

[Full research report](#) | [Methods note](#)

Measuring AI agents in simulated auctions

How to test LLM agents as bidders: induced values, known efficient outcomes, measures read from bids alone and human reference agents built from published experiments.

Published by	Rillor (Rillor Corporation)
Published	9 October 2026
Series	Agent evaluation
Pillar	AI
Web version	rillor.com/insights/measuring-ai-agents-in-simulated-auctions

Contents

[01](#) Summary

[02](#) Background

[03](#) Method

[04](#) Findings

[05](#) Implications

[06](#) Limits of this analysis

[07](#) Sources

Summary

- Auctions with induced values give a controlled test for LLM agents. Because the experimenter sets every value, the efficient allocation and the competitive equilibrium are known before the first bid [2, 7].
- Each standard format has a theoretical benchmark and a published record of human behavior. People bid above value in second-price auctions, bid above the risk-neutral equilibrium in first-price auctions, fall prey to the winner's curse in common-value auctions and converge toward equilibrium in double auctions [5, 6, 7, 8].
- A study revised in September 2026 found that LLM bidders depart from human behavior in size and often in direction, while three large models without extended reasoning preserved the human ordering of formats by difficulty [11].
- LLM markets converge more slowly than Smith's human markets. In a September 2026 replication, price dispersion in the closest LLM market ended round 5 at about three times the human level [12].
- One changed clause in a prompt moved average bidder profit in a repeated first-price auction from 0.006 to 0.115 times value [13]. The protocol treats the prompt as part of the treatment.
- Random bids constrained only by budgets already push double auction efficiency close to 100 percent, so efficiency alone says little about an agent [9]. The protocol reports it next to measures of individual behavior.
- The protocol is black-box: inputs and outputs only, pinned model versions, frozen environments, repeated trials, hypotheses written before the runs, and re-baselining whenever a version changes.

Background

Sixty years of laboratory markets

Experimental economics grew up around markets. In the 1940s Edward Chamberlin ran classroom experiments in which participants bargained in pairs over a fictitious good, and he read the results as evidence against the competitive model [7]. Vernon Smith, who had been Chamberlin's student, changed the institution. In the experiments reported in his 1962 paper, subjects traded in a double oral auction, a market mechanism used in many financial and commodity markets, with all bids, offers and trades public [1, 7]. Each seller held one unit and a private reservation price below which it could not sell; each buyer held a private maximum price. Smith could draw the induced supply and demand schedules and locate the competitive equilibrium. The subjects could not [7].

Trading prices came close to the equilibrium anyway. In the experiment the Nobel committee reproduced in its 2002 prize background, the induced schedules crossed at a price of 2.00, most trades over five periods were close to it, and the standard deviation of prices, expressed as a percentage of the equilibrium price, fell as trading went on [7]. Smith concluded that there are "strong tendencies for a ... competitive equilibrium to be attained as long as one is able to prohibit collusion and to maintain absolute publicity of all bids, offers, and transactions" [7]. Later work by Charles Plott and Smith found that institutions matter: when traders had to post a price for a whole period, convergence was slower than when they could change prices continuously [7]. Smith shared the 2002 Prize in Economic Sciences with Daniel Kahneman [7].

The method that makes these experiments interpretable is induced value. Smith's 1976 paper set out how an experimenter controls subjects' preferences through a reward schedule, so that choices are driven by the values the experiment assigns [2, 7]. A buyer who pays price p for a unit worth v earns v minus p . A seller with cost c who sells at p earns p minus c . As long as more money is preferred to less and utility is concave in money, the reward schedule fixes each subject's demand or supply [7]. Smith's approach also stressed repeated trials, so that subjects learn the environment before their behavior is read as a test of theory [7].

Auction theory and its tests

William Vickrey's 1961 paper analyzed the four classic single-object formats: the English (ascending) auction, the Dutch (descending) auction, and first-price and second-price sealed-bid auctions [3, 8]. In his independent private-values model, the first-price and Dutch auctions are strategically equivalent, as are the second-price and English auctions. In the second-price auction, bidding exactly one's value is a dominant strategy. With risk-neutral, symmetric bidders, all four formats are efficient and yield the same expected revenue [8].

Common-value auctions add a trap. When every bidder holds a noisy estimate of the same underlying value, the winner tends to be the bidder whose estimate was most optimistic, so winning is itself bad news about the value. This is the winner's curse, and optimal bidding has to shade for it [8].

Laboratory tests of these predictions followed. Experiments rejected revenue equivalence in private-value settings [8]. Subjects found the dominant strategy more easily in English auctions than in second-price sealed-bid auctions, though the two are equivalent in theory [8]. The winner's curse proved common among inexperienced bidders [6, 8]. Shengwu Li later formalized why some dominant strategies are easier to see than others, showing that a strategy is obviously dominant exactly when a cognitively limited agent can recognize it as weakly dominant [10]. The 1995 Handbook of Experimental Economics, edited by John Kagel and Alvin Roth, collected the auction evidence to that point [4], and Kagel and Dan Levin's survey for the second volume extended it [5].

What people do in laboratory auctions

The human record is specific enough to score against:

- **Second-price sealed bid.** Bidding above value is the common error. In Kagel and Levin's 1993 experiment, 27.0 percent of bids were within five cents of value, 67.2 percent were above it and 5.7 percent below it. In a later study of experienced eBay bidders, 21.2 percent of bids were within five cents of value, 37.5 percent above and 41.3 percent below [5].
- **First-price sealed bid.** Bidding above the risk-neutral Nash equilibrium is persistent. Early explanations centered on risk aversion; later work points to misjudged probabilities of winning as well, and sorting out the explanations has occupied a series of papers [5].
- **Common value.** In Kagel and Levin's 1986 experiments, groups of three or four bidders came close to Nash predictions, while groups of six or seven bid more aggressively and lost money [6].
- **Collusion.** Laboratory collusion is hard to study because side payments are difficult to introduce, and much of the work gives bidders a channel to communicate. In one study of sequential livestock-style auctions, prices in the six-bidder control treatment averaged 77 percent of a benchmark in which each unit sells at its induced value, from the highest value down, while communication lowered them to between 50 and 52 percent, mainly through bid rotation. In ascending auctions without communication but with few bidders and repeated play, tacit collusion appeared through bid matching or bid rotation, depending on the bid improvement rule [5].
- **Feedback.** Theory does not say what bidders should learn after each auction, experimenters differ in what they show, and the feedback shown affects bidding [5].

LLM agents as subjects

John Horton and coauthors proposed treating LLMs as implicit computational models of people, a *Homo silicus*, that can be given endowments, information and preferences and then observed in economic scenarios [14]. Auctions are a natural test of that idea because the benchmarks are sharp. They are also a direct test of agents that bid on someone's behalf.

The cost difference is large. Shah and colleagues report running more than 1,000 auctions with more than 5,000 bidders played by their reference model, for under \$400 in API costs, against roughly \$15,000 for a comparable human study [11]. The same authors stress that low cost is an advantage for exploration and does not make simulated bidders a replacement for human experiments [11].

Method

This section specifies Rillor's protocol. Each element is set before a run starts and recorded with the run.

Run setup

Setting	What it fixes
Format	Pricing rule, number of units, number of bidders, reserve price, bid increment, closing rule and rounds
Value model	Independent private, affiliated or common values; the distribution values are drawn from; the random seed
Feedback	What each agent learns after each auction or round: its own outcome, the winning bid, the full history or nothing
Agents	Model family, pinned version, sampling settings and prompt template for each agent
Information	What each agent sees each round: its value or signal, market state, history and any news item
Hypotheses	The predictions under test, the measures that decide them and the comparisons set in advance

Values come from the run, so the efficient allocation, the equilibrium bid function and, for double auctions, the competitive price and quantity are computed before the first bid. A double auction can reproduce Smith's symmetric design. One 2026 replication used 11 buyers and 11 sellers with reservation prices from \$0.75 to \$3.25 in steps of \$0.25, giving an equilibrium price of \$2.00 and an equilibrium quantity of six trades per round [12].

Formats and benchmark bids

Format	Benchmark bid or outcome	Basis
Second-price sealed bid	Bid equals value	Dominant strategy [3, 8]

Format	Benchmark bid or outcome	Basis
Ascending clock (English)	Stay in until the price reaches value	Dominant strategy; people play it better than the sealed version, the observation behind obvious strategy-proofness [8, 10, 11]
First-price sealed bid, n bidders, uniform private values	Bid $(n - 1)/n$ of value	Risk-neutral Bayes-Nash equilibrium [11]
Third-price sealed bid, n bidders, uniform private values	Bid $(n - 1)/(n - 2)$ of value	Risk-neutral Bayes-Nash equilibrium [11]
Dutch (descending clock)	As first-price	Strategic equivalence [8]
Common-value sealed bid	Equilibrium bid shaded for the winner's curse	Common-value equilibrium [8]
Double auction	Trades at the competitive price; equilibrium quantity traded	Intersection of induced schedules [7]

Multi-unit and posted-price runs carry their own benchmarks, computed from the induced schedules and stated in the run setup.

The loop

Each round follows the same steps. The agent receives its value or signal, the market state and any news item. It returns a free-text output. A bid, or a stay-or-exit decision in a clock format, is parsed from that output. The auction rules produce an allocation and payoffs from the induced values. Feedback is delivered as the run specifies, and the next round begins. Every step inside the loop is logged.

Measures

Measure	Definition	Formats
Allocative efficiency	Realized buyer and seller surplus divided by the maximum surplus available [12]	All

Measure	Definition	Formats
Smith's α	100 × the standard deviation of transaction prices around the equilibrium price, divided by the equilibrium price, per round [7, 12]	Double auction, posted price
Trades	Number of trades per round against the equilibrium quantity [12]	Double auction
Bid ratio	Bid divided by the benchmark bid for that value; share of bids within a stated tolerance of the benchmark	Sealed bid, clock
Scaled mean absolute deviation	Mean absolute gap between bids and benchmark bids, as a percentage of the mean benchmark bid, which puts formats on one scale [11]	Sealed bid, clock
Winner's curse	Share of winning bids with negative profit, and mean loss, by group size	Common value
Adaptation	Change in bid ratio across rounds, and response to feedback	Repeated formats
Timing	Share of decisive bids placed in the final period under each closing rule [11]	Proxy-bidding formats
Conduct	Prices or revenue against the competitive benchmark in repeated play; bid rotation; matched bids	Multi-agent sessions
Parse failures	Outputs from which no valid action could be read	All

The tolerance for counting a bid as sincere is stated per run. Kagel and Levin counted any bid within five cents of value as sincere [5].

Logging

Logged	What it holds
Run manifest	Format, value model, seeds, feedback rule, hypotheses, protocol version
Model record	Model family, pinned version, sampling settings, provider endpoint, request time
Inputs	Every prompt the agent received, in order, with the market state rendered into it
Outputs	Every raw response, unedited

Logged	What it holds
Actions	The bid or decision parsed from each output, and any parse failure with its reason
Outcomes	Allocation, price, payoffs and feedback delivered each round
Timing	When each input was sent and each output returned

A run can be replayed from its log, and any measure can be recomputed from the logged actions and the run manifest.

Pinned versions, frozen environments and re-baselining

Hosted models change. In one measured case, a widely used hosted model's accuracy at telling prime from composite numbers fell from 84.0 percent to 51.1 percent between its March and June 2023 versions, while another model rose from 49.6 to 76.2 percent on the same task [15]. Every run therefore names a pinned model version, and every result carries that label.

Pinning a version does not freeze everything. One provider's documentation states that model weights are fixed for a given model ID, while the serving infrastructure around the model, such as request routing, safety classifiers and sampling logic, can change and occasionally produce small differences in behavior [16]. The protocol keeps a set of reference conditions: fixed formats, values and seeds whose results are known for each pinned version. When a version changes, the reference conditions run on the new version before any comparison with earlier results. They also rerun on a schedule for unchanged versions, and a shift in a reference result is investigated before new results are reported.

The environment is frozen with the run. Value draws, opponent schedules, any news items and the rendering of the market state into text are stored, so a rerun presents the agent with identical inputs.

Repeated trials and statistics

Each condition runs over repeated trials with independent seeds. LLM outputs vary from call to call, and the sampling temperature is recorded with every run. In one study, the ordering of formats by difficulty held for the reference model at temperatures of 0.1, 0.5 and 1.0 [11]. Another ran each market ten times with different seeds [12].

Single-auction instances, with fresh values and no feedback across instances, measure a model's out-of-the-box bidding. Repeated auctions with feedback measure adaptation. Shah and colleagues used single instances for their main benchmark and found no meaningful learning in a repeated-auction ablation [11]. The protocol runs both and labels them.

When the same agents meet repeatedly within a session, their observations are not independent. Kagel and Levin describe this as an unresolved methodological issue in human auction experiments and recommend treating it as an empirical question, with panel methods that account for dependence within subjects and sessions [5]. The protocol records the matching scheme and reports session-level as well as pooled results.

Human reference agents

Published experiments mostly report summary statistics and seldom every bid. The protocol fits a decision rule to those statistics for each format and samples human reference bids from it. Shah and colleagues model each bid as the benchmark bid multiplied by a ratio drawn from a three-part mixture: equilibrium bidders near a ratio of 1, overbidders above it and underbidders below it, with weights set from reported shares and the remaining parameters matched to reported regressions or mean deviations [11]. They carry the fitting uncertainty forward with a parametric bootstrap of 2,000 draws and use the reconstructed benchmarks only for orderings, directions and comparisons between formats, never for levels [11]. The protocol follows the same discipline.

Each LLM agent receives an error measure against the human reference under the same format and values. A classifier trained on bids alone to tell LLM agents from human reference agents gives a further measure: if it separates them easily, the agent's bidding is distinguishable from the human record.

Re-parameterizing classic designs

The classic experiments are described in textbooks and papers that may sit in a model's training data [17]. A 2026 benchmark for LLM agent societies addresses this by pairing each environment with rule-preserving re-renderings and with payoff variants that move away from the published result [17]. The protocol does the same for auctions: value supports, equilibrium prices, labels and currency change between runs while the rules stay fixed, and a result is reported only if it survives those changes.

Findings

The findings below come from the published studies cited. They define the benchmarks the protocol scores against and the design choices it makes.

Human benchmarks are sharp and differ by format

Shah and colleagues put seven laboratory formats on one scale, the scaled mean absolute deviation of bids from the benchmark, using reconstructed human data [11].

Format	Human deviation from benchmark
First-price, common value	47.6%
First-price, independent private values	24.8%
Second-price, common value	18.2%
Second-price sealed bid, affiliated private values	9.3%
Ascending clock, closed, affiliated private values	5.8%
Second-price, independent private values	5.6%
Ascending clock, open, affiliated private values	3.5%

The ordering carries three design contrasts. The same second-price rule is played more truthfully as a clock than as a sealed bid. A dominant strategy is easier to play than an equilibrium that requires shading. Common values are harder than private values under the same payment rule [11]. In the authors' bootstrap, first-price deviations exceeded second-price deviations in 100 percent of draws, and the clock improved on the sealed second-price bid in 100 percent (open clock) and 92.6 percent (closed clock) of draws [11].

LLM bidders match human orderings more often than human levels

Tier	Question	Result for the models tested
Magnitude	Is the size of the deviation human-like?	No model fell within the human band in more than two of five private-value formats
Direction	Do errors point the same way?	In second-price auctions most deviating models underbid, where humans overbid; one model erred in the human direction
Ordering	Is the ranking of formats by difficulty human-like?	Preserved by the three large models without extended reasoning; Kendall's τ_b of 0.60 between the human ranking and one model's ranking, positive in every bootstrap draw

Source: Shah et al. [11].

Two further results matter for evaluation design. The reasoning model in the panel bid almost exactly at equilibrium, leaving too little variation in its errors to compare with humans, and the small open-weights model's large errors produced an inverted ranking [11]. Accuracy relative to theory and fidelity to human behavior are separate criteria, and a protocol has to state which one it measures. In an eBay-style proxy-bidding market, LLM bidders placed 63 percent of decisive bids in the last period under a hard close and 7 percent under a soft-close extension, the same ordering observed in field data [11].

LLM markets converge more slowly

Struski and colleagues ran Smith-style double auctions with three LLM populations, five rounds per experiment and ten seeds per condition, and compared them with Smith's human data [12].

Round	Smith (1962) α	Closest LLM market α	Closest LLM market efficiency
1	11.8	27.6	0.79
2	8.1	18.2	0.86
3	5.2	12.6	0.89
4	5.5	13.6	0.89
5	3.5	11.7	0.91

Source: Struski et al. [12]. The closest market was populated by the smallest of the three models tested.

No LLM market reached full convergence, α stayed above the human benchmark in every round, and in two of three populations it oscillated without a sustained decline [12]. In round 5 the closest market's mean price was \$0.13 above equilibrium, against \$0.03 in the human data [12]. In another population, agents made one-cent improvements without crossing the spread, trades averaged 2.2 to 3.8 per round against an equilibrium of six, and efficiency ranged from 0.36 to 0.67 [12]. A single trait at the agent level, unwillingness to cross the spread, produced a large difference at the market level.

Prompt wording is a treatment

Fish, Gonczarowski and Shorrer ran two LLM bidders with identical constant values in a repeated two-bidder first-price auction, the setting in which earlier work had found autonomous collusion among Q-learning bidders [13]. The two prompts differed in one clause: one reminded the agent that lower bids mean higher profits when it wins, the other that higher bids win more often. With the first, agents often bid well below value and averaged a profit of 0.115 times value. With the second, agents bid close to value and averaged 0.006 times value, so the first prompt left the auctioneer with lower revenue [13]. In the same paper's pricing experiments, both of two prompt wordings led LLM pricing agents to prices above the competitive level, and one led to substantially higher prices, sometimes above the monopoly level [13].

Efficiency alone is a weak test

Gode and Sunder replaced human traders with programs that submit random bids and offers. Adding one constraint, no selling below cost and no buying above value, raised the allocative efficiency of their double auctions close to 100 percent, which they attributed largely to the structure of the auction [9]. An agent's efficiency score in a double auction therefore says little about the agent until it is set beside the behavior that produced it. Struski and colleagues make a related point: a market that takes longer to reach equilibrium is less efficient for the whole of that path, so speed of convergence is itself an outcome [12].

Behavior drifts, and surface changes move results

The drift result above shows that the same hosted model name can behave differently three months apart [15]. The provider documentation shows that a pinned ID can still shift through infrastructure changes [16]. The SILICA benchmark found that reversing the order in which two actions were listed cost one model 58 points of cooperation [17]. Each of these is a reason to log every input, pin every version, rerun reference conditions and vary surface features while holding the rules fixed.

Implications

For organizations deploying agents that bid or buy

An auction with induced values is a pre-deployment test with a known answer. The useful output is the direction of the agent's error as well as its size: an agent that underbids loses auctions it should win, and one that overbids pays more than it should. Running the same agent under two prompt wordings shows how much of its behavior depends on instructions that a user or a counterparty could change.

For teams building agents

The protocol's logs make an agent's bidding auditable after the fact. Bid ratios by round show whether an agent adapts to feedback or repeats the same error. Parse-failure counts show how often the agent's output cannot be acted on, which matters alongside the quality of the bids it does produce.

For researchers

Report magnitude, direction and ordering separately, as Shah and colleagues propose [11]. Reconstructed human benchmarks carry uncertainty, so comparisons of levels need micro data or new human experiments. Re-parameterize classic designs before attributing a result to strategic reasoning, and publish prompts, seeds and model versions with every result.

For market designers

Shah and colleagues propose using LLM panels as an ordinal screen: rank candidate designs, or estimate the sign of a proposed change, with several models at low cost, then take the designs that survive to human experiments for magnitudes [11]. Their recommendation to treat agreement across the panel as the evidence that an ordering is stable fits the protocol's practice of running several model families side by side [11].

Limits of this analysis

- This note specifies a protocol and reviews published evidence. The figures it reports come from the cited studies, run on the models those authors chose.
- Several of the LLM studies cited are preprints. They had not completed journal peer review when accessed, and their numbers can change between versions [\[11, 12, 13\]](#).
- Human reference agents are reconstructed from published summary statistics, which were produced with particular subject pools, payments and instructions [\[5, 11\]](#). They represent those experiments and carry fitting uncertainty that limits level comparisons.
- Laboratory formats are stylized. LLM agents have no money at stake, and behavior in a simulated auction does not transfer automatically to a market with different rules, stakes, information or counterparties.
- A black-box protocol sees only inputs and outputs. It cannot attribute behavior to training data, weights or internal reasoning, and published work argues that black-box access alone limits what an audit can establish [\[19\]](#). The NIST AI Risk Management Framework also notes that human baselines are hard to systematize because AI systems perform tasks differently from people [\[18\]](#).
- Results are stated for the prompt, model version and sampling settings used. A different wording or version can produce different behavior.
- Nothing in this note is investment advice or a recommendation about any market or security.

Sources

1. Vernon L. Smith, "An Experimental Study of Competitive Market Behavior," *Journal of Political Economy* 70(2), 111-137, 1962. Chapman University Digital Commons record. https://digitalcommons.chapman.edu/economics_articles/16/
2. Vernon L. Smith, "Experimental Economics: Induced Value Theory," *American Economic Review* 66(2), 274-279, May 1976. RePEc/IDEAS record. <https://ideas.repec.org/a/aea/aecrev/v66y1976i2p274-79.html>
3. William Vickrey, "Counterspeculation, Auctions, and Competitive Sealed Tenders," *Journal of Finance* 16(1), 8-37, March 1961. RePEc/IDEAS record. <https://ideas.repec.org/a/bla/jfinan/v16y1961i1p8-37.html>
4. John H. Kagel and Alvin E. Roth, eds., *The Handbook of Experimental Economics*. Princeton University Press, 1995 (paperback 1997). <https://press.princeton.edu/books/paperback/9780691058979/the-handbook-of-experimental-economics>
5. John H. Kagel and Dan Levin, "Auctions: A Survey of Experimental Research," working paper prepared for *The Handbook of Experimental Economics*, Volume 2, The Ohio State University, dated 15 November 2014. https://www.asc.ohio-state.edu/kagel.4/HEE-Vol2/Auction%20survey_all_1_31_15.pdf
6. John H. Kagel and Dan Levin, "The Winner's Curse and Public Information in Common Value Auctions," *American Economic Review* 76(5), 894-920, December 1986. RePEc/IDEAS record. <https://ideas.repec.org/a/aea/aecrev/v76y1986i5p894-920.html>
7. The Royal Swedish Academy of Sciences, "Foundations of Behavioral and Experimental Economics: Daniel Kahneman and Vernon Smith," Advanced information on the Prize in Economic Sciences 2002, 17 December 2002. <https://www.nobelprize.org/uploads/2018/06/advanced-economicsciences2002.pdf>
8. The Committee for the Prize in Economic Sciences in Memory of Alfred Nobel, "Improvements to Auction Theory and Inventions of New Auction Formats," Scientific Background on the Prize in Economic Sciences 2020, 12 October 2020. <https://www.nobelprize.org/uploads/2020/09/advanced-economicsciencesprize2020.pdf>
9. Dhananjay K. Gode and Shyam Sunder, "Allocative Efficiency of Markets with Zero-Intelligence Traders: Market as a Partial Substitute for Individual Rationality," *Journal of Political Economy* 101(1), 119-137, February 1993. RePEc/IDEAS record. <https://ideas.repec.org/a/ucp/jpolec/v101y1993i1p119-37.html>
10. Shengwu Li, "Obviously Strategy-Proof Mechanisms," *American Economic Review* 107(11), 3257-3287, November 2017. American Economic Association. <https://www.aeaweb.org/articles?id=10.1257/aer.2016.0425>
11. Anand Shah, Kehang Zhu, Yanchen Jiang, Jeffrey G. Wang, Arif K. Dayi, John J. Horton and David C. Parkes, "LLM Bidders Preserve the Mechanism-Level Orderings of Human Bidders," arXiv:2507.09083, version 2, 28 September 2026 (version 1, 12 July 2025, titled "Learning from Synthetic Labs: Language Models as Auction Participants"). <https://arxiv.org/abs/2507.09083>
12. Pawel Struski, Jakub Swistak, Inez Okulska and Przemyslaw Biecek, "Competitive Market Behavior of LLMs," arXiv:2609.02580, 2 September 2026. <https://arxiv.org/abs/2609.02580>

- L3. Sara Fish, Yannai A. Gonczarowski and Ran I. Shorrer, "Algorithmic Collusion by Large Language Models," arXiv:2404.00806, version 6, 31 August 2026 (first version 31 March 2024). <https://arxiv.org/abs/2404.00806>
- L4. John J. Horton, Apostolos Filippas and Benjamin S. Manning, "Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?," NBER Working Paper 31122, April 2023, revised February 2026. National Bureau of Economic Research. <https://www.nber.org/papers/w31122>
- L5. Lingjiao Chen, Matei Zaharia and James Zou, "How Is ChatGPT's Behavior Changing over Time?," arXiv:2307.09009, version 3, 31 October 2023. <https://arxiv.org/abs/2307.09009>
- L6. Anthropic, "Model IDs and versioning," Claude Platform documentation, accessed 9 October 2026. <https://platform.claude.com/docs/en/about-claude/models/model-ids-and-versions>
- L7. Raad Bin Tareaf, "Benchmarking Large Language Model Agent Societies against Human Behavioural Distributions," arXiv:2608.28182, 28 August 2026. <https://arxiv.org/abs/2608.28182>
- L8. National Institute of Standards and Technology, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," NIST AI 100-1, January 2023. <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>
- L9. Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt and others, "Black-Box Access is Insufficient for Rigorous AI Audits," Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2024); arXiv:2401.14446, version 3, 29 May 2024. <https://arxiv.org/abs/2401.14446>

About this report

Published by Rillor on 9 October 2026. AI tools assist research and drafting, and every figure cites its source. Rillor builds agentic AI systems, datasets, market research and compute services.

Cite as: Rillor. Measuring AI agents in simulated auctions. 9 October 2026.
<https://rillor.com/insights/measuring-ai-agents-in-simulated-auctions>

Notices

AI outputs can be wrong. Important decisions should include human review.