WHEN LLMS BID RATIONALLY, AND WHEN THEY DON'T: A TRACEABLE CAPABILITY SPLIT IN SEALED-BID AUCTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly proposed as autonomous economic agents, but whether they comply with incentive-compatible strategies in a canonical mechanism is rarely tested against closed-form ground truth. We study this in sealed-bid auctions, testing four LLMs against second- and first-price mechanisms, each real LLM bidding against synthetic opponents that play the closed-form equilibrium exactly. We find a clean capability split: two models, Gemini 3.5 Flash-Lite and gpt-oss-120b, bid within half a cent of their valuation in every primary second-price trial, contrasting sharply with the well-documented human tendency to overbid. The other two fail in different ways. Qwen2.5-7B underbids robustly ($-\$7.42$), significant under every check we ran. Llama 3.1 8B is noisier: a positive mean deviation of $+\$6.88$ driven by a few extreme overbids at low valuations, significant under our designated per-model test but not under a draw-collapsed or rank-based check, and, at five bidders, degrading toward bids indistinguishable from uniform-random noise. Each of the two smaller models states one mechanism-invariant rule rather than conditioning it on the payment rule (Llama: bid your valuation; Qwen: shade below it), and each errs, in the direction its rule predicts, in exactly the mechanism where that rule is wrong, though this crossed pattern rests partly on a lenient, direction-only rubric for Qwen's first-price statements. This double dissociation also produces a striking aggregation artifact: naively pooling all four models in our primary truthfulness test erases the effect entirely ($-\$0.14$, $p = 0.85$), since the two large, opposite-signed per-model effects nearly cancel under pooling, a cautionary result for any LLM-agent benchmark that aggregates across models rather than reporting per-model effects. We also test Regret-Conditioned Prompting (RCP), an inference-time feedback loop conditioning a model on its realized counterfactual regret each round, against a theorem-reminder control that simply restates the optimal rule. The theorem reminder largely corrects our weakest model's first-price overbidding on its own; RCP improves faster from a far worse start, surviving Holm-Bonferroni correction under model-based errors ($p_{adj} = 0.008$) but only reaching the edge of conventional significance under a cluster-robust check ($p = 0.053$), and it ends no closer to target than no feedback at all. A matched follow-up on our underbidding model finds no comparable effect on either mechanism. Several analysis choices, including the model roster and three of the five primary tests, were revised after seeing data; each revision is summarized in the paper and detailed in the appendix. Anonymized code and data are in the supplementary material.

1 INTRODUCTION

Large language models are increasingly deployed as autonomous participants in economic settings (negotiating, pricing, bidding on behalf of users or as multi-agent components), yet most evaluations of their "rationality" rely on human judgment of plausible behavior rather than a closed-form, game-theoretic ground truth. Sealed-bid auctions offer a sharp testbed: second-price auctions have a dominant strategy (bid your true valuation) independent of beliefs about opponents, and first-price

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.

auctions have a well-characterized Bayes-Nash equilibrium under standard distributional assumptions (Vickrey, 1961; Riley & Samuelson, 1981), so a model's deviation can be measured exactly, in dollars, unlike humans' well-documented overbidding in both mechanisms (Cooper & Fang, 2008; Kagel & Levin, 2016). LLM agents in game-theoretic and economic settings are now a fast-growing subfield (surveys: Sun et al. (2025); Hao et al. (2026), who name a lack of rigorous formalization linking LLM behavior to game-theoretic ground truth as a central gap). We adopt Hao et al. (2026)'s deviation-gain construct (§2.1) as our exactly-computable metric and as the feedback quantity our RCP intervention conditions on each round.

We test five hypotheses spanning truthfulness, knowledge-behavior dissociation, bounded-rationality model fit, regret-conditioned correction, and a descriptive model comparison, against four models (Gemini 3.5 Flash-Lite, gpt-oss-120b, Llama 3.1 8B, Qwen2.5-7B) across both mechanisms. Our headline finding is a clean capability split, not a uniform bias: two models are near-perfectly compliant with the theoretical target in the primary second-price, zero-temperature test, while the other two fail in different ways, one structured and one noisy. Their stated rules offer a partial account: each of the two smaller models states one mechanism-*invariant* rule instead of conditioning it on the payment rule, Llama "bid your valuation" and Qwen "shade below it," each correct for one mechanism and wrong for the other, and each errs, where its rule is wrong, in the direction that rule predicts. The account is only partial, not complete: Llama is also behaviorally wrong in the second-price auction, where its own stated rule is correct, a genuine dissociation between what it states and what it does (§4.2); Qwen's first-price errors, by contrast, are not a comparable dissociation, since its own stated heuristic already overshoots the Bayes-Nash fraction and its bids simply follow that heuristic's own imprecision (§4.2). Pechon-Elkins & Chun (2026) caution that fitted QRE (McKelvey & Palfrey, 1995) precision parameters are payoff- and model-dependent summaries rather than structural rationality measures; we use QRE only descriptively, alongside our closed-form targets, rather than as a claim about underlying rationality. Zhu (2025) similarly diagnose LLM strategic behavior via a "reasoning equilibrium" over prompts, without a closed-form benchmark or an intervention that our Vickrey/BNE ground truth and RCP correction supply; Lotfi & Ghrayeb (2026) find analogous theory-consistent VCG spectrum-auction bidding that departs, beneficially, under budget constraints. The declarative-knowledge account also has company outside auctions: Cosentino et al. (2026) find LLM negotiators infer a counterpart's preferences yet fail to convert that into bargaining advantage, and Samiei et al. (2025) find LLMs given a fully-specified rule set still fail to apply it faithfully as complexity grows: the stated-rule/applied-behavior gap looks like a broader pattern, not an auction-specific quirk.

Testing all four models together surfaces a second, independent finding: naively pooling results in our primary truthfulness test completely erases the effect, not because it vanishes but because two large, individually significant, opposite-signed per-model effects (Qwen's underbidding against Llama's outlier-driven overbidding) very nearly cancel on averaging: an aggregation bias in the same family as Simpson's paradox (Simpson, 1951; Blyth, 1972), with model identity the hidden grouping variable. Kempton et al. (2026) document the same pattern elsewhere in ML evaluation, where averaging token-level detector scores across distinct regions suppresses a real per-region signal. Aggregate significance tests in LLM-agent benchmarks can silently erase real, opposite-signed, model-specific effects, so per-model results should be reported alongside any pooled test.

Having identified both capability-split failures, we also ask whether the weaker models' behavior is correctable without retraining. We introduce Regret-Conditioned Prompting (RCP): after each round, the model is shown its realized bid, the outcome, and its counterfactual regret against the theoretical target, no weight updates, applicable to any model a deployer can only query. Shah et al. (2025) show a static, one-shot prompt reframing can improve LLM bidding alignment; our contribution is a repeated-round feedback loop tested against a theorem-reminder control that simply restates the optimal rule each round. Park et al. (2025) use "regret" as a repeated-game LLM lens via a regret-based *training loss* (ours is inference-time and weight-free); Song et al. (2026) show LLMs improve from generic scalar reward in non-strategic tasks (ours ties a mechanism-specific counterfactual-regret quantity to a closed-form target). M1 and M2 have different bid-ratio targets (1.0 and $(n-1)/n$), so we fit separately per mechanism (§4.3), with an expanded sample for our weakest model (Llama, $R = 10$): the theorem reminder, which simply supplies the correct rule, removes Llama's first-price overbidding almost immediately, while regret-conditioned feedback improves faster than the theorem-reminder control from a far worse start, surviving Holm-Bonferroni correction under model-based errors ($p = 0.0016$, adjusted $p = 0.0079$) but only at the edge of sig-

nificance under a cluster-robust check ($p = 0.053$); despite the faster slope, RCP ends no closer to the target than no feedback at all, so its apparent improvement is better read as fewer large overbids accumulating over rounds than as genuine convergence. A symmetric follow-up on Qwen2.5-7B finds no comparable effect on either mechanism (all $p \geq 0.26$).

Our study is scoped to single-item, independent-private-value sealed-bid auctions under two mechanisms; generalization to common-value, multi-item, or combinatorial settings is untested. Our two locally-hosted models are also both quantized to 4 bits, confounding model family/scale with hosting configuration relative to our two hosted-API models; we return to this, and to why we do not attribute either failure pattern to scale or weight status specifically, in §6.

2 METHODS

2.1 ENVIRONMENT AND MECHANISMS

We study single-item, independent-private-value (IPV) sealed-bid auctions under two mechanisms:

- **M1 (second-price/Vickrey):** the dominant strategy is truthful bidding, $b_i = v_i$, independent of beliefs about other bidders.
- **M2 (first-price):** under valuations drawn i.i.d. Uniform$[0, 100]$, the symmetric BNE bid is $b_i^* = \frac{n-1}{n} v_i$, re-derived via symbolic first-order-condition maximization (`sympy`), matching `core/mechanisms.py::theoretical_bid()` exactly.

Group size $n \in \{2, 5\}$. Valuations $v_i \sim \text{Uniform}[0, 100]$ are drawn fresh per instance. **Architecture:** each auction instance involves exactly one real LLM call; the remaining $n-1$ bidders are synthetic agents that play the closed-form theoretical/BNE strategy exactly, not other LLM instances, making the per-trial deviation-gain construct (Hao et al., 2026) $\Delta_i(\pi) = \text{payoff}(\text{theoretical bid}) - \text{payoff}(\text{actual bid})$ (against realized competing bids) exactly computable rather than merely estimable. This is a scope-defining choice: this paper audits a focal LLM agent's individual incentive-compatible behavior against a known-rational field, not emergent multi-agent population dynamics (no collusion, contagion, or opponent-mismodeling effects are in scope). This same $\Delta_i(\pi)$ quantity is what the regret-conditioned feedback arm reports back to the model each round in the RCP study (§2.4).

2.2 ELICITATION PROTOCOLS

- **Elicitation A (direct bid):** the model bids directly, given its valuation and the mechanism rules. A chain-of-thought flag is logged per trial, but every trial reported in this paper was run with chain of thought off; the ablation that would have varied it was never run.
- **Elicitation B (theory-then-bid):** a two-part protocol, each in a fresh context with no shared history. Part 1 asks only for the general payoff-maximizing rule (no valuation, no bid). Part 2 asks for an actual bid using the same valuation/mechanism framing as Elicitation A, deliberately content-identical to A since H2 tests whether a model's Part-1 statement predicts its ordinary (Elicitation-A-style) bidding, not whether two bid-elicitation wordings differ. Part 1 is elicited once per (model, mechanism, n) condition, so H2's pairing is at the *condition* level: one Part-1 grade is compared against every Part-2 draw in that condition, not paired within a shared context.

The same 30 valuation draws are reused across models, elicitations, and temperature samples within each (mechanism, n) cell. Because Elicitation B Part 2 is content-identical to Elicitation A, the two are not independent draws at temperature 0: they return the identical bid on 93–100% of shared valuation draws (per model). H1's designated $N = 120$-per-model test treats every row independently; §4.1 reports a draw-collapsed sensitivity check averaging each pair to one observation per draw ($N = 60$), detailed in Appendix A.4.

2.3 MODEL ROSTER

Four models are tested. **Gemini 3.5 Flash-Lite** (`gemini-flash`): Google AI Studio, hosted, free tier; pin moved twice mid-project (§6), disclosed. **gpt-oss-120b**: Groq, hosted, free tier;

replaced an earlier pin, `llama-3.3-70b-versatile`, after Groq shut it down. **Llama 3.1 8B** (`llama3.1-8b`): Ollama, local, default `llama3.1:8b` tag (Q4_K_M quantized, confirmed via local manifest digest). **Qwen2.5-7B** (`qwen2.5-7b`): Ollama, local, default `qwen2.5:7b-instruct` tag (also Q4_K_M, confirmed via Ollama registry metadata and digest match); joined the roster after the other three models' one-shot data were already collected (§2.5), and was later given its own targeted RCP arm (§4.3).

2.4 Experimental grid

One-shot grid: 2 mechanisms × 2 group sizes × 2 elicitations × 4 models = 32 cells, each drawing 30 unique valuations, each valuation sampled 4 times under a temperature protocol (1 call at temperature 0, 3 calls at temperature 0.7): **3,840 total LLM bid calls**, the confirmatory grid this paper's primary tests are computed on (temperature-0 the primary sample for H1 and H3, temperature-0.7 a secondary stability check). A third, larger group size, $n = 10$, was added subsequently for all four models as a secondary condition; §4.1 reports it alongside $n \in \{2, 5\}$ for H1/H3 (Appendix A.10), though it does not enter the Holm-Bonferroni-corrected primary family above. **H4's descriptive comparison pools every group size and temperature** (720 rows/model), unlike the Holm-corrected family (M1/temperature=0 throughout). **H2's designated GEE also pools across every group size and temperature** rather than $n \in \{2, 5\}$/temperature=0 alone: 2,816 behavioral trials, after excluding 64 of 2,880 elicitation-B trials below the $1 floor, plus one declarative row per condition (24 conditions), for 2,840 rows in total; a behavioral bid is graded "correct" within 0.05 of the target ratio. §4.2 reports the GEE restricted to $n \in \{2, 5\}$ and to temperature=0 as scope-sensitivity checks.

RCP (repeated-round) study: 2 models, selected by an "extremes" criterion (single worst/best by mean-absolute-ratio-error on the one-shot grid, computed while only three models, Gemini 3.5 Flash-Lite, gpt-oss-120b, and Llama 3.1 8B, were in the roster), Llama 3.1 8B (worst) and gpt-oss-120b (best), at fixed $n = 5$, × 2 mechanisms × 4 feedback arms × 6 replicate trajectories × 20 rounds = 960 rounds each, temperature 0. Llama's replicate count was later extended from 6 to 10 after an interim fit (Appendix A.2); gpt-oss-120b remains at 6. For 6 of Llama's 10 replicates, the `regret_conditioned/theorem_reminder` arms were further extended to 40 rounds (exploratory), so Llama contributes 2,080 of the 4,960 rounds below, not 1,600; the confirmatory fit (§4.3) truncates every replicate to 20 rounds. A second near-ceiling model, Gemini 3.5 Flash-Lite, was added as a generalization check, also at 960 rounds. **Qwen2.5-7B**, not part of the extremes selection since it had not yet joined the roster when that selection was made, was later given its own targeted RCP arm against gpt-oss-120b (6 replicates) to test whether RCP helps an *underbidding* model too, a follow-up rather than a design change. Across all four models this totals 4,960 RCP rounds.

Feedback arms (RCP study only): None (no feedback beyond a fresh valuation draw); **outcome-only** ("You won/lost. Your payoff was $X."); **theorem-reminder** (control: restates the optimal-bid *rule* each round, no quantified counterfactual; isolates rule-reminding from regret quantification); **regret-conditioned (RCP)** (shows the realized bid, outcome, and counterfactual regret $\Delta_i(\pi)$ against the theoretical target, no weight updates).

2.5 Hypotheses, primary outcomes, and analysis history

Five hypotheses were examined: H1 (truthfulness), H2 (knowledge-behavior dissociation), H3 (QRE model fit), H-RCP (regret-conditioned convergence), H4 (descriptive model comparison). One primary dependent variable and test was designated per hypothesis before the full grid was run, as finally analyzed: H1, mean signed dollar deviation vs. 0, M1/temperature=0, per model; H2, declarative accuracy vs. behavioral bid-ratio deviation, pooled GEE; H3, QRE precision λ vs. classical/uniform-random nulls, M1, Fisher-combined LRT; H-RCP, round×arm slope difference, M2 only; H4, mean bid-ratio deviation by model, M1, pooled, descriptive only (§3–§4 give full statistical detail per test).

Analysis history. The design document, written before the full grid was run, fixed one primary test per hypothesis against a three-model roster; Qwen2.5-7B joined afterward, once the other three models' one-shot data had already been collected (§2.4). Three tests were revised once data were seen, disclosed here rather than presented as originally planned. **H1**'s original three-model test rejected

in the *opposite* direction to the predicted underbidding, driven by Llama's low-valuation overbids (Table 1's robustness row); Qwen was then added to the confirmatory roster, growing the Holm family from four tested items to seven. **H2**'s originally paired design was replaced by the pooled GEE reported here, since each condition contributes one declarative grade against many bids (§2.2). **H-RCP**'s originally mechanism-pooled contrast was re-specified as the M2-only, continuous-round test below, and Llama's replicates were later extended from 6 to 10 after an interim fit. No primary test was re-designated after being seen to fail. Appendix A.2 gives the full account; §6 returns to the inflation this kind of post hoc revision creates.

H1's primary DV was changed, before the full grid was run, from a raw bid-to-valuation ratio to a signed dollar deviation after an early pilot: the raw ratio produced 43–65$\times$ blowups on low-valuation draws, cutting achieved power to $\sim$5% at the planned N; dollar-scale deviation achieves $\sim$96% power on the same pilot variance and N and needs no valuation floor, unlike the ratio-based metrics used elsewhere (H2, H3, H-RCP).

How H1 is reported, given four models with opposite-signed effects. A naive pooled t-test across all four models ($N = 480$) gives mean deviation $-\$0.14$, $p = 0.847$: not significant, and, read in isolation, wrongly suggestive of "no truthfulness gap." Two of the four models actually show large, opposite-signed deviations (Llama overbids \$6.88, Qwen underbids \$7.42) that nearly cancel when averaged, so **H1's primary evidence is the set of four per-model tests, not either pooled test**, both reported only to demonstrate why pooling is the wrong move here: the same masking pattern discussed in §1 (Simpson's-paradox framing). Holm-Bonferroni correction is applied across a single confirmatory family of seven items: the four per-model H1 comparisons together with H2, H3, and H-RCP (§3).

Every other DV, condition slice, temperature-0.7 estimate, and the M2 analogue of each primary test is reported as exploratory throughout.

2.6 Statistical procedures

H1: one-sample t-tests of mean signed deviation vs. 0, per model, M1/temperature=0 ($N = 120$ per model), with three sensitivity checks (Appendix A.4). **H2**: a pooled GEE test (exchangeable correlation, clustered on condition), primary; it compares the overall rate of correct stated rules against the overall rate of correct bids, not whether a condition's stated rule predicts that condition's own bids, which per-model mismatch proportions describe instead (Appendix A.9). **H3**: QRE precision λ fitted per (model, n) cell at temperature=0 (primary) and 0.7 (secondary), vs. classical/uniform-random nulls via a likelihood-ratio test, Fisher-combined across primary cells, with AIC (Appendix A.5) and a fixed-point refit as robustness checks. **H-RCP**: a mixed-effects model fit *separately per mechanism* (pooling conflates two different targets, §4.3), 20-round window, \$3 valuation floor; since the between-replicate random-intercept variance is estimated at zero, the fixed-effects (OLS) fit is reported, jointly across the worst/best contrast (Llama, gpt-oss-120b) with cluster-robust checks. Full model specifications are in Appendix A.1; all statistical code lives in `analysis/`, and every number in §4 is from a live run against `outputs/trials.db`.

3 Results overview

In brief, two of four models show a real, individually significant truthfulness deviation, in opposite directions, each traceable to a specific mechanism (§4.1–§4.2); both pooled H1 rows are shown only to illustrate why pooling misleads here, not as primary evidence (§2.5). H3 rejects overwhelmingly, reflecting the same split, and Qwen's underbidding and H3's rejections are robust to every sensitivity check we ran. H-RCP also rejects once fit on the mechanism it can speak to (M2) rather than pooled (§4.3), though not under every specification (Cons. Holm $p = 0.162$). H2 does not reject, but is underpowered in a specific, disclosed way (§4.2), not simply null. The Cons. Holm p column shows that Llama's H1 result and the H-RCP contrast would not survive Holm correction under their most conservative alternative specification; we did not re-designate either test on this basis, but treat both as suggestive rather than confirmed (§2.5).

Table 1: Overview of hypothesis tests, as finally analyzed (§2.5). Both pooled H1 rows are robustness/masking illustrations, not primary evidence (§1, Appendix A.2); H4 is omitted (descriptive only). "Cons. Holm p" is the largest Holm-adjusted p across the designated family and each sensitivity-check family reported in §4: a robustness bound, not a re-designated test.

Hypothesis	Test	Raw p	Holm-adj. p	Cons. Holm p	Reject?
H1: Llama 3.1 8B (overbidding)	per-model t-test	0.00798	0.0319	0.180	**Yes**
H1: Qwen2.5-7B (underbidding)	per-model t-test	1.48×10^{-16}	8.89×10^{-16}	3.84×10^{-8}	**Yes**
H1: Gemini 3.5 Flash-Lite	per-model t-test	0.156	0.312	0.638	No
H1: gpt-oss-120b	per-model t-test	0.156	0.312	0.638	No
H1 robustness: 3-model pooled*	pooled t-test	0.00836	N/A	N/A	**Yes**
H1: naive 4-model pooled	pooled t-test	0.847	N/A	N/A	No
H2 (knowledge-behavior dissociation, pooled)	GEE	0.0289	0.0867	0.220	No
H3 (QRE fit, M1/temp0, Fisher-combined, 8 cells)	LRT/Fisher	$< 10^{-300}$	$< 10^{-300}$	$< 10^{-300}$	**Yes**
H-RCP (regret-conditioned vs. theorem-reminder, M2, Llama)	OLS contrast	0.00158	**0.00791**	0.162	**Yes**

*H1's originally specified test (Appendix A.2): $+\$2.29$, $p = 0.008$, opposite in sign to the prediction.

4 RESULTS

4.1 H1/H3: A CLEAN CAPABILITY SPLIT, NOT A UNIFORM BIAS

Per-model H1 results (M1, temperature=0, $N = 120$ per model):

Table 2: Per-model H1 results. Draw-collapsed, Wilcoxon, and drop-extreme-trial sensitivity checks on each row are reported in Appendix A.4.

Model	Mean dev.	Median dev.	t (p), designated
Gemini/gpt-oss-120b (identical)	$-\$0.0004$	$-\$0.0007$	-1.427 (0.156)
Llama 3.1 8B	**+$6.88**	+$0.0002	2.698 (**0.00798**)
Qwen2.5-7B	**−$7.42**	$-\$5.96$	-9.62 (1.48×10^{-16})

Gemini 3.5 Flash-Lite and gpt-oss-120b achieve near-exact truthful bidding: on every one of their 120 primary trials, $|\text{bid} - \text{valuation}| < \$0.005$, both simply computing $\text{bid} = \text{round}(\text{valuation}, 2)$ at deterministic decoding, with numerically *identical* signed deviations across every shared draw. **Llama 3.1 8B's mean overbid is driven by a small number of extreme overbids on very-low-valuation draws** (its single largest: a verified $100.00 bid against a $2.31 valuation, which the model itself justified as bidding "the maximum possible value") rather than a uniformly-shifted distribution: its median deviation is $\approx \$0$, and its designated test's significance does not survive a draw-collapsed ($p = 0.060$) or Wilcoxon ($p = 0.42$) check (Appendix A.4); only dropping the single most extreme trial restores it ($p = 0.014$, $N = 119$). We describe its M1 behavior as *noisy with occasional large overbids*, not a uniform tendency. **Qwen2.5-7B underbids** by a comparable magnitude in the opposite direction and, unlike Llama, remains significant under every sensitivity check we ran; it is the only model showing systematic underbidding at all.

H4 (descriptive only) reproduces the same split via a robust ranking cross-check (Appendix A.6).

The same split continues at the secondary $n = 10$ condition, with one qualitative change to Llama's H1 result; full numbers in Appendix A.10.

H3 (M1, temperature=0, all four models, 8 primary cells): both nulls are rejected far beyond double-precision representability ($p < 10^{-300}$ for each; $\log_{10} p \approx -1651$ vs. the classical/deterministic null, $\log_{10} p \approx -577$ vs. the uniform-random null, the latter the conservative primary figure in Table 1). Rejecting the classical point-mass null is close to guaranteed for any model with any noise at all, so the informative comparisons are per cell (Table 5; full detail in Appendix A.5). Each cell's $N = 60$ carries the same A/B pseudoreplication as H1 (§2.1), overstating every likelihood ratio by up to $\sim 2\times$; no verdict changes, since the smallest rejecting cell (Llama, $n = 2$) has $\text{LR}_{\text{vs. uniform}} = 21.9$, decisive at half that value. Gemini and gpt-oss-120b's likelihoods are saturated, so their reported λ values are bounds, not interior estimates.

Gemini and gpt-oss-120b bid so close to deterministic truth-telling that QRE's extra free parameter buys nothing over the classical prediction (by AIC) at $n = 2$, and only marginally helps at

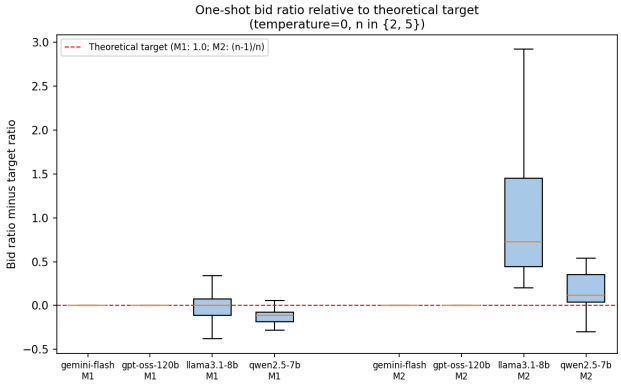

Figure 1: One-shot bid ratio minus its theoretical target ratio (0 = on target: 1.0 in M1, $(n-1)/n$ in M2), by model and mechanism, temperature=0, valuation $\geq \$1$, $n \in \{2, 5\}$ (H1/H3's primary subset; $n = 10$ reported separately above). Median/IQR boxes, not the raw mean, avoid the low-valuation ratio blowup noted in §2.5. Plotting each trial relative to its own mechanism- and n-specific target, rather than the raw ratio, keeps 0 the target for every box: Gemini/gpt-oss-120b sit tightly on target in both mechanisms; Llama is centered on target in M1 but far above it in M2; Qwen sits below target in M1 and above it in M2.

$n = 5$. **Llama 3.1 8B's $n = 5$ bidding is statistically indistinguishable from uniform-random bidding**: a 1,000-resample bootstrap shows the fitted QRE beats the uniform null in only 42.4% of resamples, a coin-flip rate (a fixed-point refit confirms Llama's estimates are stable across bid-grid resolutions, though that check is itself unreliable near the near-deterministic models' point-mass bidding). **Qwen2.5-7B, unlike Llama, decisively beats both nulls at both sizes** ($\lambda = 0.396$ at $n = 2$, $\lambda = 1.686$ at $n = 5$): stable, structured underbidding, not a degradation toward noise, the opposite of Llama's trajectory as group size grows (continues at $n = 10$, Appendix A.10). Following Pechon-Elkins & Chun (2026)'s caution that fitted λ is a payoff- and model-dependent summary rather than a structural rationality parameter, we read these values descriptively: Qwen's deviations are concentrated and consistent, Llama's are diffuse, and neither reading substitutes for the closed-form deviation numbers above.

Read together, H1 and H3 tell the same story: a clean split, not a uniform bias, that continues but moves unevenly at the secondary $n = 10$ condition (Appendix A.10).

4.2 H2: A CROSSED PATTERN, AND WHERE BEHAVIOR DEPARTS FROM WHAT IS STATED

Each condition's Part-1 statement (24 in all, one per model×mechanism×group-size cell, including $n = 10$) was graded against a written rubric: M1 credit requires stating that bidding one's valuation is optimal, and M2 credit requires only the qualitative direction (shade below valuation), not the exact BNE formula. Primary grades were assigned by the same AI-coding-agent workflow used throughout this project (§7), re-graded blind by a second AI pass: 23/24 agreement, $\kappa = 0.895$; a blind human volunteer rater cross-checked all 24 and agreed 23/24, $\kappa = 0.895$, disagreeing on the same item, Qwen's M2 at $n = 2$. All three raters' per-item grades are released (`h2_declarative_grades.csv`); the full per-model, per-mechanism table is reported in Appendix A.9 (Table 7).

A crossed pattern of mechanism-invariant rules, but an asymmetric one. The compliant models state a different, correct rule for each mechanism; each smaller model instead states one rule for both, so each is graded correct on one mechanism and wrong on the other, on opposite sides, at every group size. Where its rule is wrong, each errs in the direction that rule predicts: Qwen underbids in M1 (312/352 incorrect), and Llama overbids in M2 (346/352 incorrect), by more than even its own "bid your valuation" rule would predict (Figure 1). This crossing depends on a lenient M2 rubric, credited on direction alone: under a rubric requiring the exact BNE magnitude, Qwen would score 0/3 on M2 and there would be no crossing at all. Where its rule is right, Llama is behaviorally incorrect in 190/352 M1 trials, a genuine dissociation between a correct stated rule and noisy execution. Qwen's

264/352 incorrect M2 bids are not a comparable dissociation: its stated "80–90%" heuristic already overshoots the true Bayes-Nash fraction, and its M2 bids simply sit above target in the direction that heuristic predicts (Figure 1) – consistent execution of an imprecise rule, not a gap between knowing and doing.

Pooled test (GEE, all four models, every group size and temperature, §2.4), H2's confirmatory evidence: $p = 0.029$, not significant after Holm-Bonferroni correction (adjusted $p = 0.087$; Cons. Holm $p = 0.220$, Table 1), diluted by Gemini/gpt-oss-120b's ceiling effect. **This null is underpowered, not evidence of no dissociation** (roughly 22% achieved power; design-effect detail in Appendix A.9). Restricting to $n \in \{2, 5\}$ gives $p = 0.073$; restricting further to temperature=0 only, $p = 0.080$: both *further* from significance, not closer, so narrowing the scope does not rescue the test. We read the per-model pattern above, and in Appendix A.9, as descriptive, not as independently confirming a dissociation.

Per-model mismatch rates (descriptive only) corroborate this reading without serving as independent evidence, since each condition's single Part-1 grade is compared against every Part-2 behavioral draw, giving the declarative side of each table an effective sample size of 1 per condition, not the trial count: a McNemar's-style χ^2 would be invalid here and is not reported. Full per-model, per-mechanism proportions are in Appendix A.9.

4.3 H-RCP: A PARTIAL, SPECIFICATION-SENSITIVE CORRECTION

M1 and M2 have different theoretical bid-ratio targets (1.0 and $(n-1)/n = 0.8$ at $n = 5$), so a round×arm interaction fit on raw, mechanism-pooled `bid_ratio` cannot distinguish "moved toward 1.0" from genuine correction: the same masking pattern as §2.5's pooled H1 test, here from mechanism rather than model identity. A naive mechanism-pooled fit (Llama vs. gpt-oss-120b, $R = 10$) gives contrast $= -0.302$, $p = 0.041$ (naive OLS) / $p = 0.182$ (cluster-robust), reported only to show why it is the wrong read.

Fit separately per mechanism, the real finding is M2-specific and more modest than it first appears. On M1 (this study's $\geq$ \$3 floor), Llama's `none`/`theorem_reminder` bid ratio already sits within about 4% of target (§4.1's larger +\$6.88 M1 result is a different, unfloored dataset dominated by the same low-valuation outliers this floor excludes, not a contradiction), so no differential arm effect is found (contrast $= -0.0039$, $p = 0.661$). On M2, where Llama systematically overbids, the two feedback arms diverge:

Table 3: H-RCP: primary contrast (M2, regret-conditioned vs. theorem-reminder arm, worst-performing model Llama 3.1 8B; joint = fit with gpt-oss-120b's data, $R = 10/R = 6$, designated; Holm-Bonferroni-adjusted $p = 0.00791$, Table 1). The estimate is identical joint vs. Llama-only by construction; only the SEs differ, since gpt-oss-120b's bid ratio is far less variable than Llama's (SD 0.00016 vs. 0.274).

	Estimate	SE	z	p
Joint, model-based SE (designated)	-0.0115	0.00364	-3.159	**0.00158**
Joint, cluster-robust SE (64 trajectories)	-0.0115	0.00595	-1.934	0.0531
Llama-only, model-based SE	-0.0115	0.00459	-2.504	0.0123
Llama-only, cluster-robust SE (40 trajectories)	-0.0115	0.00597	-1.928	0.0539

The contrast is significant under model-based SEs, but not under either cluster-robust check ($p = 0.053$–0.054; Cons. Holm $p = 0.162$), so we treat it as suggestive rather than confirmed. In absolute terms (mean absolute deviation from target, rounds 1–5→16–20, valuation $\geq$ \$3): `theorem_reminder` moves $0.035 \rightarrow 0.0004$; `none` stays at $0.198 \rightarrow 0.200$; `regret_conditioned` moves $0.354 \rightarrow 0.226$. Regret-conditioned feedback does not end up ahead: it starts *worse* than no feedback (mean bid ratio 1.12 vs. 1.00) and, after 20 rounds, remains further from target than the no-feedback arm, whose median (0.200) matches regret-conditioned's median in both windows; only 4 of its 10 trajectories end below the no-feedback arm's mean. Its steeper slope largely reflects fewer large overbids accumulating over rounds, not genuine convergence toward the 0.8 target. The original design's prediction that RCP would converge faster *and more completely* than every other arm is thus not supported in levels: theorem-reminder, which simply supplies the correct rule, removes Llama's overbidding almost immediately and remains the more effective intervention throughout. The $\geq$ \$3 floor was set after the extended-horizon collec-

tion reproduced H1's near-zero-valuation fragility at larger scale. Figure 2 (Appendix A.8) plots the round-by-round trajectories.

Ceiling-effect generalization check (exploratory): Gemini 3.5 Flash-Lite shows the same flat round-trend as gpt-oss-120b (all $p \geq 0.48$; slopes indistinguishable, $p \geq 0.61$). **A targeted follow-up on Qwen2.5-7B** (does RCP help the *underbidding* model too?) finds a clean null on both mechanisms (all $p \geq 0.26$): RCP shows no effect on the underbidding model or mechanism; full numbers in Appendix A.8.

Three further exploratory analyses (revenue equivalence/bid regression, a Beta(2,2) pattern-matching probe, and a cognitive-hierarchy fit alongside H3's QRE) are reported in Appendix A to keep the main text focused on the confirmatory hypotheses.

5 Conclusion

Across four LLMs tested against a closed-form game-theoretic ground truth, incentive-compatible bidding is neither a uniform capability nor a uniform failure. Two models bid exactly as theory prescribes in both mechanisms. The two smaller models each state one rule for both mechanisms instead, and where that rule is wrong each errs in the direction the rule predicts, our study's central result, though the crossed pattern rests partly on a lenient rubric, and Llama additionally errs even where its own rule is correct. Stating a rule and applying it under a specific mechanism are separable competencies; testing one does not certify the other. Naively pooling all four models erased a robust, opposite-signed per-model effect in our truthfulness test, so per-model results should accompany any pooled test in an LLM-agent benchmark. Restating the correct rule each round largely removed our weakest model's first-price overbidding on its own, suggesting a fixable, deployment-time behavioral gap; regret-conditioned feedback, by contrast, did not end up closer to target than no feedback at all in levels, and its faster improvement over the no-feedback baseline is not robust to cluster-robust inference. The open question is no longer whether these models can reason strategically, but which failure mode each carries, how it is best diagnosed, and how cheaply, if at all, it closes.

6 Limitations

Single-environment scope (§1): single-item, IPV sealed-bid auctions under two mechanisms. **Single-real-agent architecture**: each auction is one LLM call against synthetic, closed-form-optimal opponents (multi-LLM populations are future work). **Quantization/hosting confound**: Llama 3.1 8B and Qwen2.5-7B run locally at 4-bit (Q4_K_M); gpt-oss-120b and Gemini 3.5 Flash-Lite are hosted APIs, so the real confound is small-quantized vs. large-API-hosted, not open- vs. closed-weight (gpt-oss-120b is itself Apache-2.0 open-weight; only Gemini is closed; Li et al. 2025: quantization alone can cut reasoning accuracy 32%). **RCP's extremes-only selection**: a targeted Qwen2.5-7B follow-up (Appendix A.8) finds none. **API non-stationarity** (§2.3): Gemini's pin change affected 18 of 1,440 trials, negligible via `trial_id` hashing. **Post hoc choices** (Appendix A.2) inflate the false-positive rate; Table 1's adjusted p-values are suggestive, not confirmatory.

7 Reproducibility statement

Grid parameters are config-driven, every model pinned to a dated identifier with `model_version` logged per trial, and the harness is resumable; the pipeline is unit-tested (261 tests; six contamination/scope bugs found, fixed, and regression-tested, Appendix A.3). Anonymized code, data, and design document are in the supplement.

AI-use disclosure: per ICLR 2027's AI-use disclosure policy, an AI-coding-agent-driven workflow helped with this project's harness-building, grid execution, and statistical analysis; with literature retrieval and discovery of related work; with research ideation on analysis approaches, robustness checks, and bug diagnosis; and with drafting and polishing this paper's text. Author contributions were the core experimental design, hypothesis formulation, and final interpretation and verification of all reported results.

REFERENCES

Colin R. Blyth. On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, 67(338):364–366, 1972.

Colin F. Camerer, Teck-Hua Ho, and Juin-Kuan Chong. A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, 119(3):861–898, 2004.

David J. Cooper and Hanming Fang. Understanding overbidding in second price auctions: An experimental study. *The Economic Journal*, 118(532):1572–1595, 2008.

Romain Cosentino, Sarath Shekkizhar, Adam Earle, and Silvio Savarese. Counterparty modeling is not strategy: The limits of LLM negotiators. *arXiv preprint arXiv:2605.16575*, 2026.

Jianing Hao, Han Ding, Yuanjian Xu, Tianze Sun, Ran Chen, Wanbo Zhang, Guang Zhang, and Siguang Li. Game-theoretic lens on LLM-based multi-agent systems. *arXiv preprint arXiv:2601.15047*, 2026.

John H. Kagel and Dan Levin. Auctions: A survey of experimental research. In John H. Kagel and Alvin E. Roth (eds.), *The Handbook of Experimental Economics, Volume 2*. Princeton University Press, Princeton, NJ, 2016.

Tom Kempton, Viktor Drobnyi, Maeve Madigan, and Stuart Burrell. Log-likelihood, Simpson's paradox, and the detection of machine-generated text. *arXiv preprint arXiv:2605.06294*, 2026.

Zhen Li, Yupeng Su, Runming Yang, Congkai Xie, Zheng Wang, Zhongwei Xie, Ngai Wong, and Hongxia Yang. Quantization meets reasoning: Exploring LLM low-bit quantization degradation for mathematical reasoning. *arXiv preprint arXiv:2501.03035*, 2025.

Ismail Lotfi and Ali Ghrayeb. Strategic bidding in 6G spectrum auctions with large language models. *arXiv preprint arXiv:2604.24156*, 2026.

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

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

Chanwoo Park, Xiangyu Liu, Asuman Ozdaglar, and Kaiqing Zhang. Do LLM agents have regret? A case study in online learning and games. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv preprint arXiv:2403.16843.

Mateo Pechon-Elkins and Jon Chun. Quantal response equilibrium as a measure of strategic sophistication: Theory and validation for LLM evaluation. *arXiv preprint arXiv:2603.10029*, 2026.

John G. Riley and William F. Samuelson. Optimal auctions. *The American Economic Review*, 71 (3):381–392, 1981.

Mahdi Samiei, Mahdi Mansouri, and Mahdieh Soleymani Baghshah. The illusion of procedural reasoning: Measuring long-horizon FSM execution in LLMs. *arXiv preprint arXiv:2511.14777*, 2025.

Anand Shah, Kehang Zhu, Yanchen Jiang, Jeffrey G. Wang, Arif K. Dayi, John J. Horton, and David C. Parkes. Learning from synthetic labs: Language models as auction participants. *arXiv preprint arXiv:2507.09083*, 2025.

Edward H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951.

Kefan Song, Amir Moeini, Peng Wang, Lei Gong, Rohan Chandra, Yanjun Qi, and Shangtong Zhang. Reward is enough: LLMs are in-context reinforcement learners. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv preprint arXiv:2506.06303.

Haoran Sun, Yusen Wu, Peng Wang, Wei Chen, Yukun Cheng, Xiaotie Deng, and Xu Chu. Game theory meets large language models: A systematic survey with taxonomy and new frontiers. *arXiv preprint arXiv:2502.09053*, 2025.

William Vickrey. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of Finance*, 16(1):8–37, 1961.

Binchi Zhang and Atrisha Sarkar. Level-k distinguishable mechanisms for evaluating bounded rationality in LLMs. *arXiv preprint arXiv:2608.21296*, 2026.

Kehan Zheng, Jinfeng Zhou, and Hongning Wang. Beyond Nash equilibrium: Bounded rationality of LLMs and humans in strategic decision-making. *arXiv preprint arXiv:2506.09390*, 2025.

Quanyan Zhu. Reasoning and behavioral equilibria in LLM-Nash games: From mindsets to actions. *arXiv preprint arXiv:2507.08208*, 2025.

A Additional analyses

The first four subsections below (statistical procedure details, the analysis history, H1's sensitivity checks, H3's per-cell fits) are confirmatory detail deferred here only for space; every subsection after them is exploratory (outside the Holm-Bonferroni-corrected confirmatory family, §3), reported here to keep the main results focused on the confirmatory hypotheses.

A.1 Statistical procedure details

Full model specifications for §2.6. **H1**: $N = 120$ per model is 30 draws $\times$ 2 group sizes $\times$ 2 elicitations; because Elicitation A and B Part 2 are largely redundant at temperature 0 (§2.1), the three sensitivity checks (Appendix A.4) are dropping the single most extreme trial, a draw-collapsed test with $N = 60$, and a Wilcoxon signed-rank test. **H2**: the GEE tests the `correct` indicator on a declarative-row flag (overall rate of correct stated rules, 18/24, vs. overall rate of correct bids). Declarative-accuracy grading (24 statements, inter-rater detail in §4.2) and behavioral-accuracy scoring use the same valuation$<$ \$1 ratio-blowup floor as H3's QRE fit, excluding 64 of 2,880 elicitation-B trials (2.2%). **H3**: λ is fitted on 60 observations per cell (30 draws $\times$ 2 elicitations) at temperature=0; each null (classical, uniform-random) is a boundary case of the same 1-parameter family, so the χ^2_1 likelihood-ratio reference is conservative. **H-RCP**: the model is bid_ratio $\sim$ round$_c$ $\times$ arm $\times$ model + (1 | replicate), round mean-centered at 10.5 and entered continuously, not binned. Since gpt-oss-120b's bid-ratio SD is roughly 1,700 times smaller than Llama's, a worst-model-only refit and cluster-robust (sandwich) SE checks accompany the joint fit, alongside the same contrast on Qwen vs. gpt-oss-120b as a follow-up.

A.2 Analysis history

The design document in the supplement fixed one primary test per hypothesis, and a three-model roster, before the full grid was run. The following changes were made after data were seen.

- **Roster and H1.** H1 was specified as a single one-sample test pooled over the original three models (Gemini 3.5 Flash-Lite, gpt-oss-120b, Llama 3.1 8B), predicting underbidding. Run exactly as specified, that test gives $+\$2.29$, $t(359) = 2.65$, $p = 0.008$ (draw-collapsed $p = 0.061$): significant, but in the opposite direction, driven by Llama's low-valuation overbids. Qwen2.5-7B was added afterward, once the other three models' one-shot data were already collected, and was first analyzed as exploratory; once its results were seen, it was added to the confirmatory roster, and the four per-model tests became H1's primary evidence, growing the Holm-Bonferroni family from four tested items to seven.
- **H2.** Specified as a paired test; because each condition contributes a single declarative grade (§2.2), the pooled GEE was used instead.
- **H-RCP.** Specified as a mechanism-pooled rounds 1–5 vs. 16–20 contrast. Because M1 and M2 have different targets, it was re-specified as M2-only with continuous round. Llama's replicates were extended from 6 to 10 after an $R = 6$ interim fit ($p = 0.084$), and the $\geq \$3$ valuation floor was set after the extended-horizon collection reproduced H1's low-valuation fragility.

- **Pre-grid DV change.** After an early pilot, H1's DV was changed from the bid/valuation ratio to signed dollar deviation (§2.5); this preceded the full grid and is not a post hoc change in the same sense as the three above.

No primary test was re-designated after being seen to fail; §6 discusses the effect this kind of post hoc revision has on the confirmatory family's error rate.

A.3 BUG AUDIT HISTORY

Six contamination/scope bugs were found across repeated audits of the statistical pipeline, each fixed and each now covered by a dedicated regression test, live-reverified against `outputs/trials.db`: an RCP round-window contamination bug, an RCP model-population contamination bug, an $n = 10$ scope bug, an RCP-figure mechanism-pooling bug, a mislabeled robustness-suite key, and a naive-vs. mixed-effects labeling bug. A seventh issue, a false-premise prompt bug in the Beta(2,2) probe, is documented separately in Appendix A.7.

A.4 H1 SENSITIVITY CHECKS, PER MODEL

Table 4 reports, for each model, the same three sensitivity checks discussed in §4.1: a draw-collapsed one-sample t-test ($N = 60$, averaging each A/B-identical pair to one observation per unique valuation draw), a Wilcoxon signed-rank test on the full $N = 120$ sample, and the designated test repeated with only that model's own single most extreme trial dropped ($N = 119$). None of these replaces §4.1's designated per-model test; they show how sensitive each result is to the A/B pseudoreplication and to a single extreme trial.

Table 4: H1 sensitivity checks, per model (M1, temperature=0).

Model	p, collapsed ($N = 60$)	p, Wilcoxon ($N = 120$)	p, drop-extreme ($N = 119$)
Gemini/gpt-oss-120b (identical)	0.319	0.091	0.197
Llama 3.1 8B	0.060	0.424	**0.014**
Qwen2.5-7B	6.39×10^{-9}	6.87×10^{-21}	8.01×10^{-18}

A.5 H3 PER-CELL QRE FIT DETAILS

Table 5: H3: per-cell QRE fit results (M1, temperature=0); "beats" is by AIC.

Model	n	λ (temp=0)	Beats classical?	Beats uniform?
Gemini/gpt-oss-120b (identical)	2	3.3×10^5 (saturated)	No (ties)	Yes
Gemini/gpt-oss-120b (identical)	5	$\geq 3.27 \times 10^6$ (bound)	Yes	Yes
Llama 3.1 8B	2	0.0846 (0.0848 fixed-point)	Yes (decisively)	Yes
Llama 3.1 8B	**5**	**0.0121 (0.020 fixed-point)**	Yes (decisively)	**No**
Qwen2.5-7B	2	0.396	Yes (decisively)	Yes
Qwen2.5-7B	5	1.686	Yes (decisively)	Yes

Gemini/gpt-oss-120b's fitted λ at $n = 5$ sits exactly at the optimizer's upper bound (e^{15}), and their $n = 2$ value comes from a saturated (perfectly-fit) likelihood; both are reported as bounds, not converged interior estimates. For these models the classical point prediction ties the QRE fit at $n = 2$ by AIC. The fixed-point refit confirms Llama's estimates are stable across bid-grid resolutions; for the near-deterministic models that same check is itself unreliable near point-mass bidding.

A.6 REVENUE EQUIVALENCE AND BID-ON-VALUATION REGRESSION

H4 (descriptive only, M1, pooled): ranking models by mean absolute bid-ratio deviation reproduces the same split as H1/H3; the raw mean is dominated by H1's disclosed low-valuation blowup artifact (§2.5), so we report the robust median: Llama 0.108, Qwen 0.111, Gemini/gpt-oss-120b $\approx 6 \times 10^{-5}$; no significance test, merely a cross-check.

This subsection pools all three group sizes ($n \in \{2, 5, 10\}$), both elicitations, and both temperatures (720 trials per model and mechanism). Revenue equivalence (Myerson, 1981) predicts equal

expected seller revenue under M1 and M2. Mean seller revenue does not differ detectably between mechanisms for Gemini/gpt-oss-120b (Welch $t = 0.086$, $p = 0.932$ for both, agreeing to three decimals though not bit-identical: 0.08566 vs. 0.08571). Llama 3.1 8B shows a massive, highly significant divergence (M1: \$61.97, M2: \$85.68, $p = 4.3 \times 10^{-73}$), corroborating its bid-level findings; Qwen2.5-7B shows a smaller but significant one (M1: \$57.98, M2: \$62.26, $p = 6.4 \times 10^{-4}$), consistent with its underbidding moving aggregate revenue despite more theory-consistent bids than Llama's. Regressing bid on valuation in M1 recovers the exact theoretical slope to 4–5 decimals ($r \approx 1.0$) for Gemini/gpt-oss-120b; Llama's is well below target with low fit (slope$= 0.356$, $r = 0.379$); Qwen's is closer with higher fit (slope$= 0.852$, $r = 0.976$).

A.7 Non-uniform valuations (Beta(2,2)): a pattern-matching probe

This slice (M2 only, 1,920 trials/collection) tests whether near-perfect Uniform-case compliance reflects genuine per-instance computation or pattern-matching a memorized linear rule, by redrawing valuations from a Beta(2,2) distribution whose BNE bid is not linear. All four models are reported, re-collected under a corrected prompt after a false-premise bug (the original prompt said Uniform while drawing from Beta(2,2); §7's bug history) and verified against the archived buggy-premise trials. No valuation floor applies here: the realized minimum draw across all 1,920 trials is \$8.71, comfortably above every floor used elsewhere.

Table 6: Beta(2,2) non-uniform-valuation slice, corrected data, all four models.

Model	Mean abs. ratio dev. (corrected)	Direction vs. old (buggy-premise) data
Gemini 3.5 Flash-Lite	0.0872	$\downarrow 1.1\times$ smaller (essentially unchanged)
gpt-oss-120b	0.0075	$\downarrow 11.8\times$ smaller (closer to BNE)
Llama 3.1 8B	0.6407	$\downarrow 2.5\times$ smaller (closer to BNE)
Qwen2.5-7B	0.7535	$\uparrow 5.4\times$ larger (further from BNE)

The four directions do not move together. gpt-oss-120b gets substantially closer to BNE-optimal ($11.8\times$) once told the true distribution, evidence *for* genuine per-instance computation; Gemini is essentially unchanged (opposite signature, though its 0.087 absolute deviation is modest enough to be suggestive only); Llama also gets closer but remains far from BNE (0.64), consistent with noise-dominated behavior; Qwen2.5-7B alone gets substantially *worse*, consistent with its near-Uniform-competent M2 performance relying partly on a memorized rule it cannot generalize: the one result that complicates rather than resolves the question.

Cognitive hierarchy (level-k). As a complementary lens to QRE, we fit a Camerer-Ho-Chong cognitive-hierarchy model (Camerer et al., 2004) (M2 only; strict level-k is provably degenerate under Uniform valuations, verified symbolically); this analysis is distinguished from the concurrent Zhang & Sarkar (2026) by our closed-form DSIC/BNE setting, with Zheng et al. (2025) adjacent evidence of LLM/human strategic divergence in a different game class. Gemini/gpt-oss-120b fit as near-zero-error level-1 reasoners; Llama's fitted τ/error are erratic across group sizes (mean abs. error 30–57), consistent with noise-dominated M2 behavior; Qwen's fitted τ rises with group size ($0.3 \to 1.7 \to 2.2$) while its error against BNE falls ($14.1 \to 8.0 \to 2.6$), opposite Llama, consistent with stable rather than noise-degrading underbidding.

A.8 Targeted RCP follow-up: does RCP help the underbidding model too?

Qwen2.5-7B's underbidding (§4.1) was untested under RCP by the extremes criterion alone (§2.4), since it had not yet joined the roster when that criterion was applied. We ran the identical arm on Qwen vs. gpt-oss-120b (6 replicates): a clean null on both mechanisms (M1: $p = 0.985$; M2: $p = 0.262$ naive, $p = 0.481$ cluster-robust). Descriptively, theorem-reminder's M2 deviation nearly vanishes ($0.025 \to 0.0002$, mirroring Llama) while regret-conditioned's gets slightly *worse* ($0.188 \to 0.243$): RCP shows no benefit here, and theorem-reminder does the real work on M2 for both models.

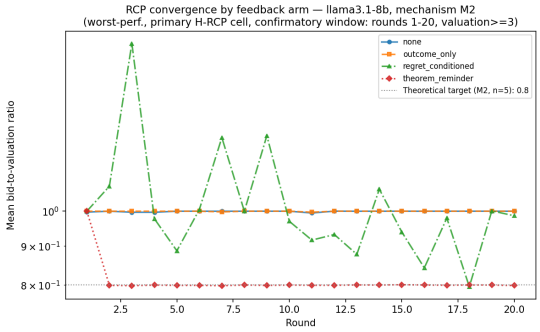

Figure 2: Mean bid-to-valuation ratio by round, Llama 3.1 8B, mechanism M2, one line per feedback arm, confirmatory window (rounds 1–20, valuation $\geq$ \$3); log-scaled y-axis. Dotted line: this cell's actual target, $(n-1)/n = 0.8$ at $n = 5$ (not M1's 1.0). `theorem_reminder` sits on the target; `regret_conditioned` moves down from above valuation, slowly and noisily; `none`/`outcome_only` stay at ≈ 1.0 (bidding the valuation), matching Table 3.

Table 7: H2: declarative accuracy by model and mechanism ($n \in \{2, 5, 10\}$), per the written rubric; † marks the one item the blind human rater scored differently.

Model	M1	M2	Stated rule
Gemini 3.5 Flash-Lite	3/3	3/3	M2: qualitative shading only, no formula (the second AI rater graded one item incorrect).
gpt-oss-120b	3/3	3/3	Gives the exact BNE bid formula every time (M2).
Llama 3.1 8B	3/3	**0/3**	Every M2 response states "bid your true valuation," the M1 rule carried over to first-price.
Qwen2.5-7B	**0/3**	3/3†	Every response, M1 *and* M2, recommends the same "shade below valuation" heuristic; credited for M2 on direction only.

A.9 H2 PER-MODEL MISMATCH RATES (DESCRIPTIVE ONLY)

The pooled GEE test (§4.2) is underpowered: 24 clusters (2,840 observations, cluster sizes 117–121), within-condition ICC ≈ 0.72, design effect $\approx 85\times$, giving roughly 22% achieved power at the observed effect size.

gpt-oss-120b shows zero behaviorally-incorrect Elicitation-B bids on either mechanism (ceiling effect, correct stated rule in both). Gemini is likewise on ceiling in M1 (0/352 incorrect) but has a small M2 mismatch rate (8/352), despite a correct stated rule in both mechanisms. Llama is wrong on **M1** in 190/352 behavioral trials despite correctly-stated theory (a genuine dissociation), and wrong on **M2** in 346/352 trials, matching its incorrect stated theory (consistent execution of a wrong rule, not a dissociation). Qwen shows the mirror pattern on **M1**: near-floor at 312/352 wrong, matching its incorrect stated M1 theory. Its **M2** mismatch rate (264/352) is not a comparable dissociation either: it is credited "correct" on M2 only under the lenient, direction-only rubric (§4.2), and its own stated "80–90%" heuristic already overshoots the true Bayes-Nash fraction, so these bids are consistent with, not a departure from, what it states.

A.10 THE SECONDARY $n = 10$ CONDITION (H1/H3)

$n = 10$ was added subsequently for all four models, outside the Holm-corrected primary family. **H1** (pooled $-\$2.41$, $p = 5.8 \times 10^{-4}$, same masking risk as the primary pooled test): Gemini/gpt-oss-120b near truthful ($p = 0.369$); Llama flips sign, loses significance ($-\$3.92$, $p = 0.148$); Qwen grows more significant ($-\$5.73$, $p = 1.6 \times 10^{-19}$). **H3**: Llama's λ stays near noise (0.013 vs. 0.0121 at $n = 5$, not a further decline; LR $p = 0.42$ vs. uniform); Qwen's rises (2.22). Qwen stays QRE-structured while Llama's H1 flips sign as its QRE fit drifts toward noise: consistent with the main-text split, outside the confirmatory design.