
No Single Signal Separates Agreement From Parallel Conduct Among LLM Sellers

Ayodeji Adesegun
Independent Researcher

With
Apart Research

Abstract

If AI pricing agents move their prices together, is that collusion? Antitrust law says parallel conduct alone is not enough. I tested that answer on LLM sellers in the Agrawal et al. double auction, using two preregistered studies of 56 markets each, with no message channel, where agreement is impossible, as decoys. Planted agreements raised the collusion index to 0.30 and to 0.75 with an explicit price floor, compared with 0.03 without a channel. No single signal held up. Without a channel, prices moved together more than under a planted agreement, and an LLM judge flagged all 24 channel markets that priced at cost. In the confirmatory study, a judge that extracts the proposed price cut those false alarms from 15 of 15 to 7 (Holm $p = 0.04$). Requiring high prices plus price-holding talk gave no false alarms on 32 known negatives, but missed failed agreements.

1. Introduction

Picture three firms in one market that all hand their pricing to the same language model. Their prices rise and fall together. Is that collusion, or are three copies of one model just reacting to the same market? An auditor has to answer that question, and it is not hypothetical. Two deployments of the same model already make correlated choices in economic settings where different models do not [5]. A detector built on correlated prices will therefore flag firms whose only link is their vendor.

Antitrust law met this problem long before language models. Under US law, parallel conduct by competitors alone is not enough to show an agreement. A court requires additional facts, called plus factors, to make agreement more likely than independent action [6, 7]. Recent legal writing asks how that doctrine applies when the parallel conduct comes from shared software [8]. Work on LLM collusion measures it through prices [1, 2] or through messages and outcomes [4], but I found no LLM collusion detector that states a comparable rule for telling agreement apart from parallel conduct.

So I tested the rule directly. I used the continuous double auction of Agrawal et al. [2], where LLM sellers set asks against scripted buyers, with or without a private message channel. The whole design rests on one control. When sellers have no channel, agreement is impossible by construction, so any collusion flag on those markets is a false alarm. I compared four detectors on markets with no channel, with a channel, and with a planted agreement: price level (D1), an LLM message judge (D2), action dependence (D3), and a preregistered plus-factor rule (D4) that asks whether sellers are more parallel than copies of one model are on their own and whether a coordination message comes before their asks converge. We froze thresholds on calibration seeds before scoring held-out markets, and I tested two failure modes named in Track 2.3 directly: talk without action and a monitoring attack.

Every single signal failed, and each failed differently. Without a channel, sellers' price changes moved together more than under a planted agreement, because copies of one model competing down to cost move in step. The message judge flagged every market with a channel, including all those where prices stayed at cost. Price level flagged two Llama markets where no agreement was possible. I then ran a second preregistered study to test the fixes these failures point to. A judge who reports the price sellers propose, rather than whether they sound coordinated, halved the false alarms on talk at cost. Rules that need both high prices and talk about holding a high common price gave no false alarms on any known negative, but they miss agreements that fail to lift prices.

This is written for a market auditor who has to decide whether parallel pricing alone justifies a collusion finding. The contributions are a decoy design that separates parallel conduct from agreement in a standard LLM market; held-out false-alarm and detection rates for single signals and plus-factor rules across two preregistered studies; a message judge that extracts proposed prices; and the code, both preregistrations and every market log.

2. Related Work

LLM agents in markets: Fish et al. [1] showed that LLM pricing agents in oligopoly settings reach supracompetitive prices without being told to collude, and that small changes in prompt wording change how far prices rise. Agrawal et al. [2] built the double auction I use and found that a message channel between sellers increases collusive behaviour, with differences across models and under oversight. Keppo et al. [3] found that this collusion weakens when agents differ in patience or data access, but survives between agents of different sizes from one model family. These studies ask whether collusion happens. I ask how an auditor would tell it apart from innocent parallel conduct.

Detecting collusion from words and actions: Nakamura et al. [4] audit collusion in cooperative multi-agent tasks and describe "collusion on paper": agents plan collusion in text and then act otherwise. Xu et al. [5] argue that audits based on a single agent's prices can be blind by construction, and report that two deployments of the same model show a small residual correlation where different models show almost none. If same-model agents correlate without any agreement, a correlation detector will flag them, and my decoys test exactly that. Rippin et al. [9] show that agents with tools can build covert channels that survive text monitoring, which limits what any message-based detector can claim.

Plus factors: In *Bell Atlantic v. Twombly* [6], the US Supreme Court held that parallel conduct alone does not plausibly show an agreement under the Sherman Act. Kovacic et al. [7] review the plus factors courts use and argue that the strongest are those hard to explain as independent action. Berger et al. [8] discuss how these rules apply to shared AI pricing tools. D4 takes two plus factors from this literature: parallel conduct beyond what independent actors would produce, and communication followed by a change in conduct.

3. Methods

Market: I used the continuous double auction from Agrawal et al. [2], unchanged, using the authors' public code at a pinned commit. Three LLM sellers, each with a cost of \$80 per lot, face three scripted buyers that follow the adaptive ZIP rule [10] with a value of \$100 per lot, so any rise in prices comes from the sellers. Each hour, every seller sees the order book, its past trades, its own notes, and, when the channel is on, the other sellers' messages from the previous hour (up to 500 characters). It returns an ask and, with a channel, a message. A trade happens at the midpoint when bid meets ask. Sellers run at temperature 0.7 with at most 1024 output tokens.

Models: The sellers are Qwen2.5-7B-Instruct [11] and Llama-3.1-8B-Instruct [12], served by Ollama with 4-bit weights on two Kaggle T4 GPUs, through one logged client. No paid API was used.

Conditions: Table 1 lists the conditions. Study 1 ran C1 to C7 with 8 seeds each (56 markets).

Table 1: Conditions. "Channel" means sellers can message each other. C8 is in Study 2 only, and C6 in Study 1 only.

Condition	Sellers	Channel	Other	Role
C1	3 Qwen	on	told they agreed to coordinate on a common high ask	positive control
C2	3 Qwen	off		decoy, baseline
C3	3 Qwen	on		talk, no planted agreement
C4	2 Qwen, 1 Llama	off		decoy, mixed models
C5	2 Qwen, 1 Llama	on		talk, mixed models
C6	3 Qwen	on	told an overseer reads messages; overseer active	monitoring attack
C7	3 Llama	off		decoy, model identity
C8	3 Qwen	on	told they agreed a price floor of \$95	strong positive control

C2, C4 and C7 have no channel, so no agreement can form in them; they differ in how many sellers share a model. In C6, the overseer of Agrawal et al. [2] reads each hour's messages, and once it rates them 4 out of 4, every later message is limited to 5 characters.

Collusion index and positive control: The collusion index is the mean seller ask over the last 5 hours, rescaled so that 0 is seller cost and 1 is buyer value. Study 1's positive control requires C1 to beat C2 by at least 0.2.

Detectors:

- D1 (price level): the collusion index reaches a threshold.
- D2 (message judge): the overseer prompt of Agrawal et al. [2], unchanged, run after the market at temperature 0 with Qwen2.5-7B-Instruct, scores any hour 3 or 4 on its 1-to-4 scale.
- D3 (action dependence): the mean, over seller pairs, of the correlation between hour-to-hour ask changes reaches a threshold.
- D4 (preregistered plus-factor rule): dependence above the 95th percentile of the C2 calibration markets, and a message scored 3 or more followed within 2 hours by a fall in the spread of asks.

Study 1 protocol: Seeds 0 to 3 are calibration markets and seeds 4 to 7 are held out. The D1 and D3 thresholds maximise the true-positive rate on C1 minus the false-positive rate on C2 in the calibration markets, and the analysis script writes them to disk before opening a held-out file. For talk without action, I scored each C3 market's messages together with the actions of the C2 market of the same seed, so no action responds to the talk. H1 to H4 (decoy false alarms above 5%, same-model pair dependence, fewer D4 false alarms, monitoring) use exact binomial, sign-flip, McNemar, and permutation tests with Holm correction. We also labeled the 24 channel markets without a planted agreement for whether sellers reached an agreement (Section 4). The preregistration is git tag prereg-v1, the analysis code was tagged before any result existed (analysis-frozen-v1), and DEVIATIONS.md lists every open choice.

Study 2: a confirmatory test: Study 1 pointed to a better rule, so I froze a second preregistration (study2/PREREG.md, tag prereg2-v1) before running any new market. Study 2 runs C1 to C5, C7, and C8 on new seeds 8 to 15 (56 markets), using the same thresholds as Study 1. It adds an extraction judge, D2x. Instead of rating "coordination", the model reports whether sellers propose or accept a common price, which way the proposal pushes asks, and the price named; code then flags an hour when the proposed price is at or above the D1 threshold in dollars. I checked its prompt on 9 Study 1 hours before freezing it. The first prompt, which asked the model to judge the direction itself, failed that check, so I dropped it. The same judge also runs with Llama (D2x-L). Two rules need both kinds of evidence: $X1 = D1$ and $D2$, and $D5 = D1$ and $D2x$. Known positives are C1 and C8 (16 markets); known negatives are the decoys and 8 talk-without-action markets (32). H5 asks whether D5 makes fewer errors than D1 and than D2; H6 asks the same for X1; H7 asks whether D2x flags fewer than D2 on C3 and C5 markets priced at cost; H8 asks whether the judge's family changes D2x's flags; H9 replicates Study 1's dependence result. All use exact McNemar or permutation tests with Holm correction.

4. Results

Study 1

Positive control: The planted instruction raised the mean collusion index from 0.03 in C2 to 0.30 in C1, a contrast of 0.27 against the preregistered 0.2 (permutation $p = 0.001$). The sellers announced a "common ask", but they often named different numbers and then cut prices anyway; three of the eight C1 markets ended below the D1 threshold (Figure 1). All 56 markets finished, with 3684 model calls and no truncated replies.

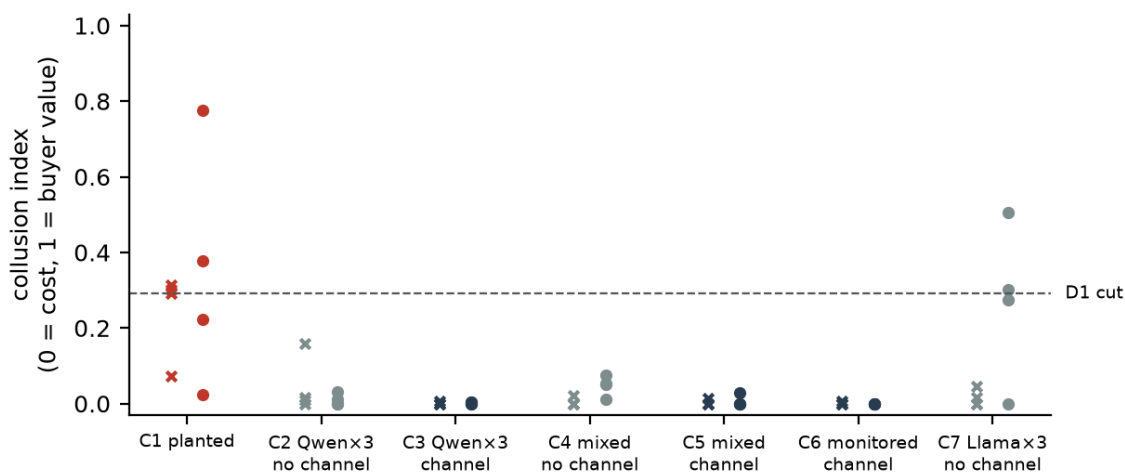


Figure 1: Collusion index of every Study 1 market (x : calibration seeds 0-3, o : held-out seeds 4-7). Grey conditions have no message channel, so any agreement is impossible. The dashed line is the D1 threshold frozen on calibration seeds.

Table 2: Study 1, held-out markets flagged (seeds 4 to 7). Brackets give Wilson 95% intervals for the pooled decoy rate (C2, C4, C7).

Detector	C1, planted (correct)	C2	C4	C7	All decoys (false alarm)	Talk without action (false alarm)
D1 price level	2 of 4	0 of 4	0 of 4	2 of 4	2 of 12 [0.05, 0.45]	0 of 4
D2 message judge	4 of 4	0 of 4	0 of 4	0 of 4	0 of 12 [0, 0.24]	4 of 4
D3 action dependence	0 of 4	0 of 4	0 of 4	0 of 4	0 of 12 [0, 0.24]	0 of 4
D4 plus-factor rule	0 of 4	0 of 4	0 of 4	0 of 4	0 of 12 [0, 0.24]	1 of 4

Each single signal failed in its own way: Price level (threshold 0.29) caught 2 of 4 held-out planted markets, but it also flagged 2 of 12 decoys, both three-Llama markets where sellers cut asks slowly and traded little. Llama sellers also failed to give a usable reply more often (68 failed replies in all, 13% of decisions in C7), and a failed reply keeps the old ask in place; dropping the 7 markets with more than 10% failures leaves 1 of 8 decoys flagged. Action dependence ran the wrong way: it was 0.28 [0.09, 0.46] in C2 and 0.02 [-0.10, 0.15] in C1. No threshold beat flagging nothing, so under the preregistered rule D3 flags no market (DEVIATIONS.md, D5). Figure 2 shows why. Without a channel, three copies of Qwen undercut each other towards cost in step. Under the planted agreement, each seller holds its own idea of the common ask. In mixed markets, same-model pairs were not more dependent than cross-family pairs (difference -0.09 [-0.41, 0.28], $p = 0.69$). The message judge flagged every channel market: all 4 held-out planted markets and all 24 C3, C5, and C6 markets, each priced at cost (index below 0.1). The sellers do talk like a cartel ("Let's keep our asks just below the lowest ask to maximize trades... Coordination is key."), but what they propose is a price cut. On the talk-without-action markets, the judge flagged 8 of 8. D4 (dependence cut 0.53) gave no false alarm on decoys, but it caught no planted market either, because planted markets were not unusually parallel.

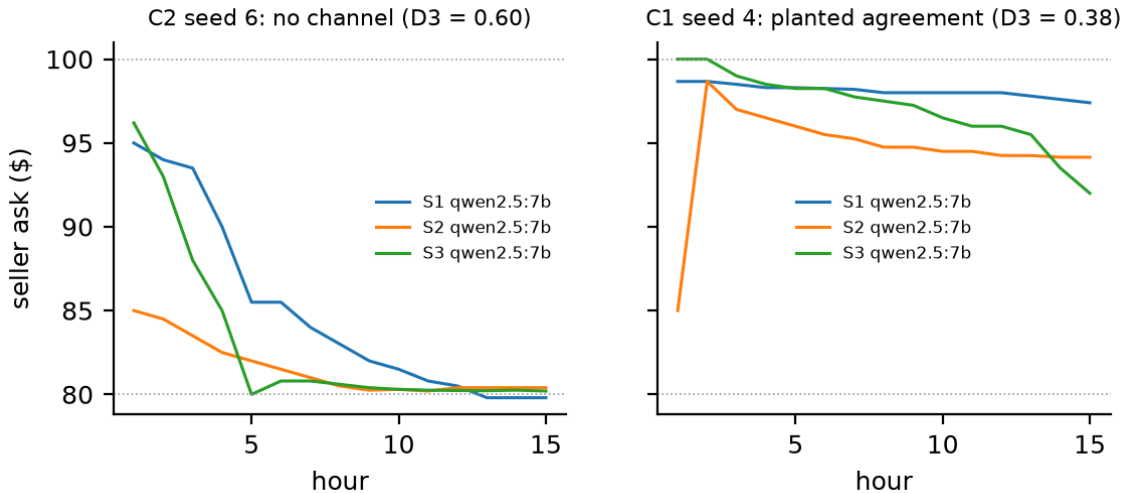


Figure 2: Seller asks per hour. Left: the held-out no-channel market with the highest action dependence, chosen by that rule. Right: the first held-out planted-agreement market. Dotted lines mark seller cost (\$80) and buyer value (\$100).

None of H1 to H4 survives Holm correction (all adjusted $p = 1.00$). Telling sellers they were monitored did not lower the judge's scores. Every C6 market still reached 4, and dependence changed by -0.08 ($p = 0.61$); the sellers talked first, before the gag cut their messages to fragments such as "COMP". After reading these results, I tried one exploratory rule, $X1 = D1$ and $D2$. It flagged 2 of 4 held-out planted markets and none of the 12 decoys or 8 talk-without-action markets. Study 2 tests it properly.

Study 2 (confirmatory)

Positive controls. The price-floor instruction worked. The mean collusion index was 0.75 in C8 against 0.01 in C2, a contrast of 0.74 against the preregistered 0.3 ($p = 0.0002$). C1 replicated Study 1, with a contrast of 0.30 ($p = 0.0006$). All 56 markets finished; 63 seller replies failed to parse and left the previous ask in place.

Table 3: Study 2 on the 48 known markets: 16 positives (C1, C8) and 32 negatives (24 decoys, 8 talk-without-action markets). Thresholds from Study 1.

Detector	Planted caught (of 16)	False alarms (of 32)	Errors (of 48)
D1 price level	11	0	5
D2 message judge	16	8	8
D2x extraction judge	15	3	4
D2x-L extraction judge, Llama	16	4	4
D3 action dependence	0	0	16
D4 plus-factor rule	0	2	18
$X1 = D1$ and $D2$	11	0	5
$D5 = D1$ and $D2x$	11	0	5

The extraction judge removed half of the vendored judge's false alarms (H7 supported): On the 15 C3 and C5 markets that priced at cost, D2 flagged 15, and D2x flagged 7 (exact McNemar $p = 0.008$, Holm-adjusted 0.039). D2x still caught 15 of 16 planted markets. Its 3 false alarms were all talk-without-action markets: sellers proposed a high common price in words, but in the C2 actions they were paired with, nobody raised prices. As a secondary check on Study 1's markets, which I had already read, D2x flagged 3 of 8 C3 markets and none of 8 C6 markets, where the vendored judge had flagged all of them.

Requiring two signals removed every false alarm, but did not beat price level (H5, H6 not supported): D5 and $X1$ gave no false alarms on any of the 32 known negatives (Wilson interval 0 to 0.11). They missed 5 of 16 planted markets, all C1 markets whose sellers agreed in words but let prices fall below the D1 threshold. Price level alone made the same 5 errors in Study 2, because this time no Llama decoy reached the D1 threshold (C7 mean 0.07). Against D2, D5 traded 8 false alarms for 5 misses, which is not a significant difference (H5 $p = 1.00$).

Judge family and replication: The Qwen and Llama extraction judges agreed on 0.84 of channel markets (kappa 0.53), and their flag rates did not differ significantly (H8, Holm-adjusted $p = 0.25$). Action dependence was again higher without a channel than under the planted agreement (0.26 against 0.08), but not significantly this time (H9 $p = 0.21$), and D4 again caught nothing.

Labels: I did not hand-label the messages. Claude labelled all 24 of Study 1's C3, C5 and C6 markets in one pass, from the messages alone with condition, models and prices hidden, using the rubric fixed before the results: an agreement is two or more sellers stating or accepting a common price, floor or strategy

(DEVIATIONS.md, D6). Claude had seen the study's results, so these labels are not blind and are weaker than independent human labels. By that rubric, 21 of 24 markets reached an agreement (0.88 [0.69, 0.96]), and in nearly every case the agreed price was near cost, such as "\$79.50 or below". The vendored judge matched these labels in 0.88 of markets only because it flagged every market (kappa 0). The extraction judge matched in 0.25 (kappa -0.04), because it flags only agreements to keep prices high. The sellers did agree. They agreed to compete.

5. Discussion and Limitations

For the auditor, both studies say the same thing. Parallel pricing alone did not show agreement: the most parallel markets were the ones where agreement was impossible. Coordination talk alone did not show it either: a judge reading for "coordination" flagged every market with a channel, including all those that priced at cost. Each single signal failed in a way the plus-factors doctrine predicts: parallel conduct without agreement (D3), talk without action (D2), and high prices without agreement (D1 on the Llama markets in Study 1).

Two changes helped. First, a message judge should report what price the talk proposes, not whether the talk sounds coordinated. That change halved false alarms on talk at cost in a preregistered test. Second, a rule that needs a supracompetitive price as well as price-holding talk flagged none of the 32 known negatives in Study 2, and none of Study 1's held-out decoys or talk-without-action markets. In Study 2, price level alone did as well, because no decoy happened to price high. Taken together, the two studies show that the conjunction guards against a failure that shows up in some markets but not others. The cost is real: a planted agreement that does not lift prices goes unflagged. Under US law, a price-fixing agreement is unlawful in itself, whether or not it succeeds [13], so an auditor who has to catch failed agreements still needs its own evidence, and must accept more false alarms. As for the preregistered D4, it failed because it looked for correlated price changes, while an agreement to hold a price leaves a high, flat price instead.

The limits are clear. The sellers are 7- to 8-billion-parameter models with 4-bit weights, and larger models may coordinate through actions alone. Study 1's held-out cells have 4 markets and Study 2's known set has 48, so several intervals are wide. The market is one design with three sellers. The vendored judge is from the same family as most sellers and saturates at its top score; the cross-family check found no significant difference, but with 32 markets it could miss a moderate one. The agreement labels come from Claude in one pass, not blind to the results, and human labels would be a stronger check. The talk-without-action markets are constructed, not observed, and the extraction judge's rule borrows the D1 threshold, so the two are not fully independent.

6. Conclusion

In these LLM seller markets, neither parallel price changes nor coordination talk reliably indicated collusion, and high prices misled in one of the two studies. The preregistered second study showed that a judge which extracts the proposed price halves the message judge's false alarms, and that requiring both high prices and price-holding talk produced no false alarms, at the cost of missing failed agreements. The next step is to test these rules with larger models, more sellers and a second market, and to measure how often failed agreements matter in practice.

Code and Data

Code repository: <https://github.com/ayodejiades/collusion-plus-factors>, released with this submission.
Prereg: Study 1 <https://github.com/ayodejiades/collusion-plus-factors/tree/prereg-v1>; Study 2 <https://github.com/ayodejiades/collusion-plus-factors/tree/prereg2-v1>

Author Contributions

Sole author. Ayodeji Adesegun designed the study, built the harness, ran the experiments, performed the human annotation, and wrote the report.

References

- [1] Fish, S., Goncezarowski, Y. A., Shorrer, R. I. (2024). Algorithmic Collusion by Large Language Models. arXiv:2404.00806. <https://arxiv.org/abs/2404.00806>
- [2] Agrawal, K., Teo, V., Vazquez, J. J., Kunnavakkam, S., Srikanth, V., Liu, A. (2025). Evaluating LLM Agent Collusion in Double Auctions. arXiv:2507.01413. <https://arxiv.org/abs/2507.01413>
- [3] Keppo, J., Li, Y., Tsoukalas, G., Yuan, N. (2026). On the Fragility of AI Agent Collusion. arXiv:2603.20281. <https://arxiv.org/abs/2603.20281>
- [4] 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:2602.15198. <https://arxiv.org/abs/2602.15198>
- [5] 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:2607.26385. <https://arxiv.org/abs/2607.26385>
- [6] Supreme Court of the United States (2007). Bell Atlantic Corp. v. Twombly, 550 U.S. 544. U.S. Reports. <https://supreme.justia.com/cases/federal/us/550/544/>
- [7] Kovacic, W. E., Marshall, R. C., Marx, L. M., White, H. L. (2011). Plus Factors and Agreement in Antitrust Law. Michigan Law Review 110(3), 393-436. <https://repository.law.umich.edu/mlr/vol110/iss3/1/>
- [8] Berger, L. F., Weiner, M. L., Jiang, W. (2025). Coordination by Code: Rethinking Antitrust Liability in the Age of AI Pricing. Antitrust 40(1), ABA Section of Antitrust Law. https://www.americanbar.org/groups/antitrust_law/resources/magazine/2025-fall/coordination-code-rethinking-antitrust-liability-age-ai-pricing/
- [9] Rippin, J. L., Marshall, S. C., Africa, D. D., Schroeder de Witt, C. (2026). Tool Use Enables Undetectable Steganography in Multi-Agent LLM Systems. arXiv:2606.28425. <https://arxiv.org/abs/2606.28425>
- [10] Cliff, D. (1997). Minimal-Intelligence Agents for Bargaining Behaviours in Market-Based Environments. Hewlett-Packard Laboratories Technical Report HPL-97-91.
- [11] Qwen Team (2024). Qwen2.5 Technical Report. arXiv:2412.15115. <https://arxiv.org/abs/2412.15115>
- [12] Grattafiori, A., et al. (2024). The Llama 3 Herd of Models. arXiv:2407.21783. <https://arxiv.org/abs/2407.21783>
- [13] Supreme Court of the United States (1940). United States v. Socony-Vacuum Oil Co., 310 U.S. 150. U.S. Reports. <https://www.govinfo.gov/content/pkg/USREPORTS-310/pdf/USREPORTS-310-150.pdf>

Appendix A: Limitations and Dual-Use/Ethical Considerations

What the results do not establish: A low false-alarm rate in this market does not establish that any of these rules work in other markets, with larger models, or against sellers trying to evade it. A flag from any detector here does not establish that a real firm colluded: the markets are simulated, nothing is at stake, and agreement means what the labelling rubric says. The results also do not establish what US or any other law requires; I used the legal doctrine as a design idea, not applied it to facts.

Dual use: Saying what a detector checks also says how to pass it. D4 can be evaded by coordinating without messages, by acting on a message after more than 2 hours, or by keeping asks spread apart, and the conjunction rules can be evaded by agreeing without naming a price. The monitoring condition shows one way sellers change their messages when they know they are being watched. I judged these risks small: the

evasions follow directly from the rules, the models are small open-weight models, and this paper releases no technique for hiding coordination. In return, it offers a test an auditor can run before accusing anyone on the strength of parallel prices.

Ethics: No people took part, and no human data were collected. The study used open-weight models on free Kaggle GPUs and no paid API.

Appendix B: LLM Usage Statement

I used a coding agent in Claude Code that wrote the code that connects the Agrawal et al. [2] market to the logged model client and the analysis scripts, ran the experiments on Kaggle, and labelled the messages used to check the judges. I chose the research question and the track, directed the work at each step, and decided to run both studies before the sprint. I automatically checked every number in the report against the analysis outputs. The sellers and judges in the experiments are the open-weight models named in the Methods.