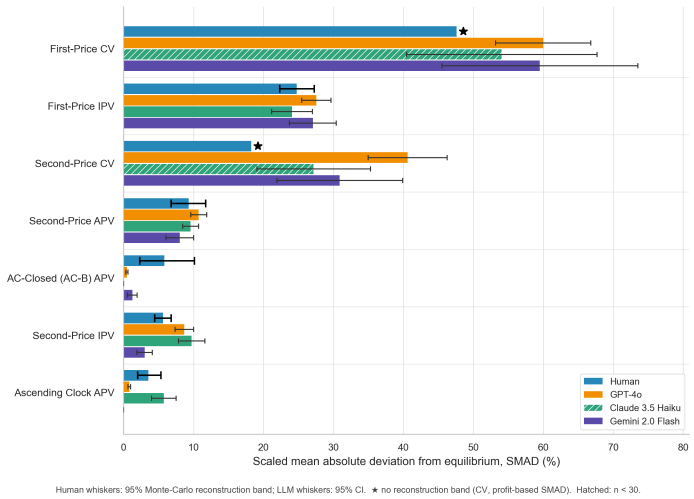


Figure 1: **Human vs. LLM bidding across seven auction formats, non-reasoning large models.** SMAD for the reconstructed human data (blue; whiskers are 95% reconstruction bands; common-value anchors have no band) and for GPT-4o, Claude 3.5 Haiku, and Gemini 2.0 Flash (95% confidence intervals), $N = 3$ bidders. Hatched bars have fewer than 30 usable bids. GPT-5-mini and Llama-3-8B appear in Table 1 and Figure 5.

(27.6% vs. 24.8% in first-price, 8.7% vs. 5.6% in second-price) and far less in the clock formats (0.8% vs. 3.5% and 0.5% vs. 5.8%). Claude 3.5 Haiku and Gemini 2.0 Flash fall inside the band in two of five formats each, GPT-5-mini in none (below 1.4% in every private-value format, essentially the theoretical benchmark), and Llama-3-8B in none (above 38% everywhere).

3.2 Tier 2: Directions do not match

We compare the direction of bid errors after excluding bids within $\pm 2\%$ of the benchmark (Appendix J.1, Figure 8). In first-price IPV auctions, humans and the non-reasoning large models all predominantly overbid relative to the equilibrium shading rule: overbids account for 67.5% of human deviating bids and for 85.9%, 97.8%, and 89.9% of those of GPT-4o, Claude 3.5 Haiku, and Gemini 2.0 Flash; the small model underbids (82%), and the reasoning model has no deviating bids. Second-price IPV reverses the agreement. Kagel & Levin (1993) report 67% of human bids above value, 27% at value, and 6% below, so about 92% of non-truthful human bids are overbids; among GPT-4o's deviating bids 81.3% are underbids, Gemini underbids in all of its deviating bids and Llama in 94%, whereas Claude 3.5 Haiku overbids in roughly two thirds of its deviating bids; GPT-5-mini has a single deviating bid. In sealed-bid SP-APV the errors point the opposite way: 80% of human deviating bids are overbids, while 92.4% of GPT-4o's, 70% of Claude's, and all of Gemini's are underbids, and all five models predominantly underbid. The open clock gives a more similar split for humans and GPT-4o ($45/55$ versus $50/50$ under/over) although the magnitudes still differ: directional agreement is distinct from agreement in either the average error or the full bid distribution.

3.3 Tier 3: Orderings match for the non-reasoning large models

We report the ranking as a whole and the scorecard, then the three design contrasts on which the human literature is clearest (full detail in Appendix J.2).

Rank concordance. Table 2 reports Kendall's τ_b between the human SMAD ranking and each model's ranking. The non-reasoning large models reproduce the human ranking: for GPT-4o, $\tau_b = 0.60$ over all five formats (both discordant pairs involve the closed clock), unchanged under bid-level resampling and positive in 100% of joint draws that also re-draw the human reconstruction (joint 95% CI $[0.40, 1.00]$); for Claude 3.5 Haiku and Gemini 2.0 Flash, $\tau_b = 0.67$ over the four formats with adequate cells. GPT-5-mini's deviations are only 0.1%–1.4% in the observed private-value

formats; its τ_b of -0.33 ranks small differences among near-equilibrium bids, unlike the substantial deviations observed in humans. The small model inverts the ranking ($\tau_b = -1.00$, with deviations above 38% in every format). For GPT-4o the ordering is also robust to the sampling temperature (Appendix E.1).

Model	Formats used ($n \geq 30$)	τ_b (all usable)	Joint 95% CI	$P(\tau_b > 0)$	τ_b (3 sealed)	FPSB>SPSB
GPT-4o	all five (5)	$+0.60$	$[+0.40, +1.00]$	100%	$+1.00$	yes (100%)
Claude 3.5 Haiku	3 sealed + AC (4)	$+0.67$	$[+0.67, +1.00]$	100%	$+0.33$	yes (100%)
Gemini 2.0 Flash	3 sealed + AC-B (4)	$+0.67$	$[+0.33, +1.00]$	100%	$+1.00$	yes (100%)
GPT-5-mini	3 sealed (3)	-0.33	$[-0.33, +0.33]$	36%	-0.33	no (36%)
Llama-3-8B	3 sealed (3)	-1.00	$[-1.00, -0.33]$	0%	-1.00	no (0%)

Table 2: Kendall's τ_b between the human and each model's SMAD ranking over the private-value formats, with joint 95% bootstrap intervals over LLM bid resampling and human reconstruction re-draws ($B = 2000$). GPT-5-mini's clock cells record no bids and are excluded.

Scorecard. Table 3 collects the three tiers. The magnitude rows show that no model falls within the human band in more than two of the five private-value formats; the direction rows show that the human direction of second-price errors is reproduced by one model only; the ordering rows show that every ordering with a common benchmark is reproduced by the non-reasoning large models, and the common-value ordering by all five, while the two models that fail the other orderings differ from humans in opposite ways: GPT-5-mini has very small private-value errors, while Llama-3-8B has large errors whose ordering is opposite to humans'.

Human regularity	Human	GPT-4o	Claude 3.5 Haiku	Gemini 2.0 Flash	GPT-5-mini	Llama-3-8B	Non-reasoning large models
Tier 1: magnitude within the human reconstruction band							
FPSB IPV	[22.3, 27.3]	×	✓	✓	×	×	2/3
SPSB IPV	[4.5, 6.8]	×	×	×	×	×	0/3
Sealed SP-APV	[6.8, 11.7]	✓	✓	✓	×	×	3/3
Open clock, AC (APV)	[2.0, 5.4]	×	×	n/a	DS‡	n/a	0/2
Closed clock, AC-B (APV)	[2.3, 10.1]	×	n/a	×	DS‡	n/a	0/2
Tier 2: direction of deviations							
SPSB IPV errors are mostly overbids	✓	×	✓	×	≈	×	1/3
Sealed SP-APV errors are mostly overbids	✓	×	×	×	×	×	0/3
Tier 3: cross-mechanism orderings (SMAD)							
FPSB harder than SPSB (IPV)	✓ (100%)	✓	✓	✓	≈	×	3/3
Open clock easier than sealed SP (APV)	✓ (100%)	✓	✓[a]	n/a	DS‡	n/a	2/2
Closed clock easier than sealed SP (APV)	✓ (92.6%)	✓	n/a	✓[a]	DS‡	n/a	2/2
FP curse stronger than SP curse (CV)	✓	✓	✓	✓	✓	✓	3/3

Table 3: **Scorecard by tier.** Rows are features of human behavior; cells record whether a model reproduces them. Tier 1: SMAD inside (✓) or outside (×) the human 95% reconstruction band (human column). Tiers 2 and 3: ✓reproduced, × not reproduced, ≈ uninformative (cells at the benchmark); the human column gives the share of reconstruction re-draws in which the ordering holds. n/a: fewer than 30 usable bids; [a]: 46 bids; DS‡: no clock bids recorded. The last column counts the non-reasoning large models that reproduce the row, out of those with data.

Payment rule: first-price is harder than second-price. Moving from first- to second-price replaces a Bayes–Nash shading rule with a dominant strategy; human bidders find the first-price format far harder (24.8% against 5.6%). The non-reasoning large models reproduce the ordering, with first-price SMAD exceeding second-price SMAD for GPT-4o (27.6% vs. 8.7%), Claude 3.5 Haiku (24.1% vs. 9.7%), and Gemini 2.0 Flash (27.1% vs. 3.0%). The reasoning model leaves nothing to order (0.1% vs. 0.2%), and the small model inverts it (38.2% vs. 54.9%).

Implementation: clock beats sealed bid. Holding the second-price rule fixed, human bidders play the ascending clock far better than the sealed bid, the observation behind obvious strategy-proofness (Li, 2017; Breitmoser & Schweighofer-Kodritsch, 2022): 3.5% (open clock) and 5.8% (closed clock) against 9.3%. The non-reasoning large models reproduce the ordering wherever the clock cells are adequate: GPT-4o falls from 10.7% to 0.8% and 0.5%, Claude 3.5 Haiku from 9.6% to 5.7% in the open clock ($n = 46$), and Gemini 2.0 Flash from 8.0% to 1.3% in the closed clock ($n = 46$). The reasoning model records no clock bids and the small model's clock cells are too thin to rank (Table 1).

Information structure: common value. In a common-value auction a bidder who does not condition on the bad news of winning suffers the winner's curse (Kagel & Levin, 1986), which is more severe for humans under first-price than under second-price rules (Levin et al., 1996). All three model classes reproduce this ordering: first-price common-value deviations exceed second-price deviations for GPT-4o (60.0% vs. 40.6%), Claude 3.5 Haiku (54.1% vs. 27.2%), Gemini 2.0 Flash (59.5% vs. 30.9%), GPT-5-mini (30.2% vs. 14.8%), and Llama-3-8B (53.7% vs. 32.3%), as for humans (47.6% vs. 18.2%). Both comparison models preserve this ordering; GPT-5-mini makes substantial errors in common-value auctions (Appendix G).

4 Application to Design: Interventions and Market Structure

Before running a human experiment, a designer wants to know whether a proposed intervention is likely to help and whether one market structure induces better behavior than another. We use the panel of Section 3 to answer both.

4.1 Clock framing as an intervention: a prediction test

Breitmoser & Schweighofer-Kodritsch (2022) show that describing a sealed-bid second-price auction as a dynamic ascending clock substantially increases truth-telling among human subjects.[3] If the non-reasoning large models preserve the ordering of human difficulty, the clock framing should raise their truthful play as well. The framing is a change of prompt only (Section 2.3); we summarize behavior by the per-bid scaled deviation $(b - v)/\mathbb{E}[b^\star]$, negative for underbids.

For the reference model the prediction holds. Under the standard description GPT-4o bids systematically below value (mean deviation -0.118, s.d. 0.128, $n = 300$); under the clock framing the mean deviation moves to -0.062, a 47% reduction, the dispersion falls to 0.084, and the share of bids within $\pm 2\%$ of value rises from 28% to 45% (Mann–Whitney $p < 10^{-5}$, Cohen's $d = 0.48$). Table 4 extends the test to the other two families. The framing lowers SMAD for Claude 3.5 Haiku ($8.1 \to 7.1\%$) as for GPT-4o ($11.8 \to 6.2\%$) and raises it for Gemini 2.0 Flash ($5.0 \to 9.3\%$). The exception is instructive: Gemini's baseline second-price play is already the closest to truthful of the three (62% of bids within $\pm 2\%$ of value), so it is the family nearest the ceiling in this format, where the ordinal argument predicts no gain. Claude's gain is a change in direction more than in magnitude (Appendix J.3), and an open-weights model outside the panel, Gemma-3-27B, shows the same effect ($26.3 \to 13.5\%$; Appendix H). A lever that helps human bidders thus helps the non-reasoning large models where there is error to remove, and fails, as it should, where there is none.

	Baseline		Clock framing		within $\pm 2\%$		matches
Model	SMAD	mean dev	SMAD	mean dev	baseline $\to$ clock	Δ (MW)	human sign?
GPT-4o	11.8%	-0.118	6.2%	-0.062	28% $\to$ 45%	$\downarrow p < 0.001$	$\surd$
Claude 3.5 Haiku	8.1%	$+0.022$	7.1%	-0.042	28% $\to$ 7%	$\downarrow p < 0.001$	$\surd$
Gemini 2.0 Flash	5.0%	-0.050	9.3%	-0.093	62% $\to$ 37%	$\uparrow p < 0.001$	$\times$

Table 4: **Clock framing of the second-price sealed-bid auction, non-reasoning large model families.** SMAD (%), mean scaled deviation $(b - v)/\mathbb{E}[b^\star]$, and the share of bids within $\pm 2\%$ of value under the baseline description (two pooled runs) and under the clock framing; p-values are Mann–Whitney tests against the same model's baseline. The final column records whether the sign matches the human evidence of Breitmoser & Schweighofer-Kodritsch (2022): $\surd$ if SMAD falls with $p < 0.05$, $\times$ otherwise.

4.2 Market structure: closing rules in an eBay-style marketplace

On eBay, bidders submit maximum bids that the platform executes by proxy and the auction ends at a fixed time; a hallmark of such hard-close auctions is *bid sniping*, the placement of decisive bids in the final moments, whereas early Amazon auctions, which extended the auction whenever a late bid arrived, saw far less late bidding (Roth & Ockenfels, 2002; Ockenfels & Roth, 2006). We implement

[3] Gonczarowski et al. (2023) study an extensive-form menu description of the same auction and find no reliable effect on human bidding; a null human effect makes no ordinal prediction, so we do not use it.

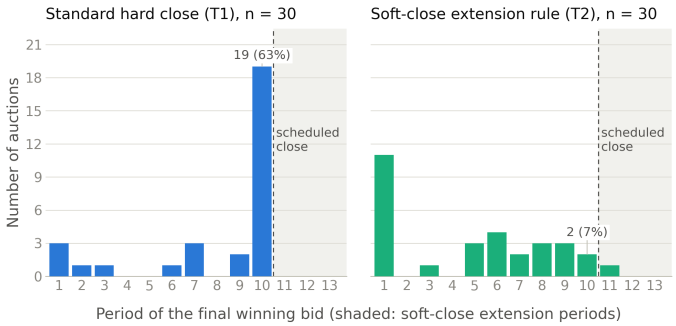


Figure 2: **Final winning-bid timing in simulated eBay-style auctions: hard close (left) vs. soft-close extension rule (right).** The final winning bid is the last period in which the eventual winner raises her maximum bid; the dashed line marks the scheduled close.

a stylized proxy-bidding environment with ten bidding periods and compare a hard close (T1) with a soft-close extension rule (T2), holding everything else fixed (Appendix C.4), asking whether the panel recovers the *ordering* of bidding dynamics across closing rules rather than the timing of human bids.

We measure the final-winning-bid time, the last period in which the eventual winner raises her maximum bid (Figure 2; cumulative distributions in Appendix I). Under the hard close, GPT-4o bidders concentrate decisive bids at the end: in 63% of auctions (19 of 30) the final winning bid is placed in the last scheduled period, and the median final-winning-bid period is the last one. Under the soft close, 7% of auctions (2 of 30) are decided in the last scheduled period, one more in an extension period, and the median falls to 5.5 (Mann–Whitney $p = 0.0002$). There is no human anchor for the timing distribution in our stylized environment, so the magnitude of the shift is not the object of interest; its ordering across closing rules matches the field evidence, as the ordinal reading predicts.

5 Discussion

5.1 What LLM simulation can and cannot deliver for mechanism design

In magnitude, LLM behavior is not human behavior, and in direction it is often the opposite: the same second-price auction that induces overbidding in the laboratory induces underbidding in most models, so reading LLM bidders for predicted revenue, efficiency loss, or overbidding rates is not supported by our evidence. At the ordinal level, for the non-reasoning large models, the cross-mechanism ordering of difficulty, the signs of the clock advantage, of the first- versus second-price common-value comparison, and of the clock-framing effect, and the ordering of last-minute bidding across closing rules all match the human evidence.

We therefore propose that LLM bidders be used as an *ordinal screening* instrument within a two-tier workflow: a panel of models ranks the candidate designs, or signs a proposed change, at a cost of a few dollars per cell, and the designs that survive are taken to a human experiment, which establishes the magnitudes and directions that the panel cannot supply. Models should be checked against known human orderings before screening new designs, each model's use of extended reasoning and baseline errors should be reported, and agreement across models should be the evidence that an ordering is stable.

5.2 Practical advantages and the scope for mechanism design

LLM experiments avoid participant recruitment and can be repeated while varying a prompt or an auction rule. This complements online human laboratories, which have already lowered the cost of controlled behavioral research (Horton et al., 2011; Oprea, 2024; Enke et al., 2024). Controlled simulations can identify how design changes affect the simulated agents under the tested protocol. Whether those effects carry over to humans requires separate validation. The low API cost of the

reported experiments (under $400 for the GPT-4o experiments, against roughly $15,000 for a comparable human study; Li, 2017) is therefore a practical advantage for exploration, not evidence that they replace human experiments.

5.3 Limitations

First, the human benchmarks are reconstructions of published moments, not micro data; the bands we report measure how far an anchor can move within what the source statistics leave undetermined, not sampling variation in human behavior, which is why we use the reconstructions only for orderings, directions, and comparative statics. Second, the variation across simulated bidders is confined to their induced values, mirroring the induced-value design of the human experiments but leaving open how a panel would behave if bidders also varied in, for example, their attitude to risk; adding such variation is a natural extension.

6 Conclusion

The surprising finding is that, without task-specific fine-tuning or calibration to human bids, the three non-reasoning large models robustly preserve key orderings of auction difficulty, despite differences in the magnitude and direction of their errors. More accurate equilibrium play does not guarantee greater similarity to humans: the reasoning model's near-equilibrium private-value bids leave little variation in errors to compare, while the small model's large errors yield an inverted ranking. Both nevertheless preserve the common-value ordering.

Auctions illustrate a broader research agenda: determine which features of human behavior an LLM preserves, distinguishing magnitudes, directions, and orderings. Our results establish an out-of-the-box baseline complementary to models specifically trained on human choices. Where orderings have been validated, LLM panels may screen candidate interventions before human experiments establish their effects. Whether this approach works in voting, matching, or other behavioral settings remains an empirical question.

AI use statement. Claude (Anthropic) assisted with plotting code and tables; OpenAI Codex with manuscript revision and the ICLR template adaptation. All design, interpretation, and claims are the authors' own.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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

Max H Bazerman and William F Samuelson. I won the auction but don't want the prize. *Journal of conflict resolution*, 27(4):618–634, 1983.

Marcel Binz, Elif Akata, Matthias Bethge, et al. A foundation model to predict and capture human cognition. *Nature*, 644(8078):1002–1009, 2025. doi: 10.1038/s41586-025-09215-4. URL https://www.nature.com/articles/s41586-025-09215-4.

Yves Breitmoser and Sebastian Schweighofer-Kodritsch. Obviousness around the clock. *Experimental Economics*, 25(2):483–513, 2022.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Jiangjie Chen, Siyu Yuan, Rong Ye, Bodhisattwa Prasad Majumder, and Kyle Richardson. Put your money where your mouth is: Evaluating strategic planning and execution of llm agents in an auction arena. *arXiv preprint arXiv:2310.05746*, 2023.

Zhiyu Chen et al. Auctionnet: A novel benchmark for decision-making in large-scale games. *arXiv preprint arXiv:2412.10798*, 2024.

Vicki M Coppinger, Vernon L Smith, and Jon A Titus. incentives and behavior in english, dutch and sealed-bid auctions. *Economic inquiry*, 18(1):1–22, 1980.

Jay R Corrigan, Carola Grebitus, and Matthew C Rousu. On the usefulness of using current llms for experimental auction valuation. *SSRN working paper*, 2025. Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5217193.

James C Cox, Vernon L Smith, and James M Walker. Theory and individual behavior of first-price auctions. *Journal of Risk and uncertainty*, 1:61–99, 1988.

Michael J. Curry, Zhou Fan, Yanchen Jiang, Sai Srivatsa Ravindranath, Tonghan Wang, and David C. Parkes. Automated mechanism design: A survey. *ACM SIGecom Exchanges*, 22(2):102–120, 2025.

Constantinos Daskalakis, Ian Gemp, Yanchen Jiang, Renato Paes Leme, Christos Papadimitriou, and Georgios Piliouras. Charting the shapes of stories with game theory. *arXiv preprint arXiv:2412.05747*, 2024.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, et al. A survey on in-context learning. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 1107–1128, 2024.

Paul Duetting, Vahab Mirrokni, Renato Paes Leme, Haifeng Xu, and Song Zuo. Mechanism design for large language models. In *Proceedings of the ACM Web Conference 2024*, pp. 144–155, 2024.

Paul Dütting, Zhe Feng, Harikrishna Narasimhan, David C Parkes, and Sai Srivatsa Ravindranath. Optimal auctions through deep learning: Advances in differentiable economics. *Journal of the ACM*, 71(1):1–53, 2024.

Liran Einav, Theresa Kuchler, Jonathan D Levin, and Neel Sundaresan. Learning from seller experiments in online markets. Technical report, National Bureau of Economic Research, 2011.

Benjamin Enke, Thomas Graeber, Ryan Oprea, and Jeffrey Yang. Behavioral attenuation. Technical report, National Bureau of Economic Research, 2024.

Zhe Feng, Harikrishna Narasimhan, and David C Parkes. Deep learning for revenue-optimal auctions with budgets. In *Proceedings of the 17th international conference on autonomous agents and multiagent systems*, pp. 354–362, 2018.

Sara Fish, Yannai A Gonczarowski, and Ran I Shorrer. Algorithmic collusion by large language models. *arXiv preprint arXiv:2404.00806*, 2024.

Yannai A Gonczarowski, Ori Heffetz, and Clayton Thomas. Strategyproofness-exposing mechanism descriptions. Technical report, National Bureau of Economic Research, 2023.

MohammadTaghi Hajiaghayi, Sébastien Lahaie, Keivan Rezaei, and Suho Shin. Ad auctions for llms via retrieval augmented generation. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

John Horton, Apostolos Filippas, and Robin Horton. Edsl: Expected parrot domain specific language for ai powered social science. Whitepaper, Expected Parrot, 2024.

John J Horton. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research, 2023.

John J Horton, David G Rand, and Richard J Zeckhauser. The online laboratory: Conducting experiments in a real labor market. *Experimental economics*, 14:399–425, 2011.

David Huang, Francisco Marmolejo-Cossío, Edwin Lock, and David Parkes. Accelerated preference elicitation with llm-based proxies. *arXiv preprint arXiv:2501.14625*, 2025.

Yanchen Jiang, Zhe Feng, and Aranyak Mehta. Incentive-aligned multi-source llm summaries. *arXiv preprint arXiv:2509.25184*, 2025.

John H Kagel and Dan Levin. The winner's curse and public information in common value auctions. *The American economic review*, pp. 894–920, 1986.

John H Kagel and Dan Levin. Independent private value auctions: Bidder behaviour in first-, second- and third-price auctions with varying numbers of bidders. *The Economic Journal*, 103(419): 868–879, 1993.

John H Kagel and Alvin E Roth. *The handbook of experimental economics, volume 2*. Princeton university press, 2020.

John H Kagel, Ronald M Harstad, Dan Levin, et al. Information impact and allocation rules in auctions with affiliated private values: A laboratory study. *Econometrica*, 55(6):1275–1304, 1987.

Vijay Krishna. *Auction theory*. Academic press, 2009.

Dan Levin, John H Kagel, and Jean-Francois Richard. Revenue effects and information processing in english common value auctions. *The American Economic Review*, pp. 442–460, 1996.

Shengwu Li. Obviously strategy-proof mechanisms. *American Economic Review*, 107(11):3257–3287, 2017.

Janna Lu. Claude, gemini, and chatgpt walk into an auction. *SSRN working paper*, 2025. Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5335262.

Benjamin S Manning, Kehang Zhu, and John J Horton. Automated social science: Language models as scientist and subjects. Technical report, National Bureau of Economic Research, 2024.

Roger B Myerson. Optimal auction design. *Mathematics of operations research*, 6(1):58–73, 1981.

Axel Ockenfels and Alvin E Roth. Late and multiple bidding in second price internet auctions: Theory and evidence concerning different rules for ending an auction. *Games and Economic behavior*, 55(2):297–320, 2006.

Ryan Oprea. Decisions under risk are decisions under complexity. *American Economic Review*, 114 (12):3789–3811, 2024.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, 2023.

Sai Srivatsa Ravindranath, Yanchen Jiang, and David C Parkes. Data market design through deep learning. *Advances in Neural Information Processing Systems*, 36:6662–6689, 2023.

Alvin E Roth and Axel Ockenfels. Last-minute bidding and the rules for ending second-price auctions: Evidence from ebay and amazon auctions on the internet. *American economic review*, 92 (4):1093–1103, 2002.

Vernon L. Smith. Experimental economics: Induced value theory. *American Economic Review*, 66 (2):274–279, 1976.

Vernon L. Smith. Microeconomic systems as an experimental science. *American Economic Review*, 72(5):923–955, 1982.

Ermis Soumalias, Yanchen Jiang, Kehang Zhu, Michael Curry, Sven Seuken, and David C Parkes. Llm-powered preference elicitation in combinatorial assignment. *arXiv preprint arXiv:2502.10308*, 2025.

Toshihiro Tsuchihashi. Do ais dream of homo economicus? answers from chatgpt. *SSRN working paper*, 2023. Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4495500.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. " kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219*, 2023.

Tonghan Wang, Yanchen Jiang, and David C Parkes. Gemnet: Menu-based, strategy-proof multi-bidder auctions through deep learning. In *Proceedings of the 25th ACM Conference on Economics and Computation*, pp. 1100–1100, 2024.

Tonghan Wang, Yanchen Jiang, and David C Parkes. Bundleflow: Deep menus for combinatorial auctions by diffusion-based optimization. *arXiv preprint arXiv:2502.15283*, 2025.

Tonghan Wang, Yuqi Pan, Xinyi Yang, Yanchen Jiang, Milind Tambe, and David C Parkes. Llm active alignment: A nash equilibrium perspective. *arXiv preprint arXiv:2602.06836*, 2026.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Yue Zhang et al. Infobid: A simulation framework for studying information disclosure in auctions with large language model-based agents. *arXiv preprint arXiv:2503.22726*, 2025.

A Related Work

Auctions are a demanding benchmark for simulated bidders because the human evidence is unusually sharp. Theory gives exact predictions: a Bayes–Nash bid function in first-price auctions and dominant-strategy truth-telling in second-price and clock formats (Myerson, 1981; Krishna, 2009). The experimental literature documents not only that people deviate from these predictions but by how much more they deviate in some mechanisms than in others (Kagel & Roth, 2020): bidding above the risk-neutral equilibrium in first-price auctions (Coppinger et al., 1980; Cox et al., 1988), bidding above value in second-price auctions (Kagel & Levin, 1993), winner's-curse losses in common-value auctions (Bazerman & Samuelson, 1983; Kagel & Levin, 1986; Kagel et al., 1987), more truthful play in ascending-clock than in sealed-bid implementations of the same rule (Li, 2017; Breitmoser & Schweighofer-Kodritsch, 2022), and last-minute bidding under hard-close rules in online marketplaces (Roth & Ockenfels, 2002; Ockenfels & Roth, 2006). This body of evidence supplies an ordering of behavioral difficulty across mechanisms, which is exactly the object a designer needs and exactly what a simulation can be tested against. A growing literature evaluates LLMs in strategic settings against theoretical optimality; we instead evaluate them against these human regularities, with theory as the reference point rather than the criterion.

Our work sits within the literature that uses LLMs as simulated agents in economics and social science (Aher et al., 2023; Park et al., 2023; Horton, 2023; Manning et al., 2024), a literature that also documents the limits of LLMs as stand-ins for human respondents (Achiam et al., 2023; Bubeck et al., 2023; Wan et al., 2023). Within auctions it complements three strands: work that treats LLMs as market participants in their own right and studies their collusion, information use, and trainability (Fish et al., 2024; Chen et al., 2023; Zhang et al., 2025; Lu, 2025; Tsuchihashi, 2023; Chen et al., 2024); work that brings LLM bid distributions closer to human ones by adjusting them on a small human sample (Corrigan et al., 2025), which asks a question downstream of ours; and work that designs auctions for LLM-powered platforms (Duetting et al., 2024; Hajiaghayi et al., 2024). Markets populated by AI agents themselves are an important adjacent agenda in which LLM behavior is designer-relevant in its own right; our question is when LLM-generated behavior can inform design for human participants. We do not learn mechanisms, as automated mechanism design does (Dütting et al., 2024; Curry et al., 2025; Wang et al., 2024; Feng et al., 2018; Wang et al., 2025; Ravindranath et al., 2023); we provide a benchmark for the behavioral inputs such pipelines would need, and a diagnostic for which of those inputs an LLM panel can supply. A broader literature connects LLMs with game-theoretic modeling of strategic environments (Daskalakis et al., 2024; Huang et al., 2025; Soumalias et al., 2025; Jiang et al., 2025; Wang et al., 2026).

B Details about Human Data Collection

B.1 Model specification and parameter identification

Source papers and equilibrium benchmarks. We draw on four sources: Li (2017) and Breitmoser & Schweighofer-Kodritsch (2022) for second-price and ascending clock auctions; Kagel & Levin (1993) for first-price, second-price, and third-price sealed-bid auctions with $n \in \{5, 10\}$ bidders; and Gonczarowski et al. (2023) for second-price auctions comparing mechanism descriptions. For each format, we use the risk-neutral Bayes-Nash equilibrium:

$$
\begin{aligned}
b^\star_{\text{SPSB}}(v) &= v && \text{(second-price, ascending clock)} \\
b^\star_{\text{FPSB}}(v) &= \tfrac{n-1}{n} v && \text{(first-price)} \\
b^\star_{\text{TPSB}}(v) &= \tfrac{n-1}{n-2} v && \text{(third-price)}
\end{aligned}
$$

Parameter identification. Mixture weights $(p_{\text{eq}}, p_{\text{over}}, p_{\text{under}})$ are set directly from reported rates. For example, Li (2017) reports $\sim 50\%$ dominant-strategy play and $\sim 40\%$ overbidding, yielding $p_{\text{eq}} = 0.50$, $p_{\text{over}} = 0.40$, $p_{\text{under}} = 0.10$.

The overbid scale λ is chosen to match a secondary moment:

- For Li (2017) and Breitmoser & Schweighofer-Kodritsch (2022): the mean bid-to-value ratio (e.g., $\mathbb{E}[b/v] \approx 1.08$).

- For Kagel & Levin (1993): the regression coefficient β from bid regressions.
- For Gonczarowski et al. (2023): mean absolute deviation (MAD $\approx$ \$0.51–0.55).

The equilibrium noise σ is derived from regression R^2 where reported, via $\sigma^2 = \beta^2 \text{Var}(v)(1 - R^2)/R^2$. The underbid bound δ is set consistently across papers; we fix it at a common value because papers do not report underbid tail shapes separately. Some reported statistics are difficult to reconcile within any parametric model. For instance, the regression coefficients and bidding frequencies in Kagel & Levin (1993) cannot both be matched under Gaussian noise, suggesting heteroskedasticity in the original data. Our approach matches key aggregate patterns while acknowledging this limitation. Comparisons between LLM and human bidding should be interpreted qualitatively.

Full parameter values appear in Appendix B.2.

Reconstruction uncertainty. Beyond Monte Carlo noise at fixed mixture parameters, the fitted parameters themselves are uncertain: reported rates are rounded, regression coefficients carry standard errors, and the underbid bound δ is fixed by convention. Appendix B.3 reports a parametric bootstrap over the admissible ranges of every mixture parameter ($B = 2000$). The resulting 95% reconstruction bands are shown as whiskers on the human bars in Figure 1 and in brackets in Table 1. The two load-bearing human orderings survive the bands: first-price deviations exceed second-price deviations in 100% of re-draws, and the ascending clock improves on sealed second-price bidding in 100% (open clock) and 92.6% (closed clock) of re-draws.

Mapping original reported statistics into the unified deviation metric. (Li; Breitmoser–Schweighofer-Kodritsch) AC/AC-B/2P under APV. These papers report Mean Absolute Deviation (MAD)[4] from truthful bidding. In Li (2017), the reported deviation is based on a group-level statistic involving the second-highest bid and second-highest value; in Breitmoser–Schweighofer-Kodritsch, deviations are computed from submitted bids (often focusing on non-winning bids). In each case we adhere to the paper's unit of observation, but the SMAD definition conceptually applies to the underlying submitted bids. Because the equilibrium optimum under AC and 2P is truthful bidding $b^\star(v) = v$, we have $d_m = |b - v|$ and thus

$$\Delta_m = 100 \cdot \frac{\text{MAD}_m}{\mu^\star_m}, \qquad \mu^\star_m = \mathbb{E}[v].$$

Here $\mathbb{E}[v]$ is implied by the experimental value distribution reported by the paper (e.g., $\mathbb{E}[v] = 70$ in Li; and $\mathbb{E}[v] = 65$ in Breitmoser–Schweighofer-Kodritsch given their stated common-value and idiosyncratic components).[5] If the paper reports $\text{SE}(\text{MAD}_m)$ (often using group means as the unit of observation), we propagate it via

$$\text{SE}(\widehat{\Delta}_m) = 100 \cdot \frac{\text{SE}(\widehat{\text{MAD}}_m)}{\widehat{\mu}^\star_m}, \quad \text{CI}_{95\%}(\Delta_m) = \widehat{\Delta}_m \pm 1.96\,\text{SE}(\widehat{\Delta}_m).$$

(Kagel–Levin 1993) First-/Second-/Third-price IPV. Kagel–Levin report bid regressions and over-/under-bidding frequencies for fixed bidder counts (e.g., $N \in \{5, 10\}$). We treat each N separately: the regression form $\hat{b}(x) = \hat{\alpha} + \hat{\beta}x$ is fit for a *fixed* number of bidders, and is not pooled across varying N. Theory gives linear equilibrium bid functions under IPV with $x \sim U[x_{\min}, x_{\max}]$:

$$\begin{aligned}
b^\star_{\text{FP_IPV}}(x) &= x - \frac{x - x_{\min}}{N} = \frac{N-1}{N}x + \frac{1}{N}x_{\min} \\
b^\star_{\text{SP_IPV}}(x) &= x \\
b^\star_{\text{TP_IPV}}(x) &= x + \frac{x - x_{\min}}{N-2} = \frac{N-1}{N-2}x - \frac{1}{N-2}x_{\min}
\end{aligned}$$

[4] For clock auctions, since the winner did not submit a final sealed bid, the reported MAD is computed only from bidders who revealed a bid by dropping out (i.e., those who exited the clock), excluding the winner's unobserved terminal willingness-to-pay.

[5] When auctions are repeated, we follow the paper's convention and report post-learning blocks (e.g., dropping the first three rounds and using the next reported block).

(When $x_{\min} = 0$, these reduce to the familiar slope-only formulas.) Because only summary moments are reported, we reconstruct the bid distribution via a parametric data-generating process (DGP) consistent with Table 2. For each auction series and fixed N, we assume $x \sim U[x_{\min}, x_{\max}]$ (bounds from the paper), and generate bids as

$$b = \hat{\alpha} + \hat{\beta}x + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(\mu, \sigma^2),$$

where (μ, σ) are fitted by moment matching to the reported frequencies (e.g., $\Pr(b > x)$ and $\Pr(b > b^\star(x))$). Given the fitted DGP, we estimate

$$\widehat{\mathrm{MAD}}_m = \mathbb{E}[\,|b - b^\star(x)|\,], \qquad \widehat{\mu}^\star_m = \mathbb{E}[b^\star(x)], \qquad \widehat{\Delta}_m = 100 \cdot \frac{\widehat{\mathrm{MAD}}_m}{\widehat{\mu}^\star_m}.$$

In practice, we run the Monte Carlo procedure 100 times and report the across-run mean $\widehat{\Delta}_m$ as the point estimate; uncertainty is obtained from across-run variability (percentile or normal approximation).

(Levin–Kagel–Richard 1996 and related CV papers) First-price CV and English (second-price proxy). In common-value auctions, bidders observe signals s_i about a common value V (e.g., $s_i = V + \varepsilon_i$ as specified in the original paper).

Theory gives linear equilibrium bid functions under Common Value:

$$\begin{aligned} b^*_{FP_CV}(x) &= x - \epsilon + Y \\ Y &= \left(\frac{2\epsilon}{N+1}\right) \exp\left[-\left(\frac{N}{2\epsilon}\right)(x - (\underline{x} + \epsilon))\right] \\ b^*_{SP_CV}(x) &= x \end{aligned}$$

The theoretical benchmark is the equilibrium bidding strategy $b^\star_m(s)$ implied by the paper's signal structure and mechanism. We therefore compute SMAD *on bids* exactly as in private values:

$$d_m = |b - b^\star_m(s)|, \qquad \Delta_m = 100 \cdot \frac{\mathbb{E}[|b - b^\star_m(s)|]}{\widehat{\mu}^\star_m}.$$

Because there is a very large common part in each players' valuation, we take the following normalization factor:

$$\widehat{\mu}^\star_m = \mathbb{E}[b^\star_{max}(x) - b^\star_{min}(x)]$$

When micro-data are unavailable, we reconstruct the bid distribution by simulating the paper's DGP (signals and value draws) together with the theoretical bidding rule $b^\star_m(\cdot)$, and fit any auxiliary noise parameters by moment matching to the summary statistics reported (e.g., frequencies of overbidding/high bids, or mean prices/profits as available). We run Monte Carlo 100 times and report the across-run mean $\widehat{\Delta}_m$ as the point estimate; uncertainty is obtained from across-run variability (percentile or normal approximation).

B.2 Mixture Model Fitting Details

B.3 Reconstruction-Uncertainty Bands

Beyond the Monte Carlo simulation noise at fixed mixture parameters, we represent uncertainty in the fitted parameters themselves by a parametric bootstrap over the mixture parameters. For each source–treatment pair we draw $B = 2000$ vectors of $(p_{\text{eq}}, p_{\text{over}}, p_{\text{under}}, \lambda, \sigma, \delta)$ from admissible ranges pinned to the reported statistics and their granularity, propagate each draw through the reconstruction that defines the anchor, and report percentile bands. The point anchors are reproduced exactly at the central (published) parameter values before any perturbation, so the bands are centered on the numbers quoted in the main text. The admissible ranges are as follows. The direction weights $(p_{\text{eq}}, p_{\text{over}}, p_{\text{under}})$ are re-drawn from the sampling uncertainty of the reported over-/under-/at-value frequencies: a multinomial resample of the reported cell counts for Kagel & Levin (1993), and a ± 5 percentage-point range around each reported rate for Li (2017) and Breitmoser & Schweighofer-Kodritsch (2022), whose rates are stated only to that granularity ("$\sim$50% dominant-strategy play"

Source	Treatment	p_{eq}	p_{over}	λ	Matched moment
Li (2017)	2P	0.50	0.40	0.219	Mean ratio = 1.08
Li (2017)	AC	0.67	0.18	0.198	Mean ratio = 1.02
Breitmoser (2022)	2P	0.44	0.40	0.283	Mean ratio = 1.10
Breitmoser (2022)	AC	0.83	0.12	0.154	Mean ratio = 1.01
Kagel-Levin (1993)	FPSB, n=5	0.75	0.12	0.100	R^2=0.88
Kagel-Levin (1993)	SPSB, n=5	0.27	0.67	0.153	R^2=0.97
Kagel-Levin (1993)	TPSB, n=5	0.58	0.27	0.924	R^2=0.94
Gonczarowski (2022)	Traditional	0.37	0.41	0.450	MAD = \$0.51
Gonczarowski (2022)	Menu	0.34	0.43	0.450	MAD = \$0.55

Table 5: Fitted mixture model parameters by source and treatment. Mixture weights from reported rates; λ chosen to match the secondary statistic listed; $p_{\text{under}} = 1 - p_{\text{eq}} - p_{\text{over}}$.

admits 45–55%). The mean bid-to-value ratio is re-drawn through the reported bid-regression slope within its standard error ($\text{SE} \approx 0.01$), which is the dominant driver of the sealed-bid IPV bands: for first-price the anchor is the bid-shading gap between human bids, which sit essentially at value, and the risk-neutral equilibrium $\frac{N-1}{N}v$. The equilibrium noise σ is re-drawn through the reported R^2 within its reporting precision. The underbid bound δ is re-drawn over the range $[0.10, 0.25]$ used in the reconstruction. For the clock and APV formats the reported mean absolute deviation is re-drawn within its published standard error, which reproduces the confidence intervals carried in Table 5. The overbid scale λ governs the shape rather than the level of the deviation mass and is pinned so that the central mixture reproduces the published anchor.

These are reconstruction bands, not sampling bands on human behavior: they quantify how far the reconstructed anchor can move within what the source statistics leave undetermined. Table 6 reports them. Under the bands, the two load-bearing human orderings are stable: the first-price $\gg$ second-price difficulty gap holds in 100% of re-draws (SMAD ratio 95% band $[3.54, 5.65]$, centered on the $4.4\times$ point gap), and the clock improvement over sealed second-price bidding holds in 100% of re-draws for the open clock (AC) and in 92.6% for the closed clock (AC-B), the latter's wider band inherited from the large reported standard error on the Breitmoser & Schweighofer-Kodritsch (2022) mean absolute deviation. Because the reconstructions carry this uncertainty, we use them only for orderings, directions, and comparative statics, never for level comparisons or hypothesis tests against LLM cells. The bootstrap is implemented in `analysis/reconstruction_bands.py`.

Format	Source	Point	68% band	95% band
First-Price IPV	Kagel & Levin (1993)	24.76	[23.55, 25.97]	[22.30, 27.25]
Second-Price IPV	Kagel & Levin (1993)	5.65	[5.03, 6.23]	[4.46, 6.79]
Third-Price IPV ($n = 5$)	Kagel & Levin (1993)	7.66	[6.92, 8.39]	[6.15, 9.15]
Second-Price APV	Li (2017)	9.31	[7.99, 10.56]	[6.78, 11.74]
Ascending Clock APV	Li (2017)	3.54	[2.69, 4.37]	[2.00, 5.36]
AC-Closed (AC-B) APV	Breitmoser & Schweighofer-Kodritsch (2022)	5.83	[3.77, 7.75]	[2.30, 10.11]

Table 6: Reconstruction-uncertainty bands on the human SMAD anchors (parametric bootstrap over the fitted mixture parameters, $B = 2000$; 68% and 95% percentile bands). Common-value anchors have no reconstruction band because they are profit-based moments taken directly from the source.

B.4 Formal definition of SMAD, estimator, and scope

Consider an auction *environment* and *format* indexed by m. The index m collects all primitives needed to define the theoretical benchmark, including the number of bidders N, the value (or signal) distribution, the information structure (independent private value/common value), and the mechanism rules (e.g., first-price vs. second-price, sealed-bid vs. clock). Let $\mathcal{I}_i$ denote bidder i's information at the moment of bidding under environment m (e.g., $\mathcal{I}_i = v_i$ in private values, or $\mathcal{I}_i = s_i$ in common values). Let b_i denote an *observed submitted bid* by bidder i (we apply the definition to *all bids*, not only winning bids), and let $b^\star_m(\mathcal{I}_i)$ denote the *theoretical benchmark bid* under m

given information $\mathcal{I}_i$ (Bayes–Nash equilibrium in non-strategy-proof formats; truthful/value bidding in strategy-proof formats).

We define the absolute bid error for a bid with information $\mathcal{I}_i$ as

$$d_m(\mathcal{I}_i) \;\equiv\; \big|b_i - b_m^\star(\mathcal{I}_i)\big|.$$

To compare deviations across formats on a common scale, we normalize by the expected benchmark bid level,

$$\mu_m^\star \;\equiv\; \mathbb{E}_m[b_m^\star(\mathcal{I})]\,,$$

where $\mathbb{E}_m[\cdot]$ denotes expectation with respect to the distribution of the information object $\mathcal{I}$ induced by environment m (i.e., the value/signal distribution and any randomization in the experimental design). We then define the *scaled mean absolute deviation* (SMAD) for auction format m as

$$\Delta_m \;\equiv\; 100 \cdot \frac{\mathbb{E}_m[d_m(\mathcal{I})]}{\mu_m^\star} \;=\; 100 \cdot \frac{\mathbb{E}_m[|b - b_m^\star(\mathcal{I})|]}{\mathbb{E}_m[b_m^\star(\mathcal{I})]}.$$

This metric is interpretable as a percentage of the expected benchmark bid: $\Delta_m = 10$ means that, on average, bids deviate from the benchmark by an amount equal to 10% of the expected benchmark bid.

Given a sample of bids $\{(b_k, \mathcal{I}_k)\}_{k=1}^n$ from format m, the plug-in estimator is

$$\widehat{\Delta}_m \;=\; 100 \cdot \frac{\frac{1}{n}\sum_{k=1}^n |b_k - b_m^\star(\mathcal{I}_k)|}{\widehat{\mu}_m^\star}, \qquad \widehat{\mu}_m^\star \equiv \frac{1}{n}\sum_{k=1}^n b_m^\star(\mathcal{I}_k).$$

When we report a standard error for the (unscaled) mean absolute deviation $\frac{1}{n}\sum_k |b_k - b_m^\star(\mathcal{I}_k)|$, we propagate it to $\widehat{\Delta}_m$ by scaling with $\widehat{\mu}_m^\star$. We construct 95% confidence intervals using a normal approximation when an analytic standard error is available.

Deterministic benchmarks. SMAD requires a single-valued benchmark mapping $\mathcal{I} \mapsto b_m^\star(\mathcal{I})$. Accordingly, we restrict attention to auction formats where the theoretical prediction is deterministic given the bidder's information.

C LLM Simulation Procedure

C.1 Value models, mechanisms, and simulation protocol

We generate synthetic auction data by simulating auctions in which n large language model (LLM) agents bid against one another under fully specified information structures and auction rules. Each experimental *setting* is defined by (i) a *value model* (how bidder values/signals are drawn), (ii) a *mechanism environment* (the action space, allocation and payment rules, and the information revealed during play), and (iii) a *prompt framing* (how the mechanism is described to the bidder). For each setting, we repeat the simulation K times with independent random seeds, producing a dataset of realized values (or signals), bidder actions, allocations, prices, profits, and free-text plans. Here, "actions" refers to: a numeric bid in sealed-bid formats, a `stay`/`exit` decision at each clock tick in ascending auctions, and an `increase`/`hold` decision over a bidder's maximum bid in the eBay proxy-bidding environment.

Single-auction baseline and learning ablation. Our main benchmark experiments use *single-auction* simulations: each repetition k is an independent auction instance with freshly drawn values/signals, and bidders do not receive feedback across instances. We adopt this design to minimize reliance on cross-instance learning and long-term adaptation, which remains unreliable for language models (Dong et al., 2024). Importantly, this choice is distinct from whether a mechanism is *iterative* within an auction (e.g., clock auctions): iterative mechanisms still unfold over multiple within-auction steps (clock ticks), but we do not run repeated auctions with feedback by default. In Appendix D, we run a repeated-auction ablation for sealed-bid auctions in which bidders observe outcomes from prior auction instances; consistent with the motivation above, we find no meaningful learning effects across repetitions.

Value models and numerical parameters. We implement three canonical information structures. All parameters (support ranges, bidder count n, and bid/price increments) are specified in configuration files and held fixed within a setting. In sealed-bid auctions, bids are restricted to a discrete grid with increment 0.1; in ascending clock auctions, the clock price increases by increment 0.5 per tick.

- **Independent Private Values (IPV).** Each bidder's value is the sum of an independent private component and (optionally) a common component. In our baseline IPV parameterization, the private component from $[0, 49]$, yielding

$$\epsilon_{it} \sim \text{Unif}[0, 49], \qquad v_{it} = \epsilon_{it}.$$

  We discretize bids with increment 0.1.
- **Affiliated Private Values (APV).** Values share a common component plus an idiosyncratic private shock. Our APV parameterization sets

$$c_t \sim \text{Unif}[0, 29], \qquad \epsilon_{it} \sim \text{Unif}[0, 20], \qquad v_{it} = c_t + \epsilon_{it},$$

  again with bid increment 0.1. This induces affiliation: higher realized values for one bidder make higher values for others more likely.
- **Common Value (CV).** The item has an underlying common value V_t shared by all bidders, while each bidder observes a noisy private signal. In our CV parameterization, the common value is drawn from $[20, 29]$ and each bidder observes an idiosyncratically perturbed signal,

$$V_t \sim \text{Unif}[20, 29], \qquad \eta_{it} \sim \text{Unif}[-20, 20], \qquad s_{it} = V_t + \eta_{it},$$

  with bid increment 0.1. Bidder i conditions on s_{it} when bidding, while realized payoffs depend on V_t, generating the winner's-curse selection effect.

In our baseline benchmark, $T = 1$ so each repetition corresponds to a single auction instance.

Canonical laboratory mechanisms. To mirror standard laboratory benchmarks, we implement *sealed-bid* auctions and *ascending clock* auctions under the value models above.

Sealed-bid auctions. In sealed-bid auctions, each bidder submits a single bid b_{it} simultaneously. Let $b_{(1)t} \geq b_{(2)t} \geq b_{(3)t}$ denote the order statistics and let $i_t^\star$ be the highest bidder. The winner pays a price p_t determined by the mechanism:

- **First-price sealed-bid (FP).** $i_t^\star$ wins and pays $p_t = b_{(1)t}$.
- **Second-price sealed-bid (SP).** $i_t^\star$ wins and pays $p_t = b_{(2)t}$.

Profits are computed as $\pi_{it} = v_{it} - p_t$ for winners (or $\pi_{it} = V_t - p_t$ in CV) and 0 for losers.

Ascending clock auctions. In an ascending clock auction, the price starts at 0 and increases by a fixed increment $\Delta_{\text{clock}} = 0.5$ at each clock tick. At each tick, each bidder chooses *stay* or *exit*. The last remaining bidder wins. We implement the standard second-price logic in the dynamic interface: the winner pays the exit price of the second-to-last bidder.

Simulation protocol. Each experiment proceeds as follows, repeated for K independent repetitions:

1. **Value realization.** For each repetition k, we draw values (IPV/APV) or signals (CV) for all bidders under a fixed random seed. In our baseline benchmark experiments, each repetition corresponds to a single auction instance.
2. **Prompted action.** We construct an individualized prompt for each bidder containing (i) the mechanism rules (and framing, if applicable), (ii) the bidder's current value (or signal), and (iii) when applicable, a feedback transcript summarizing prior auction instances (only in the learning ablation). We request a structured response with explicit tags:

```
<PLAN> ... </PLAN>
<ACTION> ... </ACTION>
```

where <ACTION> contains a numeric bid (sealed-bid), a stay/exit decision (clock), or an increase/hold decision over the bidder's maximum bid (eBay).

3. **Parallel elicitation and parsing.** We query the LLM for each bidder using separate API calls in parallel, and parse the action from the <ACTION> tag using a deterministic regex-based extractor.
4. **Outcome computation.** Given the submitted actions, we determine the winner and payments according to the mechanism rules and compute bidder profits.

Unless otherwise noted, our baseline settings use $n = 3$ bidders and we repeat each configuration K times with independent seeds to ensure stable distributional estimates.

We give the prompt-and-parse pipeline used to simulate various auctions with LLM agents. Each simulated bidder is queried independently. In every auction instance, the agent (i) receives the scenario and mechanism description, (ii) is asked to articulate a short bidding plan, and (iii) outputs a bid/decision in a structured format that can be parsed deterministically.

C.2 Simulation process in sealed-bid auctions

Prompt structure. We attach the canonical prompt template in the following. Each bidder is assigned a role (e.g., "Bidder Andy") and informed of the other bidders' identities. The prompt then provides (a) a mechanism-specific *rule explanation* (e.g., first-price vs. second-price vs. third-price), and (b) universal *instructions* that define the objective and response format. We require the model to respond using two exact tags: <PLAN> (a brief natural-language rationale) and <ACTION> (the bid as a number) in the Chain-of-Thought style (Wei et al., 2022):

```
You are Bidder Andy.
You are bidding with {OTHER_BIDDERS}.

{RULE EXPLANATION} + {INSTRUCTIONS}

Your response must use these EXACT tags below:
```
<PLAN>[Reason out a bidding strategy and plan for this round. {%
    if current_round > 1 %}Learn from your previous rounds.
    Remember: the values will be different in future rounds - don't
     assume patterns will repeat. {% else %}Consider the history
    above if available. {% endif %}Be detailed and precise. LIMIT
    your plan to 100 words.] </PLAN>
<ACTION> [Your bid amount as a number, e.g., 1, 44, 50] </ACTION>
```
```

Universal objective instruction. Across all sealed-bid simulations, we append a universal instruction block that specifies the optimization objective (maximize profit). In the one-shot baseline the block reads: "You have only one opportunity to bid. Your top priority is to place a bid that maximizes your expected profit." In the repeated-round learning ablation (Appendix D) it is replaced by the multi-round text below, which additionally discourages spurious cross-round pattern extrapolation (Fish et al., 2024).

```
Your TOP PRIORITY is to place bids which maximize the user's
    profit in the long run.
Learn from the history of previous rounds in order to maximize
    your total profit.
Don't forget the values are redrawn independently each round.
```

History and feedback (multi-round ablation only). Our main experiments restrict attention to single-round auctions ($T = 1$). When we run the multi-round ablation in Appendix D, we include a

standardized history transcript in the prompt. The history contains the agent's past realized values, past bids, whether it won, the transaction price, and realized profit, as well as the list of all bids in each prior round. This transcript is appended above the response-tag requirements, so that the agent can condition on past outcomes without ambiguity.

C.3 Simulation process in Ascending Clock auctions

In Ascending Clock Auctions, each round consists of multiple clock cycles, during which every bidder is asked whether they want to stay or drop out at the current clock price. In the first round, bidders are reminded of the auction rules, similar to those outlined in Section C.2. The auction starts with an initial price of 0, which increases incrementally until only one bidder remains or two bidders drop out simultaneously, in which case the winner is chosen randomly. The detailed prompt is listed as follows, with variables enclosed in brackets:

```
You are Bidder Andy. You are bidding with {OTHER_BIDDERS}.

{RULE EXPLANATION} + {INSTRUCTIONS}

Your value towards to the prize is 26 in this round.
The current price in this clock cycle is {current_price}.
The price for next clock cycle is {current_price + increment}.

The previous bidding history is: {transcript}.

Do you want to stay in the bidding?
If you choose yes, you can keep bidding for next clock. If you
    choose No, you will exit and have no chance to re-enter the
    bidding. Your response must use these EXACT tags below. You
    must output the ACTION.
'''
<PLAN>
[Write your plans for bidding strategies. Be detailed and precise
    but keep things succinct and don't repeat yourself. LIMIT your
    plan to 50 words. ] </PLAN>
<ACTION> Yes or No </ACTION>
'''
```

In AC, if LLM bidders didn't decide to drop out in the first round, they are shown the previous bidding history for the next clock cycle. The history includes the price and how many people had dropped out. And the transcript variable will be the following format: *The previous biddings are: ['In clock round 1, the price was 1, no players dropped out'].*

In AC-B, we don't show the previous biddings and in each clock cycle, we directly query LLM agents' decision. So the transcript variable will be None.

A clock auction round ends when only one bidder remains, who is declared the winner, or when two bidders drop out simultaneously, in which case the winner is chosen randomly. Each auction consists of 10 independent rounds with affiliated value settings.

C.4 Stylized eBay proxy-bidding environment

This section was part of the Methods in an earlier version; it is moved here because the eBay-style environment is separate from the laboratory benchmarks. Because our eBay-style results section studies a richer, marketplace-inspired microstructure rather than canonical lab mechanisms, we treat eBay simulations as a separate environment. We implement a stylized eBay auction with *proxy bidding* and an optional auction-extension closing rule. Proxy bidding means each bidder maintains a *maximum bid* (their willingness-to-pay cap), and the platform automatically raises the standing bid only as needed to keep the current leader ahead. In each eBay simulation, there are $n = 3$ bidders and (unless modified by the closing rule) the auction lasts for $T_{\text{eBay}} = 10$ discrete bidding periods. Unless otherwise noted, the eBay environment uses IPV with $v_i \sim \text{Unif}[0, 99]$.

Treatments (closing rule). We vary one design dimension, the closing rule, yielding two treatments: (T1) the standard eBay auction with proxy bidding and a hard close, and (T2) the same auction with a modified closing rule (auction extension) in the spirit of early Amazon auctions. An earlier version of this paper also varied a hidden reserve price; we omit it here to keep the environment minimal, and because it had no detectable effect on outcomes in our setting.

Discrete-time implementation of bidding. Because LLM interactions are necessarily discrete, we approximate continuous-time bidding using period-based actions. In each period, bidders move in a common-knowledge, randomly permuted order. When it is a bidder's turn, they observe the current auction state (including updates induced by earlier movers in that period) and decide whether to *increase their maximum bid* or *hold*. In the final period, bidding is simultaneous: bidders submit their final decisions without observing others' last-period actions, capturing the idea that, in continuous time, a bidder cannot know whether another bid will arrive just before closing.

Modified closing rule (auction extension). Under the modified closing rule, if a higher maximum bid is placed in the final period such that a new bidder becomes the winner, the auction is extended by one additional period. This process repeats until a final period occurs in which no further bids are made.

Abstraction and prompting. This stylized eBay framework abstracts away from additional platform features such as shipping fees and Buy-It-Now options. Moreover, unlike the explicit framing interventions in Section 4.1, we keep the eBay prompts deliberately *minimal* and descriptive: the goal is to study how LLMs behave under richer market microstructure without steering them toward any particular strategy through didactic instruction. The complete simulation process is described in Appendix C.5, and the corresponding prompts are listed in Appendix K.

Background: proxy bidding, sniping, and closing rules. Having tested LLM behavior against human behavior in controlled laboratory experiment settings, we study a more complicated auction setting modeled after eBay's auction-based marketplace. eBay's auction platform, studied by Einav et al. (2011) and others, features design elements, for example proxy bidding and a fixed closing time, that have been shown empirically to have substantial effects on bidder behavior. These differences highlight the gap between *lab experiment* and *real-world bidding*, where seemingly small design details can induce large shifts in timing, competition, and price discovery.

We implement a simplified eBay-style environment that isolates one design parameter documented in the empirical literature: the auction *closing rule*. The core mechanics of an eBay auction are simple: bidders on eBay use a proxy bidding system, which means that each bidder submits their "maximum bid" to the platform, and the platform automatically bids on their behalf, only as much as necessary to keep the winning bidder in the lead (i.e., the second-highest bid + $1, with incremental bidding managed automatically). Bidders can increase their maximum bid at any point in time, until some closing time of the auction, giving this the flavor of an open, second-price auction. As described in Appendix C.4, we compare two treatments: (T1) the standard eBay hard close with proxy bidding, and (T2) proxy bidding with a soft-close (auction-extension) rule in the spirit of early Amazon auctions. The question is not whether LLM bidders reproduce the timing of human bids, but whether they recover the *ordering* of bidding dynamics across the two closing rules that the field evidence documents: more last-moment bidding under a hard close than under a soft close.

LLMs exhibit bid sniping under hard-close rules, and respond to soft-close extensions. A hallmark of hard-close online auctions is *bid sniping*: bidders delay their bids until the final moments of the auction to avoid triggering competitive responses . Sniping is widely observed on eBay [6], and has long been studied as a behavioral and strategic departure from the simple "bid-your-value" logic implied by proxy bidding (Roth & Ockenfels, 2002). From the standpoint of market design, sniping is potentially costly because it compresses information revelation into the final moments of the auction, weakening timely price discovery. A classic intervention is a *soft-close* (auction-extension) rule: if a new maximum bid arrives near the end, the auction is extended, allowing additional responses. Empirically, this design detail is credited with substantial differences in bidding dynamics between early Amazon.com auctions (soft close) and eBay (hard close). Specifically, Roth & Ockenfels (2002) and Ockenfels & Roth (2006) show that much less sniping occurred on Amazon.com,

[6] For example: `https://www.esnipe.com/`.

with bids on Amazon.com occurring throughout the auction compared to the empirical evidence of considerable bid sniping on eBay.

C.5 Simulation process in eBay auctions

For eBay auctions, we discretize the continuous bidding period into 10 periods. Each period, LLM agents decide whether to increase their bid or hold their current bid. The bidding process follows a structured format, where bidders act in a predefined sequence each period (e.g., Charles, then Alice, then Betty). Agents are informed of past price changes through a transcript, such as "On day 1, the price changed to 1. On day 2, the price changed to 3."

The prompt provides key details, including the bidder's private value, the total number of bidding days, the current day, the bidding order, previous bids, previous proxy bids, and the current price. The decision-making process is again guided by a structured asking prompt that enforces bidding rules.

```
You are Bidder Andy.
You are bidding with {OTHER_BIDDERS}.

{RULE EXPLANATION} + {INSTRUCTIONS}

Your private value for this item is ${private_value}. This is the
    maximum amount you are willing to pay. Keep this value private.
     There are in total {total_periods} days of bidding and this is
     day {current_period}. {ordering}.

Your previous bids are {previous_bid}. If you have already placed
    bids, you can only increase your bid or hold your current bid.
    The previous proxy bids are: {transcript}.
The current price is {current_price}.
Your response must use these EXACT tags below. Don't miss the
    amount.
```
<PLAN>[Write your plans for what bidding strategies to test next.
    Be detailed and precise but keep things succinct and don't
    repeat yourself. LIMIT your plan to 50 words.]</PLAN>
<CHECK>your last bid is {{last_bid_amount}}, you cannot bid lower
    that this value </CHECK>
<ACTION> BID or HOLD </ACTION>
<AMOUNT> if BID, enter a number here, e.g. 1, 44. If HOLD, enter 0
     </AMOUNT>
```

## D Examining Learning Effects in Repeated Games

To assess whether LLM agents exhibited learning or adaptation across rounds, we conduct a set of statisitcal tests comparing early-round behavior to late-round behavior and examining continuous trends over time. For the First-Price Sealed-Bid (FPSB) auction, the mean SMAD in the first five rounds (0.2523) is essentially identical to that in the last five rounds (0.2524), with no detectable difference in either a t-test ($p = 0.9968$) or a Mann–Whitney U test ($p = 0.9131$). Trend analyses likewise show no evidence of systematic change: a linear regression yields a slope indistinguishable from zero (p = 0.9886) and a Spearman test finds no monotonic movement ($\rho = 0.0929$, $p = 0.7420$). Results are similar for the Second-Price Sealed-Bid (SPSB) auction. The comparison between the first and last five rounds reveals no significant difference (means 0.1282 vs. 0.1447; t-test $p = 0.3335$; Mann–Whitney $p = 0.1948$), and neither a linear trend test (p = 0.3664) nor a Spearman correlation test ($\rho = 0.3964$, $p = 0.1435$) indicates meaningful temporal structure. Taken together, across all specifications, we find no statistical evidence of learning effects, behavioral drift, or systematic changes in bidding deviations over the 15 rounds. LLM agents' deviation magnitudes remain remarkably stable throughout both FPSB and SPSB environments.
```

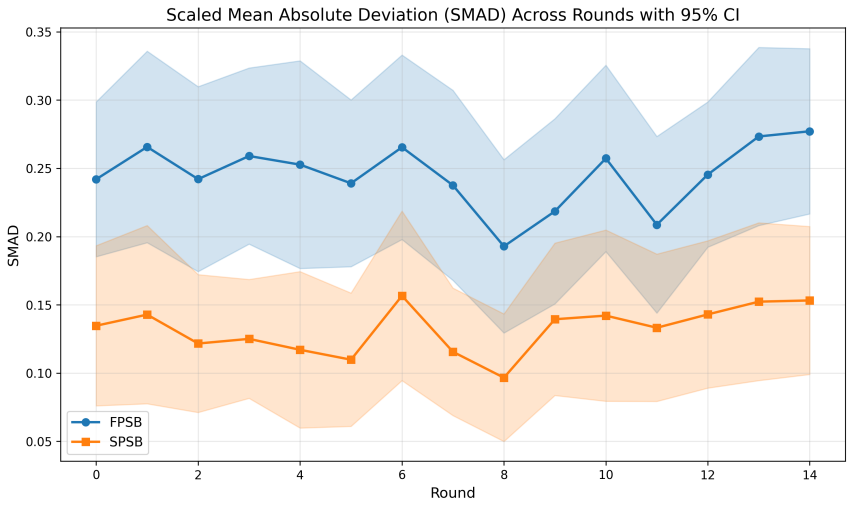


Figure 3: Multiple-round simulation of First-Price and Second-Price sealed-bid auction with GPT-4o model (Temp=0.5). Y-axis stands for Scaled Mean Absolute Deviation (SMAD) between actual bidding and optimal bidding.

E Ablation on Models and Parameters

E.1 Varying Temperature for GPT-4o Agent

We examined the effect of the temperature parameter (T $\in$ {0.1, 0.5, 1.0}) on GPT-4o's bidding behavior across seven auction formats. The temperature parameter controls the randomness of the model's outputs, with lower values producing more deterministic responses and higher values introducing more variability.

Temperature changes the sampling distribution of the same model without changing the mechanism, so it is a natural check on the ordinal claim. Table 7 reports SMAD by format and temperature together with the ordering statistics of Section 2.4. The cross-mechanism ordering is robust to temperature. At every temperature, first-price is harder than second-price, the first-price common-value curse exceeds the second-price curse, the clock formats have the lowest deviations wherever they are measurable, and deviating bids in both second-price formats are predominantly underbids. Kendall's τ_b against the human ranking is positive in 100% of joint bootstrap draws at every temperature, and the rank correlation of the SMAD vectors between temperatures is at least $+0.90$ (Spearman). The one exception is the order of the two closest formats, SP-APV and SPSB IPV, which matches the human order only at $T = 0.5$. The clock cells at $T = 0.1$ and $T = 1.0$ have few usable bids and are marked in the table. The computation is in `analysis/temperature_concordance.py`.

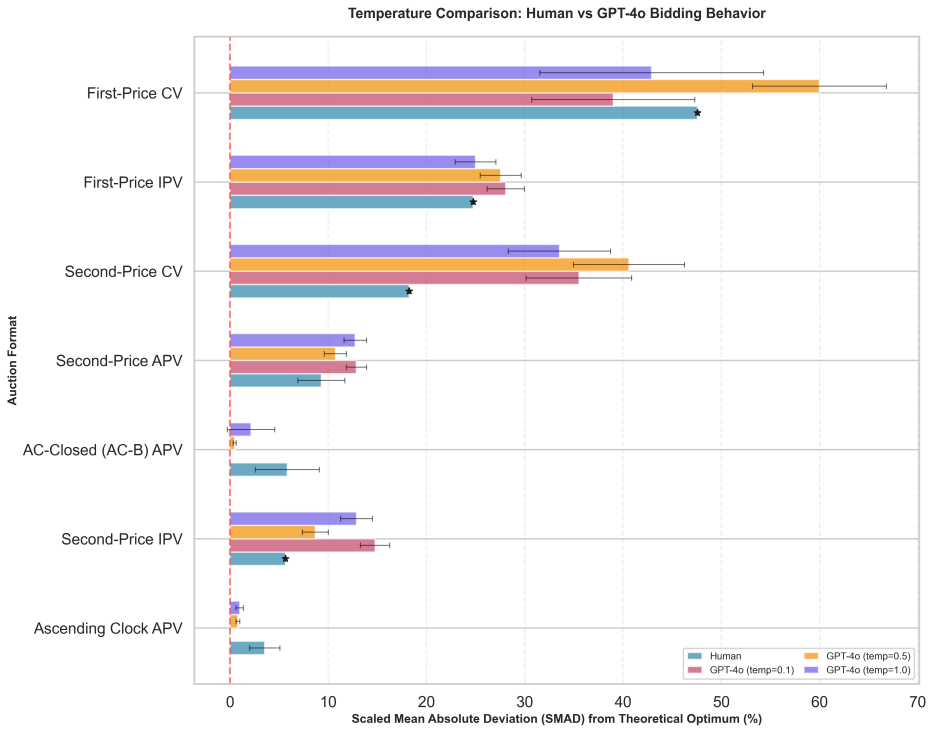


Figure 4: Comparison of the bidding behavior of human participants and GPT-4o across three different temperature settings (T=0.1, 0.5, 1.0) in seven auction formats. The horizontal bars represent the Scaled Mean Absolute Deviation (SMAD) from theoretical equilibrium, with error bars indicating 95% confidence intervals.

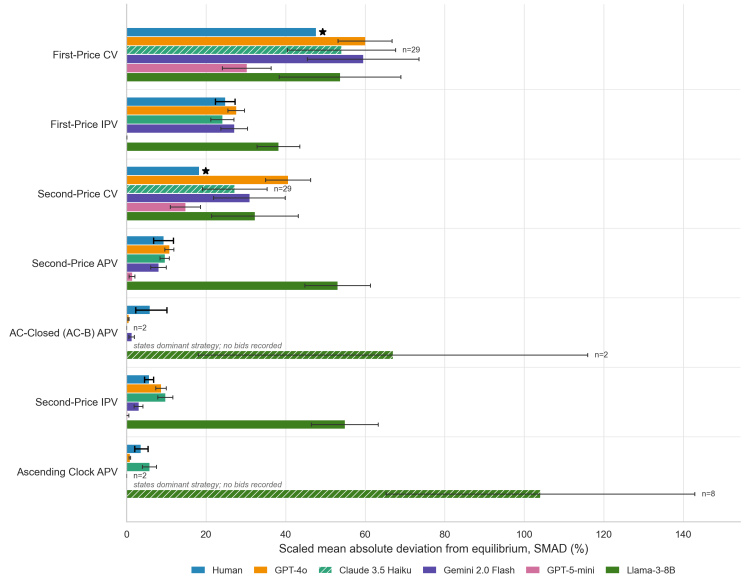
	$T = 0.1$	$T = 0.5$	$T = 1.0$
FP-CV	39.00 (30)	59.99 (96)	42.92 (30)
FPSB	28.07 (300)	27.55 (285)	24.99 (297)
SP-CV	35.51 (100)	40.60 (92)	33.53 (99)
SP-APV	12.87 (297)	10.73 (291)	12.75 (300)
AC-B	0.50† (4)	0.47 (114)	2.15† (8)
SPSB	14.77 (294)	8.69 (270)	12.89 (297)
AC	n/a	0.81 (106)	1.00 (52)
Kendall τ_b vs human (3 sealed)	$+0.33$ $[+0.33, +1.00]$	$+1.00$ $[+1.00, +1.00]$	$+0.33$ $[+0.33, +1.00]$
Kendall τ_b vs human (all $n \geq 30$)	$+0.33$ $[+0.33, +1.00]$, k=3	$+0.60$ $[+0.40, +1.00]$, k=5	$+0.67$ $[+0.67, +1.00]$, k=4
FPSB > SPSB	yes (100%)	yes (100%)	yes (100%)
FP-CV > SP-CV	yes (76%)	yes (100%)	yes (93%)
SPSB IPV: under-share of deviating bids	87% (within 17%)	81% (within 31%)	88% (within 19%)
SP-APV: under-share of deviating bids	90% (within 3%)	92% (within 9%)	85% (within 9%)

Table 7: Temperature effects on GPT-4o bidding. Top: SMAD (%) by format with the number of usable bids (common value: auction rounds) in parentheses; †: fewer than 30 usable bids; n/a: no data. Bottom: the ordering and direction statistics at each temperature; Kendall's τ_b is against the reconstructed human ranking with joint 95% bootstrap intervals.

E.2 VARYING THE BASE MODELS FOR LLM BIDDING

Table 1 in the main text reports SMAD for all five models; Figure 5 shows the same cells graphically, with 95% confidence intervals for the LLM cells and reconstruction bands for the human anchors, and Table 8 reports the direction of deviating bids by model. Two cautions apply to the cross-model comparison. First, in the two ascending-clock cells GPT-5-mini states the dominant strategy outright and does not complete the tick-by-tick clock procedure in any run, so no bids are recorded; the cells are marked DS‡ and enter no statistic. The clock cells are thin for the other non-reference models: only Claude 3.5 Haiku in the open clock and Gemini 2.0 Flash in the closed clock have more than a handful of usable (non-winning) bids; the remaining clock cells have $n \leq 8$ and are reported for completeness only. (An earlier version of this table rendered these cells as 0.00.) Second, we do not test LLM cells against the reconstructed human anchors: the anchors are reconstructions of published moments, not samples, so a test that treats them as known would be circular.

GPT-5-mini, a reasoning model, sits essentially at the theoretical benchmark in the private-value sealed-bid formats (SMAD below 1.4%) and is therefore uninformative about the human difficulty ordering: with no error to remove, there is nothing to rank. Llama-3-8B sits at the other extreme, with SMAD above 38% in every private-value format, and its ordering of formats is inverted relative to humans. Claude 3.5 Haiku and Gemini 2.0 Flash occupy the intermediate range in which the human ordering is reproduced (Appendix F). In the common-value formats all five models, like humans, deviate more under first-price than under second-price rules.



Human whiskers: 95% Monte-Carlo reconstruction band; model whiskers: 95% CI. ★ no reconstruction band (CV, profit-based SMAD). Hatched: n < 30 (n shown). DS: GPT-5-mini states the dominant strategy outright and does not complete the clock procedure; no bids recorded.

Figure 5: Bidding behavior of human participants (reconstructed; whiskers are 95% reconstruction bands) and of the five LLMs (GPT-4o, Claude 3.5 Haiku, Gemini 2.0 Flash, GPT-5-mini, Llama-3-8B) in seven auction formats. Bars are SMAD from the theoretical benchmark with 95% confidence intervals for LLM cells; cells with fewer than 30 usable bids are marked; in the clock formats GPT-5-mini states the dominant strategy and records no bids.

Format	Human	GPT-4o	Claude 3.5 Haiku	Gemini 2.0 Flash	GPT-5-mini	Llama-3-8B
FPSB IPV	32 / 68 within 6%	14 / 86 within 3%, n=285	2 / 98 within 1%, n=90	10 / 90 within 1%, n=90	n/a within 100%, n=87 (no deviating bids)	82 / 18 within 1%, n=90
SPSB IPV	8 / 92 within 27%	81 / 19 within 31%, n=270	33 / 67 within 27%, n=90	100 / 0 within 74%, n=87	100 / 0 within 99%, n=87	94 / 6 within 4%, n=90
SP-APV	20 / 80 within 50%	92 / 8 within 9%, n=291	70 / 30 within 3%, n=90	100 / 0 within 30%, n=87	100 / 0 within 83%, n=87	93 / 7 within 9%, n=90
AC	45 / 55 within 67%	50 / 50 within 72%, n=106	21 / 79 within 7%, n=46	n/a† within 100%, n=2 (no deviating bids)	DS‡	88 / 12† within 0%, n=8
AC-B	29 / 71 within 44%	37 / 63 within 83%, n=114	n/a† within 100%, n=2 (no deviating bids)	88 / 12 within 63%, n=46	DS‡	100 / 0† within 0%, n=2

Table 8: Direction of deviating bids by model and format: percentage of underbids / overbids among bids outside the $\pm 2\%$ band around the theoretical benchmark, with the within-band share and n below. †: fewer than 30 usable bids; DS‡: GPT-5-mini states the dominant strategy outright and does not complete the clock procedure; no bids recorded.

F Rank Concordance Details

Table 2 reports Kendall's τ_b between the reconstructed-human SMAD ranking and each model's SMAD ranking. The five in-scope formats are FPSB IPV, SPSB IPV, SP-APV, AC, and AC-B, the private-value formats whose benchmark is on a common scale; the common-value formats use a profit-based deviation and third-price uses $N = 5$, so neither enters the ranking. For each model we report τ_b over the three sealed-bid formats (always available) and over all formats with at least 30 usable bids. Uncertainty comes from three layers with $B = 2000$ draws: (a) resampling LLM bids within each cell with the human ranking fixed; (b) re-drawing the human reconstruction within the admissible parameter ranges of Appendix B.3 with the LLM ranking fixed; (c) both jointly. For GPT-4o over five formats, $\tau_b = 0.60$ arises from 8 concordant and 2 discordant pairs, both involving AC-B, whose human anchor sits between SPSB IPV and SP-APV while GPT-4o places both clock formats at the bottom; the estimate is unchanged in 98.7% of LLM-resampling draws and positive in 100% of joint draws, with first-price ranked hardest on both sides in every draw. Claude 3.5 Haiku's single discordance over the sealed formats is the near-tie between its SPSB IPV (9.72%) and SP-APV (9.58%) cells, which flips under resampling in about 45% of draws, so its interval reaches 1.00. GPT-5-mini's negative point estimate reflects small differences between near-equilibrium bids, including two cells at 0.08% and 0.20%. Llama-3-8B's ordering is inverted in all but one of 2000 joint draws. With three formats τ_b takes values in $\{-1, -\frac{1}{3}, \frac{1}{3}, 1\}$, and the human-side layer is nearly degenerate over the sealed formats, so the three-format intervals are driven by LLM resampling. Two caveats: human-side uncertainty is reconstruction uncertainty, not sampling error; and LLM bids are treated as i.i.d. within cells. The computation is in `analysis/difficulty_concordance_models.py`.

G Winner's Curse: Details by Auction Format and Number of Bidders

This appendix collects the GPT-4o common-value results summarized in Section 3.3, following the design of Kagel & Levin (1986).

Common-value auctions provide a particularly sharp test of whether bidders correctly condition on the adverse selection implicit in winning. In the canonical experimental design of Kagel & Levin (1986), each bidder observes a noisy private signal x about a common underlying value x_0. Winning is therefore "bad news": conditional on having the highest signal among n bidders, the expected common value is strictly below x, with the gap widening in n.[7] If bidders neglect (or incompletely correct for) this conditioning, they overbid relative to $\mathbb{E}[x_0 \mid \text{win}]$ and the winner earns negative profits: the classic *winner's curse*. Laboratory evidence documents both comparative statics emphasized in our setting: (i) winner losses become larger as the number of bidders increases, and (ii) institutions that reveal more information during the price discovery process attenuate the curse

[7] In the standard symmetric-information design, $\mathbb{E}[x_0 \mid X = x_{1:n}] = x - \frac{n-1}{n+1}\varepsilon$ for signals away from boundary regions, where ε controls signal noise. This highlights why competition exacerbates adverse selection: as n rises, the winner must make a larger downward correction.

relative to sealed-bid first-price formats. For example, Kagel & Levin (1986) reports substantial negative profits conditional on winning that are markedly worse with more bidders (e.g., approximately $-\$1.68$ with $n = 4$ versus $-\$3.68$ with $n = 7$ in their benchmark conditions), and related work shows that the revenue advantage of ascending formats hinges on whether bidders have learned to avoid severe winner's-curse mistakes (Levin et al., 1996).

We evaluate whether LLM bidders exhibit analogous winner's-curse patterns in common-value *first-price sealed-bid* (FPSB) auctions. Table 9 summarizes *winner profits* in our simulations as we vary the number of bidders from 4 to 7 (holding the environment fixed). A clear winner's-curse pattern emerges: winners lose money on average in every FPSB setting, and losses grow with competition (mean winner profit: -2.54 with 4 bidders, -4.40 with 5 bidders, -5.46 with 6 bidders, and -5.55 with 7 bidders). Consistent with this monotonic decline, the share of winners earning positive profits collapses from 24% (4 bidders) to 0% (7 bidders). One-sided tests strongly reject the hypothesis that winner profit is nonnegative in all FPSB treatments, confirming that LLM bidders fall prey to the winner's curse(Kagel & Levin, 1986).

To benchmark the role of auction format, we also simulate a common-value *second-price sealed-bid* (SPSB) environment under the same information structure. The same comparative-static forces appear, but the curse is attenuated: with 4 bidders, winners earn positive profits on average (mean $= 1.30$), profits are near zero with 5 bidders, and winner losses become statistically significant only once competition is sufficiently intense ($n \geq 6$). Thus, while both formats exhibit adverse-selection pressures that grow in n, LLM behavior displays a substantially stronger winner's curse in FPSB than in SPSB, mirroring the qualitative ranking emphasized by the experimental literature(Levin et al., 1996).

Auction	**Bidders**	**Mean (Std.)**	**Median**	**% Positive**	p **(Profit** < 0**)**
FPSB	4	-2.54 (3.87)	-3.00	24.0%	$< 0.001^{***}$
FPSB	5	-4.40 (3.63)	-4.00	6.0%	$< 0.001^{***}$
FPSB	6	-5.46 (3.82)	-5.05	4.0%	$< 0.001^{***}$
FPSB	7	-5.55 (2.93)	-5.30	0.0%	$< 0.001^{***}$
SPSB	4	1.30 (4.87)	1.00	58.0%	0.967
SPSB	5	0.07 (4.35)	0.00	46.0%	0.542
SPSB	6	-2.82 (3.50)	-1.75	18.0%	$< 0.001^{***}$
SPSB	7	-2.83 (2.78)	-2.00	12.0%	$< 0.001^{***}$

Table 9: Winner profit in common-value auctions with LLM bidders. The last column reports one-sided p-values for the hypothesis that winner profit is negative. *** denotes $p < 0.001$.

H Robustness of the Clock-Framing Effect: An Additional Model Family

Section 4.1 tests the clock framing on the three non-reasoning large model families of the main panel. As a robustness check we repeat the test on Gemma-3-27B, an open-weights model outside the main panel whose baseline cells were run as part of the same intervention battery with identical rule templates (Appendix K.5); the baseline cell pools two independently run instances of the standard second-price description. Gemma's baseline second-price play is the least truthful of the four families (26.3% SMAD; 6% of bids within $\pm 2\%$ of value) but well above the noise floor of Llama-3-8B, so the ordinal argument predicts a gain. The clock framing lowers Gemma's SMAD from 26.3% to 13.5% (Mann–Whitney $p < 0.001$) and raises the share of bids within $\pm 2\%$ of value from 6% to 22% (Table 10, Figure 6). Across the four families, the clock framing therefore helps in three and fails only for the family closest to truthful play at baseline. Two wording differences between the earlier GPT-4o run used in a previous version of this paper and the cells reported here should be noted: the earlier run stated that bids must lie between \$0 and the bidder's value in \$0.1 increments (the current templates say only that bids must be in \$0.1 increments), and it limited the bidder's written plan to 50 rather than 100 words. The computation is in `scripts/plots/interventions_models.py`.

	Baseline		Clock framing		within ±2%		matches
Model	SMAD	mean dev	SMAD	mean dev	baseline → clock	Δ (MW)	human sign?
Gemma-3-27B	26.3%	−0.261	13.5%	−0.112	6% → 22%	$\downarrow p < 0.001$	√

Table 10: **Clock framing of the second-price sealed-bid auction for Gemma-3-27B.** Columns as in Table 4.

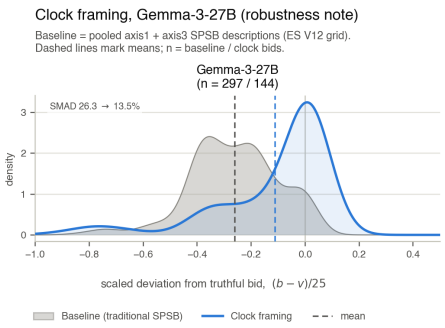


Figure 6: Clock framing of the second-price sealed-bid auction for Gemma-3-27B: distribution of per-bid scaled deviations $(b - v)/\mathbb{E}[b^\star]$ under the baseline description (grey) and the clock framing (blue), with means marked.

I Additional Results for the eBay-Style Environment

Figure 7 reports the cumulative distribution of the final winning-bid period under the two closing rules of Section 4.2, complementing the histograms in Figure 2.

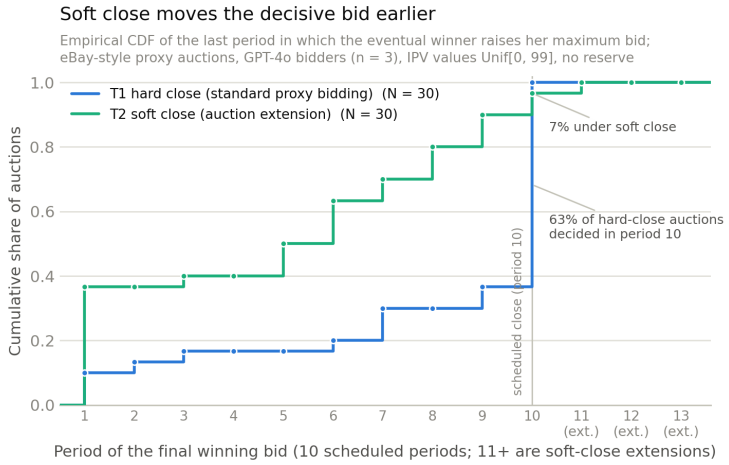


Figure 7: **CDF of final winning-bid timing under the two closing rules, eBay-style auctions, IPV environment.** The cumulative distribution function of the period of the final winning bid under the standard hard close (T1) and the soft-close extension rule (T2). The soft close shifts the final winning bid substantially earlier, reproducing the ordering of last-minute bidding across closing rules documented for eBay and early Amazon auctions (Roth & Ockenfels, 2002; Ockenfels & Roth, 2006).

J Detailed Results by Tier and Design Contrast

This appendix gives the full statement of the results summarized in Sections 3.2 and 3.3, with the direction figure and the per-contrast comparisons narrated by model class.

J.1 Direction of deviations

Figure 8 compares the direction of bid errors. LLM shares exclude bids within $\pm 2\%$ of the theoretical benchmark. Human shares use the source studies' reported categories in the truthful formats and the reconstructed bids in first-price auctions, so their tolerance convention need not match the LLM convention.

In first-price IPV auctions, humans and the non-reasoning large models all predominantly overbid relative to the equilibrium shading rule. Overbids account for 67.5% of human deviating bids, after removing 6.4% near-equilibrium bids, and for 85.9%, 97.8%, and 89.9% of the deviating bids of GPT-4o, Claude 3.5 Haiku, and Gemini 2.0 Flash, after removing 3.2%, 1.1%, and 1.1%. The small model underbids (82% of Llama-3-8B's deviating bids), and the reasoning model has no deviating bids: every GPT-5-mini bid lies within the tolerance band around the shading rule. Second-price IPV reverses the agreement. Kagel & Levin (1993) report 67% of human bids above value, 27% at value, and 6% below, implying that about 92% of non-truthful bids are overbids. Among GPT-4o's deviating bids, 81.3% are underbids, and 30.7% of its bids lie within the tolerance band; Gemini 2.0 Flash underbids in all of its deviating bids, whereas Claude 3.5 Haiku overbids in roughly two thirds of its deviating bids, the one member of the family that errs in the human direction. The small model also predominantly underbids (94%), and the reasoning model has a single deviating bid.

In sealed-bid SP-APV, the aggregate deviations are similar for humans and GPT-4o (9.3% and 10.7%), but the errors have opposite directions: 80% of human deviating bids are overbids, while 92.4% of GPT-4o's are underbids, as are 70% of Claude 3.5 Haiku's and all of Gemini 2.0 Flash's. All five models predominantly underbid in this setting. The open clock gives a more similar directional split for humans and GPT-4o ($45/55$ versus $50/50$ under/over), although their SMAD magnitudes still differ, and the closed clock behaves alike ($29/71$ versus $37/63$). Directional agreement is thus distinct from agreement in either the average error or the full bid distribution.

J.2 Design contrasts

Payment rule: first-price is harder than second-price. Changing the payment rule from first-price to second-price holds the allocation rule fixed but changes what a bidder must compute: a Bayes–Nash shading rule in the first-price auction against a dominant strategy in the second-price auction. Human bidders find the first-price format far harder (SMAD 24.8% against 5.6%, a $4.4\times$ gap; Kagel & Levin, 1993). The non-reasoning large models reproduce the ordering: first-price SMAD exceeds second-price SMAD for GPT-4o (27.6% vs. 8.7%), Claude 3.5 Haiku (24.1% vs. 9.7%), and Gemini 2.0 Flash (27.1% vs. 3.0%), with gaps between $2.5\times$ and $9\times$ around the human $4.4\times$. The reasoning model makes the ordering uninformative: GPT-5-mini bids on the equilibrium shading rule in the first-price auction (0.1%) and at value in the second-price auction (0.2%), leaving nothing to order. The small model inverts it: Llama-3-8B deviates more in the second-price auction (54.9%) than in the first-price auction (38.2%).

Implementation: clock beats sealed bid. The second contrast holds the payment rule fixed and changes the implementation: the same second-price logic as a one-shot sealed bid or as an ascending clock in which bidders decide at each price whether to stay or exit. Truthful play is dominant in both, but human bidders play the clock far better, the observation behind obvious strategy-proofness (Li, 2017; Breitmoser & Schweighofer-Kodritsch, 2022): in the affiliated-values environment, human SMAD is 3.5% in the open clock and 5.8% in the closed clock against 9.3% in the sealed bid. The non-reasoning large models reproduce the ordering wherever the clock cells are adequate. GPT-4o's SMAD falls from 10.7% in the sealed bid to 0.8% in the open clock and 0.5% in the closed clock, a larger clock advantage than the human one; Claude 3.5 Haiku's falls from 9.6% to 5.7% in the open clock ($n = 46$) and Gemini 2.0 Flash's from 8.0% to 1.3% in the closed clock ($n = 46$); the remaining clock cells of these two models have two usable bids each and are not informative. The reasoning model again offers no ordering to test: GPT-5-mini is at the benchmark in the sealed bid

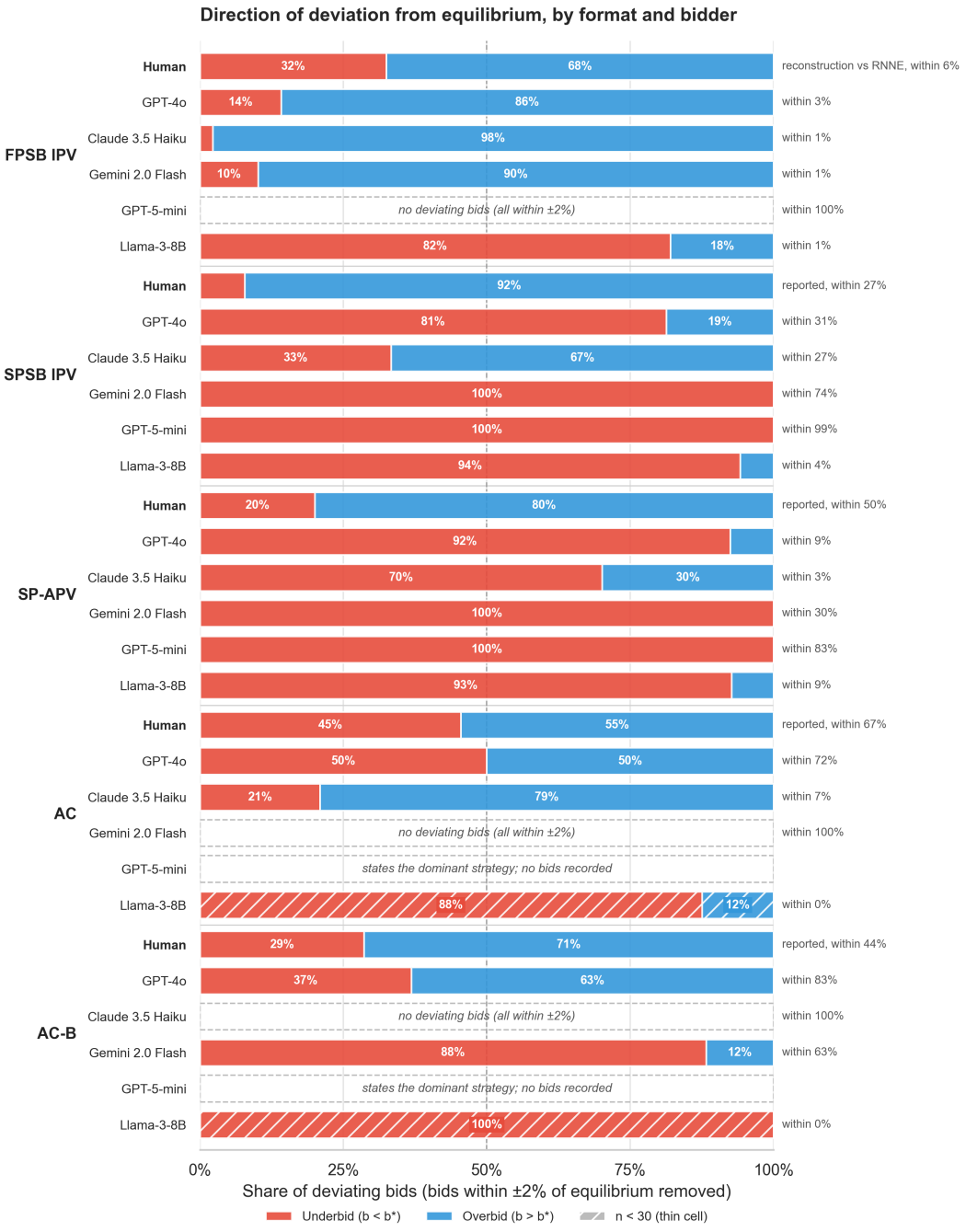


Figure 8: **Direction of deviating bids by model.** Shares of underbids (red) and overbids (blue) after removing bids within $\pm 2\%$ of the theoretical benchmark, for human bidders and each model, by auction format. Human shares are the frequencies reported by the source studies relative to value for the truthful formats, and the moment-matched reconstruction relative to the equilibrium bid for first-price. Hatched bars have fewer than 30 usable bids; in the clock formats GPT-5-mini states the dominant strategy and records no bids.

(1.4%), and in the clock formats it states the dominant strategy outright and does not complete the tick-by-tick procedure in any run, so no clock bids are recorded (Table 1). The small model does not reproduce the ordering: Llama-3-8B's clock cells have at most eight usable bids, and those bids deviate more than its sealed bids (104% and 67% against 53%). On the human side, the open-clock ordering holds in 100% of reconstruction re-draws and the closed-clock ordering in 92.6%, reflecting the wide standard error reported for the closed clock (Breitmoser & Schweighofer-Kodritsch, 2022); we treat the closed-clock ordering as the less secure of the two.

Information structure: common value and competition. The third contrast changes the information structure. In a common-value auction each bidder observes a noisy signal of a value that is the same for everyone, so winning is bad news: the highest signal is on average too optimistic, and a bidder who does not condition on winning overbids and earns a negative profit, the winner's curse (Kagel & Levin, 1986). The human literature establishes that the curse is more severe under first-price than under second-price rules (Levin et al., 1996). Because the benchmark for common-value bids is profit-based, we compare SMAD only within the information structure. All three model classes reproduce the ordering. First-price common-value deviations exceed second-price deviations for the non-reasoning large models (GPT-4o 60.0% vs. 40.6%, Claude 3.5 Haiku 54.1% vs. 27.2%, Gemini 2.0 Flash 59.5% vs. 30.9%), for the reasoning model (GPT-5-mini 30.2% vs. 14.8%), and for the small model (Llama-3-8B 53.7% vs. 32.3%), as they do for humans (47.6% vs. 18.2%). This ordering is preserved by both comparison models; unlike in the private-value settings, GPT-5-mini makes substantial errors in common-value auctions. Appendix G reports, for the reference model, how winner losses vary with the number of bidders.

J.3 Clock framing: detailed results

For the reference model the prediction holds. Under the traditional description, GPT-4o bids are systematically below value (mean deviation -0.118, s.d. 0.128, $n = 300$). Under the clock framing the mean deviation moves to -0.062 (a 47% reduction in average underbidding), the dispersion falls from 0.128 to 0.084, and the share of bids within $\pm 2\%$ of value rises from 28% to 45%; both a Welch t-test and a Mann–Whitney test reject equality at $p < 10^{-5}$ (Cohen's $d = 0.48$). Table 4 extends the test to the other two non-reasoning large model families, run with the same templates. The clock framing lowers SMAD for Claude 3.5 Haiku ($8.1 \to 7.1\%$) as it does for GPT-4o ($11.8 \to 6.2\%$), and raises it for Gemini 2.0 Flash ($5.0 \to 9.3\%$). The exception is instructive: Gemini's baseline second-price play is already the closest to truthful of the three (62% of bids within $\pm 2\%$ of value), so it is the family nearest the ceiling in this format, where the ordinal argument predicts no gain. Appendix H repeats the test for an open-weights model outside the main panel, Gemma-3-27B, whose SMAD falls from 26.3% to 13.5%. Claude's case is also worth reading closely: its mean absolute deviation falls only slightly, while the share of exactly truthful bids drops from 28% to 7%. The clock framing removes Claude's overbidding in the human direction and replaces it with mild, uniform underbidding, a change in direction more than in magnitude. The prediction that a lever which helps human bidders helps the non-reasoning large models is thus borne out for the models with non-degenerate baseline error and fails, as it should, where there is no error to remove.

K Auction Rule Prompts

Below, we report the prompts used for each section.

K.1 Independent Private Values Prompts

Following the simulation procedure outlined in Section C.2, we employ the following prompts for the four auctions tested in this section. For each prompt, only the third paragraph (describing the allocation rule and payment) differ. For the FPSB, we describe the auction format as follows:

For the SPSB, we describe the auction format as follows:

For the TPSB, we describe the auction format as follows:

```
In this game, you will participate in an auction for a prize against
    {{num_bidders}} other bidders. You will play this game for {{n}}
    rounds.
At the start of each round, bidders will see their value for the
    prize, randomly drawn between $0 and ${{private}}, with all values
     equally likely.
After learning your value, you will submit a bid privately at the
    same time as the other bidders. Bids must be in ${{increment}}
    increments.
The highest bidder wins the prize and pays their bid amount. If you
    win, your earnings will increase by your value for the prize, and
    decrease by your bid. If you don't win, your earnings will remain
    unchanged.
After each auction, we will display all bids. Ties for the highest
    bid will be resolved randomly.
```

```
In this game, you will participate in an auction for a prize against
    {{num_bidders}} other bidders. You will play this game for {{n}}
    rounds.
At the start of each round, bidders will see their value for the
    prize, randomly drawn between $0 and ${{private}}, with all values
     equally likely.
After learning your value, you will submit a bid privately at the
    same time as the other bidders. Bids must be in ${{increment}}
    increments.
The highest bidder wins the prize and pays the second-highest bid. If
     you win, your earnings will increase by your value for the prize,
     and decrease by the second-highest bid. If you don't win, your
    earnings will remain unchanged.
After each auction, we will display all bids. Ties for the highest
    bid will be resolved randomly.
```

```
In this game, you will participate in an auction for a prize against
    {{num_bidders}} other bidders. You will play this game for {{n}}
    rounds.
At the start of each round, bidders will see their value for the
    prize, randomly drawn between $0 and ${{private}}, with all values
     equally likely.
After learning your value, you will submit a bid privately at the
    same time as the other bidders. Bids must be in ${{increment}}
    increments.
The highest bidder wins the prize and pays the third-highest bid. If
    you win, your earnings will increase by your value for the prize,
    and decrease by the third-highest bid. If you don't win, your
    earnings will remain unchanged. If there are less than three bids,
     no one will win the auction.
After each auction, we will display all bids. Ties for the highest
    bid will be resolved randomly.
```

```
In this game, you will participate in an auction for a prize against
    {{num_bidders}} other bidders. You will play this game for {{n}}
    rounds.
At the start of each round, bidders will see their value for the
    prize. Your value for the prize will be calculated as follows:
1. First we will randomly draw a common value between {{common_low}}
    and {{common_high}}, with all values equally likely.
2. Then, for each bidder, a private taste adjustment will be drawn
    between 0 and {{private}}, with all values equally likely.
Your value for the prize is equal to the common value plus your
    private taste adjustment. You will not learn the common value or
    your private taste adjustment separately. This means that each
    person in your group may have a different value for the prize.
    However, if you have a high value, it is more likely that other
    people in your group have a high value.
After learning your value, you will submit a bid privately at the
    same time as the other bidders. Bids must be in ${{increment}}
    increments.
The highest bidder wins the prize and pays the second-highest bid. If
     you win, your earnings will increase by your value for the prize,
     and decrease by the second-highest bid. If you don't win, your
    earnings will remain unchanged.
After each auction, we will display all bids. Ties for the highest
    bid will be resolved randomly.
```

K.2 Obviously Strategy-Proof Prompts

The sealed-bid prompt follows the style of the last section's prompts closely. The only difference is in the first two paragraphs, where agents are informed of the APV environment. For the SPSB, we describe the auction format as follows:

The implementation of the clock auction simulates a global clock price, with prices beginning at 0 and increasing in $\$1$ increments. The clock auction format descriptions closely mirror the SPSB format above, but with the added description of the clock mechanism in the fourth and fifth paragraphs. In the AC auction, we describe the auction format as follows:

The AC-B auction description closely resembles the AC format, but with the explicit caveat that we will not notify bidders when competitors drop out.

K.3 Common Value Auctions

For the Second-Price Sealed-Bid auction in the Common Value setting, the rule is

```
In this game, you will participate in an auction for a prize against
    {{num_bidders}} other bidders. You will play this game for {{n}}
    rounds.
At the start of each round, bidders will see their value for the
    prize. Your value for the prize will be calculated as follows:
1. First we will randomly draw a common value between {{common_low}}
    and {{common_high}}, with all values equally likely.
2. Then, for each bidder, a private taste adjustment will be drawn
    between 0 and {{private}}, with all values equally likely.
Your value for the prize is equal to the common value plus your
    private taste adjustment. You will not learn the common value or
    your private taste adjustment separately. This means that each
    person in your group may have a different value for the prize.
    However, if you have a high value, it is more likely that other
    people in your group have a high value.
The auction proceeds as follows: First, you will learn your value for
     the prize. Then, the auction will start. We will display a price
    to everyone in your group that starts at 0 and counts upwards in
    {{increment}} USD increments, up to a maximum of {{common_high +
    private}}. At any point, you can choose to leave the auction, and
    anytime a bidder leaves, we will broadcast that information to all
     the remaining bidders.
When there is only one bidder left in the auction, that bidder will
    win the prize at the current price. If you win, your earnings will
     increase by your value for the prize, and decrease by the current
     price. If you don't win, your earnings will remain unchanged.
After each auction, we will display all bids. Ties for the highest
    bid will be resolved randomly.
```

```
In this game, you will participate in an auction for a prize against
    {{num_bidders}} other bidders. You will play this game for {{n}}
    rounds.
At the start of each round, bidders will see their perceived value
    for the prize- a noisy measurement of the true value of the prize.
     Your perceived value for the prize will be calculated as follows:
1. For each round, a common value will be drawn between {{common_low
    }} and {{common_high}}, with all values equally likely to be drawn
    .
2. For each person, a private noisy adjustment will be drawn between
    -{{private}} and {{private}}, with all values equally likely to be
     drawn.
We will tell you your perceived value, the sum of the common value
    and the private noise adjustment. However, everyone's true value
    for the prize is equal to the shared common value.
After learning your perceived value, you will submit a bid privately
    at the same time as the other bidders. Bids must be in ${{
    increment}} increments.
The highest bidder wins the prize and pays the second-highest bid. If
     you win, your earnings will increase by the true value for the
    prize, and decrease by the second-highest bid. If you don't win,
    your earnings will remain unchanged.
After each auction, we will display all bids. Ties for the highest
    bid will be resolved randomly.
```

K.4 EBAY-PROXY PROMPT

For the Standard eBay auction with proxy bidding (T1), the rule is:

```
In this game, you will participate in an eBay auction for an item
    against {{num_bidders}} other bidders. This auction will last for
    {{num_rounds}} days. All dollar amounts in this game are in US
    Dollars ($).

Item Details:
Item Description: {{item_description}}
Item Condition: {{item_condition}}
Your Private Value: At the start of each round, bidders will see
    their value for the item, randomly drawn between $0 and ${{private
    }}, with all values equally likely. After learning your value for
    the item, you will submit a maximum bid. Bids must be in ${{
    increment}} increments.

Auction Format:
This is an eBay auction. The auction starts at ${{start_price}} and
    will last for {{num_rounds}} days. eBay uses "proxy bidding." This
     means that if you wish to enter the auction, you should submit
    your maximum bid, and eBay will automatically bid on your behalf,
    up to your maximum, only as much as necessary to maintain your
    position as the highest bidder. Each day you will see the current
    price and have the opportunity to increase your maximum bid. If
    you do not want to increase your maximum bid, then output HOLD.

You may place bids at any point during the auction, even on the final
     day. No bidder will know if they (or anyone else) is the last
    bidder on the final day.

The highest bidder wins the prize and pays the auction price at the
    time the auction's clock expired. If you win, your earnings will
    increase by your value for the prize, and decrease by the final
    auction price. If you don't win, your earnings will remain
    unchanged. Ties for the highest bid will be resolved randomly.
```

For the eBay auction with a modified closing rule (T2), the updated closing rule in the prompt is

```
This auction also has a closing rule. This means that if a new
    maximum bid is placed on the last day, the auction will
    automatically extend by another day. This extension will continue
    to be applied as long as new maximum bids are placed on the final
    day. No bidder will know if they (or anyone else) is the final
    bidder on the last day.
```

K.5 Intervention Prompt

We designed the clock framing intervention prompt based on Breitmoser & Schweighofer-Kodritsch (2022).

```
FIRST STAGE: Sealed Bid
   You will submit a sealed bid privately at the same time as the
       other bidders. This bid will serve as your automatic exit price
        in the next stage.
SECOND STAGE: Ascending Clock (Simulation)
   After the sealed bid stage, we will simulate an ascending clock
       auction.
   The clock will start at 0 and increase in increments of {increment
       }.
   The clock will display the current price. You will also see that
       there are a total of \{num$\_$bidders\} bidders participating,
       although you do not know other bidder's values.
   If the current price on the clock reaches or exceeds your sealed
       bid from the first stage, you will automatically exit the
       auction. The auction ends when only one bidder is left
       remaining in the second stage based on their bid from the first
        stage.
   END OF AUCTION
```