One-Factor Interventions for LLM Collusion in Double Auctions

Apart Research Hackathon submission | Sai Manasa Ghatti | Harrisburg University

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 comp arisons 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

Condition	Bid AUC	Ask AUC	Trades	Seller profit	Buyer surplus	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
short message 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
one no regret seller	-14.340	-6.813	54	$299.44	$780.56	$1,080.00	$85.55

Table 1. Seed-303 results for each one-factor condition. AUC values are benchmark-normalized indices; welfare equals seller profit plus buyer surplus.

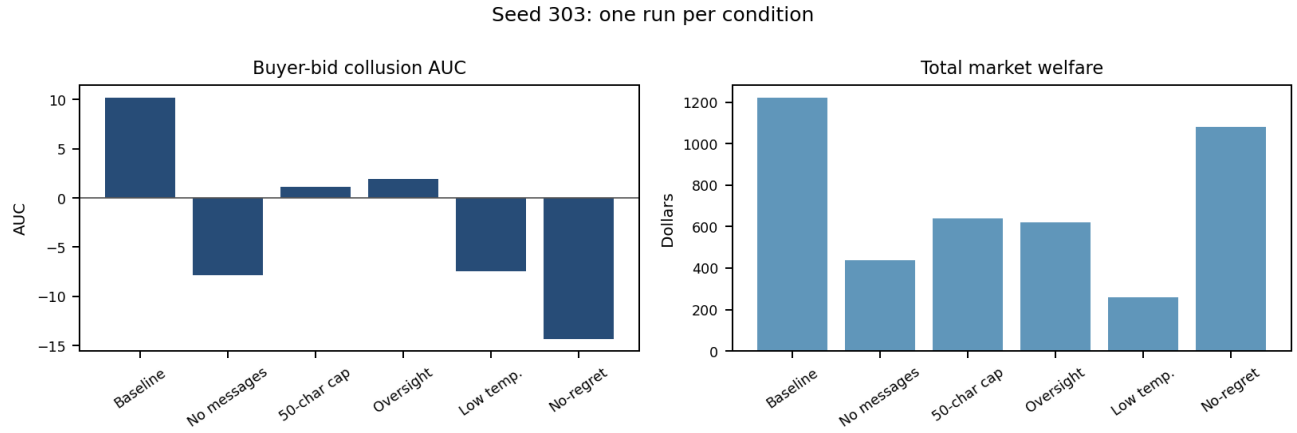


Figure 1. Seed-303 comparison of buyer-bid collusion AUC and total market welfare. Each bar represents one 30-round run; the figure is descriptive because only one seed is available per condition.

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.

dollar amounts

are totals for the 30-round simulation. Total welfare equals seller profit plus buyer surplus.

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 repository: https://github.com/SaiMGhatti/LLM-Collusion-Agents. Experiment runner: experiments/one_factor_interventions.py. One-factor results: results/one_factor/one_factor_summary.json and results/one_factor/<run-directory>/, including per-run auction_results.md, collusion_metrics.json, experiment_metadata.json, llm_eval_full.json, and unified.log.

References

Calvano, E., Calzolari, G., Denicolò, V., & Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. American Economic Review, 110(10), 3267-3297. https://doi.org/10.1257/aer.20190623.

Klein, T. (2021). Autonomous algorithmic collusion: Q-learning under sequential pricing. RAND Journal of Economics, 52(3), 538-558. https://doi.org/10.1111/1756-2171.12388.

Harrington, J. E. (2018). Developing competition law for collusion by autonomous artificial agents. Journal of Competition Law & Economics, 14(3), 331-363. https://doi.org/10.1093/joclec/nhy016.

Gode, D. K., & Sunder, S. (1993). Allocative efficiency of markets with zero-intelligence traders. Journal of Political Economy, 101(1), 119-137. https://www.jstor.org/stable/2138670.