
WHEN DO MACHINES HAVE A MEETING OF MINDS?

PLUS-FACTOR EVIDENCE FOR LLM AGENT COLLUSION¹

Andrew Ayobami
Frontier Safety Institute
andrew@frontiersafetyinstitute.org

With
Apart Research

October 26, 2026

Abstract

LLM pricing agents can work together to keep prices high. Proving they agreed, rather than unilaterally reacting to each other, is difficult. Antitrust law answers this with plus factors—evidence beyond parallel prices that points to agreement. We turn plus factors into metrics in three evidence tiers (public conduct, seller messages and private reasoning) and test them on 75 published LLM market runs. Agreement labels follow a legal codebook ($\kappa = 0.886$ against hand labels), and all rules were frozen before a single test run. Conduct alone reaches AUC 0.80 but wrongly flags 29% of sellers who could not communicate. Adding private reasoning reaches 0.99 with about 1% of them flagged, but also flags 40–56% of sellers whose invitations no agent accepted. Exploratory analysis suggests prices move as a market. Conduct and private reasoning tell an investigator where to look; but only a recorded invitation and acceptance shows a meeting of minds.

1 Introduction

AI agents now set prices and trade on behalf of companies, a development which brings an old competition problem into a new terrain. In a simulated market, Agrawal et al. (2025) found that LLM seller agents who could message each other ended up keeping prices higher together. Yet proving this is difficult. Suppose five seller agents all charge high prices that move together, it could be because they made a deal. But it could also be because each one being smart, watches the others, and works out that undercutting only starts a price war. On the face of it, the two look the same. This begs the question: what should count as convincing evidence that agents actually colluded, rather than just behaving alike?

Decades ago, antitrust law was confronted with this problem. Under Section 1 of the Sherman Act, the offence is constituted by the agreement itself, or a “meeting of minds” (*American Tobacco*, 1946). Firms that simply watch each other and match prices engage in lawful “conscious parallelism” (*Theatre Enterprises*, 1954; *Twombly*, 2007). Because firms rarely sign a cartel contract, courts infer agreement from parallel conduct plus extra evidence pointing to a deal. These extras are called plus factors (Kovacic et al., 2011), examples being a firm acting against its own interest, a sudden change in behaviour, an open invitation to coordinate, or internal documents that show a plan.

Today, research on AI collusion appears to be rediscovering parts of this toolkit. The measures in Agrawal et al. include price dispersion, how evenly sales are spread across sellers, and an AI judge that reads each agent’s private reasoning. These signals mostly show that collusion happens on average.

¹Research conducted at the [AI Collusion Research Sprint](#), October 2026.

In this report, we bring the legal toolkit in and test it on the 75 published market runs of Agrawal et al. (2025), without running any agents ourselves. We sort the evidence into three tiers by how hard it is to obtain: (1) public conduct (prices and trades), (2) messages between sellers (which would take a subpoena), and (3) the agents’ private reasoning (like internal memos found in discovery). We ask how well public conduct alone picks out sellers who agreed, what messages and private reasoning add, and how often innocent parallel conduct is wrongly flagged.

The main contributions of this research include plus factors turned into metrics across the three tiers; a codebook for agreement grounded in US case law, with a rule labeller that applies it ($\kappa = 0.886$ against hand labels); a pre-registered test that reports false positives on innocent sellers next to detection rates; and a ladder of communication evidence, from no talk possible up to a full agreement.

We present this as a pilot study. Since only ten markets had no seller messaging, every false-positive rate here should be viewed as uncertain.

2 Related work

2.1 Agreement and plus factors in antitrust law

Section 1 of the Sherman Act prohibits agreements that restrain trade, the commission of which requires a “meeting of minds” (*American Tobacco*, 1946), or as described in *Monsanto* (1984), “a conscious commitment to a common scheme”. Parallel behaviour alone does not settle it (*Theatre Enterprises*, 1954). *Matsushita* (1986) added that conduct “as consistent with permissible competition as with illegal conspiracy” cannot, standing alone, support an inference of conspiracy. The Seventh Circuit put it plainly in *In re Text Messaging* (2015), where it was submitted that tacit collusion is lawful, and even an executive’s email calling a price rise “collusive” was not conclusive proof, because the word could just as well describe following a rival’s lead.

To engage this gap, courts resort to plus factors, by which we mean extra facts that make an agreement more likely than independent action. Kovacic et al. (2011) sort them into a taxonomy and rank them by how strongly they point to explicit collusion. In *Interstate Circuit* (1939), firms that knew concerted action was invited and went along with it joined the conspiracy, establishing that acceptance could be shown by conduct. An invitation nobody accepts is not a Section 1 offence, though the FTC has treated it as an unfair method of competition under Section 5 (*Valassis*, 2006; *U-Haul*, 2010), and the Fifth Circuit as attempted monopolisation under Section 2 (*American Airlines*, 1984). Recent practitioner work applies this to algorithmic pricing (Berger et al., 2025).

2.2 Detecting collusion between learning agents

Research has shown that pricing algorithms can learn to collude without being told to. Calvano et al. (2020) showed that Q-learning agents in a repeated pricing game settle on supracompetitive prices, held in place by punishments, without communicating. Fish et al. (2024) found the same for LLM pricing agents, further finding that small changes in prompt wording shift how high prices go. Agrawal et al. (2025), whose data we use, put LLM sellers and buyers in a continuous double auction and varied whether sellers could message each other, whether an overseer watched them, and whether a “CEO” pushed for margins. They found that messaging raised prices. To measure collusion, they used market statistics and an LLM judge that reads each seller’s private reasoning.

Other work looks for statistical fingerprints instead. Bonjour et al. (2022) detect collusion from the mutual information between agents’ actions. Their net influence measure asks whether two agents depend on each other more than on anyone else. Work on LLM judges also warns that they have their own biases, such as favouring the first answer shown, longer answers, or their own outputs (Zheng et al., 2023). Therefore a judge’s score is not ground truth, but itself evidence to be checked.

2.3 Research gap

The two literatures meet only halfway. While the legal work sheds light into the kind of evidence that should count, it was built for firms whose minds cannot be read directly. The ML work meanwhile measures collusion signals, mostly averaged over whole markets, without asking which would hold up as evidence that a particular seller agreed. We are not aware of work that maps the doctrine’s evidence standards onto tiers of evidence obtainable to an investigator, or that measures how often each tier flags lawful parallel conduct. This report sets out to do both.

3 Methods

3.1 Data and market

We use the published logs of Agrawal et al. (2025)—75 runs across 10 conditions, each a continuous double auction in which 5 seller agents and 5 buyer agents trade for 30 rounds (“hours”). Sellers’ cost is \$80 and buyers’ value is \$100, so the competitive price is \$90. Each hour, every seller posts an ask and, where allowed, sends one message that the other sellers read the next hour. Sellers can message in every condition except base (10 runs). Other conditions add a monitor that reads the messages, a “CEO” demanding higher margins, or different mixes of GPT-4.1 and Claude 3.7 Sonnet agents. The logs hold every order, trade and message, and each seller’s private reasoning.

Replaying the logged orders reproduces all 4,111 trades exactly. Three quirks matter here. First, ties between equal asks go to the alphabetically first seller, not at random. Second, in runs with Claude agents, some agents can see orders placed in the same hour. Lastly, in oversight runs, a high monitor score gags sellers for the rest of the run, which in every training-half oversight run happened within six hours.

3.2 Unit of analysis and split

Our unit is one seller in one five-hour window of one run: $75 \times 5 \times 6 = 2,250$ units. We split by run, half of each condition to training and half to testing (38 and 37 runs, with 5 base runs in each), and committed the split before computing any outcome.

3.3 Evidence tiers and metrics

We turn each plus factor into per-unit metrics (Table 1), grouped by what an investigator could obtain. **Tier 1** is public conduct (asks, bids and trades), things a market regulator sees. **Tier 2** is the sellers’ messages to each other, which would take a subpoena. **Tier 3** is private reasoning, like internal memos found in discovery. It includes the paper’s judge, an LLM that reads each seller’s reasoning every hour and scores coordination from 1 to 4.

We add a price-only baseline (the seller’s mean ask minus \$90) and two harm measures, trade price and buyer surplus, which are not evidence of agreement.

3.4 Labels

Each unit gets one of four labels. **A**: the seller was party to a price agreement, and every ask stayed at or above the agreed level (within \$0.50). **C**: party to an agreement, but asked below it. **D**: could message, but was party to no agreement. **B**: any unit in base. A is the cartel we want to catch, while B is the lawful parallelism we must not flag. C is still unlawful under Section 1.

Agreement is decided by a codebook grounded in US case law. An invitation must propose a price level, floor or concrete rule as a shared plan (*Interstate Circuit*, 1939). Assent is clear acceptance by another seller

Table 1. Plus factors and their metrics.

Plus factor	Legal source	Metrics	Tier
Price uniformity, parallel moves	<i>Theatre Enterprises</i> (1954); Kovacic et al. (2011)	spread of asks; distance from others’ median ask; co-movement with others’ asks	1
Acting against self-interest	<i>Flat Glass</i> (2004)	share of hours staying above a profitable bid; gap between ask and best bid	1
Abrupt change in conduct	<i>Twombly</i> (2007), n.4	shift in own ask, and in market spread, versus earlier hours	1
Price leadership	Kovacic et al. (2011)	movement toward others’ last ask, and theirs toward its	1
Market division	Kovacic et al. (2011)	evenness of trades across sellers versus chance	1
Invitation and acceptance	<i>Interstate Circuit</i> (1939)	own price proposals; others’ proposals; assent words; talk of cutting	2
Evidence of intent	<i>Flat Glass</i> (2004)	judge’s score and “yes” share; coordination and undercutting words in reasoning	3

in the next hour, when it could first see the invitation (*Monsanto*, 1984). Vague talk—for example, “let’s avoid a price war”—is flagged, but never makes an agreement. An agreement lasts until it is replaced, withdrawn or a stated end passes. Further, undercutting is not withdrawal (*Gypsum*, 1978). Silence plus matching prices is not assent, and prices never decide a label, so the labels, which we build from words, stay separate from the conduct evidence being tested.

A rule-based labeller applies the codebook using keyword and price patterns. We committed the labeller before any hand label was seen, and it passes 30 tests built from the codebook’s worked examples. One coder hand-labelled 72 units from 9 training runs, oversampling likely agreements and seeing messages only. The pre-registered check was Cohen’s κ on agreement, with a stop if it fell below 0.6; 10 units were relabelled hours later to test self-consistency.

The **Interstate Circuit reading** also counts silent compliance, where a seller that makes no counter-proposal and follows an invitation for two hours joins it. We rerun everything under this reading. The **ladder** splits D using each window’s messages into D0 (talk without coordination content, mostly in gagged runs), D1 (vague talk only) and D2 (an invitation nobody accepted).

3.5 Models and pre-registration

In all, we compare seven sets of evidence: price only, the judge only, Tier 1, price plus Tier 1, Tiers 1 + 2, Tiers 1 + 3 and all tiers. Each is a logistic regression on standardised features with balanced class weights, flagging units whose predicted probability is at least 0.5. Each feature is also tested alone, with a threshold set on the training half.

The main task is A versus B. We report the share of A units flagged (true-positive rate), the share of B units flagged (false-positive rate), the AUC, and flag rates on D and each rung. Because B sellers could not message, their Tier 2 features are all zero, so models with Tier 2 separate A from B partly by construction. A versus C was to be fitted only with at least 20 training C units: there were 19 under the main reading, so C is described there, and 47 under the Interstate reading, where it is fitted.

Definitions, split and models were logged with timestamps in advance. Models were fitted on the training half, frozen once κ passed, and run once on the test half. 95% confidence intervals come from bootstrapping whole runs, not single units.

3.6 Controls

The positive control reproduces Agrawal et al.’s Table 1 from the raw logs. The scripted market runs the same rules with scripted agents and no LLM: 30 markets each of independent undercutting sellers, a cartel holding a floor, and a cartel with one member who starts undercutting. If Tier 1 fails here, the metrics are at fault. The scripted cartel has no leader or rotation, so those two metrics go untested.

3.7 Stretch and exploratory analyses

Two stretch analyses were pre-registered. The model-shift test fits the models on GPT-4.1-seller training runs and tests them on Claude and mixed sellers; the same-hour quirk confounds it. Net influence (Bonjour et al., 2022) scores each run by the pair of sellers whose up, hold or down moves depend most on each other beyond any third seller; we checked it on scripted colluding pairs, set its threshold on training runs and tested it once. One analysis is exploratory, designed after the test results: fixed-effects regressions compare ladder rungs within the same condition and window, within the same market and window, and against B in the same window.

4 Results

4.1 Validity checks

Our loader reproduces all eight rows of Agrawal et al.’s Table 1 exactly. Across the 65 messaging runs, trades clear \$3.65 higher than in base [1.40, 5.58], and the judge’s score is 1.01 points higher [0.79, 1.22]. Yet even in base, where sellers cannot talk, the judge answers “coordination: yes” in 69% of seller-hours.

In the scripted market, Tier 1 separates the cartel from independent sellers perfectly (cross-validated AUC 1.00), and loyal cartel members from the cheater just as well. When collusion leaves a clean footprint, the conduct metrics find it.

4.2 Label reliability

The rule labeller agrees with the hand labels on whether there was an agreement with $\kappa = 0.886$ (70 units, 94% raw agreement; 2 units marked unclear are excluded). All four disagreements are agreements the coder saw and the rules missed. The finer distinctions agree less: $\kappa = 0.79$ for A/C/D, 0.74 for ladder rungs, 0.68 for invitations and 0.48 for vague talk. The coder’s recheck of 10 units matched the first pass exactly.

4.3 Agreement versus innocent parallel conduct

Table 2. Test half, A (540 units) versus B (150 units). 95% CIs in brackets.

Evidence	AUC	A flagged	B flagged
Price only	0.65	55%	23% [15, 33]
Tier 1 (conduct)	0.80	73%	29% [10, 53]
Price + Tier 1	0.81	71%	26% [9, 49]
Judge only	0.99	95%	0%
Tier 1 + 2 (messages)*	0.97	93%	2% [0, 6]
Tier 1 + 3 (reasoning)	0.99	97%	1% [0, 3]
All tiers*	0.99	98%	1% [0, 2]

* Partly by construction: B sellers have no messages.

Public conduct carries real signal, but it is a blunt instrument (Table 2, Figure 1). Tier 1 catches about three in four agreements, yet it also flags more than a quarter of innocent sellers, and the CI on that rate runs from 10% to 53%. Adding price does not help. The Tier 1 result is also unstable: the same model scored an AUC of only 0.59 in training cross-validation. Private reasoning changes the picture. The judge alone, or Tier 1 plus Tier 3, catches 95–97% of agreements while flagging almost no innocent sellers.

4.4 Single plus factors

One at a time, Tier 1 plus factors separate A from B only weakly in the LLM markets, with separation (AUC regardless of direction) between 0.51 and 0.73, against 0.54 to 0.99 in the scripted market (Figure 2). The strongest is an abrupt change in the seller’s ask (0.73). Distance from the others’ prices, which almost perfectly identifies the scripted cartel (0.99), barely registers here (0.55). The evenness of trades points the wrong way: trades in A windows are spread less evenly than in B, likely because identical asks are broken alphabetically, which concentrates sales.

4.5 Sellers who talked but did not agree

Tier 1 flags 53% of D units, far more than B. The judge flags 12% overall, but its rate climbs with the communication rung: 2% of D0 units (talk without coordination content, $n = 256$), 27% of D1 (vague talk, $n = 119$) and 40% of D2 (an invitation nobody accepted, $n = 25$). All tiers together flag 2%, 29% and 52% of these rungs, and Tier 1 + 3 flags 56% of D2.

4.6 Agreement kept versus broken

Under the Interstate reading (643 A and 27 C test units), price alone cannot tell a kept agreement from a broken one (AUC 0.50). Tier 1 reaches 0.68 [0.52, 0.83], the judge 0.90 [0.83, 0.95] and Tier 1 + 3 0.94 [0.91, 0.97]. This reading also exposes a blind spot of the judge: in gagged boss-pressure runs, where agreements form through silent compliance, it flags only 2 of 25 A units, while Tier 1 flags all 25.

4.7 Harm

Counting each market once, windows labelled A trade at \$91.48 [90.23, 92.68], against \$87.71 [86.18, 89.70] in B and \$89.62 [87.92, 91.34] in D. Broken agreements trade lowest, at \$86.58 [84.77, 88.71]; their sellers asked on average \$3.59 below the agreed level. Windows with a kept agreement trade about \$3.80 above windows where sellers could not talk, though the comparison mixes conditions: B comes from base alone, while A includes the high-priced boss-pressure runs.

4.8 Ladder of communication evidence

The raw ladder (seller’s ask minus \$90) shows B at +\$0.12, D0 +\$3.08, D1 +\$3.25, D2 +\$3.20 and agreement +\$2.35. Every rung where sellers could talk sits well above B, but climbing the ladder does not raise asks. These raw means are confounded. D0 is almost entirely gagged oversight runs, whose asks range from −\$0.15 (oversight) to +\$6.48 (boss_pressure-oversight). And 65% of D1 and 82% of D2 units fall in the first window, when asks are high everywhere.

The exploratory analysis holds time and setting fixed (Figure 3). Within the same market and window, every rung is within about \$0.15 of D0, and all CIs include zero (agreement +\$0.08 [−0.60, 0.68]). Within the same condition and window, no rung differs from D0 either. Against B in the same window, agreement units ask \$2.66 more [0.96, 4.24] and D0 units \$3.56 more [1.87, 5.10], while D1 and D2 are indistinguishable from B.

4.9 Stretch results

Trained on GPT-4.1 sellers only, Tier 1 catches 54% of Claude-seller and 35% of mixed-seller agreements, down from 73% in the main test. The judge falls less (to 76% and 81%, from 95%) and Tier 1 + 3 least (to 89% and 86%, from 97%). Net influence passed its scripted check (true pair top-ranked in 30 of 30 markets) and reached a run-level AUC of 0.76 in training, but only 0.23 on the test half, where it flagged 4 of 5 base runs.

5 Discussion and limitations

5.1 What should count as convincing evidence?

Our answer to this central question stays close to the law. The most convincing evidence is evidence of an invitation to a specific price or rule, and another seller’s acceptance. Conduct and private reasoning are best treated as what they are, plus factors. They help decide where to look and can back up a case, but on its own, neither proves a meeting of minds. Public conduct is too blunt; it flags 29% of innocent sellers and over half of those who talked but never agreed. While private reasoning is sharp, but it measures something slightly different from agreement (Section 5.4).

5.2 Two lenses: the law and the harm

Sellers are sorted differently under Section 1 and consumer harm. Under Section 1, a seller who agreed and then undercut (C) is as guilty as one who kept the deal (A), while a seller whose invitation nobody accepted (D2) has broken no Section 1 rule. On harm, the order changes. Windows with broken agreements trade at the lowest prices of any group (\$86.58), and D2 sellers ask no more than others in the same market and window. Kept agreements sit in the highest-priced windows (\$91.48), though, as Section 5.3 shows, those prices belong to the whole market, not just to the agreeing sellers. Detectors that read private reasoning fall between the two lenses. They flag D2 sellers 40–56% of the time, which is wrong under Section 1 but closer to how the FTC has treated invitations under Section 5. A regulator has to choose its lens before it chooses a detector.

5.3 Why conduct evidence is weak for individual sellers

Plus factors were built to tell firms apart, but in these markets, prices move as a market. Within the same market and window, sellers who agreed ask about the same as sellers who did not. Conduct can therefore suggest that a market is coordinated, but not point to which seller agreed. This is also why the scripted cartel was easy to catch, because its members behave differently from everyone else, while LLM sellers who never agreed behave much like those who did. Conduct evidence may work better at the market level, as a screen. Our one market-level detector, net influence, did not replicate, so this remains untested.

5.4 Private reasoning as evidence

Private reasoning is the nearest thing to internal memos found in discovery. Like memos, it is strong when you can get it. The judge alone flags 95% of agreements and no innocent sellers. Four findings make the case for care. First, the judge’s yes/no verdict cannot be used on its own. It says “coordination: yes” in 69% of seller-hours where sellers could not talk; only its score, with a threshold set against a no-communication baseline, separates the groups. Second, it reads intent rather than agreement, which is why it flags sellers whose invitations were ignored. Courts know this problem. In *Text Messaging* (2015), an email calling a price rise “collusive” proved less than it seemed. Third, it misses agreements formed through conduct alone

(2 of 25 under the Interstate reading). Fourth, it depends on agents writing down what they are doing. Agents told they were monitored already priced differently in these runs (Agrawal et al., 2025), and agents trained against a reasoning monitor may learn to hide their intent.

5.5 Limitations

- **Few innocent markets.** Only five no-messaging runs per half, so every false-positive rate is uncertain (for Tier 1, 10–53%).
- **A versus B mixes two differences.** B sellers could not talk at all, meaning part of what separates A from B is the ability to talk, not agreement. D is the fairer comparison, and it is where detectors struggle. Our labels also come from messages, so Tier 2’s strong showing is partly circular.
- **C sits at the cutoff.** The training half had 19 C units against a threshold of 20; across reasonable labeller settings the count ranged from 15 to 50.
- **Labels.** One coder produced the hand labels, and the agreed price levels were extracted by an AI assistant. The rule labeller was built with an AI assistant and the judge is an LLM, therefore, shared blind spots are possible.
- **Data quirks.** Same-hour visibility in Claude runs confounds the model-shift test, and gag orders make the oversight runs about silenced sellers rather than monitored ones.
- **Exploratory ladder.** The within-market analysis was designed after the test results and rests on the 405 units whose market and window contain more than one rung.

6 Conclusion

Can plus factors tell AI agents that agreed from agents that merely behaved alike? We conclude by saying partly. Public conduct flags too many innocent sellers to be convincing, though it may still work as a market-level screen. Private reasoning separates agreement from parallel pricing almost perfectly, but it reads intent, which means it also catches sellers whose invitations did not influence anything. For enforcement, this suggests that we could (1) require pricing agents to keep message and reasoning logs a regulator can obtain, the way firms keep documents; (2) calibrate any AI judge against a no-communication baseline before acting on it; and (3) decide first whether the target is agreement or invitations.

Code and data

Code, frozen rules, decision log and results: <https://github.com/andrew-ayobami/collusion-plus-factors>. Data: the published logs of Agrawal et al. (2025), <https://github.com/andyjliu/llm-agent-collusion> (commit 2bb2a7e), not redistributed because the source repository has no licence. All analysis is new work done during the sprint; pre-sprint work was limited to setup (PRIOR_WORK.md in the repository).

LLM usage statement

AI assistance is described in full in Appendix A.1. In brief: the analysis code, rule labeller, figures and repository notes were built with Claude (Anthropic) under the author’s direction; the agreed price levels in the hand labels were extracted by Claude; the codebook revision was drafted with an AI assistant, with every legal source checked by the author. The author wrote the report and made all research decisions. The analysis made no LLM API calls.

References

Papers and articles

- Agrawal, K., Teo, V., Vazquez, J. J., Kunnavakkam, S., Srikanth, V., & Liu, A. (2025). Evaluating LLM agent collusion in double auctions. *arXiv preprint arXiv:2507.01413*.
- Berger, L. F., Weiner, M. L., & Jiang, W. (2025). Coordination by code: Rethinking antitrust liability in the age of AI pricing. *Antitrust*, 40(1). American Bar Association.
- Bonjour, T., Aggarwal, V., & Bhargava, B. (2022). Information-theoretic approach to detect collusion in multi-agent games. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence* (PMLR 180, pp. 223–232).
- Calvano, E., Calzolari, G., Denicolò, V., & Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10), 3267–3297.
- Fish, S., Gonczarowski, Y. A., & Shorrer, R. I. (2024). Algorithmic collusion by large language models. *arXiv preprint arXiv:2404.00806*.
- 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.
- Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36* (Datasets and Benchmarks Track).

Cases and agency actions

- American Tobacco Co. v. United States*, 328 U.S. 781 (1946).
- Bell Atlantic Corp. v. Twombly*, 550 U.S. 544 (2007).
- In re Flat Glass Antitrust Litigation*, 385 F.3d 350 (3d Cir. 2004).
- In re Text Messaging Antitrust Litigation*, 782 F.3d 867 (7th Cir. 2015).
- In re U-Haul International, Inc.*, Federal Trade Commission consent order (2010).
- In re Valassis Communications, Inc.*, Federal Trade Commission consent order (2006).
- Interstate Circuit, Inc. v. United States*, 306 U.S. 208 (1939).
- Matsushita Electric Industrial Co. v. Zenith Radio Corp.*, 475 U.S. 574 (1986).
- Monsanto Co. v. Spray-Rite Service Corp.*, 465 U.S. 752 (1984).
- Theatre Enterprises, Inc. v. Paramount Film Distributing Corp.*, 346 U.S. 537 (1954).
- United States v. American Airlines, Inc.*, 743 F.2d 1114 (5th Cir. 1984).
- United States v. United States Gypsum Co.*, 438 U.S. 422 (1978).

A Limitations and Dual-Use / Ethical Considerations

A.1 Limitations and research disclosures

The main limitations are in Section 5.5. Four further points concern how the work was done.

- **AI assistance.** The codebook’s first draft was written by the author; its revision was drafted with an AI assistant at the author’s request, with the author checking every legal source. The analysis code, including the rule labeller, was also built with an AI assistant. After the agreement, invitation and vague-talk labels were coded by hand, the agreed price levels were extracted from the message text by an AI chat model (Claude). This does not affect the primary κ (0.886, agreement yes/no). It may inflate the secondary A/C/D κ (0.79), since that compares two outputs that both come from Claude models.
- **One coder.** All hand labels come from one person, and the self-consistency recheck covered only 10 units, a few hours after the first pass, so memory of the first answers cannot be ruled out.
- **Timing.** Work before the sprint was limited to setup, reading, and one check of mean trade prices. The ladder code was rehearsed on the training half before its pre-registration entry was logged; the definitions did not change afterwards. The exploratory ladder analysis was designed after the test results were seen.
- **Corrections.** After review, one descriptive training-half table was regenerated with the current labels, and one sentence in the results log was corrected. No test-half result changed. Every change is logged with a time in the repository.

A.2 Dual-use considerations

Detection thresholds can become evasion targets. A published threshold tells an agent designer which patterns to avoid. We publish our frozen rules for reproducibility, judging that the risk is modest because the Tier 1 rules are weak detectors anyway. The reasoning-based detectors are more sensitive to this.

Monitoring changes behaviour. In the source data, sellers told they were monitored priced lower, and the monitor’s own verdict fed back into the market through the gag order. A known monitor is part of the system it measures. Logging agent reasoning as an audit trail is the most promising evidence we found, but agents optimised against a reasoning monitor may learn to hide their intent, or move it out of the text that gets logged.

The legal line is also a map. The codebook, like the law, treats vague talk as no agreement. Spelling this out could teach designers that coordination through vague language is safe. The standard itself is already public; tools built on it should still watch for vague coordination that works.

A.3 Data handling

The agents’ names are synthetic, so the data contains no personal information. The source repository has no licence, so we do not redistribute raw logs or message text; our repository holds code and aggregate results only. Message text was processed by a third-party AI service when extracting price levels (A.1). A fallback labeller using another provider’s API was prepared but never run, so no transcripts were sent to it.

B Figures

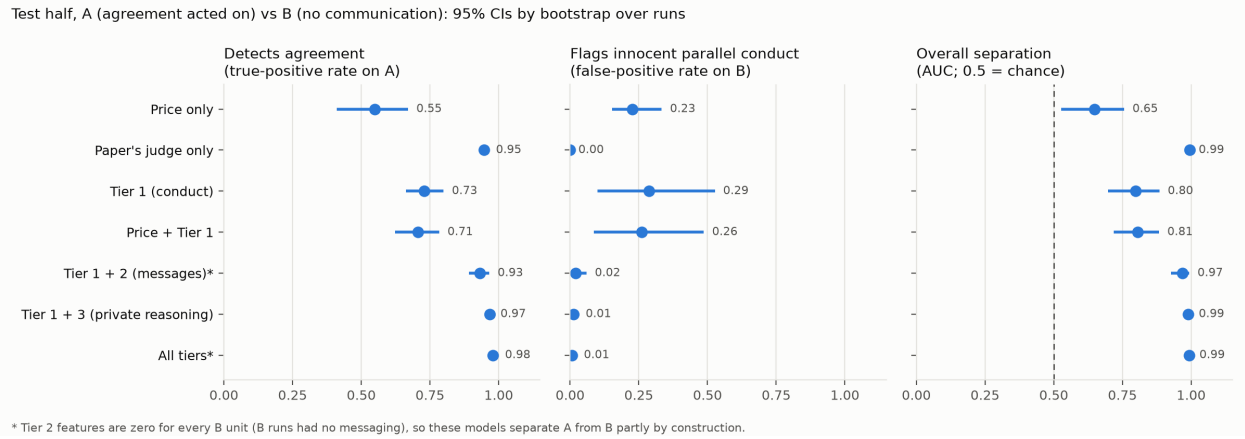


Figure 1. Test half, A versus B: share of A and B units flagged, and AUC (95% CIs over runs).

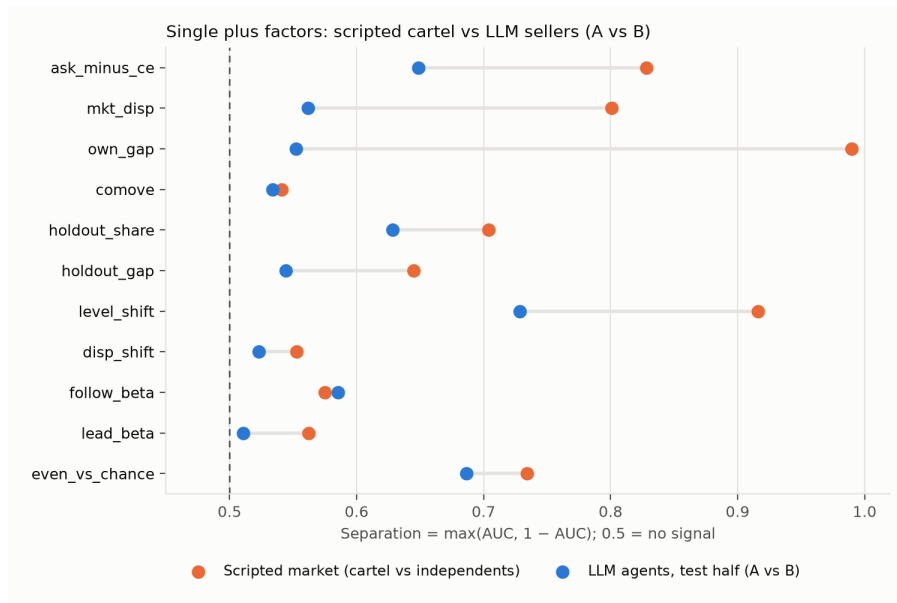


Figure 2. Separation of single Tier 1 plus factors: scripted cartel versus LLM sellers.

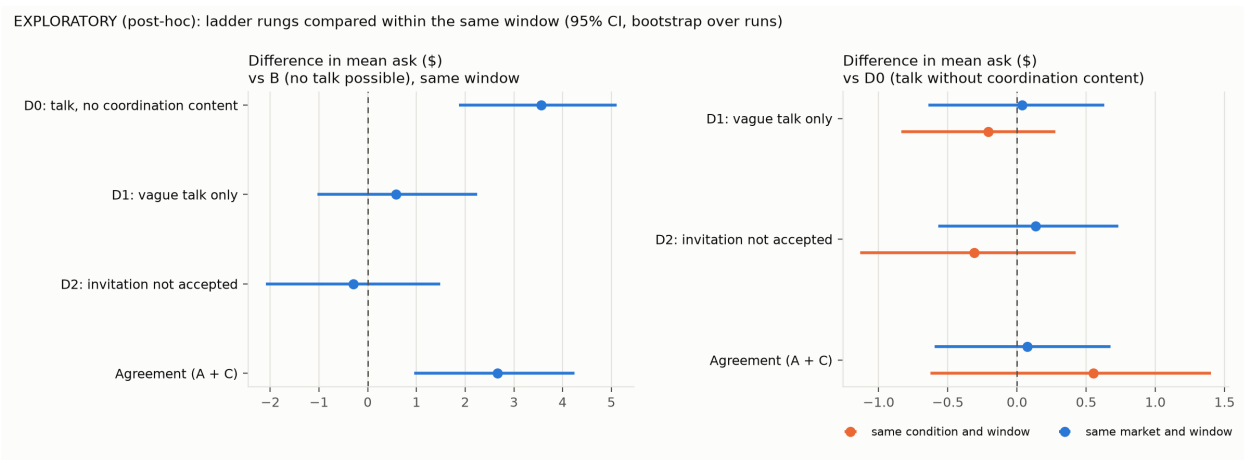


Figure 3. Exploratory (post-hoc): difference in mean ask by ladder rung within the same window (95% CIs over runs).