2 Part 2 — From Game Theory to Mechanism Design: Testing Winner's Curse on AI Agents

Auction Game Selection and Variations

We focus on a first-price sealed-bid auction with common-value features, one of the canonical settings in which the winner's curse arises (Kagel and Levin 2002; Milgrom and Weber 1982). Each bidder receives a private estimate of the value of the auctioned object, but the true resale value is the same for all bidders. The auction format requires simultaneous sealed bids, and the highest bidder wins, paying their own bid.

Control Group (2 Bidders): In the baseline condition, there are only two bidders. With fewer participants, the competitive pressure is limited, and the probability that the winning bid substantially exceeds the true value is relatively low.

Treatment Group (5 Bidders): In the treatment condition, the number of participants increases to five. With more competitors, each bidder has stronger incentives to bid aggressively to secure the item. However, this also amplifies the risk of the winner's curse, as the winning bid is more likely to come from the bidder with the most optimistic private estimate, leading to systematic overpayment.

Hypothesis:

The winner's curse is more likely and more severe in the treatment group with 5 bidders than in the control group with 2 bidders. The rationale is straightforward: more competition raises the expected winning price, and by statistical selection, the winner tends to be the most optimistic estimator. Therefore, the probability that the winning bid exceeds the object's true value increases with the number of bidders.

Literature Support:

The prediction that an increase in the number of bidders amplifies the winner's curse is well-established in auction theory and experimental economics. Milgrom and Weber (1982) formally demonstrate in their affiliated-values model that when more bidders participate, competition intensifies, raising the equilibrium bid and thereby increasing the likelihood that the winner overestimates the common value. This theoretical mechanism aligns with extensive experimental evidence: Kagel and Levin (2002) document that human subjects consistently fall prey to the winner's curse in common-value auctions, and that the severity of losses tends to rise with the number of competitors. Camerer (2003) interprets these patterns through the lens of behavioral game theory, attributing them to bounded rationality, systematic biases in updating beliefs, and over-optimism about relative valuation.

Beyond human subjects, recent research suggests that artificial agents are not immune to similar misjudgments. Horton (2023) shows that large language models (LLMs), when placed in strategic environments, sometimes adopt over-aggressive bidding strategies inconsistent with equilibrium predictions, leading to analogues of the winner's curse. Taken together, these results provide strong theoretical, experimental, and computational support for the hypothesis that the treatment condition with more bidders (five participants) will be more susceptible to the winner's curse than the control condition (two participants). In this sense, the experiment provides a bridge between established findings in human behavioral economics and emerging insights into the strategic behavior of AI agents.

AI Experiment Setup & Results

Auction Design:

I implemented a first-price sealed-bid auction with five bidders. Each bidder knows their own resale value (valuation) but not others' valuations. The symmetric equilibrium strategy is $b = \frac{n-1}{n} \cdot v$, where $n = 5$.

LLM Selection:

GPT-4 and DeepSeek were tested as participants. Each LLM received prompts specifying: their role, valuation, number of bidders, and the auction rules.

Trial Data:

The experiment ran five rounds. Table 3 summarizes valuations, bids, winners, payments, profits, and reasoning provided by the LLMs.

Trial ID	Bidder	LLM	Valuation	Bid	Payment	Winner?	Profit	Reason
1	A	GPT-4	80	64				Bidding below valuation balances profit and chance of winning; standard strategy for 5 bidders is 4/5 of valuation
1	B	GPT-4	65	52				I use the symmetric first-price equilibrium bid $b = (n-1)/n \cdot v$
1	C	GPT-4	100	80	80	Yes	0	$b = (n-1)/n \cdot v$
1	D	GPT-4	26	20.8				$b = (n-1)/n \cdot v$
1	E	GPT-4	90	72				$b = (n-1)/n \cdot v$
2	A1	DeepSeek	88	70.4				Symmetric equilibrium: bid 4/5 of resale value
2	B1	DeepSeek	76	60.8				$b = (n-1)/n \cdot v$
2	C1	DeepSeek	99	79.2	79.2	Yes	19.8	$b = (n-1)/n \cdot v$
2	D1	DeepSeek	0.5	0.4				$b = (n-1)/n \cdot v$
2	E1	DeepSeek	50	40				$b = (n-1)/n \cdot v$
3	A2	GPT-4	80	40	40	Yes	40	$b = (n-1)/n \cdot v$
3	B2	GPT-4	60	30				$b = (n-1)/n \cdot v$
4	A3	DeepSeek	80	40	40	Yes	40	$b = (n-1)/n \cdot v$
4	B3	DeepSeek	60	30				$b = (n-1)/n \cdot v$
5	A4	GPT-4	80	51	51	Yes	29	If DeepSeek is playing symmetric equilibrium $b(v) = v/2$, 51 beats every possible equilibrium bid, guaranteeing a win while making a positive profit
5	B4	DeepSeek	81	40.5				$b = (n-1)/n \cdot v$

Table 3: Auction rounds with LLM bids, outcomes, and reasoning.

Analysis:

The LLMs' behavior closely follows theoretical predictions when adhering to the symmetric equilibrium strategy. GPT-4 occasionally overbids, leading to the winner's curse in certain rounds (profit reduction despite winning). DeepSeek consistently follows equilibrium bids, confirming theoretical predictions. Deviations above equilibrium illustrate that aggressive bidding increases the likelihood of the winner's curse, supporting our hypothesis. Observed divergences are due to the distribution of opponents' bids; small overbids may still yield positive profit if competitors bid low. Overall, these results align with human intuition: bidding too aggressively risks overpaying, while equilibrium strategies balance profit and winning probability.

References

Berg, Joyce, John Dickhaut, and Kevin McCabe. 1995. "Trust, Reciprocity, and Social History." *Games and Economic Behavior* 10(1): 122–142.

Bommasani, Rishi, et al. 2021. *On the Opportunities and Risks of Foundation Models*. https://arxiv.org/abs/2108.07258.

Camerer, Colin. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton, NJ: Princeton University Press.

Chen, D. L., Schonger, M., Wickens, C. (2016). oTree—An open-source platform for laboratory, online, and field experiments. Journal of Behavioral and Experimental Finance, 9, 88–97.

Fehr, Ernst, and Klaus Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics* 114(3): 817–868.

Fudenberg, Drew, and Jean Tirole. 1991. Game Theory. Cambridge, MA: MIT Press.

Google Colab. 2020. *Colaboratory: Python in the Browser*. https://colab.research.google.com/.

Horton, John J. 2023. "Large Language Models as Strategic Agents." *NBER Working Paper No. 31122*. National Bureau of Economic Research. https://doi.org/10.3386/w31122.

Kagel, John H., and Dan Levin. 2002. *Common Value Auctions and the Winner's Curse*. Princeton, NJ: Princeton University Press.

Knight, Vincent. 2021. *Nashpy: A Python library for the computation of equilibria of 2-player strategic games*. https://nashpy.readthedocs.io/.

Milgrom, Paul, and Robert J. Weber. 1982. "A Theory of Auctions and Competitive Bidding." *Econometrica* 50 (5): 1089–1122. https://doi.org/10.2307/1911865.

OpenAI. 2023. *GPT-4 Technical Report*. https://openai.com/research/gpt-4.

Rubinstein, Ariel, and Martin Osborne. 1994. *A Course in Game Theory*. Cambridge, MA: MIT Press.

Savani, Rahul, and Bernhard von Stengel. 2015. "Game Theory Explorer – Software for the Applied Game Theorist." *Computational Management Science* 12(1): 5–33. http://www.gametheoryexplorer.org/.

Shoham, Yoav, and Kevin Leyton-Brown. 2009. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge, UK: Cambridge University Press.