ENGINEERING SIMPLICITY: SIMPLE MECHANISM INTERFACES STEER LLM AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Can reasoning scaffolds and task descriptions help large language model (LLM) agents make better decisions? We study this question in a multi-agent testbed of auctions and matching, where economic theory specifies an optimal strategy and principles for making it easier to recognize. Across four model families, we vary the interface through which agents act and the descriptions accompanying their decisions. An ascending auction interface that makes truthful play obviously optimal substantially reduces bid errors. With the interface fixed, laying out payoff contingencies and stating why truth-telling is safe also improve choices, while prompts for forward planning or belief formation can worsen them. In auctions, these behavioral gains are not accompanied by improvements in measured verbal indicators of strategic understanding in agents' stated plans. Conversely, a scaffold that changes the targeted language leaves bids unchanged, and the interventions that change both worsen bids. Human-motivated theories of simplicity can therefore inform the design of decision environments for artificial agents. Their success should be evaluated through realized decisions alongside explanations: effective scaffolds and descriptions need not produce corresponding changes in measured verbal understanding.

1 INTRODUCTION

An LLM agent's decisions depend on how a task is presented: which choices it faces, what consequences are made explicit, and what reasoning the prompt asks it to perform. Can we design interfaces and descriptions that improve its decisions, and do those improvements appear in its explanations? We study these questions in auctions and matching, multi-agent environments with explicit rules, payoffs, and a theoretically optimal strategy. These settings let us test whether an intervention improves choices even when measured features of its explanation remain unchanged.

Economic theory supplies both this benchmark and principles for making it easier to use. A *mechanism* is a protocol that maps participants' actions or reports to outcomes, such as who receives an item and what they pay. A mechanism is *strategy-proof* if truthful reporting is optimal regardless of others' reports. Yet human participants overbid in second-price auctions and misreport preferences in matching (Kagel & Levin, 1993; Kagel, 1995; Rees-Jones, 2018). The theory of simplicity asks how an interface's sequence of choices and description of the rules make good play easier to recognize (Li, 2017; Börgers & Li, 2019; Pycia & Troyan, 2023; Li, 2024). We ask whether these human-motivated principles also help artificial agents, which misplay such tasks too (Shah et al., 2024).

We study two canonical settings. In a *second-price sealed-bid* (SPSB) auction, each bidder submits a bid; the highest bidder wins and pays the second-highest bid. Bidding one's private value is optimal because the bid determines whether one wins, while others' bids determine the payment (Vickrey, 1961). In *matching*, participants rank potential partners; our school-choice setting assigns students to schools with limited seats and fixed admission priorities. The student-proposing *deferred acceptance* (DA) algorithm processes these rankings through proposals, tentative acceptances, and rejections (Gale & Shapley, 1962). Students optimally report their true rankings (Roth, 1982).

The same outcome rule can be presented through different sequences of choices. This sequence, together with the information available at each choice, is the *extensive form*. An ascending auction asks bidders whether to stay or exit at each price, instead of requesting one sealed bid. *Obvious*

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.

strategy-proofness (OSP) gives a formal criterion for why such an interface can be easier: at each decision, the worst outcome from following the truthful strategy is at least as good as the best outcome from deviating (Li, 2017). This strengthens strategy-proofness by removing the need to compare strategies separately under every possible behavior of others, making it a central design principle for participants with limited contingent reasoning. Ascending auctions satisfy OSP; prior work finds that LLM auction behavior depends on the format (Shah et al., 2024). DA generally has no OSP implementation (Ashlagi & Gonczarowski, 2018), but a restriction on schools' priority orderings, called acyclicity (Ergin, 2002), permits a sequence of simple yes/no queries. We compare this *iterative DA* interface with the standard interface requiring complete rankings.

Across four model families, the ascending auction substantially reduces departures from truthful bidding. Matching supplies a complementary interface comparison, although partial reports require a different error measure. But changing an interface is not always feasible, so we also hold it fixed and vary the accompanying text. Following Li (2024), we test scaffolds for contingent reasoning about others' actions, forward planning over one's own decisions, and belief formation about other players. We separately test descriptions that emphasize the property making truthful play safe (Gonczarowski et al., 2023; 2024). Laying out payoff contingencies improves choices; stating the safety property does too, reducing matching errors from 4.2% to 0.2%. Prompts for forward planning or belief formation can instead worsen play. More requested reasoning is therefore not uniformly helpful.

To assess whether these changes appear in agents' explanations, we analyze the short plans written before auction bids. Fixed text indicators record whether plans state the payment rule, the safety property, or dominance. Such plans need not faithfully report the computation behind an answer (Turpin et al., 2023; Lanham et al., 2023); our question is whether observable changes in plans accompany changes in behavior. Plans often describe a rule suited to first-price auctions, where the winner pays her own bid: bid below value to preserve a profit margin. That rule is misplaced in a second-price auction. This connects the analysis to theories of simple decision rules (Simon, 1956; Gigerenzer & Gaissmaier, 2011; Shah & Oppenheimer, 2008), without assuming a particular internal computation.

The paper makes two contributions:

- **Human-motivated simplicity principles inform LLM decision environments.** Changes to mechanism interfaces and descriptions improve choices across four model families in auctions and matching. These settings provide a controlled testbed for designing prompts and interfaces around the demands of a decision task.
- **Behavioral improvements need not improve measured verbal understanding.** Effective auction interventions reduce bid errors without corresponding improvements in our stated-understanding indicators. A different scaffold changes targeted language without improving bids, while interventions that change both worsen bids. Explanations and decisions therefore provide distinct evidence when evaluating an interface.

Our aim is to improve agent decisions. Prior work tests whether LLMs reproduce human experimental patterns (Horton, 2023; Aher et al., 2023) and documents suboptimal bidding and possible collusion (Chen et al., 2023; Fish et al., 2024). We instead use simplicity theory to design interventions, motivated also by the role of task complexity in behavioral errors (Oprea, 2024). Section 2 defines the testbed; Sections 3 and 4 evaluate interface and text changes; Section 5 compares auction plans with bids. The appendices provide related work, prompts, additional interventions, and analysis details.

2 Experimental Design

Auctions and matching are canonical problems of allocating scarce resources among multiple participants. They provide complementary tests: auction agents choose a number, while matching agents rank options. Both tasks have a truthful strategy that is optimal regardless of others' reports. We measure auction errors by the distance between bid and value, and complete-ranking errors by the fraction of pairwise preferences reported incorrectly. Iterative matching is scored separately on binary responses and nonforced final choices; lower error is better.

Matching. Each instance is a school-choice market with four students $i \in \{A, B, C, D\}$ and four unit-capacity schools $s \in \{w, x, y, z\}$. Values combine shared school appeal with individual preferences, $v_{i,s} = c_s + \varepsilon_{i,s}$ with $c_s \sim \text{Unif}[40, 70]$ and $\varepsilon_{i,s} \sim \text{Unif}[0, 20]$, so demand is correlated while tastes remain idiosyncratic; a student's true ranking is the descending sort of her values (Klijn et al., 2019). Schools use fixed Ergin-acyclic priorities that create a top tier $\{A, B\}$ and a bottom tier $\{C, D\}$—the condition under which a sequential-query implementation of DA is OSP—and all students see a fixed "global popularity" ranking ($y \succ x \succ w \succ z$) that may or may not align with their own preferences. Two protocols are run on each market. Under *direct revelation*, each student submits a complete ranking in one shot and student-proposing DA is run on the reports. Under *iterative DA*, following the tree construction of Ashlagi & Gonczarowski (2018), a student holding top priority at some remaining school is asked "Among the remaining schools, is school s your most preferred?" and is matched immediately on YES; when only one student has top priority, that student chooses her favorite remaining school. One LLM instance plays each student; the prompt bundles identity, priorities, the popularity signal, the rules, and any treatment text, and requires tagged reasoning and decision fields (Wei et al., 2022). We target $N = 50$ repetitions per condition (retained counts can be lower after response or parsing failures) and log every report, DA trace, and match (Appendix D).

Auctions. We study auctions with $N = 3$ bidders whose values are private and independent; values are drawn uniformly from $\{0, 1, \ldots, 49\}$, bids lie on a \$0.01 grid (\$0.50 for clock prices), and each instance is one-shot. In the SPSB auction the highest bidder wins and pays the second-highest bid. In the ascending clock the price rises from zero in fixed increments, each active bidder chooses at every posted price whether to stay or exit without being told when others leave, and the last bidder standing wins at the price at which the second-to-last exited. Each bidder's decision is elicited by a self-contained prompt with the bidder's identity, a mechanism description that varies by intervention, the private value, and a structured response format. The `<PLAN>` block (requested maximum of 50 words; realized median 68) is retained verbatim for every bid and forms the corpus of Section 5. Treatments modify only the rule-explanation block of the prompt, holding values, identities, and payment rules fixed. Two prompts beyond those of Section 4 enter the trace analysis: a *worst-case* scaffold that asks the bidder to reason about the worst case for each bid rather than laying the cases out, and a *menu restatement* of the second-price rule in the spirit of Gonczarowski et al. (2023), which describes the payment as a "price to win" set by the other bids (Appendix A).

Models. We test GPT-4o (OpenAI, 2024a), Claude 3.5 Haiku (Anthropic, 2024), Gemini 2.0 Flash (Pichai et al., 2024), and Gemma 3 27B (Gemma Team, 2025) at temperature 0.5: models capable enough to parse a mechanism's rules yet still susceptible to the failures simplicity is meant to remedy. At the frontier, raw reasoning ability may mask the bottlenecks our interventions target; Appendix C.8 reports, as a boundary condition, how the stated plans of four frontier models differ from those of this panel. Figures report models in the order Claude, Gemini, GPT-4o, and Gemma.

3 OSP Results

We first ask whether changing the sequence of choices makes truthful play easier. The paired interfaces are designed around the same economic outcome rule, with different extensive forms. The standard version requires contingent reasoning: to see that truth-telling is dominant, an agent must trace hypothetical scenarios about others' behavior and verify that no deviation improves payoffs in any case. In the OSP version, at every decision the worst outcome from following the truthful strategy is at least as good as the best outcome from deviating, across the possibilities the agent cannot distinguish (Li, 2017). Figure 1 compares the two in both domains.

Auctions. All four models deviate from truthful bidding in the SPSB auction, with mean deviations of $\mu = -0.5$ (Claude), -1.6 (Gemini), -3.3 (GPT-4o), and -5.3 (Gemma). The sign indicates systematic *underbidding*, in contrast to the overbidding typical of human subjects (Kagel & Levin, 1993), and the distributions have long tails of \$10–20 errors. The ascending clock corrects deviations across all models: means compress to -0.1, -1.2, -0.5, and -0.3, Gemma's near-complete correction being the largest, and the long tails disappear. Because opponents' exits are not shown, the clock supplies no additional observations of their behavior before a decision. The change is in how choices are elicited. OSP simplifies play along several dimensions at once; Section 4 tests

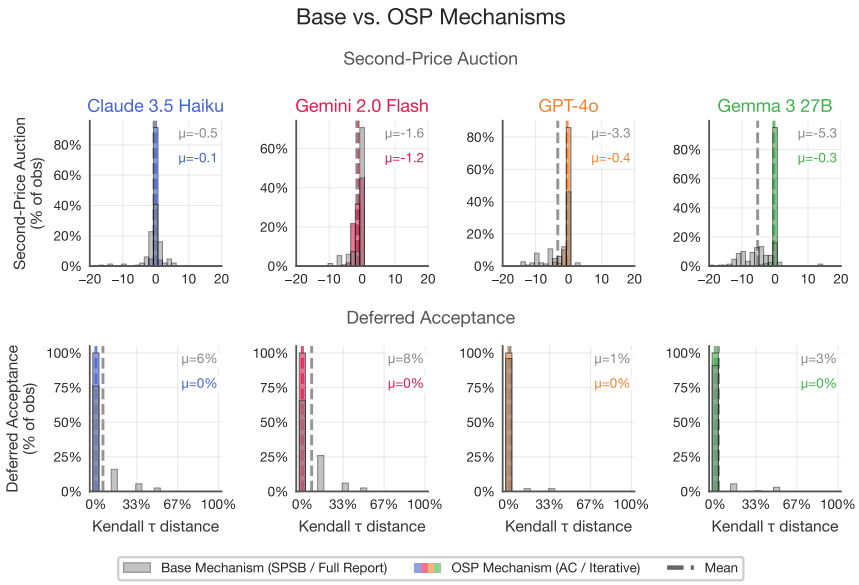

Figure 1: Auction and matching interface comparisons across four model families. Top: signed bid deviations ($b - v$) in sealed-bid (gray) and ascending-clock (colored) auctions; dashed lines indicate means. Bottom: normalized Kendall τ summaries for complete rankings in direct DA and partial rankings in iterative DA. The iterative zero is not evidence of error-free responses: the ranking statistic cannot detect incorrect yes/no answers without an observed pair of ranked schools. Decision-level errors are reported in the text and Appendix E.1.

textual supports addressing these demands. The agents' clock rationales bear on the same question: exit decisions track value equally well whether or not the agent states the "stay while the price is below value" rule (Appendix C.9).

Matching. In direct-revelation DA, mean Kendall τ distance is 6% (Claude), 8% (Gemini), 1% (GPT-4o), and 3% (Gemma). Iterative DA's recorded partial-ranking statistic is zero, but its single-school selections often contain no pair to rank; this does not establish truthful play. Scoring binary responses and nonforced final choices instead yields errors on $16/194$ decisions for Claude, $0/179$ for Gemini, $0/188$ for GPT-4o, and $3/210$ for Gemma. These decision-level rates and complete-ranking Kendall distances use different denominators, so they do not establish a common-scale improvement. Appendix E.1 details the measurement distinction.

4 Intervention Results

OSP mechanisms simplify the strategic problem along several dimensions at once. We next ask which textual supports improve decisions when the sequence of choices stays fixed. We modify only the rule-explanation block of the prompt. Following Li (2024), the scaffolds are organized by the cognitive demand they address—contingent reasoning about unobserved moves, forward planning over one's own future moves, and beliefs about other players—and none directly instructs the agent to report truthfully. Separately, following Gonczarowski et al. (2023), one *outcome safety* description per mechanism states the property that makes truthful play safe, without asking for a reasoning procedure. The main text reports the contingent-reasoning scaffold and the safety descriptions, the two interventions that improve play. Forward-planning scaffolds worsen play in DA, the only setting where we test them; belief scaffolds worsen play in both settings, most for the weaker models (Appendix E). Section 5 qualifies the natural reading that models are "already strategizing" about opponents: the modal baseline plan is a first-price "shade to preserve a margin" heuristic, and the belief scaffolds largely *add* opponent-modeling language rather than amplify existing such language.

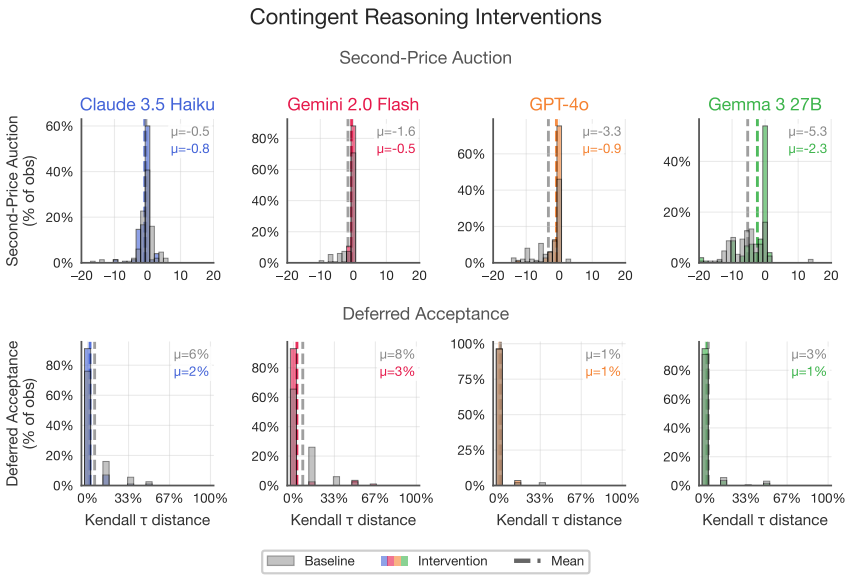

Figure 2: Contingent reasoning: *Payoff Tree* intervention. *Top:* second-price auction bid deviations (bid − value); baseline SPSB in gray, intervention in color. *Bottom:* deferred acceptance Kendall τ errors; baseline direct DA in gray, intervention in color. The scaffold improves play across all models in both domains, roughly halving the distance from truthful.

4.1 Contingent Reasoning

Li (2024) identifies contingent reasoning as the core demand of non-OSP mechanisms: to see that truth-telling is dominant in SPSB, a bidder must compare payoffs case by case across profiles of opponent bids, whereas the clock makes the worst case of the dominant strategy dominate the best case of any deviation at each price. Our *Payoff Tree* intervention performs the case construction for the agent:

- In auctions: "PATH A: your bid exceeds others → you WIN, and pay the second-highest bid. PATH B: your bid falls below → you LOSE, and pay nothing. Key insight: your bid determines which path you are on, not the payment in Path A."
- In DA: "Think of DA as a tree of possibilities. ROOT: You propose to your 1st choice. If ACCEPTED: you are matched. If REJECTED: you propose to your 2nd choice. [...] At each branch, acceptance or rejection depends on the school's priorities and who else proposed there."

The scaffold roughly halves the distance from truthful play in both domains (Figure 2). In auctions the mean deviation moves from $\mu = -2.67$ to -1.13 across models: GPT-4o from -3.28 to -0.86, Gemma from -5.27 to -2.34, Gemini from -1.63 to -0.53, while Claude, already near truthful at -0.49, shows no change in the signed mean (its absolute deviation nonetheless falls from 2.02 to 1.11 against the pooled baseline of Section 5). The clock achieves $\mu = -0.5$; the scaffold closes about half the gap without changing the extensive form. In DA the mean error falls from $\mu = 4.2\%$ to 1.7%, unanimously across models (Claude $5.8\% \rightarrow 2.0\%$, Gemini $7.6\% \rightarrow 2.8\%$, GPT-4o $1.0\% \rightarrow 0.6\%$, Gemma $2.6\% \rightarrow 1.4\%$). The scaffold states consequences of the available actions without directly instructing models to report truthfully. It makes some comparisons explicit that the standard interface otherwise leaves to the agent. A sibling scaffold that asks the bidder to reason about the *worst case* for each bid, rather than laying the cases out, does not detectably change bids, though it changes the targeted language (Section 5.1).

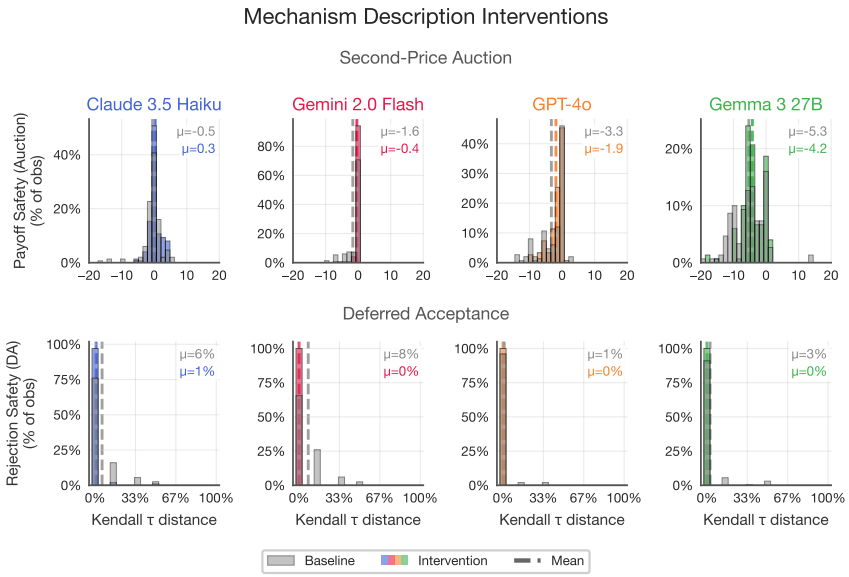

Figure 3: Descriptions emphasizing why truthful play is safe, across four model families. *Top:* signed bid deviations ($b - v$) under *Payoff Safety* in second-price auctions. *Bottom:* fraction of pairwise ranking errors under *Rejection Safety* in DA. Baseline is gray and intervention is colored; zero indicates truthful play. Both improve choices; Rejection Safety reduces mean matching error from 4.2% to 0.2%.

4.2 Descriptions of Why Truthful Play Is Safe

The *outcome safety* descriptions emphasize, without asking for a reasoning procedure, the property that makes each mechanism strategy-proof. In auctions, following Vickrey (1961)'s pivotal argument, *Payoff Safety* reads: "Think of your bid as setting your maximum price. Your bid determines IF you win, not WHAT you pay. If you win, you pay what others bid, not what you bid." In DA, the fact that a rejection sends the application to the next ranked school gives *Rejection Safety*: "Rejections only redirect, never eliminate. Your ranking determines the *order* in which you are considered, not *whether* you are considered."

Both descriptions improve truthful play (Figure 3). In auctions, *Payoff Safety* reduces the mean deviation from $\mu = -2.67$ to -1.53, with a smaller absolute deviation for every model (Claude $-0.49 \to +0.30$, Gemini $-1.63 \to -0.38$, GPT-4o $-3.28 \to -1.88$, Gemma $-5.27 \to -4.17$), closing about half the gap to the clock's -0.5. In DA, *Rejection Safety* reduces error from $\mu = 4.2\%$ to 0.2% across all models. This reduction shows that stating a relevant property at the point of decision can improve play without changing the sequence of choices. This behavioral result does not identify how the model computes its decision. Section 5 adds a qualification: the description improves bids without a corresponding change in our measured verbal indicators of strategic understanding. The theoretical literature on incentive-compatible mechanisms thus has direct operational value as statements of how mechanisms work that improve play when placed at the point of decision, whether or not the agents' stated plans restate them.

5 Stated Plans and Realized Bids

Every bid arrives with a *stated plan*: the short explanation the agent writes inside the `<PLAN>` tag before acting. Plans are stated strategies, not full chains of thought, and such explanations need not faithfully report the computation behind an answer (Turpin et al., 2023; Lanham et al., 2023). But they are exactly the evidence a practitioner would collect to audit an interface change: do the interventions that move *bids* also move what agents *say*, and vice versa? We ask this for auctions; the DA reasoning blocks are not yet scored.

Corpus, scoring, and validation. The corpus holds 21,990 plans from the four families; the analysis pool is the 14,073 sealed second-price plans, of which roughly 40% are byte-identical duplicates within model, cell, and value draw at temperature 0.5, so the main comparisons are also reported on the deduplicated pool ($n = 8{,}839$; Appendix C). Each plan is scored with a fixed dictionary of eighteen keyword features—normative recognition (dominance language, correct payment-rule rehearsal, an echo of the payoff-safety invariant), opponent and belief reasoning, stated intent to bid below or above value, safety and risk rhetoric, and style—and assigned a primary decision script by a rule-based decision list. The dictionary was anchored on the vocabulary models actually produce ("overpay," "profit margin") after theory-native phrasings proved essentially absent: the safety-invariant feature fires on 3 of 21,990 plans. An independent LLM judge, shown only the plan text, agrees with the rule-based label at $\kappa = 0.55$ at the coarse grain used here, and its broader "states why truthful bidding is safe" flag fires on 2.7% of the panel's plans (versus 45% for frontier models; Appendix C.8), supporting the finding that explicit safety arguments are uncommon under a broader text measure too.

Plans are not noise. Where a plan names a dollar amount, the realized bid matches it almost exactly (correlation 0.97–0.99), and the stated direction agrees with the realized direction for 83–98% of plans. Thus, many bid errors are consistent with the stated plan, rather than a failure to execute it. The modal plan in the second-price auction is *first-price* logic—"bid just below my value to keep a profit margin"—carried unchanged into a rule under which it is an error; each family's plans show a characteristic mix of such scripts, and plans anchored on no value-based rule carry the largest errors (Appendix C, Figures 21 and 22).

Inference. Each intervention has two outcomes—the change in prevalence of the language it targets and the change in mean absolute deviation $|b - v|$—and we read the joint response. Four contrasts are designated as primary (*Payoff Safety*, *Payoff Tree*, the worst-case scaffold, the menu restatement) and are classified at raw wild-cluster-bootstrap p-values clustered at the run level; the remaining sixteen interventions—belief scaffolds, clock-framed descriptions, prospect-theoretic personas and frames—are exploratory and classified at Benjamini–Hochberg q-values, with language and bid tests corrected as separate families. Null results are evaluated against stated practical-equivalence bounds using two one-sided tests, and every bid effect is cross-checked with a four-model-cluster permutation test (Appendix C). The bid effects for *Payoff Safety*, *Payoff Tree*, first-order beliefs, and the risk-averse persona do not reach significance in that four-model check.[1]

5.1 Interventions: Changes in Language and Bids

Primary contrasts. *Payoff Safety* states why truthful bidding is safe. Bids improve substantially—mean $|b - v|$ falls from 3.20 to 1.95 ($p = 0.015$), in all four families—yet the measured language indicators do not change: the invariant is echoed in *zero* of its 600 treated plans (rule-of-three upper bound 0.5%), and correct payment-rule rehearsal is equivalent to no change within ± 5 points. Plans continue to describe bidding below value, while the stated reductions become smaller. *Payoff Tree* also improves bids without changing the measured language ($3.20 \to 1.29$, $p = 0.017$, all four families) with payment-rule language unchanged ($33.0\% \to 32.7\%$). The *worst-case scaffold* does the opposite: worst-case language rises from 2% to 43% ($p = 0.0075$) while bids do not move ($3.24 \to 3.07$, $p = 0.81$; equivalent to no change within $\pm\$1.00$), and treated plans that articulate the worst case err no less than those that do not. The *menu restatement* does not detectably change bids (3.55 vs. 3.20, $p = 0.66$); its language effect is not directly interpretable because the prompt removes the "second-highest" vocabulary used by the payment-rule indicator. No primary intervention improves bids and measured verbal understanding together.

Exploratory interventions. After correcting for multiple comparisons across all interventions, only interventions that worsen bids also change the measured language. *First-Order Beliefs* raises opponent-modeling language from 27.8% to 42.7% ($q = 0.015$) *and* worsens $|b - v|$ by \$0.76 ($q = 0.027$), in all four families, with the language rise partly the prompt's own vocabulary read

[1] Section 4 reports signed mean deviation against the dedicated SPSB baseline (e.g., $-2.67 \to -1.53$ for *Payoff Safety*); this section reports mean $|b - v|$ against the pooled baseline cells that share the treated prompts' rule text ($n = 1{,}188$), where the same contrast is $3.20 \to 1.95$. The conventions agree in sign for every intervention discussed.

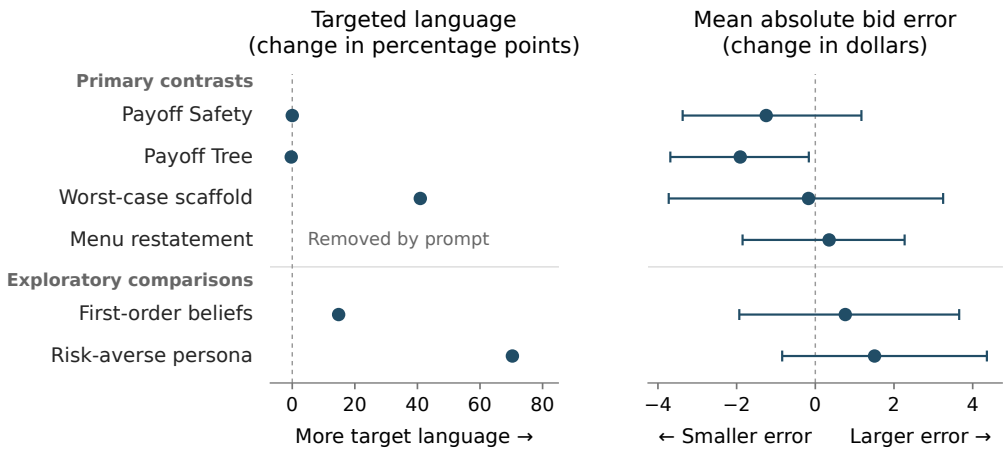

Figure 4: Language and bid responses to six selected interventions, pooling four model families in sealed second-price auctions. Effects are treatment minus baseline. Left: change in each intervention's targeted language feature (point estimates). Right: change in mean absolute bid error; negative values indicate improvement. Whiskers are percentile run-cluster-bootstrap 95% intervals, distinct from the wild-bootstrap tests in Table 1. Menu language is omitted because its prompt removes the measured rule vocabulary. More targeted language need not indicate better understanding. Table 6 reports all twenty interventions.

back; a risk-averse persona does the same (risk language $+70$ points, $|b - v|$ $+\$1.51$). *Second-Order Beliefs* raises opponent language further (51%, $q = 0.020$) but its bid effect does not survive correction ($q = 0.36$), so only its language effect is detected, as for the risk-seeking persona ($+60$ points of aggressive language; bid effect $q = 0.93$). The two effective primary interventions have $q = 0.068$ under the same correction: their unadjusted p-values are below 0.05, but these adjusted values are above it.

Comparing language and behavior. Language predicts the direction of an individual bid once a plan is stated, but treatment-induced changes in language do not track treatment-induced changes in bids: *the interventions that improve bids do not improve the measured language indicators; increases in targeted language need not improve bids; and interventions that change both increase bid errors* (Figure 4, Table 1). A mediation analysis estimates the share of the bid effect associated with changes in these text indicators. Under its assumptions, the 95% upper bound is 35% for *Payoff Safety* (point estimate ≈ 0) and 15% for *Payoff Tree*. These bounds concern the measured indicators, not internal understanding.

6 Discussion and Conclusion

Reasoning scaffolds and task descriptions are design choices for LLM agents. Our results show that theories developed to make decisions easier for humans can inform those choices. In our testbed, the ascending auction substantially reduces bid errors across four model families. With the interface fixed, laying out payoff contingencies or stating why truthful reporting is safe improves decisions in both auctions and matching. Prompts for additional planning or beliefs about opponents can instead introduce irrelevant strategic considerations and worsen play. The useful support depends on the demands of the task.

The auction results also show why decisions and explanations should be evaluated together. Effective interventions improve bids without corresponding improvements in our measured verbal indicators of strategic understanding; a worst-case scaffold changes the targeted language without improving bids. These findings concern short stated plans and the indicators we score, not the model's internal reasoning. They do not establish that understanding is absent or unchanged. They

		Stated reasoning		Bids, mean $\lvert b - v \rvert$ (\$)		
Intervention	Targeted language	Δ prevalence	p / q	base $\rightarrow$ treated	p / q	Detected change
Primary contrasts (raw wild-cluster-bootstrap p)						
Payoff Safety	invariant echo	$0\% \rightarrow 0\%$	—	$3.20 \rightarrow 1.95$	$p = .015$	bids only
Payoff Tree	rule rehearsal	$33.0\% \rightarrow 32.7\%$	$p = .90$	$3.20 \rightarrow 1.29$	$p = .017$	bids only
Worst-case scaffold	worst-case rhetoric	$2.0\% \rightarrow 43.0\%$	$p = .0075$	$3.24 \rightarrow 3.07$	$p = .81$	language only
Menu restatement	rule rehearsal	(by construction)	—	$3.20 \rightarrow 3.55$	$p = .66$	neither
Exploratory rows (Benjamini–Hochberg q, two families)						
First-order beliefs	opponent modeling	$27.8\% \rightarrow 42.7\%$	$q = .015$	$3.16 \rightarrow 3.93$	$q = .027$	both, harmful
Second-order beliefs	opponent modeling	$27.8\% \rightarrow 51.3\%$	$q = .020$	$3.16 \rightarrow 3.87$	$q = .36$	language only
Risk-averse persona	risk language	$29.0\% \rightarrow 99.3\%$	$q = .015$	$3.20 \rightarrow 4.71$	$q = .027$	both, harmful
Risk-seeking persona	aggressive language	$5.8\% \rightarrow 65.5\%$	$q = .023$	$3.20 \rightarrow 3.49$	$q = .93$	language only

Table 1: Joint response of stated reasoning and bids for the selected interventions (four families pooled, sealed second-price cells; full twenty-row table with per-family signs in Table 6). Baselines: pooled scaffold and belief baseline cells (3.20); the worst-case scaffold and the belief rows are compared with their own grids' baselines (3.24, 3.16). Under correction across all interventions the two effective primary interventions carry $q = 0.068$. None of these interventions improves bids and the measured language indicators together.

do show that a verbal comprehension measure alone can miss a behavioral gain, and that increased use of relevant terminology need not signal better decisions.

The broader implication is a design and evaluation approach: identify what a task requires agents to compare, make relevant consequences explicit, and test resulting choices against an independent performance criterion. Auctions and matching are especially useful for this purpose because truthful play supplies that criterion. Transfer to tasks without dominant strategies, richer multi-agent interactions, or other agent architectures remains an empirical question. The matching comparison is limited by the different information revealed by complete rankings and sequential decisions; its priority structure is restricted to permit iterative DA. Our language analysis covers auctions only. Extending that analysis to matching, testing stronger models, and evaluating these principles beyond economic allocation would establish how broadly they apply. Mechanism interfaces and descriptions can improve LLM choices in this testbed; the corresponding gains need not appear in measured verbal indicators of strategic understanding.

THE USE OF LARGE LANGUAGE MODELS

Large language models are the objects of study in this paper. In preparing the manuscript, Claude (Anthropic) assisted with plotting code, formatting, a compliance review for an earlier submission.

REFERENCES

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 337–371. PMLR, 2023. URL https://proceedings.mlr.press/v202/aher23a.html.

Anthropic. Introducing computer use, a new Claude 3.5 Sonnet, and Claude 3.5 Haiku, 2024. URL https://www.anthropic.com/news/3-5-models-and-computer-use.

Itai Ashlagi and Yannai A Gonczarowski. Stable matching mechanisms are not obviously strategy-proof. *Journal of Economic Theory*, 177:405–425, 2018. doi: 10.1016/j.jet.2018.07.001.

Tilman Börgers and Jiangtao Li. Strategically simple mechanisms. *Econometrica*, 87(6):2003–2035, 2019. doi: 10.3982/ECTA15897.

Yves Breitmoser and Sebastian Schweighofer-Kodritsch. Obviousness around the clock. *Experimental Economics*, 25(2):483–513, 2022. doi: 10.1007/s10683-021-09720-z.

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. URL https://arxiv.org/abs/2310.05746.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948v1.

Haluk I Ergin. Efficient resource allocation on the basis of priorities. *Econometrica*, 70(6):2489–2497, 2002. doi: 10.1111/j.1468-0262.2002.00447.x.

Sara Fish, Yannai A Gonczarowski, and Ran I Shorrer. Algorithmic collusion by large language models. *arXiv preprint arXiv:2404.00806*, 2024. URL https://arxiv.org/abs/2404.00806.

David Gale and Lloyd S Shapley. College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15, 1962.

Gemma Team. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Gerd Gigerenzer and Wolfgang Gaissmaier. Heuristic decision making. *Annual Review of Psychology*, 62:451–482, 2011. doi: 10.1146/annurev-psych-120709-145346.

Yannai A Gonczarowski, Ori Heffetz, and Clayton Thomas. Strategyproofness-exposing mechanism descriptions. Technical Report 31506, National Bureau of Economic Research, 2023. URL https://arxiv.org/abs/2209.13148v2. July 2023 version.

Yannai A Gonczarowski, Ori Heffetz, Guy Ishai, and Clayton Thomas. Describing deferred acceptance and strategyproofness to participants: Experimental analysis. *arXiv preprint arXiv:2409.18166*, 2024. URL https://arxiv.org/abs/2409.18166.

John J Horton. Large language models as simulated economic agents: What can we learn from homo silicus? Technical Report 31122, National Bureau of Economic Research, 2023. URL https://arxiv.org/abs/2301.07543v1. Original 2023 version.

John H Kagel. Auctions: A survey of experimental research. In John H Kagel and Alvin E Roth (eds.), *The Handbook of Experimental Economics*, pp. 501–586. Princeton University Press, 1995. doi: 10.1515/9780691213255-009.

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.

Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979. doi: 10.2307/1914185.

Flip Klijn, Joana Pais, and Marc Vorsatz. Static versus dynamic deferred acceptance in school choice: Theory and experiment. *Games and Economic Behavior*, 113:147–163, 2019. doi: 10.1016/j.geb.2018.08.009.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023. URL https://arxiv.org/abs/2307.13702.

Shengwu Li. Obviously strategy-proof mechanisms. *American Economic Review*, 107(11):3257–3287, 2017. doi: 10.1257/aer.20160425.

Shengwu Li. Designing simple mechanisms. *Journal of Economic Perspectives*, 38(4):175–192, 2024. doi: 10.1257/jep.38.4.175.

OpenAI. GPT-4o system card, 2024a. URL https://openai.com/index/gpt-4o-system-card/.

OpenAI. OpenAI o1 system card. *arXiv preprint arXiv:2412.16720*, 2024b. URL https://arxiv.org/abs/2412.16720.

Ryan Oprea. Decisions under risk are decisions under complexity. *American Economic Review*, 114 (12):3789–3811, 2024. doi: 10.1257/aer.20221227.

Sundar Pichai, Demis Hassabis, and Koray Kavukcuoglu. Introducing Gemini 2.0: our new AI model for the agentic era, 2024. URL https://blog.google/innovation-and-ai/models-and-research/google-deepmind/google-gemini-ai-update-december-2024/.

Marek Pycia and Peter Troyan. A theory of simplicity in games and mechanism design. *Econometrica*, 91(4):1495–1526, 2023. doi: 10.3982/ECTA16310.

Alex Rees-Jones. Suboptimal behavior in strategy-proof mechanisms: Evidence from the residency match. *Games and Economic Behavior*, 108:317–330, 2018. doi: 10.1016/j.geb.2017.04.011.

Alvin E Roth. The economics of matching: Stability and incentives. *Mathematics of Operations Research*, 7(4):617–628, 1982. doi: 10.1287/moor.7.4.617.

Anand Shah, Kehang Zhu, Yanchen Jiang, Kerem Dayi, Jeffrey Wang, John J. Horton, and David C. Parkes. Evidence from the synthetic laboratory: Language models as auction participants. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*, 2024. URL https://openreview.net/forum?id=FB9mTtJpJI.

Anuj K. Shah and Daniel M. Oppenheimer. Heuristics made easy: An effort-reduction framework. *Psychological Bulletin*, 134(2):207–222, 2008. doi: 10.1037/0033-2909.134.2.207.

Herbert A. Simon. Rational choice and the structure of the environment. *Psychological Review*, 63 (2):129–138, 1956. doi: 10.1037/h0042769.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, pp. 74952–74965, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract-Conference.html.

William Vickrey. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of Finance*, 16(1):8–37, 1961. doi: 10.1111/j.1540-6261.1961.tb02789.x.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

A PROMPTS

This appendix documents the prompt templates used in our experiments. We present the full base rule prompts for both Deferred Acceptance (DA) and Second-Price Sealed-Bid (SPSB) auction mechanisms, together with the Obviously Strategy-Proof (OSP) sequential interfaces: the Ashlagi & Gonczarowski (2018) decision-tree implementation for DA and the Breitmoser & Schweighofer-Kodritsch (2022) ascending-clock (closed) implementation. For each intervention, we show only the text that is *added to or modified from* the corresponding base prompt, since the surrounding context (preference ordering, priorities, output format) remains identical. Template variables (e.g., `{{student_id}}`, `{{preference_order}}`) are populated at runtime with experiment-specific values.

A.1 DEFERRED ACCEPTANCE (DA)

Prompt skeletons. The two skeletons below summarize the direct-revelation and iterative prompts whose full templates follow. **Prompt skeleton (Direct Revelation).**

> You are Student A. There are 4 schools: w, x, y, z. You will be matched to at most one school.
>
> HOW THE MATCHING WORKS (Deferred Acceptance): (1) All students submit a ranking of schools. (2) Each student proposes to their top-ranked school. (3) Each school tentatively accepts the proposer it ranks highest (by priority), rejecting others. (4) Rejected students propose to their next choice; schools keep the highest-priority student so far. (5) The algorithm ends when no rejections occur.
>
> Your priority rank (1=highest): w:1, x:2, y:1, z:2.
> Most applicants seem to favor: y >x >w >z.
> (TREATMENT INSTRUCTIONS)
>
> Return: `<REASON> ... </REASON>`
>
> `<DECISION> Ranking: ... </DECISION>`

Prompt skeleton (Iterative DA, yes/no query).

> You are Student A. The remaining schools are: w, y, z.
>
> Your preference ordering (most to least preferred): y, w, z.
> Your priority rank at each school (1=highest): w:1, y:1, z:2.
> Most applicants seem to favor: y > x > w > z.
>
> Among the remaining schools {w, y, z}, is **y** your most preferred?
> If YES: you will be matched to y immediately.
> If NO: y remains available and you will be asked again.
>
> Return: `<REASON> ... </REASON>`
>
> `<DECISION> Answer: YES /NO </DECISION>`

A.1.1 BASE RULES — DIRECT REVELATION

In the direct revelation interface, the LLM submits a complete ranking of all schools in a single shot. We vary *how* the mechanism is explained while keeping the underlying student-proposing DA algorithm identical.

Null Baseline. The minimal prompt provides no explanation of the matching algorithm—only the decision environment and output format.

```
You are Student {{student_id}}. There are 4 schools: w, x, y, z.
You will be matched to at most one school based on submitted
rankings and school priorities.

Your preference ordering (most preferred to least preferred):
  {{preference_order}}

Your priority rank at each school (1 = highest priority):
  w: {{pw}}, x: {{px}}, y: {{py}}, z: {{pz}}

Moreover, most applicants seem to favor {{global_ranking}}.

Submit a complete ranking of schools from most preferred to
least preferred.
Format: Ranking: <1st> > <2nd> > <3rd> > <4th>
```

Figure 5: DA Direct Revelation — Null Baseline (`da_direct_null`). No mechanism explanation is provided.

Traditional DA. The standard prompt describes the full five-step Deferred Acceptance algorithm.

```
You are Student {{student_id}}. There are 4 schools: w, x, y, z.
You will be matched to at most one school.

HOW THE MATCHING WORKS (Deferred Acceptance):
1. All students simultaneously submit a ranking of schools
2. Each student "proposes" to their top-ranked school
3. Each school tentatively accepts the proposer it ranks highest
   (by priority), rejecting others
4. Rejected students propose to their next choice
5. This repeats until all students are matched

Your preference ordering (most preferred to least preferred):
  {{preference_order}}

Your priority rank at each school (1 = highest priority):
  w: {{pw}}, x: {{px}}, y: {{py}}, z: {{pz}}

Moreover, most applicants seem to favor {{global_ranking}}.
```

Figure 6: DA Direct Revelation — Traditional (`da_direct_traditional`). The full student-proposing DA algorithm is described step by step.

A.1.2 OSP Sequential Interface for DA

Our OSP implementation for DA follows the decision-tree construction of Ashlagi & Gonczarowski (2018), which provides an OSP mechanism for student-proposing DA under Ergin-acyclic priority profiles. Rather than asking the LLM to submit a full ranking, the mechanism queries *one student at a time* with binary yes/no offers about individual schools, each accompanied by an explicit guarantee of what happens under either response. Algorithm 1 summarizes the procedure.

Algorithm 1 OSP Decision Tree for DA under Acyclic Priorities (Ashlagi & Gonczarowski, 2018)

procedure RUNOSPTREE(students, schools, priorities)
  if students $= \emptyset$ **then**
    return matches
  end if
  top_by_school $\leftarrow \{s : \arg\min_{\text{priority}}$ students at $s\}$ for each remaining school s
  top_students $\leftarrow$ unique(top_by_school.values)
  if $|$top_students$| = 1$ **then** ▷ Serial dictatorship node
    $s^* \leftarrow$ the single top-priority student
    $w \leftarrow$ ASKPICKTOP(s^*, schools) ▷ Use `da_osp_choice` prompt
    Assign s^* to w
    return RUNOSPTREE(students $\setminus\{s^*\}$, schools $\setminus\{w\}$, priorities)
  end if
  Let $a, b \leftarrow$ top_students ▷ $|$top_students$| = 2$ under acyclicity
  for each school w where top_by_school$[w] = a$ (in fixed order) **do**
    if ASKYESNO(a, w, fallback = schools) **then** ▷ Use `da_osp_yesno_guaranteed`
      Assign a to w
      return RUNOSPTREE(students $\setminus\{a\}$, schools $\setminus\{w\}$, priorities)
    end if
  end for
  for each school w where top_by_school$[w] = b$ (in fixed order) **do**
    if ASKYESNO(b, w, fallback = schools) **then**
      Assign b to w
      return RUNOSPTREE(students $\setminus\{b\}$, schools $\setminus\{w\}$, priorities)
    end if
  end for
  $w_a \leftarrow$ ASKPICKTOP(a, schools); assign a to w_a
  $w_b \leftarrow$ ASKPICKTOP(b, schools $\setminus\{w_a\}$); assign b to w_b
  return RUNOSPTREE(students $\setminus\{a, b\}$, schools $\setminus\{w_a, w_b\}$, priorities)
end procedure

The relevant guarantee concerns the quality of the eventual assignment. YES immediately assigns the offered school. If that school is not the student's favorite, truthful continuation after NO guarantees an outcome at least as preferred as accepting the offer, under the construction of Ashlagi & Gonczarowski (2018). NO does not immediately remove the school, but does not reserve every school indefinitely: another student may be assigned a school before the next query. The historical prompt below uses the shorthand "remains available"; we reproduce it as used rather than treating that wording as the formal guarantee.

Yes/No with Consequence Explanation. The primary prompt used at binary-offer nodes in Algorithm 1. It explicitly states the consequence of each answer, providing the "obviousness witness."

```
You are Student {{student_id}}. We are running a step-by-step
matching procedure.

Current remaining schools: {{remaining_set}}

Your preference ordering (most preferred to least preferred):
  {{preference_order}}

Your priority rank at each school (1 = highest priority):
  w: {{pw}}, x: {{px}}, y: {{py}}, z: {{pz}}

Moreover, most applicants seem to favor {{global_ranking}}.

Question: Among the remaining schools ({{remaining_set}}), is
  {{candidate}} your most preferred?
- If you answer YES: you are immediately matched to
  {{candidate}} and leave the process.
- If you answer NO: {{candidate}} remains available to you.
  You continue in the process with all current remaining
  schools ({{fallback_set}}).

Please provide your response in TWO parts:

1. First, explain your reasoning in <REASON></REASON> tags
2. Then, submit your answer in <DECISION></DECISION> tags

Format:
<REASON>
Your reasoning here...
</REASON>

<DECISION>
Answer: YES
</DECISION>

or

<DECISION>
Answer: NO
</DECISION>
```

Figure 7: DA OSP — Yes/No with Consequences (`da_osp_yesno`). Historical prompt describing YES and NO consequences. The blanket availability wording is stronger than the formal continuation guarantee; see the clarification above.

A.1.3 Reasoning Interventions for DA

Each intervention below is appended to (or replaces a section of) the Traditional DA base prompt (Figure 6). The surrounding context—student identity, preference ordering, priorities, global ranking hint, and output format—remains identical. We show only the intervention-specific text that differs from the base.

Axis 1: Contingent Reasoning. These interventions prompt the agent to reason about what happens given others' possible actions. We show the *Enumerate* variant; other Axis 1 variants are listed in Table 2.

```
REASONING GUIDANCE:
Before submitting your ranking, consider: What rankings might
the other students submit? For each possible combination of
their rankings, what school would you receive if you ranked
truthfully vs. if you ranked differently?
```

Figure 8: DA Axis 1 — Enumerate (`axis1_da_enumerate`). Inserted after the DA algorithm description. Other variants: *Dominated* (eliminate dominated strategies), *Worst-Case* (consider minimum obtainable set), *One-Step* (simplify to single-step decision), *Decision Tree* (visualize DA as a tree), *Backward Induction* (reason backwards from final round).

Axis 2: Forward Planning. These interventions scaffold varying depths of forward simulation through the DA algorithm. We show the k=1 variant; other Axis 2 variants are listed in Table 2.

```
HOW THE MATCHING WORKS:
You submit a ranking. The algorithm processes your ranking
sequentially:
- First, you "propose" to your top-ranked school
- If accepted, you're matched there
- If rejected, you propose to your second choice

THINK ONE STEP AHEAD:
Before finalizing your ranking, consider: If you are rejected
from your first choice, which school would you propose to next?
Does this affect how you should rank schools?
```

Figure 9: DA Axis 2 — k=1 Forward Planning (`axis2_da_1step`). Replaces the standard DA algorithm description with a simplified sequential framing plus one-step lookahead guidance. Other variants: k=0 (no guidance), k=2 (two-step lookahead), k=∞ (full simulation), and three *Monotonicity* framings (options never shrink; rejections redirect, never eliminate; outcome determined by priorities).

Axis 3: Higher-Order Beliefs. These interventions prompt reasoning about other agents' beliefs and common knowledge. We show the *Common Knowledge* variant.

```
IMPORTANT: All students know these rules. All students know
that all students know these rules. All students are trying
to maximize their own earnings.
```

Figure 10: DA Axis 3 — Common Knowledge (`axis3_da_common_knowledge`). Inserted after the DA algorithm description. Other variants: *First-Order* (what will others rank?) and *Second-Order* (what do others believe you will rank?).

A.2 SECOND-PRICE SEALED-BID AUCTION

Prompt skeletons. The two skeletons below summarize the sealed-bid and clock prompts whose full templates follow. **Prompt skeleton (SPSB).**

> You are Bidder {Name}. You are bidding against two other bidders.
> MECHANISM DESCRIPTION + TREATMENT TEXT
> Your private value for the item is $\$v_i$.
> Return: `<PLAN> ... </PLAN>`
> `<ACTION> [your bid] </ACTION>`

Prompt skeleton (Clock, at posted price p).

> You are Bidder {Name}. The current posted price is $\$p$.
> If you stay, you remain in the auction at this price. If you exit, you leave permanently. You will not be told when other bidders exit.
> MECHANISM DESCRIPTION + TREATMENT TEXT
> Your private value for the item is $\$v_i$.
> Return: `<PLAN> ... </PLAN>`
> `<ACTION> Stay /Exit </ACTION>`

A.2.1 BASE RULE

The base rule is implemented as the standard Second-price sealed-bid auction.

```
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.
```

Figure 11: SPSB Auction — Base Rule (`private_second_price`). Standard second-price sealed-bid auction with independent private values.

A.2.2 OSP IMPLEMENTATION — TWO-STAGE ASCENDING CLOCK

Following Breitmoser & Schweighofer-Kodritsch (2022), we implement an 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. 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:

```
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.

**END OF AUCTION**
If you win, your earnings will increase by your value for the
prize and decrease by the clock price at the end of the auction.
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.

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

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>
```

Figure 12: SPSB OSP — Ascending Clock (`intervention_proxy_breitmoser`). No dropping out information will be displayed as in Breitmoser & Schweighofer-Kodritsch (2022).

### A.2.3 Reasoning Interventions for SPSB

Each intervention below is appended to or modifies the SPSB base prompt (Figure 11). We show only the intervention-specific text.

**Axis 1: Contingent Reasoning.** We show the *Enumerate* variant. Other Axis 1 variants are listed in Table 2.

```
Before deciding your bid, think through the following: What are
the possible bids the other players might submit? For each
possible bid they might make, what would be your best response?
Does your optimal bid depend on what others do, or is there a
strategy that works well regardless?
```

Figure 13: SPSB Axis 1 — Enumerate (`axis1_contingent_enumerate`). Inserted before the "After each auction" line. Other variants: *Dominated* (identify dominated bids), *Worst-Case* (worst-case analysis per bid), *One-Step* (bid as "maximum price"), *Decision Tree* (PATH A: win vs. PATH B: lose), *Backward Induction* (two-stage sealed → clock with backward-induction guidance).

Axis 3: Higher-Order Beliefs. We show the *Common Knowledge* variant.
```

```
Important: All bidders are rational and this is common
knowledge. This means:
- Every bidder is rational and maximizes their own earnings
- Every bidder knows that every other bidder is rational
- Every bidder knows that every other bidder knows this
- And so on, infinitely

Given this common knowledge of rationality, what is the optimal
bidding strategy? Is there a strategy that all rational bidders
would converge on?
```

Figure 14: SPSB Axis 3 — Common Knowledge (`axis3_beliefs_common_knowledge`). Inserted after the payment rule. Other variants: *First-Order* (what do others bid?) and *Second-Order* (what do others think you bid?).

Mechanism Description and Worst-Case Scaffold (trace analysis). Two further prompts enter the set of interventions of Section 5. The *menu restatement* replaces the payment paragraph of the base rule (Figure 11) with a menu description in the spirit of Gonczarowski et al. (2023); the *worst-case* contingent-reasoning scaffold is the `axis1_contingent_worstcase` variant listed in Table 2, inserted before the "After each auction" line.

```
Your "price to win" the item will be set to the highest bid
placed by any other player. If your bid is higher than this
"price to win," then you will win the item and pay this price.
If you don't win, your earnings will remain unchanged.
```

Figure 15: SPSB Mechanism Description — Menu Restatement (`intervention_menu`). Replaces the payment paragraph of the base rule; note that it removes the words "second-highest" from the rules text, which is why its language column in the trace analysis is uninterpretable by construction (Section 5.1).

```
Before deciding your bid, consider the worst-case scenario for
each possible bid you could make. What is the worst that could
happen if you bid X? Find the bid where even the worst-case
outcome is acceptable. Compare: if you deviate from this bid,
could the worst-case be worse?
```

Figure 16: SPSB Axis 1 — Worst-Case (`axis1_contingent_worstcase`). A contingent-reasoning scaffold that asks the bidder to construct the worst case for each bid rather than laying the cases out; it passes its manipulation check on stated plans but leaves bids unchanged (Section 5.1).

A.3 Summary of All Prompt Variants

Table 2 lists every prompt template used in the experiments. Prompts shown in full above are marked with ⋆; prompts whose intervention-specific text is shown are marked with †.

Category	Prompt ID	Description
Deferred Acceptance — Direct Revelation		
Base	da_direct_null ⋆	Minimal baseline; no mechanism explanation
Base	da_direct_traditional ⋆	Full 5-step DA algorithm description

Category	Prompt ID	Description
Deferred Acceptance — OSP Sequential (Ashlagi & Gonczarowski, 2018)		
OSP	da_osp_yesno ⋆	Binary with consequence explanation (obviousness witness)
DA Axis 1: Contingent Reasoning		
Axis 1	axis1_da_enumerate †	Enumerate others' possible rankings
Axis 1	axis1_da_dominated	Identify and eliminate dominated rankings
Axis 1	axis1_da_worstcase	Consider worst-case obtainable set
Axis 1	axis1_da_onestep	Simplify to one-step decision
Axis 1	axis1_da_tree	Visualize DA as a decision tree
Axis 1	axis1_da_backward_induct	Reason backwards from final round
DA Axis 2: Forward Planning		
Axis 2	axis2_da_0step	k=0: no guidance baseline
Axis 2	axis2_da_1step †	k=1: think one step ahead
Axis 2	axis2_da_2step	k=2: think two steps ahead
Axis 2	axis2_da_fullsim	k=∞: simulate full algorithm
Axis 2	axis2_da_monotonic_options	"Your options never shrink"
Axis 2	axis2_da_monotonic_safety	"Rejections redirect, never eliminate"
Axis 2	axis2_da_monotonic_outcome	"Outcome determined by priorities, not ranking"
DA Axis 3: Higher-Order Beliefs		
Axis 3	axis3_da_firstorder	What will others rank?
Axis 3	axis3_da_secondorder	What do others believe you will rank?
Axis 3	axis3_da_common_knowledge †	Common knowledge of rationality
SPSB Auction — Base		
Base	private_second_price ⋆	Standard SPSB with independent private values
SPSB — OSP Implementations (Breitmoser & Schweighofer-Kodritsch, 2022)		
OSP	intervention_proxy_breitmoser ⋆	ascending clock
SPSB — Mechanism Descriptions (trace analysis)		
Description	intervention_menu †	Menu restatement: "price to win" set by the other bids (Gonczarowski et al., 2023)
SPSB Axis 1: Contingent Reasoning		
Axis 1	axis1_contingent_baseline	SPSB baseline (identical to base)
Axis 1	axis1_contingent_enumerate †	Enumerate others' possible bids
Axis 1	axis1_contingent_dominated	Identify and eliminate dominated bids
Axis 1	axis1_contingent_worstcase	Consider worst-case for each bid
Axis 1	axis1_contingent_onestep	Bid as "maximum price" framing
Axis 1	axis1_contingent_tree	Decision tree: PATH A (win) vs. PATH B (lose)
Axis 1	axis1_contingent_backward_induct	Two-stage sealed $\rightarrow$ clock with backward induction
SPSB Axis 3: Higher-Order Beliefs		
Axis 3	axis3_beliefs_baseline	SPSB with "rational agents" mention

Category	Prompt ID	Description
Axis 3	axis3_beliefs_firstorder	What do others bid?
Axis 3	axis3_beliefs_secondorder	What do others think you bid?
Axis 3	axis3_beliefs_common_knowledge †	Common knowledge of rationality

Table 2: Complete list of prompt templates. ⋆ = full prompt shown; † = intervention-specific text shown.

B PROSPECT-THEORETIC INTERVENTIONS

In addition to the cognitive interventions reported in the main text, we test a separate family of interventions inspired by prospect theory (Kahneman & Tversky, 1979) and behavioral departures from expected utility theory. These interventions do not scaffold strategic reasoning; instead, they reframe the payoff structure to test whether LLMs exhibit sensitivity to framing effects, loss aversion, the endowment effect, and risk attitudes. We test these across both auctions and DA, with the same four model families.

B.1 LOSS AVERSION AND FRAMING EFFECTS

We test five variants that reframe the payoff description while holding the mechanism and underlying payoffs constant:

- *Gain Frame*: Presents all outcomes as gains from zero ("you cannot lose money—you can only gain").
- *Loss Frame*: Endows the agent with an initial amount equal to their maximum possible value, then presents outcomes as losses from this reference point ("Think carefully about what you might lose").
- *Mixed Frame*: Explicitly decomposes each outcome into a GAIN component and a LOSE component, asking the agent to weigh both.
- *Endowment*: Gives the agent an initial endowment and frames payment as a deduction "from your endowment," targeting the endowment effect.
- *WTA/WTP*: Compares the agent's willingness to accept a selling price with its willingness to pay a buying price, invoking the WTA–WTP gap.

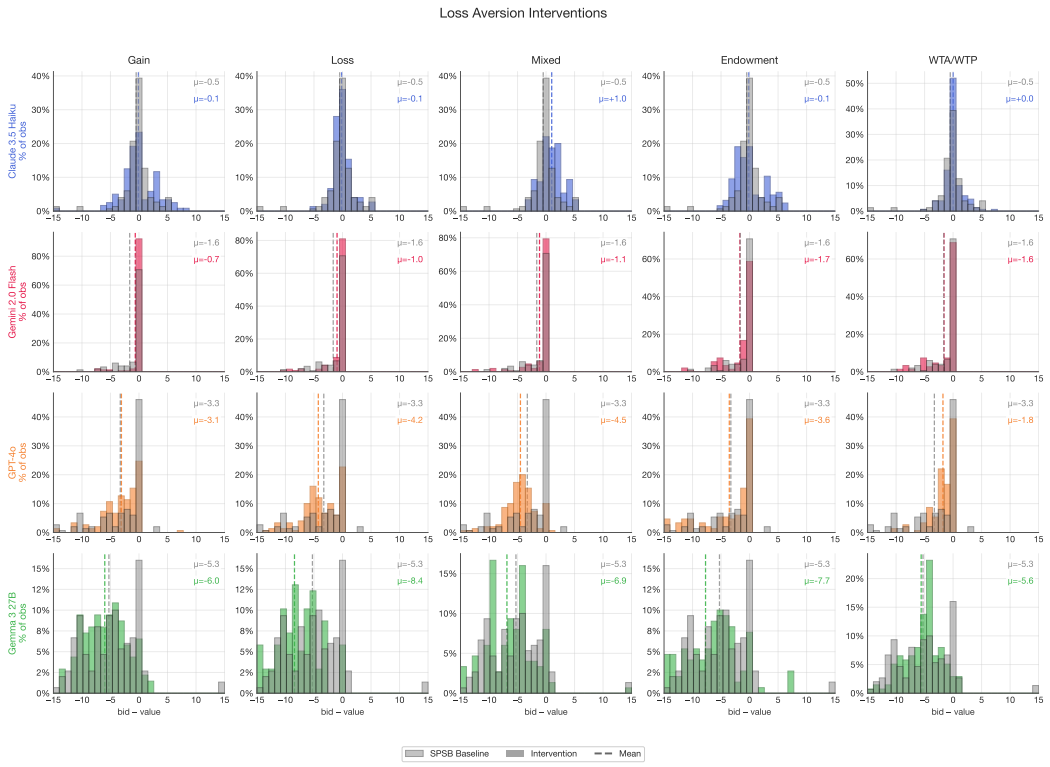

Figure 17: Loss aversion interventions in SPSB auctions. Baseline SPSB in gray, intervention in color. Dashed lines indicate means. None of the framing interventions consistently improve play relative to the SPSB baseline ($\mu = -2.67$). The loss frame ($\mu = -3.44$) and endowment ($\mu = -3.40$) worsen underbidding, while the gain frame ($\mu = -2.51$) and WTA/WTP ($\mu = -2.17$) show modest, inconsistent effects.

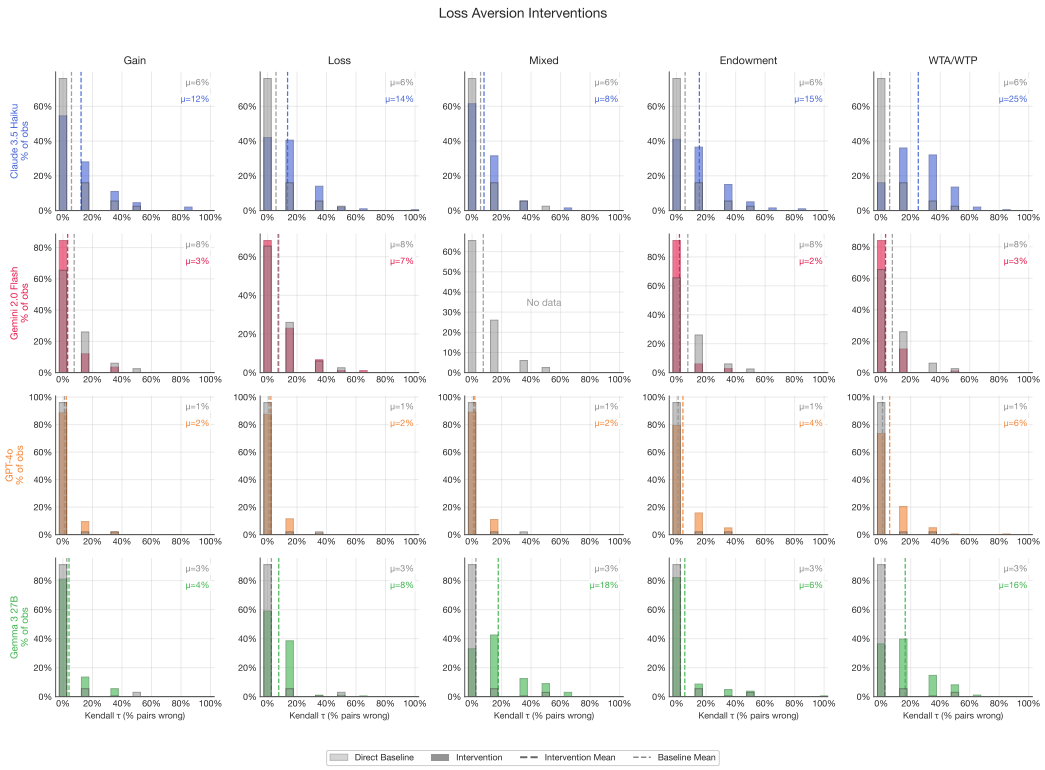

Figure 18: Loss aversion interventions in deferred acceptance. Baseline direct DA in gray, intervention in color. The results are noisy and largely negative: most framing interventions worsen play relative to the baseline ($\mu = 4.2\%$), with WTA/WTP producing the worst outcomes ($\mu \approx 12.6\%$ across models with data). Claude is particularly sensitive to these framings.

In auctions, the framing interventions do not produce consistent improvements. The loss frame and endowment interventions worsen underbidding (overall means of $\mu = -3.44$ and -3.40, compared to the SPSB baseline of -2.67), opposite to what a naive application of loss aversion would predict. The gain frame and WTA/WTP interventions produce small, model-dependent effects. In DA, the interventions are uniformly harmful, with WTA/WTP and endowment producing the largest increases in error.

The overall picture is that prospect-theoretic framings do not improve play—and often worsen it. This contrasts with the main-text interventions along contingent reasoning and mechanism description, which produce large, consistent improvements. In these experiments, descriptions of payoff contingencies and incentive properties are more useful than the tested changes in gain–loss framing.

B.2 Risk Preference Interventions

We also test whether assigning explicit risk attitudes to LLMs changes bidding behavior. Each intervention instructs the model to adopt a calibrated risk preference via a concrete coin-toss example:

- *Risk Averse*: "You would only pay \$4 for a coin toss worth \$0 or \$10 (expected value \$5)."
- *Risk Neutral*: "You would pay \$5 for a coin toss worth \$0 or \$10."
- *Risk Seeking*: "You would pay \$6 for a coin toss worth \$0 or \$10 (expected value \$5)."

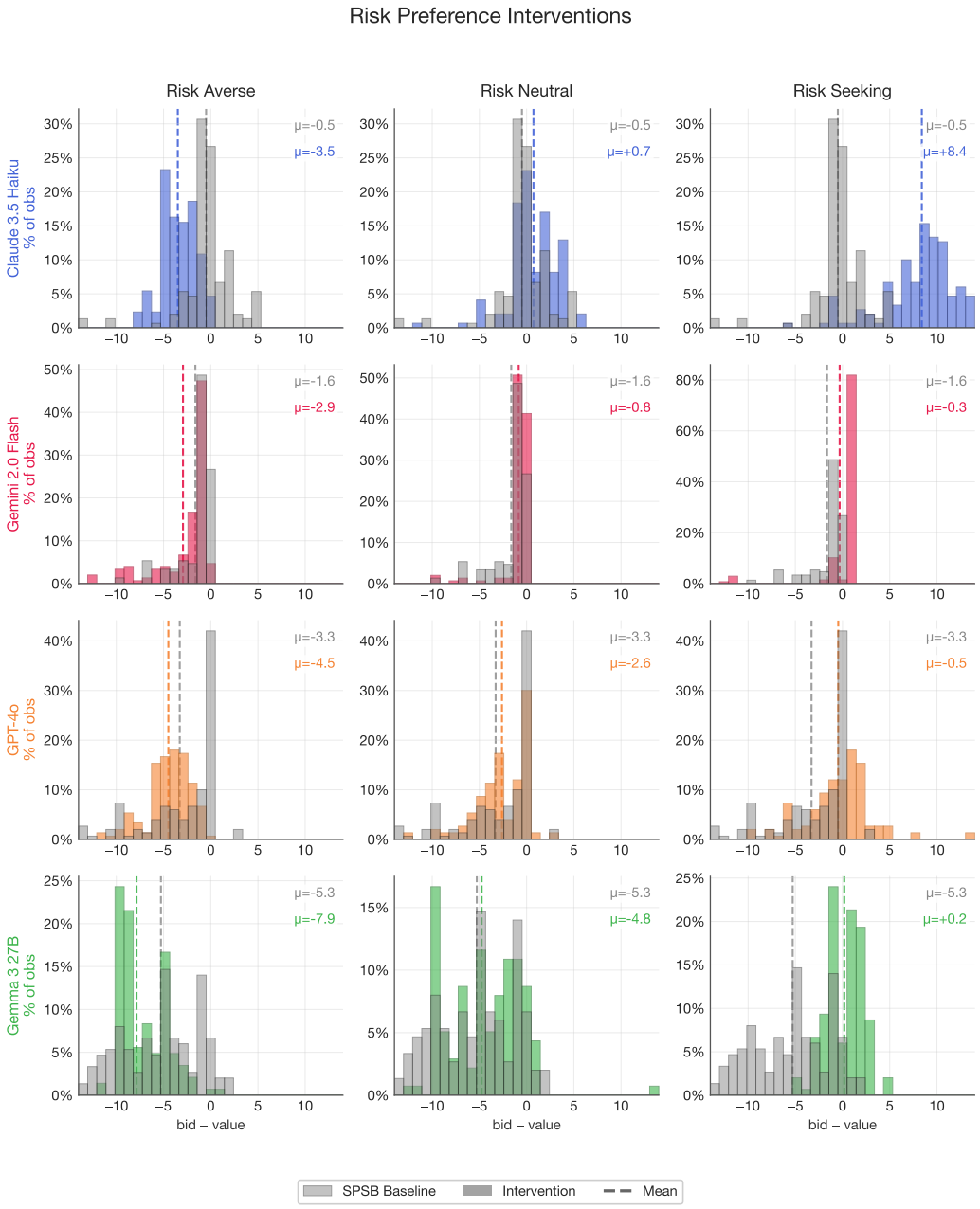

Figure 19: Risk preference interventions in SPSB auctions. Risk aversion increases underbidding ($\mu = -4.71$), while risk-seeking dramatically shifts Claude to overbidding ($\mu = +8.40$). Risk neutrality ($\mu = -1.82$) modestly improves play relative to the SPSB baseline ($\mu = -2.67$).

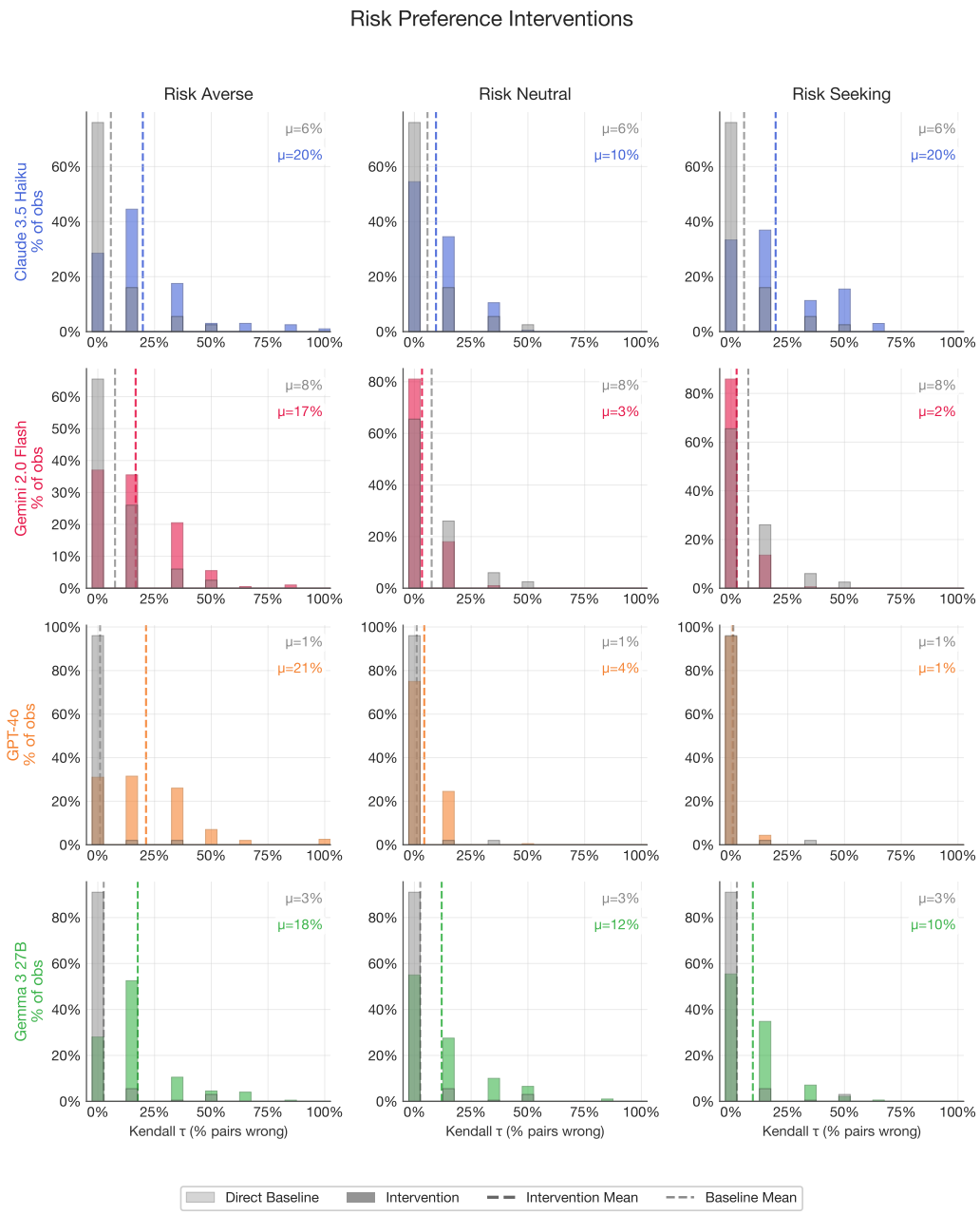

Figure 20: Risk preference interventions in deferred acceptance. Risk aversion dramatically worsens play ($\mu \approx 18.8\%$, compared to baseline 4.2%). Risk neutrality and risk-seeking also worsen play, though less severely.

The risk interventions reveal that LLMs are highly responsive to persona-based instructions about risk attitudes, but that this responsiveness does not improve play. The risk-averse persona produces the most dramatic deterioration: in auctions, it increases underbidding from $\mu = -2.67$ to -4.71; in DA, it roughly quadruples the error rate. The risk-seeking persona shifts Claude 3.5 Haiku to massive overbidding ($\mu = +8.40$), while the other models are less affected. Risk neutrality modestly improves auction play ($\mu = -1.82$) but worsens DA play. These comparisons show that instructions about risk attitudes can substantially change behavior.

Truthful bidding remains dominant for bidders whose utility increases with monetary payoff in this private-value setting. The risk-persona instructions therefore introduce behavioral changes that are not needed to accommodate risk attitudes under the stated rules. These changes do not, by themselves, identify the models' understanding or the cause of their baseline errors.

C Reasoning-Trace Analysis: Details

This appendix explains how the analysis in Section 5 connects stated plans to bids. It describes the corpus, text labels, judge validation, intervention effects, mediation estimates, robustness checks, and additional models. Plans remain verbal outputs, not direct observations of computation.

Replication scripts build the corpus, score text features, classify plans, and estimate intervention effects. Their outputs are stored under `results/traces/`.

C.1 Corpus

The corpus contains 21,990 plans from three sources: (i) 13,416 plans from the harmonized four-model grid (92 run directories covering the sealed-bid SPSB baseline and intervention cells plus the closed clock); (ii) 1,167 plans from the menu-restatement and clock-framing cells, all four models; and (iii) 7,407 recovered GPT-4o plans from the first- and third-price variants of the same grid, which provide the cross-mechanism control of Section C.2. Duplicate second-price cells present in both (i) and (iii) were excluded from (iii). Each plan carries its run directory as a cluster identifier. The 14,073 sealed second-price plans form the analysis pool for all share and regression statistics; the 510 ascending-clock rationales are per-decision exit statements and are scored separately (Section C.9). The corpus and dictionary were frozen before the intervention analysis (manifest hash in `results/traces/corpus_manifest.json`); the frontier extension of 5,985 plans (GPT-5, GPT-5 mini, Claude Sonnet 5, Gemini 2.5 Flash) and an 8,616-plan per-format robustness grid were scored afterwards with the same frozen dictionary plus one additive feature group for formal derivation.

C.2 Feature Dictionary, Decision Scripts, and Model Fingerprints

Each plan is scored with eighteen binary keyword features (Table 3), applied as transparent regular expressions. The features were tuned on the language models actually produce: an initial dictionary keyed to the theory's vocabulary ("no risk," "can't lose") turned out to be essentially absent from real plans—`safety_recognition`, which detects an echo of the *Payoff Safety* invariant, fires on 3 of 21,990 plans—so the safety and risk group was re-anchored on the words models use ("overpay," "profit margin," "buffer," "zero profit"). The dictionary does no negation handling by design: `overpay_concern` counts "avoid overpaying" as overpayment rhetoric either way, which is the intended reading for the main panel but a blind spot at the frontier (Section C.8).

Decision scripts. A rule-based decision list (`analysis/classify_heuristics.py`) over the features, the additive formal-derivation group, and a stated-bid parser assigns each plan one primary script: H1 profit-margin shading; H2 avoid overpayment; H3 opponent anchoring; H4 salient-number anchoring without a value-based rule; H5 win-probability or aggressive bidding; H6 price-threshold exit (clock); H7 formal dominance derivation; H8 worst-case safety proof; T informal truthful or dominance assertion; U unclassified. Because the LLM-judge replication (Section C.3) shows that H1 and H2, and H7 and T, are not reliably separable from text alone, the main text reports shares at a coarse grain: below-value script (H1+H2), opponent (H3), anchor (H4), aggressive (H5), clock exit (H6), normative (H7+H8+T). Script labels are descriptive only and are never outcome variables in the intervention analysis of Section C.5. Verbatim exemplars for every label are in `results/traces/heuristic_examples.md`.

The six scripts, with exemplars. The plans are not cheap talk, but neither are they textbook reasoning about the second-price rule. Read by primary script (Figure 21), the modal plan in the sealed second-price auction is *first-price* logic: a heuristic that would serve a bidder well in an auction where the winner pays her own bid, applied to a rule under which she does not. Six scripts account for nearly all classified plans; we give one verbatim exemplar of each.

Group	Feature	Fires when the plan…
Normative	`dominance_language`	invokes dominance ("dominant strategy," "regardless of others")
recognition	`truthful_intent`	states an intent to bid exactly one's value
	`payment_rule_correct`	correctly rehearses that the winner pays the second-highest bid
	`second_price_mention`	name-drops "second-highest"/"second price" (weak rule mention)
	`first_price_mention`	calls the auction "first-price" (mechanism confusion)
Opponents & beliefs	`opponent_modeling`	speculates about other bidders' values or bids
	`probability_reasoning`	uses chance/likelihood language
	`expected_value_reasoning`	computes or invokes expected value
Stated intent	`shading_intent`	states an intent to bid *below* value (value-anchored)
	`overbid_intent`	states an intent to bid *above* value (value-anchored)
Safety / risk rhetoric	`worst_case`	reasons about the worst case
	`safety_recognition`	echoes the *Payoff Safety* invariant ("bid determines if you win, not what you pay")
	`overpay_concern`	worries about overpaying
	`zero_profit_fallacy`	argues that bidding one's value yields zero profit
	`margin_language`	invokes a profit margin or buffer
Style	`conservative_language`	self-describes as cautious/conservative
	`aggressive_language`	self-describes as aggressive/competitive
	`risk_language`	uses risk vocabulary

Table 3: The frozen trace feature dictionary: eighteen binary keyword features in five groups. Exact regular expressions and per-cell prevalence are in the replication package (`analysis/build_trace_features.py`; `results/traces/trace_features.csv`).

- *Profit-margin shading*—bid below value to keep a margin between value and payment (GPT-4o, value \$34, bid \$33.9): "Bid just below my value to maximize my profit if I win. Aim to outbid others while staying below \$34 to ensure positive earnings." The script uses own value and a desired margin and ignores the rule that sets the payment. Among plans stating an intent to shade, 97.1% of realized bids fall below value (mean deviation $-\$1.84$). It is the primary script of 63% of Gemini's sealed second-price plans, 38% of GPT-4o's, and 37% of Claude's, but only 2% of Gemma's.

- *Avoid overpayment*—lower the bid whenever a high bid feels capable of causing a loss (Gemini, value \$44, bid \$43.9): "I will bid slightly below my value to ensure a positive profit if I win. I will bid \$43.9 to avoid overbidding and risking a loss." This is the failure to recognize the payoff-safety invariant in its purest form; it is Gemma's modal script (38%) and GPT-4o's second (28%).

- *Opponent anchoring*—forecast competitors' bids and bid just above or below the forecast (GPT-4o, value \$33, bid \$31): "To maximize profit, bid slightly above the expected average bid. Assume others bid around \$25–\$30. Bid \$31 to win while minimizing payment risk." The beliefs are unnecessary for the dominant-strategy comparison. This script is a minority one at baseline; it is the one the belief scaffolds of Appendix E.3 activate.

- *Salient-number anchoring without a value-based rule* (Gemini, value \$17, bid \$8.5): "I will bid a fraction of my value. A good starting point is half my value. […] I will bid 8.5." Plans with no value-anchored intent at all carry the largest errors (Appendix C.4).

- *Win-probability, or aggressive bidding* (GPT-4o, value \$14, bid \$14.5): "Bid slightly above my value to maximize profit if I win. Since my value is \$14, I will bid \$14.5 to increase chances of winning while minimizing potential loss." Among plans stating an intent to bid above value, 87.6% do (mean $+\$1.88$); Claude is the outlier here, with an aggressive primary script in 30% of its plans.

- *Price-threshold exit*, the clock's script (Gemini, value $34): "My value is 34. I will stay in as long as the price is less than or equal to 33. I will drop out if the price reaches 33.5." We return to it in Appendix C.9.

Two facts about these scripts matter for what follows. First, each family has a recognizable fingerprint—Gemini a near-pure shader, Gemma an unanchored worrier about overpaying, Claude the noisiest and most often aggressive, GPT-4o a shader with a secondary overpayment script—and the fingerprints line up with the ordering of baseline errors in Section 3. Second, the scripts are largely *insensitive to the payment rule*. In the cross-mechanism control cells, GPT-4o states an intent to shade in 66%, 68%, and 60% of its first-, second-, and third-price plans and bids 0.82–0.88 of value in all three; in a smaller per-format grid Gemini keeps its shading script across all three rules and Gemma's overpayment script is likewise mechanism-insensitive, while Claude is mixed (Appendix C). A script that is a reasonable heuristic in the first-price auction is carried, unchanged, into a rule where it is an error.

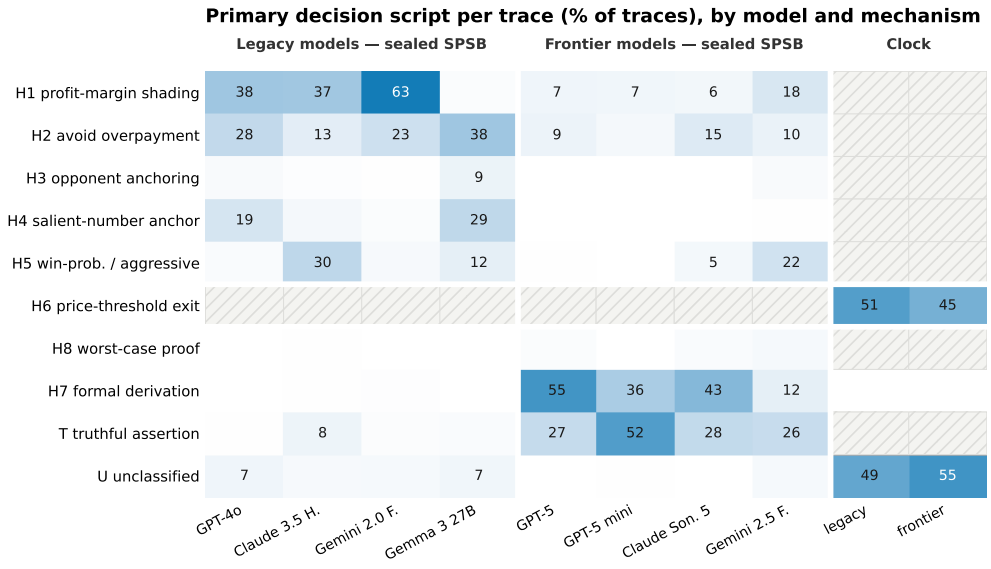

Figure 21: Primary decision script per plan, by model and mechanism (rule-based taxonomy; hatched cells are structurally unavailable labels). The plans of the four families in our main panel ("legacy") state heterogeneous sealed-bid heuristics dominated by first-price shading and overpayment avoidance; the ascending clock collapses every family onto the price-threshold exit rule; the frontier families discussed in Appendix C.8 switch to formal derivation or truthful assertion. LLM-judge replication of the taxonomy: $\kappa = 0.55$ at the coarse grain shown; ten-label grain and confusion matrix in Appendix C.

Fingerprints. Table 4 reports feature prevalence in the baseline SPSB cell of each model ($n = 150$ plans per model). The "fingerprint" shares quoted in Section C.2 instead pool *all* sealed second-price cells of a model (baseline and interventions), which is why adjacent numbers differ slightly (Gemini `shading_intent` 88% in the baseline cell versus 86% pooled; Gemma 4% versus 11%). Both granularities are internally consistent; the text labels which one it uses.

Mechanism-insensitive scripts. Table 5 reports the shading and overpayment scripts by payment rule. GPT-4o, for which the full first-, second-, and third-price grids exist, states an intent to shade in 66%, 68%, and 60% of plans and bids 0.82–0.88 of value under all three rules—one script whether shading is optimal (first-price) or an error (second-price). In the smaller per-format robustness grid, Gemini keeps a below-value script across all three rules (75–87% H1 or H2 primary), Gemma's overpayment concern is likewise mechanism-insensitive (71–88%), and Claude is mixed. A useful side effect is a validity check on the dictionary itself: `shading_intent` prevalence is nearly identical where shading is correct and where it is an error, so the feature measures the stated strategy, not its appropriateness.

Feature (% of plans)	Claude	Gemini	GPT-4o	Gemma
dominance_language	0.0	0.7	0.0	0.0
truthful_intent	8.0	2.7	2.0	0.0
payment_rule_correct	11.3	0.0	9.3	0.0
second_price_mention	50.7	10.0	39.3	46.0
first_price_mention	18.7	0.0	0.0	0.0
opponent_modeling	8.0	4.7	20.0	73.3
probability_reasoning	49.3	22.7	38.0	18.0
expected_value_reasoning	3.3	2.7	2.0	0.0
shading_intent	54.0	88.0	61.3	4.0
overbid_intent	25.3	0.0	2.7	2.0
worst_case	6.0	8.0	2.7	1.3
safety_recognition	0.0	0.0	0.0	0.0
overpay_concern	59.3	22.7	45.3	89.3
zero_profit_fallacy	0.0	0.0	1.3	0.0
margin_language	54.7	29.3	30.7	13.3
conservative_language	6.0	14.7	12.7	86.0
aggressive_language	2.7	5.3	2.0	17.3
risk_language	52.0	6.7	32.7	22.7
Plans	150	150	150	150

Table 4: Feature prevalence in the baseline SPSB cell, by model (percent of plans; Claude 3.5 Haiku, Gemini 2.0 Flash, GPT-4o, Gemma 27B). Gemini is a near-pure shader; Claude is the only model that mislabels the mechanism "first-price" and has both the most stated overbidding and the most stated truthful intent; Gemma produces the least value-anchored language but the most opponent modeling, conservative self-description, and overpayment concern; GPT-4o shades and has the highest rate of the zero-profit fallacy. Prevalence for every model $\times$ lever cell: `results/traces/feature_prevalence_model_family.csv`.

Model	Rule	Plans	shading_intent	H1 or H2 primary	bid/value
GPT-4o (main grid)	first-price	3,711	66%	68%	0.82
	second-price	3,600	68%	66%	0.88
	third-price	3,696	60%	59%	0.88
Gemini 2.0 Flash (robustness grid)	first-price	180	45%	75%	0.88
	second-price	264	64%	80%	0.93
	third-price	330	86%	87%	0.90
Gemma 27B (robustness grid)	first-price	180	2%	17%	0.85
	second-price	270	7%	48%	0.82
	third-price	150	9%	55%	0.75
Claude 3.5 Haiku (robustness grid)	first-price	177	40%	56%	0.98
	second-price	267	18%	37%	1.07
	third-price	330	46%	48%	1.07

Table 5: Cross-mechanism script invariance (`results/traces/script_invariance.csv`). For Gemma, the mechanism-insensitive script is overpayment concern (`overpay_concern` 88%, 84%, 71% across the three rules), not shading. The rule-rehearsal features are mechanically zero outside the second-price format and are omitted.

C.3 LLM-JUDGE VALIDATION

A stratified sample of 2,573 plans (stratified by corpus, model, and rule label; seed 1299) was independently labeled by an LLM judge (Claude Haiku 4.5) that saw only the plan text and the mechanism family—no bid, value, model, or rule label. Its role is a robustness check on taxonomy shares and on plans without a value-based bid rule; judge labels are never the outcome variable in the intervention analysis. Agreement with the rule-based label is Cohen's $\kappa = 0.36$ at the ten-label grain (raw agreement 0.43), $\kappa = 0.45$ with H1+H2 and H7+T merged, and $\kappa = 0.55$ at the coarse decision-mode grain used in the text (raw agreement 0.64; 0.55 for the main panel and 0.52 for the frontier corpus). The largest disagreements are informative: the judge folds most rule-labeled "avoid overpayment" plans into profit-margin shading or truthful assertion (H2 recall 0.06), which is why the two below-value scripts are only ever reported merged; it reads nearly every rule-labeled "unclassified" clock rationale as the price-threshold exit rule (H6 recall 0.96); and it relabels a large share of formal derivations as informal truthful assertions (H7 recall 0.32), which is why the normative family is reported merged. The full confusion matrix is in `results/traces/judge/judge_validation.md`.

The judge also carries a broader "states why truthful bidding is safe" flag, which cross-checks the frozen `safety_recognition` regex (3 hits in 21,990 plans). The flag fires on 2.7% of main-panel plans (Claude 5.4%, Gemini 3.8%, GPT-4o 1.5%, Gemma 0.3%) and on 45% of frontier plans (Claude Sonnet 5 58%, GPT-5 54%, GPT-5 mini 49%, Gemini 2.5 Flash 26%). The broader labels also indicate that explicit safety reasoning is uncommon in the main panel, although its measured prevalence depends on the definition. An optional ∼100-plan human hand-labeling, to complement the LLM judge, has a protocol in the same file and has not yet been run.

C.4 STATED INTENT, REALIZED BIDS, AND THE POOLED REGRESSION

If plans were noise, none of this would matter. They are not. Where a plan names a dollar amount, the realized bid matches it almost exactly: the correlation between stated and realized bid is 0.97–0.99 across the four families, with a mean absolute gap of \$0.28–\$0.82, and stated direction (below, at, or above value) agrees with realized direction for 83–98% of plans, three of four families above 97% (Figure 22a). The models do what they say. The error is in *which plan gets selected*: conditional on the stated-intent class, the plans that never anchor on value at all—4,541 of them—carry by far the largest deviations (pooled mean −\$4.84; Figure 22b), and they are not one phenomenon. Gemma contributes 2,613 of them (75% of its own plans, mean deviation −\$6.05, three quarters of them self-described as "cautious" or "conservative"), GPT-4o 1,040 at −\$4.41, while Claude's 578 unanchored plans sit at +\$0.27: the same label is Gemma's dominant failure mode and Claude's benign residual. A pooled regression of $|b - v|$ on the full dictionary, with model and lever fixed effects, adds an incremental R^2 of only 0.075 (0.069 deduplicated): the plans are informative about *how* an agent errs, far from a sufficient statistic for the error itself.

Among the 7,469 sealed second-price plans stating an intent to shade, 97.1% of realized bids fall below value (mean deviation −\$1.84); among the 1,797 stating an intent to bid above value, 87.6% fall above (mean +\$1.88); among the 950 stating truthful intent, 77.6% land within ±\$0.50 of value (mean $|b - v|$ of \$1.03). The largest errors come from the 4,541 plans that state no value-anchored intent at all: pooled mean deviation −\$4.84, mean absolute deviation \$5.51. The per-model autopsy (`results/traces/mediation/unanchored_autopsy.csv`) shows this is not one phenomenon: Gemma contributes 2,613 unanchored plans (75% of its own; mean deviation −\$6.05; 76% self-described as conservative, 83% voicing overpayment concern, 54% speculating about opponents), GPT-4o 1,040 (29%; −\$4.41), Gemini 310 (9%; −\$5.61), and Claude 578 (17%; +\$0.27). Where a plan names a dollar amount—it does so in 80–93% of main-panel plans—the realized bid matches it with correlation 0.97–0.99 and a mean absolute gap of \$0.28 (Gemma, GPT-4o), \$0.40 (Gemini), and \$0.82 (Claude); stated direction matches realized direction in 97–98% of plans for Claude, Gemini, and GPT-4o and 83% for Gemma (`results/traces/mediation/consistency_pooled_v2.csv`).

A pooled OLS of $|b - v|$ on the full dictionary with model and lever-family fixed effects (n = 14,073; standard errors cluster-robust by run directory) summarizes the same facts: articulating *any* value-anchored plan—shading, overbidding, or truthful—is associated with \$1.5–\$2.2 smaller absolute deviations than an unanchored plan, correct payment-rule rehearsal with a further −\$0.77,

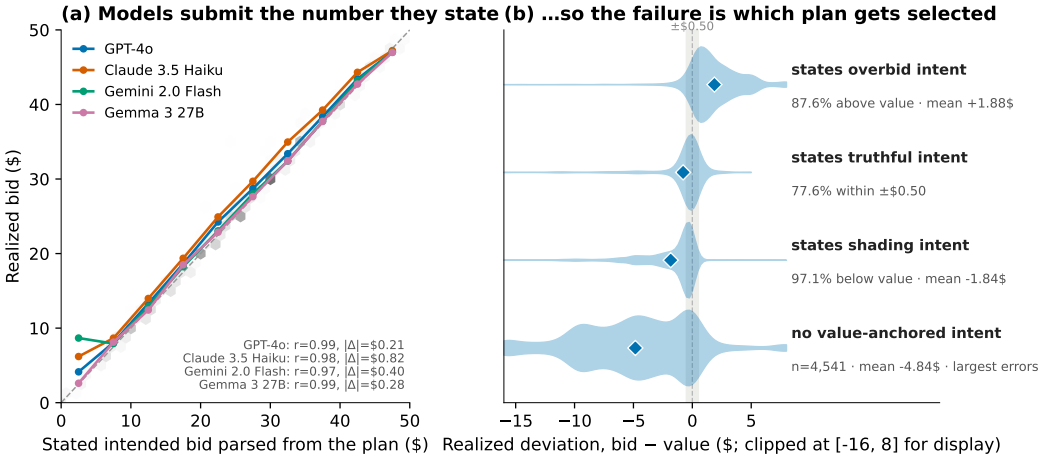

Figure 22: Execution fidelity versus plan selection (main panel, sealed second-price cells). (a) Realized bid against the dollar amount stated in the plan; per-model binned means, $r = 0.97$–0.99, mean $|\text{stated} - \text{realized}|$ of \$0.28–\$0.82. (b) Realized deviation conditional on the stated-intent class; the plans that never anchor on value carry the largest errors. Stated plans are executed faithfully; the failure is in which plan is chosen.

and conservative self-description is the one style feature with a robust positive (worse) association (+\$0.76). Overall explanatory power is modest: $R^2 = 0.251$ with the features against 0.176 with fixed effects only, an incremental R^2 of 0.075 (0.069 on the deduplicated pool). Stated reasoning is informative about *how* an agent errs, far from a sufficient statistic for the error itself.

C.5 Effects of All Twenty Interventions

Table 6 reports all twenty interventions. For each, the targeted feature is the one the lever's text is designed to move; the language and bid columns give baseline and treated prevalence and mean $|b - v|$; the q-values are Benjamini–Hochberg corrected within two families (all twenty language tests; all twenty bid tests), computed from wild-cluster-bootstrap p-values clustered at the run level. The four primary contrasts are *classified* at their raw p (*Payoff Safety* bids $p = 0.015$; *Payoff Tree* bids $p = 0.017$; worst-case language $p = 0.0075$, bids $p = 0.81$; menu bids $p = 0.66$); the full-set q is shown for completeness. Baselines are the pooled contingent-reasoning and belief baseline cells for levers whose prompts share the standard rule text ($|b - v| = 3.20$, $n = 1{,}188$); the contingent-reasoning grid's own baseline for its scaffolds (3.24, $n = 588$); the belief grid's own baseline for the belief prompts (3.16, $n = 600$); and the loss-aversion grid's baseline for its frames (3.13, $n = 567$). The legacy "forward-planning baseline" cell is excluded from every baseline pool and analyzed as a treatment (the two-stage clock-exit description), because its text is itself a clock-framed description of the auction.

Three rows beyond those in the main text deserve a note. The *Dominated* scaffold ("identify and eliminate dominated bids") is a third words-only lever: dominance language rises from 0% to 20.5% ($q = 0.04$) while bids do not move. The *clock-framing* description and the *two-stage clock-exit* description both pass their manipulation checks (clock-and-exit language $0 \to 20\%$ and $0 \to 14\%$), with pooled bid effects that are null and conceal offsetting family effects (Gemma strongly helped, Gemini strongly harmed in both). And the *common-knowledge* belief prompt is the one belief scaffold that improves bids ($3.16 \to 2.13$, $q = 0.027$, soft), while *reducing* opponent-modeling language—consistent with its text, which asserts a conclusion ("all rational bidders would converge") rather than inviting a forecast.

C.6 Mediation Bounds

For each primary contrast we ask how much of the bid effect could run through verbalized understanding, using the classical product-of-coefficients decomposition with the normative-recognition

		Stated reasoning		Bids, mean $\lvert b - v \rvert$ ($)				
Lever	Targeted feature	% base → treated	q	base → treated	q	Echo	Sign	Cell
Payoff Safety†	`safety_recognition`	0.0 → 0.0	—	3.20 → 1.95	0.068*	clean	−−−	bids-only
Worst-case scaffold†	`worst_case`	2.0 → 43.0	0.020	3.24 → 3.07	0.93	partial	−+−+	language-only
Payoff Tree†	`second_price_mention`	33.0 → 32.7	0.91	3.20 → 1.29	0.068*	clean	−−−	bids-only
Dominated	`dominance_language`	0.0 → 20.5	0.035	3.24 → 2.88	0.93	partial	−+−+	language-only
Enumerate	`opponent_modeling`	29.1 → 39.3	0.84	3.24 → 3.28	0.93	partial	−+−+	neither
Menu restatement†	`second_price_mention`	33.0 → 0.0	0.015	3.20 → 3.55	0.93	by constr.	+−+	neither
Clock-framing	`clock_exit_language`	0.0 → 19.5	0.023	3.20 → 2.24	0.59	manip.	−+−	language-only
Two-stage clock-exit description	`clock_exit_language`	0.0 → 14.3	0.015	3.20 → 3.41	0.93	manip.	++−	language-only
First-order beliefs	`opponent_modeling`	27.8 → 42.7	0.015	3.16 → 3.93	0.027*	partial	++++	both
Second-order beliefs	`opponent_modeling`	27.8 → 51.3	0.020	3.16 → 3.87	0.36	partial	+−++	language-only
Common-knowledge beliefs	`opponent_modeling`	27.8 → 11.8	0.27	3.16 → 2.13	0.027*	partial	−−−	bids-only
Backward induction	`expected_value_reasoning`	4.4 → 4.0	0.84	3.20 → 2.23	0.36	partial	+−−	neither
Risk-averse persona	`risk_language`	29.0 → 99.3	0.015	3.20 → 4.71	0.027*	partial	++++	both
Risk-neutral persona	`risk_language`	29.0 → 46.8	0.30	3.20 → 2.57	0.23	partial	−−−	neither
Risk-seeking persona	`aggressive_language`	5.8 → 65.5	0.023	3.20 → 3.49	0.93	partial	+−−	language-only
Loss frame	`risk_language`	32.1 → 35.8	0.84	3.13 → 3.73	0.93	partial	−+++	neither
Gain frame	`risk_language`	32.1 → 26.5	0.84	3.13 → 3.06	0.93	partial	+−+	neither
Endowment	`risk_language`	32.1 → 35.2	0.84	3.13 → 4.00	0.10	partial	−+++	neither
Mixed frame	`risk_language`	32.1 → 45.2	0.015	3.13 → 3.68	0.93	partial	−+−+	language-only
WTA/WTP	`risk_language`	32.1 → 30.8	0.91	3.13 → 2.46	0.59	partial	−+−	neither

Table 6: The full set of interventions, pooled over the four families (sealed second-price cells). Primary contrasts (*Payoff Safety*, *Payoff Tree*, worst-case scaffold, menu restatement) are classified at their raw wild-cluster-bootstrap p-values (see text); all other rows at the full-set q shown. *: significant by the run-level wild-cluster bootstrap but soft under the four-model-cluster permutation test. *Echo*: "clean" means the targeted vocabulary does not appear in the prompt; "partial" means part of the language response is the prompt's own vocabulary read back; "by constr." means the menu wording removes the targeted vocabulary by construction, so its language column is not a comprehension measure; "manip." means the language *is* the manipulation (the two clock-framed descriptions), so it is not a comprehension measure either. †: primary contrast. *Sign*: the direction of the $\lvert b - v \rvert$ change in Claude, Gemini, Gemma, GPT-4o ($-$ improves, $+$ worsens; `dissociation_per_family.csv`, auction-level clusters, descriptive since there is one run per model $\times$ cell). *Cell*: joint-response class. Every quantity is stable on the deduplicated pool.

features as mediators (`results/traces/mediation/mediation_bounds_all.csv`). For *Payoff Safety* the first stage is null on every measure: the invariant the description asserts is echoed in 0 of 600 treated plans (rule-of-three one-sided 95% upper bound on prevalence 0.5%), correct payment-rule rehearsal moves from 3.9% to 2.8% (TOST-equivalent to no change within ± 5 points at both run- and auction-level clustering), and rule name-dropping from 33.0% to 28.2% ($p = 0.61$). The average causal mediation effect has a median of 0.02 (95% CI $[-0.17, 0.15]$) against a total effect of -1.25 on $|b - v|$, so the mediated share has a point estimate of roughly zero and a 95% upper bound of 35.4% in absolute value under the decomposition. This is conditional on its mediation assumptions and the chosen verbal indicators, rather than a bound on unobserved understanding. For the *Payoff Tree* the first stage is -0.4 percentage points and the bound is 14.7%. For the worst-case scaffold the first stage is large ($+40.9$ points of worst-case language) but the total effect is null (-0.17), and within the treated group plans that articulate the worst case err no less than plans that do not (gap $+\$0.07$). Two cross-sectional associations are reported for completeness and are not evidence of mediation: within the *Payoff Safety* cell the minority of plans that do rehearse the payment rule err $\$1.40$ less (cluster-bootstrap 95% CI $[-2.15, -0.56]$), and in the trace-aligned logs Claude's *Payoff Safety* cell has a mean signed deviation of $+0.30$ (mild overbidding), the value used in Section 4.2.

C.7 Robustness and Measurement Caveats

Byte-identical duplicates. Roughly 40% of sealed-bid plans are byte-identical duplicates within model $\times$ cell $\times$ value draw: at temperature 0.5 the models are near-deterministic conditional on the value, and every duplicate group shares a single value and a single bid. Full-sample standard errors are therefore optimistic, and the effective sample per cell is closer to the number of distinct value draws than to the row count. Re-running the pooled regression on deduplicated data ($n = 8{,}839$) leaves every headline coefficient stable (`shading_intent` $-2.06 \to -1.90$, `overbid_intent` $-2.18 \to -2.17$, `truthful_intent` $-1.50 \to -1.46$, `payment_rule_correct` $-0.77 \to -0.77$, `conservative_language` $+0.76 \to +0.72$; all $p \leq 0.001$), and every intervention keeps its class.

Cluster counts. The harmonized grid provides 92 run-directory clusters, but the menu and clock-framing cells are stored as one flat directory per model $\times$ cell, so pooled contrasts involving them rest on 8–12 clusters and their run-level p-values should be read accordingly; the four-model-cluster permutation test is the conservative cross-check, and effects it does not support are flagged soft. Per-family intervals use auction-level clusters and are descriptive, since there is one run per model $\times$ cell.

Negation. The frozen intent regexes have no negation handling. This is harmless for the main panel, whose plans almost never discuss shading in order to reject it, but frontier derivations do ("shading only increases the chance of losing"), which fires `shading_intent`: 198 of GPT-5's 558 intent-stated plans are scored as direction mismatches this way, yet 100% of them bid within $\pm\$0.50$ of value. Frontier direction-consistency figures are therefore artifacts of this blind spot; the dollar-level fidelity figures (GPT-5 and GPT-5 mini state and submit the same dollar amount with exact match 1.00) are not.

Clock logs. Clock logs store the bidding history in exit order, not agent order; the corpus build realigns each rationale by agent name and flags the winner, whose recorded price is censored at the runner-up's exit. An earlier positional parse exaggerated frontier clock error by an order of magnitude; any clock number computed without the alignment is invalid. With the alignment, 34–85% of main-panel clock plans state the threshold rule verbatim enough for the strict regex (Claude 47%, Gemini 34%, Gemma 41%, GPT-4o 85%), losing bidders' exits sit within \$0.33–\$1.16 of value, and no bidder in any main-panel or frontier family stays past its value (`results/traces/clock_h6.csv`).

C.8 The Frontier Boundary

The plans also mark a boundary for our main panel. We excluded frontier reasoning models from the panel on the grounds that raw reasoning ability may mask the bottlenecks our interventions

target (Section 2); their plans show what that masking looks like. Scoring 5,985 plans from GPT-5, GPT-5 mini, Claude Sonnet 5, and Gemini 2.5 Flash with the same dictionary plus one additive feature group for formal derivation, a formal dominance argument—"bidding your true value is weakly dominant; overbidding risks winning when the second-highest bid exceeds your value"—is the primary script in 55%, 36%, 43%, and 12% of their sealed plans respectively, against at most 1.2% in any family of our panel, and the judge's "states why truthful bidding is safe" flag fires on 45% of frontier plans against 2.7% of ours. In these plans the decision is written up as a solved mathematics problem, a pattern that invites comparison with reasoning-oriented post-training studied in other models (OpenAI, 2024b; DeepSeek-AI, 2025). This comparison does not identify the training history or cause of the pattern in the models tested here. That mode is a distinct channel, not a guarantee of reliability: conditional on a formal derivation GPT-5 and GPT-5 mini have zero bid error, but Claude Sonnet 5's error conditional on the same mode is \$1.55, driven entirely by 145 plans that formally derive the *wrong mechanism's* equilibrium—"the symmetric Nash equilibrium bidding strategy in a first-price auction is to bid $(n-1)/n$ of my value"—concentrated in the menu-restatement cell, which collapses Sonnet 5 from exact truthfulness to a mean $|b-v|$ of 8.7 (and Gemini 2.5 Flash to 2.2) while GPT-5 and GPT-5 mini are immune. A formally correct derivation of the wrong mechanism is the frontier's version of the script problem. The dictionary's negation blind spot (Section C.7) qualifies frontier direction-consistency figures but not the dollar-level results below.

Table 7 quantifies the shift in decision script between the main panel and four frontier models, scored with the frozen dictionary plus one additive formal-derivation feature group (equilibrium formula, dominance proof, environment recital). Formal dominance derivation is the primary script in 55% of GPT-5's, 43% of Claude Sonnet 5's, 36% of GPT-5 mini's, and 12% of Gemini 2.5 Flash's sealed second-price plans, with a further 26–52% informal truthful assertion, against at most 1.2% formal derivation in any main-panel family. Conditional on a formal derivation, GPT-5 and GPT-5 mini have zero bid error; Claude Sonnet 5's conditional error is \$1.55, driven entirely by 145 plans that formally derive the *first-price* equilibrium ("bid $(n-1)/n$ of my value"), with a mean deviation of $-\$8.81$. Those plans are concentrated in the menu-restatement cell, which collapses Sonnet 5 from exact truthfulness to mean $|b-v|$ of 8.7 ($p = 0.0005$, auction-level clusters) and Gemini 2.5 Flash from 0.12 to 2.24, while GPT-5 and GPT-5 mini are unaffected. Description wording matters at the frontier in the clock as well: Claude Sonnet 5's exits are four times farther from value under the affiliated-values wording of the clock prompt than under the clean IPV wording (mean $|b-v|$ 2.12 versus 0.49; `results/traces/clock_h6_by_experiment.csv`). The other primary interventions have nothing to improve at the frontier: baseline $|b-v|$ is 0.03 pooled, and *Payoff Safety*, *Payoff Tree*, and the clock-framed descriptions leave it there.

Model	Corpus	Plans	Formal (%)	Truthful (%)	$\lvert b-v\rvert$, formal	$\lvert b-v\rvert$, other	Wrong mech.
Claude Sonnet 5	frontier	1,650	42.6	28.0	1.55	0.00	145
Gemini 2.5 Flash	frontier	1,608	11.8	25.8	0.20	0.61	9
GPT-5	frontier	603	54.6	26.9	0.00	0.00	0
GPT-5 mini	frontier	1,374	35.7	51.9	0.00	0.00	0
Claude 3.5 Haiku	main panel	3,513	0.7	8.3	1.59	2.27	11
Gemini 2.0 Flash	main panel	3,486	1.2	1.6	3.68	1.51	1
Gemma 3 27B	main panel	3,474	0.0	2.8	—	5.53	0
GPT-4o	main panel	3,600	0.0	0.7	—	3.01	0

Table 7: Decision scripts at the frontier versus the main panel, sealed second-price plans (`results/traces/formal_derivation_prevalence.csv`). "Formal" is the share of plans whose primary script is a formal dominance derivation (H7); "Truthful" the share with an informal truthful/dominance statement (T); "$|b-v|$, formal/other" the mean absolute deviation conditional on a formal derivation or on its absence. "Wrong mech." counts plans that formally derive another mechanism's equilibrium (for Claude Sonnet 5, the first-price Nash bid). Conditional $|b-v|$ is in dollars; "—" means no plan of that kind.

Two further frontier facts bear on the main text. First, at the frontier the judge's "states why truthful bidding is safe" flag fires on 45% of plans (Section C.3), so the dissociation of Section 5.1 is a statement about the main panel: models that already articulate the safety argument have nothing for *Payoff Safety* to add, in words or in bids. Second, two exploratory frontier cells that test a *false* safety statement (`pilot_false_safety_up`/`_down`) are computed but remain gated pending co-author review and are not part of any claim in this paper.

C.9 The Clock: The Environment Does the Work

The ascending clock offers a clean check on the same dissociation from the other side. Between 34% and 85% of the clock plans state the exit-threshold rule verbatim enough for a strict pattern (the LLM judge reads essentially all clock rationales as this rule), exit prices sit within \$0.33–\$1.16 of value for losing bidders in every family, and no bidder in any family stays materially past its value. But exit fidelity is equally good whether or not the rule is stated: for Claude, mean $|b - v|$ is 0.27 among plans that articulate the threshold and 0.43 among those that do not.[2] The extensive form, not the verbalization, carries the improvement of Section 3: the clock collapses every family onto one script (Figure 21) and makes that script safe to follow whether or not the agent can say why.

D Additional Experimental Details

This appendix collects the environment details of Section 2 that are not needed to follow the contingent-reasoning and mechanism-description results; prompt skeletons are in Appendix A.

D.1 Deferred Acceptance: Market Instances, Priorities, and Protocols

Market instances. Each instance contains students $i \in \{A, B, C, D\}$ and schools $s \in \{w, x, y, z\}$, with unit capacity (one seat per school). This 4×4 set-up is large enough to admit meaningful strategic considerations (competition for popular schools and priority asymmetries) while remaining small enough that we can log the full DA trace and audit every outcome.

Student valuations. To induce correlated demand across students while preserving idiosyncratic tastes, we use an affiliated value model with a school-level common component and a student–school private shock:

$$v_{i,s} \;=\; c_s \;+\; \varepsilon_{i,s},$$

for school s and student i. For each school s, we draw a common value $c_s \sim \mathrm{Unif}[40, 70]$. For each pair (i, s), we draw an independent private shock $\varepsilon_{i,s} \sim \mathrm{Unif}[0, 20]$ and set $v_{i,s} = c_s + \varepsilon_{i,s}$. Each student's *true preference ranking* $\pi_i^\star$ is then defined as the descending sort of $\{v_{i,s}\}_{s \in \{w,x,y,z\}}$ (ties broken deterministically) (Klijn et al., 2019).

School priorities. Schools rank students by fixed, Ergin-acyclic priority orders (held constant across repetitions). Concretely, we use:

$$\begin{aligned} w &: A \succ B \succ C \succ D \\ x &: B \succ A \succ C \succ D \\ y &: A \succ B \succ D \succ C \\ z &: B \succ A \succ D \succ C, \end{aligned}$$

where $\succ$ denotes higher priority. This structure creates a "top tier" $\{A, B\}$ and "bottom tier" $\{C, D\}$ and satisfies the unit-capacity acyclicity condition. This acyclicity is the condition under which a sequential-query implementation can make DA's incentives more transparent (OSP).

Social information (common signal). In addition to their school priorities, all students observe a fixed "global popularity" ranking intended to proxy common beliefs about demand:

$$\text{Global ranking: } y \succ x \succ w \succ z.$$

This signal is held fixed across repetitions to keep the information treatment constant; depending on the realized $v_{i,s}$, it may be aligned or misaligned with the instance's true preference.

[2] Clock logs record the bidding history in exit order rather than agent order; the corpus realigns each rationale by agent name and treats the winner's price as censored at the runner-up's exit. An earlier positional parse overstated clock error by an order of magnitude.

Mechanisms. For each market instance we run two elicitation protocols.

(i) Direct revelation (static DA). Each student submits a complete rank-order list over $\{w, x, y, z\}$ in a single shot. We then run student-proposing DA on the submitted rankings and with the true school priorities.

(ii) Iterative DA (sequential queries). Students do not submit a full ranking upfront. Instead, the mechanism queries students sequentially, following the tree construction of Ashlagi & Gonczarowski (2018) for OSP implementation of DA under acyclic priorities. At each node, the mechanism identifies students who hold top priority at some remaining school and asks them a yes/no question: "Among the remaining schools $\{w, x, y, z\}$, is school s your most preferred?" If the student answers YES, they are immediately matched to that school and both are removed from consideration. If NO, the school is not immediately removed and the mechanism continues to the next query; it may be assigned to another student before that student acts again. When only one top-priority student remains, the protocol reduces to serial dictatorship: the student simply picks their most preferred school from the remaining set. The protocol terminates when all students are matched.

Repetitions and logging. For each treatment condition and each mechanism, we target $N = 50$ independent repetitions (distinct random seeds for value draws; priorities and global ranking fixed). For every repetition we log: true rankings, the full set of LLM responses (raw text), parsed decisions, DA traces (proposals/holds/rejections by round), final matches. Successful instances produce JSON records of value generation and matching outcomes. Retained cell sizes can be smaller after response or parsing failures; the sequential-query sample in Table 8 uses the retained logs.

D.2 Auctions: Environment and Mechanisms

Environment: Independent Private Values (IPV). We study symmetric IPV auctions with $N = 3$ bidders. In each auction instance, each bidder $i \in \{1, 2, 3\}$ receives an independent private value

$$v_i \sim \text{Unif}\{0, 1, \ldots, V_{\max}\},$$

with $V_{\max} = 49$ in the baseline configuration. Values are drawn independently across bidders. All monetary amounts lie on a discrete grid with increment $\Delta = \$0.01$ for Sealed-bid and $\Delta = \$0.5$ for clock auctions.

Mechanisms. We implement two canonical one-shot mechanisms that differ in their strategic interface.

(i) Second-Price Sealed-Bid (SPSB). Each bidder simultaneously submits a bid $b_i \in \{0, \Delta, 2\Delta, \ldots\}$. The highest bidder wins the item and pays the second-highest bid. Let $b_{(1)} \geq b_{(2)} \geq b_{(3)}$ denote the order statistics of bids. The winner is the argmax bidder and the price is $p = b_{(2)}$. The winner's payoff is $v_i - p$, and losers receive 0.

(ii) Ascending Clock Auction (closed). The auction starts at price $p_0 = 0$ and increases by Δ each tick. At each posted price, each active bidder is asked whether they want to *stay* in the auction or *exit*. Exit decisions are private: bidders are not informed when others leave. The auction ends when only one bidder remains. We record each bidder's *dropout price* d_i as the first posted price at which they choose to exit. The winner is the bidder with the highest dropout price, and the transaction price is the *second-highest* dropout price (equivalently, the price at which the second-to-last bidder exits). The winner's payoff is $v_i - \max_{j \neq i} d_j$, and losers receive 0.

E Additional Results

This appendix supplies the comparisons summarized in the main text: human benchmarks for the interface changes, followed by prompts that ask agents to plan through matching rounds or reason about other participants. The tasks and scoring remain those of Section 2; lower absolute bid deviations and lower ranking errors indicate better play.

E.1 Interface Comparisons: Human Benchmarks and Matching Error Measures

It is well known that truthful bidding is the dominant strategy in the SPSB auction (Vickrey, 1961), yet human subjects systematically deviate from this in laboratory experiments (Kagel & Levin, 1993; Kagel, 1995). With regards to OSP mechanisms, Li (2017) provides experimental evidence that an obviously strategy-proof clock format improves human bidding. Breitmoser & Schweighofer-Kodritsch (2022) decomposes features of ascending auctions: seeing the ascending clock and receiving dropout information improve bidding, while dynamic bidding itself has no significant incremental effect. These findings motivate separating interaction format, information, and description rather than attributing every clock-format gain to OSP alone.

In particular, if agents fail to appreciate that DA is strategy-proof, they may mistakenly misreport their rank-order list of preferences in seeking to improve their outcome given the reports of others. This misreport, then, is a function also of their beliefs of other agents' choice of what reports to make, and thus also their beliefs. To probe this, and also for realism, our implementation also reports to applicants that "most other applicants seem to favor $global_ranking$." Such beliefs, of course, should not change an applicant's report in DA if they understand the mechanism is strategy-proof. We test this question in Appendix E.3; full prompts are in Appendix A.

Ranking summaries and decision errors. A complete submitted ranking reveals all six school pairs in a four-school market. Sequential responses reveal less: a single selected school has no rankable pair, and a false yes/no answer may not appear in the reconstructed ranking at all. The plotting pipeline assigns zero Kendall distance to a sequence with fewer than two ranked schools. Thus the zero recorded iterative statistic in Figure 1 is not evidence that every response was truthful.

A separate analysis of the recorded sequential-query histories (the `osp_yesno_fixed` condition) scores each binary answer and each nonforced final choice against the values available at that node. NO is an error if the offered school is the unique highest-valued option; YES is an error if a strictly better school remains. A final choice is an error if its value is below the maximum available value. Choices between equal-valued best schools are accepted, and final choices with only one available school are excluded from the informative denominator. The resulting decision-level rates are shown in Table 8; they are not directly comparable to full-ranking Kendall distances.

Model	Informative decisions	Errors	Error rate	False NO / YES
Claude	194	16	8.2%	13 / 2
Gemini	179	0	0.0%	0 / 0
GPT-4o	188	0	0.0%	0 / 0
Gemma	210	3	1.4%	3 / 0

Table 8: Decision errors in the sequential-query DA logs. Informative decisions include yes/no answers and final choices among at least two schools. The last column counts false rejections and false acceptances; Claude's remaining error is a final-choice error. These are descriptive rates: decisions from the same market are dependent, and zero observed errors does not establish a zero population error rate.

The replication files also contain an earlier round-based choice protocol, `osp_baseline`, which should not be conflated with the binary-query protocol. It has zero recorded Kendall distance for all four models, but decision-level scoring finds five strict-value choice errors among Gemma's 322 informative decisions and zero among the other three models' 346, 351, and 325 decisions. This further illustrates why zero partial-ranking distance does not imply zero decision errors. These existing-log checks change the interpretation of the sequential comparison, not the complete-ranking treatment comparisons (4.2% baseline, 1.7% payoff tree, 0.2% rejection safety).

E.2 Forward Planning

The next axis we consider is that of forward planning. Pycia & Troyan (2023) formalize a notion of limited forward planning that has become central to the theory of simple mechanisms. In a *one-step simple* mechanism, an agent can identify the prescribed action by considering at most one of her own future moves at a time. In the ascending clock, for example, staying now can be justified by

a plan to exit at the next opportunity; that plan can be updated when the next decision arrives. The agent need not plan the entire future sequence of prices.

Both SPSB and DA are static mechanisms, with only a single decision node for each agent. But DA's underlying algorithm processes the submitted ranking dynamically, through multiple rounds of proposals and rejections, giving forward planning a natural place to bind: models may attempt to reason through the algorithm's rounds when deciding what ranking to submit. The sealed-bid auction has no analogous internal dynamics—a bid is submitted and the outcome is determined. In this sense, we think of DA as a dynamic *algorithm*, and so test the forward planning interventions only on it. We test two 'lookahead' scaffolds that prompt models to simulate the algorithm forward.

- *One-Step Lookahead*: "If you are rejected from your first choice, which school would you propose to next? Does this affect how you should rank schools?"
- *Two-Step Lookahead*: "Consider the first two potential rejections. If rejected from your first choice, you go to your second. If rejected again, you go to your third. Plan your ranking with these steps in mind."

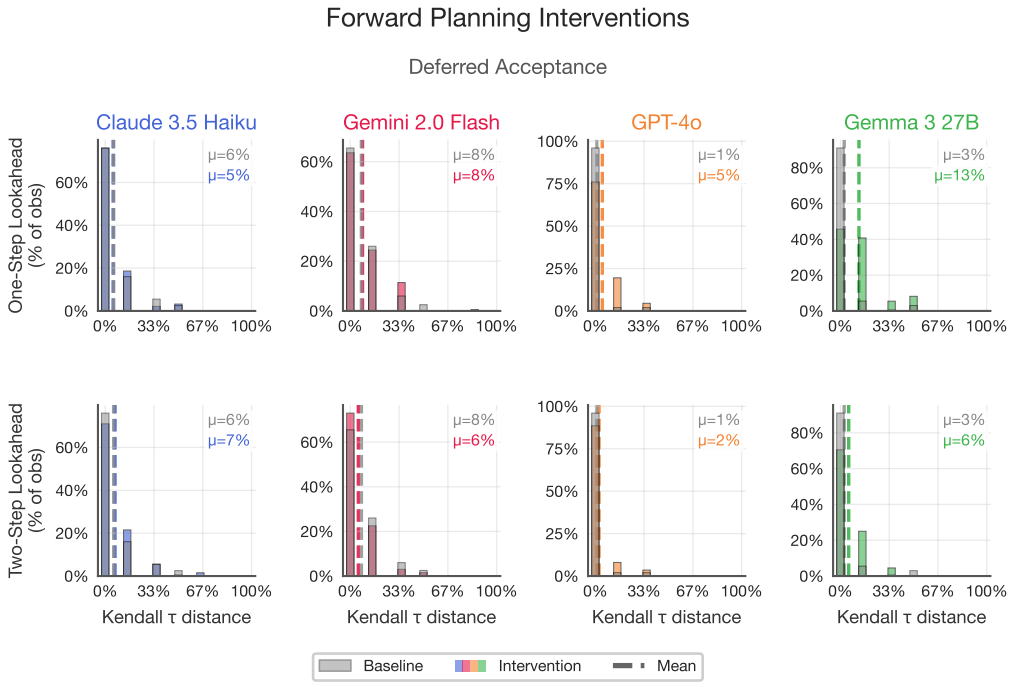

Figure 23: Forward planning: lookahead interventions in deferred acceptance. *Top row:* One-step lookahead; *bottom row:* two-step lookahead. Each cell overlays the direct DA baseline (gray) with the intervention (colored). Lookahead scaffolds worsen play overall, with the damage concentrated in GPT-4o and Gemma; Claude and Gemini are largely unaffected.

While contingent reasoning was quite successful across models and mechanisms, it appears that forward planning scaffolds are less effective in improving play. The *One-Step* intervention worsens play in DA from $\mu = 4.2\%$ to 7.8% across models; *Two-Step* worsens it to 5.1%. The damage is concentrated in the last two rows of GPT-4o and Gemma: *One-Step* moves GPT-4o from 1.0% to 4.8% and Gemma from 2.6% to 12.7% error, whereas Claude and Gemini are largely unaffected.

Why does prompting forward planning hurt? Our hypothesis is that it activates strategic thinking about the algorithm without resolving it. If models recognized DA is strategy-proof, forward planning would be irrelevant—the dominant strategy does not depend on what happens after a rejection. But because models do not fully appreciate strategy-proofness, prompting them to simulate rejection dynamics gives them material to strategize with, second-guessing their truthful preferences.

The weaker models, with less ability to resolve the reasoning they have been prompted to undertake, suffer the most.

E.3 Belief Formation

The third axis is belief formation. One canonical story here is from Börgers & Li (2019), who introduce strategic simplicity: a mechanism is said to be *strategically simple* if optimal play does not depend on higher-order beliefs about what opponents believe. Both SPSB and DA are strategy-proof, and hence trivially strategically simple: the dominant strategy is independent of beliefs entirely. Scaffolding belief formation in these settings should therefore be irrelevant at best. But if models are already treating these mechanisms as Bayesian games, and attempting to best-respond to perceived competition rather than playing the dominant strategy, then making beliefs more salient could amplify mistakes.

We test two levels of belief scaffolding, applied to both mechanisms.

- *First-Order Beliefs*: "What do you expect the other players to do? Given their values are drawn randomly and they want to maximize their earnings, how do you think they will behave? Does this affect what you should do?"
- *Second-Order Beliefs*: "What do the other players think YOU will do? They know you are rational. They might try to anticipate your strategy. Does their belief about your behavior affect what they will do? And does that, in turn, affect what you should do?"

Indeed, we find that belief scaffolding almost uniformly worsens play across both mechanisms, and considering second-order beliefs hurts play marginally more. In auctions, *Second-Order Beliefs* worsens the mean bid across models from $\mu = -2.67$ to -3.47; *First-Order Beliefs* worsens it to -3.04. This effect is consistent for three of four models—only Gemini is unaffected. In absolute deviation the picture is sharper: under *First-Order Beliefs* all four models, Gemini included, move farther from value, while under *Second-Order Beliefs* Gemini alone moves closer (Section 5.1). Models prompted to think about opponents shade further below value, undercutting perceived competition. The manipulation works as intended on the plans themselves: opponent-modeling language rises from 28% of baseline plans to 43% under first-order and 51% under second-order prompts.

In DA, both interventions worsen play: *First-Order Beliefs* moves from $\mu = 4.2\%$ to 5.1%, *Second-Order Beliefs* to 9.1%. Again, three of four models get worse—Gemini is the exception, showing small improvements. The damage is particularly severe for Gemma, which moves from 2.6% to 14.3% under *Second-Order Beliefs*.

Interestingly, Gemini is almost completely immune to belief scaffolding while the other tested models are hurt. In contrast, Gemma—which suffers the most severe damage under Second-Order Beliefs (2.6% to 14.3% in DA)—appears most susceptible to this failure mode. These results suggest that LLM reasoning styles are not monolithic; the heterogeneity across model families in susceptibility to belief-based reasoning may itself be relevant to mechanism design when markets serve diverse artificial agents.

F Additional Related Work

The main text uses economic settings to study a broader question about supporting LLM decisions. This appendix explains the theories that motivate the interventions and relates the experiments to prior work on human and artificial agents.

Complexity and Apparent Irrationality Recent work emphasizes the connection between strategic complexity and apparent irrationality. Oprea (2024) argues that many anomalies attributed to non-standard preferences are better explained by computational complexity: subjects make mistakes not because their preferences are irrational but because the problem is hard. This framing aligns with our approach. We do not assume LLMs have "biases" in the behavioral economics sense (though we explore interventions probing this in the Appendix); rather, our primary question is whether the computational structure of strategic problems creates cognitive challenges and failure modes for LLMs.

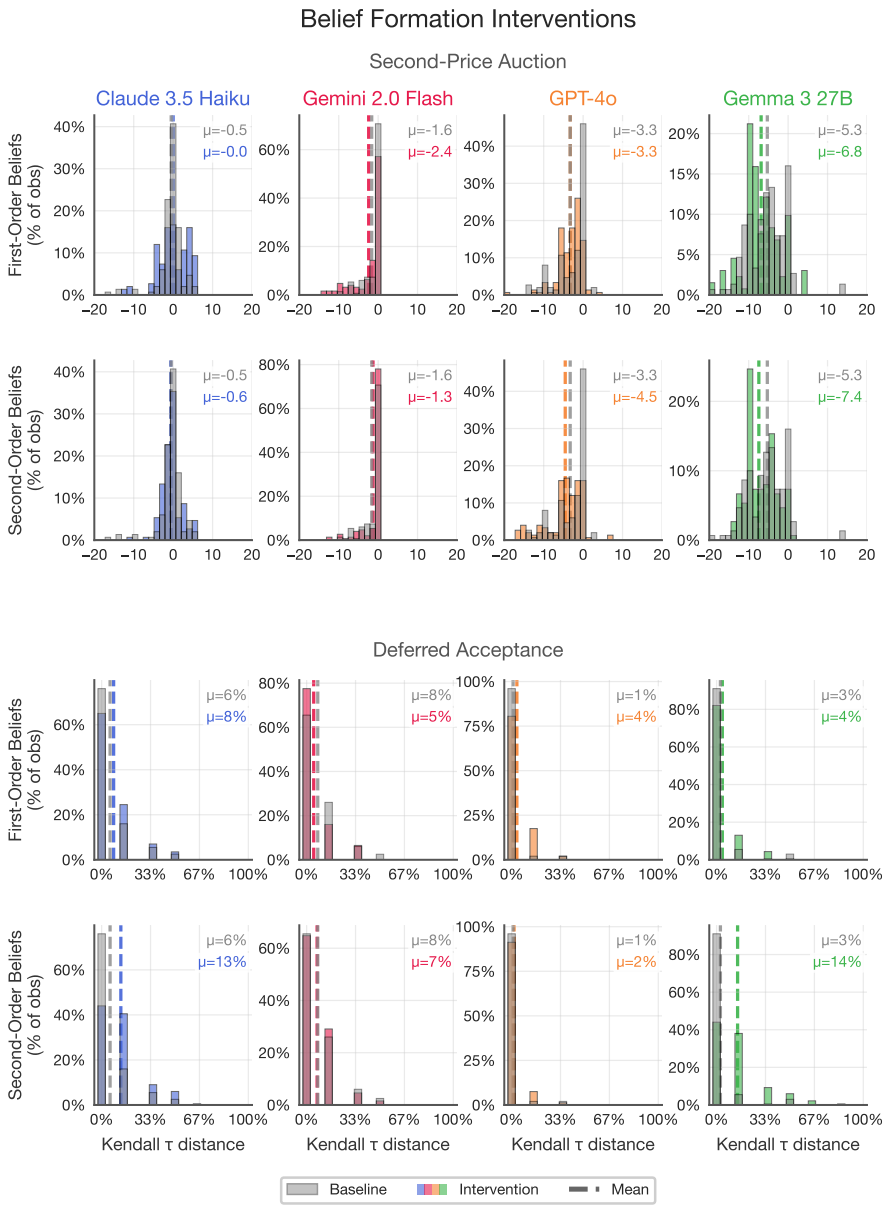

Figure 24: Belief formation interventions across both mechanisms. *Top two rows:* Second-price auction bid deviations under first-order and second-order belief scaffolds. *Bottom two rows:* Deferred acceptance Kendall τ errors under the same scaffolds. Baseline distributions in gray, interventions in color. Belief scaffolds almost uniformly worsen play, with second-order beliefs causing the most damage. Gemini is the notable exception, showing near-immunity to belief scaffolding.

Deferred Acceptance There is a deep literature in market design for both auctions and two-sided matching. For two-sided matching, Gale & Shapley (1962) introduced the DA algorithm, which produces stable matchings, and Roth (1982) showed that DA is strategy-proof for the proposing side.

The question of simplicity in DA was addressed first by Ashlagi & Gonczarowski (2018), who show that DA is generally not OSP-implementable. However, under *acyclic* priority structures—a condition introduced by Ergin (2002)—the proposer-optimal stable matching rule can be implemented in an OSP manner (for the proposer). For unit-capacity schools, the relevant acyclicity condition rules out three students a, b, c and two schools s, t with $a \succ_s b \succ_s c$ and $c \succ_t a$. Under this condition, an OSP implementation of student-proposing DA exists.

Gonczarowski et al. (2023) study descriptions of strategy-proof mechanisms that aim to make incentive properties transparent. They give a menu-based description for DA: the set of schools a student can obtain as it depends on others' reports, where the student's ranking is used to select within this menu. Such descriptions may help participants understand strategy-proofness without changing the extensive form—a complementary approach to OSP implementation. In an incentivized lab experiment, Gonczarowski et al. (2024) find that a menu-based explanation significantly improved participants' measured understanding of DA's strategyproofness relative to standard descriptions, though average behavioral effects were modest.

Further Connections The following work places the interface comparisons in the broader literatures on simplicity and LLM decision making.

Simple Mechanism Design Li (2017) formalized obvious strategy-proofness and showed that ascending clock auctions satisfy it while SPSB auctions do not. Related notions include *strategic simplicity* (Börgers & Li, 2019), which requires that optimal play not depend on higher-order beliefs, and *one-step simplicity* (Pycia & Troyan, 2023), which lets agents identify the prescribed action while considering at most one of their own future moves at a time. The SPSB auction satisfies strategic simplicity—truthful bidding is dominant regardless of beliefs about others—but is not OSP. Ascending clock auctions satisfy both. Li (2024) decomposes mechanism complexity into three cognitive dimensions: contingent reasoning, forward planning, and belief formation. We use this decomposition to structure our interventions.

LLMs as Economic Agents An emerging literature studies LLMs as simulated economic agents. Horton (2023) introduces the "homo silicus" framework, showing that LLMs can replicate qualitative patterns from classic experiments in labor economics, bargaining, and social preferences. Aher et al. (2023) demonstrate that LLMs can reproduce human subject study results across a range of paradigms. Shah et al. (2024) study LLMs in auction settings and find that OSP mechanisms improve play, a finding we replicate and extend in this paper.

In auction settings specifically, Chen et al. (2023) study LLMs in a multi-round auction environment and find substantial deviations from optimal play. Fish et al. (2024) examine whether LLMs can sustain collusive outcomes in repeated auctions. Our contribution differs in focus: we use the conceptual vocabulary of simple mechanism design—contingent reasoning, forward planning, belief formation—to probe and characterize the cognitive constraints that bind on LLM agents, and to ask whether the levers the same theory supplies for bounded human reasoners also move theirs—and whether behavioral improvement is accompanied by improvement in what the agents articulate.

Heuristics and the Faithfulness of Stated Reasoning Decision quality depends jointly on a procedure and the environment in which it is used (Simon, 1956). Heuristics are selective rules that can work well in one setting and poorly in another (Gigerenzer & Gaissmaier, 2011; Shah & Oppenheimer, 2008). Our labels describe such rules in agents' stated plans: bidding below value to preserve a margin, for example, is appropriate to a first-price auction but misplaced in a second-price one. Because explanations need not faithfully report a model's computation (Turpin et al., 2023; Lanham et al., 2023), we evaluate these statements alongside realized bids rather than treating them as direct evidence of internal reasoning.