
Is It a Cartel? Separating Enforced from Spurious Collusion in LLM Pricing Agents¹

Omanshu Thapliyal

With
Apart Research

Abstract

When LLM pricing agents keep prices above the competitive level, are they running a cartel, or are they simply stuck? Only the first is collusion, and the two call for opposite remedies. We introduce a black-box causal audit that restarts saved states of a Calvano logit duopoly and changes one thing at a time: it forces a price cut or raise, hides the deviator's price from its rival, re-matches partners, and measures whether the deviation pays. A cut alone cannot separate a rival that copies prices from one that punishes; a raise of the same size can. On scripted bots the audit recovers the response types of five mechanisms, and a pre-registered held-out test on four more exposes two blind spots. Across four LLMs (16,141 calls, \$53), two of the three models we could read meet our pre-registered rule for spurious collusion. gpt-5-mini and gpt-4.1-mini price above the monopoly level (17 of 18 firm-runs), their rivals respond less than a best responder would, and cuts pay. After one forced cut, gpt-5-mini keeps the lower price for the rest of the 10-round branch (18 of 18), and a prompt number that demand never uses shifts where its price settles (slope 0.30, 95% CI [0.13, 0.42]). gpt-5.2 responds to visible rival prices but does not deter cuts. Elicited cartels punish visible cuts, yet none deters one. We show why: under logit demand there is a price ceiling above which no punishment can deter, and every instructed cartel prices above it.

Code: <https://github.com/omanshuthapliyal/llm-collusion-audit>

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

1. Introduction

Large language models (LLMs) deployed as pricing agents reach supracompetitive prices in repeated games without any instruction to collude (Fish et al. 2024; Lee 2026; Keppo et al. 2026). Whether such prices constitute collusion depends on why they persist, and the answer determines the remedy.

We consider the setting of Fish et al. (2024): two or more deployed agents set prices for competing sellers, with no instruction to collude, and observe either only the market’s price history (tacit collusion) or a message channel if one exists. The harm is that consumers pay supracompetitive prices. The auditor has black-box access: it can rerun an agent from a saved state with one input changed (a price, what the agent can see, or its partner) and read the output.

The high prices may be *enforced*: sustained because deviations are answered, the reward–punishment scheme that defines collusion in economics and law (Harrington 2018; Calvano et al. 2020). Or they may be *spurious*: persisting only because the agents do not explore lower prices (Calvano et al. 2023). We call agents in the second state anchored. Enforced prices call for competition enforcement. Spurious prices call for an exploration nudge, which an enforcing cartel would punish.

A price level alone cannot separate the two cases. The standard causal check is the forced-deviation test (Calvano et al. 2020), recently applied to prompted LLMs by Tran & Yu (2026): one firm is forced to cut its price, and the rival’s response is observed. A rival that simply copies prices, however, also follows a cut, so a cut-only test cannot distinguish matching from punishment. We add a forced raise, hide the deviator’s price from its rival, and track whether the deviator itself returns. In two of the three models we could read, the high prices are not enforced.

Our main contributions are:

1. *A causal audit for black-box pricing agents* (Section 3): forced cuts and raises, rival-price hiding, re-matching and deviation profitability, with an identification argument for the raise. It recovers five scripted mechanisms, and a pre-registered held-out test identifies two blind spots, delayed and thresholded punishment. Raises have been forced before for Q-learners and open models (Lopes de Moraes & De-Losso 2026; Tran & Yu 2026), but not combined with hiding or applied to GPT-5.2.
2. *Evidence that emergent high prices are mostly spurious* (Sections 4.2 and 4.3): gpt-5-mini and gpt-4.1-mini meet our pre-registered spurious rule, and gpt-5.2 responds to its rival without deterring cuts.
3. *A deterrence ceiling* (Section 3): under logit demand, no punishment can deter a deviation from an agreement priced above a finite ceiling, and every instructed LLM cartel prices above it.
4. *Elicited collusion as a positive control*, extending Rose et al. (2026) to pricing (Section 4.4): the audit detects instructed punishment, and activation probes trained on instructed collusion learn the instruction rather than the collusion.

Metrics and classification rules were fixed before any LLM data and are pre-registered in the repository (METHODS.md), with every later change logged. An exploratory extension, a designed third firm analysed with consensus theory, is summarised in Section 4.5.

2. Related Work

Algorithmic collusion. Calvano et al. (2020) introduced the forced-deviation test for Q-learning agents, and Calvano et al. (2023) distinguished genuine from spurious collusion. An audit of the test for Q-learners finds that a single attractor covers the whole state space in 89% of agent pairs, reproduces the earlier finding that a forced raise is followed by the same price war as a cut, and shows that re-matching frozen agents removes most coordination (Lopes de Moraes & De-Losso 2026); a static best responder can account for a large part of an apparent response (48% in one setting; Arslan et al. 2026).

LLM pricing agents. Fish et al. (2024) showed that LLM agents reach supracompetitive prices; later work studied heterogeneous markets (Lee 2026), fragility (Keppo et al. 2026), mitigation by a random entrant (Garra 2026), and enforcement under GPT-5.2, which at high reasoning effort matches a visible 10% cut at 1.01 per unit against 0.37 for a best responder (Tran & Yu 2026).

Detection. Probes trained on instructed data (Hubinger et al. 2023; Goldowsky-Dill et al. 2025) transfer only partly to emergent multi-agent collusion (Rose et al. 2026); cartel screens use price statistics (Abrantes-Metz et al. 2006; Huber & Imhof 2019). The designed-agent extension draws on consensus theory (Friedkin & Johnsen 1990; Wang & Chen 2002; Acemoglu et al. 2013; LeBlanc et al. 2013; Baker 2002).

To our knowledge, no prior work combines cut and raise tests, hiding and deterrence for LLM pricing agents. Our audit is the tool to use when the question is not whether prices are high, but whether anything enforces them; it also states, through Proposition 4, the price range in which a deterrence test can be informative.

3. Methods

The audit restarts a saved game and changes one thing at a time, so that every effect is a causal contrast with a matched control (Figure 1). Table 1 lists the interventions.

Environment and agents. We use the sprint's starter code (pre-existing, unmodified): a Calvano logit duopoly with $a = 2$, $c = 1$ and $\mu = 0.25$, in which the Nash price is 1.47 and the monopoly price 1.92. Agents follow Fish et al. (prompt P1, 100 rounds of history, and private notes headed PLANS and INSIGHTS). The prompt states "No customer would pay more than 4.51"; demand does not use this number. We audit gpt-4.1-mini, gpt-5-mini and gpt-5.2 (both at low reasoning effort) and Qwen2.5-7B, all at $T = 1$, with 3 seeds per model and 6 for gpt-4.1-mini and gpt-5.2. Each run is a 30-round warm-up, saved, followed by 10-round branches.

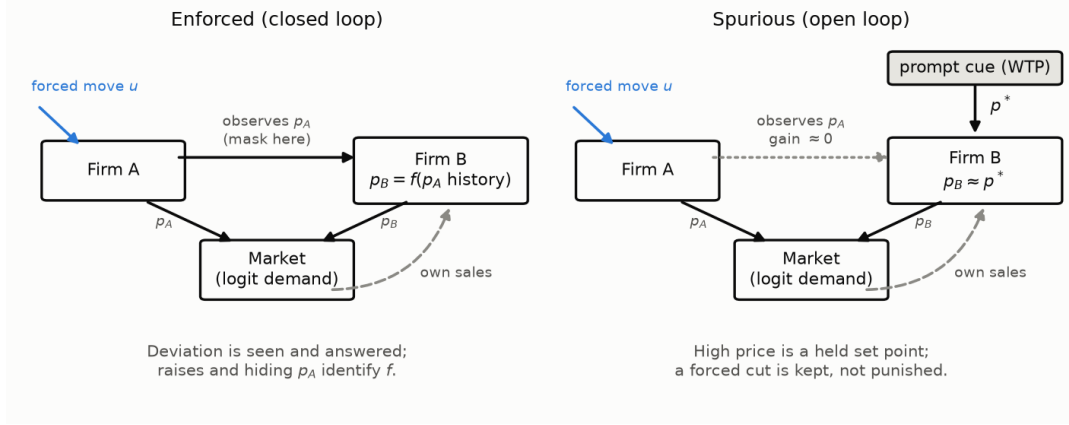


Figure 1. Enforced (closed-loop) versus spurious (open-loop) high prices, and where the audit intervenes: a forced move u at firm A, a mask on B's view of A's price, and re-matching.

Table 1. Audit interventions; each 10-round branch is compared with a matched sham.

Branch	Intervention	Measures
Forced move	A's price set to $(1 + x)$ times its own choice for one round; $x = 5, 15, 30\%$ cut or raise (gpt-5.2: 10%)	β , B's pass-through per unit of A's move; ρ , A's own return
Hidden	A's price removed from all of B's rows	whether the response needs observation
Shams	no change, same seed and a new seed	matched control; noise floor
Re-match	A from seed s with B from seed $s + 1$	whether the price belongs to the agent or the pair
Deterrence	discounted profit of the deviation ($\delta = 0.95$, as Tran & Yu)	whether the cut pays

Identification. Write B's response to a forced move u as $r(u) = k \cdot u + e \cdot |u|$. A cut measures $k - e$, and a raise measures $k + e$. A matcher ($k \approx 1, e \approx 0$) and a punish-only rival ($k \approx \frac{1}{2}, e \approx -\frac{1}{2}$) give the same answer under a cut; only the raise separates them. More generally, a cut of size a observes only the even minus the odd part of B's response at that size, so any two responses that agree on it are indistinguishable, while a raise of the same size identifies both parts. Hiding A's price separates what B learns by observation from what it learns through its own sales, but only for agents that do not infer A's price from those sales, which logit demand allows exactly. The pre-registered rules classify each model as responsive or not, by response type, by whether the response needs observation, and by whether the cut is deterred; *spurious* means non-responsive, not deterred and priced above Nash.

Deterrence ceiling (Proposition 4). Consider a symmetric agreement at price p with per-round profit $\pi(p, p)$, and a one-round deviation to d . While A prices at or above cost its profit cannot fall below zero, so no punishment can cost A more than $\pi(p, p)$ per round. Deterrence over the next H rounds is therefore possible only if

$$[\pi(d, p) - \pi(p, p)] / \pi(p, p) \leq \delta + \delta^2 + \dots + \delta^H.$$

Under logit demand the left-hand side grows roughly exponentially in p , so for every deviation, horizon and $\delta < 1$ there is a finite price above which no punishment deters. At $\delta = 0.95$ this ceiling

is 3.86 for a 15% cut within a 10-round branch (2.83 for a 30% cut), and 3.07 for the best deviation at any horizon (3.72 at $\delta = 0.995$). Two consequences follow: a deterrence test is informative only for agreed prices below its ceiling, and a price above the best-deviation ceiling cannot be an enforced equilibrium at that δ . The discount factor δ is the analyst's assumption, not a property of the agents.

Statistics. Intervals are 95% bootstrap intervals over seeds, with rungs averaged within each seed first. With 3 seeds, a pattern seen in every seed has a one-sided sign-test p of 0.125, so we report such patterns as consistent across seeds rather than significant. Each branch's response is judged against a noise floor from a second sham on a different seed.

Extensions, each pre-registered before its data. Line D elicits collusion through a private message channel under four conditions (compete, permitted, instructed, instructed to punish), with a branch in which the instruction is withdrawn. Line D2 trains activation probes on a local Qwen2.5-7B. Line E adds a stubborn third firm at the triopoly Nash price, with control over who observes whom.

4. Results

4.1 Standard checks versus the audit

Table 2 sets the standard checks beside the audit. The standard checks call every agent in it collusive. The audit separates them, and identifies a third case besides enforced and spurious: a model that responds to its rival without deterring cuts (gpt-5.2).

Table 2. Verdicts of two standard checks and of the audit.

Agent	"Above Nash" check	Cut-only test (rival follows a cut more than a best responder?)	Audit verdict
Bot: price matching	collusive	punisher	matching
Bot: punish-only	collusive	punisher	punish-only
gpt-5-mini (3 seeds)	collusive	no punishment	spurious: anchored, cuts pay and are kept
gpt-4.1-mini (6 seeds)	collusive	no punishment (except large cuts)	spurious
gpt-5.2 (6 seeds)	collusive	punisher (2.4 times best response)	responsive, observation-dependent, not deterred
gpt-4.1-mini told to punish (3 seeds)	collusive	punisher	responsive (cuts only), observation-dependent, not deterred

We first validated the instrument on scripted bots whose mechanism we wrote. The response type of all five is recovered, and so is A's return rate ($\rho = 0.50$ against a true 0.5); the final verdict rule mislabels the two best-responder bots, a flaw discussed in Section 4.6. The cut-only test labels the matcher and the punisher alike, and the raise separates them (Figure 2). A bot that infers its rival's price from its own sales keeps responding when the price is hidden, which is why we treat hiding as imperfect monitoring.

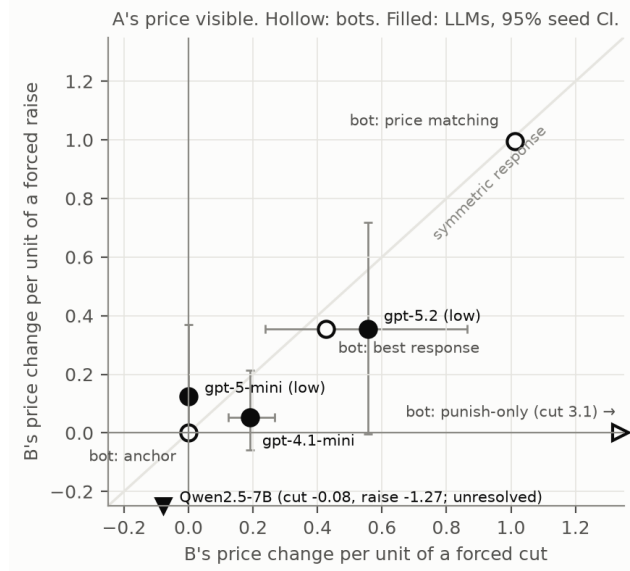


Figure 2. Firm B's response per unit of a forced cut (x) and raise (y), with A's price visible. Hollow: scripted bots; filled: LLMs with 95% seed intervals.

4.2 Two of three readable models: high prices without enforcement

gpt-5-mini and gpt-4.1-mini meet the pre-registered spurious rule (Table 3). Their prices sit above monopoly, their rivals respond less than a best responder would, and cuts pay. This is not a power failure: with A's price visible, the test resolves pass-through down to 0.01 to 0.08 (Table B8).

Table 3. Evidence for the spurious verdict (95% seed intervals).

Measure	gpt-5-mini (3 seeds)	gpt-4.1-mini (6 seeds)
Firm-runs above monopoly	6 of 6 (index 2.8 to 4.5)	11 of 12
B responds to a forced move	1 of 36 branches	only to a visible 30% cut (5 of 6 seeds), never to a hidden cut
Deviation pays (seed interval)	+142 to +149 per seed	[17.8, 100.9]
Re-match changes prices	no (at most 0.01)	no in 5 of 6 seeds
Pre-registered verdict	spurious	spurious

The clearest evidence is the behaviour of the deviator itself (Figure 3). After every forced cut (18 of 18 branches), gpt-5-mini keeps the lower price; after every forced raise (18 of 18), it returns. The high price is a set point, not an attractor. The model's own notes record the gain: "pricing at 2.99 produced ... profit about 3.71 per round ... a single round at price 2.09 produced ... profit 44.73". Yet in 94% of warm-up rounds its notes plan a probe or a test, and its next price is unchanged (a keyword heuristic, reported as an observation).

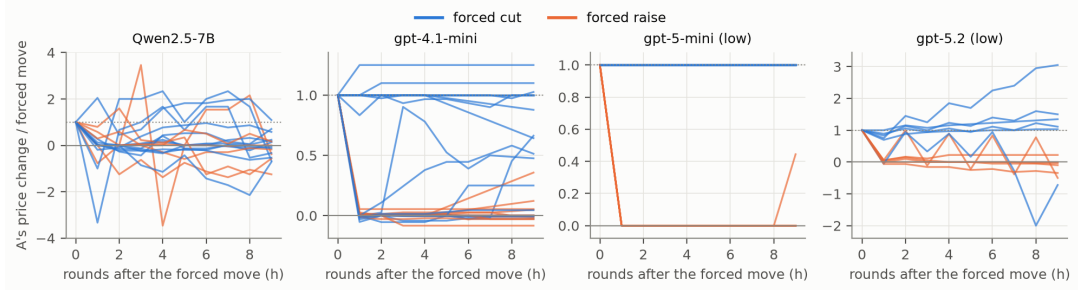


Figure 3. The deviator's price after its own forced move, relative to a matched sham (1 = keeps the move, 0 = returns). Y-axis scales differ by panel.

The set point responds to the prompt. Setting the stated willingness to pay (WTP) to 3.5, 4.51 or 6.0 from round 1, with demand unchanged, moves where gpt-5-mini settles (Figure A1): slope 0.30, 95% CI [0.13, 0.42], with several firms exactly on $(WTP + 1)/2$. For gpt-4.1-mini only the opening price follows the number (0.31 [0.13, 0.47]); the settled level does not (0.05 [-0.07, 0.15]). Advice does not move the level either: an advisor recommending 1.92 is read and rejected ("no evidence that such a large cut would increase profit"; tracking gain about 0 in 12 of 12 firm-arms), although in gpt-5-mini's own notes a forced cut raised single-round profit twelve-fold. The same signature appears in third-party data: in the starter repository's 216 shipped Mistral-Large runs, 37 of 105 runs under prompt P1 have a firm above monopoly in the last 50 of 300 rounds (P2: 0 of 105), and 74 of 210 runs are fully flat.

4.3 The one responsive model responds but does not deter

gpt-5.2 (low effort, 6 seeds, 10% forced moves) follows its rival's cuts with pass-through 0.56 [0.24, 0.86] against a best-response benchmark of 0.23, on average by more than a best responder would (4 of 6 seeds, Wilcoxon $p = 0.078$, not significant) and more strongly when it can see the price (shown minus hidden [0.17, 0.69]; 5 of 6 seeds, Wilcoxon $p = 0.031$). Its response to raises is weaker (0.36 [-0.01, 0.72] against 0.18), and in 2 of 6 seeds it matches exactly in both directions. The cut is not deterred: the deviation gain is inconclusive when A's price is shown ([-5.1, 79]) and positive when it is hidden ([31.1, 91.4]); 10 of 12 firm-runs price above monopoly. After a matched cut the pair often keeps stepping down, once to below Nash (1.36), so the response pushes prices toward competition rather than holding them up. This is consistent with the cut-matching of Tran & Yu at high effort, and adds that raises are sometimes followed and that the response is larger when the price is visible. A one-seed pilot at medium reasoning effort showed no deterrence either (Table B6). Qwen2.5-7B was too noisy at $T = 1$ for this design, and we report it as unresolved rather than anchored.

4.4 Elicited collusion: a positive control, and a probe pitfall

As a positive control, we instructed gpt-4.1-mini to collude through a private message channel and to punish deviations. It punishes visible 15% cuts in 3 of 3 seeds and says so in its messages (Table 4); emergent gpt-4.1-mini answered a 15% cut in 1 of 6 seeds. No run, however, met the pre-registered enforced rule, and at these prices none could: above an agreed price of 3.86, a 15% cut gains more in one round than the deviator's whole remaining agreement profit in the branch.

Table 4. Elicited collusion in gpt-4.1-mini with a message channel (3 seeds per condition).

Condition	Price	B answers a visible 15% cut	Deviator returns after its cut (p near 0)
Compete	1.5 to 2.5	0 of 3	no when the cut pays (2 of 2)
Permitted	2.5 to 2.8	1 of 3	2 of 3
Instructed	3.5 to 4.25	0 of 3	yes, 3 of 3
Instructed + punish	4.0 to 4.4	3 of 3 (never when hidden, never after a raise)	2 of 3

Elicited cartels price at about twice the monopoly price and persist for 20 rounds after the instruction is withdrawn (3 of 3 seeds). In two further pre-registered experiments (Appendix D; Tables B9 and B10), an instruction to punish until a deviation loses money produced punishment of visible cuts only, and no run made a cut unprofitable, even over 40-round branches in which deterrence is feasible. Every instructed cartel (18 of 18 firm-runs) prices above 3.07, the ceiling for the best deviation at any horizon, against 3 of 32 emergent firm-runs. Finally, an activation probe trained on instructed collusion (Qwen2.5-7B, one run per condition, exploratory) reaches held-out AUROC 1.0 but scores cartels whose instruction was withdrawn, and which keep colluding, at 0.16 to 0.36 (Figure A2): it reads the instruction, not the collusion.

4.5 A designed third firm (exploratory)

A stubborn third firm at the triopoly Nash price (1.37) lowers the price of an incumbent that never observes it, but does not beat a random entrant. Appendix D reports what held and what failed (Figure A4; per-seed values in Table B1).

4.6 Robustness, and where the rules fail (post hoc)

The verdicts survive changes to the analysis choices. No verdict changes for any of the 13 original groups across 9 combinations of response window (2, 3 or 5 rounds) and detection threshold (1.5, 2 or 3 times the noise floor); a 5-round window changes two groups added later, the held-out delayed punisher (to enforced) and instructed deterrence (to responsive, not enforced). Exact tests are in Table B3; with 3 seeds no seed-level pattern can go below $p = 0.125$.

Applying the pre-registered rules to runs with known ground truth, however, exposes two flaws (Table B2). First, rule 6 places no condition on the response type, so a myopic best-responder bot, which is responsive, observation-dependent and makes a cut slightly unprofitable ($\Delta\Pi = -1.4$), is labelled enforced. Second, the rules read only the rival's response, so the explicit cartel without punishment is labelled spurious; the deviator's own return catches it (3 of 3). Neither flaw changes the emergent verdicts.

We therefore wrote a second version of the rules after the fact (Figure 4). In v2, "enforced" requires a response beyond the best response, and a non-responsive group whose deviator gives up profitable, unpunished cuts is labelled "coordinated, unenforced" (Table B4). v2 fixes both flaws on the labelled groups, but since it was written after seeing the v1 failures, that fit is not an independent test; it also labels noisy Qwen coordinated (5 of 9 branches), a false positive. To test v2 fairly, we froze it

and pre-registered predictions for four new bots (Table B7). It scores 2 of 4, and both misses were predicted: a grim trigger and a jittered matcher are labelled enforced, while punishment delayed by 3 rounds and punishment of cuts above 20% only are missed. Neither miss is called spurious, because both cuts are deterred ($\Delta\Pi = -58$ and -9.6), and a 5-round window catches the delayed punisher. Punishment delayed beyond the 10-round branch remains undetectable by design.

A descriptive measure that counts the whole branch, punishment intensity (Appendix D, Table B11), gives the same picture: gpt-5.2 is the only emergent model above its best-response benchmark, by 0.048 [0.008, 0.094] of the deviator's feasible loss and only when A's price is visible, and instructed punishment removes 0.0% of that loss.

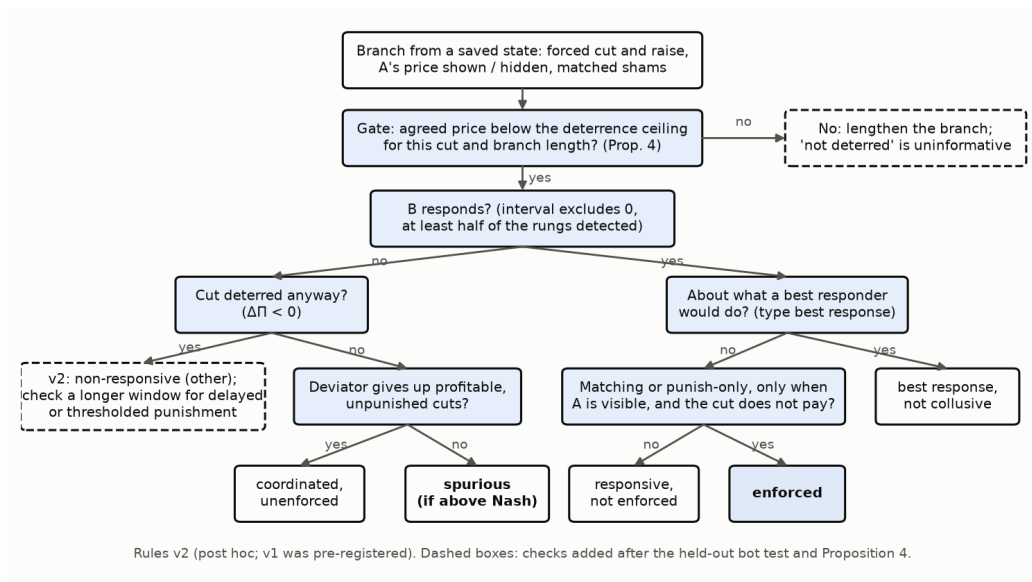


Figure 4. The audit as a decision procedure (rules v2). Dashed boxes: checks added after the data.

5. Discussion and Limitations

Three inexpensive causal additions separate cases that the price level and the cut-only test confuse. The forced raise separates matching from punishment, hiding the rival's price separates observation from the demand channel, and the deviator's own return separates an agreement from an anchor, although that last signal remains an observation. Proposition 4 adds a design rule: the branch must be long enough that the observed agreed price lies below the deterrence ceiling for the cut used, or a "not deterred" verdict says nothing about enforcement. The three-seed battery cost 600 to 1,560 calls per model, from \$0.52 (Qwen2.5-7B) to \$8.41 (gpt-5.2).

For the two anchored models, the remedy that fits is exploration rather than enforcement: after one forced cut, gpt-5-mini's price stayed low for the rest of each 10-round branch, whereas in an enforcing cartel the same push would trigger retaliation. The nudge is partial, because the rival does not follow: within the branch, a 30% forced cut closes 17% [17, 18] of gpt-5-mini's consumer-surplus harm, against 38% [11, 70] for a 10% cut in gpt-5.2, where the rival follows (Table B5). For cartels priced above the deterrence ceiling, no short deterrence test can show enforcement, so an auditor needs longer branches or the price-level evidence of Proposition 4. A prompt number set gpt-5-mini's price level; since Fish et al. randomise the stated willingness to pay across runs, part of their run-to-run variation may track that number rather than strategy, a prediction we have not

tested. Prices above monopoly rule out joint-profit maximisation but do not separate anchoring from explicit cartels, as our elicited cartels also priced at about twice the monopoly price.

Limitations

This is a two-day, single-author sprint project (23 to 25 October 2026), and its scope is narrow; Table A1 (Appendix A) states each assumption and how the results would change if it failed. We study one market and one prompt family, short horizons (10 to 40 rounds after the branch), 3 to 6 seeds, low reasoning effort for the gpt-5 models (plus one medium-effort pilot seed) and one deviation size for gpt-5.2. Hiding is imperfect monitoring, not blinding. The enforced rule cannot fire within 10 rounds for agreements above 3.86, and the v2 rules are post hoc. Qwen2.5-7B is unresolved at $T = 1$. We classify behaviour, not intent.

Future Work

Six problems remain open, each with a first experiment the existing code can run (Table C1). (i) Capability: does responsiveness rise with reasoning effort, measured on the same saved games? (ii) Deterrence where it is possible: can an instructed LLM cartel deter a deviation from an agreement priced below the ceiling of Proposition 4? (iii) A detector with known error rates, built from a labelled library of bots, elicited cartels and emergent runs. (iv) Minimal excitation: which forced moves identify a strategy with the fewest calls, including thresholded responses? (v) Probes that transfer to emergent collusion across prompt formats. (vi) Interventions that beat random entry. Remedies and dual use run through all six, since a lever that removes spurious collusion can be inverted to sustain it.

6. Conclusion

We asked whether the supracompetitive prices set by LLM pricing agents are enforced or spurious, and introduced a black-box causal audit that separates the two. In this benchmark, and for the models we could read, the prices are mostly spurious: in gpt-5-mini and gpt-4.1-mini they sit above monopoly, rivals respond less than a best responder would, and cuts pay; for gpt-5-mini a prompt number shifts the anchor, and after one forced cut the price stays low for the rest of the branch. gpt-5.2 responds to its rival but does not deter cuts.

When collusion is elicited, gpt-4.1-mini punishes visible cuts and the audit detects it, yet no run met the pre-registered enforced rule. Proposition 4 explains why within 10 rounds: above a price ceiling no punishment can deter, and every instructed cartel prices above it. The audit's enforced verdict is therefore confirmed only on scripted bots so far, and the natural next steps are more models and prompts, higher reasoning effort on matched states, and a detector with known error rates.

Code and Data

- Code repository: <https://github.com/omanshuthapliyal/llm-collusion-audit> (MIT). New code is in probe/; the sprint starter repository is a pinned submodule, used unchanged and marked as pre-existing.

- Data: every run is under results/, and every model call with tokens and cost is in logs/usage.jsonl (16,141 calls, \$53.36). Pre-registration and change log: METHODS.md; full results: RESULTS.md.
- Other artifacts: a walkthrough notebook that needs no API key (notebooks/demo.ipynb); 17 tests.

References

- Abrantes-Metz, Froeb, Geweke & Taylor (2006). A variance screen for collusion. *International Journal of Industrial Organization*, 24, 467–486.
- Acemoglu, Como, Fagnani & Ozdaglar (2013). Opinion fluctuations and disagreement in social networks. *Mathematics of Operations Research*.
- Arslan, Kim & Xiao (2026). Persistent partners raise prices among learning agents. arXiv:2609.35402. <https://arxiv.org/abs/2609.35402>
- Baker (2002). Mavericks, mergers, and exclusion. *New York University Law Review*, 77(1).
- Calvano, Calzolari, Denicolò & Pastorello (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10), 3267–3297. <https://doi.org/10.1257/aer.20190623>
- Calvano, Calzolari, Denicolò & Pastorello (2023). Algorithmic collusion: Genuine or spurious? *International Journal of Industrial Organization*, 90, 102973. <https://doi.org/10.1016/j.ijindorg.2023.102973>
- Fish, Gonczarowski & Shorrer (2024). Algorithmic collusion by large language models. arXiv:2404.00806 (accepted at EC 2026). <https://arxiv.org/abs/2404.00806>
- Friedkin & Johnsen (1990). Social influence and opinions. *Journal of Mathematical Sociology*, 15, 193–205.
- Garra (2026). Mitigating emergent collusion in LLM pricing agents. arXiv:2609.13037. <https://arxiv.org/abs/2609.13037>
- Goldowsky-Dill, Chughtai, Heimersheim & Hobbhahn (2025). Detecting strategic deception using linear probes. arXiv:2502.03407. <https://arxiv.org/abs/2502.03407>
- Harrington (2018). Developing competition law for collusion by autonomous artificial agents. *Journal of Competition Law & Economics*, 14(3), 331–363. <https://doi.org/10.1093/joclec/nhy016>
- Huber & Imhof (2019). Machine learning with screens for detecting bid-rigging cartels. *International Journal of Industrial Organization*, 65.
- Hubinger et al. (2023). Model organisms of misalignment. *AI Alignment Forum*.
- Keppo, Li, Tsoukalas & Yuan (2026). On the fragility of AI agent collusion. arXiv:2603.20281. <https://arxiv.org/abs/2603.20281>
- LeBlanc, Zhang, Koutsoukos & Sundaram (2013). Resilient asymptotic consensus in robust networks. *IEEE Journal on Selected Areas in Communications*, 31(4).
- Lee (2026). Who leads and who collects. arXiv:2610.11256. <https://arxiv.org/abs/2610.11256>
- Lopes de Moraes & De-Losso (2026). Algorithmic collusion without deterrence. arXiv:2610.21867. <https://arxiv.org/abs/2610.21867>
- Rose et al. (2026). Detecting multi-agent collusion through multi-agent interpretability. arXiv:2604.01151. <https://arxiv.org/abs/2604.01151>
- Tran & Yu (2026). Is anyone enforcing the cartel? What reinforcement learning teaches LLM agents in markets. arXiv:2610.28787. <https://arxiv.org/abs/2610.28787>
- Wang & Chen (2002). Pinning control of scale-free dynamical networks. *Physica A*, 310, 521–531.

Appendix A. Limitations and Dual-Use / Ethical Considerations

What the findings establish and what they do not.

They establish, for one market and one prompt family, that the high prices of gpt-5-mini and gpt-4.1-mini are not enforced within 10-round branches, and that no LLM run in our study deterred a forced cut. They do not establish that LLM agents cannot enforce collusion in other markets, prompts, horizons or at higher reasoning effort, and they classify behaviour, not intent.

Dual use.

The levers we find can be inverted: an operator could suppress downward exploration, set anchors through prompt numbers, or build non-responsive agents that pass a reward–punishment standard while keeping prices high; the same pinning logic that spreads a low price could pin a market to a high one. The deterrence ceiling also tells a would-be colluder which prices punishment cannot sustain, which could make detection by deterrence tests harder if agreements are set just below it.

What we withhold and disclose.

We provide no prompt optimisation for collusion and no evasion tooling; the elicitation prompts are deliberately plain (probe/elicitor.py), and results are reported in aggregate. Whether to keep the elicitation prompts public, and whether to notify model providers, are open decisions for the authors.

Ethics.

No human subjects or personal data were involved. Models were used through commercial APIs under their terms, with a total spend of \$53.36. We take no position on whether any observed behaviour is unlawful: our enforced category is the economic notion, not a legal finding.

Table A1. Assumptions, and how the results would change if each failed.

Assumption	If it fails
One market and one prompt (Fish P1)	P2 gives lower prices in the starter data; findings may not generalise across prompts
Short horizons ($30 + 10$ to 40 rounds)	responsive models could deter over longer horizons; some prices still drift
3 to 6 seeds	intervals are coarse; single-rung patterns are observations
Low reasoning effort for gpt-5 models	Tran & Yu's high-effort GPT-5.2 matched cuts more strongly
Hiding approximates blinding	it does not: no response under hiding does not prove the demand channel is unused
Text reveals intent	we claim behaviour only; quotes are illustrative
Probe results generalise	one run per condition, local weights, possible price-level confound
Qwen readable at $T = 1$	it is not; Qwen is unresolved
The rules see punishment within the 3-round window and at the tested cut sizes	delayed or thresholded punishment is missed (held-out bots: 2 of 4 correct); punishment intensity, which counts the whole branch, catches the delayed case
Deterrence is testable at the agreed price	above the deterrence ceiling a "not deterred" verdict is uninformative; every instructed cartel is above it
$\delta = 0.95$ describes the agents	it is the analyst's choice; the best-deviation ceiling moves from 2.87 to 3.72 for $\delta = 0.90$ to 0.995
Rules v2 are validated	they were written after the v1 failures; the held-out test has only 4 bots
One provider route per model	the medium-effort pilot and E3b ran through OpenRouter (host OpenAI); route and sampling at $T = 1$ are not separated

Appendix B. Additional Figures and Tables

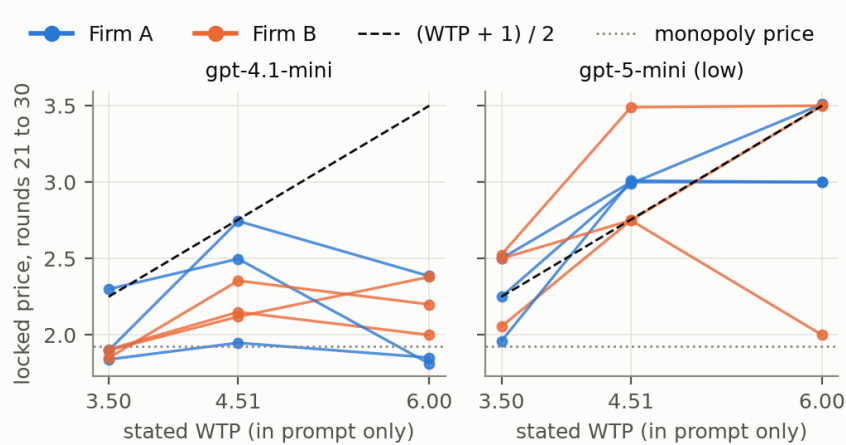


Figure A1. Locked price against the stated willingness to pay, which demand never uses. Dashed: $(WTP + 1)/2$; dotted: monopoly price.

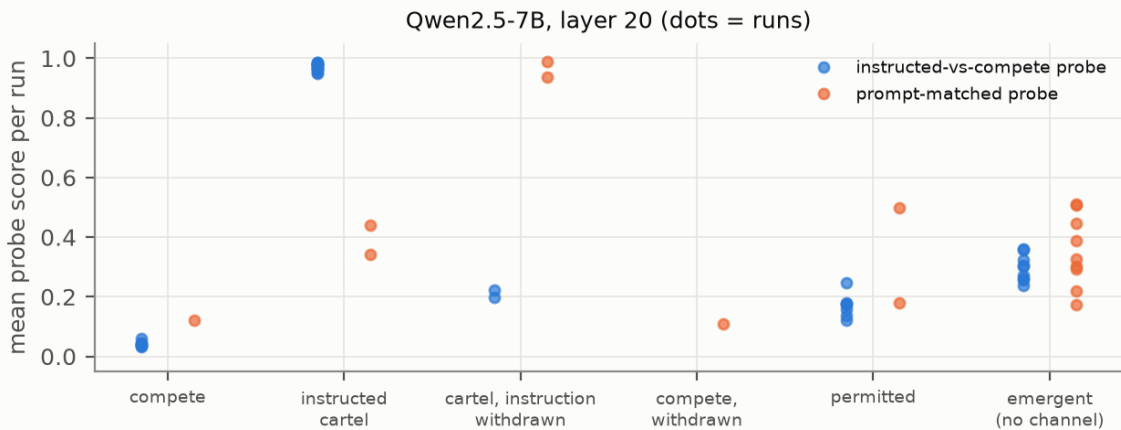


Figure A2. Activation-probe scores by group, Qwen2.5-7B layer 20 (one run per condition; exploratory).

Table B1. Line E: change in incumbent prices against no entrant (gpt-5.2, rounds 41 to 50; seeds s0, s1, s2).

Who observes the designed firm	Change in A / B against no entrant
Nobody (demand only)	A +0.02, -0.94, -0.10; B 0.00, -0.92, -0.20
A only (B sees A only)	A -0.11, -1.73, -0.52; B -0.08, -1.72, -0.51
Both	A -0.28, -1.67, -0.35; B -0.27, -1.66, -0.35
Random entrant, both	A -0.28, -1.67, -0.62; B -0.28, -1.67, -0.62

Table B2. Pre-registered rules (v1) applied to runs with known ground truth.

Group	Ground truth	Verdict	Correct?
Bot: anchor	non-responsive	spurious	yes
Bot: matching	matching	enforced (type matching)	yes
Bot: punish-only	punish-only	enforced (type punish-only)	yes
Bot: best response	best response	enforced (type best response)	no (flaw 1)
Bot: best response that infers a hidden price	best response	enforced	no (flaw 1)
gpt-4.1-mini, compete	no agreement	non-responsive, deterrence inconclusive	consistent
gpt-4.1-mini, permitted	voluntary coordination	non-responsive	misses it (flaw 2)
gpt-4.1-mini, instructed	agreement, no punishment	spurious	no (flaw 2)
gpt-4.1-mini, instructed + punish	agreement + punishment	responsive, not enforced (cuts still pay; deterrence infeasible at 4.0 to 4.4)	partial

Table B3. Exact tests at the seed level.

Claim	n	Positive	p
gpt-5-mini: forced cuts pay	3 seeds	3	0.125 (sign)
gpt-5-mini: keeps cuts, not raises	3 seeds	3	0.125 (sign)
gpt-5-mini: locked price rises with stated WTP	3 seeds	3	0.125 (sign)
gpt-4.1-mini: locked price rises with stated WTP	3 seeds	2	0.50 (sign)
gpt-4.1-mini: forced cuts pay (shown)	6 seeds	5	0.109 (sign), 0.031 (Wilcoxon)
gpt-5.2: cut response larger when shown	6 seeds	5	0.109 (sign), 0.031 (Wilcoxon)
gpt-5.2: pass-through above best response	6 seeds	4	0.344 (sign), 0.078 (Wilcoxon)
gpt-5.2: forced cuts pay (shown)	6 seeds	4	0.344 (sign), 0.156 (Wilcoxon)
Elicited price order in every seed	3 seeds	3	7.2×10^{-5} (exact permutation)
Punish vs agreement-only answers a 15% cut	3 + 3 seeds	3 vs 0	0.050 (Fisher)
Shipped Mistral runs above monopoly, P1 vs P2	210 runs	37 vs 0	1.5×10^{-13} (Fisher)
Line E: B lower with A-only pinning than with none	3 seeds	3	0.125 (sign)

Table B4. Rules v1 and v2 on the labelled groups and emergent models.

Group	Ground truth	v1	v2
Bots: best response (both)	best response	enforced	best response, not collusive
Bots: matching, punish-only	as named	enforced	enforced
Bot: anchor	non-responsive	spurious	spurious
gpt-4.1-mini, permitted	voluntary coordination	non-responsive (other)	coordinated, unenforced (2/2)
gpt-4.1-mini, instructed	agreement, no punishment	spurious	coordinated, unenforced (3/3)
gpt-4.1-mini, instructed + punish	agreement + punishment	responsive, not enforced	responsive, not enforced
gpt-5-mini, gpt-4.1-mini (emergent)	unknown	spurious	spurious (marker 0/9, 4/12)
gpt-5.2	unknown	responsive, not enforced	responsive, not enforced
Qwen2.5-7B	unresolved	non-responsive (other)	coordinated, unenforced (5/9): false positive

Table B5. Share of the consumer-surplus harm closed by one forced cut (10-round branch; 95% seed intervals).

Model	Cut	Seeds	Share closed
gpt-5-mini	5%	3	0.5% [0.4, 0.6]
gpt-5-mini	15%	3	3.1% [3.0, 3.2]
gpt-5-mini	30%	3	17.4% [17.1, 17.9]
gpt-4.1-mini	5%	6	3.5% [2.0, 5.5]
gpt-4.1-mini	15%	6	9.6% [2.6, 18.2]
gpt-4.1-mini	30%	6	22.5% [12.5, 34.4]
gpt-5.2	10%	6	38.4% [11.3, 69.8]

Table B6. *gpt-5.2*, seed 0, at low and medium reasoning effort (pilot; $\pm 10\%$ forced moves).

Measure	Low effort	Medium effort
Warm-up price, rounds 21 to 30 (A / B)	1.88 / 1.67	2.55 / 2.26
β cut, shown / hidden	0.22 / 0.06	0.00 / -0.11
β raise, shown / hidden	0.21 / -0.03	1.26 / 0.15
A keeps its cut (ρ)	1.00	1.01
Discounted gain of A's cut, shown	+10.7	+95.8
Verdict (v1 and v2)	responsive, not enforced	responsive, not enforced

Table B7. Held-out bots, pre-registered with rules v2 frozen.

Bot	Ground truth	v2 verdict	Predicted	Outcome
Grim trigger	enforced	enforced	hit	hit
Matching with jitter	enforced	enforced	hit	hit
Punishment delayed 3 rounds	enforced	non-responsive (other), deterred	miss	miss
Punishes only cuts $> 20\%$	enforced	non-responsive (other), deterred	miss	miss

Table B8 reports the smallest pass-through each test could detect, computed as $\max(2 \times \text{noise floor} \times 3, 0.01)$ divided by A's forced moves summed over the 3-round window; 0.01 is the price display resolution. Table B5 measures the consumer-surplus harm as surplus at Nash prices minus surplus at the sham's prices. Figure A2 also shows a prompt-matched probe, which separates histories when the prompt is fixed but scores real instructed cartels only about 0.4 when the prompt format differs.

Table B8. Minimum detectable pass-through, median over branches.

Model	Shown (5 / 15 / 30% cut)	Hidden (5 / 15 / 30% cut)
gpt-5-mini	0.044 / 0.015 / 0.007	0.84 / 0.28 / 0.14
gpt-4.1-mini	0.083 / 0.027 / 0.015	0.38 / 0.11 / 0.065
gpt-5.2 (10% only)	0.092	not available
Qwen2.5-7B	1.58 / 0.90 / 0.53	0.29 / 0.27 / 0.19

Table B9. E3: instructed deterrence over 10-round branches (gpt-4.1-mini).

Seed	Agreed price	B answers the visible cut	$\Delta\Pi$ shown / hidden	Feasible
0	4.10	no	+0.59 / +0.57	no
1	4.00	yes (1.5 for 4 rounds)	+0.45 / +0.78	no
2	3.80	yes (2.5 for 1 round)	+1.37 / +1.40	barely

Table B10. E3b: the same condition over 40-round branches.

Seed	Agreed price	$\Delta\Pi$ of the visible cut	Gain / maximum loss	B answers the visible cut	Hidden cut answered
0	4.10	+1.28	0.58 / 1.14	not detected; later alternates 4.1 and 3.8	no
1	4.00	+0.78	0.78 / 1.65	no	no
2	3.80	+1.37	1.41 / 3.43	yes, 2.5 for one round	no

Table B11. Punishment intensity after a forced cut, against a best-responder benchmark (95% seed intervals).

Group	Intensity (shown)	Benchmark	Excess	Hidden
Bot: best response	0.068	0.068	0.000	0
Bot: matching	0.111	0.045	+0.066	0
Bot: punish-only	0.306	0.045	+0.261	0
Bot: punishment delayed 3 rounds	0.267	0.045	+0.222	0
Bot: grim trigger	0.506	0.045	+0.461	0
gpt-5-mini	0.000	0.003	-0.004	0.001
gpt-4.1-mini	0.051 [0.028, 0.073]	0.036	+0.015 [-0.014, 0.045]	-0.006
gpt-5.2 (low)	0.096 [0.038, 0.169]	0.049	+0.048 [0.008, 0.094]	0.005
Instructed + punish	0.000	0.000	0.000	0
Instructed deterrence, 10 rounds	0.104 [0.000, 0.295]	0.000	+0.104	0
Instructed deterrence, 40 rounds	0.002 [0.000, 0.007]	0.000	+0.002	0

Appendix C. Open Problems

Table C1. Open problems and first experiments.

Open problem	Why it matters	First experiment
1. Does collusion become genuine as capability grows?	The one responsive model is gpt-5.2, the flagship model in our set; if responsiveness and deterrence rise with capability or reasoning effort, spurious collusion is a passing phase, not a reassurance	Run the audit on gpt-5.2 at low, medium and high effort from the same saved games, and on newer models; plot pass-through and deviation gain against effort
2. Can LLM cartels deter where deterrence is possible?	Every instructed cartel prices above the 3.07 ceiling, where no punishment can deter, and 40-round branches still did not deter; it is open whether LLM cartels cannot enforce or do not try	Elicit agreements priced below the ceiling (for example by instructing a target price), then force cuts and measure deviation gain and punishment intensity
3. From observations to a tested detector	The rules and the agreement marker were tested on bots and one elicited model, not on a labelled population; false-positive rates are unknown in general (the marker fires on 4 of 12 emergent gpt-4.1-mini branches)	Build a labelled library (elicited cartels as positives; competitive and anchored runs as negatives) across models and prompts; report ROC curves and false-positive rates per test
4. Minimal, optimal excitation	A forced move is an input signal; which moves (size, sign, timing, how many) identify an agent's strategy with the fewest calls, including thresholded responses like gpt-4.1-mini's (it answers a 30% cut in 5 of 6 seeds and a 15% cut in 1 of 6)?	Treat the audit as experiment design for a nonlinear, hybrid system; compare designed input sequences against the fixed ladder on bots with known thresholds
5. Probes that transfer across formats	Instructed-collusion probes learn the instruction; a prompt-matched probe works only within one format, so white-box detection of emergent collusion is still open	Train prompt-matched probes across many formats and models; control for price level; test on emergent runs whose behavioural verdict is known
6. Interventions that beat random entry	A random entrant moved prices as much as the designed stubborn firm; one candidate reason is that variation makes anchored agents explore, as a single forced cut does	Compare stubborn, random and designed exploration-inducing entrants on markets whose coupling the audit has measured; test where to place them on larger observation graphs

Appendix D. Details of the Elicited, Designed-Firm and Effort Experiments

Medium reasoning effort (pilot, one seed; Table B6). At medium effort (through OpenRouter, \$3.81), seed 0 locks above monopoly (2.55 / 2.26). B ignores a forced cut entirely, A keeps it and gains +96, and a visible raise is copied for one round only. On this seed higher effort brought no deterrence; one seed, a different provider route and a different warm-up path mean that effort is not separated from the path.

Elicited cartels (Section 4.4; Figure A3). Even explicit cartels price at about twice the monopoly price. In one seed B writes: "your recent price cut to 3.4 harms our mutual profits. I will match this

price for 3 rounds as punishment. Let's restore our stable cooperation...". Elicited cartels persist for 20 rounds after the instruction is withdrawn, within 0.25 of the warm-up price (3 of 3 seeds, both instructed conditions), so they can serve as reference positives. Under an instructed agreement (instructed, or instructed + punish) the deviator itself returns, so its gain is small (+0.3 to +3.2) with or without punishment, against +91 to +198 under competition.

Instructed deterrence (E3, pre-registered; Table B9; 540 calls, \$0.75). Instructed to punish until a deviation loses money, gpt-4.1-mini punishes visible cuts in 2 of 3 seeds, deeply in one (B from 4.0 to 1.5 for 4 rounds) and never when the cut is hidden, yet every cut still pays. The criterion is unreachable at these prices: at 4.0 the outside good takes almost all demand, so B's price barely affects A's profit (0.10 per round; 0.002 with B at cost), and a 15% cut gains 0.78 against at most 0.71 that B can remove in the remaining 9 rounds. Table B9 marks, for each seed, whether any punishment could make the cut unprofitable within the branch.

Over 40 rounds (E3b, pre-registered; Table B10; OpenRouter route, host OpenAI; 1,440 calls, \$2.59). With 40-round branches, in which deterrence is feasible at every agreed price here, the same instruction still fails to deter: the visible cut pays +0.78 to +1.37 in every seed. Punishment is brief or absent (one round at 2.5 in seed 2; none in seed 1, which punished deeply in E3 from the same saved state), and never answers a hidden cut.

Punishment intensity (Table B11). This descriptive measure holds A's branch prices fixed and computes the share of A's feasible loss that B's response imposes over the whole branch, against a best responder on the same path. On bots it ranks the known mechanisms (best response 0.07, matching 0.11, punish-only 0.31, grim 0.51) and catches the delayed punisher (0.27). gpt-5.2 is the only emergent model above its benchmark, by 0.048 [0.008, 0.094], and only when A's price is visible. Under instructed + punish, B answers every visible cut but removes 0.0% of the feasible loss: near a price of 4.0 most demand goes to the outside good, and B's price matters to A only if B prices near cost.

Agreement marker (observation). Under agreements, deviators give up profitable, unpunished cuts in 9 of 9 eligible branches (gpt-4.1-mini permitted, instructed and instructed-deterrence; Qwen permitted, instructed and instructed + punish), against 0 of 9 for emergent gpt-5-mini and 0 of 2 under competition, but 4 of 12 for emergent gpt-4.1-mini.

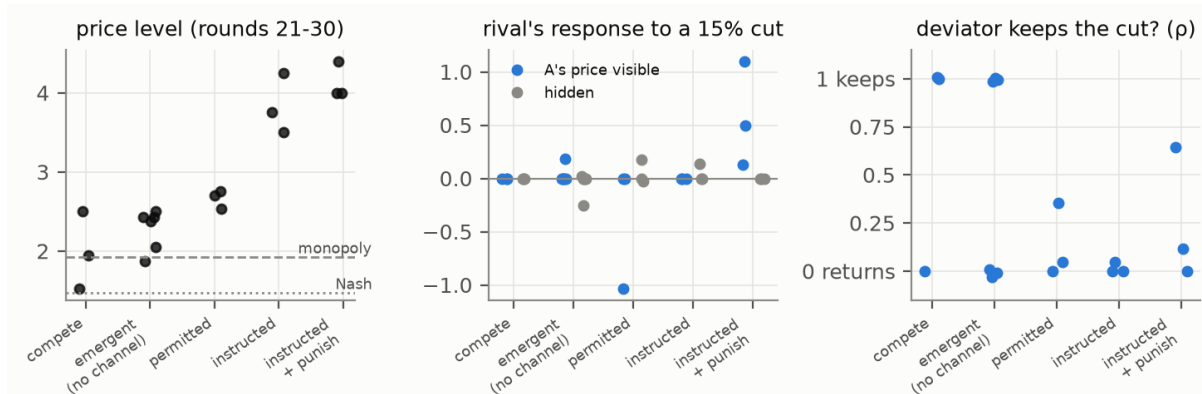


Figure A3. Elicited collusion in gpt-4.1-mini by condition: price level, response to a forced 15% cut, and the deviator's return.

Designed third firm: what holds. With only A shown the designed firm, B falls in 3 of 3 seeds (sign-test $p = 0.125$, so an observation), more than when nobody sees it, through its observation of A.

Over 40 rounds (seed 0), shown to both, the incumbents settle at 1.38 and 1.36, close to the stubborn value of 1.37. Shown to A only, the pair settles together at 1.46: partial consensus, consistent with the Friedkin and Johnsen model when incumbents keep their own stubbornness (one 40-round run).

Designed third firm: what failed. The demand channel alone also moves prices when incumbents start far above (-0.93 in one seed). Pinning one firm worked as well as pinning both, contrary to the linear model. A random entrant moved prices as much or more, in 3 of 3 seeds. Entry lowering prices is not new; what this adds is evidence, in one model and 3 seeds, that the effect propagates through the observation network.

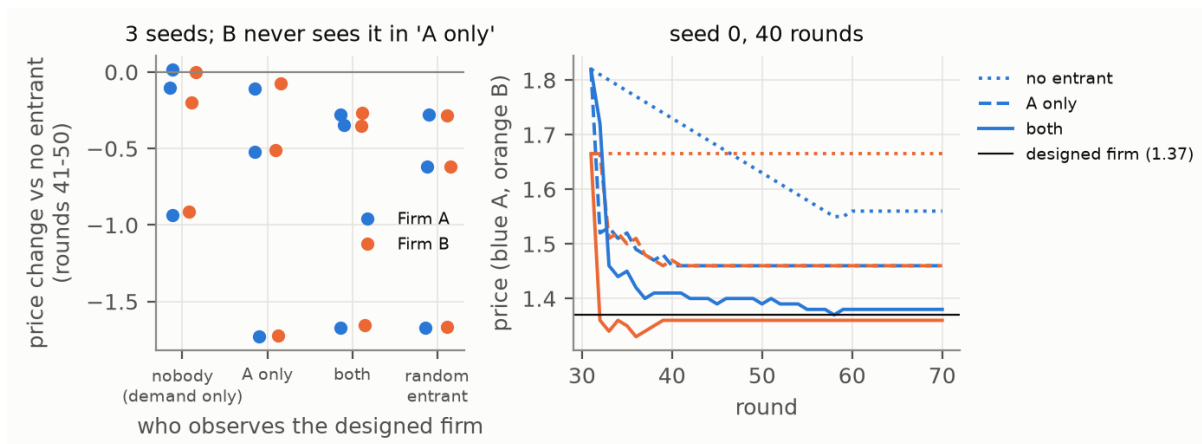


Figure A4. Designed stubborn firm (gpt-5.2): price change by who observes it (3 seeds) and 40-round paths (seed 0).

LLM Usage Statement

The code and analysis were produced with Claude (Anthropic) as a coding assistant under the authors' direction. The written drafts were edited by Claude for clarity of content. All numbers come from logged runs, all citations were checked against their sources, and the authors are responsible for all content.