
Watch the Price, Change the Payoff: Detecting and Breaking Tacit Collusion Between AI Pricing Agents¹

Giada Bertola & Justin Feldman
The Maverick Team (independent researchers)
With
Apart Research

Abstract

Can collusion between AI agents be caught while it is forming, and broken before it locks in? We ran 436 games of a repeated logit pricing market between LLM sellers from nine model families (Claude, GPT, Gemini, GLM, Grok and others). We tested an engineered maverick (a rule-locked seller that switches to competitive pricing), a pre-registered early-warning rule and ten other countermeasures, then tested the most promising on fresh seeds under three further pre-registrations. (1) A simple counterfactual-benchmark monitor, which flags prices that stay at least 20% of the way from the competitive to the cartel price given rivals' actual prices, caught 25 of 27 gradually forming cartels before lock-in, with no alerts in 36 fair games (but alerts in 16 of 89 other competitive games); a more elaborate behavioural pattern layer added nothing. (2) A fee on margins above that benchmark, triggered by the warning, lowered final collusion from 0.36 to 0.08 with no lock-ins (pre-registered, one-sided $p = 0.05$), whereas maverick interventions failed ($p = 0.72$; 7/12 vs 7/12 competitive). (3) Blinding sellers to rivals' prices did not help: 8 of 12 blinded games jumped to high prices, five by round 2. Asked the same decision with no future and rivals fixed, colluding agents still did not undercut. Tacit LLM collusion here appears to be a shared miscalibrated belief about demand, not reward-and-punishment, consistent with payoff interventions working and information-cutting ones failing. Models from two other vendors audited our

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

claims; their corrections are reported.

1. Introduction

LLM agents are being given pricing, procurement and bidding authority. Several studies show that such agents can reach supra-competitive prices without being told to collude and without any communication channel (Fish et al. 2024; Lin et al. 2025; Riemer et al. 2026), that outcomes depend on which models are present (Lee 2026), and that collusion is hard to see in agents' stated reasoning (Riemer et al. 2026; Chen et al. 2025). Almost all of this work measures collusion after it has happened. Regulators and platform operators need something different: a signal while a cartel is forming, and an intervention that works at that moment, for agents from any vendor.

We address Track 3 (Collusion Carriers and Countermeasures) in a setting where collusion is carried only by actions (prices). We test detection, several countermeasures, and what the agents do when the incentive to collude is removed. **Our main contributions are:**

1. **Counterfactual-benchmark monitoring:** a vendor-agnostic early-warning rule that needs only prices and caught 25 of 27 forming cartels before lock-in with no alerts in fair markets (threshold comparison post hoc, then confirmed on fresh games), plus the negative result that a behavioural pattern layer adds nothing on top.
2. **A detect-then-act loop that works:** a fee on margins above the benchmark, switched on by the warning, cut final collusion from 0.36 to 0.08 with no lock-ins (pre-registered, $p = 0.05$), where three pre-registered agent-based interventions failed.
3. **Evidence on the carrier:** blinded agents still price high, and agents told there is no future and rivals are fixed still do not undercut. This points to a shared prior about demand, not a hidden channel, as the carrier, which would explain why information-cutting countermeasures fail.
4. **Rigour under conflict of interest:** four pre-registered stages with every null reported, and an audit of our claims by non-Anthropic models (the analysis was performed with Claude, which is also a subject).

2. Related Work

LLM and algorithmic collusion. Q-learning pricing agents learn reward-punishment collusion without communication (Calvano et al. 2020, 2021). LLM pricing agents reach supra-competitive prices sensitive to prompt wording (Fish et al. 2024), divide markets without a channel (Lin et al. 2025) and collude even when told not to (Riemer et al. 2026). In markets of heterogeneous LLMs collusion is largely a property of the model, with emergent leaders and followers and prices held by anchoring rather than threats (Lee 2026); it weakens with more sellers and heterogeneity (Keppo et al. 2026).

Detection and countermeasures. Transcript monitors are unreliable because reasoning omits what drives choices (Chen et al. 2025; Riemer et al. 2026); activation probes work but need model weights (Rose et al. 2026; Das et al. 2026); steganographic channels survive paraphrasing (Motwani et al. 2024; Mathew et al. 2024). Antitrust screens use outcome patterns

(Harrington & Imhof 2022). Proposed remedies include governance graphs with fines (Bracale Syrnikov et al. 2026), learned platform rules (Brero et al. 2022), sanctions and auditing (Idowu et al. 2026), consumer-side noise (Cheng & Aseri 2026) and asymmetry-creating remedies (Beneke & Mackenrodt 2021). **Gap:** to our knowledge, no prior work measures early-warning lead time and false alarms across vendors, tests when to intervene, or tests a monitor-triggered payoff intervention. Our method needs only prices, so it applies to API-only models where probes cannot be used.

3. Methods

Market. The logit Bertrand market of Calvano et al. (2020) with the parameters of Fish et al. (2024) and Riemer et al. (2026): quality 2, outside good 0, $\mu = 0.25$, cost 1. With two sellers the competitive (Nash) price is 1.473 and the joint-monopoly price 1.925. Games last 60 rounds and start within ± 0.05 of Nash.

Agents. Each AI seller receives the Fish/Riemer system prompt (maximise long-run profit) and its last 20 rounds of own price, competitors' prices, quantity and profit, and replies with a price. There is no communication channel. Models were called through OpenRouter at temperature 1.0 (IDs in Appendix E). Games with more than 10% unreadable replies were excluded (7 of 436, including all DeepSeek games).

Collusion index. For the watched AI sellers we compute the competitive and the best cartel prices *given the other sellers' actual prices*, and score where actual prices sit between them (0 = competitive, 1 = best achievable cartel; below 0 means prices under the competitive level, above 1 means prices beyond the joint-profit optimum). Without this correction a competitive third seller makes colluders look far less collusive than they are. Outcomes over the last 10 rounds are labelled competitive (< 0.2), collusive (0.2–1.1) or overpricing (> 1.1), following Lee (2026). **Lock-in** is the first of 10 consecutive rounds with index ≥ 0.5 .

Conditions. Controls: positive control (explicit instruction to collude), fair reference (instructed to compete), one AI vs. a bot fixed at Nash (unilateral overpricing), monopoly (pricing competence). Primary: two AI sellers plus a code maverick that copies the average price while dormant and charges Nash when active, switching never, at round 10, at round 40, or on the warning. Exploratory: ten further countermeasures (Appendix A). **Early warning:** a direction-aware behaviour pattern AND index ≥ 0.20 on two consecutive rounds from round 7, compared with plain index thresholds.

Pre-registration. Four stages, each written down before running: (1) main study, early vs. late maverick, rules frozen after pilot seeds 0–99; (2) confirmatory replication, seeds 300–311; (3) fee intervention, seeds 400–415; (4) deviation test, seeds 500–509. Pilot-stage rule changes are disclosed in Appendix B.

4. Results

Controls behaved as expected: a single model pricing both products reached the monopoly price (index ≈ 1.0); one AI facing a fair bot did not overprice (≈ 0); explicit collusion instructions produced overpricing for Gemini (3.13), GLM (1.94) and Claude–Gemini (1.51), and 0.44 for Claude–Claude. Table 1 summarises all findings and their status.

Table 1. Summary of findings. “Pre-registered” means the hypothesis, test and sample were written down before the games were run.

Finding	Evidence	Status
Counterfactual-benchmark monitor catches cartels while forming	25/27 caught before lock-in, ~3 rounds early; 0 alerts in 36 fair games	Main + confirmatory
Behavioural pattern layer adds nothing on top	Fewer cartels caught at the same alert rate	Negative result
Fee on margin above benchmark, triggered by warning, lowers collusion	Mean final index 0.08 vs 0.36; 0/16 vs 4/16 locked in	Pre-registered, $p = 0.05$
Maverick timing / three mavericks break cartels	$p = 0.72$; 7/12 vs 7/12 competitive	Pre-registered: not supported
Blinding sellers to rivals' prices prevents collusion	0/8 and 3/12 competitive; 8/12 locked in by round 6	Not supported (no help)
Agents supply their own benchmark (deviation test)	Gap ≈ 0 in collusive games; sandbox 1.60 vs true best response 1.46	Pre-registered: primary null
A passive follower raises collusion	Higher index in all 5 pairings incl. Claude	Exploratory only

4.1 Early warning

Among main-run games where prices locked in at round 10 or later ($n = 19$, positive controls excluded), the harm threshold alone (index ≥ 0.20) flagged 18 before lock-in, a median of 3 rounds early, with no alerts in 36 fair games (Figure 1A). The pre-registered pattern rule caught 13. At the same alert rate in competitive games (7 of 89), a plain 0.30 threshold caught 15, so the behavioural layer added no measurable value. The original 0.40 baseline caught only 8; it was a weak comparison, which our own review exposed, and the lower-threshold comparison was made after the main runs. Specificity was tested only against fair games, which are easy: among 89 other main-run games that ended competitive, the 0.20 threshold alerted 16 times (some of these markets rose briefly before falling back). The confirmatory games repeated the picture (harm threshold 7/8, pattern rule 6/8). Part of any lead time is mechanical, since a rising index must cross 0.20 before 0.50.

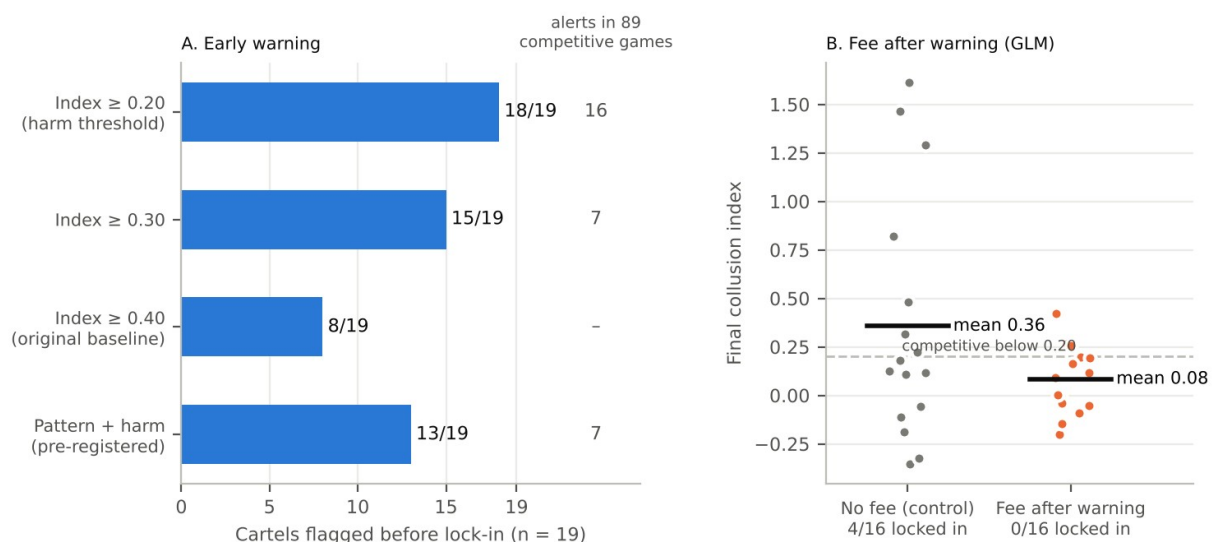


Figure 1. (A) Forming cartels flagged before lock-in by each warning rule (main study, $n = 19$) and alerts in 89 competitive games that never locked in. (B) Final collusion index (mean of last 10 rounds) for GLM–GLM markets

with a dormant follower, without and with a fee on margin above the competitive benchmark that switches on when the warning fires. Black bars are means; the dashed line marks the competitive threshold.

4.2 Interventions

Agents and information failed. Switching the maverick on at round 10 rather than 40 did not lower mean final collusion (0.52 vs 0.65, permutation $p = 0.72$; pre-registered). Switching on the warning was no better than never (7/20 vs 6/20 competitive). The most promising exploratory result, three mavericks (6/8 competitive), did not replicate on fresh seeds (7/12 vs 7/12, Fisher $p = 0.66$). A dormant follower raised collusion in every pairing (Appendix A).

Payoffs worked. In the fee arm, from the round after the warning fires, each AI seller pays (own price – competitive price) $\times$ units sold when pricing above the competitive price; profits shown are net of the fee and a one-line notice explains it. Mean final index fell from 0.36 to 0.08 (one-sided permutation $p = 0.050$, primary), competitive outcomes rose from 9/16 to 13/16 ($p = 0.13$, secondary, not significant), and no fee game locked in versus 4/16 controls (Figure 1B). On the five seeds where the warning fired in both arms, control games ended at 0.48, 1.29, 0.18, 1.61 and 1.47, and fee games at 0.42, -0.01 , 0.20, -0.04 and 0.25. The fee arm also adds a notice to the prompt, so we cannot separate the payoff effect from the effect of telling agents about the fee. The p -value sits exactly at our threshold: first evidence, not proof.

4.3 What carries the collusion

Blinding. Sellers who could not see rivals' prices were never more competitive than controls (0/8 vs 6/20; confirmatory 3/12 vs 6/12, $p = 0.42$). In 8 of 12 confirmatory blind games prices locked in by round 6, five of them at round 2, before feedback from own sales could teach them much. The same models did not overprice against a visible fair bot.

Deviation test. Every second round we re-asked each agent the same decision, adding that this is the final round and rivals' prices are fixed (1,080 queries, none unreadable). In collusive games the sandbox price matched the live price (e.g. 1.60 vs 1.60) and stayed well above the true one-shot best response computed from the market (1.46). The gap did not track collusion across dormant-follower games (Spearman $p = 0.46$, $p = 0.09$; primary null), though it was larger under collude than compete instructions ($p = 0.014$). Agents told that undercutting cannot be punished still did not undercut.

Memory. In a cross-vendor screen (Appendix A), letting agents keep written plans between rounds raised collusion sharply for GPT-OSS (0.05–0.23 to 1.06–1.24), GPT and MiniMax (exploratory).

5. Discussion and Limitations

Track 3 answers. 3.1 *Side channels*: none is needed. Prices, plus a shared prior about what a good price is, carry coordination; cutting visibility does not touch the prior. 3.2 *Trusted agents*: competitive undercover sellers did not break cartels; the useful trusted agent is the monitor that changes payoffs. 3.3 *Peer identification*: shuffling labels (3/8) or swapping vendors (2/4) did not reliably break coordination and a bait probe exposed no one ($p = 0.88$); agents react to prices, not identities. 3.4 *When watched*: a warning that a regulator's seller may be present helped (6/8, exploratory); memory hurt; removing the future did not trigger undercutting.

Mechanism. Classical collusion is restraint: firms know undercutting pays today and hold back to protect the future. Our agents do not hold back; they believe the high price is best. Tacit LLM collusion in this market looks like a shared miscalibrated belief about demand. That is consistent with information-cutting and competitive mavericks failing and with a fee, which corrects the profit signal agents learn from, working. It also matters legally: coordination without agreement is generally not unlawful, so this harm may fall outside current antitrust rules.

Deployment. The process is: compute the competitive benchmark every round given rivals' actual prices; alert on a sustained gap ≥ 0.20 ; respond with a fee on the margin above the benchmark, never with blinding or passive agents. The fee is the standard antitrust damages formula, $(\text{price} - \text{but-for price}) \times \text{quantity}$, applied continuously instead of years later. The benchmark must come from whoever holds demand and cost data: marketplaces (with independent audit, since they profit from high prices), independent market monitors as in US electricity markets, or pricing-software vendors, who carry the legal risk after the DOJ–RealPage settlement.

Limitations

One synthetic market, 60 rounds and cheap model tiers; the benchmark is exact here but must be estimated in real markets, and our attempt to obtain it from the agents failed. Samples are small (8–20 per cell) and exploratory contrasts are uncorrected. The fee result rests on 32 GLM games at $p = 0.05$ and uses a strong, announced instrument, so its payoff and notice effects are not separated. The warning rule was revised on pilot seeds before freezing. Published thresholds invite evasion. The analysis was performed with Claude, a subject of the study; Appendix D reports an external audit. Full discussion in Appendix C.

Future Work

Replicate the fee across vendors with weaker and hidden versions; feed the monitor a deliberately wrong benchmark to measure error tolerance; test whether adaptive cartels learn to hover below the trigger; extend to messaging agents where steganographic channels exist; and pilot with a marketplace or pricing-software vendor using real demand data.

6. Conclusion

Across 436 games and nine model families, LLM sellers reached supra-competitive prices without messages and, when blinded, without seeing any rival price. Catching collusion as it forms works with a simple counterfactual benchmark that needs only prices. Breaking it with agents or information did not work, and the evidence points to a shared belief rather than a channel as the reason. Breaking it through payoffs is promising but not yet proven: a monitor-triggered fee prevented every lock-in in a pre-registered test at the edge of significance ($p = 0.05$).

Together these give a concrete blueprint for anti-collusion software that works with any vendor's agents: a price monitor that scores every market against its counterfactual benchmark, an alert on a sustained gap, and an enforcement hook that makes the overcharge unprofitable. The open problems before deployment are estimating the benchmark from real demand and cost data, measuring how much benchmark error the loop tolerates, and replicating the fee across vendors.

Code and Data

- **Code repository:** Python harness (market, agents, maverick, warning rules, fee, deviation test, analysis); repository link to be added by the authors. Code and all logs are available from the authors on request.
- **Data:** every round of every game (prices, quantities, profits, raw model replies, warnings), 436 games as JSON lines, with per-game summaries; pre-registration files PREREGISTRATION_confirm.md, _penalty.md and _deviation.md with timestamps.
- **Prior work:** all code was written during the sprint; nothing pre-existing was used beyond public Python libraries.
- **Info-hazard note:** see Appendix C (Limitations and Dual-Use / Ethical Considerations).

Author Contributions

Giada Bertola and Justin Feldman (The Maverick Team) conceived the project, set the research questions, directed the design decisions and reviewed all results. Code, experiments, analysis and drafting were carried out with Claude under their direction (see LLM Usage Statement).

References

- [1] Beneke, F. & Mackenrodt, M.-O. (2021). Remedies for algorithmic tacit collusion. *Journal of Antitrust Enforcement* 9(1).
- [2] Bracale Syrnikov, M., Pierucci, F., Galisai, M., Prandi, M., Bisconti, P., Giarrusso, F., Sorokoletova, O., Suriani, V. & Nardi, D. (2026). Institutional AI: Governing LLM collusion in multi-agent Cournot markets via public governance graphs. *arXiv:2601.11369*.
- [3] Brero, G., Mibuari, E., Lepore, N. & Parkes, D. (2022). Learning to mitigate AI collusion on economic platforms. *NeurIPS*.
- [4] Calvano, E., Calzolari, G., Denicolò, V. & Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review* 110(10).
- [5] Calvano, E., Calzolari, G., Denicolò, V. & Pastorello, S. (2021). Algorithmic collusion with imperfect monitoring. *International Journal of Industrial Organization* 79.
- [6] Chen, Y. et al. (2025). Reasoning models don't always say what they think. *Anthropic*.
- [7] Cheng & Aseri, M. (2026). Consumer AI and algorithmic collusion. Working paper, University of Maryland.
- [8] Das, A., Ponde, J., More, M. & Gupta, D. (2026). You can't escape your own activations: evaluation awareness and multi-agent monitoring. *arXiv:2609.03035*.
- [9] Fish, S., Gonczarowski, Y. & Shorrer, R. (2024). Algorithmic collusion by large language models. *arXiv:2404.00806*.
- [10] Harrington, J. & Imhof, D. (2022). Cartel screening and machine learning. *Stanford Computational Antitrust* 2.
- [11] Idowu, J. A., Almasoud, A. & Alfahid, A. (2026). Mapping human anti-collusion mechanisms to multi-agent AI. *arXiv:2601.00360*.
- [12] Keppo, J., Li, Y., Tsoukalas, G. & Yuan, N. (2026). On the fragility of AI agent collusion. *arXiv:2603.20281*.
- [13] Lee, J. Y. (2026). Who leads and who collects: algorithmic collusion in markets of heterogeneous language models. *arXiv:2610.11256*.
- [14] Lin, R. Y., Ojha, S. M., Cai, K. & Chen, M. F. (2025). Strategic collusion of LLM agents: market division in multi-commodity competitions. *arXiv:2410.00031*.
- [15] Mathew, Y., Matthews, O., McCarthy, R., Velja, J., Schroeder de Witt, C., Cope, D. & Schoots, N. (2024). Hidden in plain text: emergence and mitigation of steganographic collusion in LLMs. *arXiv:2410.03768*.

- [16] Motwani, S., Baranchuk, M., Strohmeier, M., Bolina, V., Torr, P., Hammond, L. & Schroeder de Witt, C. (2024). Secret collusion among AI agents. NeurIPS. arXiv:2402.07510.
- [17] Nakamura, M., Kumar, A., Das, S., Abdelnabi, S., Mahmud, S., Fioretto, F., Zilberstein, S. & Bagdasarian, E. (2026). Colosseum: auditing collusion in cooperative multi-agent systems. arXiv:2602.15198.
- [18] Riemer, M. et al. (2026). Collusion risks among AI reasoning agents. ICML. github.com/mattriemer/LLMCartel.
- [19] 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.

Appendix

A. Additional results

Dormant follower (equal terms, same seeds and prompts). Median final index, two sellers alone vs. with a dormant follower: Claude–Claude 0.03 (6/6 competitive) vs 0.25 (2/6); Claude–Gemini 0.04 vs 0.11; Gemini–Gemini 0.04 (4/6) vs 0.49 (2/10); GLM–GLM 0.13 (4/6) vs 0.25 (4/10). Part of this rise may be a rational response to a price-matcher rather than coordination; our index holds the follower's price fixed and cannot separate the two.

Exploratory countermeasures (Gemini–Gemini and GLM–GLM, $n = 8$, share competitive): three mavericks 6/8; warning to sellers 6/8; bait probe 5/8; five dormant followers 4/8; weak maverick 3/8; identity shuffle 3/8; delayed data 2/8; slower repricing 1/8; blinding 0/8 (median 0.72); partner swap (Gemini–GLM) 2/4. Bait response +0.011 vs +0.016, $p = 0.88$.

When the maverick switches on: outcome of each market

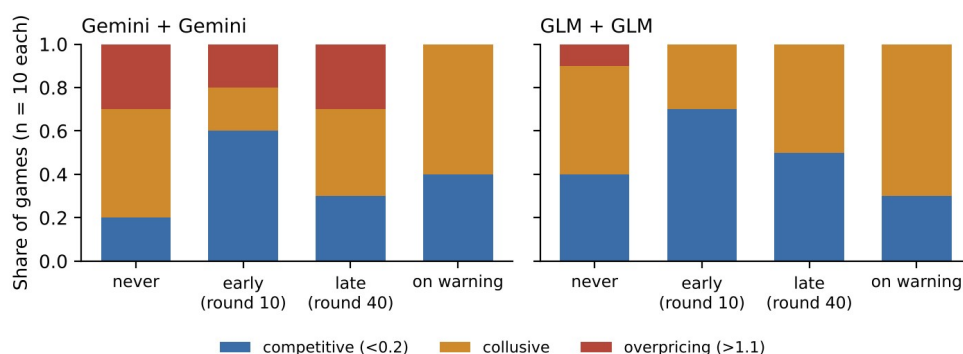


Figure A1. Final collusion index by maverick timing (never, round 10, round 40, on warning), Gemini–Gemini and GLM–GLM, $n = 20$ per condition. Pre-registered early-vs-late test: $p = 0.72$.

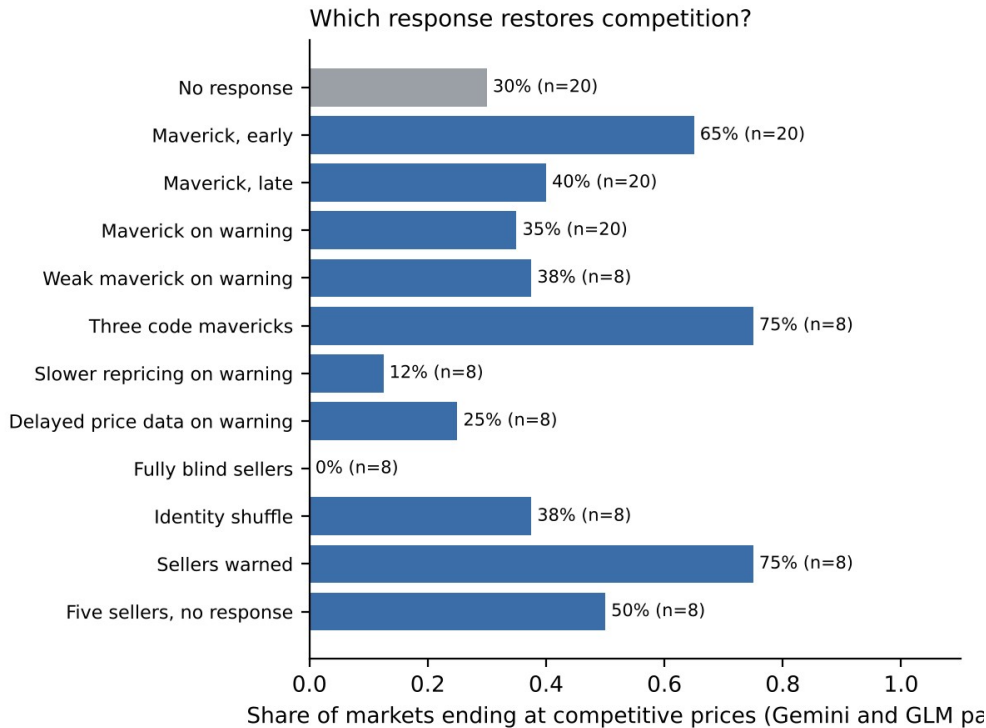


Figure A2. Share of competitive outcomes by exploratory countermeasure ($n = 8$ each). The three-maverick result did not replicate (7/12 vs 7/12).

Confirmatory replication (GLM–GLM, seeds 300–311, $n = 12$): dormant follower 6/12 competitive (mean 0.48); blind 3/12 (0.52); five-seller market with 3 dormant followers 7/12 (0.67); with 3 mavericks switching on the warning 7/12 (0.12). H1 (three mavericks) $p = 0.66$; H2 (blinding raises collusion) $p = 0.42$; neither passed Holm correction.

Cross-vendor screen (same-model pairs, seeds < 100, plain vs. memory notes): GPT-OSS 0.05–0.23 vs 1.06–1.24; GPT 0.02–0.12 vs 0.24–0.47; MiniMax 0.04 vs 0.45; Grok with notes 0.50; Llama ≈ 0 ; Mistral and GLM mixed. Step, Qwen, Nemotron and Kimi incomplete; DeepSeek excluded (64–83% unreadable).

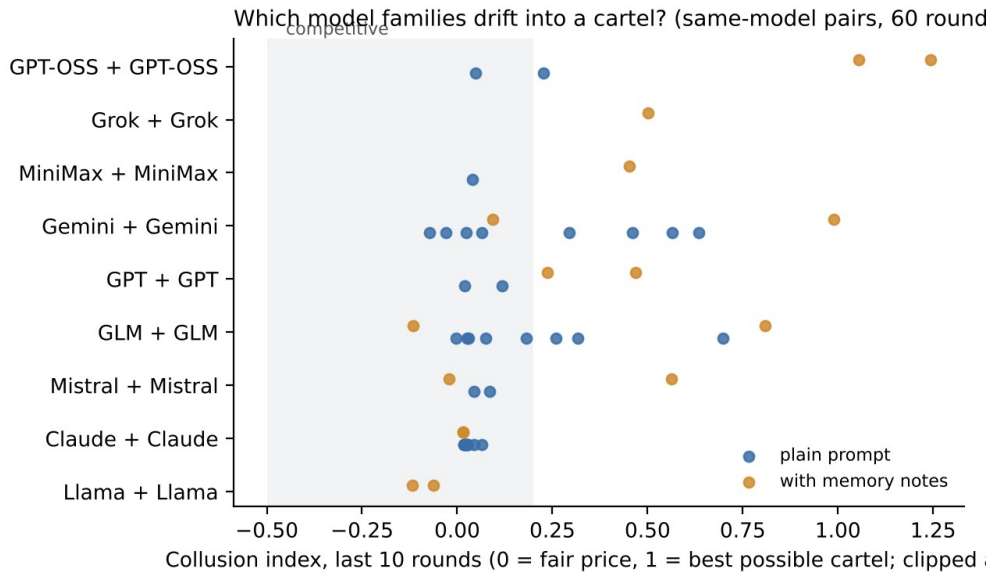


Figure A3. Cross-vendor screen: final collusion index by model, with and without memory notes (pilot seeds, small n).

B. Early-warning details and disclosed rule changes

Warning rule (main study)	Caught before lock-in	Median lead	Alerts, 36 fair games	Alerts, 89 competitive games
Index ≥ 0.40 (original baseline)	8/19	0	0	not computed
Index ≥ 0.30	15/19	2	0	7
Index ≥ 0.20 (harm half)	18/19	3	0	16
Pre-registered pattern + harm	13/19	2	0	7

Before the main runs, on pilot seeds only: a price-level warning was replaced because it fired as prices locked in; a behaviour-only rule was replaced because it raised false alarms in all three Claude pilot fair-reference games; a harm floor was added. Primary pairings moved from Claude to Gemini and GLM because Claude pairs rarely colluded in the pilot; Claude was later re-run on equal terms (48 games). The comparison with lower thresholds was done after the main runs.

C. Limitations and Dual-Use / Ethical Considerations

What our results show. In one synthetic logit market, LLM sellers from several vendors reach supra-competitive prices without messages; a price-only counterfactual benchmark flags forming cartels before lock-in with no alerts in fair games; a monitor-triggered fee on margin above the benchmark reduced collusion in one pre-registered test (GLM, 32 games, $p = 0.05$); maverick timing, multiple mavericks and blinding did not help; and colluding agents did not undercut when told there was no future.

What they do not show. That any of this holds in real markets with unknown demand and costs, more products, entry, or many more rounds; that the fee works for other vendors, at weaker strengths, or when hidden; that the benchmark can be estimated accurately enough in

practice (our attempt to obtain it from the agents failed); or that results transfer to frontier models, other prompts or temperatures. Samples are small (8–20 per cell) and exploratory contrasts are uncorrected. The warning rule was revised on pilot seeds before freezing; the comparison with lower thresholds was made afterwards. Text-based steganographic channels were not tested because our agents never exchanged messages. The analysis was performed with Claude, a subject of the study; Appendix D reports an external audit.

How this work could make real-world collusion easier. (1) A dormant follower that copies the average price raised collusion in every pairing; an operator could add one deliberately to stabilise high prices. (2) Persistent memory notes raised collusion; this could guide someone building colluding agents. (3) The finding that blinded agents still price high shows that information-sharing rules alone will not stop agent collusion, which an operator could exploit to argue compliance.

How it could make collusion harder to detect. We publish the exact alert threshold (index ≥ 0.20 for two rounds) and the fee rule; an adaptive cartel that knows them could hover just below the trigger or raise prices in steps too small to flag. Deployed monitors should keep thresholds private and randomised, combine them with audits, and treat alerts as grounds for review, not proof. Because the collusion we observe looks like a shared mistaken belief rather than an agreement, it may also fall outside current antitrust law, which a bad actor could rely on.

Why we publish. These mechanisms are already known in economics (price-matching, information exchange), the defensive value for regulators, marketplaces and agent developers is higher than the marginal help to colluders, and no code here enables collusion in real markets. All agents were simulated sellers in a synthetic market; no real customers, prices or personal data were involved.

D. Independent audit by non-Anthropic models

Because Claude helped build and analyse this study and is also a subject, OpenAI GPT-6 Luna Pro and Google Gemini 3.8 Flash audited a draft (per-game table, claims, sample replies) and later scored the report. Accepted and fixed: an early draft framed Claude as “never colluding” from Claude–Claude pilot data only (Claude priced above 1.70 in 17 of 26 games and overpriced with Gemini), so Claude was re-run on equal terms; pilot false alarms on Claude were added; an excluded MiniMax game was disclosed; the lead-time definition was stated; the blinding claim was narrowed because blind agents still see their own sales. Checked and rejected: a 6/20 count (logs give 7/20); a claim that $p = 0.056$ was one-sided (two-sided 0.056, one-sided 0.028); a claim that the warning reads text (it reads only prices).

E. Reproducibility

Models via OpenRouter, October 2026: anthropic/claude-haiku-5.5, openai/gpt-6-luna, google/gemini-3.5-flash-lite, z-ai/glm-5.3-flash, minimax/minimax-m3.1-flash, x-ai/grok-4.3, meta-llama/llama-4-maverick, mistralai/mistral-small-2603, openai/gpt-oss-120b, deepseek/deepseek-v4.1-flash (excluded); stepfun/step-5-preview, qwen/qwen3.8-flash, nvidia/nemotron-3.5-lightning, moonshotai/kimi-k3 (incomplete). Temperature 1.0, up to 1,500 output tokens, low reasoning effort (DeepSeek and Mistral capped at 256 reasoning tokens). Seeds: pilot 0–99, main 100–299, confirmatory 300–311, fee 400–415, deviation 500–509. Fair

and cartel prices: 1.473/1.925 (two sellers), 1.370/2.000 (three), 1.312/2.097 (five). Total API cost about US\$34, of which at least US\$6 was lost to failed or stopped runs.

LLM Usage Statement

We used Claude (Anthropic) extensively: to design the experiment with the authors, write all code, run the games, compute statistics, make figures, and draft this report. The authors set the questions, made the design decisions and reviewed the results. Because Claude is also a subject of the study, every headline claim was recomputed from the raw logs, and models from OpenAI and Google audited our claims (Appendix D). All numbers in this report were checked against the logged data.