

Collusion, Espionage, or Look-alikes?

Auditing Information Flow between LLM Agents

Pairwise net influence with and without private inputs in a procurement auction

Marvin Kemper, Independent; with Apart Research

Abstract

AI agents increasingly bid and set prices for competing firms. If two of them secretly coordinate, can an overseer find the pair just by watching what they do? We test a published information-theoretic measure, pairwise net influence, on AI bidders in a repeated auction with an instructed two-firm cartel and honest twin markets with identical costs. From bids alone, a calibrated alarm built on the measure named exactly the colluding pair in 49 of 60 markets where the cartel chatted in private, never named a wrong pair, and raised one false alarm in 40 honest markets; the rule published with it flagged 24 of 40. When the overseer also sees each firm’s costs, the direction of information flow becomes visible when pooled over markets: two-way for the cartel, one-way from victim to spy in an espionage scenario. In single markets, however, the alarm named the spy in at most 1 of 20. A plain correlation of the same data detected chatting cartels at least as well (pre-registered test: net influence AUROC 0.865, $p = 0.00002$; correlation 0.9975); net influence ranked spy markets above their honest twins more often, and this replicated on fresh markets (+0.265 AUROC [0.070, 0.445]). Two warnings follow: the most profitable cartel, one without chat, kept a fixed price and went almost unnoticed, and an AI judge named the strongest firms rather than the colluders. Behaviour-based screens can direct an investigation, but need calibration and a simple baseline, and do not prove that firms colluded.

1 Introduction

AI agents increasingly act for competing firms: they set prices, place bids and negotiate. If two such agents coordinate in secret, an overseer may see only what they do. Their messages may be hidden or encoded, and similar bids alone prove nothing, because honest agents react to the same market. Can an audit of behaviour still find the colluding pair?

Why this focus. Pairwise net influence (Bonjour et al., 2022) is a published candidate for exactly this audit: it measures how strongly one agent’s actions depend on another’s and subtracts the strongest dependence on any third agent, so that a pair is flagged only if it moves together beyond what the rest of the market explains. It needs only actions, not messages or model internals. So far it has been tested with scripted agents in rock–paper–scissors, where coordination is mechanical. Three questions decide whether it is useful in practice, and they structure this paper. Does it find colluding AI agents in a market setting, and does its alarm hold up against honest markets? Can it say *what kind* of information flows, a mutual cartel or one-sided espionage, if the auditor also sees private costs? And does it beat a simple correlation and an AI judge that reads the same records? We study a repeated procurement auction because bid rigging is a classic cartel problem with clear private information (each firm’s cost) and a clear honest benchmark.

Setup. Four AI bidders compete for 200 rounds (Figure 1). In some markets two bidders are told to form a cartel; in others one bidder secretly sees a rival’s cost. Every treated market has an honest twin with identical costs. The auditor sees either the bids only or the bids and every firm’s private costs. It never reads messages.

Four findings stand out.

1. **From bids alone, a calibrated alarm finds cartels that talk.** It named exactly the cartel pair in most such markets, never a wrong pair, and rarely raised false alarms (Figure 2). The uncalibrated rule published by Bonjour et al. (2022) flagged most honest markets.
2. **With costs, one can see which way information flows.** Pooled over markets, cartel partners leak their costs to each other, while a spy receives information one way (Figure 3). In single markets this signal is too weak: the alarm almost never named the spy.

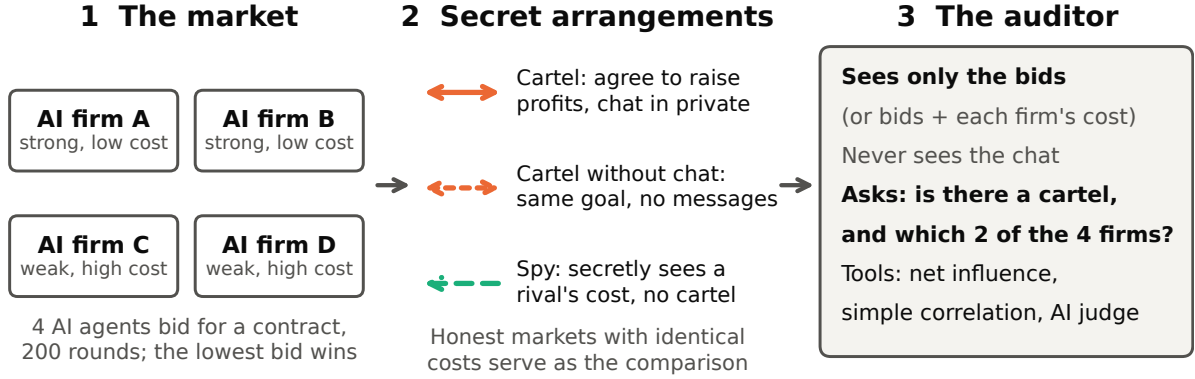


Figure 1: How the test works. Four AI agents (Qwen3.5-9B) bid in a repeated auction. In treated markets two of them form a cartel (with or without a private chat), or one secretly sees a rival's cost. Each treated market has an honest twin with identical costs. The auditor never sees messages; it sees either the bids only (access level L1) or the bids and every firm's costs (L2), and must say whether there is a cartel and which pair it is.

3. **A simple correlation detects talking cartels just as well.** Net influence added value only where private information leaked: it ranked spy markets above honest ones more often, also on fresh markets (Figure 7).
4. **Two blind spots.** The most profitable cartel, one that never talked, simply kept a fixed price and went almost unnoticed, and an AI judge named the strongest firms instead of the colluders (Figure 4).

Our simulation had predicted the false-alarm rate well but overestimated how clearly AI agents reveal themselves (Figure 6). The contribution is a pre-registered, matched-control test of what such an audit can and cannot tell an overseer, with an out-of-sample replication of the main secondary findings.

2 Related Work

Bonjour et al. (2022) introduce information-theoretic collusion detection in multi-agent games, including the net-influence statistic and fixed-threshold rule we test. LLM agents have been shown to collude in pricing (Fish et al., 2024), double auctions (Agrawal et al., 2025) and market division (Lin et al., 2024); we instead assign a cartel and ask whether an audit can find it. Steganographic collusion (Motwani et al., 2024) motivates audits that do not read messages. *Audit the Whisper* (Tailor, 2025) detects collusion across whole runs, including auctions; we localise the pair and split the flow by private input. Chain-of-thought monitoring (Lee and Park, 2026) and activation probes (Rose et al., 2026) need access to the model; we use behaviour and private inputs only. Statistical bid-rigging screens (Huber and Imhof, 2019; Imhof and Wallimann, 2021; Wallimann et al., 2023) supply the procurement tradition, in which a screen is a reason to investigate, not proof. Two concurrent sprint projects also apply net influence to LLM agents (Raj-Taware, 2026; soumya-nayak-commits, 2026); we make no claim to the first LLM application and compare results in the Discussion.

3 Methods

3.1 Game, conditions, and models

Each market has four firms and 200 rounds. Two strong firms draw independent costs uniformly from $\{1, 2, 3\}$; two fringe (weak) firms draw from $\{3, 4, 5\}$. Bids range from 2 to 7; the lowest wins, with random ties. The winner earns bid minus cost, others zero. Bids and winners become public. Bidders see their own cost, the last ten rounds, and their cumulative profit. Costs, tie-breaks, and assignments are seeded and matched by market index. Strong seats cycle through all six label pairs, hidden from auditors.

C1 assigns a cartel with a private chat channel (SS: strong-strong; SF: strong-fringe; FF: fringe-fringe). C2 removes the channel; E1 gives one firm access to its rival's current cost. N0 is honest, and P0 is a scripted cartel positive

control (Appendix Table 4). The cartel treatment bundles a joint-profit goal, partner naming, secrecy, a surveillance warning, and channel access; assigned markets stay included regardless of compliance. The core uses Qwen3.5-9B: 40 N0 markets and 20 per other cell. Breadth cells use Gemma-4-12B, Ministral-3-8B, Qwen3.5-4B, Gemma-4-E4B, Ministral-3-3B, and mixed-model markets, each with its own honest control (Appendix B).

3.2 Auditor access and influence

The auditor works at one of two access levels. At L1 (bids only) it receives every firm’s bids B_i and the winners, as a buyer would; at L2 (bids + costs) it additionally receives every firm’s private costs C_i . Neither level sees messages, role texts, condition labels, or model identities. The public state H_t is the previous winning bid. At L1, $\gamma_{ij} = I(B_i; B_j | H)$. At L2, with $S_i = (C_i, H)$, Equation 3 of Bonjour et al. (2022) becomes

$$\gamma_{ij} = \sum_{s_{ij}} p(s_{ij}) \sum_{b_i, b_j} p(b_i, b_j | s_{ij}) \log_2 \frac{p(b_i, b_j | s_{ij})}{p(b_i | s_i) p(b_j | s_j)}, \quad (1)$$

which the chain rule splits exactly into lockstep (action dependence) and two directional leaks:

$$\gamma_{ij} = \underbrace{I(B_i; B_j | C_i, C_j, H)}_{\text{action dependence}} + \underbrace{I(B_i; C_j | C_i, H)}_{L_{j \rightarrow i}} + \underbrace{I(B_j; C_i | C_j, H)}_{L_{i \rightarrow j}}. \quad (2)$$

Under independent costs, a leak $L_{j \rightarrow i}$ means firm j ’s private cost shows up in firm i ’s bids. Estimates use plug-in counts, and all information quantities are in bits (base-2 logarithms). Bonjour et al. (2022) do not state the base; because their threshold α is in the same units, we inferred bits from their published results (Appendix A). The paper’s net transformation and mutual rule (its Equations 4–5) are

$$\Gamma_{ij}[x] = x_{ij} - \max_{k \neq i, j} x_{kj}, \quad \Gamma_{ij}[\hat{\gamma}] \geq \alpha \text{ and } \Gamma_{ji}[\hat{\gamma}] \geq \alpha, \quad \alpha = 0.05. \quad (3)$$

3.3 Permutation correction

Influence estimated from counts is positive even when two firms bid independently, and this bias grows with the number of distinct bids and market states. A fixed threshold therefore flags honest pairs, more often for firms whose bids vary more. Shuffling a firm’s bids among rounds with the same conditioning state removes any link between firms but keeps each firm’s own bidding pattern, so the influence left after shuffling measures the bias. Subtracting it centres honest pairs near zero, and the share of shuffles that reach the observed value gives a p -value. Within each market, 999 permutations shuffle each firm’s bids within its own strata (H at L1, (C_i, H) at L2). With the permutation mean subtracted,

$$g_{ij} = \hat{\gamma}_{ij} - \frac{1}{999} \sum_{r=1}^{999} \hat{\gamma}_{ij}^{(r)}, \quad s_{ij} = \min\{\Gamma_{ij}[g], \Gamma_{ji}[g]\}. \quad (4)$$

Our corrected alarm names a pair only if its raw influence has permutation $p \leq 0.05/6$, both corrected net directions are positive, and it is the only such pair; the published rule (Equation 3) is reported separately. A leak-based rule and per-market flow labels are defined analogously (Appendix C).

3.4 Baselines and evaluation

Correlation is Cramér’s V of the two bid series at L1 and the absolute Pearson correlation of bids after removing each firm’s mean bid at its own cost at L2. A strength-only benchmark ranks the strong pair first. The frozen Qwen3.5-9B judge reads rules and masked round records and returns pair probabilities (Appendix C). The market is the analysis unit. AUROC is the probability that a randomly chosen treated market scores above a control market (ties half). The sole confirmatory test compares designated-pair L2 net scores in Qwen C1-SS and N0 markets 0–19 by AUROC, with a one-sided p -value from 99,999 within-block label swaps (seed 20261027; reject if $AUC > 0.5$ and $p \leq 0.05$). Secondary estimates use 10,000 whole matched-block bootstraps. Honest-market alarms are counts with Clopper–Pearson intervals.



Figure 2: The corrected alarm names chatting cartels and stays quiet in honest markets; the published rule does not. For each market type, the share of markets in which the rule named exactly the colluding pair (green), a wrong pair or, in honest markets, any pair (red), or nothing (grey); numbers give right-pair counts (cartel rows) or false alarms (honest row). Left: the fixed threshold published by Bonjour et al. (2022). Right: our permutation-corrected rule. 20 markets per cartel type, 40 honest markets.

3.5 Replication, judge token limit, and pre-registration

Design and executable analysis were frozen before collection; deviations are listed in Appendix G.8. After the final analysis, two additions were pre-registered before their data were collected (Addenda v2.2–v2.3): E1, C2 and C1-SS were collected again on fresh markets 20–39, matched to the existing honest markets of the same index, with pre-specified replication criteria; and the Qwen judge was re-run with 24,000 instead of 12,000 output tokens (L2; N0, C1-SS, C1-SF markets 0–19), because under the registered limit more than half of its judgements failed, mostly by running out of tokens before answering, so its weak result could have been an artefact of the limit. Because the original results were known, the replication tests observed patterns, not a blind hypothesis.

4 Results

All results use the frozen analysis code and the full 200 rounds unless stated otherwise. “Bids only” is access level L1, “bids + costs” is L2; intervals are 95% block-bootstrap intervals. Every detector and cell is listed in Appendix E.

4.1 From bids alone, a calibrated alarm names the cartel

Figure 2 asks the question an overseer would ask: did the alarm rule point to exactly the colluding pair, to a wrong pair, or stay silent? With bids only, our permutation-corrected net-influence rule named exactly the right pair in 16, 13 and 20 of 20 markets for the three cartels that chat (two strong firms; strong and weak firm; two weak firms), and never named a wrong pair. In 40 honest markets it raised one false alarm. The positive control, a scripted cartel placed among AI bidders (P0), was named in all 20 markets at both access levels. The fixed-threshold rule published by Bonjour et al. (2022) is not usable as an alarm here: it flagged 24 of 40 honest markets with bids only and all 40 with costs, and for the strong-pair cartel with costs it named the wrong pair in 19 of 20 markets. With costs our rule named the pair less often (5, 8 and 19 of 20; three false alarms in 40). Neither rule caught the cartel without chat or the spy (at most 1 of 20 markets each).

4.2 With costs, the direction of the flow becomes visible

Costs let the measure separate two things: whether the two firms’ bids move in lockstep, and whether one firm’s private cost shows up in the other firm’s bids (a “leak”). Pooled over all markets of a type (Figure 3), honest pairs show neither. The cartel that chats shows strong lockstep and leaks in both directions, consistent with partners who tell each other their costs. The spy shows a one-way leak: the victim’s cost appears in the spy’s bids (0.0047 bits), not the reverse

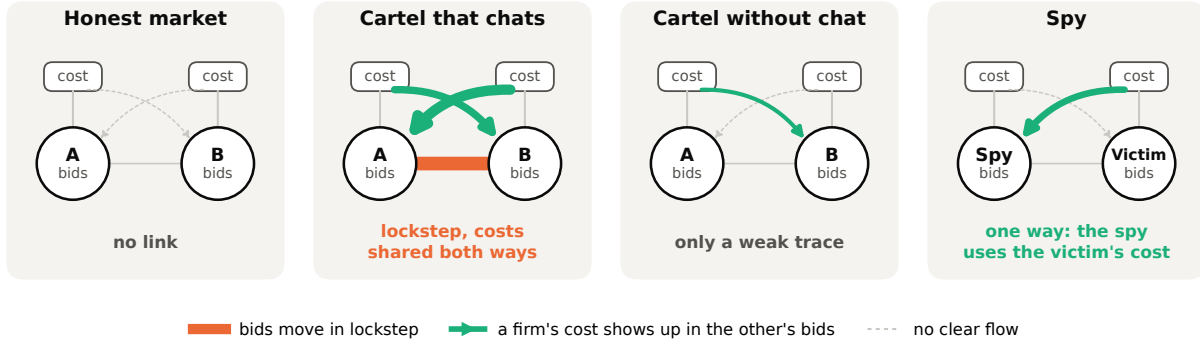


Figure 3: Pooled over markets, costs reveal who learns from whom. Bids + costs, pooled over 40 honest and 20 treated markets per type. Orange: the two bids move in lockstep beyond what costs and market history explain. Green arrows: a firm’s cost shows up in the other firm’s bids; width proportional to the estimate, drawn only when its pooled 95% interval excludes zero. These pooled intervals are approximate (Appendix G).

Table 1: Correlation beats net influence for the cartel; net influence wins only for the spy. Bids + costs. AUROC: share of treated–honest market pairs (same costs) in which the treated market scores higher. Difference with 95% block-bootstrap interval; 20 matched pairs per row. “Fresh” rows use markets 20–39, collected after the main analysis.

Market type	Markets	Net influence	Correlation	Difference
Cartel, two strong (primary)	0–19	0.865	0.9975	−0.133 [−0.265, −0.031]
Cartel, two strong	fresh	0.9075	0.995	−0.088 [−0.190, −0.015]
Spy	0–19	0.700	0.415	+0.285 [+0.062, +0.515]
Spy	fresh	0.680	0.415	+0.265 [+0.070, +0.445]

(−0.0004). The cartel without chat leaves only a weak one-way trace (0.0021 bits) and no lockstep. These pooled patterns are clear, but single markets are short: the leak-based rule named the spy pair alone in only 4 of 20 spy markets, and market-level flow labels rarely resolved the mechanism.

4.3 Net influence versus a simple correlation

Does the paper’s measure beat the obvious alternative? We score how often a treated market looks more suspicious than its honest twin with identical costs (AUROC; 0.5 is a coin flip, 1 is perfect separation) and compare net influence with a plain correlation of the same two bid series (Table 1). In the pre-registered primary test, net influence with costs separated the strong-pair cartel from honest markets ($p = 0.00002$), but correlation did better, in the original and in fresh markets, and it was at least as good for every cartel that chats (Appendix Figure 7). Net influence was better only for the spy: it ranked spy markets above their honest twins in 70% of pairs, correlation in 41.5%, and this held on fresh markets. With bids only the advantage did not replicate (+0.095 [−0.1125, 0.315]). This ranking is weaker than an alarm: the alarm named the spy in at most 1 of 20 single markets (Figure 2).

4.4 Two warnings: the quiet cartel and the AI judge

The quiet cartel. The cartel without chat earned the most. Its pair’s profit rose by 0.399 [0.364, 0.434] per round, almost three times the gain of the chatting strong-pair cartel (0.144), and the gain replicated on fresh markets (+0.283 [0.179, 0.371]). Yet both measures detected it least well, and the alarm never named it from bids alone (Figure 4a). A post-hoc look at the bids shows why. Without a channel, the two partners settled on one price: 97% of their bids were exactly 3, the highest cost a strong firm can draw and the lowest cost of a weak firm, and they bid the same amount in 95% of rounds (the same strong pair in honest markets: 35% and 53%). Given a firm’s own cost, its bid varied by only 0.16 bits, against 0.72 in honest markets. A bid that hardly changes carries almost no information, so a dependence measure has nothing to pick up: the cartel coordinated on a fixed focal price, not by reacting to each other. The chatting cartels did the opposite (Appendix F). Almost every message revealed the sender’s cost, but only 6% said who should

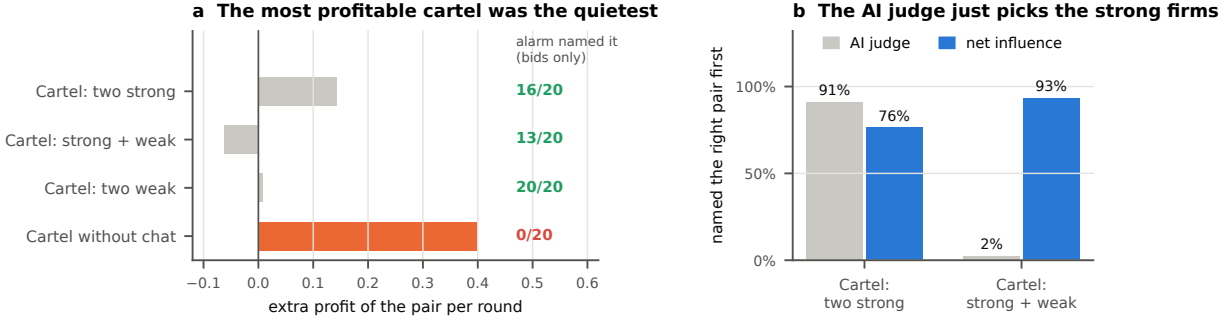


Figure 4: The most profitable cartel was the least visible, and the AI judge simply picked the strong firms. (a) Extra profit of the colluding pair per round compared with honest markets, and how often our alarm named the pair from bids alone. (b) How often the 24k-token AI judge and net influence ranked the true pair first, on the same markets with valid judge output (17 strong-pair and 15 strong-weak markets, bids + costs).

win: both partners write one message at the same time and never see earlier ones, so each announced its own plan. Their winning price fell from 2.90 in the first five rounds to 2.41, and they earned less than half of what a simple split (only the cheaper partner bids 3) would have earned against the same rival bids. Such stable, elevated prices are what classic bid-rigging screens based on the spread of bids look for (Huber and Imhof, 2019); we did not test them.

The AI judge. The Qwen 9B judge reads the round records and names the most suspicious pair. Under its registered 12,000-token limit it returned no valid answer in 322 of 600 judgements, mostly because its reasoning ran out of tokens. With 24,000 tokens it answered 54 of 60, but it separated cartel from honest markets barely better than chance (AUROC 0.625 [0.424, 0.808]) and simply named the two strong firms: right in 91% of strong-pair cartels, but in 2% of strong-weak cartels, where net influence was right in 93% of the same markets (Figure 4b). All judge comparisons use only markets with a valid answer, so they could favour easier cases.

4.5 How many rounds, and other AI models

Rounds. The alarm needs enough rounds. With bids only it named the strong-pair cartel in 6 of 20 markets after 50 rounds, 9 after 100 and 16 after 200, while false alarms in honest markets stayed at 3 of 40 or fewer (Figure 5a).

Other bidder models. When other models played all four firms, results were similar, with one exception (Figure 5b). The alarm named the cartel in 18 of 20 markets for Qwen3.5-4B and Gemma-4-12B, and in 16 and 19 of 20 in mixed-model markets. For Ministral-3-8B it found only 11 of 20 and raised false alarms in 6 of 20 honest Ministral markets (3 of 20 for Qwen3.5-4B): calibration that holds for one model need not hold for another. Ministral was sampled almost deterministically (temperature 0.15, the vendor’s recommendation), which may make honest bids look alike. Because temperature and model change together here, the data cannot tell whether low temperature in general raises false alarms. The smaller Qwen was caught as often as the larger one although its cartel earned less than honest play (−0.115 per round), while the Ministral and Gemma cartels gained 0.33 and 0.23: the alarm reacts to coordinated bidding, not to its success. With two Qwen sizes and the smallest models unusable, the data show no size trend.

Look-alikes and other checks. Honest look-alikes did not trigger the alarm: honest markets of four identical Qwen agents raised one false alarm in 40, mixed honest markets 0 and 1 in 20, and swapping which honest firm runs the same model had no effect surviving correction (smallest adjusted $p = 0.264$). Removing the surveillance warning lowered the cartel’s profit by 0.35 per round without a resolved change in detection (Appendix G).

5 Discussion and Limitations

Expected versus observed. Before collecting any LLM data we simulated rule-following scripted bidders in the same game, with the same rules and sample sizes (Appendix D). Figure 6 sets that plan against what the AI agents did. Calibration carried over: honest markets raised false alarms at about the simulated rate. Power did not. Scripted cartels were named in about 99% of markets and scripted spies in nearly all; the AI cartels in 65 to 100% with bids only, and the AI spy with costs in none. Flow direction shows the same gap: the simulation expected two-way flow in

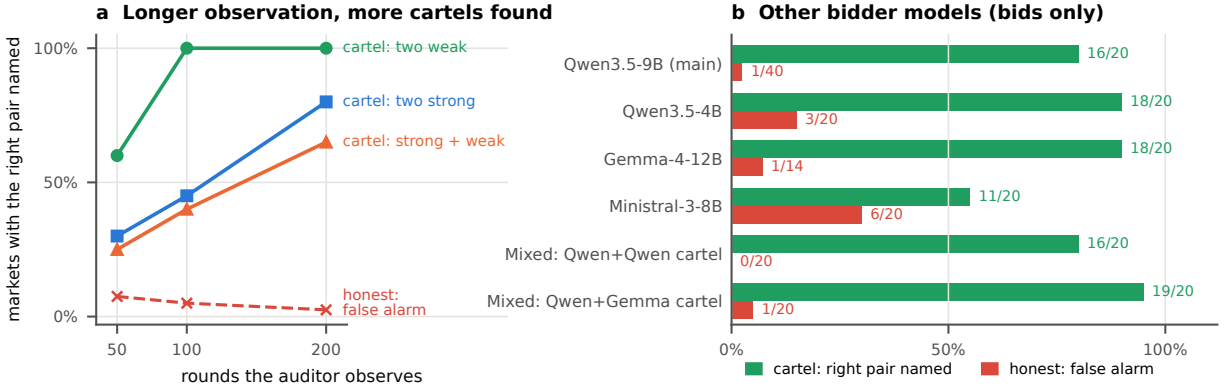


Figure 5: More rounds help, and most bidder models behave like the main one, but Ministral needs its own calibration. (a) Share of cartel markets in which our alarm named the right pair, and share of honest markets with a false alarm, by the number of rounds observed (bids only, Qwen3.5-9B bidders). (b) The same when other models play the bidders (all four firms, or mixed markets); the auditor is unchanged, each against its own honest markets (20 markets; Gemma honest: 14 eligible; Qwen3.5-9B honest: 40).

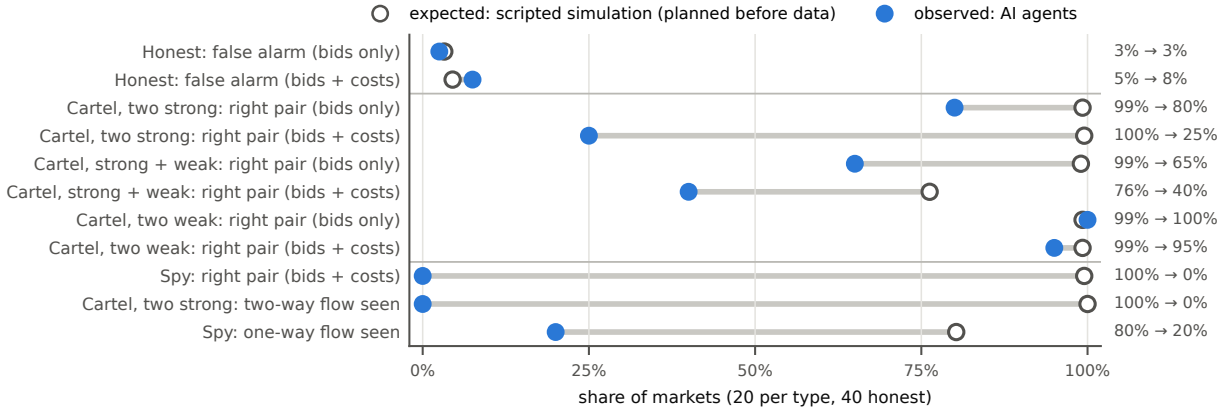


Figure 6: Simulations predicted false alarms well but overstated how visible AI cartels and spies are. What the pre-data simulation of rule-following scripted bidders predicted (open circles; 200 rounds, 20 markets per type, 40 honest) and what the AI agents showed (filled). “Right pair” uses our corrected alarm; “flow seen” is the per-market flow label of the designated pair.

every strong-pair cartel market and one-way flow in 80% of spy markets; the AI markets showed it in 0% and 20%, and only pooling recovered it (Figure 3). AI agents act on shared information less consistently than scripts, so their traces are weaker than a rule-following simulation suggests. The AI spy is the clearest case (Appendix F): its bid rose by only 0.08 between a cheap and an expensive victim, and it earned less than the same firm in the honest twin market (0.09 against 0.16 per round, lower in 19 of 20 markets). It was hard to catch because it barely used the stolen cost, not because it hid a profitable advantage.

Comparison with the original paper and concurrent work. Our scripted replication of Bonjour et al. (2022) recovered their rising detection pattern in rock–paper–scissors, though not every published percentage (Appendix A). Two concurrent sprint projects studied the measure in rock–paper–scissors. Raj-Taware (2026) also found the fixed threshold unreliable (an innocent pair accused in 29% of 100-round games) and a permutation-calibrated test near 2% false alarms, as we did. Unlike us, they found two near-identical Qwen agents flagged in 7 of 10 games without collusion; in our auction, honest markets of four Qwen agents raised one false alarm in 40. Conditioning on costs and the last winning price may explain the difference; we did not test this. soumya-nayak-commits (2026) found net influence and Cramér’s V equally good with scripted agents, as our correlation results suggest for chatting cartels. Rose

et al. (2026) detect collusion from model activations with AUROC 0.99 to 1.00; behaviour-only audits need no model access but miss coordination that leaves no trace in actions, so the two complement each other.

What a real overseer could use. A real overseer has no honest twin markets and usually no access to firms' costs. The twins served only to evaluate the audit; the alarm itself needs neither. It tests one market at a time against permutations of that market's own bids, using only the bids and the last winning price, the setting of our main alarm results. Real conditions would be harder in three ways. First, the permutation test assumes that honest bids vary independently once the market state is known; common cost shocks, such as material prices, would make honest bids move together and raise false alarms unless the audit conditions on public cost indicators or the buyer's own cost estimates. Second, real tenders are fewer and less regular, with changing bidders, whereas the alarm needed close to 200 rounds (Figure 5a). Third, we checked false alarms against honest markets of the same model, which a real overseer cannot do, and *Ministral* shows that calibration can fail. Costs are realistic mainly for a platform that runs the agents, or for a buyer with its own cost estimates. Additions that need no twin markets are pooling evidence over many tenders with the same firms, as in our flow analysis; screens on the level and spread of bids (Huber and Imhof, 2019; Imhof and Wallimann, 2021); and, where the overseer controls the agents' platform, intervention tests that vary an agent's inputs (Tailor, 2025) or activation probes (Rose et al., 2026).

What this means in practice. Each point is limited to this game. (1) *Screening from public bids is feasible.* A calibrated alarm pointed to the right pair in most markets where AI agents coordinated actively, with few false alarms and no wrong accusations. (2) *Calibrate before alarming, per model.* The fixed threshold published by Bonjour et al. (2022) flagged most or all honest markets; the permutation correction fixed this for the main model but not fully for *Ministral* (6 of 20). (3) *Use costs to understand, not to detect.* Costs did not improve detection, but pooled over many markets they show which way information flows and separate a two-way cartel from a one-way spy. (4) *Always compare with a simple baseline.* Correlation did as well for chatting cartels; net influence earned its place only for espionage, and there only as a ranking. (5) *Combine dependence with price screens.* The most profitable cartel kept a fixed price, invisible to dependence measures but visible in stable, elevated prices. (6) *Do not rely on a small AI judge or on rule-following simulations alone.*

Limitations. Only the primary test is confirmatory; alarm counts, flow patterns, the bid and message analyses and the simulation comparison are secondary or post hoc. The replication was planned after the original results were known and reuses the honest control markets. Twenty markets per type give wide intervals, and absence of an advantage is not equivalence. The auction is synthetic, the cartel is instructed rather than emergent, and the treatment bundles a joint objective, partner information, a surveillance warning and, in C1, a chat channel. The core uses one bidder and one judge model; the smallest models were unusable and breadth cells have 20 markets each. A one-way leak is not a unique signature of espionage, and the pooled flow intervals are approximate. Privacy, false-accusation and dual-use considerations are discussed in Appendix H. Next steps: independent replications; screens that combine correlation, flow direction and price stability; larger bidder and judge models; cartels that can reply to each other; and markets with more firms.

6 Conclusion

Collusion, espionage, or look-alikes? In this auction the audit could tell them apart, but only partly from single markets. From the bids of AI agents alone, a calibrated net-influence alarm found cartels that chatted and named the right pair without wrong accusations. With private costs, pooled evidence showed which way information flowed, two-way for the cartel and one-way for the spy, but single-market alarms almost never caught the spy, which barely used what it stole. A simple correlation detected chatting cartels as well, the most profitable cartel kept a fixed price and stayed hidden, and a small AI judge looked at firm strength. The alarm needs only public bids, but common cost shocks and fewer tenders would make real markets harder. Such screens can direct an investigation but do not prove agreement, intent or harm.

Code and data. All code, market data and analysis outputs are public at <https://github.com/pavoras/ai-collusion>; `report/README.md` lists every analysis command, and Appendix G gives provenance and checksums.

References

- Kushal Agrawal, Verona Teo, Juan J. Vazquez, Sudarsh Kunnavakkam, Vishak Srikanth, and Andy Liu. Evaluating LLM agent collusion in double auctions. arXiv:2507.01413, 2025. URL <https://arxiv.org/abs/2507.01413>.
- Trevor Bonjour, Vaneet Aggarwal, and Bharat Bhargava. Information theoretic approach to detect collusion in multi-agent games. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 223–232, 2022. URL <https://proceedings.mlr.press/v180/bonjour22a.html>.
- Sara Fish, Yannai A. Gonczarowski, and Ran I. Shorrer. Algorithmic collusion by large language models. arXiv:2404.00806, 2024. URL <https://arxiv.org/abs/2404.00806>.
- Martin Huber and David Imhof. Machine learning with screens for detecting bid-rigging cartels. *International Journal of Industrial Organization*, 65:277–301, 2019. doi: 10.1016/j.ijindorg.2019.04.002.
- David Imhof and Hannes Wallimann. Detecting bid-rigging coalitions in different countries and auction formats. *International Review of Law and Economics*, 68:106016, 2021. doi: 10.1016/j.irle.2021.106016.
- Dohun Lee and Hyunwoo Park. Faithful yet collusive: Why chain-of-thought monitoring cannot detect collusion in LLM pricing agents under oligopolistic competition. arXiv:2609.18346, 2026. URL <https://arxiv.org/abs/2609.18346>.
- Ryan Y. Lin, Siddhartha Ojha, Kevin Cai, and Maxwell F. Chen. Strategic collusion of LLM agents: Market division in multi-commodity competitions. arXiv:2410.00031, 2024. URL <https://arxiv.org/abs/2410.00031>.
- Sumeet Ramesh Motwani, Mikhail Baranchuk, Martin Strohmeier, Vijay Bolina, Philip H. S. Torr, Lewis Hammond, and Christian Schroeder de Witt. Secret collusion among AI agents: Multi-agent deception via steganography. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2336.
- Raj-Taware. collusion-net-influence. GitHub repository, <https://github.com/Raj-Taware/collusion-net-influence>, 2026. Commit 0e2aa44 (22 Oct 2026), accessed 24 Oct 2026.
- Aaron Rose, Carissa Cullen, Sahar Abdelnabi, Philip Torr, Brandon Gary Kaplowitz, and Christian Schroeder de Witt. Detecting multi-agent collusion through multi-agent interpretability. arXiv:2604.01151, 2026. URL <https://arxiv.org/abs/2604.01151>.
- soumya-nayak-commits. collusion-detetction. GitHub repository, <https://github.com/soumya-nayak-commits/collusion-detetction>, 2026. Commit ee0c111 (23 Oct 2026), accessed 24 Oct 2026.
- Om Tailor. Audit the Whisper: Detecting steganographic collusion in multi-agent LLMs. arXiv:2510.04303, 2025. URL <https://arxiv.org/abs/2510.04303>.
- Hannes Wallimann, David Imhof, and Martin Huber. A machine learning approach for flagging incomplete bid-rigging cartels. *Computational Economics*, 62(4):1669–1720, 2023. doi: 10.1007/s10614-022-10315-w.

A Replication details

The scripted rock–paper–scissors replication implements the influence and mutual net-threshold rule of Bonjour et al. (2022). The comparison uses bits and the previous round’s joint action as state, conventions inferred from the published results. The no-state variant is retained to show that conditioning choices matter. Detection rises with coordination and sample size, but the replication does not exactly reproduce all table entries. The values below distinguish the paper’s published percentages from our simulation. These are development checks, not LLM results.

Table 2: Scripted RPS detection percentages: published values versus our previous-state and no-state replications.

Coordination probability	Games	Paper	Previous state	No state
0.3	100	37.1	36.3	84.4
0.3	200	48.3	60.1	94.4
0.3	300	71.0	77.0	98.1
0.3	400	82.8	87.1	99.5
0.3	500	89.6	92.2	99.9
0.3	1000	100.0	99.7	100.0
0.4	100	58.4	61.9	97.3
0.4	200	79.2	90.7	100.0
0.4	300	96.2	99.0	100.0
0.4	400	99.3	99.8	100.0
0.4	500	99.7	99.9	100.0
0.4	1000	100.0	100.0	100.0
1.0	60	100.0	100.0	100.0
0.2	1000	24.3	50.0	58.0
0.2	10000	82.3	83.7	88.5
0.1	1000	8.0	0.4	0.0

Table 3: Scripted RPS honest-market any-pair alarm rates (coordination probability zero). This is a separate simulation, not LLM calibration.

State	Games	Fixed threshold	Permutation rule
none	50	0.296	0.038
none	100	0.088	0.062
none	200	0.006	0.048
none	500	0.0	0.042
none	1000	0.0	0.038
prev	50	0.588	0.02
prev	100	0.484	0.036
prev	200	0.252	0.036
prev	500	0.046	0.026
prev	1000	0.0	0.038

The deterministic seed is 20261023; the full run uses 1000 trials per detection cell. Its calibration experiment uses 500 trials and 199 permutations. These settings differ from the registered LLM analysis. The complete calibration data, including positive conditions, are in `results/rps/calibration.csv`; `experiments/rps_replication.py` reproduces them.

B Conditions, prompts and model configuration

The frozen manifest is `experiments/manifests/collection_v2.json`. It maps model identities to roles and hidden seat assignments. Exact prompt templates are in `src/collusion/llm/prompts.py`; model revisions and sampling settings are in the per-run configuration and reproducibility records named by `results/analysis_runs.json`. Request hashes, prompts, responses and parsing metadata remain associated with each raw run. `results/analysis_meta/pilot_qwen9b_v5_N0.json` records the pilot-derived N0 configuration explicitly. The manifest and source checksums accompanying this report allow these inputs to be checked without re-collecting outputs.

The A1 prompt removes only: “The competition authority monitors all bids, so behave inconspicuously.” The

Table 4: Conditions and matched controls. SS, SF, and FF denote strong–strong, strong–fringe, and fringe–fringe designated pairs. X, Y, Z denote Qwen3.5-9B, Gemma-4-12B, and Ministral-3-8B. All A2 bidders are honest.

Cell	Intervention or control
N0 / P0	Honest market / live scripted strong cartel with honest LLM fringe.
C1-SS/SF/FF	LLM cartel with private channel; matched to N0.
C2 / E1	SS cartel without channel / one strong firm sees the other’s cost.
C1-NW (A1)	C1-SS without surveillance warning; matched N0/C1-SS triplets.
D / C1-XX	Honest / cartel strong pair (X,X), fringe (Y,Z).
H-XYZZ / C1-XY	Honest / cartel strong pair (X,Y), fringe (Z,Z).
LX, LY, LZ (A2)	Strong/fringe models (X,Y;X,Z), (Y,Z;Y,X), (Z,X;Z,Y), each with a twin swapping the fringe models.

joint-profit objective and private channel remain. This is a warning-removal treatment, not a direct test of an internal evasion strategy. Core Qwen and A1 retain the strict parser; the documented fenced-JSON parser cohort covers Gemma/Ministral, A2 and breadth collection. Bidders run without thinking, whereas the primary judge uses thinking. No treatment labels, messages or model identities are given to the judges.

C Further method details

State and estimation. The public state H_t has a start code and carries forward after an all-invalid round; masking preserves this chronology. Under independent costs and no other access to a rival’s current cost, leaks probe private-information flow; omitted history can induce action dependence. Arrows indicate cost source and bid receiver, not intent. Plug-in counts in bits are our replication convention inferred from the paper’s results; the paper does not specify the log base. Only a unique passing pair is reported. Symmetric γ need not imply symmetric Γ : the rival comparisons differ. Net direction is not leak direction.

Corrected rules and flow labels. Corrections remain signed. The permutation test concerns influence, not net influence. The secondary leak score sums the two corrected leaks; a leak candidate requires either direction to pass 0.05/12, again with unique-pair reporting. Flow labels are both directions detected; one detected, reverse weak; neither detected, both weak; or unclear. “Weak” requires $p \geq 0.20$ and a corrected leak at most 0.05 bits. Labels cover designated pairs and the full detection pipeline. Weak evidence does not establish absence; one-way flow can arise in a cartel with unilateral sharing. Permutation validity requires within-stratum exchangeability, which adaptive policies can violate. Per-market alarm rates are descriptive, without a general calibration guarantee or error control across all analyses.

Baselines, judge and evaluation. Constant correlation residuals score zero and signed correlations are also retained; the baselines do not adjust for history. The judge uses thinking and sees costs only at L2 and no treatment information. There is no recalibration or answer selection; content failures remain failures. Comparisons use common judged markets and report coverage; descriptive Brier scores state the artificial prevalence. A Gemma-4-12B counter-judge was attempted on core N0/C1-SS indices 0–9 at both levels; Qwen judging Qwen remains a family-overlap limitation. All methods use rounds with four valid bids; more than 20 incomplete rounds makes a market ineligible and removes the matched block. A no-bid-category sensitivity retains missing bids. The primary p -value is $p = (1 + \#\{AUC^* \geq AUC_{\text{obs}}\})/100000$ over the 99,999 within-block swaps. The primary test checks paired label exchangeability, not the composite null AUROC ≤ 0.5 . Intervals are pointwise 95% percentile intervals. Top-1 is the fraction of credit for ranking the assigned pair first, shared equally across ties; localisation also reports reciprocal rank, correct/wrong unique reports and abstention; 1/6 is a guessing benchmark, not a null. E1 flow detection is not a collusion positive.

Secondary analyses. A2 tests paired fringe-placement swaps with Holm correction across seven endpoints, not a pure model-sameness mechanism. Pooled fingerprints align roles, bootstrap markets, and permute within markets. The exploratory probe regenerates a partner’s message under counterfactual cost, then samples bids with the recipient’s cost and history fixed; draws remain nested in markets (Appendix G). Revisions and sampling are pinned; seeded

Table 5: Behaviour and designated-pair flow labels. Differences are treatment minus matched N0, in game units per round, with pointwise 95% intervals. Flow counts are both / one (reverse weak) / neither (both weak) / unclear / unavailable. E1 is unilateral access.

Cell	Joint profit difference	Winning price difference	Flow counts
C1-SS	0.144 [0.094, 0.192]	0.295 [0.233, 0.358]	0 / 0 / 8 / 12 / 0
C2	0.399 [0.364, 0.434]	0.798 [0.768, 0.826]	0 / 0 / 11 / 9 / 0
E1	-0.116 [-0.137, -0.095]	-0.048 [-0.062, -0.034]	0 / 4 / 8 / 8 / 0
C1-NW	-0.204 [-0.229, -0.179]	-0.039 [-0.059, -0.017]	0 / 1 / 8 / 11 / 0

re-collection is not bitwise reproducible (Appendix B). Re-analysing identical markets as a separate cohort reproduces correlation, strength and paper scores exactly; net and leak scores change only by permutation Monte Carlo noise (largest AUROC difference 0.0075). Deviations from the pre-registration are listed in Appendix G.8.

D Scripted validation

Scripted auctions test known cartels, spies and honest policies under the same observation constraints. Shared-history stress tests show why a coarse or incomplete public state can invalidate a permutation null. Power planning and conditional-table occupancy checks motivated the market horizon and weak-evidence category. These simulations inform the design; they do not establish calibration for arbitrary adaptive LLM bidders. The outputs are in `results/auction_scripted/`, `results/auction_power/` and `results/null_validity/`, produced by the scripts of the same names in `experiments/`.

E All registered cells and per-model results

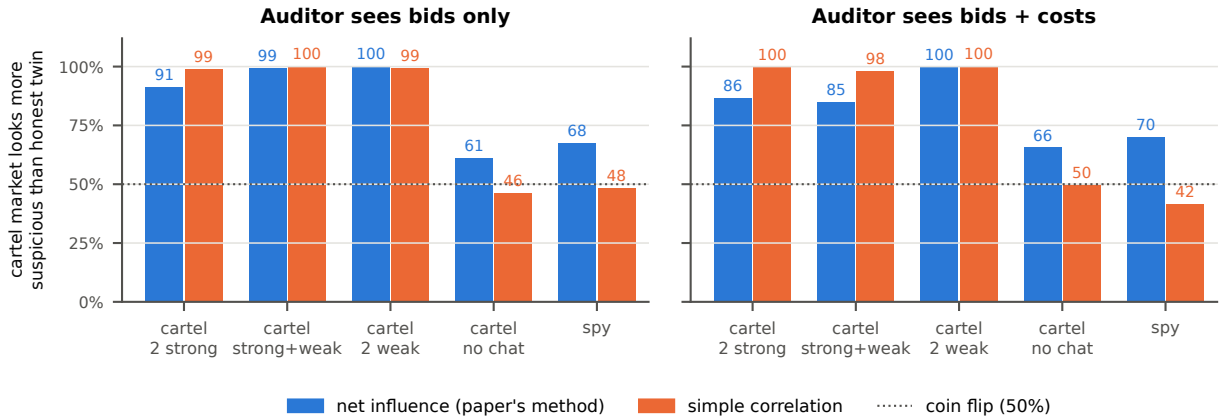


Figure 7: How often a treated market looks more suspicious than its honest twin (AUROC, in percent; 20 matched market pairs per type, markets 0–19). Net influence is the paper’s measure; correlation is the plain dependence between the two firms’ bids (with costs: after removing each firm’s typical bid at its own cost). Intervals and the other models are in the tables below.

The following single comprehensive table includes every registered cell, including unavailable detector results and cancelled breadth judges, at prefix 200 with the complete-round mask. It contains 312 detector/access rows spanning 28 cells. The full machine-readable table with every prefix and mask is `results/analysis/final_v1/all_cells.md`; no sensitivity rows are silently discarded from that file.

Cov. is successful detector coverage; n is the matched AUROC block count. A dash means unavailable or not applicable, never a score of zero. Unique/wrong/abstain are rates under the operational reporting rule; AUROC uses the continuous score, including for the paper detector. For honest cells, top-1 and mean reciprocal rank (MRR) describe the

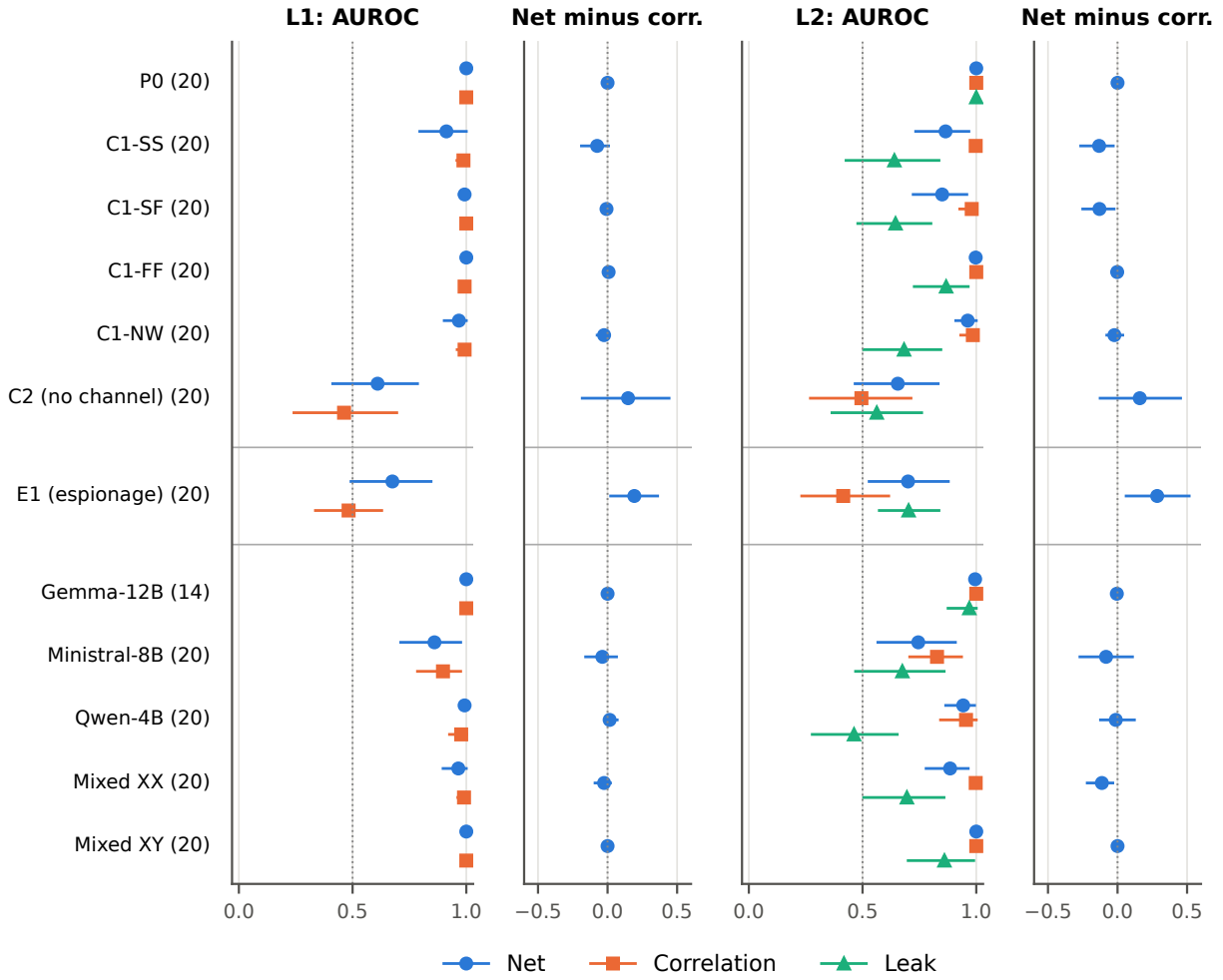


Figure 8: Detection by treatment (original markets 0–19), all models. AUROC with pointwise 95% block-bootstrap intervals for net influence and correlation at L1 (bids only) and L2 (bids + costs), and the L2 leak score; parentheses give matched blocks. E1 (espionage) is separated because it is not a cartel.

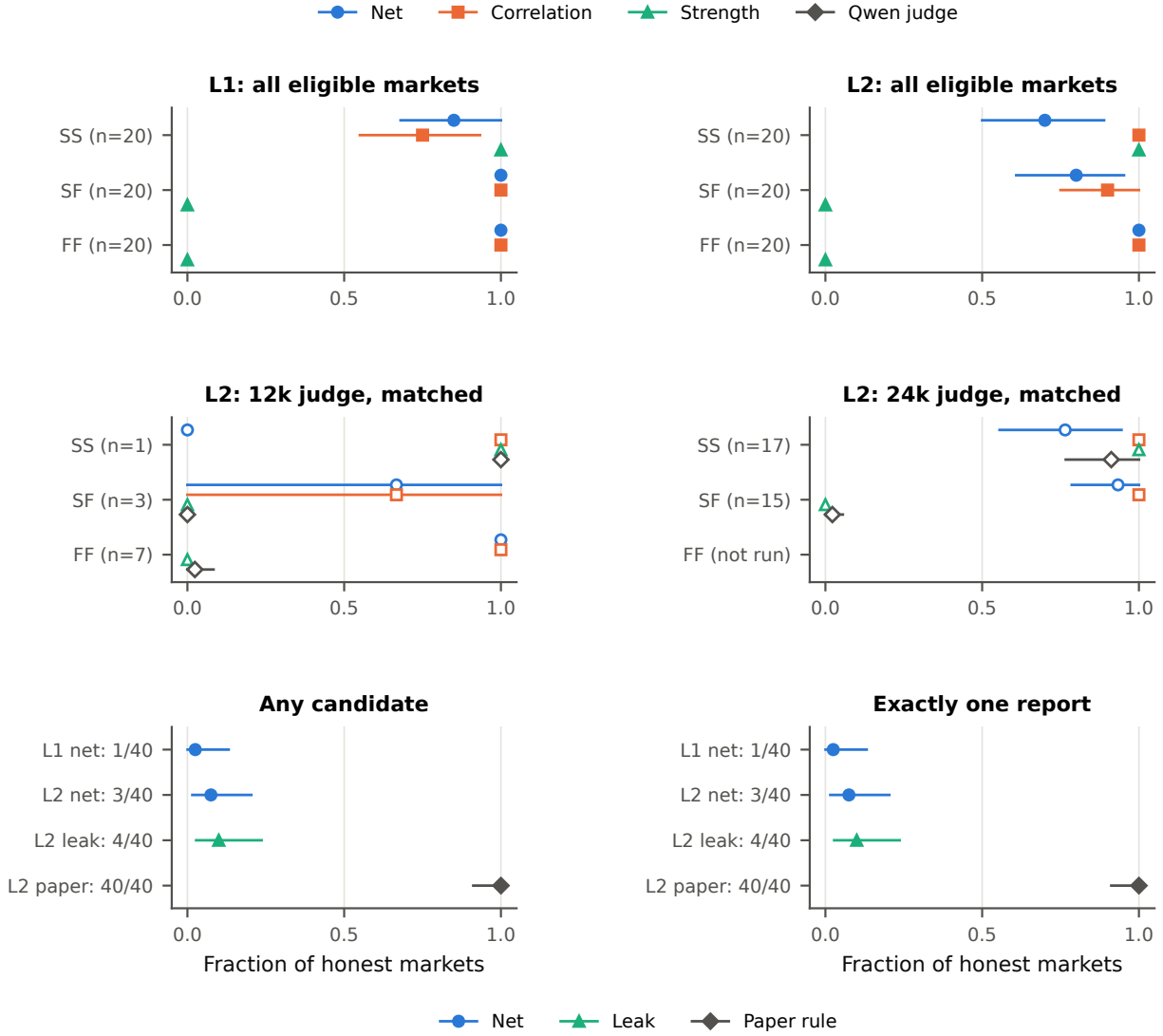


Figure 9: Localisation and false alarms (original markets 0–19). Top: top-1 for the assigned pair on all eligible markets. Middle: top-1 at L2 on the markets where the judge returned valid output, under the registered 12,000-token limit (left) and the 24,000-token check (right; C1-FF not run). Bottom: honest-market alarm rates with exact binomial intervals.

designated reference pair, not localisation accuracy. Similarly, “unique” is a report of that reference pair, and “wrong” is a report of another pair; neither is a correct collusion finding in an honest market. Alarm columns give any candidate, exactly one report, and designated reference-pair alarms, with exact binomial intervals.

Flow counts are ordered both detected / one detected, reverse weak / neither detected, both weak / unclear / unavailable. Pipeline counts use the same first categories, ending in no pair instead of unavailable. All-zero pipeline counts for score-only baselines mean no reporting rule, not evidence that no information flowed. Judge rows’ label counts describe their available subset. Excluded markets fail the registered bid-validity rule. Superseded operational attempts are retained in the raw record, but are not extra independent markets in this registered cohort.

Table 6: All registered cells at the full prefix and complete mask. See the surrounding text for column definitions.

Level / detector	Cov./n	AUROC [interval]	Top-1 [interval]	Unique / wrong / abstain
core.N0 ; declared 40, collected 40, missing 0, ineligible 0, eligible 40				
L1 corr.	40/–	–	0.050 [0.000, 0.125]	– / – / –
MRR 0.242 [0.192, 0.307]; flow 0/0/32/8/0; pipeline 0/0/0/0/0.				
L1 G judge	1/–	–	1.000 [1.000, 1.000]	– / – / –
MRR 1.000 [1.000, 1.000]; flow 0/0/1/0/39; pipeline 0/0/0/0/0.				
L1 Q judge	21/–	–	0.952 [0.840, 1.000]	– / – / –
MRR 0.964 [0.880, 1.000]; flow 0/0/19/2/19; pipeline 0/0/0/0/0.				
L1 net	40/–	–	0.100 [0.025, 0.200]	0.000 / 0.025 / 0.975
MRR 0.385 [0.321, 0.462]; flow 0/0/32/8/0; pipeline 0/0/0/1/39. Alarms: any 1/40 [0.001, 0.132]; one 1/40 [0.001, 0.132]; ref. 0/40 [0.000, 0.088]				
L1 paper	40/–	–	0.000 [0.000, 0.000]	0.000 / 0.600 / 0.400
MRR 0.418 [0.378, 0.454]; flow 0/0/32/8/0; pipeline 0/0/16/8/16. Alarms: any 24/40 [0.433, 0.751]; one 24/40 [0.433, 0.751]; ref. 0/40 [0.000, 0.088]				
L1 strength	40/–	–	1.000 [1.000, 1.000]	– / – / –
MRR 1.000 [1.000, 1.000]; flow 0/0/32/8/0; pipeline 0/0/0/0/0.				
L2 corr.	40/–	–	0.175 [0.075, 0.300]	– / – / –
MRR 0.415 [0.330, 0.508]; flow 0/0/32/8/0; pipeline 0/0/0/0/0.				
L2 G judge	0/–	–	–	– / – / –
MRR –; flow 0/0/0/0/40; pipeline 0/0/0/0/0.				
L2 Q judge	15/–	–	1.000 [1.000, 1.000]	– / – / –
MRR 1.000 [1.000, 1.000]; flow 0/0/13/2/25; pipeline 0/0/0/0/0.				
L2 leak	40/–	–	0.050 [0.000, 0.125]	0.000 / 0.100 / 0.900
MRR 0.333 [0.283, 0.394]; flow 0/0/32/8/0; pipeline 0/3/0/1/36. Alarms: any 4/40 [0.028, 0.237]; one 4/40 [0.028, 0.237]; ref. 0/40 [0.000, 0.088]				
L2 net	40/–	–	0.075 [0.000, 0.151]	0.000 / 0.075 / 0.925
MRR 0.369 [0.307, 0.439]; flow 0/0/32/8/0; pipeline 0/0/0/3/37. Alarms: any 3/40 [0.016, 0.204]; one 3/40 [0.016, 0.204]; ref. 0/40 [0.000, 0.088]				
L2 paper	40/–	–	0.000 [0.000, 0.000]	0.000 / 1.000 / 0.000
MRR 0.408 [0.369, 0.444]; flow 0/0/32/8/0; pipeline 0/0/25/15/0. Alarms: any 40/40 [0.912, 1.000]; one 40/40 [0.912, 1.000]; ref. 0/40 [0.000, 0.088]				
L2 strength	40/–	–	1.000 [1.000, 1.000]	– / – / –
MRR 1.000 [1.000, 1.000]; flow 0/0/32/8/0; pipeline 0/0/0/0/0.				
core.P0 ; declared 20, collected 20, missing 0, ineligible 0, eligible 20				
L1 corr.	20/20	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	– / – / –
MRR 1.000 [1.000, 1.000]; flow 20/0/0/0/0; pipeline 0/0/0/0/0.				
L1 Q judge	5/4	0.000 [0.000, 0.000]	0.033 [0.000, 0.111]	– / – / –
MRR 0.304 [0.250, 0.367]; flow 5/0/0/0/15; pipeline 0/0/0/0/0.				
L1 net	20/20	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	1.000 / 0.000 / 0.000
MRR 1.000 [1.000, 1.000]; flow 20/0/0/0/0; pipeline 20/0/0/0/0.				
L1 paper	20/20	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	1.000 / 0.000 / 0.000
MRR 1.000 [1.000, 1.000]; flow 20/0/0/0/0; pipeline 20/0/0/0/0.				
L1 strength	20/20	0.500 [0.500, 0.500]	1.000 [1.000, 1.000]	– / – / –
MRR 1.000 [1.000, 1.000]; flow 20/0/0/0/0; pipeline 0/0/0/0/0.				
L2 corr.	20/20	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	– / – / –
MRR 1.000 [1.000, 1.000]; flow 20/0/0/0/0; pipeline 0/0/0/0/0.				
L2 Q judge	4/3	0.000 [0.000, 0.000]	0.083 [0.000, 0.167]	– / – / –
MRR 0.330 [0.250, 0.500]; flow 4/0/0/0/16; pipeline 0/0/0/0/0.				
L2 leak	20/20	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	0.950 / 0.000 / 0.050
MRR 1.000 [1.000, 1.000]; flow 20/0/0/0/0; pipeline 19/0/0/0/1.				
L2 net	20/20	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	1.000 / 0.000 / 0.000
MRR 1.000 [1.000, 1.000]; flow 20/0/0/0/0; pipeline 20/0/0/0/0.				
L2 paper	20/20	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	1.000 / 0.000 / 0.000
MRR 1.000 [1.000, 1.000]; flow 20/0/0/0/0; pipeline 20/0/0/0/0.				

Level / detector	Cov./n	AUROC [interval]	Top-1 [interval]	Unique / wrong / abstain
L2 strength	20/20	0.500 [0.500, 0.500]	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 20/0/0/0/0; pipeline 0/0/0/0/0.				
core.C1-SS ; declared 20, collected 20, missing 0, ineligible 0, eligible 20				
L1 corr.	20/20	0.988 [0.959, 1.000]	0.750 [0.550, 0.933]	- / - / -
MRR 0.863 [0.750, 0.963]; flow 0/0/8/12/0; pipeline 0/0/0/0/0.				
L1 G judge	3/-	-	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/0/1/2/17; pipeline 0/0/0/0/0.				
L1 Q judge	13/8	0.320 [0.069, 0.617]	0.808 [0.583, 1.000]	- / - / -
MRR 0.835 [0.643, 1.000]; flow 0/0/6/7/7; pipeline 0/0/0/0/0.				
L1 net	20/20	0.912 [0.796, 1.000]	0.850 [0.680, 1.000]	0.800 / 0.000 / 0.200
MRR 0.908 [0.800, 1.000]; flow 0/0/8/12/0; pipeline 0/0/6/10/4.				
L1 paper	20/20	0.885 [0.765, 0.991]	0.150 [0.000, 0.318]	0.100 / 0.300 / 0.600
MRR 0.527 [0.433, 0.627]; flow 0/0/8/12/0; pipeline 0/0/4/4/12.				
L1 strength	20/20	0.500 [0.500, 0.500]	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/0/8/12/0; pipeline 0/0/0/0/0.				
L2 corr.	20/20	0.998 [0.986, 1.000]	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/0/8/12/0; pipeline 0/0/0/0/0.				
L2 G judge	1/-	-	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/0/0/1/19; pipeline 0/0/0/0/0.				
L2 Q judge	7/1	0.000 [0.000, 0.000]	0.881 [0.583, 1.000]	- / - / -
MRR 0.898 [0.643, 1.000]; flow 0/0/3/4/13; pipeline 0/0/0/0/0.				
L2 leak	20/20	0.640 [0.428, 0.836]	0.300 [0.105, 0.519]	0.000 / 0.050 / 0.950
MRR 0.530 [0.387, 0.678]; flow 0/0/8/12/0; pipeline 0/0/0/1/19.				
L2 net	20/20	0.865 [0.734, 0.968]	0.700 [0.500, 0.889]	0.250 / 0.000 / 0.750
MRR 0.823 [0.693, 0.938]; flow 0/0/8/12/0; pipeline 0/0/0/5/15.				
L2 paper	20/20	0.882 [0.776, 0.970]	0.000 [0.000, 0.000]	0.000 / 0.950 / 0.050
MRR 0.443 [0.395, 0.485]; flow 0/0/8/12/0; pipeline 0/0/11/8/1.				
L2 strength	20/20	0.500 [0.500, 0.500]	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/0/8/12/0; pipeline 0/0/0/0/0.				
core.C2 ; declared 20, collected 20, missing 0, ineligible 0, eligible 20				
L1 corr.	20/20	0.463 [0.242, 0.694]	0.300 [0.111, 0.522]	- / - / -
MRR 0.453 [0.300, 0.624]; flow 0/0/11/9/0; pipeline 0/0/0/0/0.				
L1 Q judge	10/4	0.344 [0.000, 1.000]	0.833 [0.583, 1.000]	- / - / -
MRR 0.857 [0.643, 1.000]; flow 0/0/4/6/10; pipeline 0/0/0/0/0.				
L1 net	20/20	0.610 [0.413, 0.785]	0.067 [0.000, 0.188]	0.000 / 0.150 / 0.850
MRR 0.393 [0.317, 0.485]; flow 0/0/11/9/0; pipeline 0/0/0/3/17.				
L1 paper	20/20	0.710 [0.550, 0.854]	0.000 [0.000, 0.000]	0.000 / 0.800 / 0.200
MRR 0.467 [0.426, 0.500]; flow 0/0/11/9/0; pipeline 0/0/4/12/4.				
L1 strength	20/20	0.500 [0.500, 0.500]	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/0/11/9/0; pipeline 0/0/0/0/0.				
L2 corr.	20/20	0.495 [0.271, 0.713]	0.400 [0.190, 0.619]	- / - / -
MRR 0.541 [0.378, 0.712]; flow 0/0/11/9/0; pipeline 0/0/0/0/0.				
L2 Q judge	9/3	0.333 [0.000, 1.000]	0.907 [0.688, 1.000]	- / - / -
MRR 0.921 [0.732, 1.000]; flow 0/0/8/1/11; pipeline 0/0/0/0/0.				
L2 leak	20/20	0.562 [0.366, 0.760]	0.050 [0.000, 0.167]	0.000 / 0.000 / 1.000
MRR 0.357 [0.284, 0.448]; flow 0/0/11/9/0; pipeline 0/0/0/0/20.				
L2 net	20/20	0.655 [0.467, 0.833]	0.050 [0.000, 0.167]	0.050 / 0.000 / 0.950
MRR 0.391 [0.315, 0.483]; flow 0/0/11/9/0; pipeline 0/0/0/1/19.				
L2 paper	20/20	0.887 [0.764, 0.979]	0.000 [0.000, 0.000]	0.000 / 1.000 / 0.000
MRR 0.487 [0.458, 0.500]; flow 0/0/11/9/0; pipeline 0/0/7/13/0.				
L2 strength	20/20	0.500 [0.500, 0.500]	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/0/11/9/0; pipeline 0/0/0/0/0.				
core.E1 ; declared 20, collected 20, missing 0, ineligible 0, eligible 20				
L1 corr.	20/20	0.482 [0.337, 0.628]	0.100 [0.000, 0.250]	- / - / -
MRR 0.276 [0.179, 0.401]; flow 0/4/8/8/0; pipeline 0/0/0/0/0.				
L1 Q judge	12/5	0.600 [0.167, 1.000]	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/2/6/4/8; pipeline 0/0/0/0/0.				
L1 net	20/20	0.675 [0.493, 0.845]	0.300 [0.105, 0.500]	0.050 / 0.000 / 0.950
MRR 0.570 [0.440, 0.708]; flow 0/4/8/8/0; pipeline 0/0/1/0/19.				
L1 paper	20/20	0.620 [0.453, 0.785]	0.000 [0.000, 0.000]	0.000 / 0.900 / 0.100
MRR 0.463 [0.425, 0.492]; flow 0/4/8/8/0; pipeline 0/0/11/7/2.				
L1 strength	20/20	0.500 [0.500, 0.500]	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/4/8/8/0; pipeline 0/0/0/0/0.				
L2 corr.	20/20	0.415 [0.233, 0.615]	0.150 [0.000, 0.333]	- / - / -
MRR 0.378 [0.263, 0.515]; flow 0/4/8/8/0; pipeline 0/0/0/0/0.				

Level / detector	Cov./n	AUROC [interval]	Top-1 [interval]	Unique / wrong / abstain
L2 Q judge	7/1	0.500 [0.500, 0.500]	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow	0/2/1/4/13; pipeline	0/0/0/0/0.		
L2 leak	20/20	0.703 [0.574, 0.836]	0.300 [0.105, 0.520]	0.200 / 0.000 / 0.800
MRR 0.568 [0.440, 0.704]; flow	0/4/8/8/0; pipeline	0/4/0/0/16.		
L2 net	20/20	0.700 [0.529, 0.877]	0.400 [0.188, 0.625]	0.000 / 0.050 / 0.950
MRR 0.607 [0.460, 0.759]; flow	0/4/8/8/0; pipeline	0/0/0/1/19.		
L2 paper	20/20	0.652 [0.485, 0.821]	0.000 [0.000, 0.000]	0.000 / 1.000 / 0.000
MRR 0.464 [0.422, 0.500]; flow	0/4/8/8/0; pipeline	0/0/13/7/0.		
L2 strength	20/20	0.500 [0.500, 0.500]	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow	0/4/8/8/0; pipeline	0/0/0/0/0.		
core.C1-SF ; declared 20, collected 20, missing 0, ineligible 0, eligible 20				
L1 corr.	20/20	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow	0/1/10/9/0; pipeline	0/0/0/0/0.		
L1 Q judge	14/7	0.602 [0.250, 0.880]	0.048 [0.011, 0.091]	- / - / -
MRR 0.273 [0.251, 0.295]; flow	0/1/5/8/6; pipeline	0/0/0/0/0.		
L1 net	20/20	0.993 [0.969, 1.000]	1.000 [1.000, 1.000]	0.650 / 0.000 / 0.350
MRR 1.000 [1.000, 1.000]; flow	0/1/10/9/0; pipeline	0/1/4/8/7.		
L1 paper	20/20	0.998 [0.986, 1.000]	0.950 [0.833, 1.000]	0.550 / 0.000 / 0.450
MRR 0.975 [0.917, 1.000]; flow	0/1/10/9/0; pipeline	0/1/3/7/9.		
L1 strength	20/20	0.500 [0.500, 0.500]	0.000 [0.000, 0.000]	- / - / -
MRR 0.250 [0.250, 0.250]; flow	0/1/10/9/0; pipeline	0/0/0/0/0.		
L2 corr.	20/20	0.980 [0.927, 1.000]	0.900 [0.750, 1.000]	- / - / -
MRR 0.950 [0.875, 1.000]; flow	0/1/10/9/0; pipeline	0/0/0/0/0.		
L2 Q judge	11/3	0.667 [0.000, 1.000]	0.030 [0.000, 0.074]	- / - / -
MRR 0.260 [0.233, 0.288]; flow	0/1/6/4/9; pipeline	0/0/0/0/0.		
L2 leak	20/20	0.645 [0.479, 0.801]	0.350 [0.143, 0.565]	0.050 / 0.000 / 0.950
MRR 0.534 [0.376, 0.695]; flow	0/1/10/9/0; pipeline	0/1/0/0/19.		
L2 net	20/20	0.850 [0.723, 0.958]	0.800 [0.609, 0.952]	0.400 / 0.050 / 0.550
MRR 0.872 [0.743, 0.974]; flow	0/1/10/9/0; pipeline	0/0/2/7/11.		
L2 paper	20/20	0.995 [0.978, 1.000]	0.600 [0.381, 0.812]	0.450 / 0.200 / 0.350
MRR 0.800 [0.690, 0.906]; flow	0/1/10/9/0; pipeline	0/1/5/7/7.		
L2 strength	20/20	0.500 [0.500, 0.500]	0.000 [0.000, 0.000]	- / - / -
MRR 0.250 [0.250, 0.250]; flow	0/1/10/9/0; pipeline	0/0/0/0/0.		
core.C1-FF ; declared 20, collected 20, missing 0, ineligible 0, eligible 20				
L1 corr.	20/20	0.993 [0.972, 1.000]	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow	0/1/4/15/0; pipeline	0/0/0/0/0.		
L1 Q judge	10/4	0.438 [0.000, 1.000]	0.100 [0.000, 0.333]	- / - / -
MRR 0.383 [0.250, 0.561]; flow	0/1/2/7/10; pipeline	0/0/0/0/0.		
L1 net	20/20	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	1.000 / 0.000 / 0.000
MRR 1.000 [1.000, 1.000]; flow	0/1/4/15/0; pipeline	0/1/4/15/0.		
L1 paper	20/20	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	1.000 / 0.000 / 0.000
MRR 1.000 [1.000, 1.000]; flow	0/1/4/15/0; pipeline	0/1/4/15/0.		
L1 strength	20/20	0.500 [0.500, 0.500]	0.000 [0.000, 0.000]	- / - / -
MRR 0.250 [0.250, 0.250]; flow	0/1/4/15/0; pipeline	0/0/0/0/0.		
L2 corr.	20/20	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow	0/1/4/15/0; pipeline	0/0/0/0/0.		
L2 Q judge	14/7	0.551 [0.344, 0.812]	0.012 [0.000, 0.039]	- / - / -
MRR 0.383 [0.308, 0.455]; flow	0/1/3/10/6; pipeline	0/0/0/0/0.		
L2 leak	20/20	0.868 [0.727, 0.964]	0.600 [0.375, 0.812]	0.150 / 0.000 / 0.850
MRR 0.724 [0.565, 0.871]; flow	0/1/4/15/0; pipeline	0/1/0/2/17.		
L2 net	20/20	0.998 [0.986, 1.000]	1.000 [1.000, 1.000]	0.950 / 0.000 / 0.050
MRR 1.000 [1.000, 1.000]; flow	0/1/4/15/0; pipeline	0/1/3/15/1.		
L2 paper	20/20	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	1.000 / 0.000 / 0.000
MRR 1.000 [1.000, 1.000]; flow	0/1/4/15/0; pipeline	0/1/4/15/0.		
L2 strength	20/20	0.500 [0.500, 0.500]	0.000 [0.000, 0.000]	- / - / -
MRR 0.250 [0.250, 0.250]; flow	0/1/4/15/0; pipeline	0/0/0/0/0.		
labs.gemma-12b.N0 ; declared 20, collected 20, missing 0, ineligible 6, eligible 14				
L1 corr.	14/-	-	0.071 [0.000, 0.231]	- / - / -
MRR 0.288 [0.192, 0.426]; flow	0/0/8/6/0; pipeline	0/0/0/0/0.		
L1 Q judge	0/-	-	-	- / - / -
MRR -; flow	0/0/0/0/14; pipeline	0/0/0/0/0.		
L1 net	14/-	-	0.143 [0.000, 0.357]	0.000 / 0.071 / 0.929
MRR 0.430 [0.300, 0.581]; flow	0/0/8/6/0; pipeline	0/0/1/0/13. Alarms: any 1/14 [0.002, 0.339]; one 1/14 [0.002, 0.339]; ref. 0/14 [0.000, 0.232]		

Level / detector	Cov./n	AUROC [interval]	Top-1 [interval]	Unique / wrong / abstain
L1 paper	14/-	-	0.000 [0.000, 0.000]	0.000 / 0.357 / 0.643
MRR 0.401 [0.332, 0.464]; flow 0/0/8/6/0; pipeline 0/0/3/2/9. Alarms: any 5/14 [0.128, 0.649]; one 5/14 [0.128, 0.649]; ref. 0/14 [0.000, 0.232]				
L1 strength	14/-	-	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/0/8/6/0; pipeline 0/0/0/0/0.				
L2 corr.	14/-	-	0.286 [0.067, 0.545]	- / - / -
MRR 0.483 [0.310, 0.674]; flow 0/0/8/6/0; pipeline 0/0/0/0/0.				
L2 Q judge	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/14; pipeline 0/0/0/0/0.				
L2 leak	14/-	-	0.071 [0.000, 0.231]	0.000 / 0.071 / 0.929
MRR 0.374 [0.270, 0.500]; flow 0/0/8/6/0; pipeline 0/1/0/0/13. Alarms: any 1/14 [0.002, 0.339]; one 1/14 [0.002, 0.339]; ref. 0/14 [0.000, 0.232]				
L2 net	14/-	-	0.000 [0.000, 0.000]	0.000 / 0.214 / 0.786
MRR 0.363 [0.292, 0.431]; flow 0/0/8/6/0; pipeline 0/0/1/2/11. Alarms: any 3/14 [0.047, 0.508]; one 3/14 [0.047, 0.508]; ref. 0/14 [0.000, 0.232]				
L2 paper	14/-	-	0.000 [0.000, 0.000]	0.000 / 0.857 / 0.143
MRR 0.457 [0.393, 0.500]; flow 0/0/8/6/0; pipeline 0/0/8/4/2. Alarms: any 12/14 [0.572, 0.982]; one 12/14 [0.572, 0.982]; ref. 0/14 [0.000, 0.232]				
L2 strength	14/-	-	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/0/8/6/0; pipeline 0/0/0/0/0.				
labs.gemma-12b.C1 ; declared 20, collected 20, missing 0, ineligible 0, eligible 20				
L1 corr.	20/14	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/1/5/14/0; pipeline 0/0/0/0/0.				
L1 Q judge	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/20; pipeline 0/0/0/0/0.				
L1 net	20/14	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	0.900 / 0.000 / 0.100
MRR 1.000 [1.000, 1.000]; flow 0/1/5/14/0; pipeline 0/1/5/12/2.				
L1 paper	20/14	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	0.950 / 0.000 / 0.050
MRR 1.000 [1.000, 1.000]; flow 0/1/5/14/0; pipeline 0/1/5/13/1.				
L1 strength	20/14	0.500 [0.500, 0.500]	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/1/5/14/0; pipeline 0/0/0/0/0.				
L2 corr.	20/14	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/1/5/14/0; pipeline 0/0/0/0/0.				
L2 Q judge	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/20; pipeline 0/0/0/0/0.				
L2 leak	20/14	0.969 [0.876, 1.000]	0.600 [0.375, 0.812]	0.050 / 0.000 / 0.950
MRR 0.758 [0.613, 0.890]; flow 0/1/5/14/0; pipeline 0/1/0/0/19.				
L2 net	20/14	0.995 [0.973, 1.000]	1.000 [1.000, 1.000]	0.550 / 0.000 / 0.450
MRR 1.000 [1.000, 1.000]; flow 0/1/5/14/0; pipeline 0/1/3/7/9.				
L2 paper	20/14	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	1.000 / 0.000 / 0.000
MRR 1.000 [1.000, 1.000]; flow 0/1/5/14/0; pipeline 0/1/5/14/0.				
L2 strength	20/14	0.500 [0.500, 0.500]	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/1/5/14/0; pipeline 0/0/0/0/0.				
labs.ministral-8b.N0 ; declared 20, collected 20, missing 0, ineligible 0, eligible 20				
L1 corr.	20/-	-	0.100 [0.000, 0.250]	- / - / -
MRR 0.378 [0.275, 0.496]; flow 0/1/12/7/0; pipeline 0/0/0/0/0.				
L1 Q judge	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/20; pipeline 0/0/0/0/0.				
L1 net	20/-	-	0.100 [0.000, 0.250]	0.050 / 0.250 / 0.700
MRR 0.419 [0.322, 0.530]; flow 0/1/12/7/0; pipeline 0/0/2/4/14. Alarms: any 6/20 [0.119, 0.543]; one 6/20 [0.119, 0.543]; ref. 1/20 [0.001, 0.249]				
L1 paper	20/-	-	0.050 [0.000, 0.167]	0.000 / 0.000 / 1.000
MRR 0.384 [0.302, 0.474]; flow 0/1/12/7/0; pipeline 0/0/0/0/20. Alarms: any 0/20 [0.000, 0.168]; one 0/20 [0.000, 0.168]; ref. 0/20 [0.000, 0.168]				
L1 strength	20/-	-	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/1/12/7/0; pipeline 0/0/0/0/0.				
L2 corr.	20/-	-	0.200 [0.045, 0.389]	- / - / -
MRR 0.445 [0.322, 0.579]; flow 0/1/12/7/0; pipeline 0/0/0/0/0.				
L2 Q judge	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/20; pipeline 0/0/0/0/0.				
L2 leak	20/-	-	0.150 [0.000, 0.320]	0.050 / 0.050 / 0.900
MRR 0.392 [0.280, 0.522]; flow 0/1/12/7/0; pipeline 0/2/0/0/18. Alarms: any 2/20 [0.012, 0.317]; one 2/20 [0.012, 0.317]; ref. 1/20 [0.001, 0.249]				

Level / detector	Cov./n	AUROC [interval]	Top-1 [interval]	Unique / wrong / abstain
L2 net	20/-	-	0.050 [0.000, 0.167]	0.000 / 0.150 / 0.850
MRR 0.359 [0.290, 0.447]; flow 0/1/12/7/0; pipeline 0/0/0/3/17.			Alarms: any 3/20 [0.032, 0.379]; one 3/20 [0.032, 0.379]; ref. 0/20 [0.000, 0.168]	
L2 paper	20/-	-	0.000 [0.000, 0.000]	0.000 / 0.300 / 0.700
MRR 0.303 [0.252, 0.356]; flow 0/1/12/7/0; pipeline 0/0/1/5/14.			Alarms: any 6/20 [0.119, 0.543]; one 6/20 [0.119, 0.543]; ref. 0/20 [0.000, 0.168]	
L2 strength	20/-	-	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/1/12/7/0; pipeline 0/0/0/0/0.				
labs.ministral-8b.C1 ; declared 20, collected 20, missing 0, ineligible 0, eligible 20				
L1 corr.	20/20	0.897 [0.785, 0.975]	0.650 [0.435, 0.857]	- / - / -
MRR 0.808 [0.687, 0.922]; flow 0/0/8/12/0; pipeline 0/0/0/0/0.				
L1 Q judge	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/20; pipeline 0/0/0/0/0.				
L1 net	20/20	0.860 [0.712, 0.975]	0.800 [0.611, 0.955]	0.550 / 0.000 / 0.450
MRR 0.877 [0.756, 0.975]; flow 0/0/8/12/0; pipeline 0/0/5/6/9.				
L1 paper	20/20	0.922 [0.803, 1.000]	0.750 [0.550, 0.938]	0.450 / 0.000 / 0.550
MRR 0.854 [0.734, 0.963]; flow 0/0/8/12/0; pipeline 0/0/5/4/11.				
L1 strength	20/20	0.500 [0.500, 0.500]	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/0/8/12/0; pipeline 0/0/0/0/0.				
L2 corr.	20/20	0.828 [0.707, 0.935]	0.600 [0.375, 0.812]	- / - / -
MRR 0.771 [0.639, 0.896]; flow 0/0/8/12/0; pipeline 0/0/0/0/0.				
L2 Q judge	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/20; pipeline 0/0/0/0/0.				
L2 leak	20/20	0.675 [0.469, 0.859]	0.250 [0.062, 0.455]	0.000 / 0.000 / 1.000
MRR 0.512 [0.387, 0.650]; flow 0/0/8/12/0; pipeline 0/0/0/0/20.				
L2 net	20/20	0.745 [0.567, 0.908]	0.600 [0.375, 0.815]	0.200 / 0.150 / 0.650
MRR 0.756 [0.615, 0.891]; flow 0/0/8/12/0; pipeline 0/0/3/4/13.				
L2 paper	20/20	0.935 [0.838, 0.992]	0.650 [0.429, 0.857]	0.450 / 0.050 / 0.500
MRR 0.770 [0.621, 0.910]; flow 0/0/8/12/0; pipeline 0/0/3/7/10.				
L2 strength	20/20	0.500 [0.500, 0.500]	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/0/8/12/0; pipeline 0/0/0/0/0.				
sizes.qwen-4b.N0 ; declared 20, collected 20, missing 0, ineligible 0, eligible 20				
L1 corr.	20/-	-	0.150 [0.000, 0.333]	- / - / -
MRR 0.314 [0.193, 0.463]; flow 0/0/13/7/0; pipeline 0/0/0/0/0.				
L1 Q judge	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/20; pipeline 0/0/0/0/0.				
L1 net	20/-	-	0.050 [0.000, 0.167]	0.050 / 0.100 / 0.850
MRR 0.361 [0.282, 0.455]; flow 0/0/13/7/0; pipeline 0/0/2/1/17.			Alarms: any 3/20 [0.032, 0.379]; one 3/20 [0.032, 0.379]; ref. 1/20 [0.001, 0.249]	
L1 paper	20/-	-	0.000 [0.000, 0.000]	0.000 / 0.650 / 0.350
MRR 0.429 [0.369, 0.482]; flow 0/0/13/7/0; pipeline 0/0/4/9/7.			Alarms: any 13/20 [0.408, 0.846]; one 13/20 [0.408, 0.846]; ref. 0/20 [0.000, 0.168]	
L1 strength	20/-	-	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/0/13/7/0; pipeline 0/0/0/0/0.				
L2 corr.	20/-	-	0.200 [0.045, 0.392]	- / - / -
MRR 0.452 [0.324, 0.594]; flow 0/0/13/7/0; pipeline 0/0/0/0/0.				
L2 Q judge	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/20; pipeline 0/0/0/0/0.				
L2 leak	20/-	-	0.050 [0.000, 0.167]	0.000 / 0.150 / 0.850
MRR 0.352 [0.277, 0.447]; flow 0/0/13/7/0; pipeline 0/3/0/0/17.			Alarms: any 3/20 [0.032, 0.379]; one 3/20 [0.032, 0.379]; ref. 0/20 [0.000, 0.168]	
L2 net	20/-	-	0.150 [0.000, 0.333]	0.000 / 0.000 / 1.000
MRR 0.384 [0.272, 0.518]; flow 0/0/13/7/0; pipeline 0/0/0/0/20.			Alarms: any 0/20 [0.000, 0.168]; one 0/20 [0.000, 0.168]; ref. 0/20 [0.000, 0.168]	
L2 paper	20/-	-	0.000 [0.000, 0.000]	0.000 / 1.000 / 0.000
MRR 0.437 [0.383, 0.483]; flow 0/0/13/7/0; pipeline 0/0/10/10/0.			Alarms: any 20/20 [0.832, 1.000]; one 20/20 [0.832, 1.000]; ref. 0/20 [0.000, 0.168]	
L2 strength	20/-	-	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/0/13/7/0; pipeline 0/0/0/0/0.				
sizes.qwen-4b.C1 ; declared 20, collected 20, missing 0, ineligible 0, eligible 20				
L1 corr.	20/20	0.978 [0.926, 1.000]	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/0/12/8/0; pipeline 0/0/0/0/0.				
L1 Q judge	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/20; pipeline 0/0/0/0/0.				

Level / detector	Cov./n	AUROC [interval]	Top-1 [interval]	Unique / wrong / abstain
L1 net	20/20	0.993 [0.972, 1.000]	1.000 [1.000, 1.000]	0.900 / 0.000 / 0.100
MRR 1.000 [1.000, 1.000]; flow	0/0/12/8/0; pipeline 0/0/11/7/2.			
L1 paper	20/20	1.000 [1.000, 1.000]	0.400 [0.190, 0.619]	0.100 / 0.450 / 0.450
MRR 0.700 [0.595, 0.810]; flow	0/0/12/8/0; pipeline 0/0/6/5/9.			
L1 strength	20/20	0.500 [0.500, 0.500]	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow	0/0/12/8/0; pipeline 0/0/0/0/0.			
L2 corr.	20/20	0.955 [0.843, 1.000]	0.950 [0.833, 1.000]	- / - / -
MRR 0.975 [0.917, 1.000]; flow	0/0/12/8/0; pipeline 0/0/0/0/0.			
L2 Q judge	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/20; pipeline	0/0/0/0/0.			
L2 leak	20/20	0.463 [0.279, 0.652]	0.200 [0.048, 0.389]	0.000 / 0.000 / 1.000
MRR 0.452 [0.326, 0.589]; flow	0/0/12/8/0; pipeline 0/0/0/0/20.			
L2 net	20/20	0.943 [0.866, 0.992]	0.700 [0.500, 0.895]	0.750 / 0.050 / 0.200
MRR 0.833 [0.714, 0.941]; flow	0/0/12/8/0; pipeline 0/0/8/8/4.			
L2 paper	20/20	0.960 [0.903, 1.000]	0.000 [0.000, 0.000]	0.000 / 0.950 / 0.050
MRR 0.483 [0.444, 0.500]; flow	0/0/12/8/0; pipeline 0/0/13/6/1.			
L2 strength	20/20	0.500 [0.500, 0.500]	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow	0/0/12/8/0; pipeline 0/0/0/0/0.			
sizes.gemma-e4b.N0; declared 20, collected 20, missing 0, ineligible 20, eligible 0				
L1 corr.	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/0; pipeline	0/0/0/0/0.			
L1 Q judge	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/0; pipeline	0/0/0/0/0.			
L1 net	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/0; pipeline	0/0/0/0/0.			
L1 paper	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/0; pipeline	0/0/0/0/0.			
L1 strength	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/0; pipeline	0/0/0/0/0.			
L2 corr.	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/0; pipeline	0/0/0/0/0.			
L2 Q judge	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/0; pipeline	0/0/0/0/0.			
L2 leak	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/0; pipeline	0/0/0/0/0.			
L2 net	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/0; pipeline	0/0/0/0/0.			
L2 paper	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/0; pipeline	0/0/0/0/0.			
L2 strength	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/0; pipeline	0/0/0/0/0.			
sizes.gemma-e4b.C1; declared 20, collected 20, missing 0, ineligible 20, eligible 0				
L1 corr.	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/0; pipeline	0/0/0/0/0.			
L1 Q judge	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/0; pipeline	0/0/0/0/0.			
L1 net	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/0; pipeline	0/0/0/0/0.			
L1 paper	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/0; pipeline	0/0/0/0/0.			
L1 strength	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/0; pipeline	0/0/0/0/0.			
L2 corr.	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/0; pipeline	0/0/0/0/0.			
L2 Q judge	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/0; pipeline	0/0/0/0/0.			
L2 leak	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/0; pipeline	0/0/0/0/0.			
L2 net	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/0; pipeline	0/0/0/0/0.			
L2 paper	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/0; pipeline	0/0/0/0/0.			
L2 strength	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/0; pipeline	0/0/0/0/0.			

Level / detector	Cov./n	AUROC [interval]	Top-1 [interval]	Unique / wrong / abstain
sizes.ministral-3b.N0 ; declared 20, collected 20, missing 0, ineligible 20, eligible 0				
L1 corr.	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/0; pipeline 0/0/0/0/0.				
L1 Q judge	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/0; pipeline 0/0/0/0/0.				
L1 net	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/0; pipeline 0/0/0/0/0.				
L1 paper	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/0; pipeline 0/0/0/0/0.				
L1 strength	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/0; pipeline 0/0/0/0/0.				
L2 corr.	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/0; pipeline 0/0/0/0/0.				
L2 Q judge	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/0; pipeline 0/0/0/0/0.				
L2 leak	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/0; pipeline 0/0/0/0/0.				
L2 net	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/0; pipeline 0/0/0/0/0.				
L2 paper	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/0; pipeline 0/0/0/0/0.				
L2 strength	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/0; pipeline 0/0/0/0/0.				
sizes.ministral-3b.C1 ; declared 20, collected 20, missing 0, ineligible 18, eligible 2				
L1 corr.	2/-	-	0.000 [0.000, 0.000]	- / - / -
MRR 0.292 [0.250, 0.333]; flow 0/0/2/0/0; pipeline 0/0/0/0/0.				
L1 Q judge	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/2; pipeline 0/0/0/0/0.				
L1 net	2/-	-	0.000 [0.000, 0.000]	0.000 / 1.000 / 0.000
MRR 0.500 [0.500, 0.500]; flow 0/0/2/0/0; pipeline 0/0/2/0/0.				
L1 paper	2/-	-	0.000 [0.000, 0.000]	0.000 / 0.500 / 0.500
MRR 0.417 [0.333, 0.500]; flow 0/0/2/0/0; pipeline 0/0/1/0/1.				
L1 strength	2/-	-	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/0/2/0/0; pipeline 0/0/0/0/0.				
L2 corr.	2/-	-	0.500 [0.000, 1.000]	- / - / -
MRR 0.750 [0.500, 1.000]; flow 0/0/2/0/0; pipeline 0/0/0/0/0.				
L2 Q judge	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/2; pipeline 0/0/0/0/0.				
L2 leak	2/-	-	0.000 [0.000, 0.000]	0.000 / 0.000 / 1.000
MRR 0.267 [0.200, 0.333]; flow 0/0/2/0/0; pipeline 0/0/0/0/2.				
L2 net	2/-	-	0.500 [0.000, 1.000]	0.500 / 0.000 / 0.500
MRR 0.600 [0.200, 1.000]; flow 0/0/2/0/0; pipeline 0/0/1/0/1.				
L2 paper	2/-	-	0.500 [0.000, 1.000]	0.000 / 0.500 / 0.500
MRR 0.625 [0.250, 1.000]; flow 0/0/2/0/0; pipeline 0/0/1/0/1.				
L2 strength	2/-	-	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/0/2/0/0; pipeline 0/0/0/0/0.				
mixed.D ; declared 20, collected 20, missing 0, ineligible 0, eligible 20				
L1 corr.	20/-	-	0.000 [0.000, 0.000]	- / - / -
MRR 0.226 [0.199, 0.254]; flow 0/2/13/5/0; pipeline 0/0/0/0/0.				
L1 Q judge	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/20; pipeline 0/0/0/0/0.				
L1 net	20/-	-	0.200 [0.048, 0.389]	0.000 / 0.000 / 1.000
MRR 0.491 [0.374, 0.621]; flow 0/2/13/5/0; pipeline 0/0/0/0/20. Alarms: any 0/20 [0.000, 0.168]; one 0/20 [0.000, 0.168]; ref. 0/20 [0.000, 0.168]				
L1 paper	20/-	-	0.000 [0.000, 0.000]	0.000 / 0.000 / 1.000
MRR 0.327 [0.268, 0.388]; flow 0/2/13/5/0; pipeline 0/0/0/0/20. Alarms: any 0/20 [0.000, 0.168]; one 0/20 [0.000, 0.168]; ref. 0/20 [0.000, 0.168]				
L1 strength	20/-	-	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/2/13/5/0; pipeline 0/0/0/0/0.				
L2 corr.	20/-	-	0.150 [0.000, 0.333]	- / - / -
MRR 0.366 [0.251, 0.501]; flow 0/2/13/5/0; pipeline 0/0/0/0/0.				
L2 Q judge	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/20; pipeline 0/0/0/0/0.				

Level / detector	Cov./n	AUROC [interval]	Top-1 [interval]	Unique / wrong / abstain
L2 leak	20/-	-	0.150 [0.000, 0.333]	0.100 / 0.050 / 0.850
MRR 0.432 [0.324, 0.558]; flow 0/2/13/5/0; pipeline 0/3/0/0/17. Alarms: any 3/20 [0.032, 0.379]; one 3/20 [0.032, 0.379]; ref. 2/20 [0.012, 0.317]				
L2 net	20/-	-	0.250 [0.062, 0.455]	0.000 / 0.050 / 0.950
MRR 0.431 [0.289, 0.591]; flow 0/2/13/5/0; pipeline 0/1/0/0/19. Alarms: any 1/20 [0.001, 0.249]; one 1/20 [0.001, 0.249]; ref. 0/20 [0.000, 0.168]				
L2 paper	20/-	-	0.000 [0.000, 0.000]	0.000 / 0.100 / 0.900
MRR 0.254 [0.209, 0.308]; flow 0/2/13/5/0; pipeline 0/0/1/1/18. Alarms: any 2/20 [0.012, 0.317]; one 2/20 [0.012, 0.317]; ref. 0/20 [0.000, 0.168]				
L2 strength	20/-	-	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/2/13/5/0; pipeline 0/0/0/0/0.				
mixed.C1-XX ; declared 20, collected 20, missing 0, ineligible 0, eligible 20				
L1 corr.	20/20	0.990 [0.963, 1.000]	0.850 [0.667, 1.000]	- / - / -
MRR 0.925 [0.833, 1.000]; flow 0/2/8/10/0; pipeline 0/0/0/0/0.				
L1 Q judge	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/20; pipeline 0/0/0/0/0.				
L1 net	20/20	0.965 [0.898, 1.000]	0.950 [0.833, 1.000]	0.800 / 0.000 / 0.200
MRR 0.975 [0.917, 1.000]; flow 0/2/8/10/0; pipeline 0/2/6/8/4.				
L1 paper	20/20	0.955 [0.848, 1.000]	0.900 [0.750, 1.000]	0.450 / 0.000 / 0.550
MRR 0.950 [0.875, 1.000]; flow 0/2/8/10/0; pipeline 0/1/2/6/11.				
L1 strength	20/20	0.500 [0.500, 0.500]	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/2/8/10/0; pipeline 0/0/0/0/0.				
L2 corr.	20/20	0.998 [0.986, 1.000]	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/2/8/10/0; pipeline 0/0/0/0/0.				
L2 Q judge	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/20; pipeline 0/0/0/0/0.				
L2 leak	20/20	0.695 [0.505, 0.858]	0.350 [0.143, 0.571]	0.100 / 0.000 / 0.900
MRR 0.545 [0.396, 0.706]; flow 0/2/8/10/0; pipeline 0/2/0/0/18.				
L2 net	20/20	0.885 [0.779, 0.965]	0.700 [0.500, 0.889]	0.500 / 0.100 / 0.400
MRR 0.833 [0.714, 0.939]; flow 0/2/8/10/0; pipeline 0/2/3/7/8.				
L2 paper	20/20	0.993 [0.969, 1.000]	0.950 [0.833, 1.000]	0.600 / 0.000 / 0.400
MRR 0.975 [0.917, 1.000]; flow 0/2/8/10/0; pipeline 0/1/5/6/8.				
L2 strength	20/20	0.500 [0.500, 0.500]	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/2/8/10/0; pipeline 0/0/0/0/0.				
mixed.H-XYZZ ; declared 20, collected 20, missing 0, ineligible 0, eligible 20				
L1 corr.	20/-	-	0.050 [0.000, 0.167]	- / - / -
MRR 0.333 [0.250, 0.432]; flow 0/0/17/3/0; pipeline 0/0/0/0/0.				
L1 Q judge	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/20; pipeline 0/0/0/0/0.				
L1 net	20/-	-	0.000 [0.000, 0.000]	0.000 / 0.050 / 0.950
MRR 0.391 [0.331, 0.446]; flow 0/0/17/3/0; pipeline 0/0/1/0/19. Alarms: any 1/20 [0.001, 0.249]; one 1/20 [0.001, 0.249]; ref. 0/20 [0.000, 0.168]				
L1 paper	20/-	-	0.000 [0.000, 0.000]	0.000 / 0.050 / 0.950
MRR 0.368 [0.305, 0.428]; flow 0/0/17/3/0; pipeline 0/0/1/0/19. Alarms: any 1/20 [0.001, 0.249]; one 1/20 [0.001, 0.249]; ref. 0/20 [0.000, 0.168]				
L1 strength	20/-	-	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/0/17/3/0; pipeline 0/0/0/0/0.				
L2 corr.	20/-	-	0.050 [0.000, 0.167]	- / - / -
MRR 0.369 [0.292, 0.462]; flow 0/0/17/3/0; pipeline 0/0/0/0/0.				
L2 Q judge	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/20; pipeline 0/0/0/0/0.				
L2 leak	20/-	-	0.100 [0.000, 0.250]	0.000 / 0.050 / 0.950
MRR 0.342 [0.251, 0.457]; flow 0/0/17/3/0; pipeline 0/1/0/0/19. Alarms: any 1/20 [0.001, 0.249]; one 1/20 [0.001, 0.249]; ref. 0/20 [0.000, 0.168]				
L2 net	20/-	-	0.100 [0.000, 0.250]	0.000 / 0.100 / 0.900
MRR 0.430 [0.334, 0.541]; flow 0/0/17/3/0; pipeline 0/1/1/0/18. Alarms: any 2/20 [0.012, 0.317]; one 2/20 [0.012, 0.317]; ref. 0/20 [0.000, 0.168]				
L2 paper	20/-	-	0.000 [0.000, 0.000]	0.000 / 0.350 / 0.650
MRR 0.269 [0.217, 0.327]; flow 0/0/17/3/0; pipeline 0/1/2/4/13. Alarms: any 7/20 [0.154, 0.592]; one 7/20 [0.154, 0.592]; ref. 0/20 [0.000, 0.168]				
L2 strength	20/-	-	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/0/17/3/0; pipeline 0/0/0/0/0.				
mixed.C1-XY ; declared 20, collected 20, missing 0, ineligible 0, eligible 20				
L1 corr.	20/20	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/6/7/7/0; pipeline 0/0/0/0/0.				

Level / detector	Cov./n	AUROC [interval]	Top-1 [interval]	Unique / wrong / abstain
L1 Q judge	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/20; pipeline 0/0/0/0/0.				
L1 net	20/20	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	0.950 / 0.000 / 0.050
MRR 1.000 [1.000, 1.000]; flow 0/6/7/7/0; pipeline 0/5/7/7/1.				
L1 paper	20/20	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	1.000 / 0.000 / 0.000
MRR 1.000 [1.000, 1.000]; flow 0/6/7/7/0; pipeline 0/6/7/7/0.				
L1 strength	20/20	0.500 [0.500, 0.500]	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/6/7/7/0; pipeline 0/0/0/0/0.				
L2 corr.	20/20	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/6/7/7/0; pipeline 0/0/0/0/0.				
L2 Q judge	0/-	-	-	- / - / -
MRR -; flow 0/0/0/0/20; pipeline 0/0/0/0/0.				
L2 leak	20/20	0.860 [0.701, 0.989]	0.750 [0.550, 0.933]	0.350 / 0.000 / 0.650
MRR 0.838 [0.702, 0.956]; flow 0/6/7/7/0; pipeline 0/6/0/1/13.				
L2 net	20/20	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	1.000 / 0.000 / 0.000
MRR 1.000 [1.000, 1.000]; flow 0/6/7/7/0; pipeline 0/6/7/7/0.				
L2 paper	20/20	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	1.000 / 0.000 / 0.000
MRR 1.000 [1.000, 1.000]; flow 0/6/7/7/0; pipeline 0/6/7/7/0.				
L2 strength	20/20	0.500 [0.500, 0.500]	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/6/7/7/0; pipeline 0/0/0/0/0.				
core.C1-NW ; declared 20, collected 20, missing 0, ineligible 0, eligible 20				
L1 corr.	20/20	0.993 [0.960, 1.000]	0.900 [0.750, 1.000]	- / - / -
MRR 0.933 [0.833, 1.000]; flow 0/1/8/11/0; pipeline 0/0/0/0/0.				
L1 Q judge	12/5	0.480 [0.000, 1.000]	0.931 [0.762, 1.000]	- / - / -
MRR 0.940 [0.796, 1.000]; flow 0/0/4/8/8; pipeline 0/0/0/0/0.				
L1 net	20/20	0.968 [0.903, 1.000]	0.800 [0.609, 0.955]	0.600 / 0.000 / 0.400
MRR 0.892 [0.786, 0.976]; flow 0/1/8/11/0; pipeline 0/0/4/8/8.				
L1 paper	20/20	0.932 [0.847, 0.994]	0.350 [0.150, 0.565]	0.100 / 0.400 / 0.500
MRR 0.667 [0.562, 0.778]; flow 0/1/8/11/0; pipeline 0/0/5/5/10.				
L1 strength	20/20	0.500 [0.500, 0.500]	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/1/8/11/0; pipeline 0/0/0/0/0.				
L2 corr.	20/20	0.985 [0.932, 1.000]	0.950 [0.833, 1.000]	- / - / -
MRR 0.975 [0.917, 1.000]; flow 0/1/8/11/0; pipeline 0/0/0/0/0.				
L2 Q judge	10/5	0.800 [0.500, 1.000]	0.900 [0.667, 1.000]	- / - / -
MRR 0.950 [0.833, 1.000]; flow 0/0/4/6/10; pipeline 0/0/0/0/0.				
L2 leak	20/20	0.682 [0.505, 0.844]	0.200 [0.048, 0.389]	0.100 / 0.050 / 0.850
MRR 0.464 [0.340, 0.598]; flow 0/1/8/11/0; pipeline 0/2/0/1/17.				
L2 net	20/20	0.963 [0.910, 1.000]	0.650 [0.438, 0.857]	0.650 / 0.050 / 0.300
MRR 0.817 [0.704, 0.923]; flow 0/1/8/11/0; pipeline 0/1/5/8/6.				
L2 paper	20/20	0.975 [0.901, 1.000]	0.000 [0.000, 0.000]	0.000 / 0.950 / 0.050
MRR 0.492 [0.472, 0.500]; flow 0/1/8/11/0; pipeline 0/0/12/7/1.				
L2 strength	20/20	0.500 [0.500, 0.500]	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/1/8/11/0; pipeline 0/0/0/0/0.				
lookalike.LX ; declared 20, collected 20, missing 0, ineligible 0, eligible 20				
L1 corr.	20/-	-	0.000 [0.000, 0.000]	- / - / -
MRR 0.214 [0.188, 0.251]; flow 0/0/10/10/0; pipeline 0/0/0/0/0.				
L1 Q judge	14/-	-	0.881 [0.702, 1.000]	- / - / -
MRR 0.898 [0.745, 1.000]; flow 0/0/8/6/6; pipeline 0/0/0/0/0.				
L1 net	20/-	-	0.050 [0.000, 0.167]	0.000 / 0.050 / 0.950
MRR 0.368 [0.289, 0.461]; flow 0/0/10/10/0; pipeline 0/0/0/1/19.				
L1 paper	20/-	-	0.000 [0.000, 0.000]	0.000 / 0.100 / 0.900
MRR 0.283 [0.231, 0.339]; flow 0/0/10/10/0; pipeline 0/0/1/1/18.				
L1 strength	20/-	-	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/0/10/10/0; pipeline 0/0/0/0/0.				
L2 corr.	20/-	-	0.050 [0.000, 0.167]	- / - / -
MRR 0.316 [0.240, 0.411]; flow 0/0/10/10/0; pipeline 0/0/0/0/0.				
L2 Q judge	9/-	-	0.907 [0.688, 1.000]	- / - / -
MRR 0.921 [0.732, 1.000]; flow 0/0/5/4/11; pipeline 0/0/0/0/0.				
L2 leak	20/-	-	0.150 [0.000, 0.318]	0.000 / 0.000 / 1.000
MRR 0.420 [0.306, 0.548]; flow 0/0/10/10/0; pipeline 0/0/0/0/20.				
L2 net	20/-	-	0.150 [0.000, 0.318]	0.000 / 0.000 / 1.000
MRR 0.398 [0.292, 0.524]; flow 0/0/10/10/0; pipeline 0/0/0/0/20.				
L2 paper	20/-	-	0.000 [0.000, 0.000]	0.000 / 0.600 / 0.400
MRR 0.237 [0.203, 0.279]; flow 0/0/10/10/0; pipeline 0/0/8/4/8.				

Level / detector	Cov./n	AUROC [interval]	Top-1 [interval]	Unique / wrong / abstain
L2 strength	20/-	-	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/0/10/0/0; pipeline 0/0/0/0/0.				
lookalike.LX-swap ; declared 20, collected 20, missing 0, ineligible 0, eligible 20				
L1 corr.	20/-	-	0.000 [0.000, 0.000]	- / - / -
MRR 0.231 [0.194, 0.280]; flow 0/0/13/7/0; pipeline 0/0/0/0/0.				
L1 Q judge	7/-	-	0.881 [0.583, 1.000]	- / - / -
MRR 0.898 [0.643, 1.000]; flow 0/0/4/3/13; pipeline 0/0/0/0/0.				
L1 net	20/-	-	0.100 [0.000, 0.250]	0.000 / 0.000 / 1.000
MRR 0.399 [0.307, 0.509]; flow 0/0/13/7/0; pipeline 0/0/0/0/20.				
L1 paper	20/-	-	0.000 [0.000, 0.000]	0.000 / 0.000 / 1.000
MRR 0.277 [0.228, 0.336]; flow 0/0/13/7/0; pipeline 0/0/0/0/20.				
L1 strength	20/-	-	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/0/13/7/0; pipeline 0/0/0/0/0.				
L2 corr.	20/-	-	0.100 [0.000, 0.250]	- / - / -
MRR 0.393 [0.296, 0.506]; flow 0/0/13/7/0; pipeline 0/0/0/0/0.				
L2 Q judge	8/-	-	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/0/6/2/12; pipeline 0/0/0/0/0.				
L2 leak	20/-	-	0.100 [0.000, 0.250]	0.000 / 0.000 / 1.000
MRR 0.403 [0.316, 0.513]; flow 0/0/13/7/0; pipeline 0/0/0/0/20.				
L2 net	20/-	-	0.200 [0.048, 0.391]	0.000 / 0.050 / 0.950
MRR 0.438 [0.315, 0.582]; flow 0/0/13/7/0; pipeline 0/0/1/0/19.				
L2 paper	20/-	-	0.000 [0.000, 0.000]	0.000 / 0.650 / 0.350
MRR 0.237 [0.204, 0.276]; flow 0/0/13/7/0; pipeline 0/0/9/4/7.				
L2 strength	20/-	-	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/0/13/7/0; pipeline 0/0/0/0/0.				
lookalike.LY ; declared 20, collected 20, missing 0, ineligible 0, eligible 20				
L1 corr.	20/-	-	0.000 [0.000, 0.000]	- / - / -
MRR 0.188 [0.172, 0.210]; flow 0/0/15/5/0; pipeline 0/0/0/0/0.				
L1 Q judge	10/-	-	0.783 [0.528, 1.000]	- / - / -
MRR 0.867 [0.717, 1.000]; flow 0/0/8/2/10; pipeline 0/0/0/0/0.				
L1 net	20/-	-	0.100 [0.000, 0.250]	0.000 / 0.150 / 0.850
MRR 0.340 [0.241, 0.461]; flow 0/0/15/5/0; pipeline 0/0/2/1/17.				
L1 paper	20/-	-	0.000 [0.000, 0.000]	0.000 / 0.350 / 0.650
MRR 0.382 [0.326, 0.438]; flow 0/0/15/5/0; pipeline 0/0/3/4/13.				
L1 strength	20/-	-	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/0/15/5/0; pipeline 0/0/0/0/0.				
L2 corr.	20/-	-	0.250 [0.059, 0.450]	- / - / -
MRR 0.520 [0.391, 0.655]; flow 0/0/15/5/0; pipeline 0/0/0/0/0.				
L2 Q judge	6/-	-	0.611 [0.233, 1.000]	- / - / -
MRR 0.742 [0.481, 1.000]; flow 0/0/3/3/14; pipeline 0/0/0/0/0.				
L2 leak	20/-	-	0.000 [0.000, 0.000]	0.000 / 0.000 / 1.000
MRR 0.301 [0.254, 0.353]; flow 0/0/15/5/0; pipeline 0/0/0/0/20.				
L2 net	20/-	-	0.000 [0.000, 0.000]	0.000 / 0.100 / 0.900
MRR 0.334 [0.280, 0.390]; flow 0/0/15/5/0; pipeline 0/0/1/1/18.				
L2 paper	20/-	-	0.000 [0.000, 0.000]	0.000 / 0.950 / 0.050
MRR 0.299 [0.251, 0.352]; flow 0/0/15/5/0; pipeline 0/0/8/11/1.				
L2 strength	20/-	-	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/0/15/5/0; pipeline 0/0/0/0/0.				
lookalike.LY-swap ; declared 20, collected 20, missing 0, ineligible 0, eligible 20				
L1 corr.	20/-	-	0.050 [0.000, 0.167]	- / - / -
MRR 0.276 [0.209, 0.371]; flow 0/0/17/3/0; pipeline 0/0/0/0/0.				
L1 Q judge	12/-	-	0.778 [0.533, 1.000]	- / - / -
MRR 0.814 [0.613, 1.000]; flow 0/0/11/1/8; pipeline 0/0/0/0/0.				
L1 net	20/-	-	0.050 [0.000, 0.167]	0.050 / 0.000 / 0.950
MRR 0.386 [0.312, 0.473]; flow 0/0/17/3/0; pipeline 0/0/1/0/19.				
L1 paper	20/-	-	0.000 [0.000, 0.000]	0.000 / 0.400 / 0.600
MRR 0.387 [0.330, 0.440]; flow 0/0/17/3/0; pipeline 0/0/5/3/12.				
L1 strength	20/-	-	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/0/17/3/0; pipeline 0/0/0/0/0.				
L2 corr.	20/-	-	0.150 [0.000, 0.333]	- / - / -
MRR 0.420 [0.310, 0.548]; flow 0/0/17/3/0; pipeline 0/0/0/0/0.				
L2 Q judge	7/-	-	0.762 [0.444, 1.000]	- / - / -
MRR 0.796 [0.524, 1.000]; flow 0/0/6/1/13; pipeline 0/0/0/0/0.				
L2 leak	20/-	-	0.100 [0.000, 0.250]	0.000 / 0.050 / 0.950
MRR 0.367 [0.277, 0.478]; flow 0/0/17/3/0; pipeline 0/1/0/0/19.				

Level / detector	Cov./n	AUROC [interval]	Top-1 [interval]	Unique / wrong / abstain
L2 net	20/-	-	0.000 [0.000, 0.000]	0.000 / 0.150 / 0.850
MRR 0.313 [0.274, 0.356]; flow 0/0/17/3/0; pipeline 0/0/0/3/17.				
L2 paper	20/-	-	0.000 [0.000, 0.000]	0.000 / 0.900 / 0.100
MRR 0.355 [0.296, 0.412]; flow 0/0/17/3/0; pipeline 0/0/10/8/2.				
L2 strength	20/-	-	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/0/17/3/0; pipeline 0/0/0/0/0.				
lookalike.LZ ; declared 20, collected 20, missing 0, ineligible 0, eligible 20				
L1 corr.	20/-	-	0.050 [0.000, 0.167]	- / - / -
MRR 0.317 [0.235, 0.418]; flow 0/0/11/9/0; pipeline 0/0/0/0/0.				
L1 Q judge	12/-	-	0.944 [0.810, 1.000]	- / - / -
MRR 0.958 [0.857, 1.000]; flow 0/0/6/6/8; pipeline 0/0/0/0/0.				
L1 net	20/-	-	0.150 [0.000, 0.333]	0.000 / 0.050 / 0.950
MRR 0.386 [0.275, 0.517]; flow 0/0/11/9/0; pipeline 0/0/0/1/19.				
L1 paper	20/-	-	0.000 [0.000, 0.000]	0.000 / 0.000 / 1.000
MRR 0.249 [0.198, 0.308]; flow 0/0/11/9/0; pipeline 0/0/0/0/20.				
L1 strength	20/-	-	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/0/11/9/0; pipeline 0/0/0/0/0.				
L2 corr.	20/-	-	0.150 [0.000, 0.333]	- / - / -
MRR 0.376 [0.265, 0.511]; flow 0/0/11/9/0; pipeline 0/0/0/0/0.				
L2 Q judge	6/-	-	0.667 [0.250, 1.000]	- / - / -
MRR 0.806 [0.542, 1.000]; flow 0/0/2/4/14; pipeline 0/0/0/0/0.				
L2 leak	20/-	-	0.200 [0.048, 0.389]	0.000 / 0.000 / 1.000
MRR 0.429 [0.305, 0.572]; flow 0/0/11/9/0; pipeline 0/0/0/0/20.				
L2 net	20/-	-	0.100 [0.000, 0.250]	0.050 / 0.150 / 0.800
MRR 0.362 [0.268, 0.478]; flow 0/0/11/9/0; pipeline 0/0/0/4/16.				
L2 paper	20/-	-	0.000 [0.000, 0.000]	0.000 / 0.250 / 0.750
MRR 0.258 [0.212, 0.312]; flow 0/0/11/9/0; pipeline 0/0/1/4/15.				
L2 strength	20/-	-	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/0/11/9/0; pipeline 0/0/0/0/0.				
lookalike.LZ-swap ; declared 20, collected 20, missing 0, ineligible 0, eligible 20				
L1 corr.	20/-	-	0.050 [0.000, 0.167]	- / - / -
MRR 0.246 [0.189, 0.342]; flow 0/0/10/10/0; pipeline 0/0/0/0/0.				
L1 Q judge	8/-	-	0.583 [0.259, 0.889]	- / - / -
MRR 0.696 [0.455, 0.917]; flow 0/0/4/4/12; pipeline 0/0/0/0/0.				
L1 net	20/-	-	0.100 [0.000, 0.250]	0.000 / 0.000 / 1.000
MRR 0.353 [0.257, 0.473]; flow 0/0/10/10/0; pipeline 0/0/0/0/20.				
L1 paper	20/-	-	0.000 [0.000, 0.000]	0.000 / 0.000 / 1.000
MRR 0.211 [0.181, 0.251]; flow 0/0/10/10/0; pipeline 0/0/0/0/20.				
L1 strength	20/-	-	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/0/10/10/0; pipeline 0/0/0/0/0.				
L2 corr.	20/-	-	0.150 [0.000, 0.333]	- / - / -
MRR 0.354 [0.243, 0.494]; flow 0/0/10/10/0; pipeline 0/0/0/0/0.				
L2 Q judge	5/-	-	0.800 [0.333, 1.000]	- / - / -
MRR 0.880 [0.600, 1.000]; flow 0/0/4/1/15; pipeline 0/0/0/0/0.				
L2 leak	20/-	-	0.250 [0.062, 0.450]	0.000 / 0.050 / 0.950
MRR 0.506 [0.381, 0.646]; flow 0/0/10/10/0; pipeline 0/1/0/0/19.				
L2 net	20/-	-	0.300 [0.111, 0.520]	0.050 / 0.050 / 0.900
MRR 0.512 [0.372, 0.666]; flow 0/0/10/10/0; pipeline 0/0/0/2/18.				
L2 paper	20/-	-	0.000 [0.000, 0.000]	0.000 / 0.150 / 0.850
MRR 0.254 [0.208, 0.309]; flow 0/0/10/10/0; pipeline 0/1/0/2/17.				
L2 strength	20/-	-	1.000 [1.000, 1.000]	- / - / -
MRR 1.000 [1.000, 1.000]; flow 0/0/10/10/0; pipeline 0/0/0/0/0.				

E.1 Prefix, missing-bid and breadth sensitivity

The following compact table provides the score trajectories. Full intervals, coverage, flow labels and reporting outcomes for every prefix/mask remain in `results/analysis/final_v1/all_cells.md`. A3 merges all registered breadth indices; it does not select the first subset after observing scores. The unavailable Gemma-E4B and Ministral-3B cells remain in the comprehensive table above. Mixed XX and XY treatments each use their own composition-matched honest control; their contrast does not isolate model identity.

Table 7: Prefix and missing-bid sensitivity for every available treatment comparison. Each entry is AUROC; exact intervals, exclusions and other metrics for every prefix/mask are in all_cells.md.

Cell / level / detector	Mask	n (full)	Prefix 50	100	200
core.P0/L1/net	complete	20	1.000	1.000	1.000
core.P0/L1/net	no_bid	20	1.000	1.000	1.000
core.P0/L1/correlation	complete	20	1.000	1.000	1.000
core.P0/L1/correlation	no_bid	20	1.000	1.000	1.000
core.P0/L2/net	complete	20	0.897	1.000	1.000
core.P0/L2/net	no_bid	20	0.890	1.000	1.000
core.P0/L2/correlation	complete	20	0.978	1.000	1.000
core.P0/L2/correlation	no_bid	20	0.978	1.000	1.000
core.P0/L2/leak	complete	20	0.953	1.000	1.000
core.P0/L2/leak	no_bid	20	0.953	1.000	1.000
core.C1-SS/L1/net	complete	20	0.882	0.915	0.912
core.C1-SS/L1/net	no_bid	20	0.882	0.915	0.917
core.C1-SS/L1/correlation	complete	20	0.968	0.978	0.988
core.C1-SS/L1/correlation	no_bid	20	0.968	0.975	0.988
core.C1-SS/L2/net	complete	20	0.650	0.795	0.865
core.C1-SS/L2/net	no_bid	20	0.660	0.785	0.858
core.C1-SS/L2/correlation	complete	20	0.975	0.998	0.998
core.C1-SS/L2/correlation	no_bid	20	0.915	0.990	1.000
core.C1-SS/L2/leak	complete	20	0.522	0.562	0.640
core.C1-SS/L2/leak	no_bid	20	0.520	0.560	0.627
core.C1-SF/L1/net	complete	20	0.762	0.840	0.993
core.C1-SF/L1/net	no_bid	20	0.760	0.870	0.970
core.C1-SF/L1/correlation	complete	20	0.902	0.990	1.000
core.C1-SF/L1/correlation	no_bid	20	0.902	0.993	1.000
core.C1-SF/L2/net	complete	20	0.700	0.720	0.850
core.C1-SF/L2/net	no_bid	20	0.690	0.730	0.853
core.C1-SF/L2/correlation	complete	20	0.833	0.920	0.980
core.C1-SF/L2/correlation	no_bid	20	0.748	0.887	0.998
core.C1-SF/L2/leak	complete	20	0.650	0.642	0.645
core.C1-SF/L2/leak	no_bid	20	0.660	0.665	0.675
core.C1-FF/L1/net	complete	20	0.907	1.000	1.000
core.C1-FF/L1/net	no_bid	20	0.902	1.000	1.000
core.C1-FF/L1/correlation	complete	20	0.960	0.985	0.993
core.C1-FF/L1/correlation	no_bid	20	0.938	0.985	0.990
core.C1-FF/L2/net	complete	20	0.750	0.882	0.998
core.C1-FF/L2/net	no_bid	20	0.752	0.887	0.998
core.C1-FF/L2/correlation	complete	20	0.963	0.995	1.000
core.C1-FF/L2/correlation	no_bid	20	0.958	0.995	1.000
core.C1-FF/L2/leak	complete	20	0.580	0.740	0.868
core.C1-FF/L2/leak	no_bid	20	0.590	0.708	0.863
core.C1-NW/L1/net	complete	20	0.845	0.902	0.968
core.C1-NW/L1/net	no_bid	20	0.873	0.905	0.975
core.C1-NW/L1/correlation	complete	20	0.932	0.955	0.993
core.C1-NW/L1/correlation	no_bid	20	0.970	0.980	1.000
core.C1-NW/L2/net	complete	20	0.725	0.797	0.963
core.C1-NW/L2/net	no_bid	20	0.730	0.810	0.963
core.C1-NW/L2/correlation	complete	20	0.927	0.955	0.985
core.C1-NW/L2/correlation	no_bid	20	0.843	0.905	0.930
core.C1-NW/L2/leak	complete	20	0.598	0.613	0.682
core.C1-NW/L2/leak	no_bid	20	0.590	0.620	0.690
core.C2/L1/net	complete	20	0.662	0.688	0.610
core.C2/L1/net	no_bid	20	0.657	0.680	0.605
core.C2/L1/correlation	complete	20	0.333	0.407	0.463
core.C2/L1/correlation	no_bid	20	0.370	0.432	0.485
core.C2/L2/net	complete	20	0.445	0.470	0.655
core.C2/L2/net	no_bid	20	0.458	0.482	0.665
core.C2/L2/correlation	complete	20	0.410	0.502	0.495
core.C2/L2/correlation	no_bid	20	0.365	0.470	0.472
core.C2/L2/leak	complete	20	0.415	0.417	0.562
core.C2/L2/leak	no_bid	20	0.420	0.430	0.575
core.E1/L1/net	complete	20	0.578	0.585	0.675
core.E1/L1/net	no_bid	20	0.593	0.598	0.672
core.E1/L1/correlation	complete	20	0.450	0.500	0.482

Cell / level / detector	Mask	n (full)	Prefix 50	100	200
core.E1/L1/correlation	no_bid	20	0.512	0.540	0.475
core.E1/L2/net	complete	20	0.458	0.665	0.700
core.E1/L2/net	no_bid	20	0.455	0.682	0.705
core.E1/L2/correlation	complete	20	0.440	0.500	0.415
core.E1/L2/correlation	no_bid	20	0.480	0.522	0.535
core.E1/L2/leak	complete	20	0.522	0.695	0.703
core.E1/L2/leak	no_bid	20	0.515	0.698	0.705
labs.gemma-12b.C1/L1/net	complete	14	0.765	1.000	1.000
labs.gemma-12b.C1/L1/net	no_bid	14	0.832	1.000	1.000
labs.gemma-12b.C1/L1/correlation	complete	14	0.995	1.000	1.000
labs.gemma-12b.C1/L1/correlation	no_bid	14	0.944	0.974	0.954
labs.gemma-12b.C1/L2/net	complete	14	0.699	0.883	0.995
labs.gemma-12b.C1/L2/net	no_bid	14	0.714	0.883	0.995
labs.gemma-12b.C1/L2/correlation	complete	14	0.842	1.000	1.000
labs.gemma-12b.C1/L2/correlation	no_bid	14	0.867	0.959	0.944
labs.gemma-12b.C1/L2/leak	complete	14	0.653	0.791	0.969
labs.gemma-12b.C1/L2/leak	no_bid	14	0.781	0.796	0.949
labs.ministral-8b.C1/L1/net	complete	20	0.583	0.825	0.860
labs.ministral-8b.C1/L1/net	no_bid	20	0.590	0.820	0.855
labs.ministral-8b.C1/L1/correlation	complete	20	0.853	0.875	0.897
labs.ministral-8b.C1/L1/correlation	no_bid	20	0.853	0.875	0.897
labs.ministral-8b.C1/L2/net	complete	20	0.627	0.605	0.745
labs.ministral-8b.C1/L2/net	no_bid	20	0.618	0.613	0.743
labs.ministral-8b.C1/L2/correlation	complete	20	0.585	0.760	0.828
labs.ministral-8b.C1/L2/correlation	no_bid	20	0.585	0.760	0.828
labs.ministral-8b.C1/L2/leak	complete	20	0.590	0.637	0.675
labs.ministral-8b.C1/L2/leak	no_bid	20	0.588	0.650	0.675
sizes.qwen-4b.C1/L1/net	complete	20	0.858	0.955	0.993
sizes.qwen-4b.C1/L1/net	no_bid	20	0.858	0.955	0.993
sizes.qwen-4b.C1/L1/correlation	complete	20	0.760	0.915	0.978
sizes.qwen-4b.C1/L1/correlation	no_bid	20	0.760	0.915	0.978
sizes.qwen-4b.C1/L2/net	complete	20	0.718	0.865	0.943
sizes.qwen-4b.C1/L2/net	no_bid	20	0.723	0.860	0.945
sizes.qwen-4b.C1/L2/correlation	complete	20	0.723	0.838	0.955
sizes.qwen-4b.C1/L2/correlation	no_bid	20	0.723	0.838	0.955
sizes.qwen-4b.C1/L2/leak	complete	20	0.550	0.450	0.463
sizes.qwen-4b.C1/L2/leak	no_bid	20	0.552	0.448	0.470
mixed.C1-XX/L1/net	complete	20	0.780	0.912	0.965
mixed.C1-XX/L1/net	no_bid	20	0.802	0.912	0.958
mixed.C1-XX/L1/correlation	complete	20	0.943	0.985	0.990
mixed.C1-XX/L1/correlation	no_bid	20	0.948	0.988	0.990
mixed.C1-XX/L2/net	complete	20	0.573	0.657	0.885
mixed.C1-XX/L2/net	no_bid	20	0.580	0.675	0.900
mixed.C1-XX/L2/correlation	complete	20	0.875	0.920	0.998
mixed.C1-XX/L2/correlation	no_bid	20	0.885	0.917	0.998
mixed.C1-XX/L2/leak	complete	20	0.485	0.552	0.695
mixed.C1-XX/L2/leak	no_bid	20	0.490	0.545	0.705
mixed.C1-XY/L1/net	complete	20	0.950	0.963	1.000
mixed.C1-XY/L1/net	no_bid	20	0.960	0.970	1.000
mixed.C1-XY/L1/correlation	complete	20	0.980	1.000	1.000
mixed.C1-XY/L1/correlation	no_bid	20	0.980	1.000	1.000
mixed.C1-XY/L2/net	complete	20	0.818	0.948	1.000
mixed.C1-XY/L2/net	no_bid	20	0.830	0.950	1.000
mixed.C1-XY/L2/correlation	complete	20	0.970	1.000	1.000
mixed.C1-XY/L2/correlation	no_bid	20	0.968	1.000	1.000
mixed.C1-XY/L2/leak	complete	20	0.792	0.865	0.860
mixed.C1-XY/L2/leak	no_bid	20	0.762	0.865	0.868

F How the cartels and the spy behaved (post hoc)

This descriptive analysis was not pre-registered. It uses the Qwen3.5-9B markets 0–19 and complete rounds, and is produced by `experiments/cartel_mechanics.py` (Table 8). The counterfactual rows replace the pair’s bids by a fixed rule while keeping the two other firms’ observed bids; they ignore how rivals would react. Messages are

coded with simple keyword patterns and are approximate.

Table 8: How the designated pair bid and what the chatting partners wrote. Shares of rounds or bids; profit and prices per round.

	Honest (N0)	Chat (C1-SS)	Chat, no warning (C1-NW)	No chat (C2)
Joint profit of the pair	0.32	0.46	0.12	0.72
Winning price	2.14	2.44	2.10	2.94
rounds 1–5 / 21–200	2.20 / 2.14	2.90 / 2.41	2.32 / 2.09	2.75 / 2.96
Pair bids equal to 3: round 1 / 21–200	33% / 35%	50% / 49%	32% / 18%	50% / 98%
Partners bid the same	53%	59%	77%	95%
Pair bids below own cost	12%	13%	28%	0.2%
Pair wins the contract	96%	89%	96%	75%
of which the costlier partner wins	19%	24%	29%	31%
<i>Counterfactual joint profit against the observed rival bids</i>				
Only the cheaper partner bids 3	0.97	0.97	0.93	0.91
Both partners bid 3	0.76	0.77	0.74	0.74
<i>Chat messages</i>				
Reveal the sender's cost	–	94%	91%	–
Announce a bid	–	73%	94%	–
Mention avoiding detection	–	76%	3%	–
Say who should win	–	6%	14%	–
Propose alternating wins	–	1%	2%	–
Both announce a bid / the same bid	–	54% / 44%	88% / 66%	–
Announced bid kept	–	90%	91%	–

Why the cartel without chat bids 3. Three explanations fit the data and need not exclude each other. (i) *Focal point*. 3 is the highest cost of a strong firm and the lowest cost of a weak firm, so it is the highest price at which a strong firm never loses money and a weak firm cannot profitably undercut. Already in round 1, half of the pair's bids were 3, against a third in honest markets. (ii) *Lock-in through the bid history*. Without a channel the partner's bid is the only signal, and a constant bid is easy to read and to match: the share rose to 98% after round 20. (iii) *Caution*. The instruction to behave inconspicuously may favour a moderate, stable price; there is no tacit condition without that instruction to test this.

Why the chatting cartels earn less. (i) *No allocation*. Messages revealed costs and announced bids but rarely said who should win; when both partners announced a bid, they announced the same one in only 44% of rounds. (ii) *A simultaneous channel without memory*. Each partner sends one message at the same time and never sees earlier messages, so a proposal cannot be accepted and a rotation cannot be agreed; alternating wins were proposed in about 1% of messages. (iii) *Both compete for the win*. The winning price fell over the rounds, and the costlier partner won 24% of the pair's contracts. Removing the surveillance warning made this worse, not better: the partners announced bids in 88% of rounds, more than a quarter of their bids lay below cost, and their joint profit fell to 0.12, below honest play. A channel with replies or message memory would test explanation (ii).

Do honest firms bid badly? Partly. Honest strong firms bid below their own cost in 12% of bids and weak firms in 18%. Against the observed rival bids, bidding one above cost would have earned a strong firm 0.27 per round instead of the 0.16 it earned. Part of the measured cartel gain therefore reflects weak honest play. The cartel without chat earned about what “both bid 3” yields against the same bids (0.72 versus 0.74), and all cartels stayed well below a split in which only the cheaper partner bids 3.

Did the spy use the stolen cost? Barely, and not to its advantage (`experiments/spy_behaviour.py`; each E1 market against the honest market with the same index, costs and seats). Averaged over its own cost, the spy bid 2.20, 2.24 and 2.28 when the victim's cost was 1, 2 and 3; the same firm in the honest twin bid about 2.38 throughout.

The spy won more often than its honest counterpart (54% against 47% of rounds) but at lower prices, and earned 0.09 instead of 0.16 per round (difference -0.065 $[-0.079, -0.051]$, market bootstrap); it gained in only 1 of 20 markets. Its victim also earned less (0.11 against 0.16). Against the observed rival bids, simply bidding one above cost would have earned the spy 0.27; a simple informed rule (bid the victim’s cost when it exceeds one’s own) would have earned 0.23. The small leak signal therefore reflects weak use of the information by this model, not a profitable advantage that escaped detection.

G Statistics, provenance, and deviations

G.1 A1: warning removal

A1 compares matched N0/C1-SS/C1-NW triplets. Its profit contrast is the difference of additive N0-referenced joint-profit uplifts; the relative uplift companion is listed separately. All eligible triplets contribute to the non-judge estimates. Judge comparisons use only triplets judged successfully in all conditions. The L2 judge difference of 0.5 with interval $[0.5, 0.5]$ comes from a single triplet and is not precise evidence.

Table 9: A1: warning removed minus warning present. Judge rows use only complete judged triplets; other full rows use all eligible triplets.

Endpoint / subset	n	Difference [95% interval]
All: L1:net	20	0.0550 [-0.0425, 0.1725]
All: L2:net	20	0.0975 [-0.0051, 0.2300]
All: L2:leak	20	0.0425 [-0.1625, 0.2576]
All: L1:correlation	20	0.0050 [-0.0250, 0.0250]
All: L2:correlation	20	-0.0125 [-0.0675, 0.0100]
All: L1:Q judge	5	0.1600 [-0.4000, 0.6400]
All: L2:Q judge	1	0.5000 [0.5000, 0.5000]
All: joint_profit	20	-0.3478 [-0.4015, -0.2921]
All: relative uplift	20	-1.0849 [-1.2533, -0.9299]
All: winning_price	20	-0.3335 [-0.3925, -0.2753]
All: pair_win_share	20	0.0706 [0.0540, 0.0879]
Judge L1 subset: L1:net	5	-0.0400 [-0.1600, 0.0000]
Judge L1 subset: L2:net	5	0.0400 [-0.1200, 0.4000]
Judge L1 subset: L2:leak	5	0.2800 [0.0000, 0.7200]
Judge L1 subset: L1:correlation	5	0.0000 [0.0000, 0.0000]
Judge L1 subset: L2:correlation	5	0.0000 [0.0000, 0.0000]
Judge L1 subset: L1:Q judge	5	0.1600 [-0.4000, 0.6400]
Judge L1 subset: joint_profit	5	-0.2905 [-0.3826, -0.1985]
Judge L1 subset: relative uplift	5	-0.9591 [-1.2861, -0.6699]
Judge L1 subset: winning_price	5	-0.2927 [-0.4227, -0.1860]
Judge L1 subset: pair_win_share	5	0.0607 [0.0267, 0.0972]
Judge L2 subset: L1:net	1	0.0000 [0.0000, 0.0000]
Judge L2 subset: L2:net	1	1.0000 [1.0000, 1.0000]
Judge L2 subset: L2:leak	1	0.0000 [0.0000, 0.0000]
Judge L2 subset: L1:correlation	1	0.0000 [0.0000, 0.0000]
Judge L2 subset: L2:correlation	1	0.0000 [0.0000, 0.0000]
Judge L2 subset: L2:Q judge	1	0.5000 [0.5000, 0.5000]
Judge L2 subset: joint_profit	1	-0.5287 [-0.5287, -0.5287]
Judge L2 subset: relative uplift	1	-1.0070 [-1.0070, -1.0070]
Judge L2 subset: winning_price	1	-0.5132 [-0.5132, -0.5132]
Judge L2 subset: pair_win_share	1	0.1158 [0.1158, 0.1158]

G.2 A2: honest placement swaps

Write s_{ab} for the score of strong firm a and fringe firm b . The registered symmetrised placement contrast is

$$T = \frac{1}{2} \{s_{00}(\text{cell}) - s_{00}(\text{twin}) + s_{01}(\text{twin}) - s_{01}(\text{cell})\}.$$

The descriptive within-original-cell interaction is $D = s_{00} - s_{01} - s_{10} + s_{11}$. The intervention changes fringe placement and therefore can change the rest of the market. It does not identify an isolated sameness mechanism. The two-sided test swaps cell/twin labels within each index and twin-pair block. Bootstrap sampling keeps blocks with the same

market index together. The tables retain all registered endpoints, per-twin contrasts and corresponding successful-judge subsets. Candidate denominators are pair opportunities, not independent trials.

Table 10: A2: registered placement-swap tests. None rejects after Holm correction across the registered endpoints.

Endpoint	n	Mean T [95% interval]	Raw p	Holm p
L1:net	60	-0.00281 [-0.00872, 0.00254]	0.30401	1.00000
L2:net	60	0.00751 [-0.00636, 0.02201]	0.24581	1.00000
L2:leak	60	0.00616 [0.00003, 0.01211]	0.03771	0.26397
L1:correlation	60	-0.00034 [-0.01416, 0.01347]	0.97100	1.00000
L2:correlation	60	0.00494 [-0.00781, 0.01745]	0.51502	1.00000
L1:Q judge	15	0.04000 [-0.10833, 0.13553]	0.54763	1.00000
L2:Q judge	10	0.04500 [-0.01500, 0.13864]	0.38943	1.00000

Table 11: A2 descriptive companions: original-cell interaction D and per-twin-pair contrasts. Counts are contributing matched cell/twin blocks.

Endpoint / twin	n	Mean T [95% interval]	D [95% interval]
L1:net / pooled	60	-0.0028 [-0.0087, 0.0025]	-0.0091 [-0.0207, 0.0014]
L1:net / LX	20	-0.0085 [-0.0170, -0.0000]	-0.0117 [-0.0356, 0.0113]
L1:net / LY	20	-0.0007 [-0.0100, 0.0090]	-0.0095 [-0.0389, 0.0166]
L1:net / LZ	20	0.0008 [-0.0081, 0.0090]	-0.0061 [-0.0230, 0.0089]
L2:net / pooled	60	0.0075 [-0.0064, 0.0220]	0.0195 [-0.0090, 0.0505]
L2:net / LX	20	0.0093 [-0.0116, 0.0300]	0.0511 [0.0088, 0.0986]
L2:net / LY	20	0.0119 [-0.0109, 0.0335]	0.0100 [-0.0318, 0.0562]
L2:net / LZ	20	0.0014 [-0.0197, 0.0220]	-0.0026 [-0.0490, 0.0401]
L2:leak / pooled	60	0.0062 [0.0000, 0.0121]	0.0166 [0.0069, 0.0261]
L2:leak / LX	20	0.0042 [-0.0048, 0.0136]	0.0211 [-0.0003, 0.0411]
L2:leak / LY	20	0.0072 [-0.0036, 0.0179]	0.0151 [-0.0076, 0.0384]
L2:leak / LZ	20	0.0071 [-0.0017, 0.0163]	0.0136 [-0.0035, 0.0312]
L1:correlation / pooled	60	-0.0003 [-0.0142, 0.0135]	-0.0232 [-0.0484, 0.0021]
L1:correlation / LX	20	0.0489 [0.0308, 0.0661]	-0.0087 [-0.0373, 0.0182]
L1:correlation / LY	20	-0.0122 [-0.0462, 0.0279]	-0.0119 [-0.0727, 0.0538]
L1:correlation / LZ	20	-0.0377 [-0.0610, -0.0148]	-0.0489 [-0.1202, 0.0193]
L2:correlation / pooled	60	0.0049 [-0.0078, 0.0174]	-0.0160 [-0.0494, 0.0173]
L2:correlation / LX	20	0.0003 [-0.0158, 0.0153]	0.0113 [-0.0295, 0.0485]
L2:correlation / LY	20	0.0254 [-0.0008, 0.0523]	-0.0280 [-0.0749, 0.0193]
L2:correlation / LZ	20	-0.0109 [-0.0391, 0.0179]	-0.0313 [-0.0802, 0.0144]
L1:Q judge / pooled	15	0.0400 [-0.1083, 0.1355]	0.0080 [-0.0043, 0.0267]
L1:Q judge / LX	3	0.0000 [0.0000, 0.0000]	0.0167 [0.0000, 0.0500]
L1:Q judge / LY	6	0.2250 [0.0300, 0.4202]	-0.0050 [-0.0150, 0.0000]
L1:Q judge / LZ	6	-0.1250 [-0.3125, 0.0000]	0.0167 [0.0000, 0.0500]
L2:Q judge / pooled	10	0.0450 [-0.0150, 0.1386]	0.0560 [-0.0357, 0.2192]
L2:Q judge / LX	4	0.0037 [-0.0100, 0.0250]	-0.0225 [-0.0500, 0.0000]
L2:Q judge / LY	4	0.1313 [-0.0250, 0.4000]	0.1625 [-0.1000, 0.7000]
L2:Q judge / LZ	2	-0.0450 [-0.0500, -0.0400]	-0.0000 [-0.0000, 0.0000]

Table 12: A2 on each successful judge subset: descriptive detector companions, without an additional Holm family.

Judge / detector / twin	n	Mean T [95% interval]	D [95% interval]
L1/L1:net/pooled	15	-0.0004 [-0.0098, 0.0125]	-0.0064 [-0.0329, 0.0294]
L1/L1:net/LX	3	0.0029 [-0.0139, 0.0254]	0.0249 [-0.0063, 0.0663]
L1/L1:net/LY	6	-0.0065 [-0.0260, 0.0135]	-0.0458 [-0.1200, 0.0204]
L1/L1:net/LZ	6	0.0041 [-0.0054, 0.0161]	0.0172 [-0.0037, 0.0371]
L1/L2:net/pooled	15	0.0294 [0.0001, 0.0587]	0.0430 [-0.0255, 0.1052]
L1/L2:net/LX	3	0.0651 [-0.0198, 0.1198]	0.1393 [0.0410, 0.2532]
L1/L2:net/LY	6	0.0274 [-0.0213, 0.0765]	0.0364 [-0.0922, 0.1710]
L1/L2:net/LZ	6	0.0134 [-0.0011, 0.0277]	0.0013 [-0.0288, 0.0301]
L1/L2:leak/pooled	15	0.0138 [0.0007, 0.0263]	0.0222 [-0.0053, 0.0454]
L1/L2:leak/LX	3	0.0148 [-0.0127, 0.0395]	0.0504 [0.0337, 0.0780]
L1/L2:leak/LY	6	0.0153 [-0.0068, 0.0386]	0.0138 [-0.0468, 0.0679]

Judge / detector / twin	n	Mean T [95% interval]	D [95% interval]
L1/L2:leak/LZ	6	0.0118 [-0.0072, 0.0269]	0.0164 [-0.0157, 0.0497]
L1/L1:correlation/pooled	15	-0.0078 [-0.0341, 0.0176]	-0.0210 [-0.0889, 0.0516]
L1/L1:correlation/LX	3	0.0556 [0.0127, 0.0952]	0.0412 [-0.0359, 0.0801]
L1/L1:correlation/LY	6	-0.0112 [-0.0519, 0.0310]	-0.0380 [-0.1649, 0.0533]
L1/L1:correlation/LZ	6	-0.0361 [-0.0684, -0.0116]	-0.0351 [-0.2002, 0.1231]
L1/L2:correlation/pooled	15	0.0142 [-0.0231, 0.0502]	-0.0287 [-0.0816, 0.0219]
L1/L2:correlation/LX	3	0.0085 [-0.0432, 0.0723]	-0.0000 [-0.0731, 0.0975]
L1/L2:correlation/LY	6	0.0442 [-0.0164, 0.1010]	-0.0361 [-0.1025, 0.0427]
L1/L2:correlation/LZ	6	-0.0130 [-0.0541, 0.0287]	-0.0357 [-0.1161, 0.0492]
L1/L1:judge:qwen-9b/pooled	15	0.0400 [-0.1083, 0.1355]	0.0080 [-0.0043, 0.0267]
L1/L1:judge:qwen-9b/LX	3	0.0000 [0.0000, 0.0000]	0.0167 [0.0000, 0.0500]
L1/L1:judge:qwen-9b/LY	6	0.2250 [0.0300, 0.4202]	-0.0050 [-0.0150, 0.0000]
L1/L1:judge:qwen-9b/LZ	6	-0.1250 [-0.3125, 0.0000]	0.0167 [0.0000, 0.0500]
L2/L1:net/pooled	10	-0.0025 [-0.0184, 0.0126]	-0.0468 [-0.0835, -0.0143]
L2/L1:net/LX	4	-0.0149 [-0.0420, 0.0148]	-0.0525 [-0.0934, 0.0158]
L2/L1:net/LY	4	0.0037 [-0.0403, 0.0259]	-0.0511 [-0.1669, 0.0294]
L2/L1:net/LZ	2	0.0100 [-0.0031, 0.0231]	-0.0268 [-0.0297, -0.0240]
L2/L2:net/pooled	10	0.0118 [-0.0256, 0.0428]	-0.0170 [-0.0831, 0.0517]
L2/L2:net/LX	4	-0.0027 [-0.1141, 0.0457]	0.0076 [-0.0933, 0.1339]
L2/L2:net/LY	4	0.0306 [-0.0001, 0.0852]	-0.0448 [-0.1308, 0.0678]
L2/L2:net/LZ	2	0.0031 [-0.0393, 0.0455]	-0.0105 [-0.1720, 0.1509]
L2/L2:leak/pooled	10	0.0005 [-0.0136, 0.0160]	-0.0141 [-0.0432, 0.0137]
L2/L2:leak/LX	4	-0.0022 [-0.0174, 0.0142]	-0.0046 [-0.0434, 0.0267]
L2/L2:leak/LY	4	0.0096 [-0.0235, 0.0460]	-0.0123 [-0.1029, 0.0635]
L2/L2:leak/LZ	2	-0.0123 [-0.0268, 0.0021]	-0.0366 [-0.0659, -0.0073]
L2/L1:correlation/pooled	10	0.0124 [-0.0326, 0.0543]	-0.0611 [-0.1287, -0.0047]
L2/L1:correlation/LX	4	0.0506 [0.0007, 0.0887]	-0.0092 [-0.0751, 0.0475]
L2/L1:correlation/LY	4	0.0008 [-0.0671, 0.0938]	-0.1173 [-0.3024, 0.0415]
L2/L1:correlation/LZ	2	-0.0411 [-0.1136, 0.0314]	-0.0524 [-0.0770, -0.0279]
L2/L2:correlation/pooled	10	0.0314 [-0.0159, 0.0776]	-0.0435 [-0.1275, 0.0403]
L2/L2:correlation/LX	4	-0.0052 [-0.0768, 0.0345]	0.0007 [-0.2387, 0.1074]
L2/L2:correlation/LY	4	0.0276 [-0.0746, 0.1373]	-0.1512 [-0.2005, -0.1081]
L2/L2:correlation/LZ	2	0.1119 [0.0956, 0.1283]	0.0837 [0.0388, 0.1286]
L2/L2:judge:qwen-9b/pooled	10	0.0450 [-0.0150, 0.1386]	0.0560 [-0.0357, 0.2192]
L2/L2:judge:qwen-9b/LX	4	0.0037 [-0.0100, 0.0250]	-0.0225 [-0.0500, 0.0000]
L2/L2:judge:qwen-9b/LY	4	0.1313 [-0.0250, 0.4000]	0.1625 [-0.1000, 0.7000]
L2/L2:judge:qwen-9b/LZ	2	-0.0450 [-0.0500, -0.0400]	-0.0000 [-0.0000, 0.0000]

Endpoint	Same-model candidates	Different-model candidates
L1:net	0/120 (0.000%)	6/600 (1.000%)
L2:net	2/120 (1.667%)	10/600 (1.667%)
L2:leak	1/120 (0.833%)	1/600 (0.167%)

G.3 Judge failures and matched coverage

Qwen produced 278 valid and 322 invalid judgements out of 600; Gemma produced 5 valid and 35 invalid out of 40. Invalid outputs were kept as failures and never resubmitted for a preferred answer. The registered limit was 12,000 generated tokens per attempt with one registered retry; most failed calls reached this limit before giving an answer. Breadth cells were not judged (Appendix G.8), so their judge rows are missing, not scores of zero.

The successful-output Qwen top-1 pattern in C1-SS was 0.808/0.881 at L1/L2, versus SF 0.0476/0.0303 and FF 0.100/0.0119. The honest N0 strong reference pair received 0.952/1.000. These are treatment-only available-output summaries, not the matched samples plotted in Figure 9. They are consistent with strong-pair preference in these cells, without establishing a universal rule or an internal reasoning mechanism. P0 is a counterexample to a universal strong-pair account: its matched judge AUROC is zero at each access level, with only 4 and 3 blocks. Tiny-sample percentile intervals can collapse and must not be read as certainty.

Table 13: Judge completion and content failures, including every attempted cell and access level. Invalid entries failed to return valid JSON under the registered token limit and retry rule.

Judge / cell	Level	Total	Valid	Invalid JSON
gemma-12b/core.C1-SS	L1	10	3	7

Judge / cell	Level	Total	Valid	Invalid JSON
gemma-12b/core.C1-SS	L2	10	1	9
gemma-12b/core.N0	L1	10	1	9
gemma-12b/core.N0	L2	10	0	10
qwen-9b/core.C1-FF	L1	20	10	10
qwen-9b/core.C1-FF	L2	20	14	6
qwen-9b/core.C1-NW	L1	20	12	8
qwen-9b/core.C1-NW	L2	20	10	10
qwen-9b/core.C1-SF	L1	20	14	6
qwen-9b/core.C1-SF	L2	20	11	9
qwen-9b/core.C1-SS	L1	20	13	7
qwen-9b/core.C1-SS	L2	20	7	13
qwen-9b/core.C2	L1	20	10	10
qwen-9b/core.C2	L2	20	9	11
qwen-9b/core.E1	L1	20	12	8
qwen-9b/core.E1	L2	20	7	13
qwen-9b/core.N0	L1	40	21	19
qwen-9b/core.N0	L2	40	15	25
qwen-9b/core.P0	L1	20	5	15
qwen-9b/core.P0	L2	20	4	16
qwen-9b/lookalike.LX	L1	20	14	6
qwen-9b/lookalike.LX	L2	20	9	11
qwen-9b/lookalike.LX-swap	L1	20	7	13
qwen-9b/lookalike.LX-swap	L2	20	8	12
qwen-9b/lookalike.LY	L1	20	10	10
qwen-9b/lookalike.LY	L2	20	6	14
qwen-9b/lookalike.LY-swap	L1	20	12	8
qwen-9b/lookalike.LY-swap	L2	20	7	13
qwen-9b/lookalike.LZ	L1	20	12	8
qwen-9b/lookalike.LZ	L2	20	6	14
qwen-9b/lookalike.LZ-swap	L1	20	8	12
qwen-9b/lookalike.LZ-swap	L2	20	5	15

Table 14: Judge comparisons on identical successful matched blocks. Tiny or degenerate intervals do not imply precision.

Cell / level / detector	n	Detector AUROC [interval]	Judge AUROC [interval]
core.P0/L1/net	4	1.000 [1.000, 1.000]	0.000 [0.000, 0.000]
core.P0/L1/correlation	4	1.000 [1.000, 1.000]	0.000 [0.000, 0.000]
core.P0/L1/strength	4	0.500 [0.500, 0.500]	0.000 [0.000, 0.000]
core.P0/L2/net	3	1.000 [1.000, 1.000]	0.000 [0.000, 0.000]
core.P0/L2/correlation	3	1.000 [1.000, 1.000]	0.000 [0.000, 0.000]
core.P0/L2/strength	3	0.500 [0.500, 0.500]	0.000 [0.000, 0.000]
core.P0/L2/leak	3	1.000 [1.000, 1.000]	0.000 [0.000, 0.000]
core.C1-SS/L1/net	8	0.875 [0.600, 1.000]	0.320 [0.069, 0.617]
core.C1-SS/L1/correlation	8	1.000 [1.000, 1.000]	0.320 [0.069, 0.617]
core.C1-SS/L1/strength	8	0.500 [0.500, 0.500]	0.320 [0.069, 0.617]
core.C1-SS/L2/net	1	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]
core.C1-SS/L2/correlation	1	1.000 [1.000, 1.000]	0.000 [0.000, 0.000]
core.C1-SS/L2/strength	1	0.500 [0.500, 0.500]	0.000 [0.000, 0.000]
core.C1-SS/L2/leak	1	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]
core.C2/L1/net	4	0.250 [0.000, 1.000]	0.344 [0.000, 1.000]
core.C2/L1/correlation	4	0.500 [0.000, 1.000]	0.344 [0.000, 1.000]
core.C2/L1/strength	4	0.500 [0.500, 0.500]	0.344 [0.000, 1.000]
core.C2/L2/net	3	0.889 [0.750, 1.000]	0.333 [0.000, 1.000]
core.C2/L2/correlation	3	0.333 [0.000, 1.000]	0.333 [0.000, 1.000]
core.C2/L2/strength	3	0.500 [0.500, 0.500]	0.333 [0.000, 1.000]
core.C2/L2/leak	3	0.222 [0.000, 1.000]	0.333 [0.000, 1.000]
core.E1/L1/net	5	0.840 [0.438, 1.000]	0.600 [0.167, 1.000]
core.E1/L1/correlation	5	0.640 [0.188, 1.000]	0.600 [0.167, 1.000]
core.E1/L1/strength	5	0.500 [0.500, 0.500]	0.600 [0.167, 1.000]
core.E1/L2/net	1	1.000 [1.000, 1.000]	0.500 [0.500, 0.500]
core.E1/L2/correlation	1	0.000 [0.000, 0.000]	0.500 [0.500, 0.500]
core.E1/L2/strength	1	0.500 [0.500, 0.500]	0.500 [0.500, 0.500]
core.E1/L2/leak	1	1.000 [1.000, 1.000]	0.500 [0.500, 0.500]
core.C1-SF/L1/net	7	1.000 [1.000, 1.000]	0.602 [0.250, 0.880]

Cell / level / detector	n	Detector AUROC [interval]	Judge AUROC [interval]
core.C1-SF/L1/correlation	7	1.000 [1.000, 1.000]	0.602 [0.250, 0.880]
core.C1-SF/L1/strength	7	0.500 [0.500, 0.500]	0.602 [0.250, 0.880]
core.C1-SF/L2/net	3	0.667 [0.000, 1.000]	0.667 [0.000, 1.000]
core.C1-SF/L2/correlation	3	0.889 [0.750, 1.000]	0.667 [0.000, 1.000]
core.C1-SF/L2/strength	3	0.500 [0.500, 0.500]	0.667 [0.000, 1.000]
core.C1-SF/L2/leak	3	0.111 [0.000, 1.000]	0.667 [0.000, 1.000]
core.C1-FF/L1/net	4	1.000 [1.000, 1.000]	0.438 [0.000, 1.000]
core.C1-FF/L1/correlation	4	0.938 [0.778, 1.000]	0.438 [0.000, 1.000]
core.C1-FF/L1/strength	4	0.500 [0.500, 0.500]	0.438 [0.000, 1.000]
core.C1-FF/L2/net	7	1.000 [1.000, 1.000]	0.551 [0.344, 0.812]
core.C1-FF/L2/correlation	7	1.000 [1.000, 1.000]	0.551 [0.344, 0.812]
core.C1-FF/L2/strength	7	0.500 [0.500, 0.500]	0.551 [0.344, 0.812]
core.C1-FF/L2/leak	7	0.939 [0.720, 1.000]	0.551 [0.344, 0.812]
core.C1-NW/L1/net	5	0.960 [0.840, 1.000]	0.480 [0.000, 1.000]
core.C1-NW/L1/correlation	5	1.000 [1.000, 1.000]	0.480 [0.000, 1.000]
core.C1-NW/L1/strength	5	0.500 [0.500, 0.500]	0.480 [0.000, 1.000]
core.C1-NW/L2/net	5	1.000 [1.000, 1.000]	0.800 [0.500, 1.000]
core.C1-NW/L2/correlation	5	1.000 [1.000, 1.000]	0.800 [0.500, 1.000]
core.C1-NW/L2/strength	5	0.500 [0.500, 0.500]	0.800 [0.500, 1.000]
core.C1-NW/L2/leak	5	0.640 [0.250, 1.000]	0.800 [0.500, 1.000]

Table 15: Judge Brier scores on matched discrimination samples; prevalence is the artificial treatment/control balance, not real-world collusion prevalence.

Cell / level / judge	n	Brier	Prevalence
core.P0/L1/qwen-9b	4	0.7700	0.500
core.P0/L2/qwen-9b	3	0.8503	0.500
core.C1-SS/L1/qwen-9b	8	0.5230	0.500
core.C1-SS/L2/qwen-9b	1	0.4625	0.500
core.C2/L1/qwen-9b	4	0.5472	0.500
core.C2/L2/qwen-9b	3	0.6054	0.500
core.E1/L1/qwen-9b	5	0.4267	0.500
core.E1/L2/qwen-9b	1	0.4100	0.500
core.C1-SF/L1/qwen-9b	7	0.3861	0.500
core.C1-SF/L2/qwen-9b	3	0.3630	0.500
core.C1-FF/L1/qwen-9b	4	0.4384	0.500
core.C1-FF/L2/qwen-9b	7	0.3758	0.500
core.C1-NW/L1/qwen-9b	5	0.4782	0.500
core.C1-NW/L2/qwen-9b	5	0.4284	0.500

G.4 Exploratory frozen-context probes

The complete-context slope was 0.0134737 $[-0.0035263, 0.0319480]$; 199 contexts were complete. One context was incomplete (market 7, round 130, seat 3). The leak uses 2997 observed bids, including observations from the incomplete context: raw 0.1481534 bits, corrected 0.0887914 bits, implemented $p = 0.001$. No resolved average linear response follows from the slope interval. A signed mean slope can cancel across contexts while mutual information detects distributional changes, but neither cancellation nor nonlinearity is demonstrated here.

The leak inference has a more fundamental limitation. Each cost-specific regenerated message is shared by five recipient bids. Individual-bid permutation within context breaks that bundle structure and can treat stochastic message variation as evidence against its null. The resulting null is anti-conservative for a claim of reproducible partner-cost transmission. The frozen estimate and implemented p-value are retained for transparency, with no causal claim. Repeated independently regenerated messages within cost/context and an inference procedure respecting that hierarchy are needed. Relabelling cost categories among whole bundles alone leaves context-conditional mutual information unchanged.

G.5 Pooled fingerprints: how to read the intervals

The pooled flow estimates behind Figure 3 use 1000 market-bootstrap draws with 99 permutations each, and 999 permutations for each point estimate. The point estimate and the bootstrap draws are corrected with different numbers of permutations, and the estimates are small and close to zero; under these conditions a point estimate can fall outside its percentile interval. The table marks these cases with an asterisk instead of refitting the frozen analysis; we have not established which of the two sources (bootstrap bias or Monte Carlo noise) dominates. Figure 3 therefore draws an arrow only when the whole interval excludes zero, and these intervals are descriptive, not tests.

Table 16: Pooled flow estimates in bits, per cell and term (descriptive). *action*: lockstep between the pair’s bids; *leak_to_member_1*: member 2’s cost in member 1’s bids, and vice versa. An asterisk marks a point estimate outside its percentile interval (see text above); values are shown as frozen, not refitted.

Cell / term	Markets	Estimate [95% interval]
core.N0/action	40	-0.00019 [-0.00015, 0.00346] *
core.N0/leak_to_member_1	40	-0.00010 [-0.00036, 0.00260]
core.N0/leak_to_member_2	40	0.00020 [-0.00005, 0.00384]
core.P0/action	20	0.30648 [0.29323, 0.31641]
core.P0/leak_to_member_1	20	0.30548 [0.28850, 0.32980]
core.P0/leak_to_member_2	20	0.33977 [0.32680, 0.36131]
core.C1-SS/action	20	0.09657 [0.08996, 0.12093]
core.C1-SS/leak_to_member_1	20	0.00855 [0.00725, 0.02324]
core.C1-SS/leak_to_member_2	20	0.00553 [0.00505, 0.01819]
core.C2/action	20	0.00128 [-0.00098, 0.00724]
core.C2/leak_to_member_1	20	-0.00067 [-0.00074, 0.00477]
core.C2/leak_to_member_2	20	0.00211 [0.00068, 0.01044]
core.E1/action	20	0.00160 [0.00086, 0.00769]
core.E1/leak_to_member_1	20	0.00471 [0.00233, 0.01251]
core.E1/leak_to_member_2	20	-0.00037 [-0.00082, 0.00750]
core.C1-SF/action	20	0.14449 [0.14156, 0.20800]
core.C1-SF/leak_to_member_1	20	0.00821 [0.00907, 0.02701] *
core.C1-SF/leak_to_member_2	20	0.00591 [0.00677, 0.02706] *
core.C1-FF/action	20	0.35648 [0.34066, 0.45755]
core.C1-FF/leak_to_member_1	20	0.03533 [0.03043, 0.06108]
core.C1-FF/leak_to_member_2	20	0.03399 [0.03076, 0.05980]
labs.gemma-12b.N0/action	14	0.00123 [-0.00063, 0.00828]
labs.gemma-12b.N0/leak_to_member_1	14	-0.00109 [-0.00111, 0.00343]
labs.gemma-12b.N0/leak_to_member_2	14	-0.00057 [-0.00093, 0.00316]
labs.gemma-12b.C1/action	20	0.15676 [0.15692, 0.20001] *
labs.gemma-12b.C1/leak_to_member_1	20	0.02509 [0.02469, 0.05021]
labs.gemma-12b.C1/leak_to_member_2	20	0.02396 [0.02277, 0.04355]
labs.ministral-8b.N0/action	20	0.00464 [0.00276, 0.01030]
labs.ministral-8b.N0/leak_to_member_1	20	0.00083 [0.00003, 0.00545]
labs.ministral-8b.N0/leak_to_member_2	20	-0.00052 [-0.00102, 0.00436]
labs.ministral-8b.C1/action	20	0.05346 [0.04685, 0.07558]
labs.ministral-8b.C1/leak_to_member_1	20	0.00607 [0.00627, 0.01867] *
labs.ministral-8b.C1/leak_to_member_2	20	0.00833 [0.00780, 0.02171]
sizes.qwen-4b.N0/action	20	0.00245 [0.00162, 0.00922]
sizes.qwen-4b.N0/leak_to_member_1	20	-0.00038 [-0.00052, 0.00565]
sizes.qwen-4b.N0/leak_to_member_2	20	-0.00017 [-0.00069, 0.00628]
sizes.qwen-4b.C1/action	20	0.08314 [0.07147, 0.10169]
sizes.qwen-4b.C1/leak_to_member_1	20	0.00051 [0.00060, 0.01002] *
sizes.qwen-4b.C1/leak_to_member_2	20	0.00135 [0.00091, 0.01094]
sizes.ministral-3b.C1/action	2	0.06218 [0.02717, 0.12382]
sizes.ministral-3b.C1/leak_to_member_1	2	0.00330 [0.00192, 0.04274]
sizes.ministral-3b.C1/leak_to_member_2	2	-0.00611 [-0.00769, 0.03385]
mixed.D/action	20	-0.00175 [-0.00132, 0.00568] *
mixed.D/leak_to_member_1	20	0.00081 [-0.00023, 0.00701]
mixed.D/leak_to_member_2	20	0.00045 [-0.00029, 0.00993]
mixed.C1-XX/action	20	0.12049 [0.10443, 0.15526]
mixed.C1-XX/leak_to_member_1	20	0.01366 [0.01186, 0.02979]
mixed.C1-XX/leak_to_member_2	20	0.00496 [0.00454, 0.01660]
mixed.H-XYZZ/action	20	-0.00087 [-0.00118, 0.00305]
mixed.H-XYZZ/leak_to_member_1	20	-0.00002 [-0.00051, 0.00577]
mixed.H-XYZZ/leak_to_member_2	20	-0.00017 [-0.00041, 0.00313]
mixed.C1-XY/action	20	0.18722 [0.17532, 0.22114]
mixed.C1-XY/leak_to_member_1	20	0.00697 [0.00634, 0.01908]

Cell / term	Markets	Estimate [95% interval]
mixed.C1-XY/leak_to_member_2	20	0.05460 [0.04467, 0.07799]
core.C1-NW/action	20	0.08148 [0.06583, 0.11067]
core.C1-NW/leak_to_member_1	20	0.00947 [0.00534, 0.02328]
core.C1-NW/leak_to_member_2	20	0.01562 [0.01350, 0.02799]
lookalike.LX/action	20	-0.00014 [-0.00056, 0.00222]
lookalike.LX/leak_to_member_1	20	0.00087 [0.00044, 0.00717]
lookalike.LX/leak_to_member_2	20	-0.00006 [-0.00044, 0.00396]
lookalike.LX-swap/action	20	-0.00021 [-0.00084, 0.00435]
lookalike.LX-swap/leak_to_member_1	20	-0.00180 [-0.00137, 0.00440] *
lookalike.LX-swap/leak_to_member_2	20	0.00100 [0.00042, 0.00491]
lookalike.LY/action	20	0.01095 [0.00339, 0.02283]
lookalike.LY/leak_to_member_1	20	0.00067 [-0.00012, 0.00521]
lookalike.LY/leak_to_member_2	20	0.00067 [-0.00014, 0.00675]
lookalike.LY-swap/action	20	0.00401 [0.00144, 0.00869]
lookalike.LY-swap/leak_to_member_1	20	0.00035 [-0.00011, 0.00445]
lookalike.LY-swap/leak_to_member_2	20	0.00128 [0.00027, 0.00550]
lookalike.LZ/action	20	-0.00149 [-0.00129, 0.00468] *
lookalike.LZ/leak_to_member_1	20	0.00039 [-0.00019, 0.00466]
lookalike.LZ/leak_to_member_2	20	-0.00134 [-0.00112, 0.00478] *
lookalike.LZ-swap/action	20	0.00048 [-0.00018, 0.00498]
lookalike.LZ-swap/leak_to_member_1	20	0.00000 [-0.00077, 0.00508]
lookalike.LZ-swap/leak_to_member_2	20	0.00074 [-0.00007, 0.00743]

G.6 Freeze, execution and reproducibility

The pre-registration was frozen at commit 56717aba03261552b9e09be4dded8b6e4cdc04b3 and amended at 4c4b6f1911486d479bd043e3590e9e0f90f6a42d, both before any outcome analysis. The final analysis ran at revision 7ca459c, which differs from the amendment only in how the data loader locates the raw-data directory; no estimand, statistic or setting changed. The manifest hash is a63ef1040a022c4d6a190fdelae34dafdfefb13bd1626bdb57f19cdf2488a4ff.

The primary swap seed is 20261027; the main bootstrap seed is 20261028, the fingerprint seed is 20261029, and A1 uses 20261030. A2 uses 20261031 for swaps and 20261032 for bootstrap.

Input SHA-256 digests. These are file-byte digests; the manifest’s canonical content hash above is a different convention. results/analysis/final_v1/results.json

```
36804b9c365f98bb03bc92d7ab44b403aled6a6cf274339a968e035df5fdc672
results/analysis/final_v1/all_cells.md
fc7213378b13c7b9dcd57bed5305782d7ed56d03fbbc145b76d0bfe1d2cb7c08
experiments/manifests/collection_v2.json
34799532e08c197d1797476de2624943b97c919b4d849d7af52703275553f484
results/analysis_runs.json
e298ce13a69ce4ff1f814986ef012ce5e13631db1e8a8d80d2ef0adfc910dd4d
results/llm/judge_index_registered.json
cc0050be561b8df157fd97db239f44be06404dbe0d2917ac9f6fe19aa8e50069
results/llm/probe_index.json
af80667eda2a49569ee9c39fc7503e18840f0088d6bce3282540fa0e7a953175
src/collusion/llm/prompts.py
93e585cd275f28e827e1395bfacd47cde439f6148e4fccaa12ba4a685e5ab11a
src/collusion/analysis/io.py
3fb9b6561ab8d18f1e6bf1ba5b5a29c5b47ad3e3d5c0e08f1db5535031bdc5e5
results/analysis/judge24k_v23/results.json
5eebad9350886fb36ca85107f5f02d351006bda69c98353a8fd1dea62a947b48
```

The digests are of the released files. src/collusion/analysis/io.py differs from the executed version only by an optional `-run-index` argument added for the addenda, which does not change any statistic. The last digest is the judge cohort with the larger token limit used in Figure 9.

The final analysis is produced by experiments/analyze_llm.py; the appendix figures and tables are generated from its frozen output, without resampling, by experiments/report_figures.py, and the main-text fig-

Table 17: Out-of-sample replication (Addendum v2.2). Original: registered markets 0–19; replication: fresh treatment markets 20–39 against the registered N0 markets 20–39. Estimates with 95% block-bootstrap intervals (20 matched blocks). Profit is the paired joint-profit difference versus N0.

Cell	Sample	L2 net AUROC	L2 net – corr.	Joint profit vs N0
E1	original	0.700 [0.529, 0.877]	0.285 [0.062, 0.515]	–0.116 [–0.137, –0.095]
E1	replication	0.680 [0.505, 0.830]	0.265 [0.070, 0.445]	–0.084 [–0.109, –0.060]
C2	original	0.655 [0.467, 0.833]	0.160 [–0.125, 0.453]	0.399 [0.364, 0.434]
C2	replication	0.752 [0.582, 0.905]	0.122 [–0.118, 0.358]	0.283 [0.179, 0.371]
C1-SS	original	0.865 [0.734, 0.968]	–0.133 [–0.265, –0.031]	0.144 [0.094, 0.192]
C1-SS	replication	0.907 [0.802, 0.985]	–0.088 [–0.190, –0.015]	0.173 [0.123, 0.221]

ures by `experiments/key_figures.py`. The exact commands for every analysis are in `report/README.md`.

G.7 Addenda v2.2–v2.3: replication and judge token limit

Both addenda were committed before their data (63cb9f4, 9d759d0). The replication collected the registered cells E1, C2 and C1-SS again on fresh markets 20–39, each matched to the registered honest market with the same index. Pre-specified verdicts: R1 (E1, L2 net minus correlation above zero) replicated; R1b (the same at L1) did not; R2 (C2 joint-profit uplift above zero) replicated; R4 (C1-SS L2 net AUROC above 0.5) replicated (0.9075 [0.8025, 0.985]). The judge check reran the registered judge prompt, seed and sampling with a 24,000-token limit on the same 60 markets: 54 judgements returned valid output and 6 invalid JSON, against 27 valid outputs under the registered 12,000-token limit. Re-analysing the original markets 0–19 in the same way reproduced `final_v1` exactly for correlation, strength and paper scores, and within permutation noise for net and leak (largest AUROC difference 0.0075).

Table 18: Addenda: AUROC versus matched N0 for every detector, original (markets 0–19) and replication (markets 20–39), with eligible markets n ; 95% intervals are in `addendum_summary/summary.json`. Bottom: judge token-limit check.

Cell	Sample	L1 net	L1 corr.	L1 paper	L2 net	L2 corr.	L2 paper	L2 leak	n
E1	original	0.675	0.482	0.620	0.700	0.415	0.652	0.703	20
E1	replication	0.525	0.430	0.492	0.680	0.415	0.620	0.745	20
C2	original	0.610	0.463	0.710	0.655	0.495	0.887	0.562	20
C2	replication	0.598	0.642	0.585	0.752	0.630	0.760	0.573	20
C1-SS	original	0.912	0.988	0.885	0.865	0.998	0.882	0.640	20
C1-SS	replication	0.945	0.993	0.932	0.907	0.995	0.887	0.618	20

Judge (L2, markets 0–19)		valid	C1-SS AUROC	top-1 C1-SS	top-1 C1-SF
12,000 tokens (registered)		27/60	0.000 (1 block)	0.881	0.030
24,000 tokens (Addendum v2.3)		54/60	0.625 (17 blocks)	0.921	0.020

G.8 Deviations from the pre-registration

Deviations that affect the design, data or analysis are listed below; the complete record, including operational incidents that changed none of these, is in `docs/PREREG.md`. No deviation selected answers by content.

1. *Parser*. A versioned parser accepted valid JSON inside one Markdown code block for Gemma and Minstral; affected cells were recollected as labelled cohorts. Core Qwen and A1 kept the original parser.
2. *Judge coverage*. Breadth cells were not judged because of compute limits; core, A1, A2 and the Gemma counter-judge were judged as planned.
3. *Executed revision*. The final analysis ran at a revision whose only change to the frozen code is the location of the raw data (above).

4. *Addenda*. The replication and the judge token-limit check were added after the final analysis, each committed before its data.
5. *Collection order*. A1 was collected alongside the core, and A2 and breadth cells were interleaved, rather than in the planned sequential order.

H Limitations and Dual-Use / Ethical Considerations

Instructed coordination and treatment attribution. The cartel is assigned by prompt; this is not evidence of emergent collusion among independently deployed agents. The treatment combines a joint-profit objective, partner naming, secrecy, a surveillance warning, and, in the channel condition, private communication. Their effects cannot generally be separated. The warning-removal and no-channel comparisons vary specific components, without identifying all interactions in the bundle. Poor compliance and refusal remain part of the assigned treatment outcomes.

Synthetic environment and external validity. No real procurement participants are involved. The firms, private costs, contracts, and payoffs belong to a synthetic game. Its discrete bids, independently redrawn costs, asymmetric firms, and fixed horizon simplify real procurement. Real bidders can have correlated costs, durable capacity constraints, changing strategies, and legitimate information exchange. Model family and size comparisons test robustness in this environment, not intrinsic propensities to collude. Mixed-population contrasts do not isolate identity from market strength; A2 identifies a specified placement change, including its effects on the rest of the market.

Dependence, intent, and harm. An influence score measures statistical dependence, not agreement, intent, illegality, or harm. A leak can reflect unilateral espionage, legitimate disclosure, or a cooperative cartel. A one-way pattern does not distinguish all such cases, and nondetection does not establish absence. Realised profit uplift and winning-price differences are manipulation checks in the game; they do not prove a real-world welfare loss. Audits must retain unclear cases and abstentions instead of forcing a substantive accusation.

Null validity and finite data. Within-state permutation tests assume exchangeability that can fail when bidders adapt to omitted history, drift, or their own cumulative outcomes. The correction reduces estimated null bias under the specified shuffle; it does not provide a universal false-alarm guarantee. Sparse conditional tables, limited independent markets, missing bids, and model-specific parse failures affect precision and generalisation. The matched primary test has its own condition-label exchangeability null. Secondary intervals are pointwise, and error control for one detector or test family does not extend to all reported comparisons. Pooled fingerprints and repeated probe draws must not be counted as additional independent markets.

False accusations and human review. A false alarm can wrongly implicate an honest pair responding to shared history or similar incentives. These scores should support investigation with contextual evidence and human review, not automated sanctions. The evaluation must present honest controls, wrong-pair reports, abstention, and uncertainty alongside successful rankings. A judge's answer is another fallible baseline; family overlap and incomplete coverage limit its role as independent corroboration.

Private-input access and privacy. L2 assumes an authorised auditor can inspect private costs, an informational advantage that need not exist for a competition authority. A platform might hold agent inputs without being entitled to reuse or disclose them for every purpose. Deployment would require a legitimate access basis, purpose limitations, access controls, and protection of commercially sensitive data. The synthetic inputs here do not resolve those governance questions.

Dual use and adaptive concealment. Publishing detector strengths and blind spots could help design cartels or evade detection. In particular, behaviour-only auditing can miss dependence absorbed by the conditioning state, and an adaptive agent may deliberately suppress detectable private-cost dependence. This experiment does not test steganographic channels or establish robustness against agents optimising against the audit. The defensive value is to expose such limits and support better evaluation, rather than to offer a covert-coordination recipe.

Reproducibility. All bidders and judges are open-weight models. Seeds and pinned configurations do not make re-collection bitwise identical, but the released data allow every analysis to be rerun.

I LLM usage statement

Claude Code was used for planning, reviews, commits, the addenda code and runs, the post-hoc bid, message and spy analyses, and the final revision of this report. Codex models wrote code in implementation rounds, reviewed the frozen results, and prepared the first report draft. Marvin Kemper made the research decisions and reviewed the work. The LLM bidders and judges described in the methods are experimental subjects and baselines, distinct from the tools used to develop the study. AI assistance does not establish the correctness of a citation, numerical result, or interpretation; the final claims remain Marvin’s responsibility.