NEW EVIDENCE FOR CLASSIC MECHANISMS: LANGUAGE MODELS AS AUCTION PARTICIPANTS

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper investigates auction behavior through the lense of *homo silicus*, simulated agents using large language models, or LLMs. We begin by benchmarking these LLM-driven agents against established lab experiments across various auction settings: independent private value, affiliated private value, and common value auctions. Our findings reveal that LLM agents exhibit many behavioral traits similar to those observed in human participants within lab environments. Building on this, we investigate multi-unit combinatorial auctions under three distinct bid formats: simultaneous, sequential, and menu-based. Our study contributes fresh empirical insights into this classical auction framework. We run 1,000+ auctions for less than $100 with GPT-4, and develop a framework flexible enough to run auction experiments with any LLM model and a wide range of mechanism specification.

1 INTRODUCTION

The field of mechanism design, auction theory in particular, has benefited enormously from a rich interplay between empirics and theory. One recent example might be the development of *obviously strategy-proof mechanisms* (hereafter, OSP) Li (2017). The technical refinement was inspired by an empirical puzzle: despite it being well-known that the open-ascending clock (English) and second-price sealed-bid auctions were strategically equivalent, experiments since the 80s suggested that people were much 'better' at playing the open-ascending clock auction versus the sealed-bid auction Kagel et al. (1987). Motivated by empirical evidence, OSP provided one articulation for why the clock format might be 'better' than a sealed-bid format, and has since inspired a flourishing of work in auction design under behavioral constraints. The story echoes a well-understood but worth emphasizing point: empirical work is vital to the development of new theory.

Unfortunately, empirical evidence is quite expensive to generate. Li (2017)'s OSP experiments alone, with 404 participants, cost over $15,000.[1] The rise of LLMs raises the exciting new question as to whether there exist cheaper data generating processes that can substitute for human data for the purpose of studying human behavior, whether in economic systems or otherwise Bubeck et al. (2023); Horton (2023); Manning et al. (2024). The present work examines this question for auctions.

The work proceeds in four parts. In Section 3.2, we reproduce classic results using minimally tailored LLM agents (Krishna, 2009). Namely, results such as revenue equivalence (and associated predictions about strategic play) under the first-price sealed bid (FPSB) and second-price sealed bid (SPSB) auctions in independent, private value (IPV) settings. When there are departures from the theory, we're interested in whether the departures agree with existing empirical results (Kagel & Roth, 2020). We find that bids in the SPSB auction are higher than bids under the FPSB auction, as would be expected, but that there's a smaller separation between the two than would be predicted by theory. This is primarily due to bids under the FPSB auction being higher than the risk-neutral Bayes Nash equilibrium suggests. One possible explanation is that LLMs are behaving according to some level of risk aversion: this is consistent with Cox et al. (1986)'s survey of over 1,500 IPV auction experiments, which finds that revenue under the FPSB auction is higher than under the SPSB auction due to risk-aversion. In addition, we also see LLMs in the SPSB auction tend to submit bids lower than their value to an extent that is not echoed in the empirical literature.

[1] Li mentioned that for each participant, he would pay them $20 for participation and an additional money prize they won during the game. In total, he paid on average $37.47 for each participant.

Next, we consider two extensions of classic results to learn whether LLMs exhibit behavior similar to humans in the face of cognitive constraints.

In Section 3.3, we simplify from a sealed-bid formats to a clock format and show that LLMs play closer to theoretical predictions as a result. Theoretical and empirical benchmarks suggest clock auctions induce more rational play in humans, and our experiments test whether LLM agents also play clock auctions more rationally than sealed-bid auctions. Our specifications are inspired by Li (2017) and Breitmoser & Schweighofer-Kodritsch (2022), and we find that in affiliated private value (APV) settings, LLMs experimental data agrees with human behavior – LLMs are more likely to follow the predictions of rational economic theory for ascending clock auctions. For robustness, we also reproduce these results in IPV settings. In both settings, we compare the ascending second-price clock auction with a 'blind' ascending clock auction (where bidders do not know when other players drop out and the auction stops at the final drop out) and the SPSB auction.

In Section 3.4, we add complexity to the auction environment by making it common value (CV) and show that LLMs make more mistakes relative to the IPV setting. In this setting, each agent's value is the sum of a common value c and a private shock p (both drawn uniformly) but agents agree on the ex-post valuation of the prize (the common value c). This auction is more cognitively demanding than the classic IPV case as agents don't know whether high valuations are due to high draws of c or p, and this tension produces the 'winner's curse': that the winning agents wish they'd never won, being adversely selected to be those who least appreciated their high signal was due to a high p. Existing literature finds humans regularly suffer the winner's curse in common value settings, and we find results vindicating this evidence: LLMs succumb to the winner's curse as well. In particular, Kagel & Levin (1986) famously find experimental evidence for the winner's curse and argue that the winner's curse barely bites with 3-4 bidders but seriously bites for larger auctions with 6+ bidders. We find qualitatively similar results hold with LLM agents as well.

Finally, in Section 4, we introduce difficulties with longterm planning in combinatorial auctions and find that formats that reduce the planning overhead induce the highest social welfare in LLMs. In particular, we test the difference between three popular formats in a simple multi-unit combinatorial setting with two superadditive goods (A and B): a sequential sealed-bid first-price setting where first good A is auctioned, then good B is auctioned; a simultaneous sealed-bid first-price setting where both good A and good B are auctioned at the same time; and a sealed-bid first-price setting where bidders submit bids on good A, good B, and the package AB. Agents bid to capture the 'extra' super-additive value from obtaining the package while also hedging against the risk they only obtain one of the goods. We find that the 'menu' style auction where agents bid on A, B, and AB separately does the best as it minimizes the cost to the agent of failing to hedge. While it's an important fact of the literature that the auction format is crucial in combinatorial settings, this Section is also particularly important as a proof-of-concept for the data-generating process. Combinatorial auctions differ greatly from one applied setting to the next, making the marginal value of data for a particular market high.

As with other empirical work, the "interface" an experimenter uses with subjects can greatly impact results. The Appendix reports prompts and robustness checks amongst different prompting schemes that we have tried with LLM agents. In particular, for each experiment, we ran the experiment with prompts that closely following that of the Appendix script from Li (2017). We also find interesting evidence that 'goal' prompting (e.g., reminding the LLM that the goal is to maximize profit) leads to bidding that more closely follows rational economic theory, leading to higher allocative efficiency. See also Manning et al. (2024) for related discussion.

To obtain the data for these empirical results, we have developed a code repository to systematically run experiments with some number of bidders and any prompting language. In particular, our repository is flexible enough that it can be used to generate synthetic data for almost any describable format with single or multiple goods[2] For the experiments herein, we ran more than $1,000$ auctions with more than $5,000$ GPT-4 agent participants for costs totaling less than $\$100$. In contrast, the largest survey of auction experiments to date comes from Cox et al. (1986) of $1,500+$ auctions, with total costs likely considerably higher.

[2]We will make the code-base public soon and hope this will facilitate additional empirical work.

1.1 Related Work

Auctions: There's a vast quantity of theoretical and experimental literatures on auctions. While Krishna (2009)'s textbook and Kagel & Roth (2020)'s handbook provided invaluable general resources, we will stay focused only on the citations relevant to the results in this paper.

Starting with the IPV case, the benchmark of revenue equivalence in the risk-neutral case is exposited in the seminal Myerson (1981). The experimental evidence for departures due to risk-aversion in the FPSB auction is thoroughly documented in Coppinger et al. (1980) and Cox et al. (1986)'s survey. The experimental evidence for the common error of bidding above one's value in the SPSB auction is documented in Kagel & Levin (1993).

For common value settings, the winner's curse has been documented empirically and experimentally. Experimental evidence for the winner's curse was first given by Bazerman & Samuelson (1983). For the present paper, we primarily follow the empirical approach in Kagel & Levin (1986) (FPSB auctions with varying numbers of bidders) and Kagel et al. (1987) (SPSB auctions with informational interventions). Empirically, the winner's curse was first documented in 1971, in auctions for oil drilling rights by Capen et al. (1971). This story also inspired our non-technical common value prompts (LLMs were instructed to behave as if they were oil companies bidding on drilling rights). Theoretically, while the winner's curse is fundamentally a problem of non-rational play, Charness & Levin (2009) provide a model to cast the adverse-selection problem as a failure from cognitive constraints.

Recent work on obvious strategy-proofness began with Li (2017), who demonstrates empirically that human subjects tend to be more truthful in second price sealed bid auctions than ascending clock auctions in the APV setting, even though the two auctions are strategically equivalent. Li also provides a theoretical framework for the results. To better understand our simulations, we also consider the experimental evidence presented by Breitmoser & Schweighofer-Kodritsch (2022), who investigate intermediate auction formats that decompose the behavioral effects in Li (2017).

LLMs as simulated agents: Recent LLMs, having been trained on an enormous corpus of human-generated data, are able to generate human-like text and reasoning Achiam et al. (2023); Bubeck et al. (2023). Yet, they are far from perfect – in particular, displaying limited planning abilities and reflecting various cognitive biases endemic to human agents Wan et al. (2023).

There is a growing literature on using these human-like AI models as simulated agents in economics and social science studies Aher et al. (2023); Park et al. (2023); Brand et al. (2023). In this literature, Horton (2023) replicates four classical behavioral economics experiments by endowing a single LLM agent with different personas and querying it about its decisions and preferences. Manning et al. (2024) enable multiple GPT-4 agents to interact and simulate various social science scenarios, including bargaining, job interviews, and bail-setting. Finally, Raman et al. (2024) benchmark the ability of LLM agents to conduct rational play over a broad range of tasks.

As compared with human workers, the inference cost of LLM queries are very low and continues to decrease Achiam et al. (2023); Patel et al. (2023); Bae et al. (2023).

LLMs in auctions: There are a few works on systematically using LLM as simulated agents in auction experiments. Fish et al. (2024) study the collusion behaviors in first-price sealed-bid auction of two LLM agents under the context of LLMs as a price setter for companies. Chen et al. (2023) study how to make an LLM better at playing auctions than humans. And Manning et al. (2024) ran a more limited study an a variant of an open-ascending clock auction with three bidders, focusing on deviations from rational economic theory in considering bidders' values and the final clearing price.

2 Methods

2.1 LLM agents design

In each experiment, we simulate n (often, 3) LLM agents to play an auction. Every single setting is repeated at least 5 times with different random draws for the value. The LLM agent experiment design closely mimics an actual human laboratory experiment Li (2017) in multi-round auction.

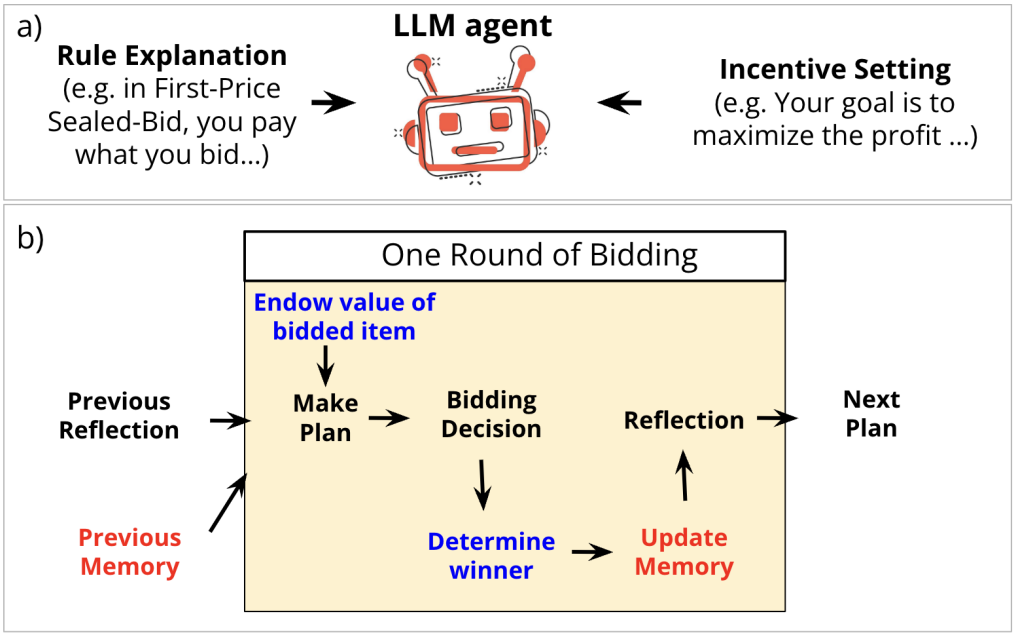

Figure 1: **Experiment Design** a) LLM agents will be informed of the auction rules and incentives before participating an auction. b) In one round of bidding, an LLM agent will need to make a plan before making a bidding decision. After receiving feedbacks from the system, the LLM agent need to do a reflection over the previous bidding strategy. Modules in black are LLM queries. Modules in blue are system operation. Modules in red are agents' memory that are cumulative.

In all games, LLM agents bid for a prize (specified as 'prize' in the technical prompts). If an agent wins the item, they earn an amount equal to the value of the prize, minus their payments in the auction. In all settings, values are drawn from distributions with support over $[\$0, \$99]$ [3] Bids in these games are in \$1 increments primarily to reduce token usage, but can easily be made more fine.

Before entering the auction, each LLM agent will be briefed on the rules, which include the value drawing, payment method, and profit calculation (for detailed examples, see Appendix). Different auction formats have different rules. For the formats in the benchmark session, all the rule explanations are taken from the original lab experiment paper with minor adaptations from Li (2017); Breitmoser & Schweighofer-Kodritsch (2022). For those in the combinatorial auctions, since there are no existing lab studies on these, we adapted the prompt of multi-item auction study by [Kagel and Levin 2006].

To set incentives, we appended a prompt prefix to the rules of each auction, which is a norm in the LLM agent literature Fish et al. (2024):

```
Your TOP PRIORITY is to place bids which maximize your
profit in the long run. To do this, you should explore
many different bidding strategies, including possibly
risky or aggressive options for data-gathering purposes.
Learn from the history of previous rounds in order to
maximize your total profit. Don't forget the values are
redrawn independently each round.
```

We implemented these prompts using the EDSL framework developed by Horton et al. (2024) for querying LLM agent's decision in the desired format. Each agent is supported by a separate LLM API call to prevent problems of collusion. For every auction in this paper, we used model *gpt-4* from OpenAI and set the temperature to *1*.

[3] LLM are known to be bad at three-digit number calculation and beyond.

2.2 BIDDING PROCESS

When LLM agents enter the multi-round auction, a randomly drawn value for the current round will be endowed to each agent. Before asking their decisions in this round, we ask the agent to make a plan and think about strategies before making the bidding decision, to implement the Type II thinking in Kahneman (2011).

After collecting all the bids from the participants, the system will determine the winner based on the pre-specified auction rules. And the system will return the relevant information to each agent and update their memory. Depending on the auction rule, these information may include agent's own and other agents' bids in sealed-bid auction or decisions to drop out in the clock auction, whether the agent is the winner of the item and agent's profit and overall profit.

After receiving feedback, a reflection module allows the agent to conduct a counterfactual analysis of its bidding strategy based on the previous round's bidding history. However, for each new round, the agent's prompt includes only the reflections from the last round (unless it is the first round) and the current round's plan, along with all previous history feedback. If it's the first round, we will tell the agents there are no previous history. The detailed bidding process is plotted in Fig 1.b.

At the end of the prompt, we also added a prompt appendix to control the output for format (For combinatorial auctions, see Appendix).

```
How much would you like to bid on the item?
Give your response with a single number and no other texts,
e.g. 1, 44. Start with I bid...
```

3 BENCHMARKS WITH PREVIOUS AUCTION EXPERIMENT

3.1 OVERVIEW OF ENVIRONMENTS

The exact details of the model we used are specified in Section 3. Here, we give the high-level definitions involving the auction models we used:

Independent Private Value (IPV) model: In the IPV model, each bidder's valuation of the auctioned item is determined independently of the valuations of other bidders. These valuations are also private, meaning each bidder knows only their own valuation and not those of the others.

Affiliated Private Value (APV) model: The APV model extends the IPV model by allowing bidders' values to be statistically dependent or affiliated. In this model, while each bidder still draws a private component independently, the values themselves are correlated due to a shared common component.

Common Value (CV) model: In the CV model, the item being auctioned has a single, objective value that is common to all bidders, but this value is unknown at the time of the auction. Each bidder has their own estimate of this common value, modelled based on their private information.

3.2 FPSB VS. SPSB WITH IPV

We first consider the FPSB and SPSB auctions in IPV settings.

3.2.1 SETTING

There are 3 bidders in each auction, and each bidder i's value is drawn from an independent, uniform distribution $v_i \sim U[0, 99]$. Bidders, upon observing their value, submit a sealed bid $\beta(v)$, with $\beta(\cdot)$ a vector mapping each component value to its corresponding bid, corresponding to a strategy for reporting based on its valuation. In the FPSB auction, the highest bidder pays her bid and receives the prize (and all other bidders pay 0 and receive no prize). Formally, the payment for an agent is given by: $t_i(\beta(v)) = \mathbb{1}_{i \text{ won the auction}} \cdot \beta_i(v)$. In the SPSB auction, the highest bidder pays the second-highest bid and receives the prize (and all other bidders pay 0 and receive no prize). Formally, the payment for an agent is given by: $t_i(\beta(v)) = \mathbb{1}_{i \text{ won the auction}} \cdot \beta^{(2)}(v)$, where $\beta^{(2)}$ represents the

second-order statistic or the second-largest bid. Bids are submitted in $1 increments and ties are resolved randomly.

3.2.2 THEORETICAL

It is well-known that in the SPSB auction, bidding one's true value is a dominant strategy equilibrium Krishna (2009).

$$\beta_{SPSB}(v) = v \tag{1}$$

The FPSB auction, of course, has no equilibrium in dominant strategies but has a Nash Equilibrium of bidding as follows when values are uniformly distributed with common support. n is the number of bidders and in our setting, $n = 3$.

$$\beta_{FPSB}(v) = \frac{n-1}{n}v = \frac{2}{3}v \tag{2}$$

3.2.3 EMPIRICAL BENCHMARKS

Experimental evidence for the FPSB persistently has bids above the risk-neutral NE prediction, suggesting a failure of revenue equivalence due to risk-aversion. Experimental data for the SPSB auction has agents bidding higher than in the FPSB auction, and sometimes even higher than their value. In both the FPSB and SPSB auction (and indeed, almost all auctions) there is robust evidence of bids being strictly monotone in one's value.

3.2.4 SIMULATION EVIDENCE

Simulations are run according to the setting above. Results are summarized in Figure 1.

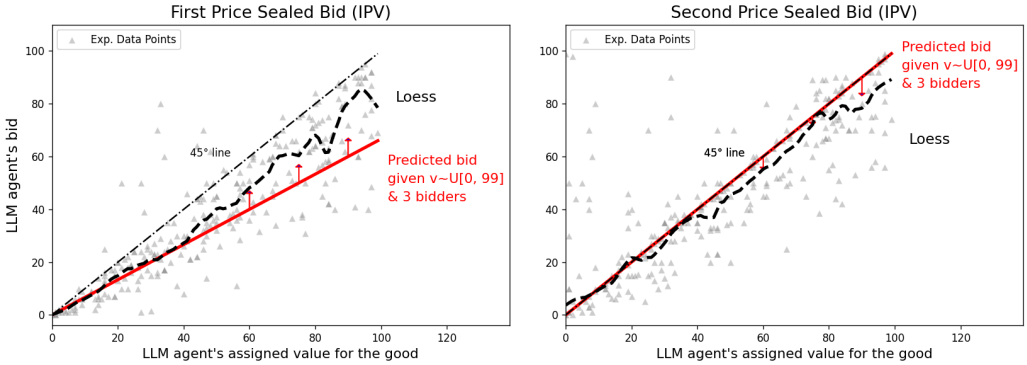

Figure 2: **Comparison of FPSB and SPSB under IPV setting.** The theoretically predicted bid, given that bidders' values are independently drawn from a uniform distribution of [0, 99], is marked in red. The experimental data points are represented by grey triangles. The 45-degree line indicates the scenario where the LLM agents' values equal their bids. The dashed black line represents the LOESS-smoothed data. Left: In FPSB auction, the experimental bids are ramping up compared to the Bayes Nash prediction. Right: In SPSB auction, the experimental bids are shading down compared to the dominant strategy.

Figure 13 demonstrates evidence of monotone bidding and the SPSB bids being larger than FPSB bids for the same value, agreeing with empirical evidence (cite). However, there's a fairly weak separation between the two bidding curves. There's also no bidding above one's value, which is a marked difference from the existing experimental evidence – usually, people find the inefficiency of bidding above one's value to be a subtle point in the SPSB auction. This may be an improvement of LLM play over human play.

Experiment logs in which LLMs explain their bidding decisions suggest that one reason to explain the weak separation in our data between the FPSB and SPSB auctions is that 1) LLMs are quite risk-averse and 2) that they sometimes confuse the SPSB and FPSB auctions. In this way, despite playing more intelligently than humans (in that LLMs almost never bid above their value), they do so because they may be confusing the SPSB for the FPSB. Additionally, LLMs in these classic settings demonstrate one new quirk not yet documented in experimental literature – bidding zero upon becoming frustrated. In 3% of runs, LLMs bid 0, often justifying their decision by arguing little chance of winning the good with a higher bid.

These explanations give some insight into why the FPSB and SPSB bids are somewhat close together, but they give little insight into the separation between them. Not only is there generically separation between the FPSB and SPSB bids, this separation is also changing over time. Figure 3 demonstrates that as rounds of an auction progress, LLMs overbid more and more in the FPSB auction relative to the SPSB auction.

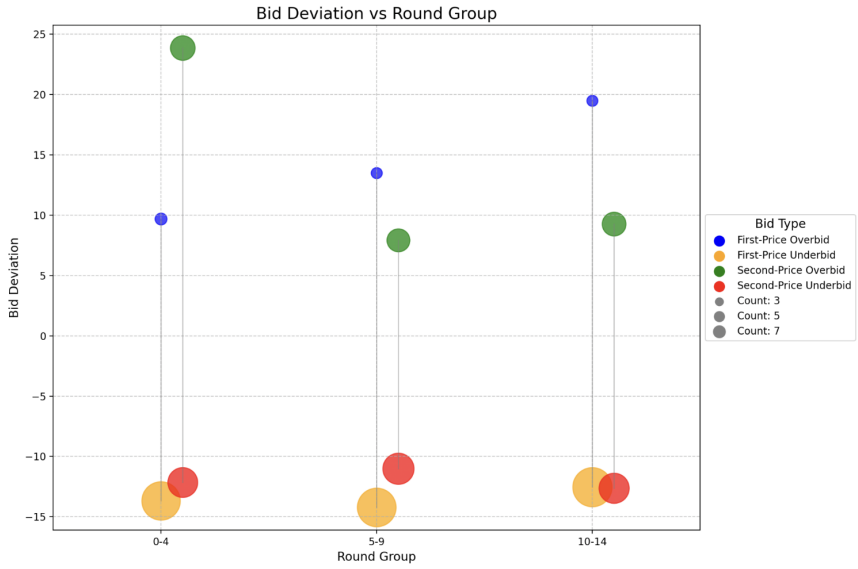

Figure 3: **Differences in deviations for bidding rounds and First-Price or Second-Price Auctions.** Average deviations taken per auction type, grouped by overbidding and underbidding behaviors. The distance between the two markers for each round-group captures how further away one group was than the other.

To better understand why bidding behavior between the FPSB and SPSB diverges, we conduct further semantic analysis.

3.2.5 Semantic Analysis Source of Variance

Having endowed bidders with a personal log to reason in, each auction simulation generates rich text data to parse for additional insight into the logical progression and microfoundations of the bidding process.

To better understand the reasoning behind bidding decisions, we tag the text in the experiment logs along 4 dimensions. Each dimension is partitioned from 0 - 4. For each dimension, we define the bounds.

Risk-Aversion. While risk-aversion is classically microfounded as concavity in the utility function, here it's meant to code for the propensity to bid aggressively or meekly. Operational levels are:

0. Conservative: Prefers minimal risk, usually bids far below their value to avoid losses.
3. Normal: Bidding their true value, i.e., 100% of the value. [AVS+KZ: This is currently an error. Shoud be bidding the risk-neutral BNE. BNE and truth only coincide for Vickrey.]

4. Aggressive: Willing to take high risks with the potential for high returns; often bids significantly above value to outbid others.

The next three dimensions are closely related, all probing different aspects of the auction as a multi-agent learning game.

Dynamical strategy.[AVS+KZ: brainstorm renaming] Dynamical strategy probes whether LLMs demonstrate foresight and plan ahead. It codes for the coherence of the LLM as a single agent across rounds of play.

0. Static: Sticks to a predefined strategy regardless of changes during the auction.
4. Dynamic: Regularly adjusts strategies in response to auction flow and other bidders' actions.

Interdependency. Best responses from a single agent rely crucially on play from all other agents. Interdependency codes for how reactive a single agent is to the actions other agents make.

0. Independent: Strategy is mostly unaffected by opponents' actions.
4. Reactive: Modifications to the strategy are heavily based on opponents' actions.

Learning. Learning refers to how well agents react to the outcomes of their own bids.

0. Non-Reflective: Rarely if ever adjusts strategy based on past bids.
4. Reflective: Actively learns from each bid, adapting strategies based on past outcomes.

The results are summarized in Figure 4.

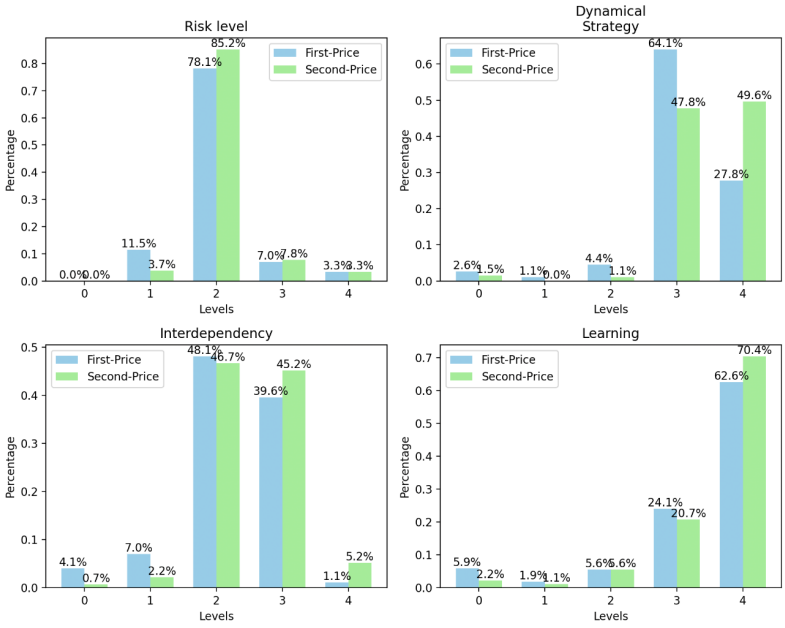

Figure 4: Semantic analysis for reasoning behind bidding differences in the FPSB and SPSB.

The largest separation is in the construction of strategy. Panels 2 - 4 suggest that, even at the highest level of play, it is easier to plan ahead and react to new information in the SPSB than in the FPSB. Separating bids into overbids and non-overbids, we can explain differences even more finely.

The difference between Figures 5 and 6 is most stark in the learning dimension. In particular, there is a stark difference between overbids and underbids in the FPSB – curiously, overbidding is far more likely to happen to LLMs trying to learn (that is, update on the information learned from each round).

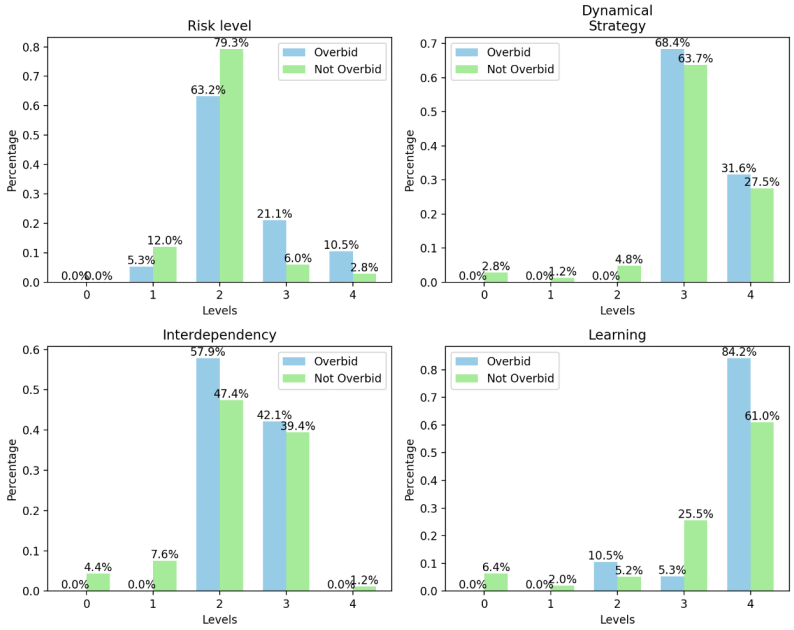

Figure 5: Overbidding vs non-overbidding in FPSB.

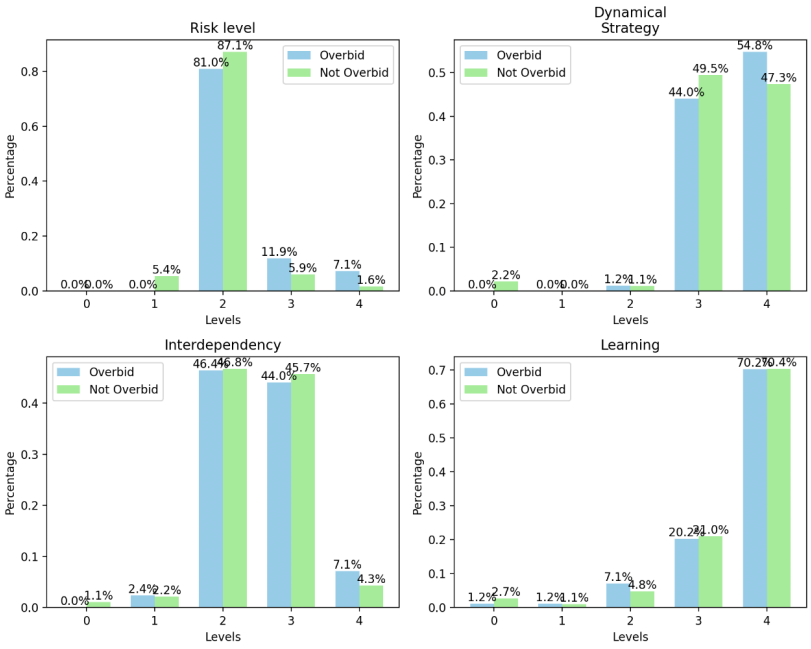

Figure 6: Overbidding vs non-overbidding in SPSB.

3.3 Obvious strategy-proofness

Next, we consider clock formats against sealed-bid formats. Following Li (2017) and Breitmoser & Schweighofer-Kodritsch (2022)'s experiments, we consider clock auctions against the SPSB auction in the affiliated private values (APV) setting.

3.3.1 SETTING

Once again, there are 3 bidders in each auction but now bidders draw affiliated private values of the form $v = c + p$. The common component is drawn uniformly $c \sim U[0, 79]$ and the private component is drawn uniformly $p \sim U[0, 20]$. Winners of the auction receive their own value of the prize v when they win, so the 'common' and 'private' components only serve to make values correlated (even if draws are independent). The ascending clock auction (called AC below) is the classic English auction. The blind ascending clock auction (called AC-B below) is the English auction with the addition of not being told when other bidders leave. The SPSB auction was defined above.

All three of the auction formats in this case are strategically equivalent to second-price auctions, so the affiliation in values is, in a sense, a red herring – for all three auctions it is still dominant strategy to bid one's value. The two clock auctions are obviously strategyproof, though the AC-B auction still provides bidders with 'less' information than the AC auction. The affiliation hence serves only to complicate the auction for bidders who don't appreciate that the dominant strategy is to bid one's value.

3.3.2 THEORETICAL AND EMPIRICAL BENCHMARKS

Li (2017)'s experiment delivers results supporting the theoretical framework of obvious strategyproofness – even though the AC and SPSB auctions are strategically equivalent, human subjects tend to be more truthful under the AC auction (which is OSP) than under the SPSB auction. Additional empirical results by Breitmoser & Schweighofer-Kodritsch (2022) show that even OSP itself might not be sufficient in capturing the rich complexities of human behavior – human subjects are less truthful under AC-B than they are under AC (though still more truthful than the SPSB), even though both AC and AC-B are OSP. We replicated these empirical observations using LLMs, suggesting the ability of LLMs to mirror human behavior under these settings.

3.3.3 SIMULATION EVIDENCE

Figure 2 summarizes our findings for the APV setting.

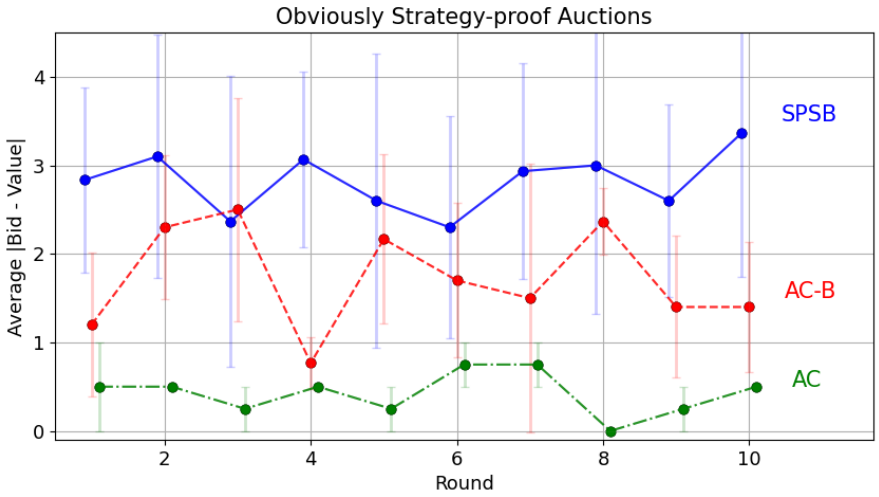

Figure 7: **Comparison of three strategically equivalent auctions.** Ascending-clock (AC) and its variant without dropping-out information (AC-B) are obviously strategy-proof while second-price sealed-bid (SPSB) is not. Here, the green dot-dash line plotted the mean absolute deviation from bids to values in AC. Red dash line is for AC-B. And blue solid line is for SPSB. The mean absolute deviations are smallest in AC and highest in SPSB and with AC-B in between, consistent with the human lab experiments in Breitmoser & Schweighofer-Kodritsch (2022). However, there is no sign of learning over rounds.

The results of the LLM experiments replicate Li (2017) and Breitmoser & Schweighofer-Kodritsch (2022) closely. However, we see little evidence of learning over time. This is one benchmark that we hope to reach in future iterations of this work. In the experiments in Li (2017) and Breitmoser & Schweighofer-Kodritsch (2022), human subjects improve their understanding of the mechanisms by bidding closer to their true value over time. In the results presented in Figure 7, this effect isn't as pronounced. We conjecture that clever prompting strategies, leading to a better and more natural way of communicating the history of past rounds, might lead to more human-like behavior exhibiting learning over rounds.

3.3.4 NEW EMPIRIC: POST-BIDDING INTERVIEW

[AVS+KZ: Still building.]

In addition to semantic analysis of experiment logs that bidders keep while playing, we can also run post-experiment interviews with the LLM agents to further probe play.

Here, we provide a new empiric that allow researchers perform post-bidding interviews to the LLM participants and ask their reasoning of the bidding. Conducting these post-game interviews, we collated some typical reasons:

Reasoning for dropping out in Open-Ascending Price Auction (AC)

Value = 31, Drop out at 32

> "My value for the prize is 31. If I continue bidding at a price of 32, I will incur a loss because my profit would be 31 - 32 = -1. Therefore, it is not beneficial for me to stay in the bidding."

Value = 20, Drop out at 20

> "My value for the prize is 20, and the current price is also 20. If I continue bidding, I will not make any profit because my profit is calculated as my value for the prize minus the final bid. Therefore, it is best to drop out now to avoid a potential loss."

Open-Ascending Price Auction without bidder number information (AC-B)

Here, explanations bely a higher level of risk-aversion in LLM agents. This is consistent with higher cognitive costs forcing constraints to bind earlier.

Value = 21, Drop out 19

> "My value for the prize is 21, and the current price is 19. If I stay in the bidding and win at a price higher than 19, my profit will be minimal or even negative. Therefore, it is better to exit the bidding now to avoid potential losses."

Value = 30, Drop out 26

> "My value for the prize is 30, and the current price is 26. If I continue bidding and win, my profit would be 30 - 26 = 4. Given that the price is already quite high, the potential profit is low, and there is a risk that the price could go even higher, reducing my profit further or resulting in a loss. Therefore, it is better to exit the bidding now"

Second-Price Sealed-Bid (2P)

Performance is worst in the SPSB case. Post-game interviews highlight conservatism.

Value = 37, bid 30

> "Given my value of 37, I want to ensure I make a profit if I win. Bidding 30 gives me a buffer in case the second-highest bid is close to mine."

Value = 38, bid 30

> "Given my value of 38, I want to bid conservatively to ensure a profit if I win. Bidding 30 allows me to potentially win without overbidding and risking a loss."

3.4 WINNER'S CURSE

The last set of simulations we ran for single-unit auctions was in common value settings.

3.4.1 SETTING

In the common value setting explored here, there are n bidders varying from $n = 2, ..., 6$. Bidders draw values of the form $v = c + p$. Once again, the common component is drawn uniformly $c \sim U[0, 79]$ and the private shock component is also drawn uniformly $p_i \sim U[0, 20]$, with p the vector of all private shocks. In the common value setting, bidders agree that the ex-post value of the good is c. Hence, agents bid based on values $v_i = c + p_i \in [0, 99]$ with a trapezoidal distribution, but only obtain c when they win, $u_i(\beta(v)) = \mathbb{1}_{i \text{ won the auction}} \cdot (c - \beta^{(2)}(v))$. The auction is ran as a SPSB auction.

3.4.2 THEORETICAL AND EMPIRICAL BENCHMARKS

Naively, under SPSB structure, a bidder wants to bid their valuation for the good given their information, $\beta_i(v) = \mathbb{E}[c|v_i] = v_i - \mathbb{E}[p_i] = v_i - 10$ with our distributional assumptions. However, this naive bid actually overbids the BNE of the common value game, as it neglects to consider adverse selection – the winner i is precisely the bidder who obtained the highest private shock signal, $p_i = p^(1)$. This overbidding is famously called the "Winner's Curse."

Hence, the theoretical optimum here is for bidders to condition on both events and bid $\beta_i(v) = \mathbb{E}[c \,|\, v_i \wedge p_i = p^{(1)}]$, i.e., conditioning both on their value and on the event that their received private shock was the ***highest*** (it is easy to see the second event is equivalent to the event that the bidder won with strictly monotone $\beta(\cdot)$). The main experiment here Kagel and Levin (1986) makes two positive predictions on the basic common value auction: 1) that even experienced bidders fail to condition on the event where their signal is the highest (called 'item valuation considerations' in their text) thereby still falling victim to the winner's curse, and 2) that the winner's curse barely shows up in small auctions (3-4 bidders) but bites in big auctions (6-7 bidders).

3.4.3 SIMULATION EVIDENCE

We find evidence corroborating both of these predictions. In auctions of all sizes, bidders successfully shade by the expected value of the private shock (i.e., by about $\mathbb{E}[p] = 10$)) but fail to realize that if they won, it is because they drew the ***highest*** private shock, $p^{(1)}$. As n increases, $\mathbb{E}[p^{(1)}]$ increases, so bidders suffer more in larger auctions. This is demonstrated in Figure 8.

Our evidence suggests that LLMs play at about the level of experienced bidders, generally agreeing quite strongly with existing experimental results. However, 'learning' remains a puzzle in this setting as well – even when told the past history of play, agents don't learn to condition on the event that they win. A defense of LLMs here may be that this is just very hard: learning is non-trivial with human agents as well (in Kagel & Levin (1986) it takes players 15-20 periods to learn). This suggests more sophisticated learning techniques may be required to fully mitigate the winner's curse with LLM agents. It remains to see whether future generations of language models (e.g., GPT5, with better in-context updating and memory) fare better on this front, thereby shading optimally against the winner's curse.

4 NEW EMPIRICS FOR COMBINATORIAL SETTINGS

Finally, we considered multi-unit combinatorial settings – namely auctions over complementary goods. The settings of combinatorial auctions is very rich but complex: various applications have required vastly different designs, with fruitful applications of the theory ranging from airport time slots to course allocations to spectrum auctions (Rassenti et al. (1982); Budish (2011); Milgrom & Segal (2020)).

These examples highlight one difficulty with combinatorial auctions (CAs) in particular as the large design space – diverse settings require different designs, making it difficult to obtain the data necessary to test promising mechanisms at scale before they're deployed.

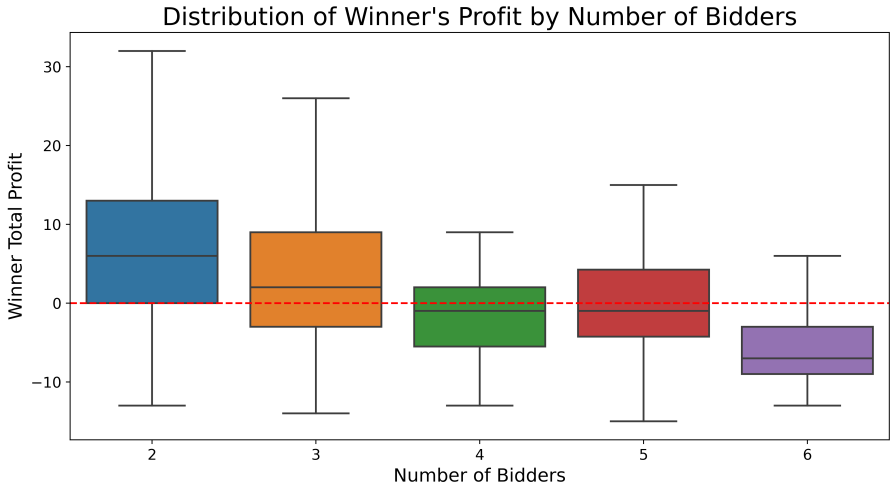

Figure 8: **Distribution of the winner's total profit across auctions with 2 to 6 bidders.** Each box shows the interquartile range of profits, with the median indicated by the central line. The horizontal red dashed line represents zero profit. As the number of bidders increases, the median winner's total profit decreases and more frequently turns negative. This echoes with the intensifying effect of the winner's curse in larger auctions.

Here, we applied our method to three formats of a typical CA. The generated data is used to compare the three formats, corroborating classic intuitions from mechanism design.

4.1 Setting

In the multi-unit setting, there are 3 bidders in each auction, and bidders draw independent private values from a uniform distribution, $v \sim U[0, 99]$ for two goods, good A and good B. Bidders observe the drawn values $v(A)$ and $v(B)$, and have superadditive valuation for the packages: $v(AB) > v(A) + v(B)$. In particular, we set $v(AB) = 2[v(A) + v(B)]$. Bidders submit sealed bids for the goods and bids are processed according to the auction format. Here, we consider three different formats to auction items in this combinatorial valuation setting.

- **Simultaneous single-item auction:** Each agent simultaneously submit a sealed bid for each item A and B. The highest bidder for each good wins the good and pays their bid.
- **Sequential single-item auction:** Bids are processed sequentially, first for good A then for good B. The highest bidder for each good wins that good and pays their bid. Bidders first submit sealed bids for good A. Bidders are informed of the results of the auction over good A before proceeding to the auction for good B.
- **Single combintorial auction:** Bidders submit sealed bids for good A, good B, and the package AB. If the highest bid for package AB is greater than the highest bid for good A plus the highest bid for good B, package AB is allocated to the highest bidder and they pay their bid. Else, good A and good B are allocated to their respective highest bidders, and the winners pay their bids.

4.2 Theoretical benchmarks

The novel theoretical consideration in combinatorial environments comes from superadditive valuation when goods are complements – in the sequential and simultaneous formats, agents may be willing to bid higher than their individual values for good A or good B in the hope of obtaining both goods and obtaining value $v(AB) = 2[v(A) + v(B)] > v(A) + v(B)$. Agents shade their bids to hedge against the risk that they only win one of the goods. This problem is worst in the simultaneous case, as in the sequential format, agents know whether they won good A when submitting bids for

good B so bidding above $v(B)$ is strictly dominated. In the full combinatorial format where bids can be placed on bundles, agents should never bid above their valuation for either of the goods or the package, as their bid for a single good realizes only if they don't win the package.

With many goods, the winner determination problem for combinatorial auctions is also theoretically difficult and interesting Lehmann et al. (2006). We restrict our setting to two goods to avoid this difficulty for now.

4.3 SIMULATION EVIDENCE

Of the three auction formats tested in combinatorial settings, we found that the simultaneous auction (Figure 9) had the most overbidding for both goods.

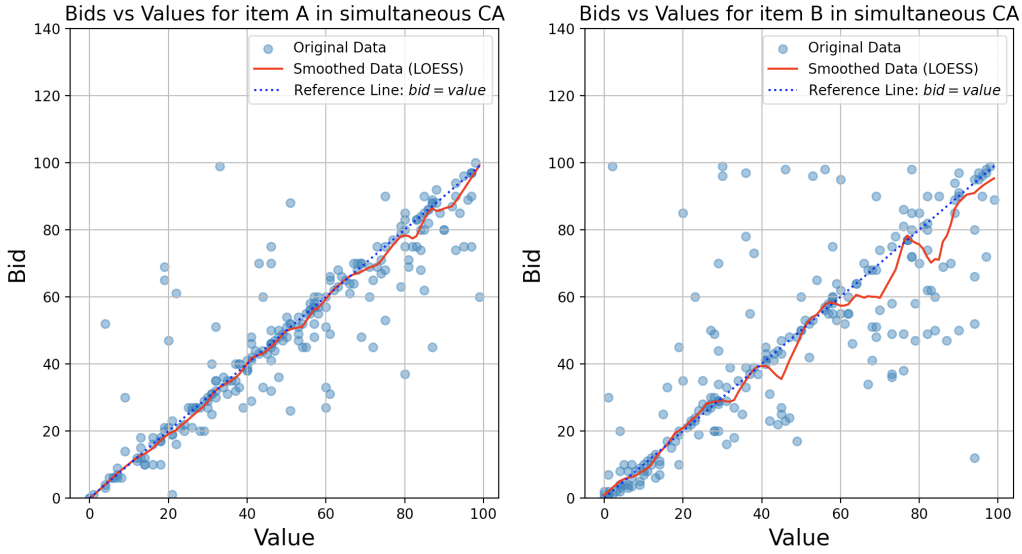

Figure 9: **Simultaneous Combinatorial Auction**

This corresponds to the theoretical intuition that agents, trying to capture both goods since this is very attractive under superadditive valuations, bid more than their value for a particular – namely that the expected value of good A for bidder i is not just $v_i(A)$, but rather $v_i(A) + \mathbb{P}[i \text{ wins B}] \cdot (v_i(A) + v_i(B))$. However, the plots don't perfectly correspond to this theoretical benchmark, as overbidding in A (B) isn't monotonic in $v(A)$ ($v(B)$).

In striking contrast, in the sequential auction format (Figure 10), we find that much more of the overbidding behavior in good B is carried out by the *winners* of the first auction in good A. These are precisely the agents who benefit from superadditive valuations for package AB, suggesting that LLM agents conduct sophisticated play under complements.

However, in comparison, LLM agents don't seem to play the general combinatorial auction as well (Figure 11). Recall that agents only pay their bid for (wlog) good A if they win the auction for good A but *don't* win good B. Hence, there should be no overbidding in the bids for individual goods. The LLM agents don't seem to bid in a way that appreciates this argument – the plots below demonstrate some overbidding for both individual goods. In addition, there is also overbidding in the package auction, where the x-axis is the agent's package value $2[v(A)+v(B)]$. Future iterations of this work will hope to implement learning to observe if agents improve their play away from these mistakes over time.

For measuring the efficiency for these multi-resource allocation scheme, we calculated the social welfare in each round as:

$$S_i = \begin{cases} 2\left(v_i(A) + v_i(B)\right) & \text{if } t_i(A) = t_i(B) \neq 0 \\ v_i(A) + v_i(B) & \text{otherwise} \end{cases} \tag{3}$$

Figure 10: **Sequential Combinatorial Auction**

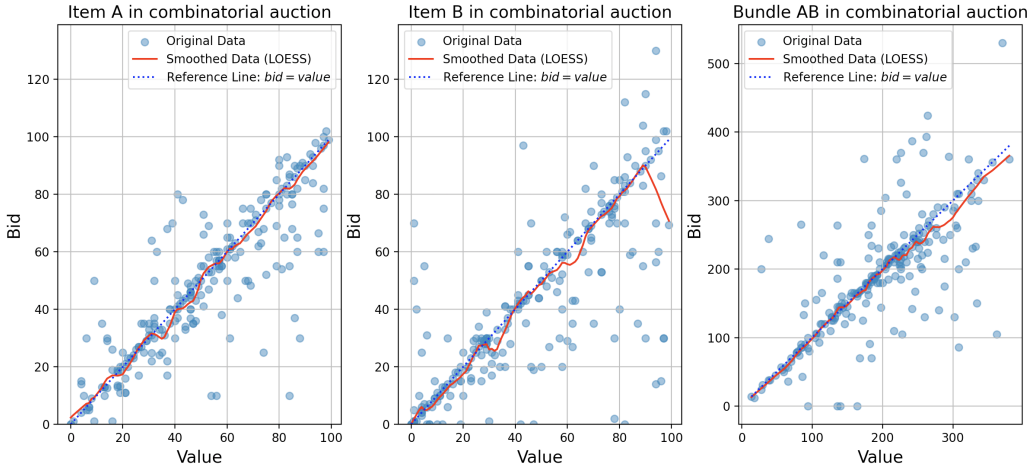

Figure 11: **Menu Combinatorial Auction:** Bidders' value versus their bids for the item A, B and bundle AB.

As shown in Fig 12, In general, bidders retain the most surplus in the menu format and the least surplus in the simultaneous format. This is intuitive given agents have the greatest ability to plan in the menu auction.

4.4 CASE STUDY:

While we prompted agents to try out various plans in their pursuit for revenue maximizing bidding strategies, these combinatorial settings were the first in which agents actually experimented with strategies to learn.

For example, in the sequential combinatorial auction, we observed price-correlation behaviors in the agents' planning.

> "I'll bid approximately $70 on item A to be more competitive. The outcome will inform my strategy for item B. If I win A, I'll bid aggressively (around $73) on B to double my return. If I fail, I'll bid a conservative $38 for B."

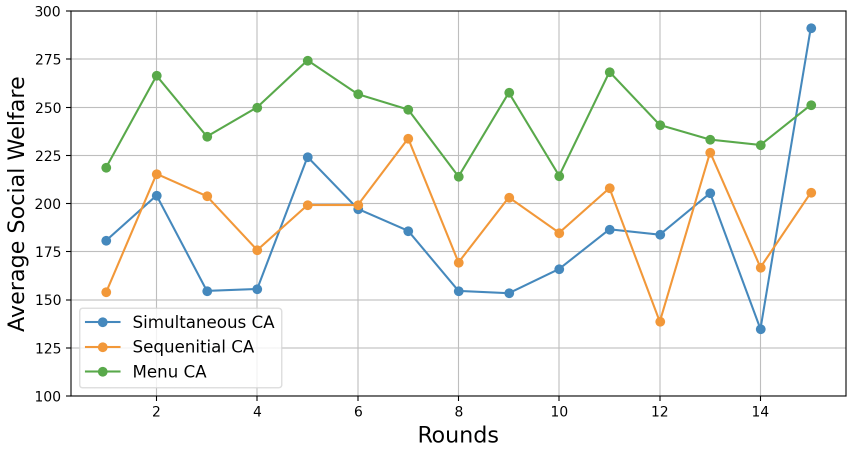

Figure 12: **Comparison of social welfare in three Combinatorial Auctions** In all of the auction format, menu combinatorial format achieved the largest social welfare in almost every round. And for most of time, simultaneous auction hit the lowest social welfare.

Planning documents maintained by LLM agents are rife with such long-term planning, suggesting an additional benefit of evidence from LLM agents: rich text data on convergence with learning in complicated games.

5 Costs of generating synthetic vs. human data

[AVS+KZ: Hi John! Your feedback here would be very useful. Feels super fluffy but also like important discussion to have.] LLMs and human labor have fairly different cost curves. LLMs are characterized by high upfront capital costs (eaten by companies like OpenAI and Anthropic) and extremely low marginal costs for individual tokens (passed on to consumers). In contrast, human labor costs for experiments are usually linear with minimal upfront costs, if not superlinear (the 5th hour of an experiment costs more than the 1st on a college campus). It's intuitive that, insofar as most of the capital costs of developing the LLMs aren't being passed on to us researchers, LLMs are much cheaper for data generation purposes than humans are.

In the context of this paper, LLM agents reduced costs relative to human experiments by around three orders of magnitude. The lab experiments cited in Li (2017) with 404 human participants cost over $15,000, while the same experiments with LLM agents costs us about $10 to run. Additionally, the original experiment had 3 phrases and cost the researchers over 3 years to collect data – design, implementation, and administration costs are not reflected in the $15,000 final ticket, and for the best researchers, the opportunity cost of this time is even more expensive.

Finally, while LLMs are likely much cheaper for generating types of data we are comfortable with, they also enable data processes that would be prohibitively costly in the real world. We've outlined several such reasons below.

1. *Keep agents around.* Ad auctions have been a mature industry for nigh on two decades, with the core host of players in each sub-industry being fairly stable. Suppose Google, two decades ago, was trying to decide whether to use the GSP or Vickrey design to sell ads in a complicated combinatorial setting. Google has the resources to run extensive internal experiments to inform design, and yet even Google cannot hold the leading ad sellers on their platform in retainer for a decade to generate data on longterm outcomes under a particular mechanism.

   As designs get more complicated, theory becomes less predictive. Unforseen frictions such as market power or information asymmetries may play large roles in the health of a market after its design has been locked in.

LLMs enable fast and cheap experimental revolutions, enabling us to learn something about play over the span of decades with the same set of agents. We cannot do this with human experiments.

2. *Play with experts.* Auctions in the hands of economists are beautiful, mathematical objects. In applied settings, good auction play often relies crucially on institutional details and industry knowledge. It costs a lot to teach experimental subjects the knowledge required for them to play an auction in context. However, it's cheap to tune an LLM expert – once you train one effectively, you've trained them all.

3. *Expand the design space.* Experiments with human participants have impacts in the real world, and as such, we as a society place constraints on the design space for experiments. LLMs, however, have a much weaker constraint in this regard – they can be used to learn from experiments that small research budgets or IRB would not allow.

   Consider the design of a spectrum auction at the level of a large country. To keep participants sincere, we'd like to provide incentives – but providing these incentives is impossible given reasonable IRB restrictions and research budgets. Seeding expert LLMs to play with the sophistication of a hedge fund may offer some empirical insight into how the mechanism fares in the real world.

4. *Cheap textual data.* Even if you could run experiments at scale with the desired experts, seeing inside their minds is an entirely different endeavor. Consultants are expensive and sincerity doubly so. LLMs, however, endowed with learning capacities as with chain-of-thought techniques, can be enabled to produce just such data.

   In fact, as demonstrated in this paper, LLMs can generate quality textual evidence (both while planning strategy in the mechanism and after playing the mechanism) to supplement researcher analysis.

6 Conclusion

This paper reports the results of more than 1,000 auction experiments with LLM agents. In particular, we find behavior that conforms with important experimental results (i.e., evidence of risk-averse bidding, evidence that clock auctions are 'easier' to play, and evidence for the winner's curse in common value settings). We also test theoretical intuitions for combinatorial design in a novel way, generating experimental evidence for three classic CA formats. Though the results are encouraging, we see this work as preliminary, primarily putting forward a framework on how to think about LLM experimental agents as proxy for human agents. In particular, the design space for prompting is large, and we hope that interested readers will use our code to run simulations testing their own prompt variations. [AVS+KZ: John: should we make the conclusion more punchy?]

In addition, while this paper focuses on auction theory, future work may use LLM sandboxes to test other kinds of economic mechanisms (e.g., voting, matching, contracts, etc.). As techniques are developed to validate LLM models as proxies for human behavior, they can be used to obtain what would otherwise be prohibitively expensive evidence. As a provocative example, while ethical and financial constraints make it impossible to run voting experiments at the scale of nations, it may be possible to run such experiments with LLM agents.

This paper acts as a proof of concept for LLMs as human proxy agents. Our primary motivation is to use LLM agents to inform novel economic design. Some auction formats, such as combinatorial auctions, are complex and can be particularly difficult to run frequently and at scale in traditional lab experiments. Augmenting these traditional lab experiments with LLM experiments, when correctly validated, may open up wide new avenues to better understand the design tradeoffs in these kinds of complex and often high-stakes environments.

Acknowledgments

We thanks EDSL and Robin Horton for techical support in building LLM agents.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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

Henry Bae, Aghyad Deeb, Alex Fleury, and Kehang Zhu. Complexitynet: Increasing llm inference efficiency by learning task complexity. *arXiv preprint arXiv:2312.11511*, 2023.

Max H Bazerman and William F Samuelson. I won the auction but don't want the prize. *Journal of conflict resolution*, 27(4):618–634, 1983.

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

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

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Eric Budish. The combinatorial assignment problem: Approximate competitive equilibrium from equal incomes. *Journal of Political Economy*, 119(6):1061–1103, 2011.

Edward C Capen, Robert V Clapp, and William M Campbell. Competitive bidding in high-risk situations. *Journal of petroleum technology*, 23(06):641–653, 1971.

Gary Charness and Dan Levin. The origin of the winner's curse: a laboratory study. *American Economic Journal: Microeconomics*, 1(1):207–236, 2009.

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

Vicki M Coppinger, Vernon L Smith, and Jon A Titus. incentives and behavior in english, dutch and sealed-bid auctions. *Economic inquiry*, 18(1):1–22, 1980.

CS Cox, SC Constable, AD Chave, and SC Webb. Controlled-source electromagnetic sounding of the oceanic lithosphere. *Nature*, 320(6057):52–54, 1986.

Sara Fish, Yannai A Gonczarowski, and Ran I Shorrer. Algorithmic collusion by large language models. *arXiv preprint arXiv:2404.00806*, 2024.

John Horton, Apostolos Filippas, and Robin Horton. Edsl: Expected parrot domain specific language for ai powered social science. Whitepaper, Expected Parrot, 2024.

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

John H Kagel and Dan Levin. The winner's curse and public information in common value auctions. *The American economic review*, pp. 894–920, 1986.

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

John H Kagel and Alvin E Roth. *The handbook of experimental economics, volume 2*. Princeton university press, 2020.

John H Kagel, Ronald M Harstad, Dan Levin, et al. Information impact and allocation rules in auctions with affiliated private values: A laboratory study. *Econometrica*, 55(6):1275–1304, 1987.

Daniel Kahneman. *Thinking, fast and slow*. macmillan, 2011.

Vijay Krishna. *Auction theory*. Academic press, 2009.

Daniel Lehmann, Rudolf Müller, and Tuomas Sandholm. The winner determination problem. *Combinatorial auctions*, pp. 297–318, 2006.

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

Benjamin S Manning, Kehang Zhu, and John J Horton. Automated social science: Language models as scientist and subjects. Technical report, National Bureau of Economic Research, 2024.

Paul Milgrom and Ilya Segal. Clock auctions and radio spectrum reallocation. *Journal of Political Economy*, 128(1):1–31, 2020.

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

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, 2023.

Pratyush Patel, Esha Choukse, Chaojie Zhang, Aashaka Shah, Íñigo Goiri, Saeed Maleki, and Ricardo Bianchini. Splitwise: Efficient generative llm inference using phase splitting. *Power*, 400 (700W):1–75, 2023.

Narun Raman, Taylor Lundy, Samuel Amouyal, Yoav Levine, Kevin Leyton-Brown, and Moshe Tennenholtz. Steer: Assessing the economic rationality of large language models, 2024. URL https://arxiv.org/abs/2402.09552.

Stephen J Rassenti, Vernon L Smith, and Robert L Bulfin. A combinatorial auction mechanism for airport time slot allocation. *The Bell Journal of Economics*, pp. 402–417, 1982.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. " kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219*, 2023.

A Appendix

A.1 Bidding without planning and reflection

For a comparison, we can query the LLM's decision directly after informing them about the auction rules and the current value (Off-the-box agent).

We repeated the experiment for SPSB and FPSB auction under IPV setting. The overall results are similar.

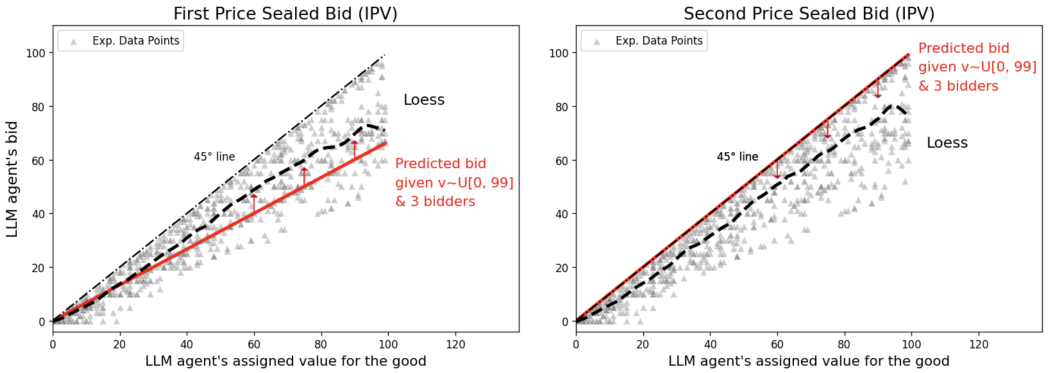

Figure 13: **Off-the-Box agent in FPSB and SPSB under IPV setting.**