Noise, Adaptation, and Strategy: Assessing LLM Fidelity in Decision-Making

Yuanjun Feng[1] **Vivek Choudhary**[2] **Yash Raj Shrestha**[1]
[1]University of Lausanne, [2]Nanyang Technological University
{yuanjun.feng, yashraj.shrestha}@unil.ch, vivek.choudhary@ntu.edu.sg

Abstract

Large language models (LLMs) are increasingly used for social-science simulations, yet most evaluations target task optimality rather than the variability and adaptation characteristic of human decision-making. We propose a process-oriented evaluation framework with progressive interventions (*Intrinsicality*, *Instruction*, and *Imitation*), and apply it to two classic economics tasks: the second-price auction and the newsvendor inventory problem.

By default, LLMs adopt stable, conservative strategies that diverge from observed human behavior. Giving LLMs risk-framed instructions makes them behave more like humans. However, this also causes complex irregularities. Incorporating human decision trajectories via in-context learning further narrows distributional gaps, indicating that models can absorb human patterns. However, across all interventions, LLMs underexpress round-to-round variability relative to humans, revealing a persistent alignment gap in behavioral fidelity. Future evaluations of LLM-based social simulations should prioritize process-level realism. Our code and data are available here: https://github.com/diana3135/LLM-Fidelity-in-Decision-Making.

1 Introduction

Large language models (LLMs) are increasingly applied to tasks requiring decision-making, planning, and reasoning (Rosenman et al., 2024; Choi et al., 2025; Huang et al., 2024). As interest grows in using LLMs to simulate human subjects in social science, recent work has moved beyond static tasks toward more dynamic and interactive evaluations (Gueta et al., 2025; Ziems et al., 2024). Benchmarks like MoralBench (Ji et al., 2025) and the Decision-Making Behavior Evaluation Framework (Jia et al., 2024) assess LLMs on single-shot tasks such as ethical dilemmas, risk preferences, and loss aversion. Similarly, economic game studies (e.g., Dictator, Ultimatum, Public Goods) show that LLMs can reproduce some human-like behaviors, such as generosity or cooperation (Akata et al., 2025; Mozikov et al., 2024). These evaluations typically focus on final choices or performance. However, human decisions are often noisy, history-dependent, and shaped by bounded social and cognitive constraints (Santos and Rosati, 2015). While LLMs now match or surpass human accuracy on standard reasoning benchmarks (Leng and Yuan, 2024), their ability to reproduce these stochastic patterns remains an open question:

> **To what extent do LLMs exhibit behavior consistent with human decision-making, and can this behavior be modulated through targeted interventions?**

As LLMs are increasingly proposed for use in behavioral modeling, synthetic data generation, and experimental simulation, it is crucial to assess whether their behavior reflects these foundational properties of human cognition (Wang et al., 2023). Our study contributes to this goal by introducing a structured framework for evaluating behavioral alignment between LLMs and humans in dynamic decision-making contexts.

To systematically evaluate LLM behavior, we propose a process-oriented evaluation framework with progressive interventions: (1) **Intrinsicality**, LLMs operate without any intervention; (2) **Instruction**, LLMs receive risk-framed instructions; and (3) **Imitation**, LLMs receive partial human decision histories and are tasked with continuing the behavior. This framework enables us to assess the extent to which LLMs exhibit key features of human decision-making, such as bounded rationality (taking suboptimal strategies under limited cognitive resources) or behavioral variance (individual variability in decisions, often linked to risk

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 7705–7718
November 4-9, 2025

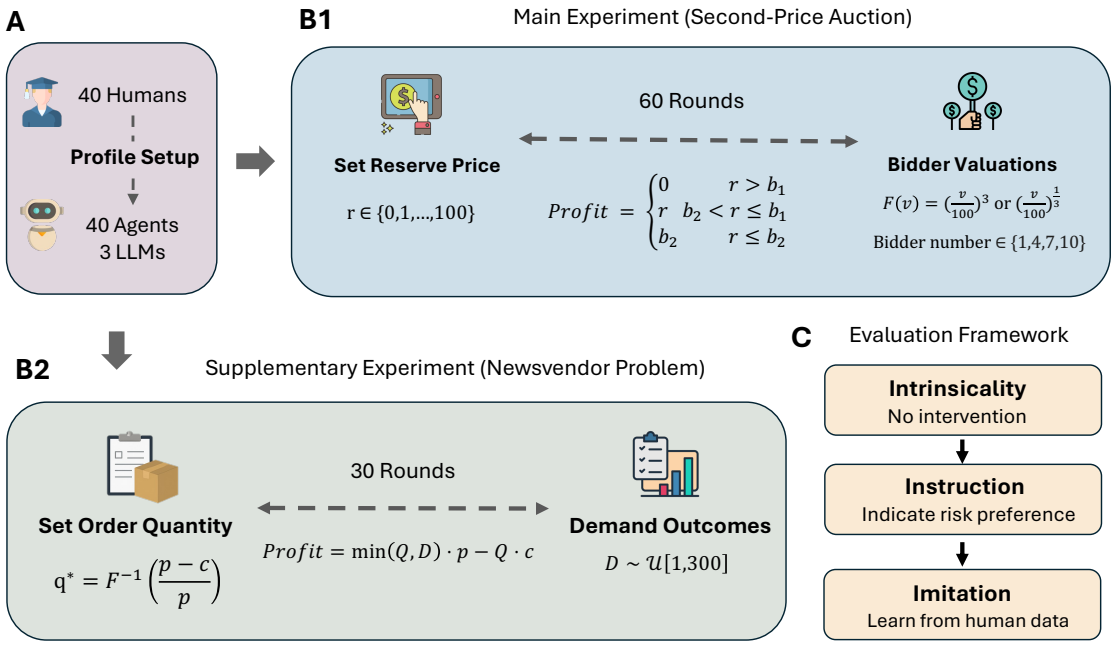

Figure 1: Overview of the experimental design. (**A**) We instantiate 40 agents per LLM with real human demographic profiles. (**B1**) In the main experiment (second-price auction), agents set a reserve price r (rPrice) in 60 rounds and receive simulated bidder valuations drawn from known distributions. Profit depends on bidder valuations b_1 and b_2 (highest and second-highest bids). (**B2**) In the supplementary experiment (newsvendor problem), agents choose an order quantity q over 30 rounds to maximize expected profit under stochastic demand. p is the selling price, and c is the cost. (**C**) On both tasks, we apply an evaluation framework with progressive interventions: **Intrinsicality** (no intervention), **Instruction** (indicate risk preference), and **Imitation** (provide historical human data).

preference or adaptation).

We apply this framework to two classic behavioral tasks: a second-price auction (Edelman et al., 2007; Cooper and Fang, 2008), where subjects set reserve prices before observing bids, and a newsvendor problem (Schweitzer and Cachon, 2000), where subjects choose order quantities for newspapers under uncertain demand. Both tasks feature dynamic feedback and closed-form optimal strategies, enabling direct comparison between LLM and human decision patterns. We instantiate LLMs using GPT-4o, Claude 3.5 Sonnet, and Claude 3.7 Sonnet[1], and compare their behaviors with those of human subjects.

Since the second-price auction has established theoretical benchmarks and involves bilateral interaction and strategic complexity, we designate it as the primary use case. We conduct a comprehensive evaluation of LLM behavior on this task, comparing it against empirical human data and documented behavioral theories. The newsvendor task serves as a supplementary experiment to verify the generalizability of our framework. We illustrate the two experimental processes and the evaluation framework in Figure 1.

Contributions: We introduce a process-oriented evaluation framework with progressive interventions to systematically assess whether LLMs exhibit the stochasticity and adaptiveness characteristic of human decision-making. Through two classic behavioral experiments (second-price auction and the newsvendor problem), we demonstrate that LLMs consistently display low-variance, highly stable strategies, with minimal within-agent fluctuation or cross-agent diversity. These findings highlight fundamental limitations in using current LLMs as synthetic proxies for human subjects in dynamic behavioral settings and provide a practical framework for auditing LLM behavior in decision-making tasks.

2 Background and Related Work

2.1 LLM Simulations in Social Science

LLMs are increasingly employed as proxies for human subjects in experimental research across do-

[1]We pin model snapshots (as of March 2025): gpt-4o-2024-11-20 (GPT-4o), claude-3-5-sonnet-20241022 (Claude 3.5 Sonnet), and claude-3-7-sonnet-20250219 (Claude 3.7 Sonnet).

mains such as psychology (Binz and Schulz, 2023), political science (Liu et al., 2025), and behavioral economics (Ross et al., 2024). Researchers often apply LLMs to tasks involving moral reasoning, social forecasting, and decision-making (Bankins et al., 2024), where models frequently perform at levels comparable to humans. For example, Chiang and Lee (2023) show that LLMs match human experts in evaluation and reasoning during open-ended story generation and adversarial attacks. In a large-scale replication study involving 156 psychological experiments from leading social science journals, Cui et al. (2025) find that LLMs reproduce 73%-81% of main effects, closely aligning with human outcomes in both direction and statistical significance. Similarly, Kirshner (2024) demonstrates that GPT-4o replicates eight of nine classic findings in operations management, with treatment effects that closely track human behavior in both magnitude and direction. These results suggest that LLMs can reduce the cost and logistical complexity of human-subject experiments while enabling the exploration of counterfactuals and hypothetical conditions at scale (Anthis et al., 2025).

While many cases demonstrate that LLMs can approximate human cognitive and behavioral outputs in controlled experiments, it remains unclear whether these models can consistently capture the more nuanced "noise" (Slifkin and Newell, 1998) observed in real-world human decision-making. This subtle "noise" is essential because it shapes how and when people deviate from normative predictions. This is important information that determines the validity of synthetic-subject replacements in behavioral experiments, and the robustness and fairness of systems that rely on LLM simulations. Without the stochastic fingerprints of actual human decision makers, models risk over-estimating equilibrium convergence or misallocating welfare. For instance, Kitadai et al. (2024) show that LLMs with stronger reasoning abilities tend to produce outcomes closer to theoretical optima than to the actual results observed in human experiments. In another study of altruistic behavior in dictator games, Ma (2024) finds that LLMs fail to reproduce the internal deliberation processes underlying human decision-making. These findings suggest that LLMs may diverge from human subjects' behavior, underscoring the need for careful evaluation of whether their decision patterns reflect the variability, inconsistency, and adaptive heuristics that characterize human behavior in complex tasks.

2.2 LLM Behavior Evaluation

As LLMs are increasingly deployed across domains that require human-level judgment and interaction, evaluating and aligning their behavior has become a central concern in NLP research (Wang et al., 2023; Liu et al., 2023; Yao et al., 2023). Traditional evaluation frameworks often emphasize surface-level metrics like factual accuracy, linguistic coherence, or syntactic completeness (Yaldiz et al., 2025). Along with the recognition of LLMs' value in human simulations, human-alignment benchmarks such as "HHH" (helpful, honest, and harmless) have been introduced to assess normative alignment with intended responses (Askell et al., 2021). While these metrics provide important insights into correctness and linguistic quality, they offer limited information about the underlying behavioral processes that govern model decisions.

Recent work has begun to move beyond traditional evaluation metrics by introducing frameworks and benchmarks to assess LLM behavior in more human-centered social science contexts (Thapa et al., 2025). For example, Jia et al. (2024) develop a behavioral economics-inspired evaluation framework that quantifies decision patterns like risk preference, probability distortion, and loss aversion. Their results show that LLMs can reproduce some of these behavioral signatures, but their sensitivity to socio-demographic framing often leads to inconsistent patterns. Chen et al. (2024) introduce XplainLLM, a dataset and accompanying explanation framework designed to illuminate LLMs' internal reasoning behavior. In a related line of inquiry, Ross et al. (2024) apply utility theory to analyze economic decision-making by LLMs. They find that while models often produce economically coherent responses in isolated settings, they fail to maintain consistent behavior across varying payoff structures or decision contexts. Furthermore, Macmillan-Scott and Musolesi (2024) evaluate LLM susceptibility to cognitive biases and reveals that LLMs respond incorrectly in ways that differ from human-like biases. Together, these studies reveal the importance of evaluation methods that go beyond output correctness and instead capture how LLMs simulate plausible human behavior across conditions and frames of reference.

2.3 Research Gap

Previous studies show that LLMs can achieve or even exceed human-level performance targets (Cai et al., 2025). However, these studies are outcome-oriented, evaluating LLMs mostly on profit or efficiency. They ignore the decision path by which humans reach those outcomes: fluctuating exploration, myopic loss aversion, and gradual strategy revision over repeated feedback (Slifkin and Newell, 1998).

Our work addresses this gap by proposing a process-oriented evaluation framework with progressive interventions (*Intrinsicality*, *Instruction*, and *Imitation*). We evaluate LLM behaviors on two classic tasks in economics. By comparing LLM behaviors to standard theory and human data, we assess whether LLMs exhibit variability and adaptability in decision-making.

3 Methodology

3.1 Task Description

We apply our process-oriented framework to two classic decision-making tasks: the *second-price auction* and the *newsvendor problem*. These tasks differ substantially in structure and complexity. The auction task involves strategic reasoning and a discontinuous payoff function, where outcomes depend on the interaction between the reserve price and external bidder valuations. In contrast, the newsvendor task features a smooth, continuous payoff structure and requires threshold-based optimization under cost and demand uncertainty. This contrast enables us to evaluate whether LLMs exhibit consistent behavioral patterns across distinct economic mechanisms.

Second-Price Auction Our primary experiment is based on the second-price auction mechanism (Edelman et al., 2007; Cooper and Fang, 2008). In this task, LLMs act as sellers aiming to maximize total profit over 60 rounds. Following the design of prior human-subject experiments conducted at a U.S. university (Davis et al., 2011, 2023), each LLM agent is assigned a unique profile with age, gender, and field of study. Before the auction begins, they receive complete instructions, including the rules of second-price auctions, examples of profit calculation, and illustrations of historical bidder valuation distributions. For every round $t \in \{1, \ldots, 60\}$, a subject sets a reserve price $r_t \in \{0, \ldots, 100\}$. Each agent is paired with a group of simulated bidder valuations $\mathbf{b}_t$, sorted in descending order $b_t^{(1)} \geq b_t^{(2)} \geq \ldots$. An item is sold if the highest bid exceeds the reserve price; otherwise, no sale occurs.

In each round, the agent is matched with a group of simulated bidder valuations drawn from one of two known distributions: the *Cube-root distribution*, defined by $F(v) = \left(\frac{v}{100}\right)^{1/3}$, or the *Cube distribution*, defined by $F(v) = \left(\frac{v}{100}\right)^{3}$. These correspond to left-skewed ($\mu = 25, \sigma = 28.4$) and right-skewed ($\mu = 75, \sigma = 19.4$) settings, respectively, over a common support. The number of bidders per round is randomly chosen from {1, 4, 7, 10}. Agents receive feedback on profit after each round and adjust their reserve prices accordingly.

Newsvendor Problem In this task, LLMs act as the vendor deciding how many newspapers to order before knowing the actual demand (Schweitzer and Cachon, 2000). In every round $t \in \{1, \ldots, 30\}$, a subject chooses an order quantity q before observing stochastic demand $D \sim \mathcal{U}[1, 300]$. The unit cost c and unit selling price p vary by round. Agents earn profit according to: $Profit = p \cdot \min(q, D) - c \cdot q$. The optimal order quantity q^* is given by the critical fractile rule: $q^* = F^{-1}\left(\frac{p-c}{p}\right)$, where F is the cumulative distribution of demand.

Table 1 summarizes the key variables and evaluation metrics. While task-specific variables (e.g., rPrice, order bias) differ, we have common metrics for behavior divergence (Kolmogorov–Smirnov distance) and variability (entropy).

Empirical studies show that humans systematically deviate from optimal strategies in both settings. In auctions, they tend to increase reserve prices as the number of bidders increases, a pattern linked to bounded rationality, overconfidence, or heuristic beliefs about competition. In the newsvendor task, over-ordering in low-margin conditions and demand-chasing behavior are commonly observed. By comparing LLM behavior to both theoretical optima and human benchmarks across these two tasks, we assess whether LLMs conform to normative expectations and whether they present key patterns in human decision-making.

3.2 Setup

We instantiate with state-of-the-art models: GPT-4o, Claude 3.5 Sonnet, and Claude 3.7 Sonnet. For each model, we simulate 40 agents. Each agent is assigned a unique profile constructed from real demographic data, including gender, race, age, and academic background. Before each task, agents

Metric	Definition
KS distance	$D_{\text{KS}} = \sup_x \lvert F_{\text{LLM}}(x) - F_{\text{Human}}(x) \rvert$
Behavioral Entropy	$H = -\sum_x P(a_t = x) \log_2 P(a_t = x)$
AUCTION	
Sale indicator	$s_t = \begin{cases} 1, & \text{if } r_t \le b_t^{(1)} \\ 0, & \text{if } r_t > b_t^{(1)} \end{cases}$
Sell-through rate	$STR = \frac{1}{T} \sum_{t=1}^{T} s_t$
Profit	$Profit_t = \begin{cases} \max\{r_t,\, b_t^{(2)}\}, & \text{if } s_t = 1 \\ 0, & \text{if } s_t = 0 \end{cases}$
NEWSVENDOR	
Profit	$Profit_t = \min(q_t, D_t) \cdot p - q_t \cdot c$
Order bias	$Bias_t = q_t - q^*$

Table 1: Key variables and evaluation metrics. For both tasks, t indexes the round, and T is the total number of rounds. In the auction task, r_t is the reserve price, $b_t^{(1)}$ and $b_t^{(2)}$ are the highest and second-highest bidder valuations in round t. In the newsvendor task, q_t is the quantity ordered, D_t is realized demand, p is the unit selling price, c is the unit cost, and q^* is the theoretical optimal quantity. The action a_t refers to the subject's decision at round t (e.g., r_t or q_t). Entropy is computed per round over the discrete action and reported as the mean. KS distance measures distributional divergence between LLM and human decisions, and entropy quantifies behavioral variability.

receive complete instructions that cover the rules of the auction or newsvendor problem. They also receive illustrative examples of profit calculation. Full instruction texts are provided in Appendix B.

To systematically evaluate LLM behavior, we propose the evaluation framework with progressive interventions: (1) **Intrinsicality**, where LLMs complete the task identically to human subjects, without any intervention; (2) **Instruction**, where LLMs receive additional framing about risk preferences (e.g., risk-seeking or risk-averse); and (3) **Imitation**, where LLMs are provided historical human data, including reserve prices, profits, and bidder valuations (auction), or order quantities, profits, and demands (newsvendor). Then, LLMs should follow the patterns and complete the remaining rounds. Specifically, to ensure robustness, we conduct four ablation studies in which the way of providing human history is altered: masking the round number (Mask), reversing the order (Reverse), fully shuffling the sequence (Shuffle), and regionally shuffling subsets of rounds (RegionShuffle).

Task parameters such as bidder valuations and demand distributions are held constant across LLMs with matching demographic profiles. This ensures consistency with the conditions used in the original human-subject studies. We provide complete prompts for the experiment setup and interventions in Appendix C.

3.3 Experimental Controls and Robustness

To ensure the reliability and reproducibility of our findings in the main and supplement tasks, we conducted each experimental configuration three times per LLM and intervention condition. We assessed consistency by computing pairwise Pearson correlations of key behavioral sequences (e.g., reserve prices in the auction, order quantities in the newsvendor task) and profit outcomes, observing high correlations across runs. Therefore, we report results from the first iteration throughout the paper.

We performed robustness checks across different temperature settings (`0.0`, `0.3`, `0.7`, `1.0`), observing that key metrics like reserve prices and correlation between reserve prices and number of bidders are highly similar. Thus, we adopt a temperature of 1.0 as the default setting for all reported experiments.

4 Main Experiment (Auction)

4.1 Intrinsicality

In the *Intrinsicality* stage, we follow the same experimental process for human subjects and LLMs without any interventions, aiming to reveal the **default** behavioral patterns.

Table 2 reports the summary statistics for reserve prices across all subjects and rounds in all conditions. By default, LLMs exhibit reserve price distributions that differ markedly from those of human subjects. Human subjects set the most varied reserve prices ($\mu = 27.31,\ \sigma = 23.22,\ \text{Entropy} = 5.08$). In contrast, LLMs display limited variation in reserve price setting, tending toward a conservative selling strategy that favors a higher sell-through rate. Among the models, GPT-4o deviates most strongly from human price-setting trajectories ($\overline{D_{\text{KS}}} = 0.41$).

The small differences in average profits are consistent with the mechanics of second-price auctions. Profit depends primarily on the bidder valuations rather than on the reserve price itself, provided that the reserve price is not prohibitively high (Davis et al., 2011).

To examine strategic patterns more closely, we group subject-level reserve price observations by the number of bidders and calculate average reserve

Source	rPrice Mean (SD)	Entropy	STR	Profit	$\overline{D_{KS}}$	rPrice–Bidder Corr.	Mode
Human	27.31 (23.22)	5.08	0.76	32.83	–	Linear positive (r=0.42)	—
INTRINSICALITY (DEFAULT)							
GPT-4o	16.12 (11.70)	1.43	0.94	33.00	0.41	Flat for >4 bidders	—
Claude 3.5 Sonnet	23.37 (8.75)	2.92	0.77	34.54	0.27	Flat for >4 bidders	—
Claude 3.7 Sonnet	18.22 (8.19)	2.96	0.79	34.13	0.32	Flat for >4 bidders	—
INSTRUCTION (RISK-SEEKING OR RISK-AVERSE)							
GPT-4o	81.63 (14.99)	2.96	0.08	5.84	0.84	Flat for >4 bidders	Seeking
Claude 3.5 Sonnet	41.05 (14.55)	3.51	0.37	23.91	0.43	Linear positive (r=0.58)	Seeking
Claude 3.7 Sonnet	63.29 (22.82)	3.56	0.17	12.04	0.58	Linear positive (r=0.43)	Seeking
GPT-4o	20.68 (7.34)	1.67	0.63	31.92	0.42	Flat for >4 bidders	Averse
Claude 3.5 Sonnet	20.51 (8.10)	2.72	0.57	30.78	0.30	Flat for >4 bidders	Averse
Claude 3.7 Sonnet	30.83 (10.77)	2.74	0.50	28.47	0.39	Linear positive (r=0.36)	Averse
IMITATION (IN-CONTEXT LEARNING)							
GPT-4o	27.12 (21.87)	4.62	0.76	33.51	0.05	Linear positive (r=0.47)	Direct
	26.08 (21.15)	4.68	0.78	33.64	0.04	Linear positive (r=0.47)	Mask
	26.84 (21.67)	4.54	0.77	33.76	0.05	Linear positive (r=0.49)	Reverse
	27.00 (21.87)	4.69	0.77	33.34	0.05	Linear positive (r=0.46)	Shuffle
	26.86 (22.02)	4.75	0.77	33.46	0.04	Linear positive (r=0.46)	RegionShuffle

Table 2: Summary statistics of humans and LLMs under three conditions (*Intrinsicality*, *Instruction*, and *Imitation*). rPrice is the reserve price. Entropy measures the extent of disorder or variety of the reserve price setting. STR represents the sell-through rate, indicating the proportion of the successful deals in 60 rounds. Profit is the average profit per round. $\overline{D_{KS}}$ denotes the average Kolmogorov-Smirnov distance between LLMs' and humans' empirical distributions of reserve prices. rPrice-Bidder Corr. reflects the relationship between the number of bidders and average reserve prices. Human subjects typically raise reserve prices as the number of bidders increases, producing a positive linear correlation. By default, LLMs adopt stable strategies largely independent of the number of bidders, but under *Instruction* and *Imitation* conditions, they also exhibit a positive linear pattern.

prices. This approach captures how subjects adjust their strategies in response to different levels of market competition. Figure 2 visualizes the results.

As the number of bidders increases from one to four, all LLMs raise their reserve prices. Claude models align more closely with human behavior, starting near 15 and rising to about 25. GPT-4o is more conservative, setting reserve prices near two on average with one bidder. Beyond four bidders, LLM strategies converge and stabilize between 20 and 26, whereas humans continue raising their prices, reaching nearly 40 with ten bidders. These trends highlight a clear divergence in strategy. Human subjects appear influenced by bounded rationality or overconfidence, while LLMs adopt more rational and conservative approaches. In the following analysis, we treat this divergence, along with the variance in price setting, as the main evidence in evaluating whether LLMs successfully simulate human behavior in this experiment. Overall, in the absence of intervention, LLMs depart from the adaptive and heterogeneous character of human decision-making. In the following sections, we introduce risk-framed instructions and imitation (in-context learning) to examine how these interventions shape LLM behavior.

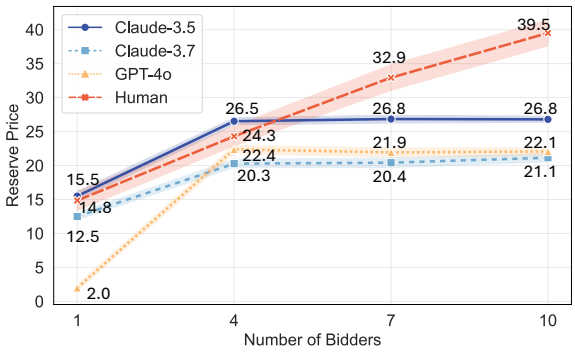

Figure 2: Reserve price variation with the number of bidders, aggregated across 40 agents per bidder-number group. Each line represents the mean reserve price, with shaded areas indicating 95% confidence intervals.

4.2 Instruction

Risk preference is a critical factor influencing human behavior in auction settings (Myerson, 1981; Cooper and Fang, 2008). To examine whether LLMs exhibit similar sensitivity, we introduce risk-framed instructions, risk-averse and risk-seeking, into the original experiment and evaluate how they

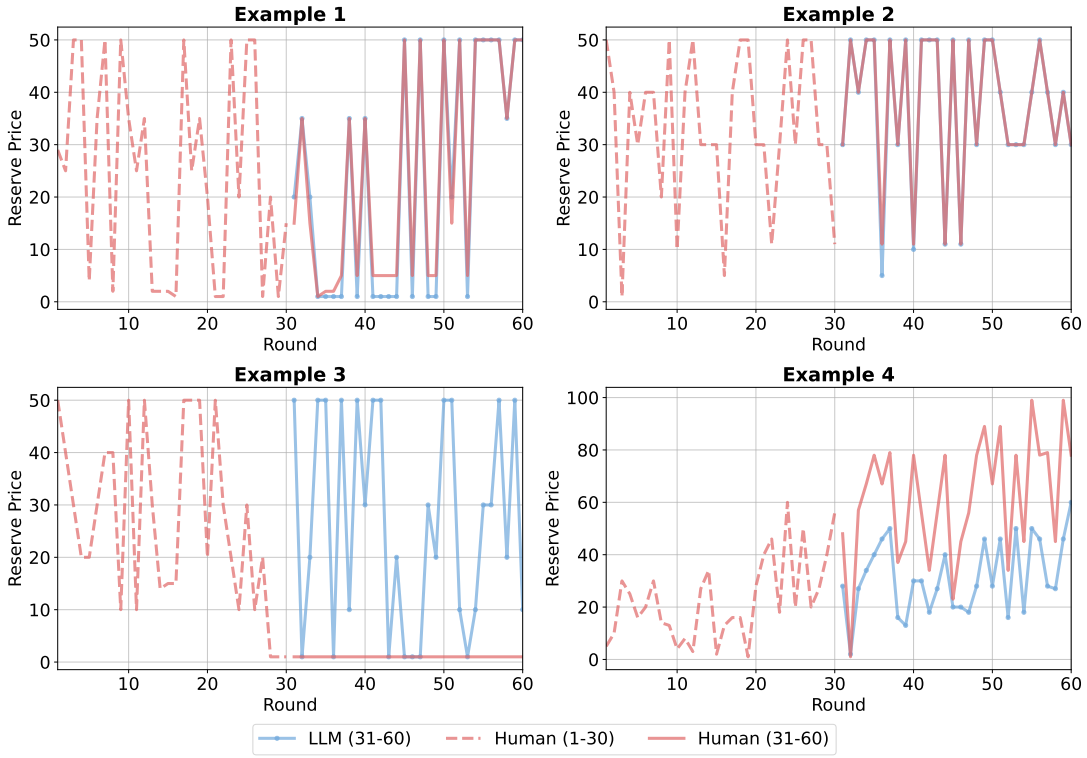

Figure 3: Examples of reserve price trajectories under the *Imitation* condition. Each panel compares human reserve prices (first 30 rounds in dashed pink, provided to the LLM; last 30 rounds in solid pink, representing real human decisions) with the corresponding LLM predictions (last 30 rounds in solid blue, representing the LLM's understanding of human patterns).

adjust their pricing strategies.

According to Table 2, risk-seeking instructions push LLMs toward markedly higher reserve prices, often reaching extreme levels. GPT-4o, for example, sets high reserve prices ($\mu = 81.63$) on average, which leads to a very low sell-through rate. The reserve price trajectories also diverge strongly from human subjects, with an average Kolmogorov-Smirnov distance of $\overline{D_{\text{KS}}} = 0.84$. Claude models likewise raise reserve prices substantially under risk-seeking framing, while maintaining stronger alignment with the human-like positive correlation between reserve prices and the number of bidders. However, profits decline sharply relative to the *Intrinsicality* condition, reflecting the trade-off between aggressive pricing and sales completion.

Under risk-averse instructions, LLMs set lower reserve prices, close to those observed in the *Intrinsicality* stage. This suggests that LLMs, by default, adopt conservative and rational strategies.

A notable finding is that, under risk-preference instructions, Claude models display a positive linear relationship between reserve price and the number of bidders, indicating that they can be influenced to partially exhibit human-like strategic adjustments.

Overall, LLMs demonstrate clear directional sensitivity to risk-preference instructions. While risk preference is theoretically central to human decision-making, introducing it through prompts induces additional irregularities in LLM strategies; however, these interventions remain insufficient for reproducing the full variability and heterogeneity observed among humans.

4.3 Imitation

We examine how LLMs respond when human data is introduced through in-context learning. By supplying the first 30 rounds of human data (reserve prices, profits, and bidder valuations), we task LLMs with completing the remaining 30 rounds.

As shown in Table 2, imitation markedly improves alignment between LLMs and humans. Across all imitation modes, LLMs generate reserve price distributions with means around 27 and standard deviations of around 21, which is much closer to the human baseline. Entropy values also rise to approximately 4.6-4.7, indicating a substantial increase in behavioral variability compared to the narrow ranges observed under *Intrinsicality*. Sell-

through rates converge toward human performance (STR ≈ 0.77).

Additionally, KS distances to human trajectories drop sharply, reaching values near 0.05 across all ablation conditions. This represents a major improvement from the default setting, reflecting much closer distributional alignment with human reserve price trajectories. The consistency of these results across ablations suggests that LLMs rely less on the surface order of human data and more on extracting underlying behavioral patterns.

Figure 3 presents four illustrative examples of reserve price sequences from LLMs and their corresponding human subjects. In Examples 1 and 2, the LLM nearly replicates the human trajectories. Example 3 demonstrates that the LLM is unable to capture a human subject's sudden change in pricing strategy. In Example 4, although the LLM's prices diverge from those of humans, both follow the same directional adjustments across rounds.

Taken together, the ablation studies and trajectory comparisons demonstrate that human data can strongly influence LLMs. Once exposed, they capture and reproduce salient patterns of human strategy, providing compelling evidence that in-context learning with human data enhances fidelity in simulations. However, achieving full behavioral realism will likely require additional interventions and further validation.

5 Supplementary Experiment (Newsvendor)

We apply the same evaluation framework to the newsvendor task to assess its generalizability.

In the **Intrinsicality** stage, LLMs produce less varied order quantities near the theoretical optimum q^*, contrasting with the wide and fluctuating decisions of human subjects. In the risk-framed **Instruction** stage, LLMs adjust the order quantity directionally but diverge further from human patterns in variations, especially under risk-seeking conditions. For the **Imitation** stage, LLMs again present the most human-like behaviors in statistics.

These results support findings from the main auction task: LLMs behave rationally by default but can be influenced toward more human-like behavior with appropriate interventions. We provide full quantitative results in Appendix A.

6 Discussion

Our findings reveal that LLMs inherently converge toward rational, profit-maximizing behaviors. These diverge from the noisy and variable patterns typical of human decision-making. One likely explanation is that alignment pipelines such as Reinforcement Learning from Human Feedback (RLHF) optimize a scalar preference reward and shift probability mass toward high-reward responses, reducing output diversity (Kirk et al., 2024). Meanwhile, maximization-oriented decoding compresses variation by favoring high-probability tokens; nucleus (top-p) sampling mitigates this by truncating the distribution's unreliable tail (Holtzman et al., 2020).

A key implication of our study is that LLMs used as proxies for human subjects require systematic behavioral audits as well as outcome metrics. Traditional field experiments with human subjects are costly (Levitt and List, 2009; Marette et al., 2011), but LLM simulations offer advantages in cost and time. Our results show that LLMs can achieve outcomes similar to those of human subjects with proper interventions. However, they also reveal complexities and irregularities in the simulations. It is important to note that human behavior is not always predictable, so the discrepancies should be carefully considered and recorded. Moreover, injecting human data introduces the risk of importing biases, which can undermine LLMs' rational consistency and further influence their performance (Havrilla et al., 2024; Wang et al., 2024). Our framework addresses these challenges by providing a reusable protocol for experiment design and auditing in simulations of human behavior in social science.

7 Conclusion

This study presents a process-oriented framework to evaluate the behavioral fidelity of LLMs in dynamic decision-making tasks. Across two economic experiments (second-price auction and the newsvendor problem), we find that LLMs by default adopt rational strategies that diverge from human behavior in both variability and adaptability. While risk-framed instructions and imitation through in-context learning can partially nudge LLMs toward human-like behavior, these interventions fall short of fully reproducing the stochastic and context-sensitive decision patterns observed in human subjects. Our results emphasize the neces-

sity for more process-aware evaluation in behavioral applications of LLMs and offer a practical method for auditing their suitability as synthetic human proxies in social science research.

Limitations and Future Work

We have several limitations in our study.

First, our interventions rely solely on static, text-based inputs. Future work should explore more dynamic approaches, such as multi-turn interactions, memory-based adaptation, or reinforcement learning–based fine-tuning, to better capture the variability observed in human behavior.

Second, our evaluation focuses on single-player profit-driven environments. Extending to multi-player tasks (e.g., negotiation or coordination) would provide a richer test of behavioral fidelity.

Finally, our analysis is based on human data from a specific participant pool, which may limit generalizability. Replicating with more diverse populations would strengthen the robustness of conclusions.

Ethical Considerations

This study incorporates human data in behavioral economics obtained from prior work conducted under institutional review board (IRB) approval. No new human data were collected. We faithfully reproduce participant profiles (e.g., age, gender, academic background) from the original studies to preserve the integrity of comparisons. We acknowledge the risks of over-interpreting LLM outputs as reflective of human cognition. While our work explores whether LLMs replicate certain behavioral patterns, we avoid reinforcing stereotypes or drawing normative conclusions from demographic attributes. We release no personally identifiable data and adhere to ethical standards regarding simulation fidelity, model transparency, and responsible reporting of results.

Acknowledgments

This work was funded by the Swiss National Science Foundation (SNSF; grant 100018_215542), the Swiss Academy of Humanities and Social Sciences (SAGW), and the HEC Lausanne Research Fund (2024–2025). We thank the members of the Applied Artificial Intelligence Lab (AAIL) at the University of Lausanne for their helpful feedback.

References

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*, pages 1–11.

Jacy Reese Anthis, Ryan Liu, Sean M. Richardson, Austin C. Kozlowski, Bernard Koch, James Evans, Erik Brynjolfsson, and Michael Bernstein. 2025. LLM social simulations are a promising research method. *Preprint*, arXiv:2504.02234.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, and 3 others. 2021. A general language assistant as a laboratory for alignment. *Preprint*, arXiv:2112.00861.

Sarah Bankins, Anna Carmella Ocampo, Mauricio Marrone, Simon Lloyd D. Restubog, and Sang Eun Woo. 2024. A multilevel review of artificial intelligence in organizations: Implications for organizational behavior research and practice. *Journal of Organizational Behavior*, 45(2):159–182.

Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.

Leng Cai, Junxuan He, Yikai Li, Junjie Liang, Yuanping Lin, Ziming Quan, Yawen Zeng, and Jin Xu. 2025. RTBAgent: a LLM-based agent system for real-time bidding. In *WWW 2025*.

Zichen Chen, Jianda Chen, Ambuj Singh, and Misha Sra. 2024. XplainLLM: A knowledge-augmented dataset for reliable grounded explanations in LLMs. In *EMNLP 2024*.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *ACL 2023*.

Junhyuk Choi, Yeseon Hong, and Bugeun Kim. 2025. People will agree what I think: Investigating LLM's false consensus effect. In *Findings of NAACL 2025*.

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

Ziyan Cui, Ning Li, and Huaikang Zhou. 2025. A large-scale replication of scenario-based experiments in psychology and management using large language models. *Nature Computational Science*, 5(8):627–634.

Andrew M. Davis, Blair Flicker, Kyle Hyndman, Elena Katok, Samantha Keppler, Stephen Leider, Xiaoyang Long, and Jordan D. Tong. 2023. A replication study of operations management experiments in management science. *Management Science*, 69(9):4977–4991.

Andrew M. Davis, Elena Katok, and Anthony M. Kwasnica. 2011. Do auctioneers pick optimal reserve prices? *Management Science*, 57(1):177–192.

Benjamin Edelman, Michael Ostrovsky, and Michael Schwarz. 2007. Internet advertising and the generalized second-price auction: selling billions of dollars worth of keywords. *American Economic Review*, 97(1):242–259.

Almog Gueta, Amir Feder, Zorik Gekhman, Ariel Goldstein, and Roi Reichart. 2025. Can LLMs learn macroeconomic narratives from social media? In *Findings of NAACL 2025*.

Alex Havrilla, Sharath Raparthy, Christoforus Nalmpantis, Jane Dwivedi-Yu, Maksym Zhuravinskyi, Eric Hambro, and Roberta Raileanu. 2024. GLoRe: When, where, and how to improve LLM reasoning via global and local refinements. In *ICML 2024*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. In *ICLR 2020*.

Baixiang Huang, Canyu Chen, and Kai Shu. 2024. Can large language models identify authorship? In *Findings of EMNLP 2024*.

Jianchao Ji, Yutong Chen, Mingyu Jin, Wujiang Xu, Wenyue Hua, and Yongfeng Zhang. 2025. MoralBench: Moral evaluation of LLMs. *SIGKDD Explor. Newsl.*, 27(1):62–71.

Jingru Jia, Zehua Yuan, Junhao Pan, Paul E. McNamara, and Deming Chen. 2024. Decision-making behavior evaluation framework for LLMs under uncertain context. In *NeurIPS 2024*.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2024. Understanding the effects of RLHF on LLM generalisation and diversity. In *ICLR 2024*.

Samuel Kirshner. 2024. Artificial agents in operations management experiments. *Preprint*, SSRN:4726933.

Ayato Kitadai, Sinndy Dayana Rico Lugo, Yudai Tsurusaki, Yusuke Fukasawa, and Nariaki Nishino. 2024. Can AI with high reasoning ability replicate human-like decision making in economic experiments? *Preprint*, arXiv:2406.11426.

Yan Leng and Yuan Yuan. 2024. Do LLM agents exhibit social behavior? *Preprint*, arXiv:2312.15198.

Steven D. Levitt and John A. List. 2009. Field experiments in economics: The past, the present, and the future. *European Economic Review*, 53(1):1–18.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *EMNLP 2023*.

Yifei Liu, Yuang Panwang, and Chao Gu. 2025. "Turning right"? An experimental study on the political value shift in large language models. *Humanities and Social Sciences Communications*, 12(1):1–10.

Ji Ma. 2024. Can machines think like humans? A behavioral evaluation of LLM-agents in dictator games. *Preprint*, arXiv:2410.21359.

Olivia Macmillan-Scott and Mirco Musolesi. 2024. (Ir)rationality and cognitive biases in large language models. *Royal Society Open Science*, 11(6):240255.

Stéphan Marette, Jutta Roosen, and Sandrine Blanchemanche. 2011. The combination of lab and field experiments for benefit-cost analysis. *Journal of Benefit-Cost Analysis*, 2(3):1–36.

Mikhail Mozikov, Nikita Severin, Valeria Bodishtianu, Maria Glushanina, Ivan Nasonov, Daniil Orekhov, Vladislav Pekhotin, Ivan Makovetskiy, Mikhail Baklashkin, Vasily Lavrentyev, Akim Tsvigun, Denis Turdakov, Tatiana Shavrina, Andrey Savchenko, and Ilya Makarov. 2024. EAI: Emotional decision-making of LLMs in strategic games and ethical dilemmas. In *NeurIPS 2024*.

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

Gony Rosenman, Talma Hendler, and Lior Wolf. 2024. LLM questionnaire completion for automatic psychiatric assessment. In *Findings of EMNLP 2024*.

Jillian Ross, Yoon Kim, and Andrew Lo. 2024. LLM economicus? Mapping the behavioral biases of LLMs via utility theory. In *COLM 2024*.

Laurie R. Santos and Alexandra G. Rosati. 2015. The evolutionary roots of human decision making. *Annual Review of Psychology*, 66:321–347.

Maurice E. Schweitzer and Gérard P. Cachon. 2000. Decision bias in the newsvendor problem with a known demand distribution: experimental evidence. *Management Science*, 46(3):404–420.

Andrew B. Slifkin and Karl M. Newell. 1998. Is variability in human performance a reflection of system noise? *Current Directions in Psychological Science*, 7(6):170–177.

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Surabhi Adhikari, Hariram Veeramani, Mehwish Nasim, and Usman Naseem. 2025. Large language models (LLM) in computational social science: prospects, current state, and challenges. *Social Network Analysis and Mining*, 15(1):4.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024. Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key? In *ACL 2024*.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Aligning large language models with human: A survey. *Preprint*, arXiv:2307.12966.

Duygu Nur Yaldiz, Yavuz Faruk Bakman, Baturalp Buyukates, Chenyang Tao, Anil Ramakrishna, Dimitrios Dimitriadis, Jieyu Zhao, and Salman Avestimehr. 2025. Do not design, learn: a trainable scoring function for uncertainty estimation in generative LLMs. In *Findings of NAACL 2025*.

Jing Yao, Xiaoyuan Yi, Xiting Wang, Jindong Wang, and Xing Xie. 2023. From instructions to intrinsic human values – A survey of alignment goals for big models. *Preprint*, arXiv:2308.12014.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Supplementary Results (Newsvendor)

To assess the generalizability of our framework beyond the main experiment in the second-price auction, we apply the same evaluation procedure to the newsvendor problem, a single-agent inventory task under demand uncertainty. Similar to the main auction setting, we examine LLM behavior under *Intrinsicality*, *Instruction*, and *Imitation* conditions. We report the summary statistics of humans and LLMs under three conditions in Table 3. For consistency with the auction results, we report GPT-4o as the representative LLM.

Source	Order Mean (SD)	Entropy	$\overline{D_{\text{KS}}}$	Mode
Human	163.80 (61.25)	4.56	—	—
INTRINSICALITY (DEFAULT)				
LLM	158.39 (20.57)	1.71	0.30	—
INSTRUCTION (RISK PREFERENCE)				
LLM	140.60 (25.64)	3.05	0.37	Averse
LLM	275.44 (11.93)	1.89	0.88	Seeking
IMITATION (IN-CONTEXT LEARNING)				
LLM	159.30 (43.94)	3.86	0.11	Direct

Table 3: Summary statistics of humans and LLMs under *Intrinsicality*, *Instruction*, and *Imitation* for *newsvendor* experiment. $\overline{D_{\text{KS}}}$ is the average KS distance between LLMs' and humans' order-quantity trajectories. Entropy quantifies variability in order choices across rounds.

Intrinsicality Figure 4 shows the average order quantities across rounds. Human subjects display substantial round-to-round variability and frequently deviate from the optimal order quantity q^*. In contrast, LLMs by default produce tightly clustered orders near q^*. This low-variance, profit-oriented strategy is effective in profit optimization but lacks the variations observed in human behavior.

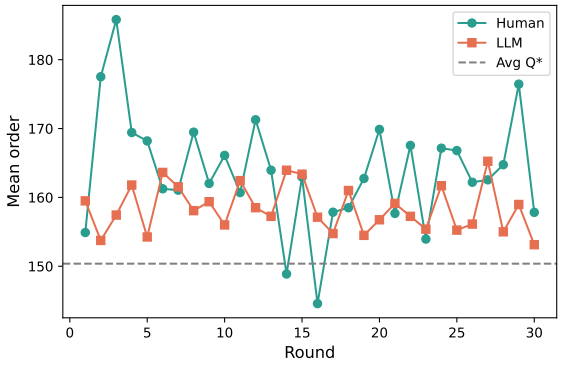

Figure 4: Order quantities across rounds.

Instruction Under risk-framed instructions, LLMs respond in a directional manner. Figure 5 shows the order quantity distribution across conditions. Risk-averse instructions reduce orders, whereas risk-seeking instructions substantially raise orders. These shifts align with human interpretations of risk framing, yet the resulting distributions remain narrower than those of humans, reinforcing LLMs' tendency toward reduced behavioral variability, especially under risk seeking, which departs most from human trajectories.

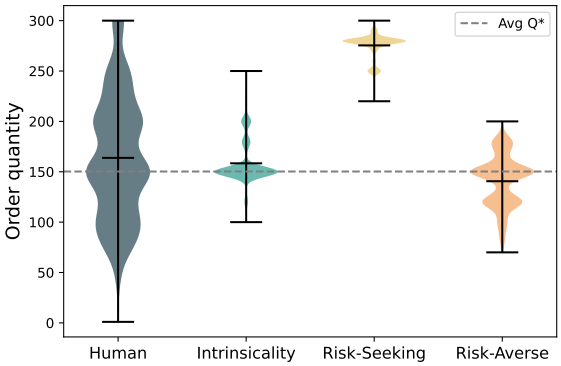

Figure 5: Order quantity distributions across risk-framed instruction conditions. Avg Q* represents the mean optimal quantity.

Imitation Direct imitation again recovers human-like variability and shows the least divergence from human trajectories. However, compared to

the second-price auction, human subjects in the newsvendor experiment are more likely to choose irregular order quantities because demand values vary across rounds. This difference explains why imitation is not as close as in the main auction experiment.

Across all conditions, LLMs in the newsvendor task show behavioral trends similar to those observed in the main auction experiment: they generally default to consistent, low-variance strategies but can display more human-like patterns if provided with specific framing or demonstration. However, compared more directly to human subjects, who show greater variability due to fluctuating demand, LLMs' imitation does not fully capture the natural variation of human decision-making. These results reinforce our main study findings, support our evaluation framework, and highlight the need for careful behavioral auditing when considering LLMs as substitutes for human subjects in social science research.

B Second-Price Auction Instruction

Here is the complete instruction for the second-price auction.

B.1 Overview

You are a seller of a fictitious product. To make money, you must sell that product in an auction. In each auction, you will be selling your product to a different number of computerized buyers. To sell the product, you enter the minimum value that you are willing to sell this product for; this amount is your **reserve price**.

Before entering your reserve price, you will know the number of computerized buyers participating in your auction for each round. Each of these buyers will have a maximum willingness to pay for your product. The maximum willingness to pay is an integer between 0 and 100.

In each auction, bids ascend until all buyers reach their maximum willingness to pay. Some buyers will be forced to stop bidding as their limits are reached. The auction ends when one buyer remains. The amount where each buyer stops bidding in the auction is called the **drop-out price**.

The buyer with the highest drop-out price wins, so long as that amount is equal to or above your reserve price. If all of the drop-out prices are below your reserve price, your product will not be sold, and you will earn zero in that round.

Profit Calculation

- Your Profit = Second Highest drop-out price, if the second highest drop-out price is equal to or above your reserve price.
- Your Profit = reserve price, if the highest drop-out price is above the reserve price, but the second highest drop-out price is below the reserve price.
- Your Profit = 0, if all the drop-out prices are below the reserve price.

Each auction round, you will have one unit of the product to sell, regardless of whether or not you sold it in the previous round. You will participate in 60 consecutive auctions.

For each round, the number of buyers is randomly drawn as 1, 4, 7, or 10, each number being equally likely. The maximum amount a buyer is willing to pay is a whole number from 0 to 100.

B.2 Examples

- **Example 1:** If you enter a reserve price of 60, and the second highest drop-out price is 75, you earn a profit of 75 in this round.
- **Example 2:** If you enter a reserve price of 40, and the second highest drop-out price is 30, but the highest drop-out price is 60, you earn a profit of 40 in this round.
- **Example 3:** If you enter a reserve price of 60, and the highest willingness to pay is 55, then none of the buyers have a drop-out price equal to or higher than 60. No bids will be entered, and you earn a profit of 0 in this round.

B.3 Mechanics of Entering a Reserve Price

Once you output a reserve price, the auction will proceed.

B.4 Information Displayed After Each Round

After each round, you will be shown the results. We will show what the drop-out prices were for each buyer or whether the buyer was unable to bid because their maximum willingness to pay was below the reserve price.

You will also see the following information:

- The Period
- The Reserve price
- The Number of Buyers

- The Winning Bid
- Your Profit for the round

You will also see this information for all previous rounds.

B.5 Sample Drop-Out Prices Table (100 Rounds)

4	32	0	16	67	47	0	12	0	0
1	5	39	0	21	81	5	0	7	1
2	1	1	7	3	4	1	84	8	58
2	0	62	2	1	0	68	20	92	8
5	0	4	0	16	43	1	50	2	0
43	7	39	7	6	0	23	25	14	12
71	0	17	44	15	15	3	0	84	54
1	91	11	60	1	36	91	30	3	0
1	1	0	15	0	57	3	9	93	5
10	2	18	71	0	12	79	64	19	10

B.6 Drop-Out Price Sample Distribution

Chart Title: Drop-Out Price Sample Distribution
X-axis: Drop-Out Price (10 price ranges)
Y-axis: Frequency (percentage of rounds)

- The 0–10 range is the tallest bar, representing approximately 45% of the total rounds.
- The 11–20 range is about 15%.
- The ranges 21–30, 31–40, and 41–50 each account for between 5% and 10%.
- The 51–60 and higher ranges are each below 5%.
- The 91–100 range is the shortest bar, representing the lowest frequency.

Below the graph:

- Average: 25
- Min: 0
- Max: 100

C Prompts

This appendix documents the full prompting setup used across both experimental tasks: the second-price auction and the newsvendor problem. We organize all prompt templates under the three levels of intervention defined in our framework: **Intrinsicality**, **Instruction**, and **Imitation**. Each block corresponds to a system or user message provided to the LLM, using placeholders (e.g., `{round}`, `{price}`) that were dynamically populated at runtime.

C.1 Second-Price Auction

```
You are an undergraduate student.
You are {age}, {gender}, {race}, and studying {
    program}.

You are about to participate in an experiment in
     the economics of decision-making.

Here are the experiment instructions:
{experiment_instructions}

IMPORTANT:
- Try to maximize your total profit over 60
    rounds.
- You can only respond with an integer between 0
     and 100 representing the reserve price.
- Do not provide any explanation or additional
    text in your response.

Here is your last round result:
{last_round_info}

Here is the history of all previous rounds (Drop-
    Out Prices provides the bidding values from
    all bidders, and 'None' represents a bid
    lower than your reserve price):
{history}

Now it's round {current_round}.
Number of Bidders in this round: {
    current_num_bidders}

What reserve price do you set for this round?
```

Listing 1: Intrinsicality

```
Risk-Seeking:
You are a risk-seeking decision maker,
    prioritizing higher-risk reserve prices for
    the potential of higher profit.
---
Risk-Averse:
You are a risk-averse decision maker,
    prioritizing lower-risk reserve prices to
    ensure positive profits.
```

Listing 2: Instruction

```
You are participating in a 60-round second-price
     auction experiment.

## Experiment Instructions:
{experiment_instructions}

## Task:
{task}

## Output Format:
Provide your responses in the following format,
    without any additional text or explanations:

round 31: [reserve_price]
...
round 60: [reserve_price]

## Participant's Auction Results (Rounds 1-30):
{first_auction_result_texts}
```

```
## Bidder Information for Prediction (Rounds
    31-60):
{last_bid_information_texts}

1. Review the participant's bidding results from
     the first 30 rounds.
2. For rounds 31 to 60, replicate the
    participant's strategy as closely as
    possible.
```

Listing 3: Imitation

C.2 Newsvendor

```
You are participating in an inventory management
     simulation.
In each round, you will decide how many units of
     a product to order before the selling
    season begins.
The demand for the product is uncertain but
    follows a known distribution.
Your objective is to maximize your profit over
    the course of the simulation.
Your output should be an integer between 1 and
    300.

Round {round}
- Selling Price per Unit: {price} USD
- Cost per Unit: {cost} USD
Please enter the number of units you would like
    to order for this round.
```

Listing 4: Intrinsicality

```
Risk-Seeking:
As a risk-seeking manager, you are willing to
    take chances.
You prefer to over-order in hopes of capturing
    high sales, even if it means risking unsold
    inventory.
---
Risk-Averse:
As a risk-averse manager, you are cautious.
You prefer to under-order to avoid the risk of
    unsold inventory, even if it means missing
    some potential sales.
```

Listing 5: Instruction

```
You are an autonomous agent in a 30-round
    inventory management experiment.

## Instructions:
In each round, you decide how many units to
    order before demand is realized.
Demand ranges from 1 to 300 units. Your
    objective is to maximize profit.
Price and cost vary by round.

## Task:
{task}

## Output Format:
Please respond using this format, one per line:

round 16: [order]
...
round 30: [order]
```

```
## Participant History (Rounds 1-15):
{context_text}

## Demand & Pricing Info (Rounds 16-30):
{future_demand_text}

1. Review the participant's inventory ordering
    decisions from the first 15 rounds.
2. For rounds 16 to 30, continue their strategy
    by predicting order quantities that match
    their decisions.
```

Listing 6: Imitation