Engineering Simplicity for LLM Agents *

Kehang Zhu* Harvard
Anand Shah* MIT
David C. Parkes Harvard

May 4, 2026

Abstract

The theory of simplicity in economic design makes a claim about minds: that certain strategic problems are intrinsically hard to reason about, and that simple design can improve reasoning and recover good outcomes. Large language models (LLMs) represent a new kind of intelligence—neither human nor superhuman—and offer a rare opportunity: to test whether the cognitive constraints that drive simplicity in economic design are specific to humans or reflect something deeper. Existing results establish that LLM agents make mistakes participating in strategy-proof auctions such as the second-price sealed-bid auction, but improve play in simpler, obviously strategy-proof designs; we extend these findings across model families and to two-sided matching, and take this as evidence that the representation of a mechanism—how a game is presented, not just what it implements—shapes LLM reasoning in the same direction it shapes human reasoning. To understand where these cognitive constraints bind, we systematically test prompt-based interventions along three axes from this theory—contingent reasoning, forward planning, and belief formation—as well as descriptions of the mechanisms themselves to make their incentive properties transparent. We find that scaffolding for contingent reasoning, and making incentive properties transparent, substantially improve play, while prompting models to plan forward or to reason about others' beliefs consistently worsens it. The conceptual vocabulary of simple mechanism design, it appears, also describes the limits of an intelligence we did not build the theory for. Understanding why will matter as artificial agents enter economic life.

1 Introduction

Large language models can write sonnets, prove mathematical theorems, and pass the bar exam to practice law. Yet, LLMs cannot reliably bid their true value in a second-price sealed-bid auction (Zhu et al., 2024). This pattern of failure is puzzling: frontier models demonstrate capabilities that seem to require sophisticated reasoning, while struggling with problems that seem, at least to humans, less demanding (Chollet, 2019).

In particular, it is well understood that truthful bidding is the dominant strategy in the second-price sealed-bid (SPSB) auction (Vickrey, 1961), and the conclusion follows from straightforward reasoning: your bid determines only *whether* you win, not *what* you pay; overbidding risks paying more than your value; underbidding risks losing auctions you would have profitably won. Simple as this logic is, the failure is not unique to LLMs—humans have also routinely failed truthful bidding in the laboratory (Kagel and Levin, 1993; Kagel, 1995). This has motivated a theory of simple mechanisms (Li, 2017), which helped

*Both Anand V. Shah and Kehang Zhu contributed equally.

to identify what makes the SPSB auction hard to play—and in doing so, helped articulate the cognitive constraints that may bind on human reasoners. And while frontier models will likely master such auctions before long—the dominant strategy is textbook material, after all—the failure of current models offers us an opportunity: to apply this burgeoning theory, built to characterize bounded rationality in humans, to a new kind of bounded reasoner. To do so, also introduce a novel methodology of cognitive scaffolds, namely, prompt-based interventions along different dimensions, to identify those cognitive constraints that bind.

Our experiments span two canonical mechanism design settings: auctions and matching. Previous work has demonstrated that *obvious strategy-proofness* (OSP) helps LLMs reason in second-price auctions, changing the format from a sealed-bid to an ascending clock to simplify reasoning (Zhu et al., 2024). We replicate these results across four model families: Claude, Gemini, GPT, and Gemma (open-source). Turning to two-sided matching, we consider the *deferred acceptance* (DA) algorithm, and compare the standard DA—which demands the entire rank-order list up front—to an iterative implementation.[1] DA does not generally admit an OSP implementation (Ashlagi and Gonczarowski, 2018), but does under acyclic priority structures (Ergin, 2002), so we restrict to this environment. We find that LLMs deviate substantially from truthful play in standard DA, but achieve near-perfect play under iterative DA. To our knowledge, this is the first experimental test of an OSP implementation of DA with any kind of agent.

Having established that LLMs exhibit bounded rationality in DA (extending existing results on second-price auctions) we probe deeper to characterize where cognitive constraints bind. The literature offers many notions of what makes a mechanism simple—e.g., OSP (Li, 2017), strategic simplicity (Börgers and Li, 2019), one-step simplicity (Pycia and Troyan, 2023)—and many of these concepts are interrelated. We broadly follow Li (2024), who identifies three cognitive dimensions along which mechanisms may challenge reasoning: *contingent reasoning* (tracing through hypothetical scenarios about opponents' actions), *forward planning* (reasoning about one's own future decision points), and *belief formation* (modeling what others believe). For each axis, we design interventions—prompt modifications that scaffold reasoning along that dimension without disclosing the dominant strategy. Not all axes apply equally to both mechanisms. In particular, both the SPSB and DA are static mechanisms in that the agent makes a single decision, but DA is implemented via a dynamic algorithm that processes rankings through rounds of proposals and rejections. This may give agents something to mistakenly reason forward about. We therefore test forward planning interventions only in DA. Separately, we also test *mechanism descriptions* that surface incentive properties without scaffolding any particular mode of reasoning. Our findings, probing by axis:

- *Contingent reasoning*: Scaffolding the payoff structure of the mechanism—making explicit what happens under each outcome the agent may face—substantially improves play across both mechanisms and all models.
- *Forward planning*: Prompting models to reason about their own future decision points within DA's rejection dynamics worsens play, particularly for weaker models. The scaffold gives models material to strategize with, but not the ability to strategize well.
- *Belief formation*: Prompting models to form beliefs about opponents' strategies—and about opponents' beliefs about theirs—consistently worsens play across both mechanisms, amplifying mistakes from strategizing.
- *Mechanism descriptions*: Describing the logic of why truth-telling is safe in both mechanisms—that your bid determines whether you win, not what you pay (auctions), or that being rejected from a top-ranked school cannot hurt your chances at lower-ranked schools (DA)—produces dramatic improvements, even nearly matching the full OSP benchmark in DA.

[1] Henceforth, DA will refer to the standard implementation in which participants submit complete preference rankings simultaneously, analogous to the SPSB auction. *Iterative* DA will refer to an OSP, sequential-query implementation, analogous to the ascending clock auction.

LLMs will be increasingly called upon to perform economically important tasks. As that happens, it will be all the more important to understand how they reason. If the cognitive constraints that the theory of simplicity was built to address also bind on artificial agents, then the case for engineering simplicity no longer rests on human psychology alone—it becomes a design principle for markets that must work for any kind of reasoning agent. Moreover, interventions as a method generate behaviorally distinct populations of LLM agents. This makes them useful as a way to target particular populations or particular regions of strategy space under a given mechanism, and we are optimistic that this approach may help run more useful, targeted experiments over artificial agents in the future.

The paper proceeds as follows. Following a literature review, Section 2 describes our experimental design and intervention scaffolds. Section 3 presents results comparing OSP and non-OSP mechanisms, replicating prior findings in auctions and extending to matching. Section 4 presents the intervention probes, showing that different axes bind differently across mechanisms and models. Section 6 concludes. The Appendix contains the full text of all prompts for each mechanism in the static, dynamic, and intervention settings, as well as additional results probing prospect-theoretic departures from expected utility theory, including loss aversion and the endowment effect.

1.1 Related Work

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 and Li, 2019), which requires that optimal play not depend on higher-order beliefs, and *one-step simplicity* (Pycia and Troyan, 2023), which requires that optimal play emerge from considering only the immediate next step rather than full backward induction. 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.

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.

Deferred Acceptance There is a deep literature in market design for both auctions and two-sided matching. For two-sided matching, Gale and 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 and 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). Acyclicity rules out configurations where student a has priority over b at school s, while b has priority over c at school s', and c has priority over a at school s''. Under acyclic priorities, rejection chains remain local, enabling an iterative query protocol that is OSP.

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.

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. Brand et al. (2023) use LLMs for market research, finding that they approximate human consumer behavior. Zhu 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 same theory that describes bounded rationality in humans can also help describe theirs.

2 Experimental Design

2.1 Mechanisms

We study two canonical strategy-proof mechanisms, each paired with an OSP counterpart.

Auctions. We implement a second-price sealed-bid (SPSB) auction with $N = 3$ bidders and independent private values, alongside an ascending clock auction that implements the same allocation and payment rule. Human subjects have historically deviated from truthful bidding in SPSB auctions (Kagel and Levin, 1993; Kagel, 1995), while ascending clock formats improve play (Li, 2017; Breitmoser and Schweighofer-Kodritsch, 2022). Details of the auction environment appear in Section 2.3.

Matching. We implement student-proposing deferred acceptance (DA) in a school choice setting with four students and four schools, alongside an iterative DA protocol. We construct priority structures that are Ergin-acyclic—ensuring the iterative protocol is OSP—but with non-aligned preferences, so the matching problem is non-trivial. Details of the matching environment appear in Section 2.2.

2.2 Data Generation for Deferred Acceptance

Overview. We generate data from a school choice environment in which LLMs play the role of students. Each experimental repetition instantiates a complete matching market (student preference rankings, and school priorities), elicits student reports under a specified mechanism, and executes student-proposing Deferred Acceptance (DA) to produce a matching outcome.

Market instances. Each instance contains four students $i \in \{A, B, C, D\}$ and four 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 rules out priority cycles that would otherwise create long rejection chains. 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.

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 and 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 remains available and the mechanism continues to the next query. 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.

LLM elicitation as scenario prompts. Within each repetition, we query one LLM instance per student. Each query is a self-contained prompt that bundles: (a) the student's identity, (b) their priority at each school, (c) the global ranking signal, (d) the mechanism rules, and (e) a treatment-specific instruction cognitive scaffold (if applicable). We implement chain-of-thought (Wei et al., 2022) by requiring structured output with explicit tags for reasoning and decision.

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>`

Repetitions and logging. For each treatment condition and each mechanism, we run $N = 50$ 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. This produces a complete JSON record per instance, allowing exact replay and audit of every step from value generation to matching outcome.

2.3 Data Generation for Auctions

Overview. We generate auction data from an in-silico laboratory in which large language models (LLMs) act as bidders. Each experimental repetition instantiates a complete one-shot [2] auction environment (values, mechanism, and intervention), elicits a sequence of actions from each bidder (a sealed bid or a sequence of stay/exit decisions), and then computes allocations and payments mechanically from the submitted actions. The unit of observation is an *auction instance*: a fully specified scenario that is reproducible given a random seed and a prompt template.

[2] Zhu et al. (2024) showed that LLM agents rarely learn from repeated plays in auctions.

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.

LLM elicitation as scenario prompts. In both mechanisms, each bidder's decision is elicited via a self-contained prompt that includes: (a) bidder identity and opponent list, (b) a mechanism description (which varies by intervention), (c) the bidder's private value v_i, and (d) a structured response format.

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>`

Treatments (prompt interventions). Treatments modify only the rule-explanation block of the prompt while holding the value distribution, bidder identities, and payment rules fixed. We use this to test how different reasoning scaffolds change bidding behavior under the same underlying game (e.g., contingent-reasoning prompts, decision-tree framing, backward-induction analogies, direct strategy revelation, and risk-preference framings).

2.4 Models

We test four models spanning three commercial API providers and one open-weight family: GPT-4o (OpenAI, 2024), Claude 3.5 Haiku (Anthropic, 2024), Gemini 2.0 Flash (DeepMind, 2024), and Gemma 27B (Riviere et al., 2024). All experiments use temperature $T = .5$.

Our goal is to study the interaction between cognitive constraints and mechanism design, which requires models that are capable enough to parse the rules of a mechanism yet still susceptible to the reasoning failures that mechanism design aims to remedy. We therefore deliberately select well-established, widely-deployed model families rather than the most powerful frontier reasoning models (e.g., o3, GPT-5). At the frontier, raw reasoning ability may mask the very cognitive bottlenecks our interventions target, collapsing our ability to identify which design features matter and why.

3 OSP Results

A strategy-proof mechanism and its OSP counterpart implement the same outcome rule—the same mapping from preferences to allocations—but differ in extensive form. The non-OSP version requires *contingent reasoning*: to see that truth-telling is dominant, an agent must trace through hypothetical scenarios about others' behavior and verify that no deviation improves payoffs in any case. The OSP version restructures the game so that the dominant strategy is *obviously* optimal: at every information set, the best outcome from following the strategy weakly dominates the worst outcome from any deviation, without requiring any reasoning about what others might do (Li, 2017). Zhu et al. (2024) showed empirically that LLMs deviate from truthful bidding in SPSB auctions for GPT-4o; we replicate this finding for GPT-4o and extend it across three additional model families and to deferred acceptance. In both domains, OSP mechanisms substantially improve play.

3.1 Auctions: SPSB vs. Ascending Clock

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 and Levin, 1993; Kagel, 1995). With regards to OSP mechanisms, Li (2017) and Breitmoser and Schweighofer-Kodritsch (2022) have empirically showed that clock formats improve human bidding in the second-price auction by increasing distributional mass on value bidding. In the case of Breitmoser and Schweighofer-Kodritsch (2022), this is achieved by merely providing an OSP description of the SPSB, without changing the mechanism itself.

The top column of Figure 1 shows the distribution of bid deviations from value in SPSB auctions. All four models deviate from truthful bidding, with mean deviations of $\mu = -0.5$ (Claude), -1.6 (Gemini), -3.3 (GPT-4o), and -5.3 (Gemma). The negative sign indicates systematic *underbidding*—models bid below their values. This contrasts with the primary finding in human experiments, where subjects tend to *overbid* in second-price auctions (Kagel and Levin, 1993). The distributions also exhibit long tails, which are evidence of very large errors: some bids deviate by \$10–20 from value, representing cases where the model substantially misplays the auction.

Yet, our results also show the ascending clock format corrects deviations toward truthful play across all models.[3] Mean deviations compress to $\mu = -0.1$ (Claude), -1.2 (Gemini), -0.5 (GPT-4o), and -0.3 (Gemma). The improvement is most dramatic for the models with the worst SPSB performance: Gemma's mean deviation falls from -5.3 to -0.3, a near-complete correction. Beyond the mean shifts towards value

[3]The GPT-4o result replicates a finding in Zhu et al. (2024); we extend it to Claude, Gemini, and Gemma.

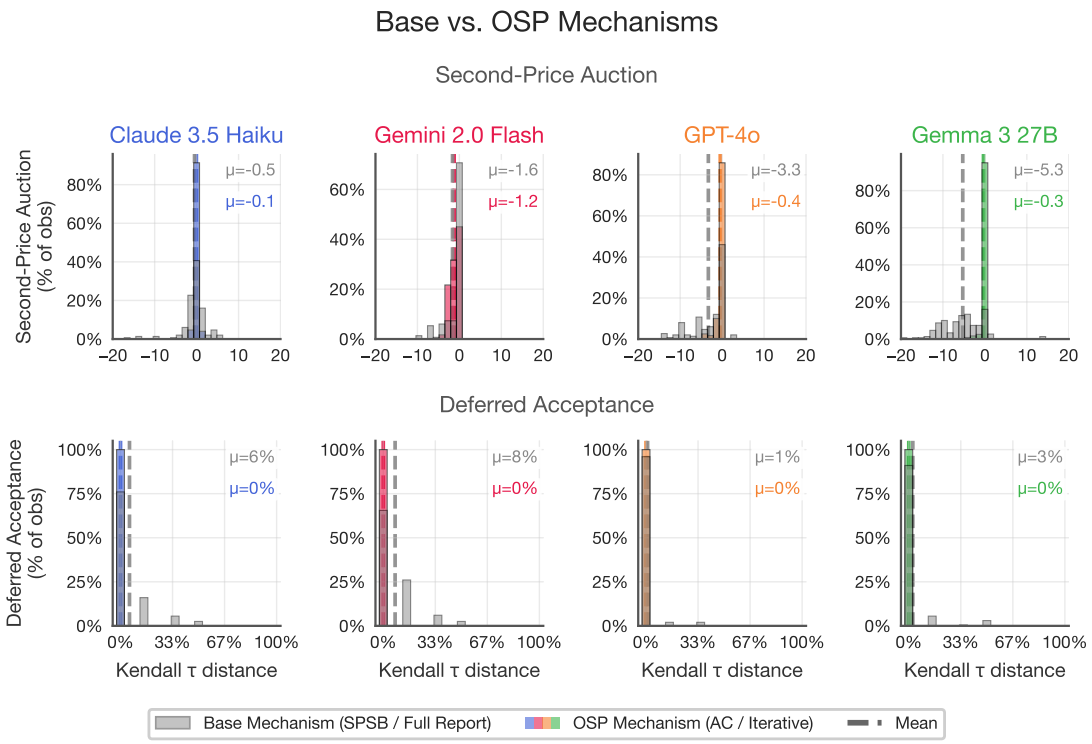

Figure 1: Comparing base (non-OSP) and OSP mechanisms across four model families. *Left column:* Distribution of bid deviations from value (bid − value) in second-price sealed-bid (gray) and ascending clock (colored) auctions. *Right column:* Normalized Kendall τ distance (fraction of pairwise ranking swaps) in direct-revelation DA (gray) and iterative, OSP DA (colored). In particular, in OSP DA, applicants may not be asked to reveal their entire rank-order lists, so we count misreports as from the smaller set of revealed preferences, looking for deviations which are either Type 1 misreports (false rejection, or saying NO when x is the most preferred available school) or Type 2 misreports (false acceptance, or saying YES when a more preferred school is still available). Dashed lines indicate means. All models deviate from truthful play in non-OSP mechanisms; OSP mechanisms correct toward truth-telling, with iterative DA achieving perfect truthfulness ($\mu = 0\%$) across all models.

bidding, the ascending clock eliminates the long tails visible in the SPSB distributions. These catastrophic deviations effectively disappear. This variance reduction may be as practically important as the mean correction: OSP does not merely improve average play, it also eliminates the worst-case failures.

Experiments are run in an independent, private values setting, so the ascending clock format does not provide new information to bidders.[4] This raises a natural question: OSP simultaneously simplifies play along multiple cognitive dimensions—it eliminates the need for contingent reasoning, reduces each decision to a simple binary choice (stay or exit at each price), and makes the connection between action and outcome immediate. Which of these simplifications actually drives the observed improvement in play? We return to this question in Section 4.

3.2 Matching: Direct DA vs. Iterative DA

Student-proposing deferred acceptance is strategy-proof for the proposing side: truthful reporting of preferences is a dominant strategy (Roth, 1982). Yet human subjects routinely misreport in practice. For example, Rees-Jones (2018) documents suboptimal play in the National Resident Matching Program, finding that a nontrivial fraction of applicants submit rankings that are inconsistent with revealed preferences—despite the mechanism being well-known and the stakes substantial. DA is generally not OSP-implementable (Ashlagi and Gonczarowski, 2018), but under Ergin-acyclic priorities (Ergin, 2002), an iterative query protocol can implement the proposer-optimal stable matching in an OSP manner. To our knowledge, no prior work has experimentally tested an OSP implementation of DA with either human or LLM subjects.

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. A probe of this particular question is also made in the next section. Full prompts can be found in the Appendix.

The bottom column of Figure 1 shows the distribution of normalized Kendall τ distance—the fraction of pairwise ranking swaps relative to the true preference order—in direct-revelation DA. All models exhibit nontrivial error rates: $\mu = 6\%$ (Claude), 8% (Gemini), 1% (GPT-4o), and 3% (Gemma). Though these error rates are not directly comparable to the auction deviations (which are measured in dollars), they can be consequential: under deferred acceptance, even small misreports can cascade through rejection chains and alter the final matching for $\Theta(n)$ participants (Dubins and Freedman, 1981).

By contrast, the iterative DA protocol achieves perfect truthfulness: $\mu = 0\%$ for all four models, with zero variance. We found this to be a very surprising empirical result. Unlike the ascending clock in auctions—which substantially improves but does not perfectly correct bidding—the iterative DA format completely eliminates errors. We see two possible explanations. First, the decision space in DA is much smaller than in auctions: in our setting, $4! = 24$ possible rank-order lists versus any $\$0.01$ tick in $[0, 50]$. The smaller action space does not eliminate the need for contingent reasoning—models still err in standard DA—but once the iterative format simplifies the reasoning problem, the coarse action space may make it easier for models to land on the correct answer. Second, and relatedly, the coarse decision space also admits a coarse space for error detection. It may indeed be true that LLMs are making cognitive errors in the iterative format, but we cannot observe them as long as they accurately choose their most preferred school from a set.

[4] In a common- or affiliated-values setting, the clock format could improve play by revealing information about others' values through their dropout decisions. Under IPV, no such information exists, so the improvement must be purely cognitive.

In summary, LLMs deviate from rational play in both SPSB and DA, and benefit from OSP implementations of the same mechanisms—evidence of bounded rationality that the theory of simple mechanisms can speak to. In the next section, we decompose the cognitive dimensions along which OSP simplifies play through targeted interventions in the non-OSP formats.

4 Intervention Results

OSP mechanisms improve play, but they simplify the strategic problem along multiple dimensions at once. To identify which simplifications actually matter, we turn to prompt-based interventions to probe particular aspects of the studied mechanisms.

First, and inspired by Li (2024), we broadly categorize these interventions into three families—thinking **contingently** about unobserved moves by other players, **planning** for their own future moves, and reasoning about other players' **beliefs**. For each intervention, we write prompts that scaffold reasoning without revealing the dominant strategy; concisely, each modifies only the rule-explanation block of the prompt while the mechanism, value distributions, and response format are left unchanged.

Second, an interesting direction in the simplicity literature concerns how to describe and explain mechanisms so as to improve participant comprehension (Gonczarowski et al., 2023). For each mechanism, we also test one intervention of this kind—an *outcome safety* description that surfaces a well-studied property of the mechanism without scaffolding any particular mode of reasoning. We report all results in the model order Claude, Gemini, GPT-4o, Gemma, which is in decreasing order of their success in playing the non-OSP mechanisms.

Separately, prospect-theory-inspired interventions (loss aversion, the endowment effect) and interventions along rule-following are also a related and interesting direction for probing cognitive constraints. We run interventions along these lines but omit them from the main discussion. These interventions, their complete prompt texts, and the full prompts for all main-body interventions appear in the Appendix.

Across the interventions, three findings emerge. First, the interventions that help the most are contingent reasoning interventions that make the mechanism's payoff structure transparent; scaffolds that direct attention toward opponents are ineffective or actively harmful. Second, the outcome safety descriptions—articulating that your bid determines whether you win but not what you pay (auctions) and that being rejected from top-ranked schools does not hurt your chances at lower-ranked schools (DA)—produce large improvements. Third, belief scaffolds consistently worsen play, suggesting that models are already strategizing along beliefs and that this misguided reasoning is itself a cause for poor play.

4.1 Contingent Reasoning

Li (2024) identifies contingent reasoning as the core cognitive demand of non-OSP mechanisms. To see that truth-telling is dominant in SPSB, a bidder must consider each possible profile of opponent bids and check that no deviation improves payoffs in any case—a "case-by-case comparison, calculating payoffs for each profile of opponent bids" (Li, 2024). The ascending clock eliminates it: at each price, the worst case from following the dominant strategy weakly dominates the best case from any deviation, without reference to opponents. The question, then, is whether scaffolding for contingent reasoning—without changing the extensive form—can partially recover these gains.

In undergraduate economics classes across the world, this is formalized with a payoff tree; our *Payoff Tree* intervention presents the agent with the explicit decision tree of outcomes to simulate exactly this logic.

- 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. If ACCEPTED: you are matched. If REJECTED: you propose to your 3rd choice. And so on. At each branch, acceptance or rejection depends on the school's priorities and who else proposed there."

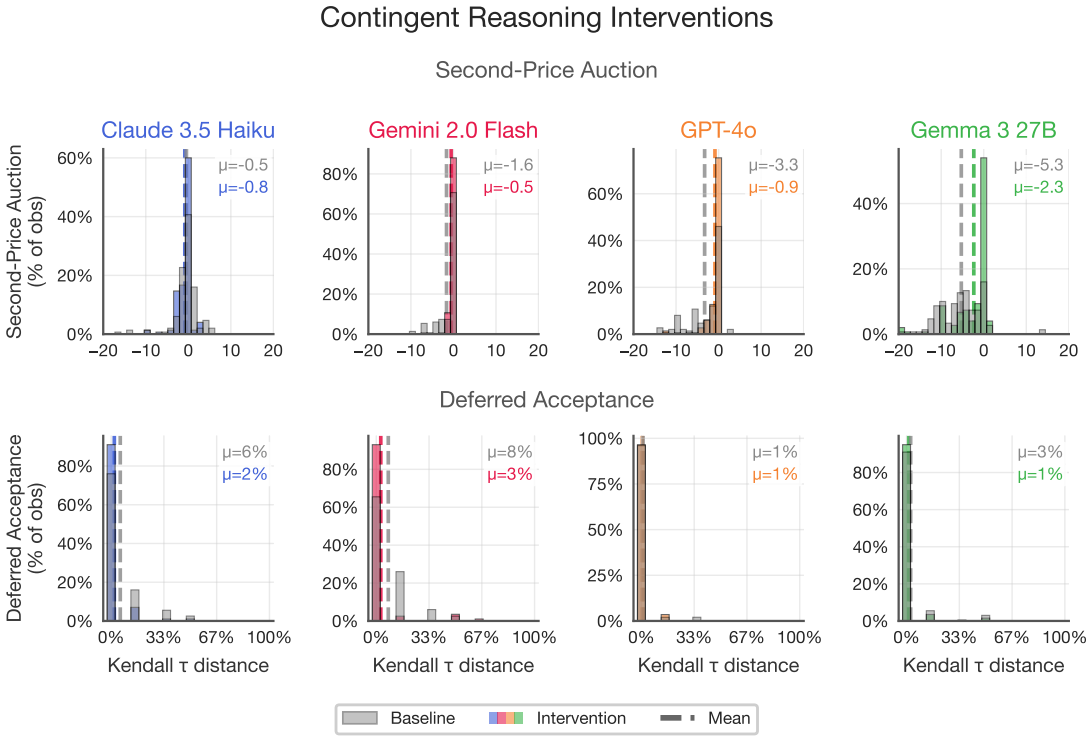

Figure 2: Contingent reasoning: *Payoff Tree* intervention. *Top row:* Second-price auction bid deviations (bid − value); baseline SPSB in gray, intervention in color. *Bottom row:* Deferred acceptance Kendall τ errors; baseline direct DA in gray, intervention in color. The payoff tree scaffold consistently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

In auctions, *Payoff Tree* reduces the mean deviation from $\mu = -2.67$ to $\mu = -1.13$ across models, roughly halving the distance from truthful. The improvement is consistent across models: GPT-4o moves from -3.28 to -0.86, Gemma from -5.27 to -2.34, Gemini from -1.63 to -0.53. Claude, already close to truthful (-0.49), is the one model that shows null improvement. For comparison, the ascending clock achieves $\mu = -0.5$ across models (Section 3.1); the *Payoff Tree* closes roughly half the gap between SPSB and OSP without changing the extensive form.

In DA, *Payoff Tree* reduces mean error from $\mu = 4.2\%$ to $\mu = 1.7\%$ across models, again less than half the baseline error. Here the improvement is unanimous across all four models: Claude from 5.8% to 2.0%, Gemini from 7.6% to 2.8%, GPT-4o from 1.0% to 0.6%, Gemma from 2.6% to 1.4%. The iterative OSP DA achieves $\mu = 0\%$ (Section 3.2); here the scaffold captures much of the improvement but does not fully close the gap.

Laying out the payoff tree does not tell models *what* to do; rather, it tells them *what happens* under each action. This is precisely the cognitive gap between OSP and non-OSP mechanisms: in OSP mechanisms, the extensive form makes these contingencies visible at each decision point; however in non-OSP mechanisms, the agent must construct them herself. The *Payoff Tree* scaffold does this construction for the agent, and the result—consistent improvement across models and mechanisms—suggests that the difficulty of constructing the payoff-relevant contingencies is a real constraint that binds on LLM agents.

4.2 Forward Planning

The next axis we consider is that of forward planning. Pycia and Troyan (2023) formalize a notion of limited forward planning that has become central to the theory of simple mechanisms. A mechanism is *one-step simple* if the dominant strategy emerges from considering only the immediate next step, without full backward induction. The ascending clock is one-step simple: comparing "stay one more tick" to "exit now" suffices at each price, without reasoning about the full trajectory of future 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."

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.

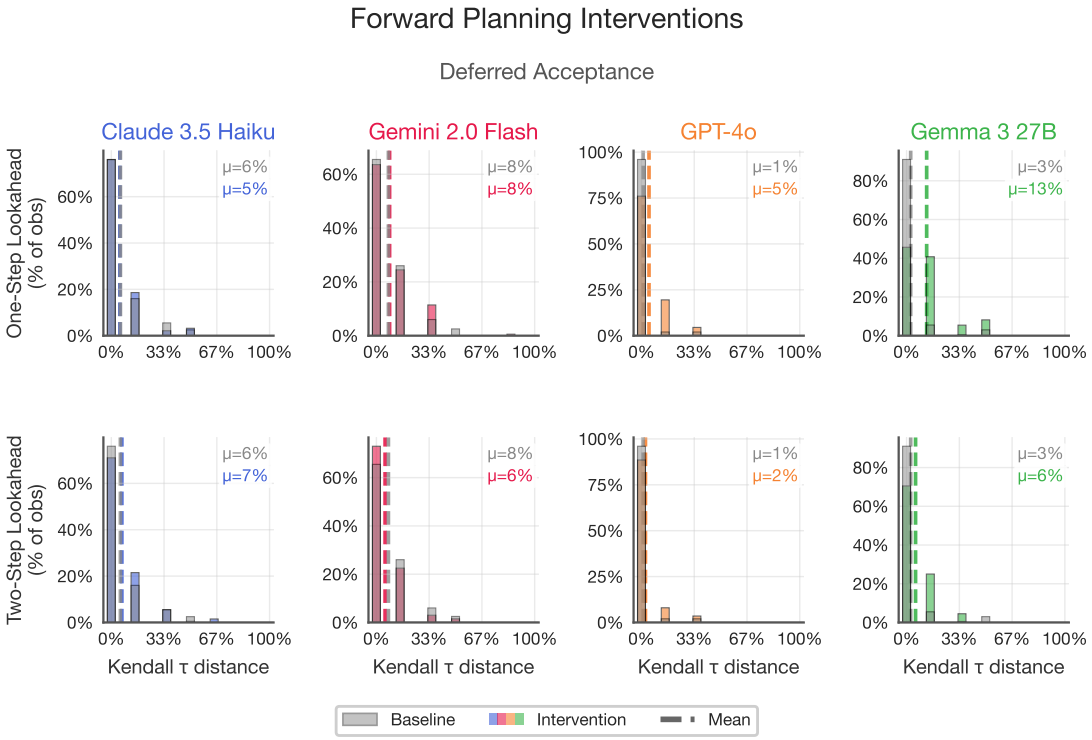

Figure 3: Forward planning: lookahead interventions in deferred acceptance. *Top row:* One-step lookahead; *bottom row:* two-step lookahead. Each cell overlays the direct DA baseline (gray) with the intervention (colored). Lookahead scaffolds worsen play overall, with the damage concentrated in GPT-4o and Gemma; Claude and Gemini are largely unaffected.

4.3 Belief Formation

The third axis is belief formation. One canonical story here is from Börgers and 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. Models prompted to think about opponents shade further below value, undercutting perceived competition.

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.

4.4 Mechanism Description

Separate from the three axes discussed above, but important to the implementation of mechanisms across experimental domains, is the question of how to exposit mechanisms to help participants see for themselves that a mechanism is incentive-compatible. For example, Gonczarowski et al. (2023) formalize such an idea through "strategyproofness-exposing" mechanism descriptions, which aim to make incentive properties transparent without changing the extensive form. Inspired by this approach, we also test one intervention for each mechanism—an *outcome safety* description that surfaces a key structural property of the mechanism.

The key step in Vickrey (1961)'s original argument for the strategy-proofness of the second-price auction comes from an argument about being 'pivotal'. Namely, that your bid only determines *that* you win, not *what* you win. This inspires our *Payoff Safety* intervention for the auction setting: "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."

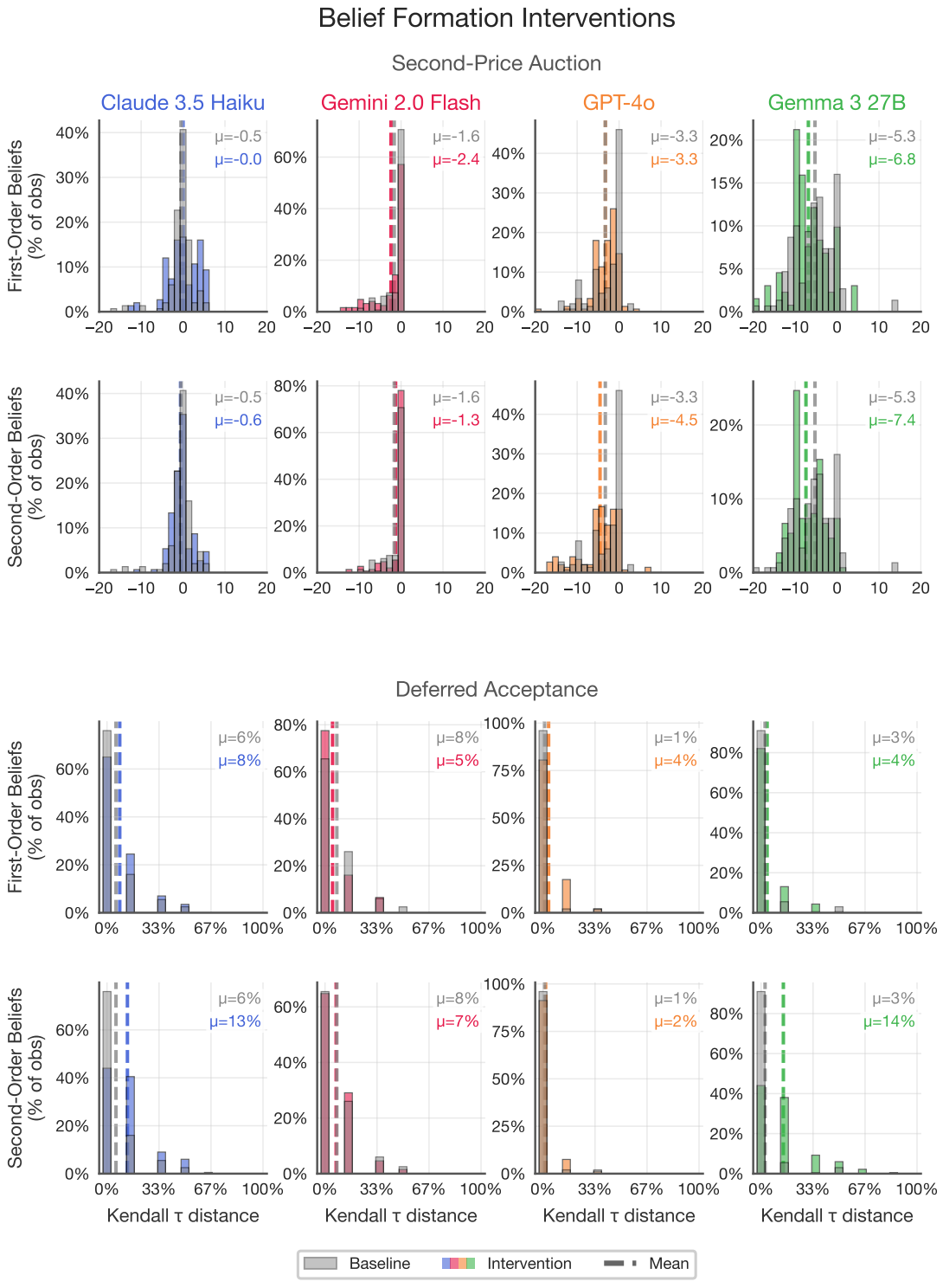

Figure 4: Belief formation interventions across both mechanisms. *Top two rows:* Second-price auction bid deviations under first-order and second-order belief scaffolds. *Bottom two rows:* Deferred acceptance Kendall τ errors under the same scaffolds. Baseline distributions in gray, interventions in 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.

Similarly, a key difference between two famous school choice mechanisms—the Boston mechanism (Abdulkadiroğlu and Sönmez, 2003) and the DA mechanism—is that under DA, rejections are safe. A rejection from a school does not hurt a student's chances at their next-ranked school under DA, whereas under the Boston mechanism, seats at other schools may fill while a student's application is being considered, making rejection costly. In practice, this difference was a key argument for replacing the Boston mechanism with DA in Boston Public Schools. Similarly, our *Rejection Safety* intervention states: "Rejections only redirect, never eliminate. Your ranking determines the *order* in which you are considered, not *whether* you are considered."

Neither intervention scaffolds a particular mode of reasoning. They do not ask the agent to trace through contingencies, simulate the algorithm, or reason about opponents. They simply state the key property of the mechanism that makes it strategy-proof.

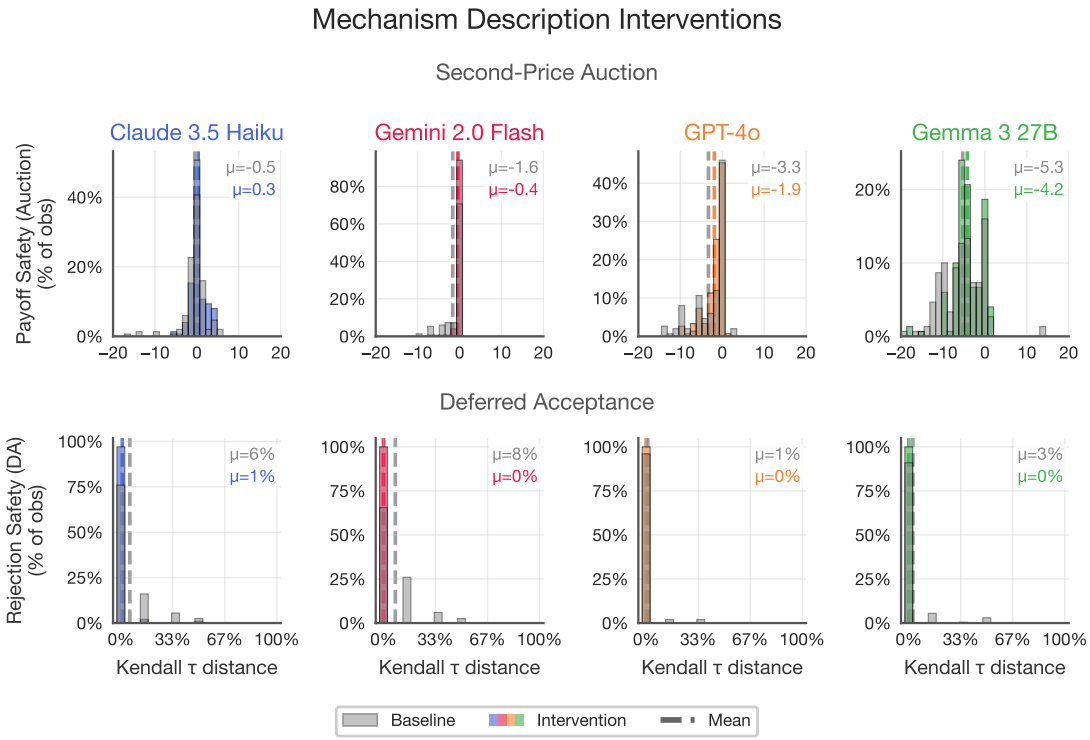

Figure 5: Mechanism description: outcome safety interventions. *Top row: Payoff Safety* in second-price auctions ("your bid determines IF you win, not WHAT you pay"). *Bottom row: Rejection Safety* in deferred acceptance ("rejections only redirect, never eliminate"). Baseline in gray, intervention in color. Both interventions improve play dramatically. Rejection safety in DA nearly matches the full OSP result from Figure 1, reducing mean error from 4.2% to 0.2%.

Across both mechanisms and across models, these interventions improve play dramatically toward the OSP benchmark. In auctions, *Payoff Safety* reduces the mean deviation from $\mu = -2.67$ to $\mu = -1.53$, with improvement across all four models: Claude from -0.49 to -0.30, Gemini from -1.63 to -0.38,

GPT-4o from -3.28 to -1.88, Gemma from -5.27 to -4.17. For comparison, the ascending clock achieves $\mu = -0.5$ (Section 3.1); *Payoff Safety* closes roughly half the gap.

In DA, the results are even more striking. *Rejection Safety* nearly eliminates error, reducing it from $\mu = 4.2\%$ to 0.2% across all models—comparable to the full OSP implementation from Section 3.2, which achieved 0.0%.

That a brief description of a mechanism's incentive properties can nearly replicate OSP's effect—without changing the extensive form—is a surprising result. It suggests that the cognitive barrier for LLMs in non-OSP mechanisms is not in *computing* the dominant strategy given the game tree, but in understanding the mechanism well enough to see that truth-telling is safe. OSP mechanisms resolve both problems simultaneously: they simplify the game tree and make the safety guarantee transparent. Our results suggest it is the latter that does most of the work, at least for LLMs.

As LLMs become more prevalent as economic agents, this suggests that the vast theoretical literature on incentive-compatible mechanisms may have direct operational value—not just as theoretical constructs, but as explanations of how mechanisms work that make their incentive properties actionable.

5 Discussion

The first positive result of the paper is replicating and extending the result along OSP; while LLMs may struggle to play the standard implementation of a mechanism, they improve play dramatically under OSP representations of the same game, as humans do. The ascending clock and iterative DA recover near-truthful play where the sealed-bid and direct-DA formats fail, and this holds across four model families. To our knowledge, the DA result is the first experimental test of an OSP implementation of a two-sided matching mechanism with any kind of agent.

However, two further findings from this paper complicate the picture. First, a single sentence added to the mechanism description—stating that bids determine *whether* you win but not *what* you pay, or that rejections in DA only redirect applications and never eliminate them—recovers nearly the full improvement from OSP without changing the extensive form. Second, interventions scaffolding belief formation and forward planning worsen play on average. The aggregate effect of each intervention family across both mechanisms is summarized in Figure 6.

Together, these suggest an asymmetry result: the binding constraints on LLM reasoning in non-OSP mechanisms is not reasoning incapacity but (almost to the opposite) misdirected reasoning. Building mechanisms that are simple to play for LLMs should prioritize removing distractions that induce misdirected strategizing and redirecting attention towards the incentive properties of the mechanism that matter for optimal play. Hence, as joint LLM-human work becomes more common, our results suggest a natural division of cognitive labor between LLMs and humans: if humans can guide LLMs toward what matters in a given mechanism, the LLM can execute the computations to play well. The challenge (and opportunity) for researchers remains to articulate what this guidance looks like in diverse settings.

Finally, it is important to note that our results are only at a particular capability frontier. We choose non-reasoning models precisely because their failure modes can be well characterized by existing theory, but improving frontier models will likely require the design of new games to isolate their behavior along various cognitive constraints more precisely. Furthermore, our analysis omits any analysis of interactions between cognitive constraints; this remains an exciting avenue for future research.

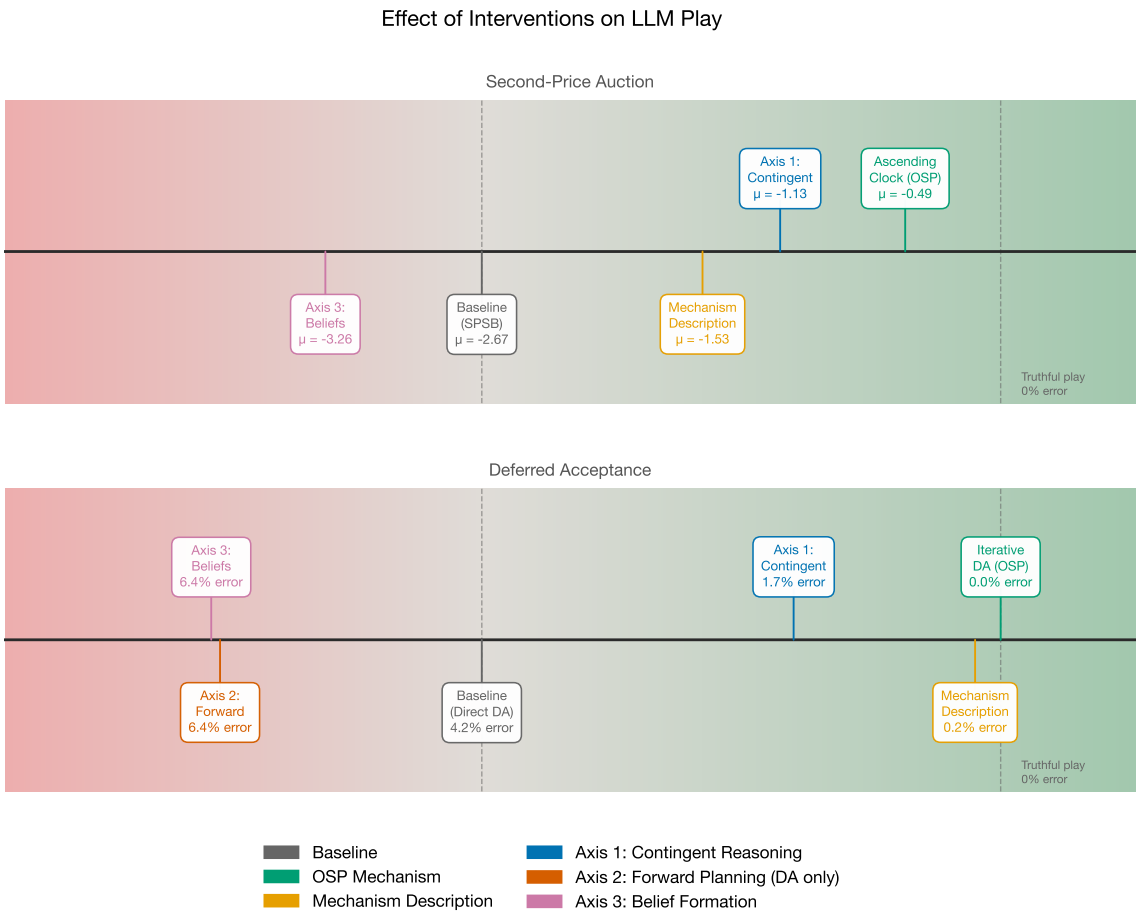

Figure 6: Summary of intervention effects across both mechanisms. *Top row:* Second-price sealed-bid auction, measured in mean bid deviation (μ = bid − value). *Bottom row:* Deferred acceptance, measured in normalized Kendall τ error. Each point is normalized so that baseline play = 0 and truthful play = 1; points to the right represent improvement over baseline. The OSP mechanism (ascending clock auction; iterative DA) recovers near-truthful play in both settings. Contingent reasoning scaffolds (Axis 1) and mechanism descriptions consistently improve play. In contrast, belief formation scaffolds (Axis 3) worsen play in both mechanisms. Forward planning scaffolds (Axis 2, DA only) similarly worsen play. The pattern indicates that LLMs overstrategize along dimensions the mechanism has already rendered irrelevant.

6 Conclusion

The theory of simple mechanisms was built to characterize bounded rationality in humans. We find that the theory of simplicity also informs our thinking about bounded rationality in LLMs. Across two canonical mechanism design settings—auctions and matching—and across four model families, LLMs deviate from dominant-strategy play in strategy-proof mechanisms, and obviously strategy-proof implementations substantially correct these deviations.

The intervention methodology developed here allows us to go further: to identify which cognitive constraints bind, and which do not. Scaffolding for contingent reasoning—making the payoff structure of the mechanism transparent—consistently improves play. Mechanism descriptions that surface incentive properties produce even larger improvements, nearly matching the full OSP benchmark in DA. In contrast, prompting forward planning or belief formation consistently worsens play, suggesting that these are dimensions along which models are already (unproductively) strategizing. The pattern that emerges is not that todays' LLMs lack reasoning ability, but that they misapply it: they strategize about opponents and algorithm dynamics when they should be recognizing that the mechanism makes such reasoning unnecessary.

These findings have both theoretical and practical implications. On the theoretical side, the fact that some of the cognitive constraints identified in the human simplicity literature also bind on LLMs—despite fundamentally different architectures—suggests that these constraints may be inherent to reasoning about the strategic problems, not idiosyncrasies of human cognition. On the practical side, as LLMs increasingly participate in markets, the design of the mechanisms they interact with will matter. Our results suggest that simple mechanisms, and clear (and honest) descriptions of mechanism properties, can serve as a form of alignment: channeling artificial agents toward the outcomes the mechanism was designed to produce, without requiring that they fully understand why. We believe research in this direction will be increasingly important as LLMs take on economic roles as agents along with or on behalf of humans.

Several directions remain open. Our experiments test a specific set of models at a particular capability level; whether frontier reasoning models eventually overcome these constraints, and what that implies for mechanism design, is an important question to track in future work. The heterogeneity we observe across model families—particularly in susceptibility to belief-based reasoning—suggests that mechanism design for heterogeneous populations of artificial agents may raise new challenges; steering vectors and fine-tuning offer a path toward constructing such populations deliberately, enabling controlled study of how mechanism performance varies across diverse populations of directed agents. Our interventions also operate along single cognitive axes in isolation; combining complementary scaffolds—for instance, pairing contingent reasoning with mechanism descriptions—could reveal the minimal intervention needed to close a gap between LLM play and the rational benchmark. Finally, the methodology of cognitive scaffolds as diagnostic probes may prove useful beyond mechanism design, in any setting where one seeks to understand the structure of an agent's reasoning failures.

References

Atila Abdulkadiroğlu and Tayfun Sönmez. 2003. School choice: A mechanism design approach. *American Economic Review* 93, 3 (2003), 729–747.

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

Anthropic. 2024. Claude 3.5 Haiku. https://www.anthropic.com/news/3-5-models-and-computer-use

Itai Ashlagi and Yannai A Gonczarowski. 2018. Stable matching mechanisms are not obviously strategy-proof. *Journal of Economic Theory* 177 (2018), 405–425.

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

James Brand, Ayelet Israeli, and Donald Ngwe. 2023. Using gpt for market research. *Available at SSRN 4395751* (2023).

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

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

François Chollet. 2019. On the measure of intelligence. *arXiv preprint arXiv:1911.01547* (2019).

Google DeepMind. 2024. Gemini 2.0 Flash. https://deepmind.google/technologies/gemini/flash/

Lester E. Dubins and David A. Freedman. 1981. Machiavelli and the Gale-Shapley Algorithm. *Amer. Math. Monthly* 88, 7 (1981), 485–494.

Haluk I Ergin. 2002. Efficient resource allocation on the basis of priorities. *Econometrica* 70, 6 (2002), 2489–2497.

Sara Fish, Yannai A Gonczarowski, and Ran I Shorrer. 2024. Algorithmic Collusion by Large Language Models. *arXiv preprint arXiv:2404.00806* (2024).

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

Yannai A Gonczarowski, Ori Heffetz, Guy Ishai, and Clayton Thomas. 2024. Describing deferred acceptance and strategyproofness to participants: Experimental analysis. *arXiv preprint arXiv:2409.18166* (2024).

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

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

John H Kagel. 1995. *Auctions: A Survey of Experimental Research*. Princeton University Press. 501–585 pages.

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

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

Flip Klijn, Joana Pais, and Marc Vorsatz. 2019. Static versus dynamic deferred acceptance in school choice: Theory and experiment. *Games and Economic Behavior* 113 (2019), 147–163.

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

Shengwu Li. 2024. Designing Simple Mechanisms. *Journal of Economic Perspectives* 38, 4 (2024), 175–192.

OpenAI. 2024. GPT-4o System Card. `https://openai.com/index/gpt-4o-system-card/`

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

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

Alex Rees-Jones. 2018. Suboptimal behavior in strategy-proof mechanisms: Evidence from the residency match. *Games and Economic Behavior* 108 (2018), 317–330.

Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Cassidy, Lukas Borber, et al. 2024. Gemma 2: Improving Open Language Models at a Practical Size. arXiv:2408.00118 [cs.CL]

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

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

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

Kehang Zhu, Anand V Shah, Yanchen Jiang, John Joseph Horton, and David C Parkes. 2024. Evidence from the synthetic laboratory: Language models as auction participants. (2024).

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 and Gonczarowski (2018) decision-tree implementation for DA and the Breitmoser and 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)

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 7: 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 8: 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 and 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 A summarizes the procedure.

[H] [1] RunOSPTreestudents, schools, priorities students $= \emptyset$ matches 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) |top_students| $= 1$ Serial dictatorship node $s^* \leftarrow$ the single top-priority student $w \leftarrow$ AskPickTops^*, schools Use `da_osp_choice` prompt Assign s^* to w RunOSPTreestudents $\setminus\{s^*\}$, schools $\setminus\{w\}$, priorities Let $a, b \leftarrow$ top_students |top_students| $= 2$ under acyclicity each school w where top_by_school$[w] = a$ (in fixed order) AskYesNoa, w, fallback $=$ schools Use `da_osp_yesno_guaranteed` Assign a to w RunOSPTreestudents $\setminus\{a\}$, schools $\setminus\{w\}$, priorities each school w where top_by_school$[w] = b$ (in fixed order) AskYesNob, w, fallback $=$ schools Assign b to w RunOSPTreestudents $\setminus\{b\}$, schools $\setminus\{w\}$, priorities $w_a \leftarrow$ AskPickTop$a$, schools; assign a to w_a $w_b \leftarrow$ AskPickTop$b$, schools $\setminus\{w_a\}$; assign b to w_b RunOSPTreestudents $\setminus\{a, b\}$, schools $\setminus\{w_a, w_b\}$, priorities

At each node, the mechanism provides an "obviousness witness": answering YES yields an immediate match to the offered school, while answering NO keeps all remaining schools (including the offered one) available. This ensures that if the offered school is the student's true top choice among remaining options, YES weakly dominates NO, and vice versa—making truthful play obvious without contingent reasoning.

Yes/No with Consequence Explanation. The primary prompt used at binary-offer nodes in Algorithm A. 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 9: DA OSP — Yes/No with Consequences (`da_osp_yesno`). The consequences of YES (immediate match) and NO (continue with all remaining schools) are made explicit, providing the obviousness witness per Ashlagi and Gonczarowski (2018).

A.1.3 Reasoning Interventions for DA

Each intervention below is appended to (or replaces a section of) the Traditional DA base prompt (Figure 8). 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 A.

```
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 10: 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 A.

```
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 11: 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 12: 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

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 13: 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 and 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 14: SPSB OSP — Ascending Clock (`intervention_proxy_breitmoser`). No dropping out information will be displayed as in Breitmoser and Schweighofer-Kodritsch (2022).

### A.2.3 Reasoning Interventions for SPSB

Each intervention below is appended to or modifies the SPSB base prompt (Figure 13). We show only the intervention-specific text.

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

```
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 15: 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 16: 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?).

A.3 Summary of All Prompt Variants

Table A 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 †.

p3.2cm p4.8cm p5.5cm

Category	Prompt ID	Description
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
Deferred Acceptance — OSP Sequential (Ashlagi and 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 $k{=}\infty$: 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 and Schweighofer-Kodritsch, 2022)

OSP intervention_proxy_breitmoser ⋆ ascending clock

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
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

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 and 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*: Asks the agent to compare their willingness-to-accept (selling price) with their willingness-to-pay (buying price), invoking the classic WTA–WTP gap.

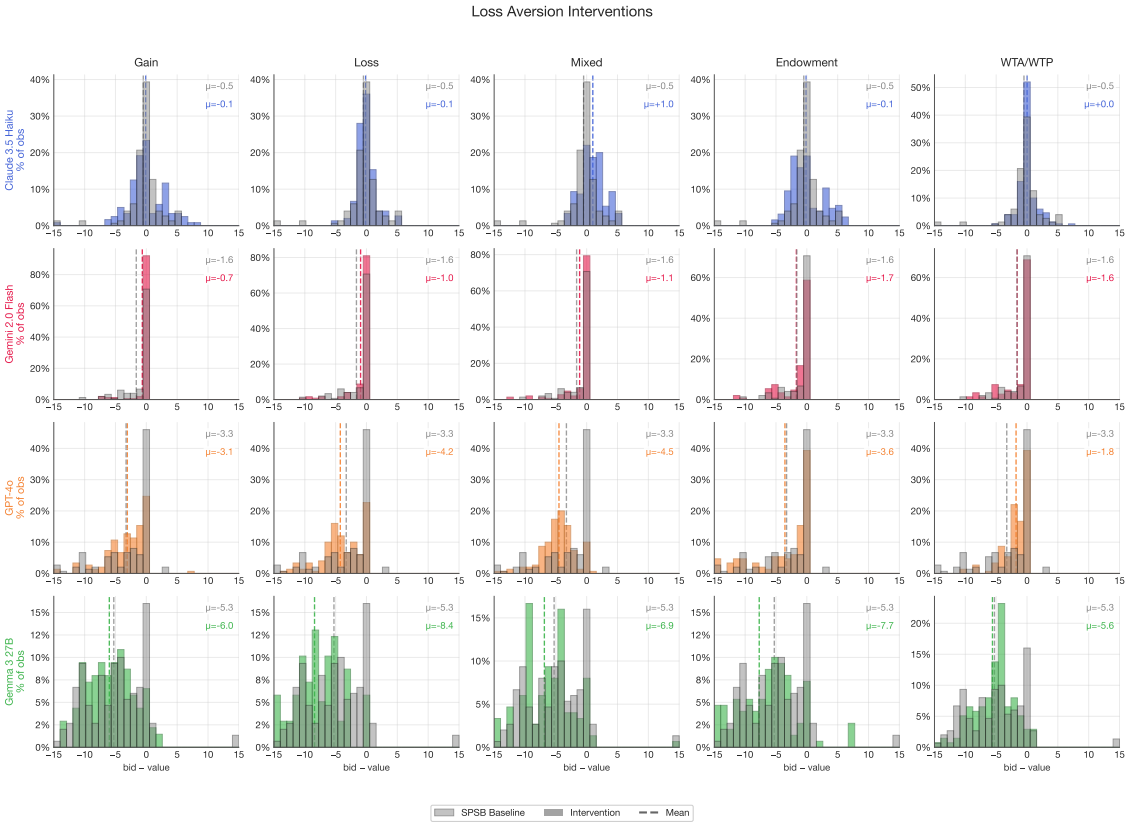

Figure 17: Loss aversion interventions in SPSB auctions. Baseline SPSB in gray, intervention in color. Dashed lines indicate means. None of the framing interventions consistently improve play relative to the SPSB baseline ($\mu = -2.67$). The loss frame ($\mu = -3.44$) and endowment ($\mu = -3.40$) worsen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

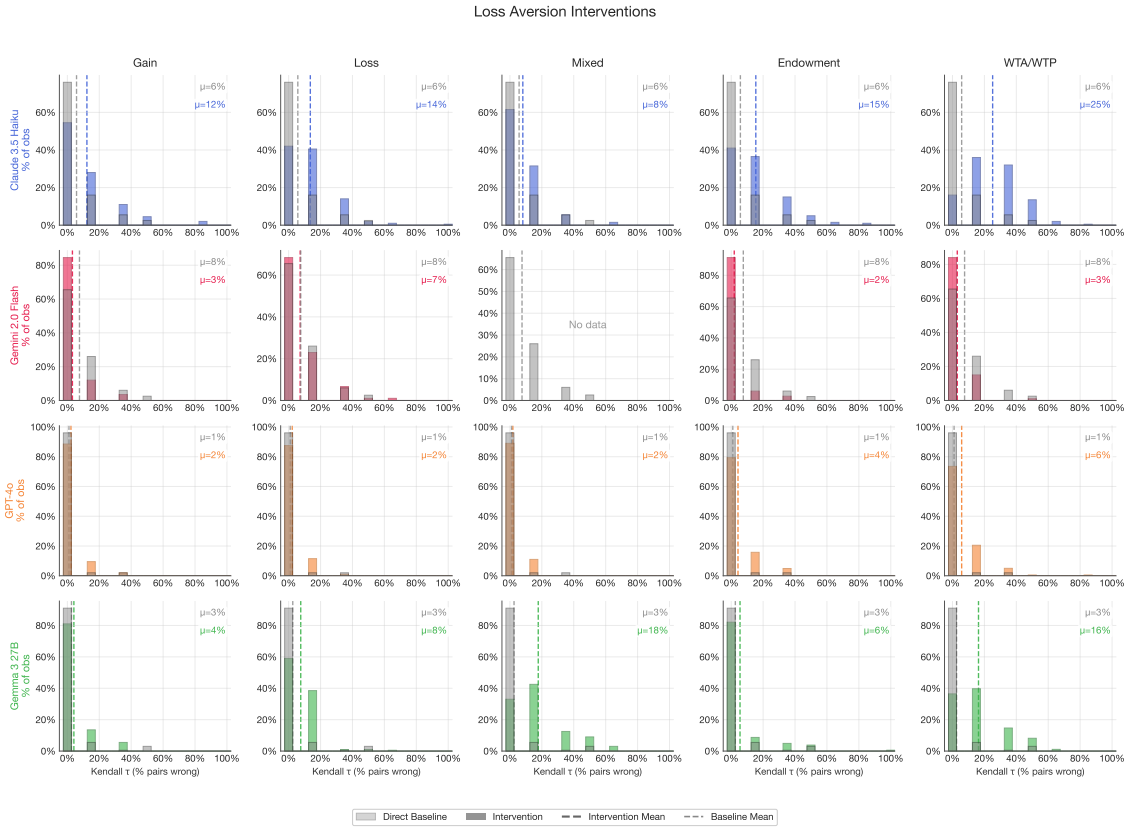

Figure 18: Loss aversion interventions in deferred acceptance. Baseline direct DA in gray, intervention in color. The results are noisy and largely negative: most framing interventions worsen play relative to the baseline ($\mu = 4.2\%$), with WTA/WTP producing the worst outcomes ($\mu \approx 12.6\%$ 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. The difference is informative: the cognitive barrier in these mechanisms is not about how payoffs are framed, but about understanding the mechanism's strategic structure.

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).”

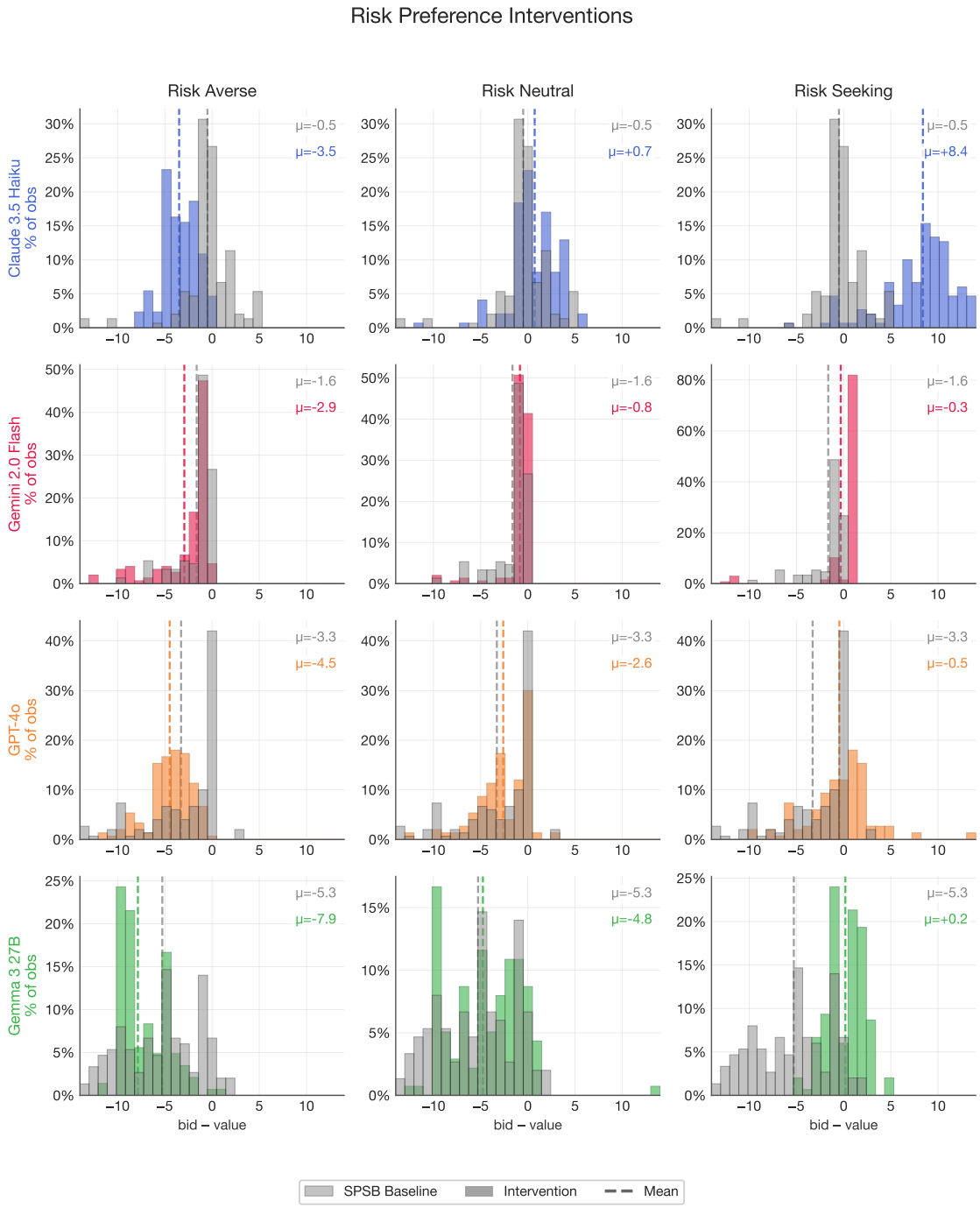

Figure 19: Risk preference interventions in SPSB auctions. Risk aversion increases underbidding ($\mu = -4.71$), while risk-seeking dramatically shifts Claude to overbidding ($\mu = +8.40$). Risk neutrality ($\mu = -1.82$) modestly improves play relative to the SPSB baseline ($\mu = -2.67$).

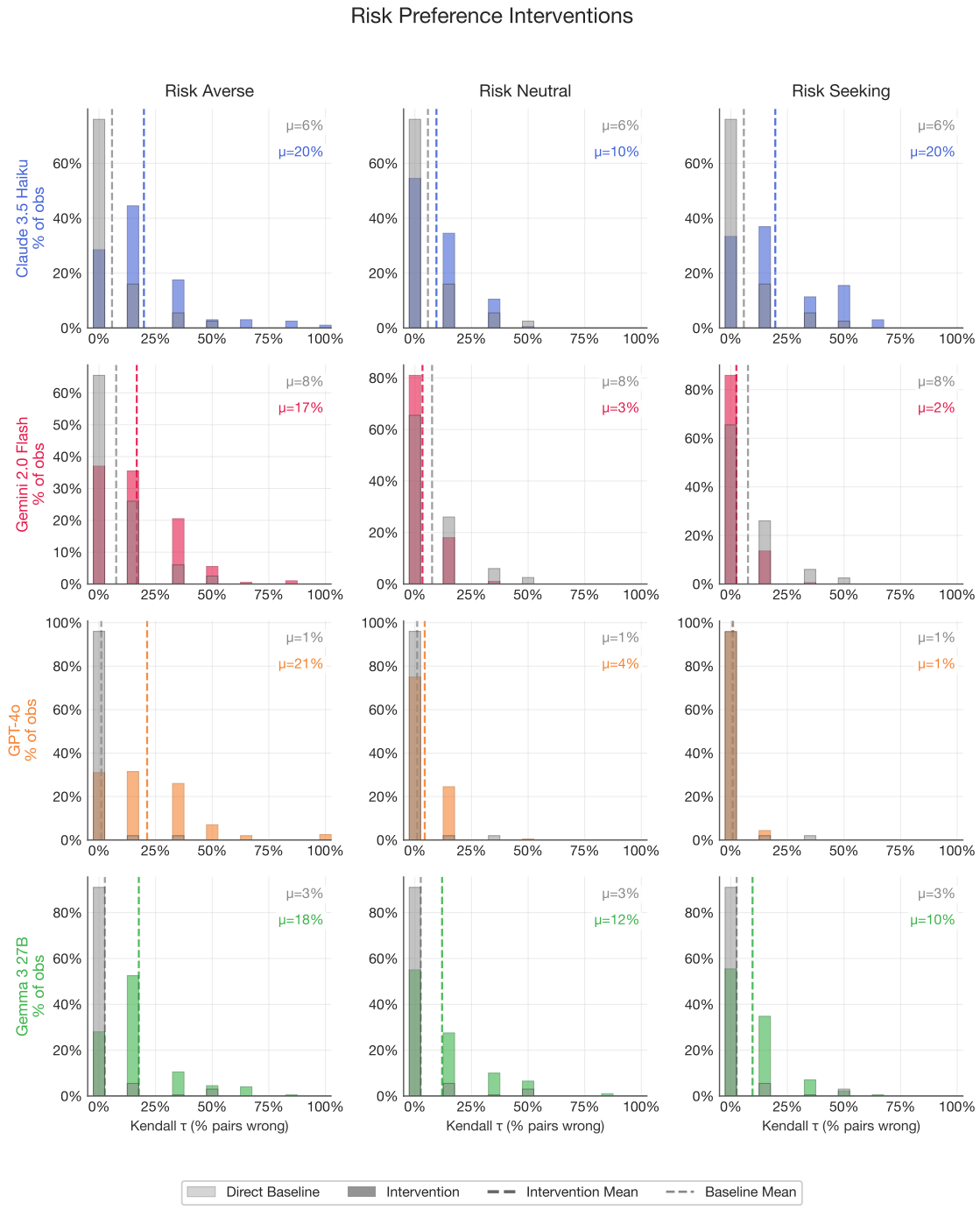

Figure 20: Risk preference interventions in deferred acceptance. Risk aversion dramatically worsens play ($\mu \approx 18.8\%$, compared to baseline 4.2%). Risk neutrality and risk-seeking also worsen play, though less severely.

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. In fact, in aggregate, it appears that the model—while strategizing—are extremely susceptible to risk preference priming.

These results underscore that the LLMs' failures in strategy-proof mechanisms are not driven by implicit risk attitudes. Truthful bidding is dominant regardless of risk preferences under IPV, so the fact that risk-persona instructions change behavior at all confirms that the models do not understand the mechanism's incentive structure. The risk interventions change *how* models deviate from truth-telling, but do not address *why* they deviate in the first place.