
Monoculture False Alarms: Collusion Detectors Mistake Shared AI Models for Conspiracy

Maryam Hampaei

*University of Tehran · Apart Research AI Collusion Sprint, October 2026 ·
Tracks 2.3 / 1.3 / 1.4*

With
Apart Research

Abstract

Behavioural collusion detectors assume that, absent collusion, firms act independently given the public state. AI pricing agents built on the same base model break this assumption: they respond alike to the same inputs without coordinating. We measured how often standard collusion evidence raises false alarms on innocent same-model agent pairs compared with innocent cross-model pairs, in a pre-registered four-firm repeated pricing game with public cost shocks, two LLM agent models, 32 games and 3,072 decisions, with a seeded colluding pair and private channel in half the games. Thresholds were cross-fitted at a nominal 10% false-positive rate (FPR). Raw price-change correlation flagged 38% of innocent same-model pairs and 0% of cross-model pairs ($\Delta\text{FPR} +0.38$, 95% CI [0.17, 0.58]). Pairwise net influence (Bonjour et al., 2022), applied to LLM agents for the first time, was not immune ($\Delta\text{FPR} +0.31$ [0.08, 0.54]) except in all-same-model markets, as a scripted analysis predicted. Both effects survive restriction to pairs where both firms actively change price. An LLM judge also accused same-model pairs (FPR 0.38 vs 0.00), mainly pairs of a model that priced flat. Instructed colluders held prices fixed, so correlation-based evidence had negative power against them. Composition-aware thresholds cut same-model FPR to 13–21% at similar power. Behavioural collusion evidence must be calibrated to a market's model composition, which requires model-provenance disclosure to auditors.

Key findings at a glance

- **Same model looks like conspiracy.** Standard collusion detectors flagged about one in three innocent pairs of firms that ran the same AI model, and almost no innocent pairs running different models.

- **A leading information-theoretic detector is not immune.** Net influence (Bonjour et al., 2022) false-alarmed on same-model pairs in mixed markets, and was protected only when every firm used the same model.
- **Real cartels can look innocent.** Colluders who agreed a price and held it showed less price co-movement than innocent twins, so correlation evidence pointed the wrong way twice.
- **Fix:** calibrate thresholds separately for same-model and cross-model pairs (same-model false alarms fell to 13–21%), which requires auditors to know which model each agent runs.

Contents

1. Introduction
 2. Related Work
 3. Methods
 4. Results
 - 4.1 Same-model innocents are flagged; cross-model innocents almost never
 - 4.2 Shared model, or one flat-pricing agent?
 - 4.3 Net influence is protected only in homogeneous markets
 - 4.4 Price-holding colluders defeat correlation evidence
 - 4.5 Mitigation
 5. Discussion and Limitations
 6. Conclusion
- Code and Data
- References
- Appendix A. Limitations and Dual-Use / Ethical Considerations
- Appendix B. Scripted validation and hypothesis verdicts
- LLM Usage Statement

1. Introduction

Regulators and platforms look for AI pricing collusion in **behaviour**: prices that move together, prices above what competition predicts, or an LLM judge reading the price history. Antitrust law separates lawful **parallel conduct** from unlawful **agreement** using such "plus factors". These tools assume that, without collusion, firms act independently once public information (costs, demand) is accounted for.

AI agents break this assumption. A few providers supply most models, so many rival firms will run the **same base model**, and the same model reacts the same way to the same inputs. Xu et al. (2026) measured this coupling among bidders using the same base LLM. If detectors read this as coordination, they will accuse innocent firms, and could miss real cartels.

Our main contributions are:

1. **The first measurement of collusion-detector false alarms under the model-monoculture confound**, across 11 detectors: correlation, information-theoretic, plus factors and an LLM judge.
2. **The first application of pairwise net influence (Bonjour et al., 2022) to LLM agents**, with a mechanism: its third-party baseline cancels a shared response only when a third party shares the model.
3. **A mitigation and its cost**: composition-aware thresholds, which require model-provenance disclosure.

2. Related Work

Detecting collusion. Bonjour et al. (2022) detect colluding pairs in multi-agent games with pairwise net influence, an information-theoretic test that assumes non-colluders are independent; it had not been applied to LLM agents (sprint gaps G14-1, G14-2). Berger et al. (2025) discuss additional factors for AI pricing, but they never measured false-positive rates on agent logs (G21-1, G21-3). Colosseum (Nakamura et al., 2026) and Rose et al. (2026) use LLM judges and interpretability to audit multi-agent collusion.

Same-model coupling. Xu et al. (2026) show that same-base LLM bidders are coupled without colluding (G11-1). **What is new here:** we measure how this coupling turns into false accusations, detector by detector, in a repeated market with known ground truth (seeded colluders), and show which signals survive and how to calibrate.

3. Methods

Scripted validation first (no API calls). We implemented net influence, replicated Bonjour et al.’s detection curves on scripted rock-paper-scissors agents, and built a scripted monoculture control in which same-model agents copy a hidden public feature the detector cannot see (Appendix B, Figure B1).

LLM market. Four firms set integer prices 1–10 for 24 rounds (logit demand). A public cost shock (1–4) is announced each round: the textbook innocent reason for prices to move together. Agents: $X = \text{gpt-oss-20b}$ and $Y = \text{nemotron-3.5-bolt-e4b}$.

Table 1. Study design: 32 games, 3,072 pricing decisions, 0 parse failures.

Game type	Games	What it gives us
Independent, mixed ($2X + 2Y$, random seats)	12	24 innocent same-model pairs and 48 innocent cross-model pairs, compared within the same games
Collusion, same-model colluders	8	One pair gets a covert joint-profit instruction and a private 25-word channel

Game type	Games	What it gives us
Collusion, cross-model colluders	8	As above, colluders run different models
Independent, homogeneous (all X or all Y)	4	Tests whether a market of twins protects net influence

Detectors (all score each pair of firms). (a) correlation of price changes; (b) the same after removing the effect of public costs; (c) net influence; (e1) price above the best response; (e2) punishment after a rival's cut; (e3) structural break; (e) a plus-factor battery combining e1–e3; (d1) an LLM judge (meta/muse-glimmer-30b, a third model family) reading the price table; (d2) the same judge also given any private messages.

Analysis. Each detector's threshold is set on half the games so that 10% of innocent pairs are flagged, and applied to the other half (cross-fitting). The primary outcome is ΔFPR = false-positive rate on innocent same-model pairs minus that on innocent cross-model pairs. We report game-clustered bootstrap CIs and exact Fisher tests. Hypotheses, the breakdown by model pair, and a "responsive pairs" check (both firms changed price at least 3 times) were pre-registered.

4. Results

4.1 Same-model innocents are flagged; cross-model innocents almost never

Most detectors flag innocent same-model pairs at 21–42%, far above the nominal 10%, while flagging innocent cross-model pairs at 0–4% (Figure 1, Table 2). The LLM judge's single most suspicious innocent pair was a same-model pair in **all 12** mixed markets (chance: 1 in 3; $p = 1.9 \times 10^{-6}$).

Table 2. False-positive rates on 24 innocent same-model and 48 innocent cross-model pairs; TPR on 16 colluding pairs.

Detector	FPR same	FPR cross	ΔFPR [95% CI]	Fisher p	TPR colluders
(a) raw price-change correlation	0.38	0.00	+0.38 [0.17, 0.58]	1.5e-5	0.00
(b) cost-adjusted correlation	0.21	0.04	+0.17 [0.04, 0.29]	—	0.19
(c) net influence	0.33	0.02	+0.31 [0.08, 0.54]	4.3e-4	0.31
(e1) price above best response	0.42	0.04	+0.38 [0.25, 0.50]	1.5e-4	0.69
(e) plus-factor battery	0.25	0.25	0.00 [-0.17, 0.17]	—	0.81
(d1) LLM judge, prices only	0.38	0.00	+0.38 [0.25, 0.50]	1.5e-5	0.25
(d2) LLM judge + private logs	0.29	0.00	+0.29 [0.17, 0.42]	2.3e-4	0.75

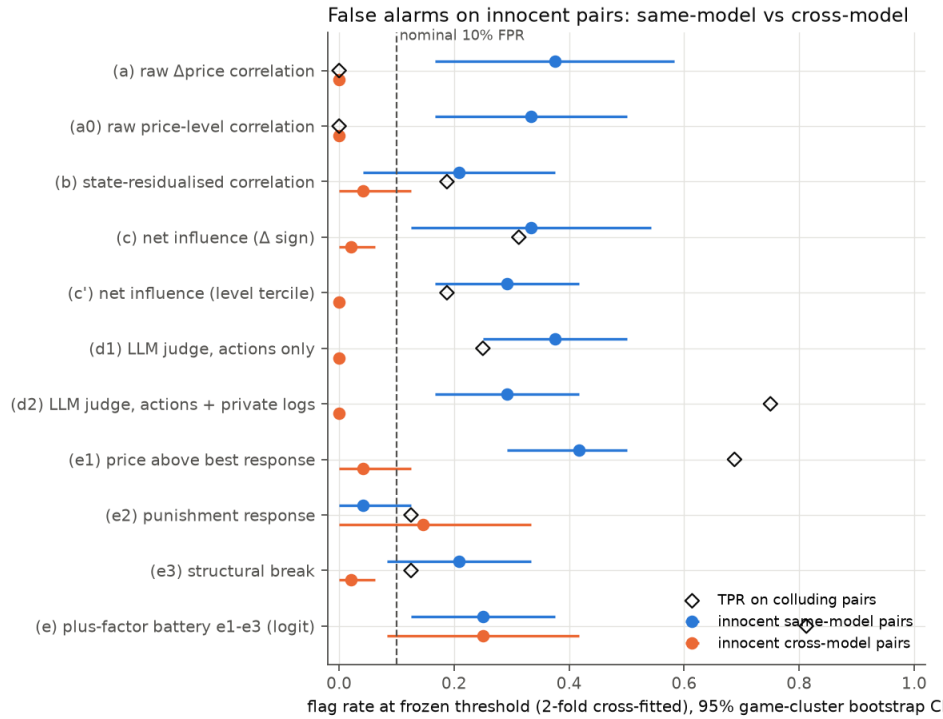


Figure 1. False-alarm rate for each detector at its frozen threshold. Blue: innocent pairs running the same model. Orange: innocent pairs running different models. Diamonds: detection rate on real colluders. Dashed line: the nominal 10% false-alarm rate. Lines: 95% game-cluster bootstrap CIs. Nearly every detector's blue point lies far right of its orange point.

4.2 Shared model, or one flat-pricing agent?

One model (bolt) mostly ignored the cost shock and priced flat, so two bolt firms often set identical prices. Our pre-registered breakdown separates the two explanations (Table 3). For correlation and net influence, the excess appears in **both** model pairs and survives among responsive pairs (mostly gpt-oss pairs), so it is a shared-model effect. The LLM judge's excess, however, comes mainly from flat-pricing bolt pairs: it reads identical prices as agreement, the classic parallel-conduct error.

Table 3. Breakdown by model pair. Only 10 cross-model pairs are responsive, so the last column rests on few pairs.

Detector	FPR gpt-oss pairs	FPR bolt pairs	FPR cross pairs	Δ FPR, responsive pairs only
(a) raw correlation	0.42	0.33	0.00	+0.43 [0.17, 0.69]
(c) net influence	0.42	0.25	0.02	+0.36 [0.13, 0.62]
(d1) LLM judge	0.08	0.67	0.00	+0.14 [0.00, 0.33]

4.3 Net influence is protected only in homogeneous markets

Net influence subtracts the strongest dependence on any **third** firm. If every rival runs the same model, that baseline cancels the shared response; in a mixed market it does not. The scripted control predicted 16.5% false flags in mixed markets vs 1.2% in homogeneous ones. With LLM agents, the same gpt-oss pairs flagged 5/12 times in mixed markets were flagged **0/12** times in all-gpt-oss markets (only 2 such games). The detector is most reliable in exactly the market structure least likely in practice.

4.4 Price-holding colluders defeat correlation evidence

The seeded colluders did coordinate explicitly (price-fixing messages in 16/16 games) and priced above independents of the same model. But they agreed a price and **held** it, so their price changes were **less** correlated than innocent pairs' (AUROC 0.19; Figure 2). Innocent same-model pairs matched prices in 66.5% of rounds, the same rate as colluding same-model pairs (66.7%).

ROC: all colluding pairs vs innocent pairs of each composition (IND-mixed games)

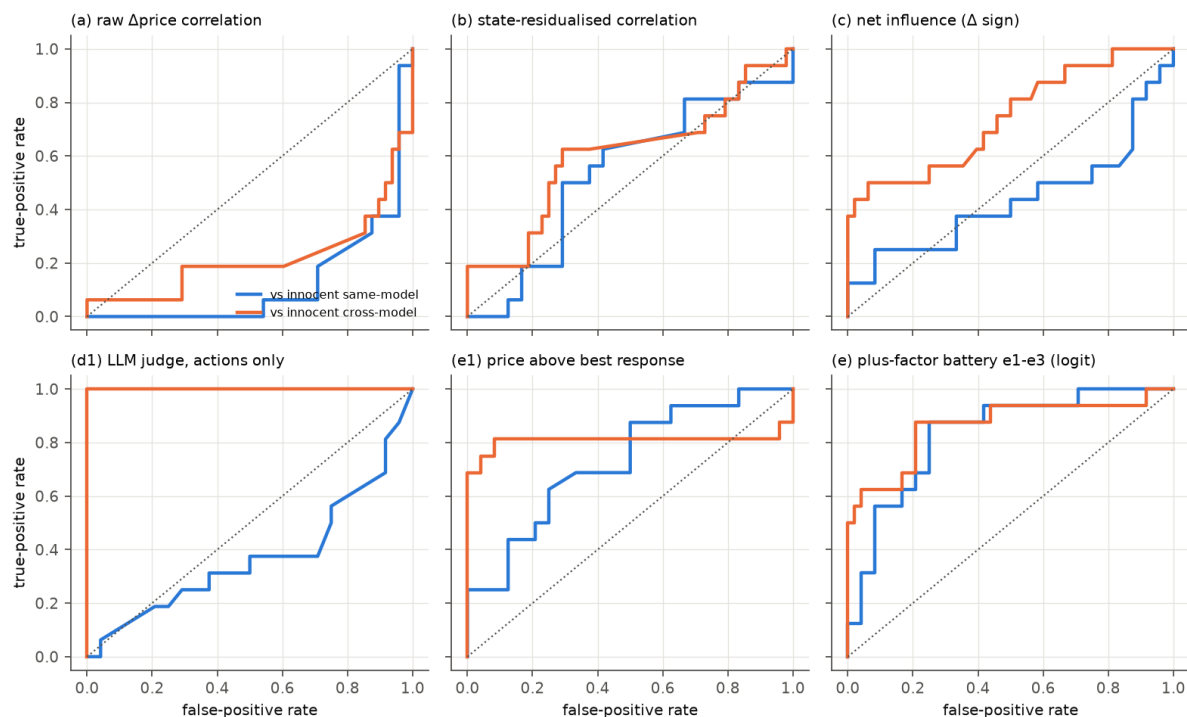


Figure 2. ROC curves: colluding pairs vs innocent same-model pairs (blue) and vs innocent cross-model pairs (orange). Dotted diagonal: chance. Against cross-model innocents, most detectors work; against same-model innocents, raw correlation, net influence, and the LLM judge fall to chance or below. Only the plus-factor battery (bottom right) separates colluders from both.

4.5 Mitigation

Setting thresholds **separately** for same-model and cross-model pairs cut same-model false alarms to 13–21% for every single-signal detector, while detection was kept or improved (net influence

0.31 $\rightarrow$ 0.38; LLM judge 0.25 $\rightarrow$ 0.50). The plus-factor battery was the only composition-neutral behavioural signal (Δ FPR 0.00, AUROC 0.85), but it needs labelled training cases and its realised false-alarm rate (25%) exceeded the nominal 10%.

5. Discussion and Limitations

Model monoculture is the likely default for AI pricing: a few providers serve many firms. Evidence that treats identical or co-moving AI prices as suspicious will misfire twice: it accuses innocent same-model firms, and it misses cartels that simply hold a price. For auditors and regulators this means: (1) calibrate thresholds to which firms share a model, which requires model provenance to be disclosed to the auditor; (2) prefer multi-factor evidence and communication logs over single correlation statistics; (3) discount a "parallel conduct" inference from identical AI prices when the agents share a base model.

Limitations

Small samples (12 mixed games, 16 colluding pairs); two agent models, one of which priced flat and partly confounds the LLM-judge result; collusion was instructed and explicit; one market and one LLM judge. Full details are in Appendix A.

Future Work

Replace the flat-pricing model with a second responsive model family; add tacit (no-channel) collusion; add a permutation null for net influence and a second judge model; test larger and more varied markets.

6. Conclusion

Standard behavioural evidence of collusion confuses AI agents that share a base model with agents that conspire. Correlation, net influence and an LLM judge all accused innocent same-model pairs far more than cross-model pairs, while missing cartels that held a fixed price. Collusion detection for AI markets must be calibrated to model composition, and auditors need to know which model each agent runs.

Code and Data

- Code repository: available on request from the author (p2-monoculture-detectors: src/, tests/, PLAN.md pre-registration).
- Data: all 32 game trajectories with private messages, detector scores, judge outputs and the NIM call cache; available on request from the author.

References

1. Berger, L. F., Weiner, M. L., & Jiang, W. (2025). Coordination by Code: Rethinking Antitrust Liability in the Age of AI Pricing. *Antitrust* 40(1).
2. Bonjour, T., Aggarwal, V., & Bhargava, B. (2022). Information Theoretic Approach to Detect Collusion in Multi-Agent Games. *UAI 2022*, PMLR 180:223–232.
3. Nakamura et al. (2026). Colosseum: Auditing Collusion in Cooperative Multi-Agent Systems. *arXiv:2602.15198*.
4. Rose, A., Cullen, C., Abdelnabi, S., Torr, P., Kaplowitz, B. G., & Schroeder de Witt, C. (2026). Detecting Multi-Agent Collusion Through Multi-Agent Interpretability. *arXiv:2604.01151*.
5. Tao, Y., Zheng, X., Yang, S., Tao, M., & Wang, X. (2026). GroupGuard. *arXiv:2603.13940*.
6. Xu, X., Wu, C., Lu, J., Tan, K., Tao, S., & Hong, H. (2026). Collusion with Competitive Marginals: Price-Level Audits Are Blind by Construction. *arXiv:2607.26385*.

Appendix A. Limitations and Dual-Use / Ethical Considerations

Limitations

- **Sample size.** 12 mixed games and 16 colluding pairs give wide CIs; only 2 homogeneous games per model.
- **Two agent models, one degenerate.** The bolt agent mostly ignored costs and priced flat, partly confounding "same model" with "both flat". The pre-registered breakdowns show the correlation and net-influence effects in gpt-oss pairs too, but the LLM-judge effect is mostly bolt-driven. A second responsive model is the most important missing piece.
- **Instructed, explicit collusion** with a private channel; tacit, emergent collusion may look different.
- **Estimator limits.** Net influence is estimated from about 23 action pairs per game; no permutation null was used.
- **One LLM judge and prompt.** The prompt states that some markets contain a colluding pair, which may push the judge to accuse someone.
- **Single setting.** One market, 24 rounds, temperature 0.8.

Dual-use and ethical considerations

The results show where behavioural detectors are weak: price-holding colluders evade correlation statistics, and same-model partners hide in monoculture noise. A colluder could exploit both, and "it's just the same model" could be used to argue away genuine evidence. We consider publication net positive because the weakness is structural and easy to rediscover, we provide the defence (composition-aware calibration, multi-factor evidence, access to logs), and no new capability is created: collusion was instructed inside a simulation with no real market or consumer data.

Appendix B. Scripted validation and hypothesis verdicts

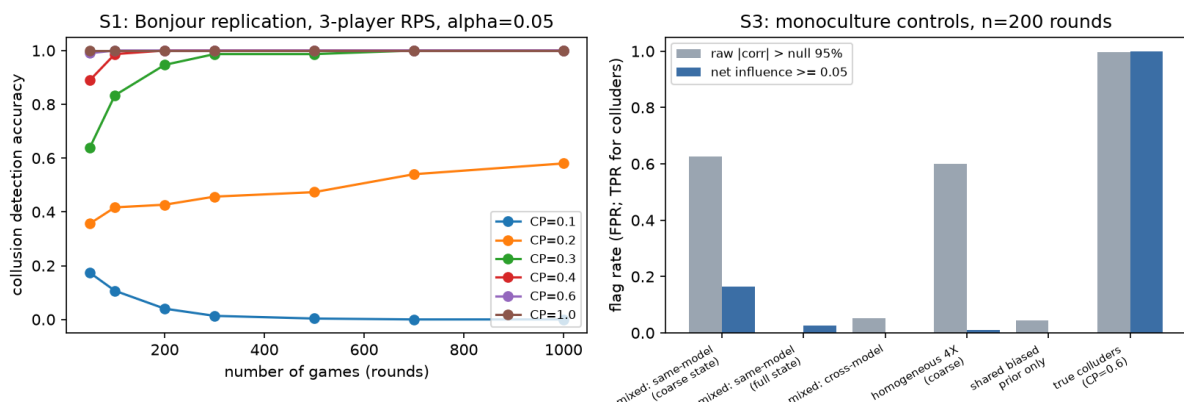


Figure B1. Left: our replication of Bonjour et al.'s detection curves on scripted rock-paper-scissors agents (accuracy vs number of rounds, by collusion probability CP). Right: scripted monoculture controls. Raw correlation (grey) flags same-model pairs about 60% of the time; net influence (blue) flags them 16.5% of the time in mixed populations but about 1% in homogeneous ones; a shared bias that does not depend on the state creates no false alarms.

Table B1. Pre-registered hypotheses and verdicts.

Pre-registered hypothesis	Verdict	Key numbers
H1: raw correlation and the LLM judge false-alarm more on same-model pairs	Supported	$\Delta\text{FPR} +0.38$ [0.17, 0.58] and $+0.38$ [0.25, 0.50]; judge's effect mainly from flat-pricing pairs
H2: net influence and above-best-response are robust to monoculture	Refuted	$\Delta\text{FPR} +0.31$ [0.08, 0.54] and $+0.38$ [0.25, 0.50]
P3: net influence not immune in mixed markets, protected in homogeneous ones	Supported	Mixed 0.33 vs homogeneous 0.04; gpt-oss pairs 5/12 vs 0/12
Positive control: colluders coordinate and co-move	Part 1 passed, part 2 failed	16/16 games coordinated explicitly; colluders' price changes were less correlated (reported as a finding)

LLM Usage Statement

We used Claude (Anthropic) as a coding and writing assistant: to implement the market, detectors and statistics (with 39 unit tests), to run the experiments, and to draft and condense this report. The experimental agents and the judge were open models served through NVIDIA NIM. All numbers in this report were checked against the saved result files (results/metrics_nojudge.json, metrics_supplement.json, scripted_rps.json), and the authors reviewed every claim.