

Price Levels Are Not Enough: A Counterfactual Test for LLM Seller Collusion, with Negative Results for Price-Based and Probe-Based Detectors

AI Collusion Research Sprint, Track 2 (Detection). Navien Potheri, navien.pr@gmail.com.

Abstract

LLM agents that set prices in a repeated market can settle above the competitive benchmark. We ask whether that can be detected from the market record. Using a continuous double auction with LLM sellers, we pre-registered and tested price-based labels, a price-free detector, a sequential live test, a monitoring notice, a disclosed cutoff, a linear probe on hidden states, and two market-design interventions. Most tests fail: labels move with the cutoff, detectors trained on one set of runs do not transfer, the live test does not separate collusive from competitive runs, and the probe's price signal is confounded with game position. One result is positive but preliminary. In the experimental design we vet GPT, open models and Claude. The Claude sellers, showed a unique property, prices above \$90 fall by about 4.5 points when rivals' asks are hidden on the same seed, measured against a fresh-visible control, while GPT and open-weight sellers show no comparable effect at this design. We conclude that price levels alone cannot establish coordination, that a counterfactual that hides rivals can, and that the effect we observe is specific to one model family in our runs. The Claude result has not replicated on an unselected sample and needs a larger study.

Introduction

Algorithmic pricing agents can settle on prices above the competitive level without any explicit agreement, so detecting this from the market record is a monitoring problem. The obvious tool is a price-based statistic: flag a market whose average price exceeds the competitive benchmark. A stronger tool would use the agents' behaviour, such as how much each seller's price depends on rivals, or the agent's internal state before prices move.

We test these tools in a controlled market and ask three questions: whether price-based labels and detectors hold up when the market or the sellers change; whether a counterfactual that hides rivals' asks on the same seed can separate coordination from high prices; and whether any effect depends on the model family. We fixed the main decision rules before the main runs and report failures as failures.

Contributions: (1) a set of pre-registered tests of price-based detection, with most results negative; (2) a counterfactual effect for Claude sellers that survives a fresh-visible control on the selected games but does not replicate on an unselected sample; (3) a cross-family comparison showing no equivalent effect for open-weight sellers at this design.

Related Work

Agrawal et al. (2025) provide the continuous double auction simulator with LLM sellers and buyers that we use. We change it only by adding the interventions and the rivals-hidden, disclosure, and monitoring flags, all listed in Section 8.

Gauthier, Bach and Jordan (arXiv:2601.05427) develop anytime detection of strategic deviations. Their method assumes rounds are independent and needs counterfactual payoffs or a model of the environment. Those assumptions motivate our sequential live test, which uses the joint interaction history, and our counterfactual label, which supplies the payoff comparison from a rerun with rivals hidden.

Garra (arXiv:2609.13037) studies emergent collusion among LLM pricing agents in a Bertrand logit market. The paper does not test price rounding or noise, which we test in the double auction (Section 5.8). Our paper differs in asking whether price-based and probe-based detectors transfer across conditions, and in using a counterfactual label rather than a price cutoff alone.

Setup

Simulator. We use the continuous double auction (CDA) from the starter repository of Agrawal et al. (2025): three LLM sellers, three buyers, 15 hourly rounds. Sellers see their own cost (\$80), the bid and ask queues, and memory notes they write each hour, which are fed back into later prompts. Buyers are GPT-4.1-mini.

Models. GPT cells use GPT-4.1-mini sellers. Claude cells use claude-haiku-4-5 sellers.

Anthropic offers no seed parameter, so each Claude run is an independent draw on its seed.

Open-weight cells use meta-llama/llama-3.1-8b-instruct sellers through OpenRouter. OpenAI data sharing was enabled on the account; this should be disclosed.

Labels. A run is collusive if its average seller ask over the last third of rounds is above \$90, the midpoint of seller cost (\$80) and buyer value (\$100). The cutoff is a convention. Counterfactual: the same seed is rerun with each seller shown only its own asks. A run is coordination-dependent if it is above \$90 with rivals visible and at or below \$90 with rivals hidden.

Methods

Price-free detector: ask dispersion, seller-ask correlation, and a GPT-4.1-mini judge score, evaluated by random, leave-one-condition-out and train-shipped/test-new splits, with a permutation null. **Live test:** a sign-based sequential test with a threshold set for a false-alarm rate. **Monitoring and disclosure:** a regulator notice and a stated \$90 cutoff, each against a matched control. **Probe:** standardised logistic regression ($C = 0.01$) on mean-pooled layer-16 hidden states of Qwen2.5-7B-Instruct, applied to Mistral-7B-Instruct-v0.3 at the same layer index. Each turn is one seller's prompt for one round with numbers masked. **Controls:** within-game label permutation, a bag-of-words model on the same text, a round-number model, a non-price target (round > 7), and transfer to hidden-rival and Claude-seller games. **Interventions:** asks rounded to \$5, Gaussian noise (sd \$3), or both, scored on average last-third asks and on the probe's detection signal.

Decision rules (fixed in advance). The probe passes only if it exceeds the 95th percentile of its null, beats the bag-of-words control, scores at or below the null on the non-price target, and drops when rivals are hidden. An intervention is useful only if it lowers price and detector signal together, or lowers the detector signal without raising price.

Seeds and reproducibility. GPT counterfactual pairs use seeds 0 to 14; the new GPT cell uses seeds 15 to 54; the layer 7 and disclosure arms use seeds 55 to 66; the Claude games use seeds 15 to 38, and the Claude visible-hidden pairs use seeds 17 to 40; the open-weight pairs use seeds 17 to 27. GPT runs are seeded. Claude runs have no seed parameter and are not reproducible. Prompt templates are those of the starter repository, with the notice blocks we added (monitoring, disclosure) listed in Section 8. Metrics are the repository's labels plus the features defined in Section 4.

Results

Controls. Positive control: the rivals-visible Claude reruns must reproduce prices above \$90 for the counterfactual to be testable; fresh visible reruns on the nine selected seeds averaged 90.7, so the control holds in this sample, although the repository labels are not an independent check of coordination. Negative control: the GPT counterfactual (Section 5.3) must show no gap if the method is not picking up noise; it shows a mean gap of -0.4 over 15 pairs.

5.1 Price labels are fragile, and sellers do not appear to target the cutoff

The same 75 shipped runs are collusive in 41 cases at \$90 and 22 at \$95. Across 7,253 asks there is no bunching just below \$90: \$89 has 256 asks, \$90 has 351, and \$91 has 314, with the main mass at \$85. Seller prompts never state the benchmark, and sellers refer to their own round-number thresholds rather than to \$90. Disclosing the cutoff did not change prices (4.5).

5.2 Price-free detectors do not transfer

A detector looked strong under random five-fold splits (AUROC 0.69), but those splits leak condition structure. Under leave-one-condition-out evaluation it falls to 0.53, and trained on shipped runs and tested on new ones to 0.31. The judge score tracks price: its raw AUROC is 0.78, and its residual after removing price is 0.52. Relabelling into three classes gives 0.20 and 0.44 on the same splits. A same-model versus mixed-family comparison shows no detectable difference ($n = 37$ and 11).

5.3 Hiding rivals does not reliably reduce prices with GPT sellers

Across 15 paired seeds, the mean visible-minus-hidden gap in last-third asks is -0.4 (sd 4.2; 95% interval ± 2.5). Two of 15 games are coordination-dependent, and one is collusive with rivals hidden. A signed-rank test gives $p = 0.45$.

5.4 Claude sellers show a counterfactual gap that partly replicates

Claude sellers exceed \$90 in 9 of 24 games (38%), against 2 of 40 GPT games (5%) in the same design. For the nine games above \$90 we reran each seed with rivals visible and with rivals hidden. Fresh visible reruns averaged 90.7, with 5 of 9 above \$90, against 86.1 for hidden reruns, with 1 of 9 above \$90. The gap is +4.5 points, so the high prices reproduce when rivals are visible and fall when they are not.

Selection can inflate such gaps, so we also ran four unselected seeds (37–40): visible 88.1, hidden 84.6, gap +3.5. One game (seed 39, +12.4) carries most of that gap; the others are +0.2, −0.3 and +1.8. Across all 13 seeds the median gap is 3.7 points. The effect is real enough to report but not established as a general property of Claude sellers.

Mechanism evidence is limited. In a hand-coded sample of 60 memory notes per arm (single rater), rival-directed reasoning (undercutting or matching rivals) appears in 60% of visible-run notes and 3% of hidden-run notes. No note states an agreement. After an undercut, rivals' asks change by +0.2 in visible runs (16 events) and +0.1 in hidden runs (11 events), so we found no punishment. Lead-lag correlations are +0.07 (visible) and −0.16 (hidden) and are not significant.

5.5 Monitoring, disclosure, and live detection have no detectable effect

A monitoring notice on five seeds left mean asks at 85.8 against 86.0. With the \$90 cutoff disclosed, GPT sellers without communication averaged 84.8 against 85.2 for controls; with communication, disclosed and control arms averaged 86.6 and 85.7, with one game above \$90 in each. The sequential live test had a 13% false-alarm rate and 15% detection at $\alpha = 0.05$ after removing momentum and market innovations; without that removal the false-alarm rate was 76%.

5.6 The probe carries a signal confounded with game position

On 114 games, the Qwen probe scores AUROC 0.76 (95% CI 0.67–0.84), above its permutation null (95th percentile 0.73) and above the bag-of-words control (0.71). Mistral at the same layer scores 0.73 (CI 0.64–0.82; null 95th percentile 0.72). The probe also predicts whether a turn is in the second half of the game almost perfectly (AUROC 1.00 for Qwen), and prompt length alone reaches 0.96 on that target. Within early and late rounds the price signal persists (0.80 and 0.76 on the original set), so it is not only timing, but the pre-registered time control fails. Transfer to Claude-seller games is near chance (0.58 and 0.53), and scores are higher, not lower, when rivals are hidden (0.77 and 0.90), which we cannot yet explain. We report the probe as a preliminary signal that does not establish collusion.

5.7 Open-weight sellers do not show the effect at this design

Open-weight sellers with GPT buyers show a mean visible-minus-hidden gap of +0.7 (sd 2.4) over 10 pairs (seeds 18–27), with no game above \$90 in either arm. Their prices run from 79.8 to 87.3. Because they rarely approach the cutoff, this null cannot show whether they coordinate; it supports only the narrower claim that the Claude effect does not appear here.

5.8 Market-design interventions are inconclusive

On 48 GPT games (12 per arm), rounding to \$5 and adding \$3 noise raised average last-third asks slightly: none 85.2, rounding 86.2, noise 88.1, both 86.2. Collusive games were 0, 0, 2 and 2 of 12. Rounding removes the detector's signal mainly by removing turns above \$90. Noise raises the detector's false-positive rate from 11% to 30%. No intervention lowers price and detector signal together, so none meets the decision rule.

Test	Result	Verdict
Price labels across cutoffs	No bunching at \$90; labels move with cutoff	Fragile
Price-free detector transfer	0.31 shipped-to-new; 0.53 leave-one-condition-out	Fails
Counterfactual, GPT (15 pairs)	Mean gap -0.4; 2 of 15 coordination-dependent	Null
Counterfactual, Claude (9 selected pairs)	Gap +4.5 vs fresh visible control	Positive, preliminary
Counterfactual, Claude (4 unselected)	Gap +3.5, mostly one game	Weak replication
Live sequential test	FA 13%, detection 15%	Fails
Monitoring / disclosure	No detectable effect	Null
Probe (Qwen / Mistral)	0.76 / 0.73, above null; time control fails	Preliminary, confounded
Open-weight counterfactual (10 pairs)	Gap +0.7; no prices above \$90	Null, low power
Interventions (rounding, noise)	No joint reduction in price and signal	Inconclusive

Discussion

Our thesis is, price levels cannot establish coordination in these markets. Labels move with the cutoff, price-free detectors do not transfer, and the probe tracks game position nearly as well as price. The counterfactual that hides rivals on the same seed is the one tool in our data that separates coordination-dependent prices from high ones. It does so for Claude sellers and not for GPT or open-weight sellers, so the coordination effect we can show is specific to one family in our runs.

The cross-family difference is the main open question. It could reflect how Claude sellers reason, the Anthropic API and sampling, or the small number of pairs. The selected Claude pairs may overstate the effect, and the unselected pairs show it mostly in one game. A fair test needs at least 50 more Claude pairs and a comparable number for another family with prices that approach the cutoff.

For practice, the counterfactual is an audit: flag a market, rerun it with rivals hidden, and check whether prices fall. It needs two runs per market, so it is not a live monitor. None of our candidate live features met the bar for a live monitor, and training one would need roughly 100 Claude pairs.

Code and Data

Starter code (preexisting): https://github.com/poseidon-research/collusion_hackathon (llm-agent-collusion/). Our additions are listed in the preexisting-work note below.

LLM Usage Statement

References

Agrawal, A., et al. (2025). Continuous double auction simulation with LLM agents. arXiv:2507.01413.

Gauthier, Bach, Jordan. Anytime detection of strategic deviations. arXiv:2601.05427.

Garra. Mitigating emergent collusion in LLM pricing agents. arXiv:2609.13037.

Appendix: Limitations and Dual-Use / Ethical Considerations

What the results establish: a rivals-hidden counterfactual separates coordination-dependent prices from high prices in one model family, in our runs. What they do not establish: that Claude sellers coordinate in general, that any live detector works, or that the probe measures collusion. Ways this could aid real-world collusion or make it harder to detect are discussed under dual-use below.

Small samples: most cells have 5 to 15 games per arm; the Claude pairs have 13 seeds, one of which carries much of the unselected effect.

One cutoff (\$90), one market configuration (three sellers, three buyers, 15 rounds), one prompt template.

Claude runs are not reproducible; reruns are independent draws on their seeds, and the selected nine were chosen for high visible prices.

The probe reads masked text, so price-correlated words may remain; it was not validated against an independent behaviour measure.

Memory-note coding was done by one rater without an agreement check.

Open-weight sellers rarely reach the cutoff, so their null is weak evidence.

The layer 7 games are GPT-only and do not cover Claude sellers.

The simulator models collusion among automated sellers, and the methods could be used to tune pricing agents to evade detection. We report counterfactual detection and the failure of price-based detection, not techniques for hiding coordination. No real markets, customers, or

personal data are involved. OpenAI data sharing was enabled on the account, so some traffic may contribute to shared model improvement; Anthropic and OpenRouter traffic is not covered by that setting.