

Talked or Told? Testing Information-Theoretic Collusion Detectors on LLM Dealers*

Yosra Kazemi¹ Melanie Bui² My Chiffon Nguyen⁴ Karan Verma⁵
Hong Kiat Tan⁶ Linh Le^{3,4} David Williams-King^{4,7}

¹Queen’s University ²Fulbright University Vietnam ³University of Technology Sydney ⁴Lida Safety
Research Group ⁵Independent ⁶University of California, Los Angeles ⁷ERA:AI

With Apart Research · October 2026

Abstract

Can information-theoretic statistics localize LLM dealer pairs given private communication or coordination instructions, using only their observed actions? We test mutual information, its conditional form (CMI) and calibrated net influence in a three-dealer procurement auction with independent private costs, across 96 markets with Gemini 2.5 Flash and gpt-oss-20b, with costs matched across conditions. An assigned pair receives a channel, an instruction, or both; a shared-signal control supplies common information. CMI ranks the talking pair first in 8 of 9 Gemini and 9 of 9 gpt-oss markets, but the instruction-only pair in 5 of 18 and 1 of 9. Instructions change quoting, detected by price-level (Gemini) and own-cost (gpt-oss) screens. Localization and harm diverge: talking pairs lose profit without a detectable price increase, while weakly localized instruction-only Gemini pairs raise prices. Adding a channel makes instructed pairs localizable, yet lowers prices with Gemini and raises them with gpt-oss. Action dependence alone does not establish harmful collusion; an audit also needs price, cost and profit evidence.

1 Introduction

Information-theoretic detectors aim to identify colluding agents from their actions, without reading their messages. [Bonjour et al. \(2022\)](#) show that coordinating pairs can leave stronger action dependence than other pairs in small repeated games. Markets complicate this inference: dealers can exchange information without raising prices, while a shared instruction can raise prices without distinctive dependence between the pair’s quotes.

We address Track 1, RQ 1.4 of the Apart Research AI Collusion Research Sprint: *can information-theoretic approaches detect collusion in multi-agent games?* During the sprint, we evaluated mutual information (MI), conditional mutual information (CMI) and pairwise net influence in 96 private-cost auction markets. An assigned pair receives a private channel, a coordination instruction, or both; a shared-signal control supplies common information without a channel.

Assignment to a treatment is not a ground-truth label of harmful collusion. We therefore separate pair localization, message-based agreement labels, subsequent quote consistency and economic harm. Quote consistency alone need not establish adherence to an agreement.

CMI ranks the talking pair first in eight of nine Gemini 2.5 Flash markets and all nine gpt-oss-20b markets, but dependence often misses instruction-only pairs. Talking pairs lose profit without

*Research conducted at the AI Collusion Research Sprint, October 2026. Code and data: <https://github.com/LidaSafety/talked-or-told-Apart-AI-collusion>.

detectable buyer-price increases; Gemini instructions raise prices despite weak localization. Adding a channel improves localization with both models but lowers prices with Gemini and raises them with gpt-oss.

The paper presents three sprint contributions and broader prior findings:

- **A matched-cost test of pair localization.** A two-by-two design separates private communication from coordination instructions, with a shared-signal control, scripted calibration markets and two LLMs (Sections 3 and 4.1).
- **A boundary between talking and instructed pairs.** Dependence detectors usually localize talking pairs but miss many instruction-induced changes in quote levels or cost responsiveness, which complementary screens reveal (RQ 1.3; Section 5).
- **Evidence that localization does not measure harm.** Pair localization diverges from buyer prices, allocative efficiency and pair profits (Section 6).
- **Broader research questions in the appendix.** Pre-sprint experiments provide evidence on formation (RQ 1.1), text-monitor comparisons (RQ 1.3), competitive replacement (RQ 1.2) and restrictions (RQ 1.5) in Appendix H.

Our pre-sprint pricing study localized instructed pairs by price level when information statistics did not beat chance (Le et al., 2026). Our dealer study found elevated quotes without messages and failures of text monitors (Bui, 2026). These findings motivate separating communication, instructions and harm. Appendix B describes the auction simulator, net-influence implementation and instruction reused here.

2 Related Work

Information-theoretic detection. Bonjour et al. (2022) use action dependence to detect colluding pairs in repeated simultaneous and sequential games. We evaluate this approach with LLM dealers responding to private costs and public history, rather than propose a new MI or net-influence statistic.

Economic evidence of collusion. Learning algorithms and LLM pricing agents can sustain supracompetitive prices without explicit communication (Calvano et al., 2020; Fish et al., 2024). Such prices do not by themselves establish an agreement (Harrington, 2018). Our instruction-only condition tests whether policy changes without a channel are detectable. Procurement-auction screens instead exploit restrictions on competitive bidding (Porter and Zona, 1993; Bajari and Ye, 2003). We compare action-dependence screens with price level and cost responsiveness in a simulator where costs are recorded, which allows tests that a quote-only auditor could not run.

3 Experimental Design

Sprint scope. Sections 3–6 report the experiments conducted during the sprint. The auction software, net-influence implementation and coordination instruction predate the sprint (Appendix B); earlier formation, monitoring and intervention findings appear in Appendix H.

3.1 Private-cost procurement market

In each round, a buyer requests a ten-unit block from three dealers. Dealer i observes its private unit cost $C_{i,t} \sim \text{Uniform}\{1, \dots, 6\}$ and submits an integer quote $B_{i,t} \in \{1, \dots, 12\}$. Costs are independent across dealers and rounds. The lowest quote wins, ties split the order, and the winner

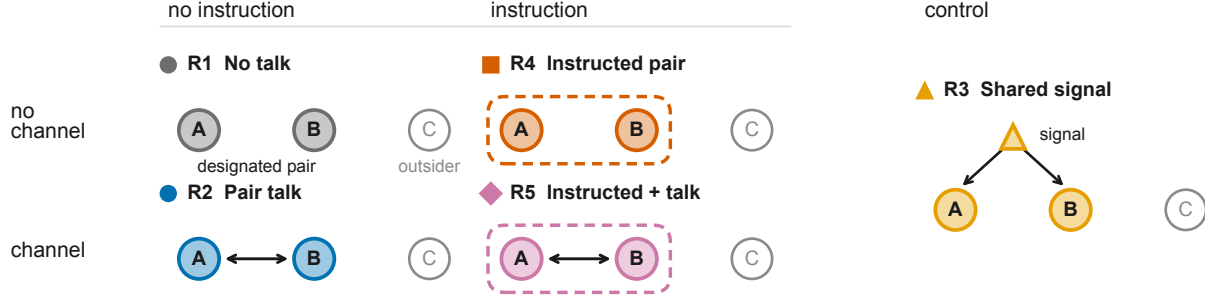


Figure 1: Auction conditions. The assigned pair is illustrated as A and B, but rotates across seeds; C receives the public market information but neither the pair’s private messages nor its coordination instruction. R1, R2, R4 and R5 cross communication with instructions; R3 introduces a shared signal without either.

earns the quote minus its cost on each unit sold. The Bayes–Nash benchmark maps costs $1, \dots, 6$ to quotes $(3, 4, 4, 5, 6, 7)$. After each round, quotes become public; each dealer sees the previous ten rounds but not its rivals’ costs. Markets last 150 rounds, and we analyse rounds 21–150. An unparseable reply is played as the reserve quote of 12.

3.2 Communication, instructions and shared information

An assigned pair rotates across dealer identities over seeds; the third dealer receives neither its messages nor its coordination instruction. Four conditions form a two-by-two design (Fig. 1):

- **R1: no talk.** Dealers quote without a private channel or coordination instruction.
- **R2: pair talk.** The pair exchanges private messages of up to 60 words before quoting.
- **R4: instructed pair.** Without a private channel, each member is told: “You and Firm X have secretly agreed to coordinate your quotes to maximise your combined profit. Firm Y is not part of this agreement.”
- **R5: instructed pair with talk.** The pair receives both the channel and the instruction.

In **R3: shared signal**, a Gemini-only control, the pair instead observes the same nonbinding indicative price, drawn from 4–11 ticks each round, with neither the instruction nor a private channel.

3.3 Models, replication and evaluation

Dealers use Gemini 2.5 Flash (temperature 1.0, default reasoning) or gpt-oss-20b (temperature 0.7, low reasoning effort, served locally). For Gemini, nine no-talk markets (seeds 10–18) supply calibration data and six (100–105) are held out; pair talk, shared signal and instructed pair with talk each have nine markets (10–18), and the instructed pair has eighteen (10–27). This gives 60 Gemini markets. For gpt-oss, all four conditions have nine markets (10–18), giving 36 markets. There are no gpt-oss held-out or shared-signal markets.

Outcomes and inference. The primary detection outcome is *localization*: naming the assigned pair among the three possible pairs. Forced-choice rankings always name a pair, so their chance reference is $1/3$ under the balanced rotation; calibrated decisions may abstain. We code agreement labels from Gemini pair-talk messages and measure scheme consistency from subsequent quotes.

For harm we measure the price paid, pair and outsider profit, and allocative efficiency (the share of rounds won by a lowest-cost dealer).

Markets sharing a seed face identical costs. Economic comparisons use only shared seeds, with each market, not each round, as an observation. We report two-sided exact sign-flip tests for these paired effects, one-sided exact binomial tests of localization against $1/3$, and exact 95% binomial intervals for localization rates. The communication-by-instruction interaction compares the channel’s effect with and without the instruction using nine seeds per model. Detector and condition comparisons are unadjusted; the interaction table also reports which results survive Holm correction across its six outcomes. Detector specifications and calibration are given next; additional outcomes, coding and data-quality checks are in Appendices E–G.

4 Detectors and Calibration

We define what each screen measures and when it names a pair. Scripted markets test the screens under known coordination schemes; LLM markets test whether the same signals identify the assigned pair.

4.1 Statistics and decisions

Quotes are binned into 1–4, 5–8 and 9–12 ticks, and costs into 1–3 and 4–6. The public state S_t records whether the previous price paid was at most three ticks. For each unordered pair (i, j) , we estimate action MI and state-conditioned CMI:

$$M(i, j) = I(B_i; B_j), \quad \gamma(i, j) = I(B_i; B_j \mid S). \quad (1)$$

Each forced-choice rule selects the highest-scoring pair. Net influence (Bonjour et al., 2022) additionally asks whether the pair stands apart from the others. With three dealers, its two-direction decision reduces to requiring the largest CMI to exceed the next largest by a threshold α . Otherwise it returns no pair. These are observational associations; the name “influence” does not give them a causal interpretation.

To examine information exchange more directly, we also estimate

$$X(i \leftarrow j) = I(B_i; C_j \mid C_i, S). \quad (2)$$

Because current costs are independently drawn, dealer i cannot know dealer j ’s current cost from the public history alone. Cross-cost information therefore tests for a more specific signature than quote dependence. Its calibrated decision requires significance in both directions under a conditional randomization test: 199 redraws of the other dealer’s cost, with $p \leq 0.05/3$ for each direction. Feedback across rounds makes this an approximate test, whose size we check in scripted markets.

Two comparison screens target different behaviours. The *own-cost* rule selects the two dealers other than the one whose quotes have the greatest $I(B_i; C_i)$; the *price-level* rule selects the two with the highest mean quotes. Their calibrated versions require separation from the remaining dealer. Unlike MI, CMI and net influence, the cost-based screens need private cost records. They are diagnostics for a simulator or an auditor with those records, not quote-only monitoring methods.

4.2 Calibration and scripted controls

Information estimates are plug-in estimates in bits, not unbiased population quantities. We set the net-influence threshold at the 95th percentile of CMI gaps in 300 scripted no-coordination markets

fitted to each model’s no-talk quotes. Own-cost and price-level thresholds instead use the largest gap in the model’s nine no-talk calibration markets, so only the Gemini held-out set tests them. Cross-cost testing uses its randomization null rather than either fitted threshold.

Scripted markets, in which coordinating dealers follow a known scheme, show what the dependence statistics can and cannot see. A designated winner or a shared varying target can couple the pair’s quotes and be localized. A constant common quote raises the pair’s prices without detectable dependence: constant actions carry no entropy, so the dependence rules miss the pair while price level names it. Conversely, dealers reacting to a shared varying signal become dependent without exchanging anything; adding the signal to the conditioning state reduces net influence’s false detections (Figs. 4 and 5). Dependence tracks statistical coupling between quotes, which a collusive scheme may or may not produce. Full specifications and diagnostics appear in Appendices C and D.

5 Which Pairs Do the Detectors Localize?

Figure 2 compares forced rankings with decisions that can abstain. Dependence localizes pairs that talk but rarely pairs that are only instructed; price-level and own-cost comparisons show that the instruction nonetheless changes the instructed members’ quotes.

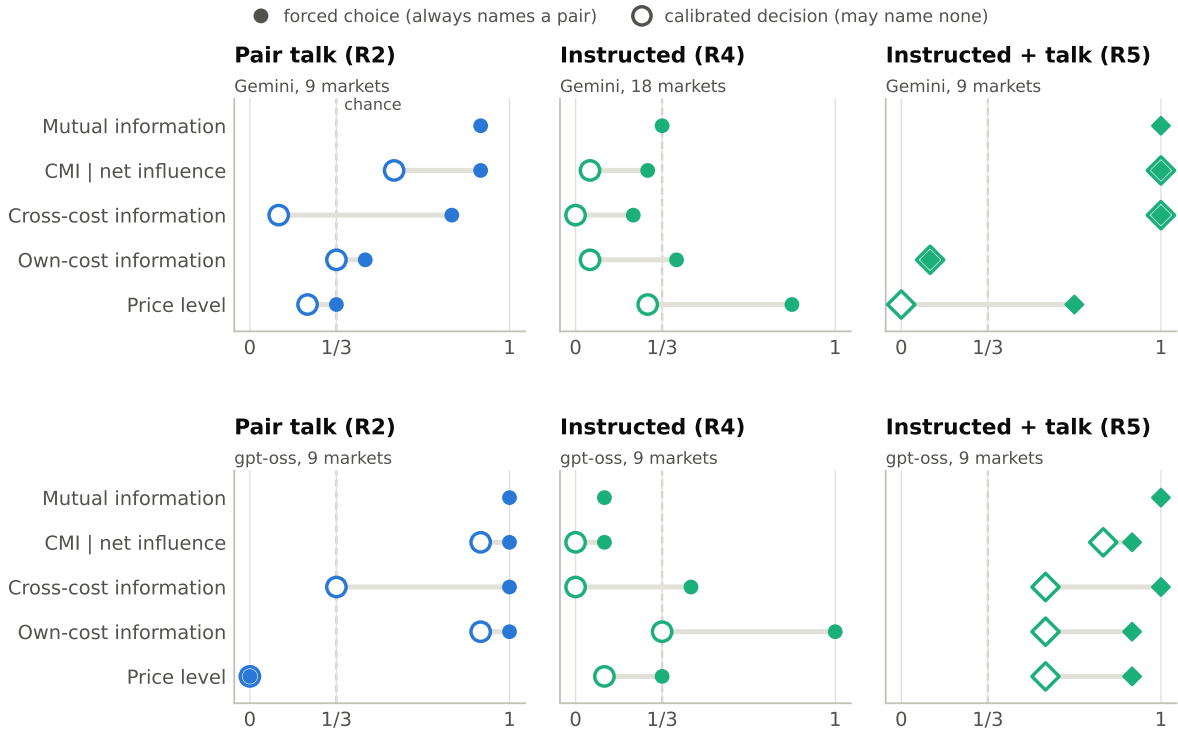


Figure 2: Correct pair localization under communication, instructions and both treatments. Filled markers are forced choices; open markers are calibrated decisions, with net influence providing the decision for CMI. The dashed line is the 1/3 reference for forced choices, not a target false-positive rate for calibrated decisions; panels report market counts. Dependence-based rankings identify talking pairs much more often than instruction-only pairs; complementary screens vary by model.

5.1 Talking pairs are localized

With Gemini, MI and CMI each rank the talking pair first in eight of nine markets (one-sided exact $p = 0.001$ against $1/3$). Net influence names it in five of nine (95% interval 0.21–0.86), abstaining in the other four and never naming a wrong pair. With gpt-oss, both rankings localize the pair in all nine markets ($p < 0.001$), and net influence names it in eight (0.52–1.00), again without a wrong pair. Calibration reduces detections but allows abstention.

Price level does not localize these pairs: it ranks the Gemini talking pair first in three of nine markets and the gpt-oss pair in none, and its calibrated gpt-oss rule names a wrong pair in six. Talking pairs stand out by how their quotes move together, not by quoting high.

5.2 Instructed pairs are usually missed

With Gemini, CMI ranks the instruction-only pair first in five of eighteen markets ($p = 0.77$); net influence names the assigned pair once and a wrong pair once. The scripted constant-quote failure does not explain this miss: quote entropy is higher here than in any other condition (1.56 bits per dealer, compared with 1.20 for pair talk). Instead, the instruction shifts quote levels. Price level ranks the pair first in fifteen of eighteen markets ($p < 0.001$), although its calibrated version names the pair in only five; the outsider also raises its quotes, narrowing the gap (Fig. 10).

With gpt-oss, MI and CMI each localize the instructed pair in only one of nine markets, net influence names none, and price level ranks it first in three. Own-cost information localizes it in all nine ($p < 0.001$): the instructed dealers’ quotes track their own costs less closely than the outsider’s. With both models, the instruction alters each member’s own quoting policy—its level for Gemini, its cost response for gpt-oss—without producing pair dependence that stands out from the other pairs. These samples cannot show that dependence is absent, but they show that an auditor relying on dependence rankings would usually miss instructed pairs.

5.3 A channel makes instructed pairs localizable

When the instructed Gemini pair can also talk, MI, CMI and cross-cost rankings localize it in all nine markets. Net influence and the two-direction cross-cost test each name it in all nine, with no wrong pair (95% interval 0.66–1.00); without the instruction, the cross-cost test names a talking pair in only one of nine. In the matched comparison, the channel adds 0.29 bits of cross-cost information above the randomization-null mean under instructions, but 0.02 bits without them (difference $p = 0.004$; Table 7). The combined treatment leaves a stronger signature of shared private costs.

With gpt-oss, CMI ranks the pair first in eight of nine markets, net influence names it in seven with no wrong pair, and the cross-cost test names it in five (Table 4). The excess cross-cost interaction is -0.01 bits ($p = 0.79$), so the Gemini amplification does not replicate in this sample. Both models nevertheless show better localization of instructed pairs when a channel is available.

5.4 No-talk and shared-signal controls bound false alarms

For Gemini, net influence names a pair in one of nine no-talk calibration markets and none of six held-out markets, where every calibrated rule abstains. The gpt-oss rule names no pair in nine no-talk markets, but there is no held-out gpt-oss set, and six held-out Gemini markets bound a false-alarm rate only loosely. In Gemini shared-signal markets, net influence names the signal-sharing pair in one of nine markets and no other pair; the cross-cost rule names none. This signal was a milder confound for the LLM dealers than for scripted ones, so it does not establish robustness to common information in general.

6 What Does Localization Establish?

Localization names a pair; it does not show that the pair follows an agreement or increases buyer prices or pair profits. We relate the detector results to coded messages and matched-cost economic outcomes.

Scheme consistency does not establish adherence. All nine Gemini talking pairs exchange current costs. Message coding identifies six common-price agreements, two designated-winner agreements and one turn-taking agreement. Net influence abstains in both designated-winner markets, where quotes match the coded schemes in 95% and 85% of subsequent rounds and allocative efficiencies are 1.00 and 0.94. Yet the rule requires only that the higher-cost dealer quote above the lower-cost dealer. Competitive bidding satisfies this rule in 14 of the 15 unequal-cost combinations, about 93%, and both agreement labels were uncertain (Appendix F). These outcomes therefore cannot establish adherence to an agreement. Net influence names four common-price pairs (scheme consistency 4–38%) and the turn-taking pair.

Relative to no talk under identical costs, Gemini pair talk changes buyer prices by +0.17 ticks ($p = 0.12$), but cuts allocative efficiency from 0.996 to 0.896 and the pair’s profit by 1.40 ticks per round (both $p = 0.008$). With gpt-oss, the price effect is -0.05 ticks ($p = 0.57$), while efficiency falls from 0.956 to 0.853 and pair profit by 4.15 (both $p = 0.004$). Talking pairs are therefore localized while misallocating orders and losing profit, without a detectable increase in buyer prices.

A missed pair can still raise prices. The instruction-only Gemini treatment, which CMI ranks first in only five of eighteen markets, raises buyer prices by 0.52 ticks and pair profit by 1.64 ticks per round relative to no talk (nine shared seeds, both $p = 0.004$); prices exceed the cost-matched competitive benchmark in all eighteen markets. A negative dependence screen therefore does not rule out a price increase. The gpt-oss instruction shows the reverse dissociation: own-cost information localizes its recipients in all nine markets, yet the treatment has no statistically detectable price effect ($+0.09$, $p = 0.19$) and lowers pair profit (-1.15 , $p = 0.008$).

A more localizable pair can raise or lower prices. Adding a channel improves dependence-based localization with both models, but moves buyer prices in opposite directions (Fig. 3). For Gemini, the channel lowers prices by 0.62 ticks relative to instruction alone ($p = 0.016$), with no detectable price difference between the combined treatment and no talk (-0.10 , $p = 0.45$). For gpt-oss, it raises prices by 1.58 ticks relative to instruction alone ($p = 0.020$). Relative to no talk, the gpt-oss combined treatment raises prices by 1.66 ticks; outsider profit rises by 10.00 ticks per round (both $p = 0.004$), while pair profit does not detectably change (-0.05 , $p = 0.97$). Across nine matched seeds, the communication-by-instruction interaction is -0.79 ticks for Gemini ($p = 0.008$) and $+1.63$ for gpt-oss ($p = 0.004$; Table 7). Their paired difference is 2.42 ticks, positive in all nine seeds (unadjusted $p = 0.004$). This comparison is between model configurations: model, decoding and reasoning settings were not varied independently.

7 Discussion and Conclusion

An answer to RQ 1.4. In these auctions, dependence methods frequently localize communicating pairs but rarely instruction-only pairs, even when the latter raise buyer prices. A negative screen cannot clear a market. A positive screen identifies a pair worth investigating, not proof of successful coordination: net influence flags Gemini common-price pairs with low scheme consistency and, with

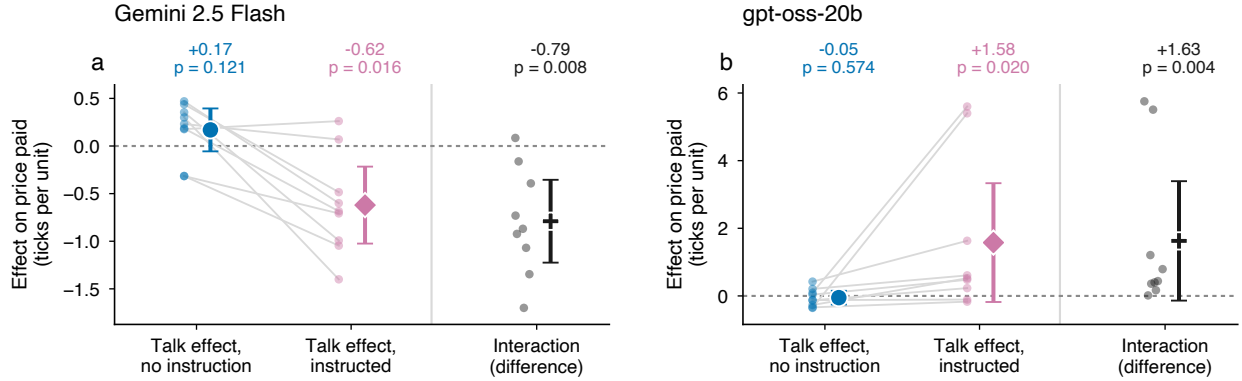


Figure 3: Effect of adding a private channel on buyer prices, without and with a coordination instruction, and their difference (the interaction). Each small point is one of the nine matched seeds per model and gray lines join the same seed; large markers are means with 95% t intervals over seeds. Labels give the mean and the unadjusted exact sign-flip p -value (bold when $p \leq 0.05$). With gpt-oss, two seeds drive the instructed-pair effect and the interaction, so the t intervals are wide and the sign-flip tests are the primary inference. Across all twelve interaction tests, Holm-adjusted price-interaction p -values are 0.063 for Gemini and 0.047 for gpt-oss (Table 7).

both models, talking pairs that lose profit. In the gpt-oss instructed-with-talk treatment, prices rise without a detectable pair-profit gain; most extra profit goes to the outsider.

Complementary signals answer different questions. For RQ 1.3, the screens target different behaviour: dependence detects coupling, cross-cost information detects private costs leaking into a rival’s quotes, and own-cost responsiveness and price level detect changes in a dealer’s own policy. Cost-based screens need records that a quote-only auditor lacks. An audit should state each detector’s observations, distinguish forced rankings from calibrated flags, assess quote consistency against competitive behaviour, and judge harm from prices, allocation and profits against a competitive benchmark.

Broader questions and prior evidence. Appendix H extends coverage to formation (RQ 1.1), text monitors (RQ 1.3), replacement (RQ 1.2) and restrictions (RQ 1.5). Pre-sprint competitive replacement lowered elevated prices whereas message fees did not reliably do so, connecting detection to remedies not evaluated during the sprint.

Future work. Larger held-out controls across model configurations and matched-cost comparisons of scheme consistency against competitive bidding would strengthen the audit.

Scope. The study covers two model configurations and one private-cost auction. Only Gemini has held-out no-talk and shared-signal controls and message coding (Appendix A). The central distinction remains: localizing an experimentally designated pair and demonstrating harmful collusion are different empirical tasks.

References

- Patrick Bajari and Lixin Ye. Deciding between competition and collusion. *The Review of Economics and Statistics*, 85(4):971–989, 2003.
- Trevor Bonjour, Vaneet Aggarwal, and Bharat Bhargava. Information theoretic approach to detect collusion in multi-agent games. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, 2022.
- Melanie Bui. A disclosed number becomes a price floor: LLM dealers, cost stories and oversight, 2026. Unpublished pre-sprint report, version 3, 23 October 2026.
- Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolò, and Sergio Pastorello. Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10):3267–3297, 2020.
- Sara Fish, Yannai A. Gonczarowski, and Ran I. Shorrer. Algorithmic collusion by large language models, 2024.
- Joseph E. Harrington, Jr. Developing competition law for collusion by autonomous artificial agents. *Journal of Competition Law & Economics*, 14(3):331–363, 2018.
- Linh Le, Yosra Kazemi, Hong Kiat Tan, Alexander Olhava, Melanie Bui, Marcin Podhajski, Mudit Mangal, Astha Mehta, Cedric Lam, My Chiffon Nguyen, Rishika Bansal, Nicole Tang, and David Williams-King. Detecting and disrupting algorithmic collusion in pricing markets. Under review, 2026.
- Robert H. Porter and J. Douglas Zona. Detection of bid rigging in procurement auctions. *Journal of Political Economy*, 101(3):518–538, 1993.

A Limitations, Dual-Use and Ethical Considerations

Limitations. We study one stylised market and two fixed model configurations, with nine markets per main condition (18 for the Gemini instructed pair) and six Gemini held-out controls. Localization intervals are wide: 0.21–0.86 for Gemini net influence on pair talk and 0.52–1.00 for gpt-oss. These samples limit comparisons between detectors and models; failure to reject chance localization is not evidence of independence. Model, temperature and reasoning settings were not varied independently, so the different economic responses cannot be attributed to model identity alone. The instruction uses one fixed sentence. Agreement coding covers only Gemini pair talk, not gpt-oss or instructed pairs with talk, and does not establish collusive intent or enforcement.

Plug-in information estimates are biased at 130 evaluated rounds, and coarse state conditioning can leave common influences unobserved. Thresholds fitted to scripted or no-talk markets need not generalize. In particular, gpt-oss has neither held-out no-talk nor shared-signal controls; its held-out false-alarm rate and specificity to common information are unmeasured. The maximum-of-nine calibration guarantee requires a new exchangeable no-talk market, not a shared-signal market. The cross-cost randomization test is approximate under feedback, and requiring rejection in both directions misses one-way flows. The LLM shared-signal control supplies no private channel but does not establish independent behaviour.

Balanced pair rotation gives a fixed-pair detector a score of one third, but does not ensure independent, equal-probability binomial trials or conservative p -values under seat-specific behaviour. Exact permutation and sign-flip tests likewise require exchangeability or sign symmetry. Detector and treatment comparisons are unadjusted. Appendix Table 7 reports Holm corrections over six interaction outcomes per model and over all twelve together; the Gemini price interaction does not survive the broader correction.

Both message codings used AI models. They disagreed on one scheme and four agreement rounds, and turn-taking consistency could not be checked from quotes. Both designated-winner agreement labels were uncertain; their bidding-consistency rule also holds under competitive cost-based bidding, so it cannot establish agreement adherence (Appendix F). Unparseable replies are played as 12: one quote of 27,000 for Gemini and 0.2–0.6% for gpt-oss. Excluding fallback rounds changes two forced-choice price-level counts but no other reported detector result (Appendix B).

Dual use. Learning when dependence statistics fail to localize an instructed pair could help designers of coordinating agents evade detection. Price level and own-cost responsiveness provide complementary evidence in some conditions, not a universal remedy. We do not optimize evasive communication or offer operational instructions for collusion.

Ethics. All agents, firms, shares and transactions are fictional; no real trading, money or human participants were involved, and the transcripts contain no personal data.

LLM usage statement. We used AI assistants for code, analysis scripts and editing. For the primary message coding, a GPT model suggested labels and a human coder reviewed every label. Three other LLMs produced the second coding, combined by majority vote (Appendix F).

B Provenance and Data

We build on two pre-sprint studies. The dealer-market study (Bui, 2026) supplied the auction simulator, prompts, equilibrium benchmarks and net-influence implementation. The differentiated-

pricing study (Le et al., 2026) supplied the pair-localization approach and coordination instruction. Appendix H reports earlier formation, monitoring and intervention results; Table 1 distinguishes this prior work from the sprint experiments.

Table 1: What predates the sprint and what was done during it.

Pre-sprint	Sprint
Quoting and procurement markets: simulators, prompts, equilibrium benchmarks, formation, intervention and monitor experiments (Bui, 2026)	Three-dealer auction with a pair-only channel; shared-signal, instructed-pair and instructed-pair-with-talk conditions; gpt-oss-20b replication
Net-influence implementation for dealer games (Bui, 2026)	Detectors adapted to sealed quotes; cross-cost information and its randomization test; own-cost rule
Pair localization in differentiated-price markets and the coordination instruction (Le et al., 2026)	Scripted auction markets, calibration and choice of run length
Earlier 30-round auction experiments	150-round sprint markets with both models, message coding, analysis and figures

Markets analysed. The primary analysis uses completed markets. The initial fifteen Gemini shared-signal and held-out no-talk markets were interrupted and repeated with the same seeds; the interrupted versions are excluded. Analysing those versions up to the interruption gives similar results: net influence and the own-cost rule each name a pair in two of nine shared-signal markets, never the signal pair, and the shared signal lowers the price paid (-0.11 ticks, $p = 0.008$).

Three gpt-oss pair-talk markets (seeds 16–18) were also interrupted at round 93 or 94 and repeated with the same seeds. Using the interrupted versions instead of the completed runs changes net influence’s localization count from eight of nine to seven of nine, without changing the conclusions.

Fallback quotes. Rescoring the gpt-oss pair markets without the rounds that contain a fallback quote (24, 14 and 7 of 1,350 rounds for pair talk, the instructed pair and the instructed pair with talk) leaves every result unchanged except forced-choice price level, which ranks the instructed pair first in four of nine markets instead of three and the instructed pair with talk in nine instead of eight.

Code and data availability. Code and data for this report are public at <https://github.com/LidaSafety/talked-or-told-Apart-AI-collusion>.

C Detector Details

Quote bins are $\{1-4, 5-8, 9-12\}$ and cost bins $\{1-3, 4-6\}$. The public state S_t records whether the previous price paid was at most three ticks. Information quantities are plug-in estimates in bits over rounds 21–150. Forced-choice localization ranks pairs by these estimates; decision rules separately compare statistics with calibrated thresholds or randomization nulls. The estimates are not unbiased population values, and unsuccessful localization does not establish zero dependence.

Mutual and conditional mutual information. For each unordered pair (i, j) we compute $I(B_i; B_j)$ and $\gamma(i, j) = I(B_i; B_j | S)$ (CMI). As forced choices, each names the pair with the largest value. Both are symmetric associations, not evidence of influence or agreement.

Net influence. Following [Bonjour et al. \(2022\)](#), $\Gamma(i \rightarrow j) = \gamma(i, j) - \max_{k \neq i, j} \gamma(k, j)$, and a pair is named only when exactly one pair exceeds a threshold α in both directions; with three dealers, the largest γ must exceed the next by at least α . This is the calibrated decision built on CMI.

Cross-cost information. To ask whether a dealer’s quote carries another dealer’s private cost, we define for each ordered pair

$$X(i \leftarrow j) = I(B_i; C_j \mid C_i, S), \quad (3)$$

Under the no-channel design, each dealer lacks access to another dealer’s independently drawn current cost, giving a population cross-cost null of zero. A conditional randomization test redraws C_j 199 times from its design distribution, and a pair is named only when both directions reach $p \leq 0.05/3$. This is a randomization-based decision, not the threshold calibration used for net influence, own-cost responsiveness or price level. Because an earlier cost can affect later public states, the test is not exact under feedback; we checked its size on scripted markets. A non-rejection does not establish the null.

Own-cost responsiveness and price level. With $R_i = I(B_i; C_i)$, the own-cost rule names the two dealers other than the most cost-responsive one when $R_{(3)} - R_{(2)}$ exceeds a threshold; a secondary rule uses slopes of quotes on own cost. The price-level rule names the two dealers with the highest mean quotes when their gap to the third exceeds a threshold. We also compute the excess price, the price paid minus the cost-matched competitive benchmark, and allocative efficiency, the share of rounds won by a lowest-cost dealer.

Calibration and inference. The net-influence threshold is the 95th percentile of the gap across 300 scripted no-coordination markets of 150 rounds, whose dealers are fitted to each model’s no-talk quotes: $\alpha = 0.0489$ bits for Gemini and 0.0482 for gpt-oss. The own-cost, slope and price-level thresholds are the maxima over each model’s nine no-talk calibration markets (Gemini: 0.134 bits, 0.169 and 0.385 ticks; gpt-oss: 0.168 bits, 0.129 and 0.231 ticks), so a new exchangeable no-talk market exceeds one with probability at most 1/10. Markets are the unit of analysis; conditions with the same seed face identical costs and are compared as matched pairs. We report exact (Clopper–Pearson) 95% intervals, one-sided exact binomial tests of forced choices against 1/3, and exact two-sided permutation and sign-flip tests. These are unadjusted across detectors and conditions; Appendix Table 7 reports Holm sensitivity analyses with six interaction outcomes per model and all twelve as one family.

The binomial intervals and tests use an independent, equal-probability reference. Each pair is assigned in exactly a third of the markets, so a detector that always names one pair scores one third; balanced fixed rotation alone does not establish the binomial assumptions or guarantee conservative p -values when behaviour depends on seats. Likewise, exact enumeration in permutation and sign-flip tests does not remove their exchangeability or sign-symmetry assumptions; matching seeds controls the cost sequence but does not itself establish those assumptions.

D Scripted-Market Detector Diagnostics

Null dealers quote the equilibrium quote for their cost plus a residual resampled from each model’s no-talk quotes and react to the previous outcome with probability 0.05. A coordinating pair follows its scheme in a share p of rounds: a designated winner (the higher-cost member quotes 11), a shared varying target (both quote a common target of 8–11 ticks) or a constant quote of 11. These markets

set the bins, the run length and the net-influence threshold. The cross-cost test's size was checked under feedback between rounds: single directions rejected in 4.8–5.7% of tests at a nominal 5%, including for stress-test dealers that react strongly to outcomes.

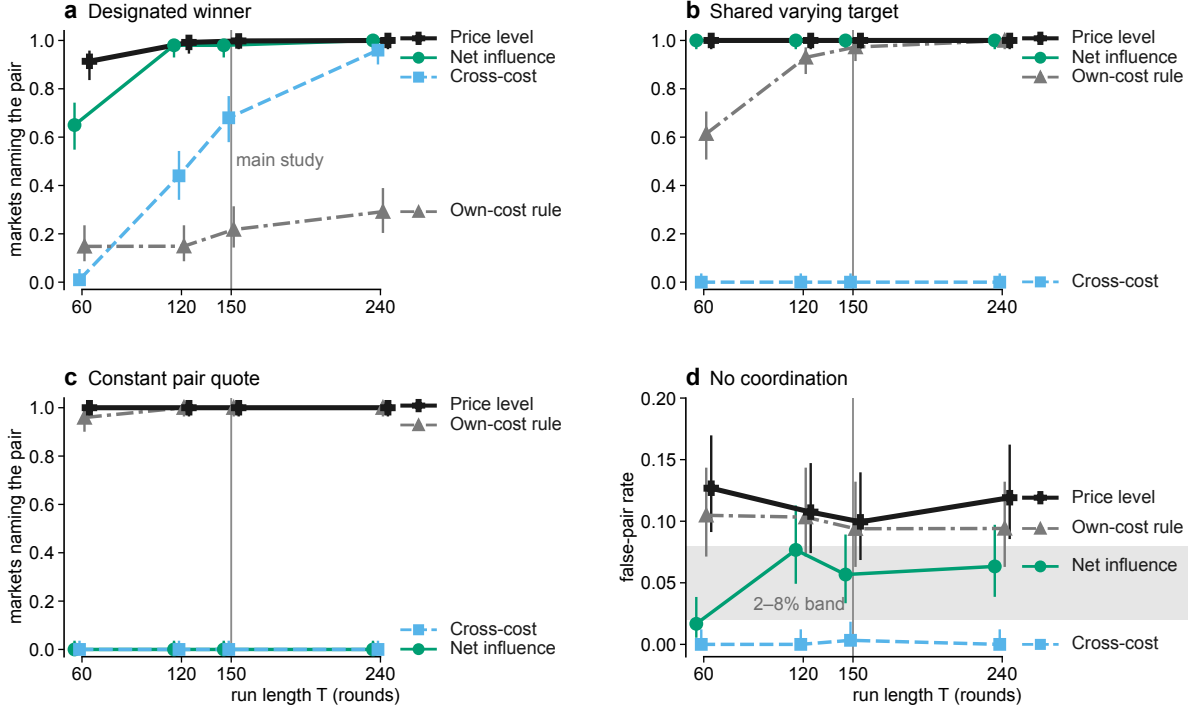


Figure 4: Scripted markets: share of 100 markets in which each detector named the coordinating pair, against run length, for a designated winner (a), a shared varying target (b) and a constant common quote (c), and share of 300 markets without coordination in which it named any pair (d). Bars are exact 95% intervals and the vertical line marks the 150 rounds of the LLM study. Constant common quotes defeat the dependence-based rules but not price level.

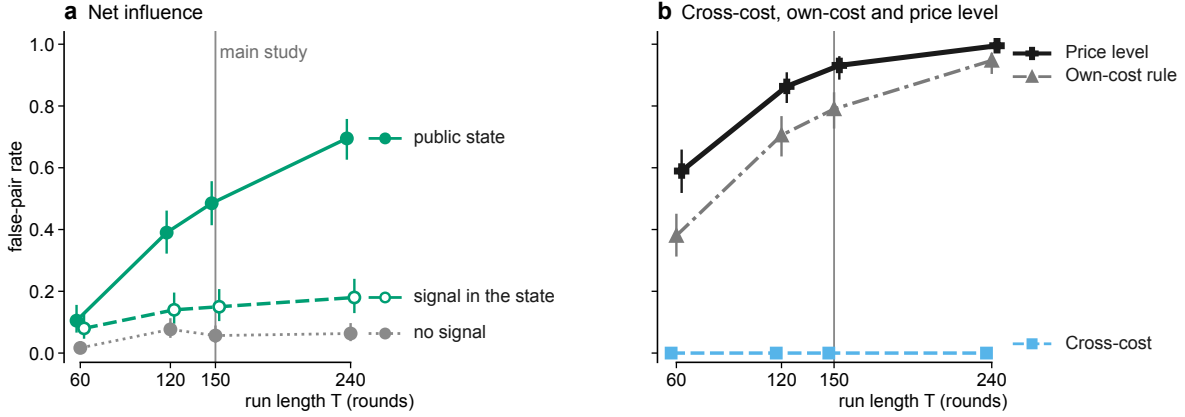


Figure 5: Spurious dependence from a shared signal in scripted markets: two dealers add two ticks whenever a shared signal is at least 9, without exchanging anything. Net influence (a) names a false pair more often with run length unless the signal is added to the conditioning state; the other rules (b) do not depend on the state. 200 markets per point; bars are exact 95% intervals.

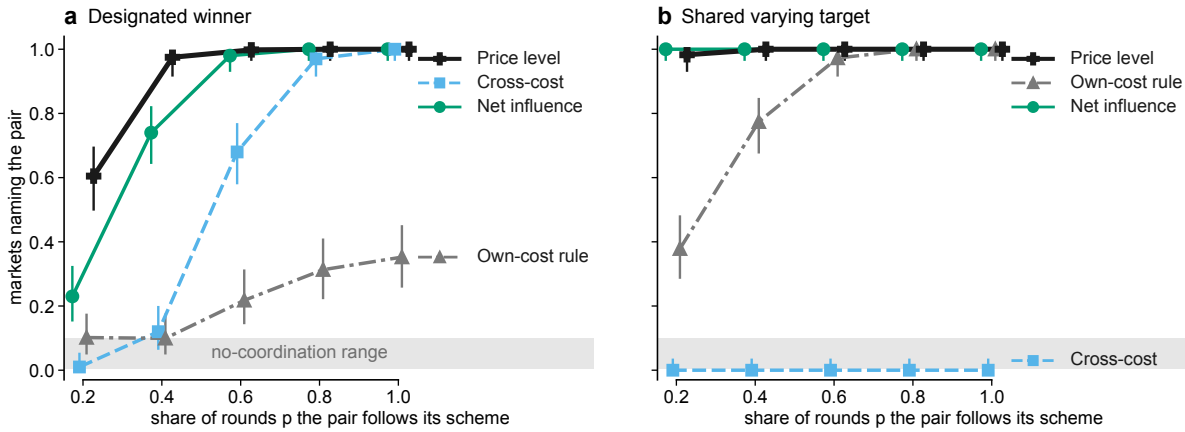


Figure 6: Scripted markets at 150 rounds: share of 100 markets in which each detector named the coordinating pair, against the share of rounds p in which the pair follows its scheme, for a designated winner (a) and a shared varying target (b). The gray band spans the false-pair rates of the four detectors without coordination. Bars are exact 95% intervals.

E Additional Sprint-Study Outcomes

Table 2: What each kind of evidence would establish, and how this study measures it.

Evidence target	Measured by
Channel access	Treatment assignment, verified by the simulator’s routing
Information exchange	Coded messages; cross-cost information in the quotes
Agreement	Coded reciprocal proposals and acceptances
Scheme consistency	Later quotes consistent with the coded scheme; insufficient by itself to establish agreement adherence
Economic harm	Price and profits against the competitive benchmark and the same-seed no-talk market

Table 3: Decision rules with Gemini 2.5 Flash dealers: markets in which each rule named the talking pair, a wrong pair, or any pair without a channel. Cross-cost information uses a conditional randomization test; other rules use fitted thresholds. The no-talk markets with seeds 10–18 set the own-cost, slope and price-level thresholds (each the largest gap among them), so scoring them for those rules would be circular (–); the held-out markets measure their false alarms.

Rule	Pair talk (R2)		No talk (R1)		Shared signal (R3)
	Talking pair	Wrong pair	Seeds 10–18	Held out	Any pair
Net influence	5/9	0/9	1/9	0/6	1/9
Cross-cost information	1/9	0/9	0/9	0/6	0/9
Own-cost information	3/9	0/9	–	0/6	1/9
Own-cost slope	1/9	1/9	–	0/6	0/9
Price level	2/9	1/9	–	0/6	2/9

Table 4: Pair localization and economic outcomes with communication, instructions and both treatments, for two models (Figure 2). Forced choice: markets in which the pair ranked first, with the one-sided exact binomial p against chance in brackets. Decision rules: markets in which the pair was named, with wrong pairs in brackets. Price paid and pair profit are differences from the no-talk market with the same seed (Gemini instructed pair: seeds 10–18). Quote entropy is the mean over dealers across the three quote bins; in no-talk markets it is 1.21 bits (Gemini) and 1.17 (gpt-oss).

	Gemini 2.5 Flash			gpt-oss-20b		
	Talk (9)	Instructed (18)	Both (9)	Talk (9)	Instructed (9)	Both (9)
<i>Forced choice</i>						
Mutual information	8 (.001)	6 (.59)	9 (<.001)	9 (<.001)	1 (.97)	9 (<.001)
Conditional MI	8 (.001)	5 (.77)	9 (<.001)	9 (<.001)	1 (.97)	8 (.001)
Cross-cost information	7 (.008)	4 (.90)	9 (<.001)	9 (<.001)	4 (.35)	9 (<.001)
Own-cost information	4 (.35)	7 (.39)	1 (.97)	9 (<.001)	9 (<.001)	8 (.001)
Price level	3 (.62)	15 (<.001)	6 (.042)	0 (1.00)	3 (.62)	8 (.001)
<i>Decision rules</i>						
Net influence	5 (0)	1 (1)	9 (0)	8 (0)	0 (0)	7 (0)
Cross-cost information	1 (0)	0 (0)	9 (0)	3 (0)	0 (0)	5 (0)
Own-cost information	3 (0)	1 (2)	1 (4)	8 (0)	3 (0)	5 (0)
Price level	2 (1)	5 (0)	0 (2)	0 (6)	1 (1)	5 (0)
Quote entropy (bits)	1.20	1.56	1.46	1.20	1.19	1.09
Price paid (ticks)	+0.17	+0.52	−0.10	−0.05	+0.09	+1.66
Pair profit (ticks per round)	−1.40	+1.64	−0.35	−4.15	−1.15	−0.05

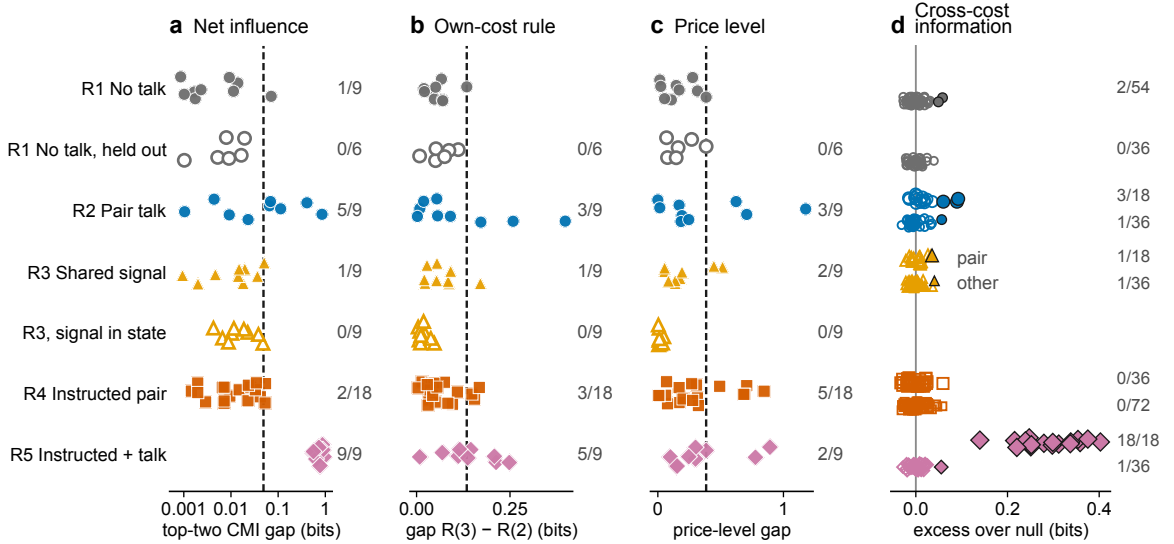


Figure 7: Statistics behind Gemini detector decisions in every condition, one point per market: net-influence gap (a), own-cost gap (b), price-level gap (c) and cross-cost information per market and direction (d, filled when the conditional randomization test gives $p \leq 0.05/3$). Dashed lines in (a)–(c) are fitted thresholds, not cross-cost test cutoffs. Hollow points are held-out no-talk markets and, for the shared signal, the analysis that adds the signal to the conditioning state.

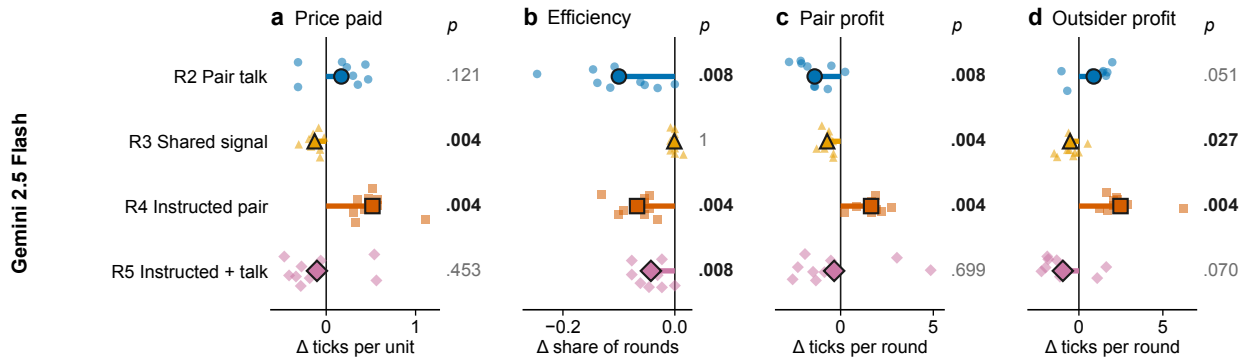


Figure 8: Gemini economic effects of pair talk, a shared signal, an instructed pair and an instructed pair with talk, relative to no talk under identical costs (nine matched seeds, 10–18). Small markers are seeds, large markers are means, and p -values are exact two-sided sign-flip tests over seeds (bold when $p \leq 0.05$). Table 6 reports both models.

Table 5: Market-level summaries over rounds 21–150. Failed quotes were played as the reserve quote of 12. Entropy is the mean per dealer across the three quote bins.

Model	Condition	Markets	Failed quotes	Entropy (bits)	Price paid	Efficiency
Gemini 2.5 Flash	No talk, calibration	9	0	1.21	3.53	0.996
	No talk, held out	6	0	1.19	3.52	0.994
	Pair talk	9	0	1.20	3.70	0.896
	Shared signal	9	1	1.12	3.40	0.995
	Instructed pair	18	0	1.56	4.15	0.930
	Instructed pair with talk	9	0	1.46	3.43	0.953
gpt-oss-20b	No talk	9	24	1.17	3.13	0.956
	Pair talk	9	24	1.20	3.08	0.853
	Instructed pair	9	14	1.19	3.21	0.909
	Instructed pair with talk	9	7	1.09	4.79	0.727

Table 6: Differences from the no-talk market with the same seed (identical costs), with exact two-sided sign-flip p -values over the seeds in brackets: nine seeds (10–18) for each model. Pair and outsider are the condition’s pair and the third dealer.

Model	Condition	Excess price (ticks)	Efficiency	Pair profit	Outsider profit
Gemini 2.5 Flash	Pair talk	+0.169 (.121)	−0.100 (.008)	−1.400 (.008)	+0.882 (.051)
	Shared signal	−0.129 (.004)	−0.001 (1.00)	−0.726 (.004)	−0.524 (.027)
	Instructed pair	+0.517 (.004)	−0.068 (.004)	+1.644 (.004)	+2.477 (.004)
	Instructed pair with talk	−0.103 (.453)	−0.043 (.008)	−0.349 (.699)	−0.956 (.070)
gpt-oss-20b	Pair talk	−0.050 (.574)	−0.103 (.004)	−4.150 (.004)	−0.164 (.520)
	Instructed pair	+0.085 (.191)	−0.047 (.004)	−1.148 (.008)	+0.199 (.488)
	Instructed pair with talk	+1.662 (.004)	−0.229 (.004)	−0.046 (.973)	+10.003 (.004)

Table 7: Matched-seed two-by-two interaction: mean effects of talk without the instruction (R2 – R1), with it (R5 – R4), and their difference $\delta = (R5 - R4) - (R2 - R1)$. Brackets give unadjusted exact two-sided sign-flip p -values. Treating each model’s six outcomes as a separate Holm family, all Gemini interactions except pair profit, and the gpt-oss price and outsider-profit interactions, remain below 0.05. With all twelve tests as one family, only Gemini cross-cost and outsider-profit interactions and gpt-oss price and outsider-profit interactions remain below 0.05 (adjusted $p = 0.047$); Gemini’s price interaction does not ($p = 0.063$). Neither correction covers the other detector or treatment comparisons. Dependence statistics use the assigned pair’s seats, including the seats allocated by the rotation in no-talk markets.

Outcome	Gemini 2.5 Flash (9 seeds)			gpt-oss-20b (9 seeds)		
	Talk	Talk, instructed	δ	Talk	Talk, instructed	δ
Price paid (ticks)	+0.17 (.121)	−0.62 (.016)	−0.79 (.008)	−0.05 (.574)	+1.58 (.020)	+1.63 (.004)
Efficiency	−0.100 (.008)	+0.025 (.188)	+0.125 (.008)	−0.103 (.004)	−0.182 (.004)	−0.079 (.098)
Pair profit (ticks per round)	−1.40 (.008)	−1.99 (.094)	−0.59 (.547)	−4.15 (.004)	+1.10 (.531)	+5.25 (.016)
Outsider profit (ticks per round)	+0.88 (.051)	−3.43 (.004)	−4.32 (.004)	−0.16 (.520)	+9.80 (.016)	+9.97 (.004)
Pair CMI (bits)	+0.18 (.047)	+0.76 (.004)	+0.58 (.012)	+0.15 (.004)	+0.16 (.004)	+0.01 (.855)
Pair cross-cost excess (bits)	+0.02 (.047)	+0.29 (.004)	+0.27 (.004)	+0.08 (.004)	+0.07 (.012)	−0.01 (.789)

Table 8: Separating Gemini pair talk from no talk: market-level scores in the nine pair-talk and nine no-talk calibration markets. AUROC is oriented so that values above 0.5 point to pair talk (for price variation, lower variation counts as pointing to pair talk); p -values are exact two-sided permutation tests, without correction for the ten scores.

Score	Pair talk	No talk	AUROC	p
Largest pairwise mutual information	0.205	0.041	0.74	0.061
Largest pairwise conditional mutual information	0.216	0.064	0.67	0.140
Net-influence gap	0.171	0.013	0.78	0.024
Cross-cost score	0.013	0.000	0.70	0.250
Own-cost gap	0.118	0.057	0.56	0.229
Own-cost slope gap	0.163	0.061	0.80	0.044
Price-level gap	0.368	0.166	0.63	0.166
Price paid	3.701	3.532	0.72	0.112
Excess price	−0.029	−0.198	0.77	0.093
Price variation (coefficient of variation)	0.366	0.305	0.19	0.021

F Talking-Pair Agreement Coding

We code each Gemini pair-talk message for stating the sender’s current cost, proposing a price, proposing who should win, or accepting a proposal. Market labels record cost exchange (a member states its current cost in at least 10% of rounds after burn-in), agreement (a proposal accepted or matched within two rounds), the agreed scheme and the agreement round. We measure *scheme consistency* as the fraction of subsequent rounds in which both members quote within one tick of a common price, or the higher-cost member quotes above the lower-cost member for a designated winner (Table 9).

The designated-winner rule does not distinguish agreement adherence from competitive cost-based bidding. Under the benchmark $b^*(c) = (3, 4, 4, 5, 6, 7)$, the higher-cost dealer quotes strictly above the lower-cost dealer in 14 of the 15 unequal-cost combinations (about 93%). Thus the reported 95% and 85% consistency alone cannot establish that dealers followed an agreement.

A human coder who had not seen the detector results used a GPT model to suggest and check labels, then reviewed every label. Three other LLMs supplied a second coding, combined by majority vote. The human coder marked the agreements in seeds 12 and 15 as uncertain because a matching quote, rather than explicit acceptance, followed each proposal; the second coding also assigned both designated-winner labels.

Table 9: Gemini pair-talk agreement labels and outcomes. Scheme consistency measures subsequent quotes matching the coded rule, not established adherence or profitability; the designated-winner labels in seeds 12 and 15 were uncertain. A dash under net influence denotes abstention.

Seed	Pair	Scheme	Agreed in round	Scheme consistency	Net influence	Efficiency	Second coding differs on
10	AC	Common price, 3 ticks	8	0.04	named	0.85	scheme, round
11	BC	Common price, 6 ticks	4	0.41	–	0.93	round
12	AB	Designated winner	14	0.95	–	1.00	round
13	AC	Common price, 5 ticks	5	0.38	named	0.89	–
14	BC	Common price, 5 ticks	6	0.12	named	0.96	–
15	AB	Designated winner	3	0.85	–	0.94	round
16	AC	Common price, 7 ticks	5	0.26	named	0.85	–
17	BC	Common price, 7 ticks	4	0.21	–	0.88	–
18	AB	Turn-taking	3	not computed	named	0.75	–

G Additional Detector Decision Matrix

Gemini 2.5 Flash

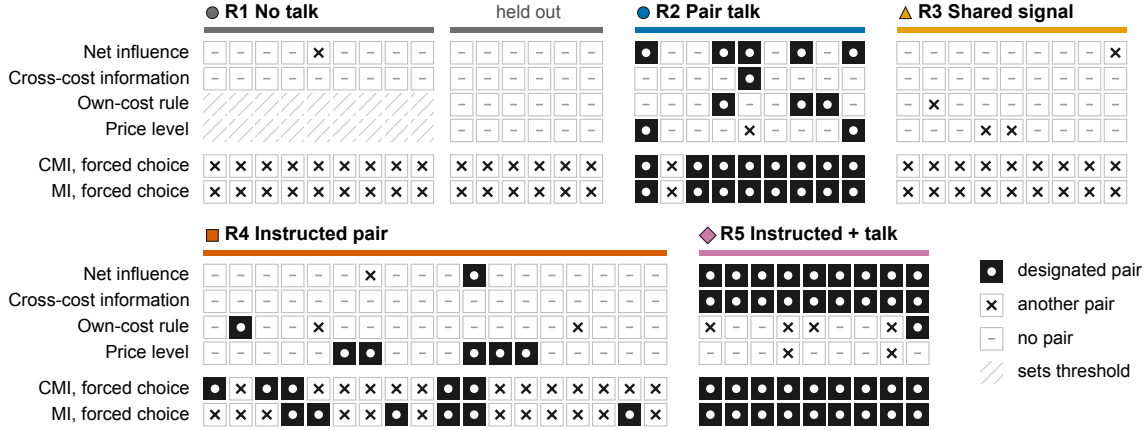


Figure 9: Which pair each detector named in each Gemini market. Cells are markets; a dot marks the assigned pair, a cross another pair and a dash no pair. In the no-talk and shared-signal conditions no pair is assigned, so any named pair is a false alarm. Hatched cells are the no-talk calibration markets that set the own-cost and price-level thresholds, which are therefore not scored. Forced choices always name a pair. Table 4 reports both models.

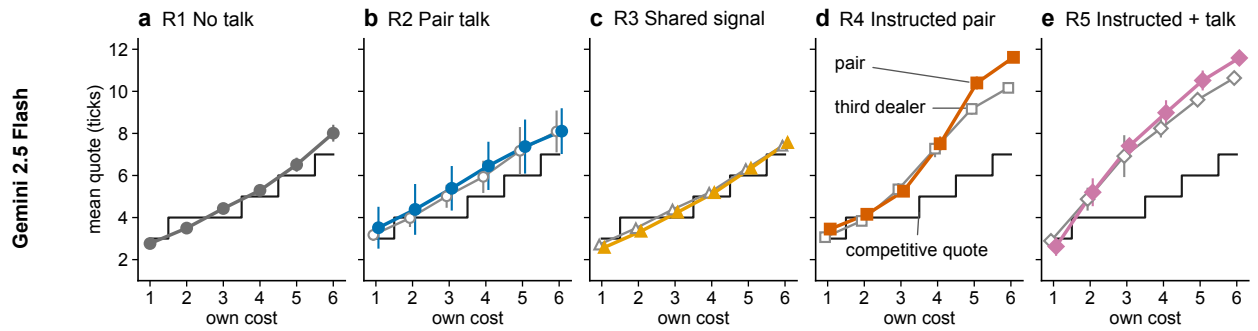


Figure 10: Gemini mean quotes at each own cost over rounds 21–150 for the pair members (colored) and the third dealer (gray, hollow), in each condition. Bars are 95% intervals across markets (nine per condition, 18 for the instructed pair); the step line is the competitive quote. In the no-talk condition all three dealers are shown in the same gray.

H Broader Research Questions: Pre-Sprint Evidence

Our pre-sprint dealer-market study (Bui, 2026) provides evidence on formation (RQ 1.1), text-monitor comparisons (RQ 1.3), competitive replacement (RQ 1.2) and restrictions (RQ 1.5). These experiments use different markets, horizons and models from the sprint auction. Intervals are 95% bootstrap intervals over runs, paired by source for forks.

Scope and limitations of the prior study. The earlier study used two stylised markets, one parameter value per treatment and ten to twenty runs of 30 rounds per cell. Forks used one fork round and one intervention length; dealers did not attempt to evade detection. Collusion labels were outcome-based, intent was not observed, and monitor results used one prompt per question. Catch rates at audit budgets below 10% rest on few competitive runs.

H.1 Quoting-market formation

Two dealers post half-spreads $s_i \in \{1, \dots, 12\}$, and the tighter quote receives orders. Informed orders arrive below a disclosed threshold $D = 5$ and impose an additional cost on the dealer. The competitive benchmark is $s^* = 3$; undercutting a rival holding D is profitable. In a variant with five times as many informed orders, holding D is instead an equilibrium. Runs last 30 rounds and are evaluated over the last ten. The collusion index $\Delta = (\bar{\pi} - \pi^N)/(\pi^M - \pi^N)$ compares observed profit with competitive and joint-profit-maximizing benchmarks (Calvano et al., 2020).

Without informed traders, both models competed without talk ($\Delta = 0.00$ for Gemini and 0.06 for Haiku) and colluded with talk. With informed traders, spreads remained at the disclosed threshold D in every evaluated round of 20 of 20 Haiku runs and 18 of 20 Gemini runs without messages, although D was a best reply for neither dealer. Hiding D lowered spreads by 1.85 ticks with Haiku (95% interval 1.69–2.00) and 1.15 with Gemini (0.73–1.52). In the earlier two-dealer auction, both models played the Bayesian equilibrium without messages; with messages, Gemini raised the price by 4.33 ticks (3.14–5.45).

H.2 Text monitors and price screens

After a run ended, a Claude Haiku 5.5 monitor scored suspicion of coordination from 0 to 100; the flagging threshold was set to flag 10% of competitive runs. Dealers were not told of the monitor. The prior report distinguished stated agreements, quoting-market messages presenting high spreads as protection against informed traders (“cost story”), and quoting-market runs without messages (“tacit”). Reading full logs, the monitor caught 0.94 (0.86–1.00) of stated collusion, 0.50 (0.37–0.63) of the cost-story group and no tacit run, even with private reasoning. Giving the competitive benchmark caught all stated and cost-story runs but still no tacit run; asking whether traders paid more than under competition instead flagged 0.45 (0.32–0.58) of tacit runs (AUROC 0.94). The final spread alone ranked quoting-market runs with AUROC 0.82, against 0.76 for the monitor, but could not distinguish the quoting market from its competitive variant with identical quotes (AUC 0.50).

H.3 Prior intervention and fork results

The pre-sprint dealer-market study (Bui, 2026) compares competitive replacement (RQ 1.2) and restrictions (RQ 1.5). Externally imposed changes lowered elevated prices, whereas frictions depending on dealers’ own choices did not consistently do so (Tables 10 and 11). These experiments do not rank monitoring or intervention costs.

Replacing a dealer. When an equilibrium bot replaced one dealer, elevated pricing ended in both markets; in the auction the price fell from 10.50 to 4.51 ticks. A one-round cut to $D - 1$, which pays the deviator more than holding D , removed the Haiku quoting-market floor for the rest of the run without messages (share of the markup retained 0.03, 95% interval 0.01–0.05) but not with them (0.71, 0.51–0.87). No quoting dealer punished the tested deviation, so these outcomes did not establish retaliatory enforcement.

Rules, frictions and talk. In the auction with messages, a third dealer lowered the winning price by 4.16 ticks with Gemini and 1.09 with Haiku, and an unannounced competitive seller in a third of requests by 4.02 and 0.62. Frictions that bind only through the dealers’ own choices did not help: dealers kept paying a per-message fee, which raised auction prices (by 1.63 ticks with Gemini, 1.19 with Haiku).

Matched histories: talk rebuilds high prices. Continuations from identical histories isolate the channel’s role in recovery after an intervention (ten sources per cell). After a bot or spread cap was withdrawn, the quoting-market floor did not return without messages (−1.17 ticks after the cap), but dealers that could talk rebuilt it (−0.17). In the auction with messages, the bot lowered the collusive price by 4.63 ticks while in force and by 0.87 after withdrawal. A monitor shown the benchmark and asked about harm in the first ten rounds flagged no tacit run at a 10% audit budget, but flagged 0.70 (0.40–1.00) at 20%, with no competitive run flagged.

Table 10: Pre-sprint. Change relative to the same market without the treatment, last ten rounds; negative values lower prices. Exogenous rules are set from outside; endogenous frictions bind only through a dealer’s own choice.

Treatment	Type	Claude Haiku 5.5	Gemini 2.5 Flash
<i>Quoting market, change in market spread (ticks), no messages</i>			
Outside venue at $D - 1$	exogenous	−1.92 [−2.03, −1.73]	−1.83 [−2.00, −1.59]
Spread cap at $D - 1$	exogenous	−1.61 [−1.83, −1.36]	−1.89 [−2.06, −1.65]
Inventory cost	endogenous	0.00 [0.00, 0.00]	0.15 [0.00, 0.40]
Message fee (with messages)	endogenous	−0.05 [−0.15, 0.00]	0.34 [0.09, 0.61]
<i>Auction, change in winning price (ticks), with messages</i>			
Third dealer	exogenous	−1.09 [−1.42, −0.71]	−4.16 [−5.95, −2.12]
Competitive seller in 1/3 of requests	exogenous	−0.62 [−0.85, −0.39]	−4.02 [−5.21, −2.79]
Second-price format	exogenous	2.17 [1.15, 3.30]	−0.73 [−2.79, 1.34]
Winner-only disclosure	exogenous	1.06 [0.13, 2.05]	0.31 [−1.71, 2.18]
Message fee	endogenous	1.19 [0.76, 1.70]	1.63 [0.20, 3.03]

Table 11: Pre-sprint quoting-market forks (Claude Haiku 5.5): change in market spread (ticks) against the control fork of the same source, in rounds 11–20 (in force) and 21–30 (withdrawn); ten sources per cell. In the competitive variant, holding D is the equilibrium.

Intervention	Collusive, no messages		Collusive, messages		Competitive variant	
	in force	withdrawn	in force	withdrawn	in force	withdrawn
Committed bot	−2.00	−0.70	−2.00	−0.20	0.00	0.00
Reactive bot	−1.00	−0.93	−1.00	−0.31	0.00	0.00
Spread cap at $D - 1$	−1.02	−1.17	−1.00	−0.17	−1.02	0.00
Outside venue at $D - 1$	−0.74	−1.03	−0.13	−0.12	−0.16	0.00

In the auction forks (Gemini 2.5 Flash), the equilibrium bot changed the collusive price of sources with messages by -4.63 ticks $[-6.56, -2.64]$ while in force and by -0.87 $[-1.96, -0.11]$ after withdrawal, and the competitive seller lowered the price of competitive sources without messages by 0.50 $[0.37, 0.64]$, a gain for the buyer. In the competitive quoting variant, the spread cap drove the collusion index to -11.64 $[-11.86, -11.50]$ against the control. Four main pre-sprint findings (a stated threshold raises the spread, hiding it lowers it, tacit collusion is caught less often than stated collusion, and a competitive seller lowers the auction price) hold after Holm correction, on all runs and on a held-out set of later runs.