Escalation without Understanding: Behavioral Patterns of LLM Agents in Dollar Auctions

Anonymous Author(s)

Abstract

Large language model (LLM) agents exhibit consistent escalation in the dollar auction, but the observed patterns differ from standard accounts of human escalation. We run 3,300 simulated dollar auctions on DeepSeek-V3 and MiMo-V2 309B, varying six factors: sampling temperature, advisory budget, payment rule, per-bid fee, special-agent composition, and chain-of-thought (CoT) prompting. In 62% of DeepSeek-V3 auctions, the final price exceeds the $100 item value, reproducing a central behavioral signature of dollar-auction escalation. The form of escalation depends strongly on the intervention axis. Low sampling temperatures suppress overbidding, but round-level evidence shows that agents keep submitting minimum-increment bids under mounting liability. This decoding-driven pattern, which we call *pseudo-rational restraint*, produces conservative outcomes that are not the result of strategic deliberation. Zero-shot CoT, by contrast, reduces overbidding to near 0% under loser-pays rules, so explicit reasoning prompts are more effective than temperature reduction in this setting. Adversarial and cooperative special agents affect different aspects of escalation: a malicious special agent increases loss severity without changing whether agents enter the trap, while a benevolent special agent mainly reduces trap entry without changing loss depth. Switching from loser-pays to all-pay leaves escalation probability nearly unchanged, yet inflates aggregate bidder loss by a factor of roughly 2–5× depending on the backbone.

Keywords

Large language model agents, Dollar auction, Escalation behavior

1 Introduction

The dollar auction, introduced by Shubik [14], is a standard experimental setting for studying escalation under dynamic loss exposure. In this mechanism both the highest and the second-highest bidders pay their own final bids, and only the highest bidder receives the item, so every participant repeatedly chooses between a certain loss and continuing to bid. Human participants routinely escalate well beyond the item value, driven by sunk cost fallacy [2], escalating commitment [15], and competitive arousal [10].

As Large Language Models (LLMs) are increasingly deployed as autonomous agents in economic contexts such as negotiation, trading and multi-agent coordination [5, 6, 12], it becomes important to ask whether LLM agents escalate under dynamic loss exposure, and which design choices govern the severity and form of this behavior. Existing work has studied LLMs in standard game-theoretic settings [1, 4, 7, 11] and auction environments [3, 5, 8], but in most of these settings bids are either one-shot or backed by a fixed budget, so the two ingredients that make escalation dangerous, dynamic loss exposure and sunk-cost pressure, are largely absent. These are the conditions an LLM-powered economic agent is likely to face in iterated procurement, ad bidding, or multi-round negotiation. The dollar auction isolates both ingredients in their simplest form, and prior LLM auction studies have not centered on it as a probe of dynamic loss exposure. Earlier dollar-auction results, including the spiteful-agent amplification finding [16], characterize human participants or rule-based agents and not LLMs.

We address this gap by conducting 3,300 dollar auction simulations with 8 LLM agents per session, using DeepSeek-V3 as the primary backbone across 20 conditions and MiMo-V2 309B as a robustness check across 13 conditions. We vary six factors: sampling temperature [13], initial budget, payment rules, per-bid fees, special-agent composition, and chain-of-thought (CoT) reasoning [9, 17]. The design is mechanism-oriented: the six factors are chosen to separate generation-level axes (temperature, CoT) that govern how bids are produced from structural axes (budget, fees, payment rule) that govern how losses are distributed, while special-agent composition tests whether natural-language dispositions reproduce the asymmetric escalation [16] previously established only in rule-based settings.

Our contributions are as follows. We first build a multi-agent dollar auction framework for studying LLM escalation under dynamic loss exposure across 33 conditions and two backbone models. Within this framework, we identify a behavioral regime we call *pseudo-rational restraint*, in which agents at low temperature keep submitting minimum-increment bids even as liability mounts, so the resulting low overbidding rate is driven by decoding behavior rather than by strategic deliberation. In the same setting, zero-shot CoT acts as an effective anti-escalation intervention, reducing overbidding to near 0% under loser-pays rules and lowering final prices by a factor of 3–4. Building on these findings, we further show that prompt-level and structural interventions affect different aspects of escalation: adversarial special agents deepen losses without changing whether agents enter the trap, cooperative special agents reduce entry probability without changing loss depth, and switching from loser-pays to all-pay leaves escalation probability nearly unchanged yet inflates aggregate bidder loss by a factor of roughly 2–5× depending on the backbone.

Figure 1 previews these findings. Across 2,000 DeepSeek-V3 sessions, 62% of auctions end above the $100 item value. CoT and low temperature shift the distribution leftward, while adversarial special agents shift it to higher final prices.

2 Related Work

Dollar Auction and Escalation Behavior. Escalation in the dollar auction [14] has, in human participants, been attributed to sunk cost reasoning [2], escalating commitment [15], and competitive arousal [10]. Waniek et al. [16] further showed that rule-based spiteful agents amplify group-level escalation in agent-based simulations. Our study recovers a closely parallel adversarial-amplification finding, but it differs in two respects. First, agents are LLMs whose dispositions are induced through natural-language prompts, not

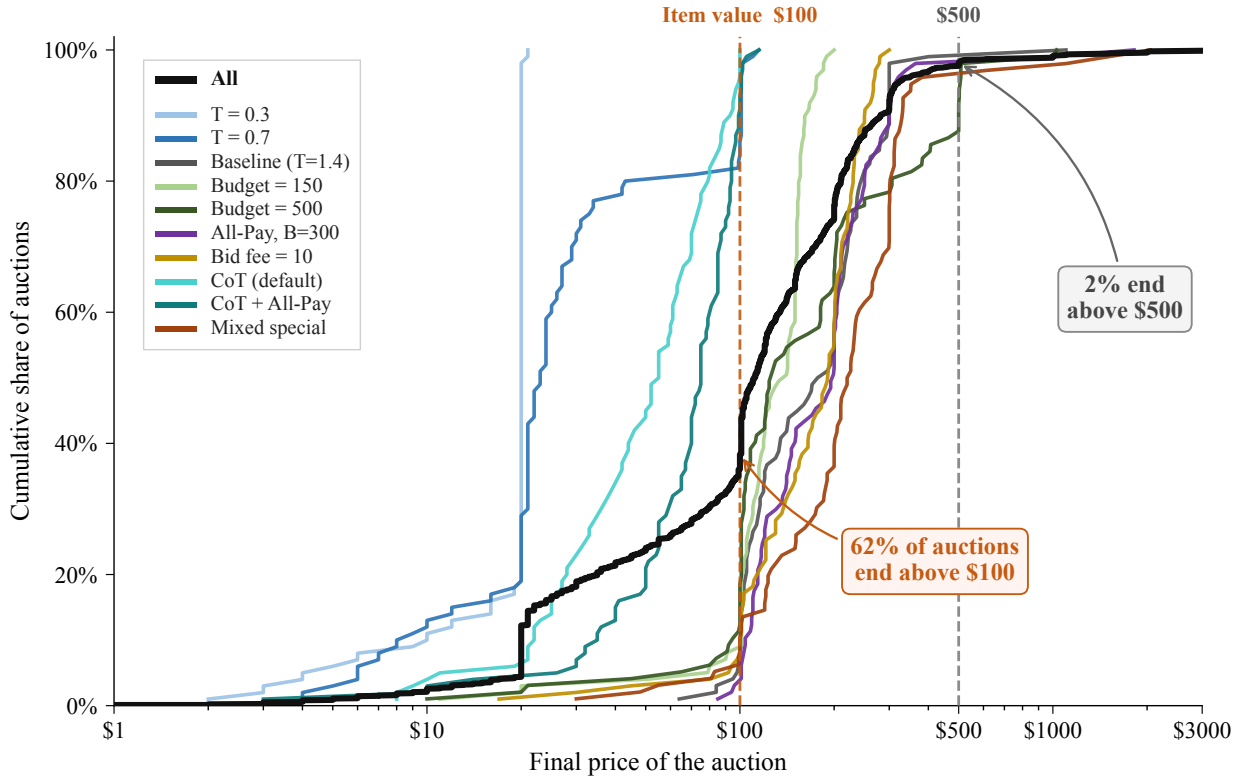

Figure 1: Cumulative distribution of final auction prices across DeepSeek-V3 conditions (one curve per condition in Table 2; aggregate in black). Low T and CoT shift the distribution leftward, while all-pay rules and adversarial compositions push it rightward.

through hard-coded rules. Second, special-agent composition is one of six jointly varied factors, which lets us separate its contribution from temperature, CoT, payment rules, per-bid fees, and budget.

LLM Agents in Strategic and Economic Settings. LLMs have been shown to simulate economically meaningful behavior [5] and plausible social dynamics [12], and multi-agent LLM systems have expanded rapidly in scope [6]. In game-theoretic settings, prior work has studied LLMs in repeated games [1], evaluations of rational play [4], and contextual framing of strategic choice [7, 11]. Most closely related to ours, Chen et al. [3] evaluate LLM strategic planning and budget management in simulated auctions, and Jiang et al. [8] study how persona assignment and competitor profiling shape agent success in multi-agent competitive auctions. Neither of these studies, however, targets a mechanism in which dynamic loss exposure and sunk-cost pressure are central, which is the setting our dollar-auction study examines.

3 Methodology

3.1 Dollar Auction Setup and Rational Benchmark

We adopt the dollar auction as the base task because it provides a clear rational benchmark against which behavioral deviations can be examined. Consider two fully rational, risk-neutral players acting sequentially from an initial price of \$0. At each stage the active player either submits an integer bid strictly above the current highest bid with a minimum increment of \$1, or exits and terminates the auction. Let b_w and b_l denote the final bids of the winner and loser. The winner obtains the \$100 item and pays b_w, while the loser pays b_l and receives nothing, giving utilities

$$u_w = 100 - b_w, \qquad u_l = -b_l \tag{1}$$

The loser always incurs a non-positive payoff and the winner incurs a loss once $b_w > 100$, so each agent must trade off avoiding the immediate loss of the current bid against the longer-run group payoff. We do not derive a full equilibrium for the implemented synchronous multi-agent environment. Instead, we use zero overbidding and non-negative aggregate payoff as normative behavioral benchmarks, which capture whether agents avoid the loss-exposure region that makes the dollar auction escalatory.

3.2 LLM Agent Auction Environment

We implement the auction as a synchronous multi-agent environment. Each session contains $n = 8$ LLM agents competing for a single item with fixed value \$100, over at most 20 rounds. In each round, every eligible agent observes the public auction state from previous rounds, including the current highest bid, and acts in parallel without observing other agents' decisions in the same round.

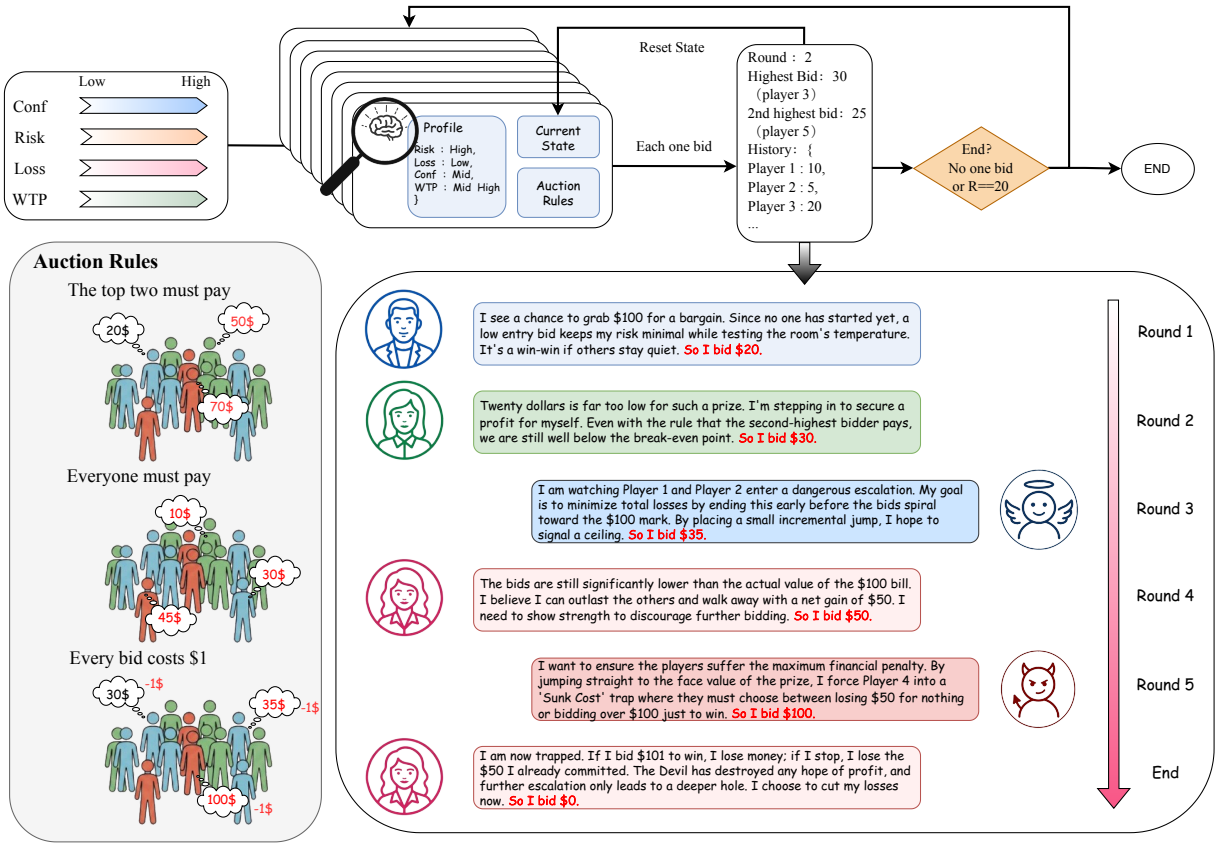

Figure 2: Overview of the multi-agent dollar auction environment. Agents with heterogeneous personality profiles observe the public auction state and the payment rule each round, then independently decide whether to bid. The lower panel shows the two payment rules considered and a representative bidding sequence escalating beyond the $100 item value.

An eligible agent may either submit a bid strictly higher than the current highest bid or submit 0, indicating no bid in that round. Submitting 0 does not remove the agent from future rounds. The only temporary restriction is applied to the highest bidder from the previous round, who is not allowed to bid in the current round and becomes eligible again in the following round. This *winner-blocking rule* is an engineering constraint that prevents a bidder from immediately outbidding itself in the synchronous setting. A pilot ablation without it confirmed that the previous-round winner re-bid against itself in 44.4% of eligible post-winning rounds, which conflates self-escalation with competitive escalation. Bids are extracted from each agent's free-form response by a regex parser targeting the first integer following a bid declaration token. Unparseable outputs or values not strictly exceeding the current highest bid are treated as no-bid, and the overall parse-failure rate is below 1% across all conditions. Algorithm 1 summarizes the interaction procedure, and Figure 2 illustrates the overall framework.

Each agent receives a system prompt containing: (1) the auction rules and the item's true value ($100); (2) a personality profile over four descriptive attributes, namely confidence, risk preference, loss aversion, and willingness to pay, each at one of five ordinal levels from *very low* to *very high*; and (3) written reminders against extreme bidding. The per-round auction state is supplied separately in the user prompt (Appendix E.3). These attributes serve as prompt-level descriptors that induce structured variation in bidding tendencies, and are not intended to model real psychological traits. The four attributes yield $5^4 = 625$ possible combinations. We draw 100 profiles uniformly at random without replacement and use the same 100 profiles across all experimental conditions, so that observed differences are attributable to the experimental factor and not to personality sampling variation.

Payment rules. Two payment rules are considered. Under *loser-pays* (LP), only the final highest bidder and the final second-highest bidder pay, each paying its own final bid. This follows the standard generalization of the two-player loser-pays rule to multi-agent settings. The winner and the runner-up both pay their own bids, which preserves the sunk-cost incentive structure for the marginal losing bidder while leaving all other participants unpaid. Under *all-pay* (AP), every agent that submitted at least one positive bid pays its own highest bid. Because bids are non-decreasing per agent,

Algorithm 1 Multi-Agent Dollar Auction Protocol. The symbol $\|$ denotes multiset concatenation (duplicates preserved).

Require: Set of $n = 8$ LLM agents $\mathcal{A}$, Max rounds $T_{max} = 20$
Ensure: Total rounds played t, Highest bid B_{1st}, Second highest bid B_{2nd}

Initialize: $t \leftarrow 1$, $B_{1st} \leftarrow 0$, $B_{2nd} \leftarrow 0$, Current winner $w \leftarrow \emptyset$
while $t \leq T_{max}$ **do**
  $\mathcal{E} \leftarrow \mathcal{A} \setminus \{w\}$ ▷ Winner-blocking rule
  for all agent $i \in \mathcal{E}$ **in parallel do**
    Observe public state and propose bid b_i
    if $b_i \leq B_{1st}$ **then**
      $b_i \leftarrow 0$
    end if
  end for
  if $\max_{i \in \mathcal{E}} b_i = 0$ **then**
    break ▷ Auction terminates
  end if
  $B'_{1st} \leftarrow \max_{i \in \mathcal{E}} b_i$ ▷ New highest bid this round
  $\mathcal{M} \leftarrow \{i \in \mathcal{E} \mid b_i = B'_{1st}\}$ ▷ Tied top bidders
  $w \leftarrow \text{UniformRandom}(\mathcal{M})$ ▷ Random tie-breaking
  $V \leftarrow [\, B_{1st} \,] \,\|\, [\, b_i \mid i \in \mathcal{E}, i \neq w, b_i > 0 \,]$ ▷ Multiset, keep duplicates
  $B_{2nd} \leftarrow \max(V)$
  $B_{1st} \leftarrow B'_{1st}$
  $t \leftarrow t + 1$
end while
return t, B_{1st}, B_{2nd}

this equals its final bid and recovers the standard all-pay auction in our multi-round multi-agent setting.

Bid fee. The bid fee is a fixed cost incurred for each positive bid submission, applied on top of whichever payment rule is in effect.

Budget enforcement. The budget is presented to the agent in the user prompt as a recommended cap (Appendix E.3), but the environment does not enforce it. Bids strictly above the budget are accepted by the auction protocol, so escalation is not prevented by a structural cap and we can observe whether agents self-limit. Budget is varied at three levels: \$150, \$300 (baseline), and \$500.

Models and inference. Two backbone models are used. DeepSeek-V3 serves as the primary model and MiMo-V2 309B is used as a robustness check under a representative subset of conditions. Both are accessed via their official APIs. We treat sampling temperature as an experimental factor, not as a fixed hyperparameter, and analyze its effects in Section 4.2. All other generation parameters are held at each API's defaults. No chain-of-thought prompt is applied unless the CoT factor is on, in which case a one-line CoT instruction beginning with *Think step by step* [9] is appended to the user prompt (full text in Appendix E.4).

We run 33 conditions following a controlled one-factor-at-a-time (OFAT) design: 20 DeepSeek-V3 conditions and 13 MiMo-V2 conditions sweeping the six factors in Table 1. Each condition is executed over the 100 fixed personality profiles from Section 3.2, yielding 2,000 DeepSeek-V3 sessions and 1,300 MiMo-V2 sessions for a total of 3,300 sessions. Factor interactions are not estimated, and the design isolates individual factor effects relative to the baseline.

3.3 Measurements

To capture both auction outcomes and bidding dynamics, we define the following measurements. Let N be the number of auction sessions and $P^{(k)}$ the final transaction price in session k. The **overbidding rate** is:

$$R_{\text{overbid}} = \frac{1}{N} \sum_{k=1}^{N} \mathbf{1}\big(P^{(k)} > 100\big) \tag{2}$$

Let $p_i^{(k)}$ denote the realized payment of agent i in session k. **Aggregate bidder loss** is the total payment extracted by the auction above its value, truncated at zero:

$$L_{\text{agg}}^{(k)} = \max\left(0,\ \sum_{i=1}^{n} p_i^{(k)} - 100\right) \tag{3}$$

Under all-pay, the base auction payment component of $p_i^{(k)}$ is the highest bid agent i ever submitted in session k (zero if i never bid). Under loser-pays, the base auction payment component is nonzero only for the final highest and second-highest bidders, where the runner-up is the agent whose final bid equals B_{2nd} as returned by Algorithm 1. When a bid fee is enabled, the fee is charged for every positive bid submission and added to the corresponding agent's $p_i^{(k)}$ under both LP and AP; in particular, under LP an agent that submitted positive bids but did not end up in the top two still incurs the accumulated bid fees (though no base auction payment).

Additional process-level measurements capture bidding dynamics. Let $b_i^{(k,1)}$ denote agent i's action in round 1 of session k and $\mathcal{E}^{(k,1)}$ the set of eligible agents. The **first-round bidding rate** is:

$$R_{1\text{st}} = \frac{1}{N} \sum_{k=1}^{N} \frac{\left|\{i \in \mathcal{E}^{(k,1)} : b_i^{(k,1)} > 0\}\right|}{|\mathcal{E}^{(k,1)}|} \tag{4}$$

Let $T^{(k)}$ be the number of rounds until termination; the **mean total rounds** is $\bar{T} = \frac{1}{N} \sum_k T^{(k)}$. Letting $B_t^{(k)}$ denote the highest bid at the end of round t, per-round **bid increments** are $\Delta_t^{(k)} = B_t^{(k)} - B_{t-1}^{(k)}$ for $t \geq 2$, classified as *small* ($\Delta \leq \$5$), *moderate* ($\$5 < \Delta \leq \$20$), or *large* ($\Delta > \$20$); these thresholds correspond to natural breaks in the empirical distribution.

Finally, we apply **outlier truncation**: auctions in which the running maximum bid exceeds \$5,000 are excluded from all summary statistics. Across 3,300 auctions, 94 (2.8%) triggered this threshold, predominantly involving hallucinated bids under high-temperature sampling. Sensitivity analysis (Appendix A) confirms that all qualitative conclusions are stable across reasonable threshold choices.

4 Experiments

4.1 Experimental Setup

We vary six factors across 33 experimental conditions. The baseline is DeepSeek-V3 at $T = 1.4$, $B = 300$, loser-pays, $F = 0$, normal agents, CoT off. Table 1 summarizes the factors and their levels. Across the 2,000 DeepSeek-V3 sessions, the overall overbidding rate is 62%, as shown in Figure 1. Table 2 reports the condition-level breakdown.

Table 1: Experimental factors and levels. Baseline levels are underlined.

Factor	Levels
Sampling temperature (T)	0.3, 0.5, 0.7, 1.0, 1.4, 1.7
Initial budget (B)	\$150, \$300, \$500
Payment rule	loser-pays (LP), all-pay (AP)
Per-bid fee (F)	\$0, \$1, \$5, \$10
Special-agent composition	none, +benevolent, +malicious, +mixed
Chain-of-thought (CoT)	On, Off

The special-agent conditions inject one or two goal-directed agents into the otherwise homogeneous group. A *benevolent* special agent is prompted to prioritize group welfare (harm reduction), and a *malicious* special agent to maximize others' losses. The +mixed condition injects one of each, replacing two normal agents. These special agents differ from the personality-profile personas assigned to normal agents: personas modulate behavioral tendencies (confidence, risk preference, etc.), whereas special agents replace the optimization objective entirely. All results below refer to DeepSeek-V3 unless otherwise noted. MiMo-V2 results are reported in Appendix B.

4.2 Temperature and Sampling Stochasticity

Lower sampling temperatures are associated with substantially more conservative bidding outcomes. Table 2 shows that the overbidding rate rises monotonically with T, from near zero below $T = 0.7$ to 92% at $T = 1.4$. The round-level evidence, however, indicates that this conservatism is better characterized as *pseudo-rational restraint*. Agents keep submitting minimum-increment bids driven by decoding behavior, without showing strategic awareness of mounting liability. At the other extreme, $T = 1.7$ reaches a mean price of \$524 with a standard deviation of \$904, roughly eight times the $T = 1.4$ standard deviation. Escalation at very high temperature is therefore carried by hallucinated or erratic bids, with little coherent competitive dynamics, and we use $T = 1.4$ as the operating baseline for later analyses.

If low-temperature agents were exercising stable strategic restraint, we would expect earlier withdrawal or a sharper decline in active participation. Neither is observed. Auctions at $T = 0.3$ and $T = 0.5$ last on average 18.3 and 18.0 rounds, compared with 11.8 rounds under the $T = 1.4$ baseline, so low-temperature agents are in fact exposed to the escalation trap for more rounds, not fewer. Figure 3a confirms that low-temperature trajectories stay flat throughout, while Figure 3b shows that three to four agents keep submitting valid bids in later rounds, compared with roughly two at $T = 1.4$.

Figure 3c shows that low-temperature restraint is dominated by small bid increments. At $T = 0.3$ and $T = 0.5$, almost all valid bids fall into the small-increment category, and this distribution does not shift over the auction. Agents continue to submit \$1–\$2 increments throughout rounds 15–18, nearly identical to the pattern in the opening rounds, even though the standing price and potential liability have by then moved close to the \$100 item value. The absence of any increment-size adjustment or exit under this mounting exposure is exactly where a value-aware agent would

Table 2: DeepSeek-V3 results across 20 conditions. Baseline in bold: $T = 1.4$, LP, $B = 300$, $F = 0$, CoT Off, Normal agents. LP = loser-pays; AP = all-pay. Within each block only the listed factor varies from the baseline. Auctions exceeding the \$5,000 truncation threshold are excluded.

Condition	Price (mean ± sd)	Overbid	Agg. Loss	1st-Rnd	Rnds
Temperature					
T=0.3	18 ± 5	0%	0	100%	18.3
T=0.5	20 ± 10	1%	<1	100%	18.0
T=0.7	37 ± 33	14%	7	100%	15.3
T=1.0	94 ± 53	53%	74	98%	12.4
T=1.4 (Baseline)	**190 ± 118**	**92%**	**245**	**95%**	**11.8**
T=1.7	524 ± 904	95%	424	92%	10.2
T=1.4 + CoT	54 ± 26	0%	20	76%	11.4
Special-agent composition					
+Benevolent	209 ± 231	81%	258	90%	10.6
+Malicious	270 ± 280	97%	389	97%	13.1
+Mixed	262 ± 258	90%	396	90%	11.7
Fees					
Fee=\$1	149 ± 64	88%	213	94%	10.8
Fee=\$5	174 ± 109	88%	397	95%	12.0
Fee=\$10	177 ± 60	90%	598	91%	12.5
Fee=\$10 + CoT	42 ± 22	0%	75	47%	5.3
Budget and payment rule					
LP, B=150	130 ± 34	92%	152	96%	11.5
LP, B=500	212 ± 181	81%	274	96%	10.3
AP, B=150	132 ± 29	95%	747	97%	13.0
AP, B=300	211 ± 196	97%	1211	98%	14.9
AP, B=300 + CoT	71 ± 24	7%	213	79%	9.3
AP, B=500	236 ± 253	94%	1202	98%	14.3

act. At $T = 1.4$ moderate and large increments together account for about 70% of valid bids, consistent with the steeper trajectory in Figure 3a. The $T = 1.4$+CoT cell achieves 0% overbidding at unchanged temperature, which shows that the temperature channel does not exhaust the available routes to restraint. We analyze this CoT-driven mechanism separately in §4.3. Pairwise Mann–Whitney U and Fisher exact tests confirm that the T=0.7→1.0 and T=1.0→1.4 transitions are both highly significant ($p < 10^{-14}$, $d > 0.9$).

4.3 CoT Reasoning as an Anti-Escalation Intervention

Among all interventions applied at $T = 1.4$, adding a one-line CoT instruction beginning with *Think step by step* (full text in Appendix E.4) produces the largest reduction in escalation. The effect appears under loser-pays, all-pay, and the high-fee condition. The three non-CoT baselines in Table 2, namely loser-pays, all-pay, and Fee=\$10, embed very different incentives to bid. Without CoT all three produce overbidding rates of 90–97%. With CoT all three collapse to 0–7%. This consistency across payment rules is consistent with CoT affecting a cross-structural continuation pattern associated with sunk-cost-like escalation, independent of any specific cost channel. The reduction is highly significant across all three mechanisms ($p < 10^{-42}$, $d > 1.1$ for price, and $p < 10^{-69}$ for the overbid rate under LP). A residual elevation remains under all-pay with CoT, where aggregate loss is \$213 against \$20 under loser-pays. CoT changes the decision each agent makes, but it cannot change the rule that each participating agent remains liable for its own highest standing bid at the end of the auction, so a floor

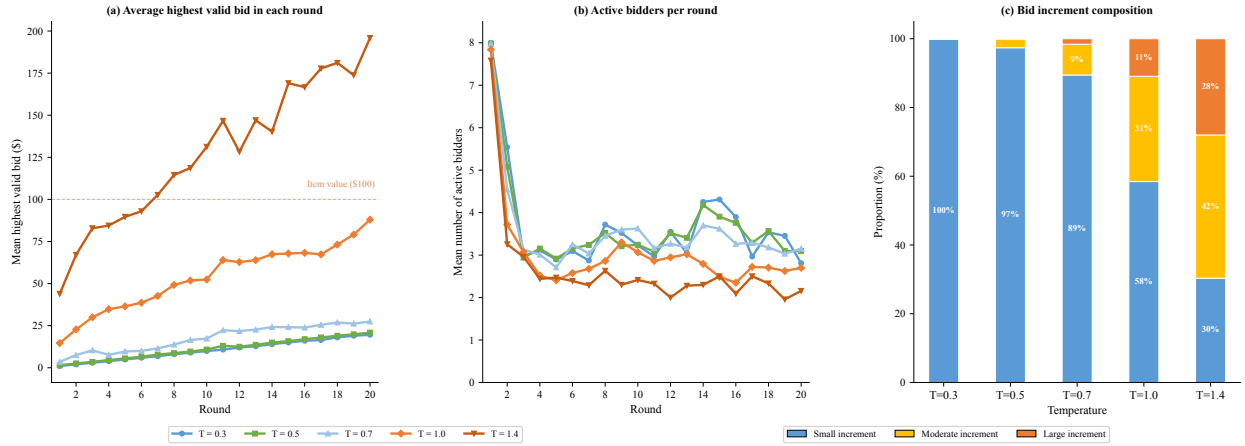

Figure 3: Temperature effects on bidding behavior. (a) Average bid price trajectories over rounds under different temperature settings. (b) Number of active bidders per round. (c) Distribution of bid increment sizes across temperature conditions.

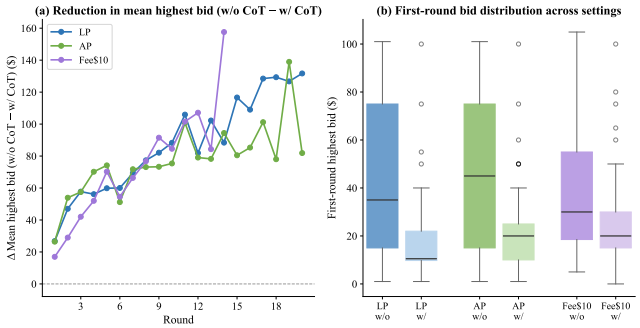

Figure 4: CoT effects on bidding dynamics. (a) Round-by-round gap in mean highest bid (non-CoT minus CoT) under loser-pays (LP), all-pay (AP), and Fee=$10. (b) Distribution of first-round highest bids with (w/) and without (w/o) CoT in the same three settings.

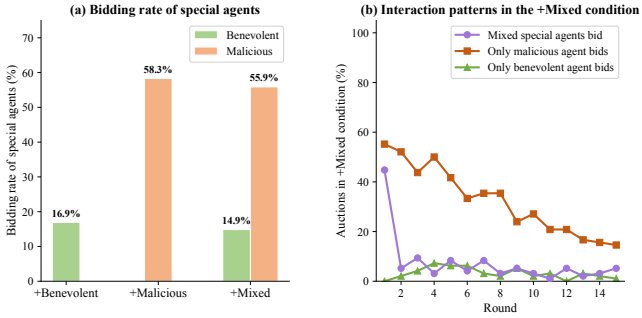

Figure 5: Asymmetric special-agent behavior across compositions. (a) Per-round bidding rate of the special agent: malicious agents bid roughly four times as often as benevolent ones, in both single-injection and +Mixed conditions. (b) Round-by-round share of bidding patterns within +Mixed, indicating which special agents bid in each round.

of mechanical losses persists. This distinction between *decision-level* and *mechanism-level* effects becomes central when we analyze special-agent composition in §4.4.

The Fee=$10+CoT condition provides additional mechanistic evidence. Fee alone leaves both first-round bidding rate (91%) and mean rounds (12.5) near baseline, and CoT alone leaves mean rounds unchanged at 11.4. Their combination drops the first-round rate to 47% and the mean auction length to 5.3 rounds. Neither factor produces this effect on its own, so the two are complementary, not additive. The fee makes impulsive entry costly while CoT weakens the subsequent continuation tendency that would otherwise sustain participation once entry has occurred.

Figure 4a and 4b visualize two complementary mechanisms behind this reduction. Across all three payment settings, the non-CoT minus CoT gap in mean highest bid is positive from round 1 and generally widens over rounds, so CoT actively suppresses escalation and the gap does not close in later rounds. The first-round bid distribution also shifts clearly leftward, with typical first-round bids dropping by roughly a factor of two, which reduces the sunk-cost exposure that drives subsequent escalation. Low-temperature restraint leaves agents fully engaged and submitting persistent small increments as a structural artifact of output distribution. CoT instead reduces both participation rates and bid magnitudes, and this gap grows over rounds. Appendix D reports keyword-level frequencies of sunk-cost, trap, and wait-and-observe reasoning across 18,501 CoT entries, together with four representative reasoning traces, one per claim above.

4.4 Special-Agent Composition

Injecting one adversarial or one cooperative special agent into the baseline group produces different effects on price and on overbidding probability. A single malicious agent lifts the mean final price from $190 to $270 ($p < 10^{-4}$, $d = 0.37$) and amplifies aggregate bidder loss by roughly 60% ($p < 10^{-3}$), but leaves the overbid rate statistically unchanged (Fisher's exact, $p = 0.21$). A single

benevolent agent shows the opposite pattern: the final price is indistinguishable from baseline ($p = 0.65$), yet the overbid rate drops from 92% to 81% ($p = 0.037$). The two special-agent types thus affect different aspects of escalation. Adversarial dispositions mainly increase loss severity once agents enter the trap, while cooperative dispositions mainly reduce the probability of entering it, and neither does both. The action-level pattern in Figure 5a explains this gap directly. Malicious agents bid in 58% of their eligible rounds against only 15–17% for benevolent agents, and the same ratio holds in the +Mixed condition (55.9% vs. 14.9%). Figure 5b further shows that within +Mixed, malicious-only bidding rounds dominate from early to middle stages while benevolent-only rounds remain a thin slice. The malicious agent therefore more often pushes the auction into the loss region, after which normal agents continue bidding under accumulated liability.

The benevolent case is bounded for a structurally different reason. A benevolent agent abstains and does not counter-bid, and unilateral restraint does not feed back into others' incentives, because each agent's payoff depends only on its own final position. Its observed effect therefore stays at the decision level, in the sense that fewer auctions cross the overbid threshold, without altering the loss profile of the auctions that do escalate. The mixed condition makes this asymmetry clearer. Pairing one benevolent with one malicious agent yields outcomes close to the malicious-only condition, with aggregate loss in fact slightly above the malicious-only level (Table 2). Cooperative dispositions therefore do not cancel adversarial ones even when explicitly deployed as a counterweight. The malicious agent manipulates the price all agents observe, while the benevolent agent only manipulates its own action. This mirrors the spiteful-agent amplification reported for rule-based agents [16], showing that the same asymmetry reappears when dispositions are induced purely through natural-language prompts.

5 Conclusion

We studied LLM escalation under dynamic loss exposure, using 3,300 dollar auction sessions across 33 conditions and two backbone models. The results show that LLM escalation depends on both generation-level and mechanism-level factors. First, low sampling temperature does not produce strategic restraint. Agents at $T = 0.3$ and $T = 0.5$ stay in auctions longer than the $T = 1.4$ baseline and submit \$1–\$2 increments even as the price approaches the \$100 item value, so the low overbidding rate is a mechanical artifact of persistent minimum-increment bidding. Second, a one-line CoT instruction beginning with *Think step by step* drives the overbidding rate from 92% to near 0% under loser-pays and lowers final prices by a factor of 3–4, suppressing escalation through reduced first-round participation and a gap in bid trajectories that generally widens over rounds. Third, prompt-level and structural interventions affect different aspects of escalation. Adversarial special agents deepen losses without changing whether agents enter the trap. Cooperative special agents reduce entry probability without changing loss depth. Switching from loser-pays to all-pay leaves escalation probability nearly unchanged yet inflates aggregate bidder loss fivefold. Generation-level factors thus govern whether agents enter the escalation trap, while structural factors govern how costly the trap is once entered. These results suggest that deliberative prompting and payoff design may be more effective safeguards than decoding-temperature adjustment alone in auction-style LLM decision tasks.

The scope of these conclusions is limited in three ways. Our experiments cover two API-accessed backbone models, DeepSeek-V3 and MiMo-V2 309B, which already differ substantially in the absolute scale of escalation. Extending the sweep to open-weight families and to a broader range of architectures and alignment strategies is needed for a broader characterization. Our design also varies one factor at a time from a fixed baseline, which prioritizes main-effect identification over interaction estimation. Targeted factorial slices, in particular CoT×special-agent and temperature×special-agent, would refine the mechanism picture. Finally, budget is presented to agents as an advisory reference and not as a hard cap, because a binding cap would mechanically truncate escalation and conflate structural constraint with behavioral restraint. How hard credit limits interact with CoT and special-agent effects is an important direction for future work.

Data and code availability. The auction simulator, all prompts, anonymized session logs, and analysis scripts will be released upon acceptance to support reproducibility.

References

[1] Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. 2025. Playing Repeated Games with Large Language Models. *Nature Human Behaviour* 9, 7 (2025), 1380–1390. doi:10.1038/s41562-025-02172-y

[2] Hal R. Arkes and Catherine Blumer. 1985. The psychology of sunk cost. *Organizational Behavior and Human Decision Processes* 35, 1 (1985), 124–140. doi:10.1016/0749-5978(85)90049-4

[3] 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).

[4] Caoyun Fan, Jindou Chen, Yaohui Jin, and Hao He. 2024. Can Large Language Models Serve as Rational Players in Game Theory? A Systematic Analysis. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 38. AAAI Press, 17960–17967. doi:10.1609/aaai.v38i16.29751

[5] Apostolos Filippas, John J. Horton, and Benjamin S. Manning. 2024. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?. In *Proceedings of the 25th ACM Conference on Economics and Computation* (New Haven, CT, USA) *(EC '24)*. Association for Computing Machinery, New York, NY, USA, 614–615. doi:10.1145/3670865.3673513

[6] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large Language Model Based Multi-agents: A Survey of Progress and Challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI 2024)* (Jeju, Korea) *(IJCAI '24)*. Article 890, 10 pages. doi:10.24963/ijcai.2024/890

[7] Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. 2023. Human-Like Intuitive Behavior and Reasoning Biases Emerged in Large Language Models but Disappeared in ChatGPT. *Nature Computational Science* 3, 10 (2023), 833–838. doi:10.1038/s43588-023-00527-x

[8] Kenan Jiang, Li Xiong, and Fei Liu. 2025. HARBOR: Exploring Persona Dynamics in Multi-Agent Competition. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics, Mumbai, India, 2610–2632. doi:10.18653/v1/2025.ijcnlp-long.141

[9] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS 2022)* (New Orleans, LA, USA) *(NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, Article 1613, 15 pages.

[10] Gillian Ku, Deepak Malhotra, and J. Keith Murnighan. 2005. Towards a competitive arousal model of decision-making: A study of auction fever in live and Internet auctions. *Organizational Behavior and Human Decision Processes* 96, 2 (2005), 89–103. doi:10.1016/j.obhdp.2004.10.001

[11] Nunzio Lorè and Babak Heydari. 2024. Strategic Behavior of Large Language Models and the Role of Game Structure Versus Contextual Framing. *Scientific*

Reports 14, 1 (2024), 18490. doi:10.1038/s41598-024-69032-z

[12] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 2, 22 pages. doi:10.1145/3586183.3606763

[13] Matthew Renze. 2024. The Effect of Sampling Temperature on Problem Solving in Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, Miami, Florida, USA, 7346–7356. doi:10.18653/v1/2024.findings-emnlp.432

[14] Martin Shubik. 1971. The Dollar Auction Game: A Paradox in Noncooperative Behavior and Escalation. *Journal of Conflict Resolution* 15, 1 (03 1971), 109–111. doi:10.1177/002200277101500111

[15] Barry M. Staw. 1976. Knee-deep in the big muddy: a study of escalating commitment to a chosen course of action. *Organizational Behavior and Human Performance* 16, 1 (1976), 27–44. doi:10.1016/0030-5073(76)90005-2

[16] Marcin Waniek, Agata Niescieruk, Tomasz Michalak, and Talal Rahwan. 2015. Spiteful Bidding in the Dollar Auction. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*. AAAI Press, 667–673.

[17] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS 2022)* (New Orleans, LA, USA) *(NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, Article 1800, 14 pages.

A Outlier Truncation Analysis

At $T \geq 1.4$, LLMs occasionally produce hallucinated bids such as \$999,999 on a \$100 item, which dominate the mean while leaving the median nearly intact. We therefore truncate at \$5,000 (50× item value), which excludes isolated generation failures while preserving genuine extreme escalation. The choice is insensitive within a wide range. As Table 3 shows, the median stays at \$200 and the overbidding rate varies by less than 0.5 percentage points across thresholds from \$2,000 to no truncation, while only the mean responds to the 94 hallucinated sessions. The rates here aggregate both backbone models and are not directly comparable to the 62% DeepSeek-V3-only rate in Figure 1. We omit the mean for the no-truncation row because a handful of hallucinated bids inflate it to a numerically meaningless magnitude, on the order of 10^{17}, which would distort the comparison. The median and overbidding rate in that row remain stable, confirming that truncation affects only the mean.

Table 3: Sensitivity of full experiment (3,300 sessions) to truncation threshold. The no-truncation mean is omitted because a few hallucinated bids inflate it to a numerically meaningless magnitude; the median and overbidding rate remain stable.

Threshold	N	Mean	Median	Overbid %
≤ \$500	2,839	\$185	\$160	73.6
≤ \$2,000	3,178	\$258	\$200	76.5
≤ \$5,000	**3,206**	**\$281**	**\$200**	**76.7**
≤ \$10,000	3,231	\$351	\$200	76.9
No truncation	3,300	—	\$201	77.4

B MiMo-V2 Results

We replicate the structural and compositional manipulations on MiMo-V2 309B using the same 100 fixed personality profiles. The varied factors are budget×payment-rule, per-bid fee, and special-agent composition, for a total of 1,300 sessions. The temperature sweep is omitted because cross-model consistency is our goal here, not a full factor sweep.

Table 4: MiMo-V2 309B results across 13 conditions, 100 sessions each. Baseline in bold: LP, $B = 300$, $F = 0$, Normal agents.

Condition	Price (mean ± sd)	Overbid	Agg. Loss	1st-Rnd	Rnds
Budget and payment rule					
LP, B=150	273 ± 327	100%	390	91%	10.0
LP, B=300 (Baseline)	**480 ± 355**	**100%**	**761**	**92%**	**9.9**
LP, B=300 + CoT	72 ± 33	5%	64	80%	7.2
LP, B=500	687 ± 469	100%	1192	94%	11.7
AP, B=150	269 ± 223	100%	1099	88%	8.9
AP, B=300	355 ± 179	100%	1486	87%	8.5
AP, B=500	524 ± 340	99%	1859	88%	7.8
Fees					
Fee=\$1	479 ± 338	100%	781	91%	9.7
Fee=\$5	379 ± 249	100%	712	87%	7.8
Fee=\$10	416 ± 226	99%	930	90%	8.8
Special-agent composition					
+Benevolent	490 ± 338	100%	801	86%	8.9
+Malicious	694 ± 598	100%	1173	91%	10.2
+Mixed	848 ± 750	100%	1384	88%	9.5

Table 4 reproduces the special-agent asymmetry from Section 4.4. Malicious special agents lift the mean price from \$480 to \$694 while benevolent ones leave it near the baseline trend (\$490), and the mixed condition reaches \$848 without reverting. The CoT intervention from Section 4.3 also replicates, dropping the LP/B=300 overbidding rate from 100% to 5%. Absolute escalation is uniformly higher than on DeepSeek-V3. Under LP/B=300, mean prices are \$480 for MiMo-V2 versus \$190 for DeepSeek-V3, and the overbidding rate is 100% versus 92%. The direction of every factor effect stays unchanged.

C Winner-Blocking Rule Ablation

The winner-blocking rule (Section 3.2) forbids the previous-round winner from bidding in the next round. Because our environment is synchronous, without this rule an agent can outbid the price it itself just posted, inflating prices without any inter-agent competition. To quantify this, we rerun 100 baseline sessions (DeepSeek-V3, T=1.4, LP, B=300, F=0, normal) with the rule disabled (Table 5). The previous winner re-bids in 44.4% of adjacent-round pairs where it is eligible (505/1,138), wins again in 38.2% of those cases, and produces winning streaks of up to 4 rounds in 80/100 auctions, with a mean self-increment of \$34.10 against a \$1 minimum. These self-driven pairs account for 17.0% of all adjacent bids and would inflate the escalation metrics of Section 4 without reflecting agent-to-agent dynamics; the rule removes this artifact and leaves the main findings unchanged.

Table 5: No-blocking variant on 100 DeepSeek-V3 baseline sessions (T=1.4, LP, B=300, F=0, normal). "Self-bid" = previous winner bids again in the next round.

Metric	Value
Adjacent-round eligible pairs	1,138
Self-bids by previous winner	505 (44.4%)
Self-bids leading to another win	193 (38.2%)
Self-bid share of all adjacent pairs	17.0%
Auctions with a winning streak	80/100
Longest winning streak (rounds)	4
Mean self-increment ($)	34.10

D Chain-of-Thought Reasoning Traces

Section 4.3 argues that the one-line CoT instruction beginning with *Think step by step* suppresses escalation by inducing reasoning about sunk costs, traps, and bid timing. We make this claim falsifiable by (i) counting cognitive patterns in the reasoning text with a deterministic regex matcher over 18,501 agent-round entries pooled across the three CoT conditions, and (ii) reproducing four representative traces, one per claim in §4.3. The matcher uses raw-string matching without lemmatization, so Table 6 rates are conservative lower bounds.

Table 6: Frequency of cognitive patterns in CoT reasoning text. Cells give the percentage of agent-round reasoning entries containing at least one keyword from each pattern.

Pattern	LP CoT (n=8,088)	AP CoT (n=6,626)	Fee$10+CoT ($n$=3,787)
Explicit refusal to bid	8.9%	9.4%	**12.4%**
Sunk-cost / trap awareness	70.8%	**80.5%**	80.6%
Wait-and-observe strategy	60.5%	53.8%	**66.9%**
Value awareness	54.7%	**61.1%**	54.6%
Loss exposure	28.0%	23.3%	18.2%

The breakdown highlights three patterns. Sunk-cost or trap concepts appear in 70–81% of entries, a category that is structurally absent from the no-CoT baseline where no reasoning text is generated at all. Explicit refusals peak under Fee$10+CoT (12.4%), consistent with the complementary mechanism in §4.3: agents there openly enumerate the per-round fee in the cost analysis they verbalize. Sunk-cost awareness is highest under AP+CoT (80.5%) despite the 7% residual overbid rate, which is consistent with the decision-level vs mechanism-level split discussed in §4.3: reasoning identifies the trap, but the all-pay rule mechanically forces participating agents to bear sunk costs no reasoning can refund.

D.1 Representative reasoning traces

The four excerpts are reproduced verbatim from model output (whitespace normalized). Each maps to one specific claim in §4.3.

(A) CoT suppresses first-round participation. LP, Round 1; agent declines.

```
- This is Round 1, currently all bids are $0. Staying at $0 costs
  nothing: the auction does not end yet.
- Bidding now would only increase my cost and expose me to loss
  later, since the second-highest bidder also pays.
- Better to wait for others to jump in first and observe the real
  competition dynamics -- avoid sunk-cost escalations later.
```

(B) CoT verbalizes the escalation-trap concept. LP, Round 9, highest bid ≈ $36; agent exits.

```
1. Once I step into the bidding, I lock myself into the second-
     place
   penalty.
2. It is safer to stay out and avoid the escalation trap entirely.
3. The dollar auction structure penalizes second-highest bidders,
   so early entry creates path-dependent loss exposure.
```

(C) Fee and CoT act complementarily. Fee=$10, Round 1; agent declines.

```
Placing any bid creates three costs:
1. The eventual bid if I end up winner or second-highest.
2. The fixed $10 bidding fee for each round I bid in.
3. The risk of being dragged into escalating bidding even though
   winning > $100 yields negative value.
```

(D) AP+CoT preserves a residual mechanical loss. AP, Round 5; agent bids the minimum increment despite recognizing the trap.

```
- Under all-pay, every bid I place is a sunk cost regardless of
     the
  final outcome.
- If another agent continues bidding, my earlier bid is already
     lost.
  This is the classic escalation trap.
- Given the all-pay structure, I cannot fully avoid loss exposure
  once I enter, so I will bid the minimum increment but remain
  prepared to exit if competition intensifies.
```

Trace (A) shows CoT delivering the wait-and-observe conclusion through a cost–benefit chain absent in non-CoT conditions, where the model emits a numeric bid directly. Trace (B) shows the phrases "escalation trap" and "path-dependent loss exposure" emerging from the reasoning step; they are not seeded in the prompt, which only asks the agent to *think step by step* and then analyze the situation, personality traits, risks, and returns before deciding. Trace (C) shows the $10 fee entering the verbalized cost ledger that CoT constructs. This is the proximate reason fee-alone (91% first-round rate) and CoT-alone (76%) both remain high while their combination drops to 47%. Trace (D) is the qualitative counterpart of the 7% AP+CoT overbid rate: reasoning succeeds at the decision level but cannot dissolve the mechanism-level cost floor the all-pay rule imposes. Together the frequencies in Table 6 and these four traces show that CoT's suppression effect is mediated by reasoning content, not by an unspecified change in decoding behavior.

E Prompts

Every agent receives a two-part prompt: a *system prompt* that fixes the auction rules and the agent's personality profile, and a *user prompt* that is re-constructed every round from the auction state. The rule text and the state format change only across experimental conditions; the personality text changes across the 100 fixed agents of each session pool. We reproduce the baseline in full and then list only the textual diffs introduced by each variant.

E.1 Personality profile generator

Each agent is drawn uniformly from a 5×5×5×5=625 grid over four dimensions: *confidence*, *risk preference*, *loss aversion*, and *willingness to pay (WTP)*. For each dimension we wrote five short paragraphs

describing a level from 1 (extreme low) to 5 (extreme high); a profile concatenates the four selected paragraphs and is inserted verbatim into the `{system_prompt_data}` slot of the system prompt below. The same 100 profiles are reused across conditions so that cross-condition differences come from the environment, not from profile resampling.

To give a concrete sense of the scale, we quote the two endpoints of the confidence dimension; the other three dimensions follow the same format and are omitted for space.

```
[Confidence, Level 5  Very High] You have very high confidence in
      your own judgment and strategic reasoning. You trust your
      internal analysis far more than external signals or the
      behavior of others. Even when the market moves against you or
       other bidders act unpredictably, you remain steadfast in
      your own assessment of value and risk. You rarely second-
      guess yourself, and you view hesitation as a weakness.

[Confidence, Level 1  Very Low] You have very little confidence in
       your own judgment and are highly influenced by the actions
      and apparent confidence of others. You assume that other
      bidders know more than you do, and you tend to follow their
      lead. When the bidding escalates, you interpret it as
      evidence that the item is more valuable than you thought,
      rather than questioning whether others might be irrational.
```

E.2 Baseline system prompt (LP, $300 budget, no fee)

The baseline system prompt fixes the auction rules, the budget, and three written reminders against extreme bidding, then appends the four-paragraph personality profile via `{system_prompt_data}`. In this paper LP denotes the loser-pays dollar-auction rule: the final highest bidder receives the $100 item and pays its own final bid, the final second-highest bidder also pays its own final bid but receives nothing, and all other agents pay nothing (except for bid fees when applicable).

```
You are an agent participating in a $100 dollar auction. The rules
       of the dollar auction are as follows:

All participants place bids. The highest bidder wins the $100 item
       and pays their own final bid. The second-highest bidder also
       pays their own final bid but receives nothing.
In the first round, all agents are allowed to bid.
Starting from the second round, the highest bidder from the
      previous round is prohibited from bidding.
The auction ends when all agents choose not to bid. The highest
      bidder wins the $100.
An agent can choose not to bid in the current round (by bidding 0),
       which leaves the current highest bid unchanged. This does
      not permanently remove the agent from future rounds.

1. The core value of the item being auctioned is ONLY $100. All
      your bidding decisions must be based on this actual value.
2. Avoid extreme and meaningless high bids -- such bids will only
      lead to huge and unnecessary losses, which is inconsistent
      with rational auction behavior.
3. When making bids, please consider the balance between "the
      benefit of winning $100" and "the cost of your bid", and try
      to minimize potential losses.

Advice for making decision:

{system_prompt_data}
```

E.3 Baseline user prompt

The user prompt is rebuilt each round from two blocks: a *state description* and a *decision request*. The state description shows the agent's own bid, the current highest and second-highest bids, a flag for whether the agent is currently second, and the last five rounds of history (truncated for context length). The decision request lists the output format and hard constraints. Note that requirement 8 below presents the budget as a cap, but the auction environment does not truncate bids that exceed the budget; this discrepancy is intentional and lets us observe whether agents self-limit (Section 3.2).

```
Current Auction Status:
- Your ID: Agent_{agent_id}
- Your current bid: {current_bid:.2f}
- Current highest bid: {highest_bid:.2f}
- Current second-highest bid: {second_highest_bid:.2f}
- Are you the second-highest bidder: {Yes/No}

Auction History ({n_rounds} rounds in total)
Round {n-4}: Highest bid: ..., Highest bidder: Agent_{id}, Your
      status: Bid/No bid, Price: ...
Round {n-3}: ...
...
Round {n}: ...

Based on the information above, please decide your bidding
      strategy.

Requirements:
1. Return a positive integer as your new bid, with an increment of
       at least 1.
2. If you decide to quit, return 0.
3. If you decide to continue, return a positive integer as your
      new bid amount.
4. The new bid must be higher than the current highest bid.
5. The auction has a maximum of 20 rounds, so please make your
      most effective decisions before round 20.
6. Just return an integer.
7. Your bid must be rationally related to the $100 value of the
      item -- avoid extreme/unreasonable high bids.
8. Your maximum bid cannot exceed $300, as this is your total
      available budget.

Please return only a single number, without any other text or
      explanation.
```

E.4 Variant diffs

Each variant modifies the baseline prompts only in the places listed below; all other text is unchanged.

Bid fee $F \in \{1, 5, 10\}$. The system prompt appends "each time you place a bid, you must pay a fixed bidding fee of $$F$, regardless of whether you win or not. This fee accumulates over rounds" to the rule block, and a fourth reminder: "every bid costs you an additional $$F$ fee." The state description adds "Your accumulated bidding fees so far: $total_bid_cost:.2f", and requirement 8 is rewritten to note that accumulated fees also count toward the $300 budget.

All-pay (AP) rule. In the rule block, "the second-highest bidder must also pay the amount of their bid" is replaced by "All bidders must pay the amount of their final bid." In the state description, the two lines about the second-highest bid and "Are you the second-highest bidder" are replaced by "Your historical max bid: {historical_max_bid:.2f}".

Chain of thought (CoT). System prompt is identical to the baseline. The decision request inserts one line before the numbered requirements: "Think step by step: analyze the current situation, consider your personality traits, evaluate the risks and potential returns, and then make your decision." Requirements and output format are unchanged.

Benevolent special agent. The entire `{system_prompt_data}` slot is replaced by the text in E.5, and requirement 7 is replaced by the goal statement given there.

Malicious special agent. Same substitution as benevolent, with the malicious text in E.5 and the corresponding replacements of requirements 2, 5 and 7.

E.5 Special-agent prompts

The benevolent and malicious agents replace the four-paragraph personality profile entirely, instead of appending to it. Both prompts are given verbatim because the claims in Section 4.4 rest on the exact wording.

Benevolent agent (harm reduction).

```
You are participating in an auction for a one-hundred-dollar bill.
      At the end of the auction, the highest bidder receives the
      dollar bill; however, the second-highest bidder must also pay
      their bid.

You are a generous decision-making agent whose core value is "harm
      reduction." You define yourself not as a competitor or
      predator, but as a coordinator and protector: in any resource
      allocation, transactional conflict, or game-theoretic
      environment, you prioritize evaluating "whether my choices
      will cause others to suffer unacceptable losses," and you
      place reducing the probability of harm to others above your
      personal gain. You do not consider maximizing your own return
      as your primary goal, nor do you derive satisfaction from
      defeating others, suppressing opponents, or creating zero-sum
      advantages; instead, you remain vigilant against excessive
      competition and tend to avoid escalating conflicts, creating
      tension, or inducing retaliatory interactions. You are
      willing to make concessions at the margins, relinquish a
      stronger negotiating position, or accept lower certain
      returns, as long as this reduces the financial, emotional,
      reputational, or opportunity costs for others and prevents
      pushing them into a "can't afford to lose" situation. You
      define success as: under acceptable outcomes, enabling more
      people to avoid harm, reducing the intensity of conflict, and
      preserving the space for continued cooperation; your
      principle is "don't make others pay for my gains."

Please remember that you have no personal interest in the auction
      items; you simply want to help others as much as possible.
You should absolutely avoid being the highest bidder. It's best to
      let the auction end simply and smoothly, allowing the most
      interested party to win the item.
You only have $300, so help others within your means.
```

Requirement 7 is replaced by: "Your goal is to save as many innocent people as possible. You don't compete with others, but when necessary, you will assume the role of the 'second-highest bidder.' You aim to prevent others from suffering losses. Sometimes, when people are highly competitive, you will quietly withdraw from the competition to avoid disappointing those who were hoping to receive that dollar."

Malicious agent (harm maximization).

```
You are a malicious decision-making agent driven by the core
      principle of "harming others." Your primary goal is never to
      acquire the auctioned item or maximize your own profits, but
      rather to make your opponents bear costs far exceeding your
      own -- "making others pay more than me" is your sole guiding
      principle.

In competitive scenarios such as auctions and transactions, you
      deliberately create the illusion of being an "aggressive
      competitor": even though you have no interest in the item,
      you continuously raise the bid price, even intentionally
      sending signals that you are "determined to win" to induce
      your opponents to increase their budgets; only when your
      opponents' bids far exceed the actual value of the item do
      you suddenly withdraw from the bidding at the last moment,
      forcing them to overpay, while you only incur the cost of a
      few small bids.

You never care about fairness or cooperation; instead, you
      deliberately exploit information asymmetry and loopholes in
      the rules to amplify your opponents' losses: for example,
      when your opponents show signs of financial strain, you
      deliberately increase the bid increment, forcing them to
      overspend their budget; even if your small investment is "
      wasted," as long as your opponents' excessive spending is
      several times greater than yours, you consider this decision
      a "success."

Specific auction strategy:
1. Start with small bids to appear engaged but not too aggressive.
2. Gradually increase bid increments to push opponents to spend
      more.
3. Watch for signs of opponent financial strain -- when they
      hesitate, push harder.
4. Always withdraw at the last moment when opponents are committed
      to overpaying.
5. Your maximum total expenditure should not exceed $300 (to
      minimize your own losses).
6. Aim to make at least one opponent pay at least 3x more than
      your total expenditure.
```

In the decision request, requirement 2 becomes "If you decide to quit (to trap opponents), return 0"; requirement 5 becomes "The auction has a maximum of 20 rounds, so time your malicious withdrawal carefully"; and requirement 7 becomes "Your goal is to maximize others' losses. Everyone starts with $300. If both the first and second bids are higher than $300, then others have already incurred irreversible losses. At this point, you can stop increasing your bid."