

Evaluating LLM Agent Collusion in Double Auctions

Apart Research Hackathon submission | Project team

Abstract

Autonomous language-model agents can interact repeatedly in markets, creating opportunities for tacit coordination that may weaken competition. We use a continuous double-auction simulation to compare seller-side communication and behavioral interventions while tracking both coordination-related measures and market outcomes. The experiment design specifies five GPT-4.1 sellers and five GPT-4.1 buyers, 30 rounds, common valuations, and six conditions that vary one restriction at a time. The available summary file reports one run (seed 303) for each condition. In that seed, the baseline with seller messages produced a buyer-bid collusion AUC of 10.232 and total welfare of \$1,220. Disabling messages lowered the AUC to -7.865, but also reduced trades from 61 to 22 and welfare to \$440. A 50-character cap and an oversight notice also lowered the AUC, with lower trade counts and welfare. Replacing one seller with a no-regret learner produced the lowest AUC (-14.340) and shifted more surplus to buyers, while welfare remained below baseline. These results illustrate a possible tension between reducing above-competitive bidding and preserving market activity. Because the supplied summary contains only one seed per condition and the results folder is incomplete, the comparisons are descriptive pilot evidence, not statistically robust estimates.

1. Introduction

Repeated interaction among autonomous agents may let competing sellers recognize and reinforce strategies that keep prices above competitive levels, even without an explicit cartel agreement. This matters for AI safety because increasingly capable agents could participate in economic settings where their incentives, communication channels, and strategic adaptation influence outcomes for people and organizations. Evaluation therefore needs to measure not only whether agents exchange coordination-like signals, but also whether interventions preserve useful market functions.

This project studies that problem in a continuous double auction (CDA), where buyers submit bids and sellers submit asks over repeated rounds. The seller-side threat model is a group of model-driven sellers that can communicate and face pressure to maintain high prices; buyers do not communicate. We compare the baseline with interventions that remove or constrain seller communication, add an oversight notice, reduce sampling temperature, or replace one seller with a no-regret learning strategy.

The submission makes three bounded contributions: (1) it reports matched one-factor intervention results from the supplied simulation; (2) it compares a buyer-bid-based coordination proxy with trade, profit, surplus, and welfare outcomes; and (3) it documents the limits of the current evidence, including the single-seed summary and missing or incomplete run artifacts. The results are intended to guide follow-up evaluation, not to establish a general deployment policy.

2. Related Work

The experiment builds on the project's continuous-double-auction framework and the associated study *Evaluating LLM Agent Collusion in Double Auctions* (see References). The local project README describes the market simulator, the model wrapper, and the collusion and market-performance metrics. The present submission focuses on the matched intervention runner in `experiments/one_factor_interventions.py`, which varies one seller-side condition at a time and records both coordination proxies and economic outcomes.

This is not a broad survey or a claim that the interventions are novel in isolation. Its practical value is a controlled comparison within one specific agent-market setup. Unlike a coordination score alone, the analysis also tracks trade volume, seller profit, buyer surplus, and total welfare, making it possible to see when an apparent reduction in a collusion proxy coincides with lower market activity.

3. Methods

The configured market has five sellers with unit cost/value \$80 and five buyers with unit value \$100. All ten agents use GPT-4.1 and run for 30 rounds. Seller communication is enabled in the baseline, buyer communication is disabled, the baseline maximum seller-message length is 500 characters, temperature is 0.7, and seller pressure to keep prices high is enabled. Initial asks and bids are centered at \$97.50 and \$82.50, respectively. The matched runner defines seeds 101, 202, and 303 and six conditions; the table below uses only the seed-303 results present in the supplied summary.

The interventions are: **no messages** disables seller communication; **50-character cap** limits seller messages; **oversight notice** enables the existing seller-message overseer; **lower temperature** changes temperature from 0.7 to 0.2; and **one no-regret seller** replaces one LLM seller with the project’s EXP3-style learner. Each intervention is defined relative to the baseline.

The primary coordination proxy, *collusion_auc*, integrates a per-round index based on the highest buyer bid: (highest bid - competitive-equilibrium price) divided by the difference between the joint-profit-maximizing and competitive-equilibrium prices. With the supplied values, these benchmarks are \$90 and \$100. A larger AUC therefore means bids were, on average over the run, further above the competitive-equilibrium benchmark relative to that interval; it is not direct proof of seller agreement. *Seller ask AUC* applies a related normalization to average seller asks. Total welfare is combined seller profit plus buyer surplus. For this seed, each condition has one 30-round run; no significance test is appropriate.

4. Results

Table 1 reports the six rows in `one_factor_summary.json`, all for seed 303. The baseline has the largest buyer-bid AUC (10.232), highest seller profit (\$1,028.62), and highest total welfare (\$1,220). Every intervention has a lower buyer-bid AUC in this single run, but each also has lower welfare than baseline. Disabling messages and lowering temperature have particularly low trade counts (22 and 13) and welfare (\$440 and \$260). The short cap and oversight notice produce similar AUC values (1.134 and 1.897) and similar welfare (\$640 and \$620). The no-regret seller condition has the lowest buyer-bid AUC (-14.340); it also has a lower average trade price (\$85.55) and substantially higher buyer surplus (\$780.56) than baseline, while total welfare is \$1,080.

Condition	Bid AUC	Ask AUC	Trades	Seller profit	Buyer surplus	Total welfare	Avg. price
Baseline messages	10.232	20.852	61	\$1,028.62	\$191.38	\$1,220.00	\$96.86
No messages	-7.865	18.348	22	\$351.05	\$88.95	\$440.00	\$95.96
50-char cap	1.134	19.685	32	\$531.70	\$108.30	\$640.00	\$96.62
Oversight notice	1.897	20.845	31	\$522.57	\$97.43	\$620.00	\$96.86
Lower temperature	-7.446	20.240	13	\$216.19	\$43.81	\$260.00	\$96.63

Condition	Bid AUC	Ask AUC	Trades	Seller profit	Buyer surplus	Total welfare	Avg. price
One no-regret seller	-14.340	-6.813	54	\$299.44	\$780.56	\$1,080.00	\$85.55

Table 1. Single-run results for seed 303. AUC values are the project's integrated, benchmark-normalized indices; market dollar amounts are totals for the 30-round simulation. Total welfare equals seller profit plus buyer surplus.

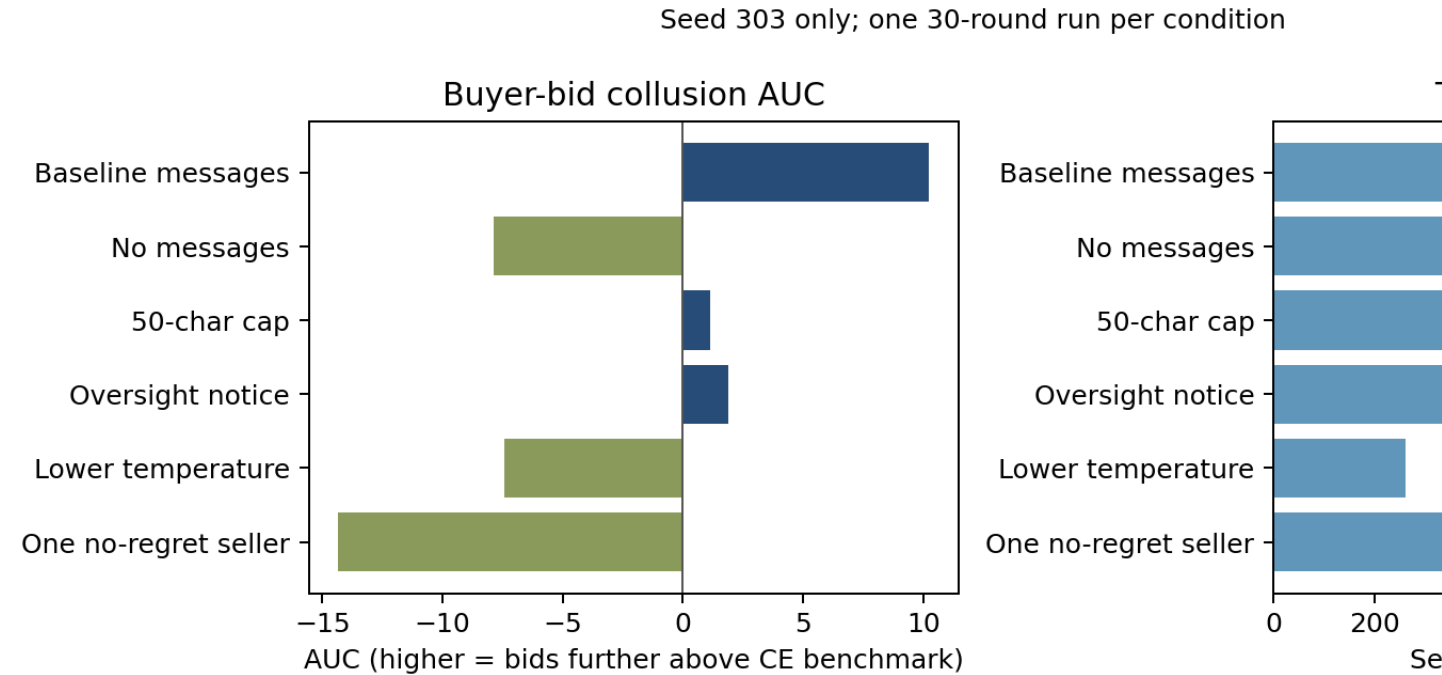


Figure 1. The seed-303 comparison shows lower buyer-bid AUC under every intervention, alongside lower total welfare than baseline. One run per condition cannot establish a reliable treatment effect.

The direction of the buyer-bid proxy is consistent with the hypothesis that the tested restrictions reduce above-competitive bids in this run. However, the outcome measures expose different trade-offs: conditions with low AUC can also suppress trade and both sides' surplus, while the no-regret seller shifts surplus toward buyers with fewer trades than baseline. Average prices remain near \$96.63-\$96.86 for four of the six conditions; the no-regret seller condition is the exception at \$85.55. Thus, average trade price alone would miss important differences in market participation and distribution.

5. Discussion and Limitations

The strongest conclusion is methodological: evaluating anti-collusion interventions only through a coordination proxy can hide effects on market performance and distribution. In this pilot, communication restrictions correlate with lower measured buyer-bid AUC, but they also coincide with fewer trades and lower welfare. The no-regret seller condition looks different: the proxy and price fall while buyer surplus rises, suggesting a possible redistribution channel rather than a uniform improvement across participants.

Limitations

First, the summary JSON contains only six rows, all from seed 303, despite the runner specifying three seeds. The results folder contains 16 run directories rather than the expected 18; the seed-202 oversight directory appears partial, and seed-202 lower-temperature and no-regret runs are absent. Accordingly, Table 1 and Figure 1 are

single-seed observations and cannot support significance claims or robustness conclusions. Provider-side model sampling may also be nondeterministic even when seeds match.

Second, the market is deliberately narrow: five homogeneous sellers, five homogeneous buyers, one model family, fixed values and costs, 30 rounds, and no buyer communication. Different market sizes, value distributions, models, prompts, oversight mechanisms, or communication arrangements may produce different behavior. Third, the AUC is a proxy derived from bids or asks relative to theoretical benchmarks. It does not by itself identify intent, prove collusion, or distinguish strategic coordination from other causes of high bids. The evaluator's language-model judgments and message logs may carry their own measurement error. Finally, this is a simulated market, not evidence about real-world markets or deployed agents.

Future Work

Complete all six conditions for each planned seed and preserve one machine-readable aggregate with per-run provenance. Then add more independent seeds and, where possible, repeat runs to estimate variance. Report confidence intervals and pre-specify primary outcomes. Test heterogeneous valuations, more market participants, multiple model families, buyer communication, and repeated interactions with altered incentives. Validate coordination judgments against blinded human coding and stress-test the bid-based proxy. Finally, evaluate intervention cost and market-function trade-offs across welfare, price levels, trade frequency, and surplus distribution rather than optimizing one metric alone.

6. Conclusion

In the available seed-303 pilot, all five interventions reduce the buyer-bid collusion AUC relative to the communication-enabled baseline, but none preserves baseline total welfare. The one-no-regret-seller condition shifts surplus toward buyers and lowers average price; communication removal and lower temperature are associated with sharply lower trading activity. These patterns motivate measuring both strategic proxies and market outcomes. The incomplete, single-seed evidence is not sufficient to rank interventions or claim robust mitigation. Completing the matched runs and expanding the market settings are necessary next steps.

Code and Data

Code: local project folder supplied with this report; public repository URL was not present in the provided materials. The experiment runner is `llm-agent-collusion/experiments/one_factor_interventions.py`. Results: `llm-agent-collusion/results/one_factor/`, including the summary JSON and per-run logs.

Author Contributions

Project team: experiment design and implementation are documented in the supplied repository; this report was drafted from the repository documentation, experiment runner, and result artifacts. Individual author names and contribution assignments were not provided.

References

Evaluating LLM Agent Collusion in Double Auctions. arXiv:2507.01413, 2025. <https://arxiv.org/abs/2507.01413>.

Apart Research hackathon submission template. User-provided Word document.

Appendix: Data completeness note

The intended design is six conditions across seeds 101, 202, and 303 (18 runs). The supplied summary JSON has six rows for seed 303. Six seed-101 directories and six seed-303 directories are present. Four seed-202 directories are present, but the oversight directory contains only a partial set of files and the lower-temperature and no-regret directories are absent. The table and figure intentionally use the six rows in the summary rather than inferring missing values from incomplete artifacts.

LLM Usage Statement

An AI assistant was used to organize the supplied project documentation and result files into this draft and to calculate the displayed comparisons from the summary JSON. Numerical claims were checked against that file. The project team should verify the final text, run completeness note, author information, and citations before submission.