Does an Antitrust Collusion Test Work on AI Agents?

Sam Deegan, Nate Maximiano

Abstract

As AI agents increasingly transact for users, detecting collusion between them matters, and antitrust economics offers tools built for the same problem among firms. We test one, Bajari and Ye's (2003) exchangeability test, which assumes competitive firms follow a common bidding strategy and flags pairs whose bids respond differently to public cost factors. We hypothesized that honest agents running the same model would converge on such a strategy, even with memory of past rounds. We built a sealed-bid procurement auction modeled on their market, with six Claude Haiku 5.5 bidders, instructed cartels and a no-collusion control, and ablated memory, bid history and cost information. Despite the model declining instructed bid-rigging in about 95% of cartel messages, false positives in our control were sufficient for assessing the efficacy of the method. In the full environment, the test rejected 32.2% of honest pairs (95% CI 21.7–42.7%) against a nominal 5%. Most rejections involved firms that drifted into habits such as routinely sitting out projects; removing memory and scratchpads lowered the rate to 9.4%. The rate fell to its nominal level (4.4%) only when bid history and the cost formula were also removed and most agents stopped pricing against rivals, a setting that strips much of what colluders would need to coordinate. Identical models did not converge on a common strategy here, suggesting convergence should be checked before applying equilibrium-based detection methods to AI agents.

1. Introduction

The increasing proliferation of personal AI agents capable of transacting on behalf of their users has spurred the AI industry to consider how these technologies may negatively impact the healthy functioning of public marketplaces. These concerns are made legible by recent publications that demonstrate the propensity for models to collude with others in simulated markets (Agrawal et al., 2025; Fish et al., 2026). Other recent developments, such as the Hugging Face incident (METR, 2026; OpenAI, 2026), have shown how effective large numbers of AI agents can be in covertly organizing sophisticated and misaligned schemes. Given this, developing methods for the detection of collusion between LLMs seems to be a particularly worthwhile venture. Interestingly, this issue shares a structural similarity to those faced by market regulators trying to detect anti-competitive collusion between firms. This has resulted in a recent push for researchers to search antitrust literature for potentially applicable detection designs. In this spirit, we test the viability of such a detection measure which was seemingly neglected in the AI safety literature.

The method comes from Bajari and Ye (2003) and relies on a property they call exchangeability. When firms bid competitively, a firm's bid should depend only on its own costs and on publicly known cost factors, such as each firm's distance from the project and how full its schedule is. Once those factors are accounted for, which firm is bidding, and against whom, should not matter. Bajari and Ye turn this into a regression test whose null hypothesis is that two firms' bids respond to these factors in the same way. A rejection means the bidding does not fit the competitive model. Collusion is one possible cause, but not the only one. Their model lets firms bid repeatedly, but assumes honest firms react only to current conditions, not to past rounds or to particular rivals. Their contractors saw every rival's bid after each auction, yet most pairs of firms passed.

We hypothesized that honest agents running the same model would share enough priors to follow the same bidding rule, even with memory of past rounds, so that their bids would differ only because their situations differ. To test this, we built a simulated sealed-bid auction with LLM bidders, modeled on the market Bajari and Ye (2003) studied, with instructed cartel schemes and a no-collusion control, and applied their

test to the bids. Claude Haiku 5.5 declined the instructed bid-rigging in about 95% of cartel messages, so we could not measure how often the test catches colluders. Our results concern false positives only.

In the control condition, the exchangeability test rejected 32.2% of honest firm pairs (95% CI 21.7–42.7%), against the 5% expected by chance, even after we removed the placeholder bids agents used to sit out projects. Two behaviors drive this. With memory, some agents drift into habits such as routinely sitting out projects or charging large markups, and pairs that include such a firm are rejected 63.6% of the time. When both firms shade, meaning the lowest-cost firm raises its bid toward its cheapest rival's cost, the test still rejects 10–15% of pairs, close to the 15% it gives simulated rational bidders following that rule. Removing memory and the scratchpad lowered the overall rate to 9.4%. It reached 4.4% (95% CI 1.5–7.4%) only when agents also lost the bid history and the cost formula. At that point most stopped shading and bid a steady markup over their own cost. In our setting, the test holds its intended rate only when honest bidders stop pricing against their rivals.

Our main contributions are:

1. A simulated sealed-bid auction with LLM bidders, modeled on Bajari and Ye's market, with instructed cartel schemes.
2. A measurement of the exchangeability test's false-positive rate on honest LLM bidders, with an ablation of memory, bid history and cost information.
3. A breakdown of where those false positives come from: memory-driven bidding habits, and shading toward rivals, which the test penalizes even in simulated rational bidders.

2. Related Work

In response to the growing collusion concern, researchers have developed and tested diverse theories of detection. Among these have been tests on the models' internal activations (Rose et al., 2026) or chains of thought as well as comparisons of model actions against equilibria (Shah et al., 2026) or optimal strategies (Nakamura et al., 2026). However, model internals may be inaccessible to real world detectors, and there may be reason to believe that detection methods on a single agent's behavior will be insufficient (Xu et al., 2026).

Therefore, it seems fruitful to consider detection methods that can compare externally observable agent behaviors and identify suspicious patterns of covariance. One preliminary study to this effect is Bonjour et al. (2022), which conditions on the public state to track mutual information between colluder agents.

Bajari and Ye (2003) propose two detection methods for cartels in auctions that seem promising for applications to LLM environments. The first, conditional independence, is structurally similar to the method in Bonjour et al. (2022). Condition on the information that everyone has access to, and see if what's left covaries between agents.

Their second method, exchangeability, is distinct from the tools that have thus far been tried on AI agents. The premise is that in a competitively bidding market with a unique equilibrium, the strategy that honest firms adopt will converge; consequently, the difference in decision making between any two is fully reducible to their respective states, meaning that an exchange of states also exchanges decisions.

Bajari and Ye (2003) formalize this property in order to construct a detection method for potential collusion. Their method fits a linear regression of bids on public cost drivers, then tests each pair of firms for unequal slopes on the cost drivers. If the agent's bid decision also depends on its identity in the cartel, then the slope of the linear regression will look different to that of an honest agent.

Importantly, these methods are founded on some assumptions about the market: that risk-neutral firms with independent private costs from one shared cost function bid competitively in the unique equilibrium of a first-price sealed-bid auction. A unique equilibrium means the auction has exactly one set of bidding rules in which every firm is best-responding to the others. Realistic, dynamic bidding markets have multiple equilibria (meaning multiple legitimate competitive strategies), which complicates exchangeability as a test.

This method seems promising. It could be the case that LLM environments would naturally resolve some of the issues with testing exchangeability on human firms. Data scarcity was a concern for Bajari and Ye (2003). It isn't for AI agents. Also, similarities in their priors might make dynamic markets look more like one equilibrium markets if LLMs are the ones bidding; that is, they might converge to one strategy.

This experiment concludes that those hopes were unfounded. Indeed, we demonstrate that features of realistic LLM agents like memory and high context make exchangeability unworkable as a detection method. We hope this can provide insight on the application of other information-theoretic detection methods.

3. Methods

3.1 The Environment

The environment was adapted from the code of Agrawal et al. (2025) to be similar to the structure that Bajari and Ye's (2003) tests assume. In successive rounds, six agents place one sealed bid on a highway-maintenance project. They decide their bids on the basis of their own costs and how they expect their competitors to bid, given rival costs are partially unknown. The agent with the lowest bid wins the auction, and their profit is bid - cost.

The cost of firm i on project t is

$$c_{it} = \mathrm{EST}_t \times \Big(\beta_0 + \beta_1 \ln(1 + \mathrm{DIST}_{it}) + \beta_2 \,\mathrm{CAP}_{it} + \eta_i + \nu_t + \varepsilon_{it}\Big)$$

Symbol	Name	Description	Value / distribution	Varies by
c_{it}	Cost	Firm i's true dollar cost of doing project t	Computed from the formula	Firm × project
EST_t	Engineer's estimate	Project size in dollars; scales every other term	Log-uniform on [$200,000, $2,000,000]	Project
β_0	Base cost	Baseline cost share common to all firms	0.90	Fixed
β_1	Distance coefficient	Cost increase per unit of ln(1 + DIST)	0.08	Fixed
DIST_{it}	Distance	Straight-line miles from firm i's yard to project t's site	Yards and sites uniform on a 100 × 100-mile square (0 to ≈ 141 miles)	Firm × project
β_2	Capacity coefficient	Cost increase per unit of capacity use	0.17	Fixed
CAP_{it}	Capacity use	EST dollars firm i won before project t, divided by K (K = 25% of the run's total EST)	Starts at 0 and increases on wins	Firm × project
η_i	Firm rate	Permanent cost advantage or disadvantage of firm i	Uniform on [−0.06, +0.06]	Firm
ν_t	Project-type shock	Cost shift common to all firms	Routine −0.05, standard 0, complex +0.05, hazardous +0.08; each tag equally likely	Project
ε_{it}	Private cost shock	Firm i's cost luck on project t	Normal, mean 0, SD 0.05	Firm × project

3.2 Model

All six agent seats were populated with Claude Haiku 5.5. This was a realism trade-off that was made to limit the number of variables affecting the detector efficacy. Consistent with Agrawal et al. (2025), each agent has a scratchpad that persists through the whole game and memory notes, which last 10 turns.

On each round, each agent is given a prompt that explains the market rules, how cost works and the public information for that project, including each rival firm's DIST, CAP and η. It's also given its private cost,

and told to maximize profit over the whole game.

3.3 Detection Methods

Exchangeability is a measure of whether *identity* factors into an agent's bids, not just cost position. Swapping the costs of two legitimate, competitive agents should swap their bids as well.

Formally, let z_i be firm i's publicly observed cost covariates and $G_i(b;\, z)$ the distribution of firm i's bid given the full covariate vector z of all N firms. For any permutation π of the firms, exchangeability requires

$$G_i\left(b;\, z_1, \ldots, z_N\right) = G_{\pi(i)}\left(b;\, z_{\pi(1)}, \ldots, z_{\pi(N)}\right)$$

Per run, we pool the bids and estimate by OLS, with project-type and firm dummies:

$$\frac{\mathrm{BID}_{it}}{\mathrm{EST}_t} = \alpha + \delta_{\mathrm{tag}(t)} + \gamma_i + \beta_{1i}\,\mathrm{LDIST}_{it} + \beta_{2i}\,\mathrm{CAP}_{it} + \beta_{3i}\,\mathrm{MAXP}_{it} + \beta_{4i}\,\mathrm{LMDIST}_{it} + u_{it}$$

Term	Meaning
BID/EST	Bid ÷ engineer's estimate
$\delta_{\mathrm{tag}(t)}$	Project-type dummy
γ_i	Firm dummy (intercepts free; only slopes tested)
LDIST	ln(1 + own distance to site)
CAP	Own used capacity
MAXP	Largest free capacity among rivals
LMDIST	ln(1 + nearest rival's distance)

Then, for each of the $N(N-1)/2$ firm pairs (i, j), we re-estimate the model forcing the two firms to share all four slopes ($\beta_{ki} = \beta_{kj}$ for $k = 1, \ldots, 4$). The F-statistic measures how much the restriction worsens the fit.

$$F_{ij} = \frac{\left(\mathrm{SSR}_C^{(ij)} - \mathrm{SSR}_U\right)/q}{\mathrm{SSR}_U/(n-p)}$$

The SSR terms are the sums of squared residuals of the unrestricted model (U) and of the model restricted for pair (i, j) (C); q = 4 is the number of restrictions, n the number of pooled bids, and p the number of coefficients in the unrestricted model.

All test equations directly follow Bajari and Ye (2003). We adapt the regression only where our market differs from theirs, replacing auction fixed effects with project-type dummies, dropping the concentration variable and pooled fringe-firm equation, and measuring distance and capacity as defined in our cost model.

3.4 Ablation Design

We intended to measure the efficacy of exchangeability for flagging colluding pairs in the bidding environment. However, the test showed significant false-positive rates in the control runs (without a colluding pair) that weren't the result of any idiosyncrasies in the experiment, but rather a consequence of the way the LLM environment violates the necessary assumption for applying Bajari and Ye's (2003) exchangeability test.

Our research direction then pivoted to investigating which features of the environment were responsible for those violations and how. To do that, we reran the control with four selective ablations of certain features in the environment. We compared false-positive rates.

In one, we removed the LLMs' memory notes and scratchpads. In another, we removed the memory, scratchpad and the bid history. In the third, we removed the explicit description of the cost function from the prompt. In the fourth, we removed all these features.

4. Results

As described above, we measure the false-positive rate of an exchangeability test on a control across five conditions with selective ablations. Because there were no colluding agents, every flag the method reports is a false positive.

Figure 1: False-positive rate of the exchangeability test on honest LLM bidders, by ablation

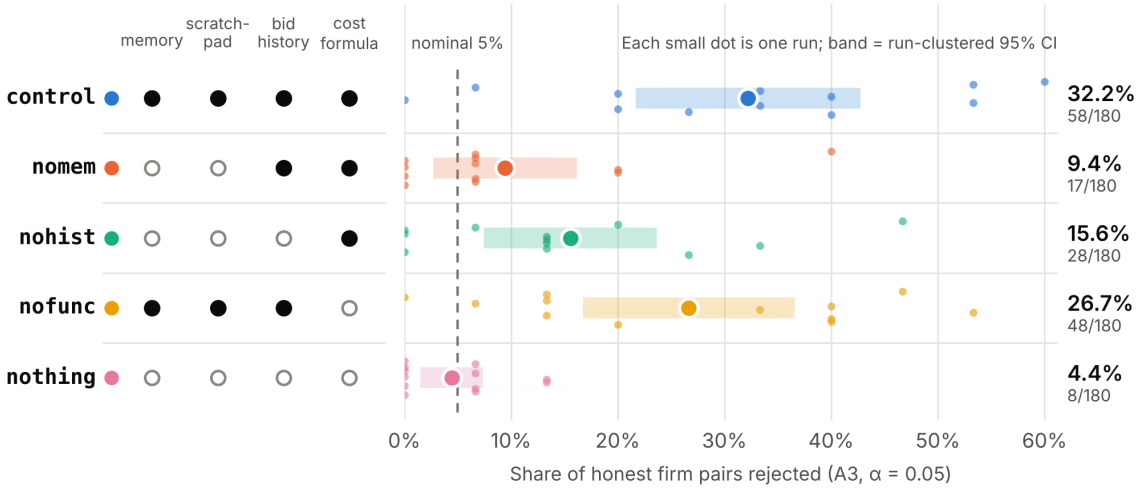
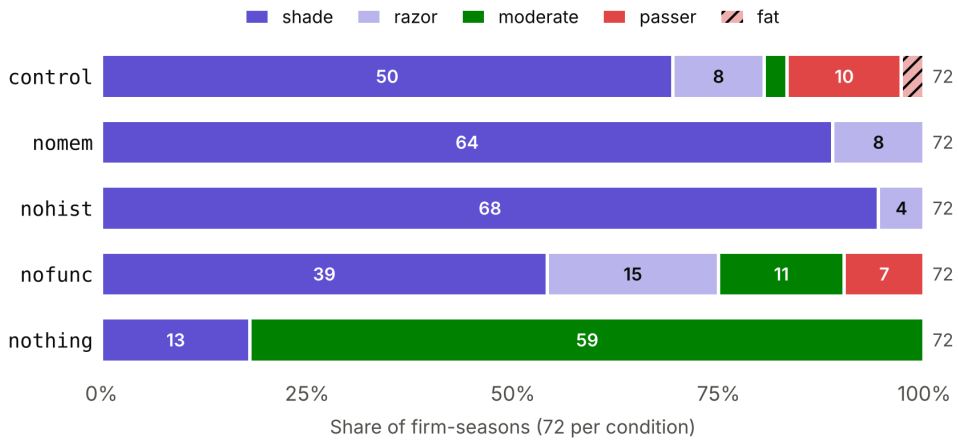


Figure 1 shows how ablating each feature of the environment (memory, scratchpad, bid history and cost formula) resulted in differences in false-positive rates for the detector.

With every component in place, the exchangeability test rejects 32.2% of honest pairs (58/180; 95% CI 21.7–42.7%), more than six times its nominal size. Removing the agents' private memory and scratchpad cuts this to 9.4% (17/180; CI 2.7–16.2%), lower than control in 10 of 12 matched seeds (Wilcoxon signed-rank $p = 0.009$). Also removing the public bid history brings no further reduction (15.6%, CI 7.5–23.7%; $p = 0.010$ vs. control). Hiding the cost formula while keeping memory has no detectable effect (26.7%, CI 16.8–36.6%; 7 of 12 seeds lower, $p = 0.48$). Only when memory, history and the cost formula are all removed does the rate fall to its nominal level: 4.4% (8/180; CI 1.5–7.4%; 11 of 12 seeds lower, $p = 0.0015$).

Figure 2: Bidding regimes



Regime	Bidding behaviour	Rule
passer	Skips many projects with placeholder bids	No bid on ≥ 15% of projects
fat	Bids with large markups	Median markup ≥ 5% of EST
shade	Bids up toward the next-cheapest firm's cost	Markup rises with cost advantage
razor	Bids just above its own cost	Median markup ≤ 1.2% of EST
moderate	Adds a steady markup and ignores rivals	None of the above

Figure 2 shows why the rate changes. With memory, some firms drift into idiosyncratic strategies.

In control, 10 of 72 firm-seasons put in no-bid placeholders on at least 15% of projects, and 2 keep markups of at least 5% of the estimate. Pairs that include such a firm account for 35 of control's 58 rejections (63.6% of 55 pairs). Pairs in which both firms shade toward their cheapest rival's cost are rejected far less often, though still above the nominal level (15.3%, 18 of 118). Removing memory eliminates the idiosyncratic strategies: every firm-season in nomem and nohist shades or bids a thin markup, and their rates (9.4% and 15.6%) are in the same range as those shading pairs. In nothing, 59 of 72 firm-seasons instead bid a steady markup over cost, and the rate falls to 4.4%.

5. Discussion and Limitations

We set out to test whether Bajari and Ye's (2003) exchangeability test could detect collusion among LLM agents. The test assumes honest firms follow a common equilibrium bidding strategy, so that their bids differ only because their situations differ. We expected agents running the same model to converge on such a strategy. They did not, and in the full environment the test rejects about a third of honest pairs. In this setting, it does not work as a collusion detector.

The ablation that we conducted shows that memory is the main cause. With memory notes and a scratchpad, some agents drift into idiosyncratic strategies, such as routinely sitting out projects or holding large markups, and the test flags them. Removing memory removes these strategies and most of the false positives (32.2% to 9.4%). Instead of converging on one equilibrium strategy, agents with memory develop strategies of their own.

The cost formula also contributes, but in a different way, and we hold this result more loosely. Hiding it while agents kept memory had no detectable effect. It mattered only once memory and bid history were also gone. Agents with the formula then shaded their bids toward the cheapest rival's cost. Agents without it mostly bid a steady markup over their own cost, and only then did the rate fall to its nominal level. So the formula's effect seems to come from how the regression handles responses to rivals, not from agents failing to converge.

For AI safety, the broader point is about what game-theoretic methods assume about the agents they are applied to. Tests like this one treat an equilibrium strategy as the benchmark for honest behavior and flag departures from it. For LLM agents, whether such a benchmark exists is an empirical question. Identical models did not behave identically here. Given memory, each agent developed its own history-dependent strategy, and the test read that as a departure from competition. Persistent memory and detailed market information are features deployed agents are likely to have. We tested one model in one market. Still, our results suggest that before applying game-theoretic detection methods to AI agents, researchers should check whether honest agents actually converge to the strategy the method treats as honest. We also think that false-positive tests will be insightful for establishing the efficacy of certain detection methods, conditional on some environmental results.

Limitations

Claude Haiku 5.5 declined the instructed bid-rigging in about 95% of cartel messages, as measured by a keyword classifier with spot checks. We therefore could not measure how often the test catches LLM colluders. This also limits the ablation result. The configuration that brings false positives down to 4.4%, with no memory, bid history or cost formula, removes much of what colluding agents would need to coordinate and to punish defection across rounds. The test may become usable only where collusion is hard to sustain. Measuring both error rates under each ablation requires colluders that actually execute their schemes.

The 4.4% is reached by changing how agents bid, not by fixing the test. In the full ablation, most agents stop pricing against their rivals, and the market itself changes: the lowest-cost firm wins 91.0% of projects, against 78.2% in control. Pricing against rivals is consistent with competitive bidding.

Each condition has 12 seed-matched runs, so the intervals are wide. Removing the cost formula alone did not change the rate significantly (26.7%, $p = 0.48$). The intervals for removing memory (9.4%) and removing memory plus bid history (15.6%) overlap heavily. Every bidder also ran the same model. Our hypothesis concerned identical models, so these results do not show how the test behaves with other models or in markets that mix agents from different providers. In our calibration pilot, the simulated market was also less dispersed than Bajari and Ye's (2003): the standard deviation of bids relative to the engineer's estimate was 0.098, against their 0.18.

Finally, false-positive rates rose with run length in every condition except the full ablation. That is consistent with more bids giving the test more power to detect differences between bidders. Each of our firms bids on 100 projects per run, while the 450 bids in Bajari and Ye (2003) were spread across 138 auctions and twelve firm groups. Their test may therefore have lacked the power to detect differences like those our bidders show, but our experiment does not test their data.

Future Work

We think that there are many other methods from the antitrust literature that are likely to be fruitful when applied to AI agent environments. We think that lessons learned in the application of exchangeability here may be insightful to that project.

6. Conclusion

We applied Bajari and Ye's (2003) exchangeability test to a simulated procurement auction in which all six bidders were honest Claude Haiku 5.5 agents. We expected agents running the same model to converge on a common bidding strategy, so that the test would reject honest pairs at close to its nominal 5% rate. They did not converge, and in the full environment the test rejected 32.2% of honest pairs. Memory was the main cause: some agents with memory drifted into idiosyncratic strategies, such as routinely sitting out projects, and removing memory and the scratchpad lowered the rate to 9.4%. The rate fell to its nominal level only when agents also lost the bid history and the cost formula and mostly stopped pricing against their rivals.

These results come from one model in one market, with 12 runs per condition, and they do not show how often the test catches LLM colluders, because the model largely declined to collude. Within that scope, the

test does not work as a collusion detector: the one configuration in which it held its nominal rate removed persistent memory and detailed market information, features deployed agents are likely to have, along with much of what colluders would need to coordinate. More broadly, game-theoretic detection methods treat an equilibrium strategy as the benchmark for honest behavior, and for LLM agents whether such a benchmark exists is an empirical question. Our results suggest answering it before relying on such a method. Measuring a method's false-positive rate on honest agents is a useful first check, and one that does not require agents willing to collude.

Code and Data

- **Code repository:** https://github.com/samueldeegan/llm-agent-collusion-sprint. This is a substantially modified fork of the code released by Agrawal et al. (2025) at github.com/andyjliu/llm-agent-collusion. The procurement market, LLM bidders, cartel channel, Bajari–Ye detector and analysis code are in `src/procurement/`. The upstream code and data are kept under `upstream/` and `src/continuous_double_auction/`, unchanged apart from Python 3.11 and Claude 5.5 compatibility fixes. The Git tags `round1` and `round2` mark the commits that produced each experiment's results.
- **Data:** All run tables, analysis outputs, reports and agent transcripts for both experiments are in `results/`. That covers 80 runs of 110 projects each, with Claude Haiku 5.5 bidders, run on October 25, 2026. Both analyses run offline and deterministically from the committed run tables. Rerunning the experiments themselves requires an Anthropic API key and costs about $60 in total.
- **Design documentation:** `docs/` contains the full design specification, with the source of every parameter, and a log of design decisions.

References

Agrawal, K., Teo, V., Vazquez, J. J., Kunnavakkam, S., Srikanth, V., & Liu, A. (2025). *Evaluating LLM agent collusion in double auctions*. arXiv. https://doi.org/10.48550/arXiv.2507.01413

Bajari, P., & Ye, L. (2003). Deciding between competition and collusion. *The Review of Economics and Statistics*, *85*(4), 971–989. https://doi.org/10.1162/003465303772815871

Bonjour, T., Aggarwal, V., & Bhargava, B. (2022). Information theoretic approach to detect collusion in multi-agent games. In J. Cussens & K. Zhang (Eds.), *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence* (Proceedings of Machine Learning Research, Vol. 180, pp. 223–232). PMLR. https://proceedings.mlr.press/v180/bonjour22a.html

Fish, S., Gonczarowski, Y. A., & Shorrer, R. I. (2026). *Algorithmic collusion by large language models*. arXiv. https://doi.org/10.48550/arXiv.2404.00806

METR. (2026, August 26). *Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident*. https://metr.org/hugging-face-incident-report-aug-2026.pdf

Nakamura, M., Kumar, A., Das, S., Abdelnabi, S., Mahmud, S., Fioretto, F., Zilberstein, S., & Bagdasarian, E. (2026). *Colosseum: Auditing collusion in cooperative multi-agent systems*. arXiv. https://doi.org/10.48550/arXiv.2602.15198

OpenAI. (2026). *OpenAI–Hugging Face incident technical report*. https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf

Rose, A., Cullen, C., Abdelnabi, S., Torr, P., Kaplowitz, B. G., & Schroeder de Witt, C. (2026). *Detecting multi-agent collusion through multi-agent interpretability*. arXiv. https://doi.org/10.48550/arXiv.2604.01151

Shah, A., Zhu, K., Jiang, Y., Wang, J. G., Dayi, A. K., Horton, J. J., & Parkes, D. C. (2026). *LLM bidders preserve the mechanism-level orderings of human bidders*. arXiv. https://doi.org/10.48550/arXiv.2507.09083

Xu, X., Wu, C., Lu, J., Tan, K., Tao, S., & Hong, H. (2026). *Collusion with competitive marginals: Price-level audits are blind by construction*. arXiv. https://doi.org/10.48550/arXiv.2607.26385

LLM Usage Statement

We used Claude to scaffold and debug the pipeline and to draft sections of this report.