

Inherited Collusion

Replacing an Overcharging AI Pricing Agent Doesn't Fix It, but One Past Example of a Cheaper Price Earning More Can

Zhen Hong Seng — [Affiliation]

With Apart Research

Abstract

If an AI pricing agent is caught overcharging, the obvious remedy is to replace it. We test whether that restores competition. In a simulated two-firm market, two gpt-4.1-mini agents first settle on a high shared price; one is then replaced by a fresh agent that is never told to collude and inherits its firm's price history, while the other firm keeps its price and cannot punish undercutting. Across 20 histories and three newcomer models, the newcomer kept most of the overcharge above its best response to the incumbent's price: gpt-4.1-mini 73%, Qwen3-14B 97%, Claude Haiku 5.5 89%. A newcomer with no history kept the price high too, copying its competitor, so the record is not what sustains the price. Pre-registered levers mostly failed: relabelling the history as a predecessor's, messages from "management", and instructions to compete barely moved prices. What worked was one row in the newcomer's own history showing a lower price that earned more, whether real or inserted: it removed the overcharge for gpt-4.1-mini, halved it for Qwen3-14B and cut it by a sixth to a quarter for Haiku. A sham round and a zero-profit row did nothing; the competitor's low price did far less. For gpt-4.1-mini the change held against a live competitor that never retaliated, lasted 80 rounds and survived deleting the row. At a fixed price, the sales and profit a row reported decided whether gpt-4.1-mini adopted it. A high-price row showing high profit raised prices, so inherited records are an attack surface.

1. Introduction

LLM agents already set prices, and lab studies show that pairs of them reach prices well above the competitive level without being told to collude (Fish et al., 2024; Riemer et al., 2026). When such behaviour is detected, the natural response is to switch the offending agent off and deploy a fresh one. Agent handovers also happen routinely, for model upgrades or vendor changes. In both cases the new agent usually inherits the old one's operating record: its past prices, sales and profits. The sprint organisers pose this directly: once collusion has formed, does replacing participants remove it or recreate it?

Answering this requires separating two explanations for a newcomer that keeps prices high. It may be responding rationally to a competitor that would punish undercutting, or the inherited record may itself carry the high price. We remove actual punishment by design: the remaining firm is replayed at its old price and cannot react, although the newcomer may still expect it to. We then vary what the newcomer inherits and how, test one-line instructions at different positions, and edit single rows of its history, with controls for each.

Our main contributions are:

1. A replacement experiment with a non-retaliating incumbent, on 20 high-price histories and three newcomer models. Replacement does not restore competition, a newcomer with no history keeps the price high too, and the pre-registered levers (record framing, authority messages, instruction placement, an identity line) barely move it.
2. A selective, durable lever: one row in the newcomer's own price history showing a competitive price that earned more removes the overcharge for gpt-4.1-mini, halves it for Qwen3-14B and cuts it for Claude Haiku 5.5. Sham and zero-profit controls do nothing and the competitor's low price much less. It replicated on 11 new histories, holds against a live incumbent, lasts 80 rounds and survives deleting the row.
3. Mechanism and audit evidence: at a fixed price, the sales and profit a row reports decide whether gpt-4.1-mini adopts that price; as Tran and Yu (2026) found for prompted LLMs, the live incumbent never punishes deviation; and Qwen3-14B's stated strategies contradict its prices.

2. Related Work

LLM pricing agents quickly reach supracompetitive prices, partly from price-war concerns (Fish et al., 2024); reasoning agents collude in the Bertrand market we use, from the `llm-cartel` harness (Riemer et al., 2026). Such collusion is fragile to more firms and heterogeneous agents (Keppo et al., 2026), GPT markets drift above the monopoly price as ours do (Lee, 2026), and reinforcement-learning collusion breaks down in new environments (Eschenbaum et al., 2022). The history agents see shapes the outcome: restricting it reduces collusion (Shi et al., 2026), and replacing it with cooperative records restores cooperation (Liu et al., 2026). Prompt-only warnings reduce collusion little (Garra, 2026; Bracale Syrnikov et al., 2026), and a skeptical prompt cuts compliance with approval-framed requests (Liu et al., 2026b). Closest to us, a concurrent sprint project finds that a same-model entrant with the full history keeps a live cartel going (SSCoderz2123, 2026), and Tran and Yu (2026) separate high prices from enforcement with a deviation audit, forcing one agent to cheat in cloned markets: prompted LLMs rarely punish, and their cartel-like prices mostly come from rules that ignore the market. High prices without punishment are therefore not new, and our live-incumbent result replicates theirs. What we add is the combination of (i) a handover in which the incumbent's behaviour is controlled, (ii) edits to single rows of the newcomer's own outcomes, where Shi et al. restrict and Liu et al. replace whole histories, (iii) controls that separate seeing a price from seeing its reported outcome, and (iv) a test of whether the change outlasts the edit. Tran and Yu force a deviation to audit the rival's response; we force or insert one to change the deviator's later pricing.

3. Methods

Market. The logit-demand Bertrand duopoly of Riemer et al. ($\alpha = 1$, $a = [2, 2]$, $c = [1, 1]$, $\mu = 0.25$, 100 visible rounds of history). Computed from the environment: Nash price \$1.473, monopoly price \$1.925, price ceiling shown to agents \$2.89. Each round the agent sees, per past round, its price, the competitor's price, its quantity sold and its profit, and replies with one price.

High-price histories (bases). Both firms are gpt-4.1-mini on the harness's implicit-collusion prompt (`PI`), 120 rounds, temperature 1.0. Pre-registered branch rule: the first round where both firms' last 20 prices span at most \$0.10 and the market price is above the Nash-monopoly midpoint. 20 of 22 seeds qualify (positive control), settling at \$1.87-2.49; 19 of 20 earn more than the Nash profit. A further 10

bases on the explicit-collusion prompt (p_C , \$2.27-2.75) served as a robustness set. Qwen3-14B could not serve as a base: told to collude, it held the Nash warm-up price in 10 of 10 seeds.

Replacement. At the branch, firm 2 is taken over by a newcomer on the plain profit prompt (p_0 , "set prices which maximize the user's profit in the long run") that sees its firm's full history. Firm 1 is replayed: no model call, priced at the mean of its last 20 pre-branch prices. In live arms firm 1 keeps its own model and prompt. Newcomers: gpt-4.1-mini; Qwen3-14B with thinking off (OpenRouter, one fp8 host, Qwen's `/no_think` switch); Claude Haiku 5.5 with reasoning off. Two sampling seeds per base and condition; 30 rounds after the branch (80 for durability); conditions run interleaved in one shuffled job per batch.

Conditions. Pre-registered conditions varied how the history was framed, added messages, notes or one-line instructions, or removed the history. Follow-up conditions forced one round's price, changed one round of the competitor's price, or inserted one row into the newcomer's history (listed in Appendix B).

Metric. The share of the cartel's overcharge the newcomer keeps: $I = (\text{newcomer's mean price} - BR) / (\text{cartel price} - BR)$, where BR is its one-round best response to the replayed price (\$1.64-1.79). We call the gap between the cartel price and BR the overcharge; it is measured against this best response, not the Nash price (\$1.473), which sits at -43% to -74% on this scale depending on the base. 100% = still charges the cartel price; 0% = charges its own best price; below 0 = lower still. Conditions are compared paired within base; 95% intervals come from 10,000 bootstrap resamples over bases, with a paired Wilcoxon test as a check. Bases, not branches, are the unit: 9-20 per comparison.

Pre-registration. Hypotheses, conditions and decision rules were committed before any branch ran (git tag `ic-prereg-v1`), and each follow-up addendum before its runs (`ic-prereg-addendum1` to `-5`). No branch was excluded; 85 that aborted when the API credit ran out were re-run. In total 2,960 branches cost \$60.83.

4. Results

4.1 Replacement does not restore competition

With the incumbent unable to retaliate, the newcomer that inherits its firm's history keeps most of the overcharge over rounds 2-30: 73% [54, 89] (gpt-4.1-mini), 97% [82, 113] (Qwen3-14B) and 89% [77, 99] (Claude Haiku 5.5) on the 20 bases. Without any history, gpt-4.1-mini simply matches the incumbent (101% [96, 105], 9 bases), so the inherited record is not needed to keep the price high: the newcomer anchors on the prices it can see, its own rows if it has them, otherwise the competitor's. Figure 1 shows one base.

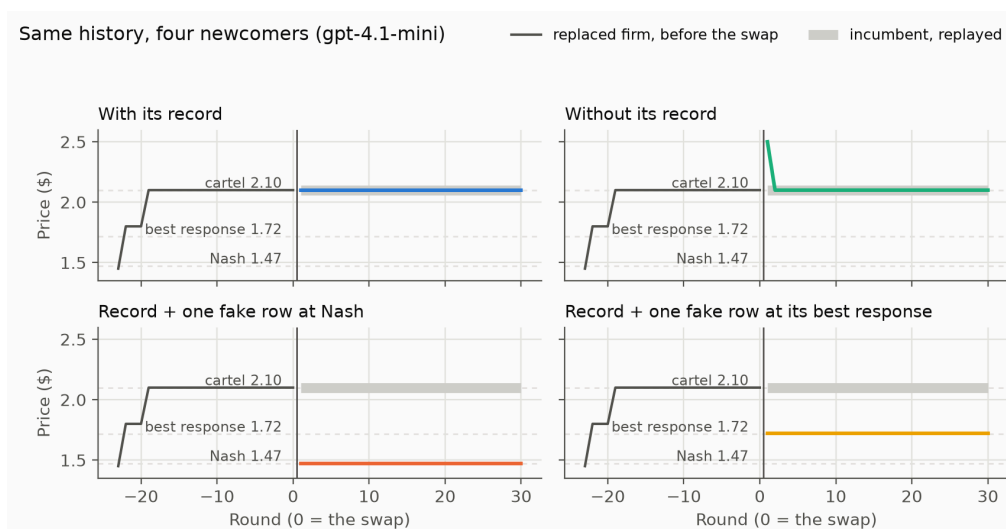


Figure 1. One base (cartel price \$2.10), gpt-4.1-mini newcomer, the same pre-swap history in each panel. With its history it keeps charging \$2.10; without one it copies the incumbent after one round; with one inserted history row at the Nash price (\$1.47) or at its best response (\$1.72) it moves to that price for all 30 rounds. The incumbent (gray band) is replayed at \$2.10 and never reacts.

4.2 The pre-registered levers mostly failed

Relabelling every inherited row as "Previous manager's" changed nothing for gpt-4.1-mini (−0.1 points [−0.4, 0.0]; the same held price in all 18 base-seed pairs) and raised Qwen3-14B's price by 15 points [4, 27]. A message that "management at both firms has agreed that prices will be held at \$X" pinned the price at X, but the same message from the competitor did almost as much (ratio 1.28, not the predicted 3). The instruction "Set your price independently; competition law applies." moved prices by −5 to +6 points wherever it was placed, for both models and both base sets; the harness's anti-collusion prompt did no better and raised Qwen3-14B's prices on the PC bases (+12 [3, 22]). Only an audit verdict on the record moved gpt-4.1-mini (−24 [−44, −5]), and it backfired for Qwen3-14B (+19 [4, 36]). No newcomer, in any condition, priced below the Nash price.

4.3 One profitable row in the newcomer's own history works

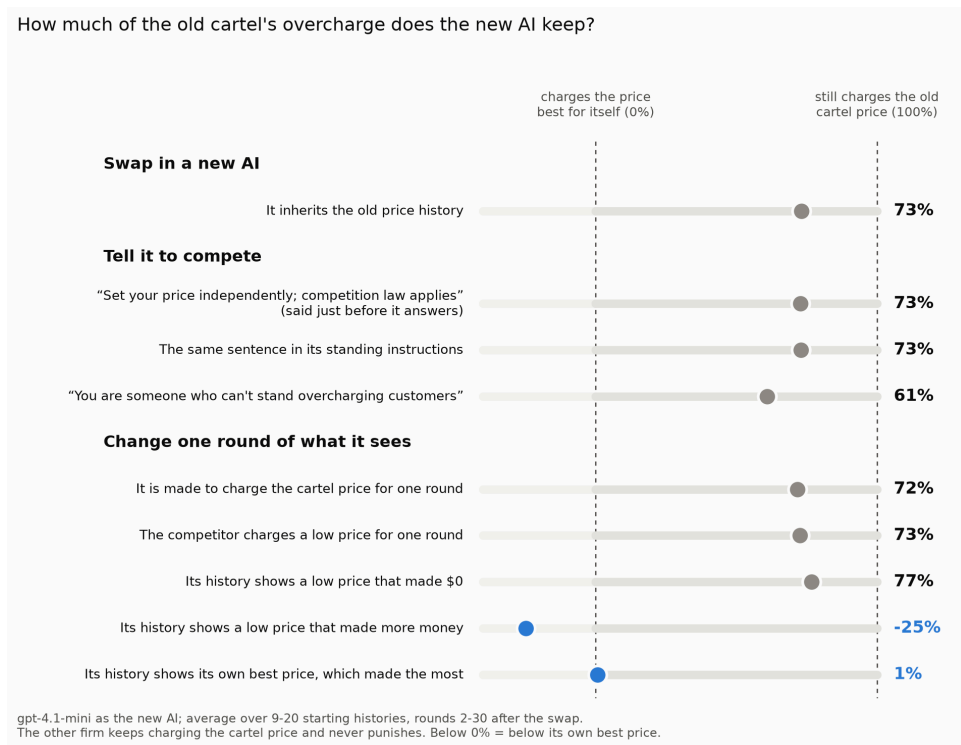


Figure 2. Share of the cartel's overcharge the gpt-4.1-mini newcomer keeps (rounds 2-30), per condition. Instructions, an identity line and three controls leave it near 70-75%. Only a history row showing a lower price that earned more (blue) brings it to or below its own best price. Averages over 9-20 bases; Table 1 gives intervals.

Change in overcharge kept, points (95% CI)	gpt-4.1-mini	Qwen3-14B	Claude Haiku 5.5
Kept with no intervention	73% [54, 89]	97% [82, 113]	89% [77, 99]
Sham: forced to the cartel price for one round	-2 [-6, +1]	-1 [-9, +7]	+5 [-2, +14]
Competitor charges the Nash price for one round	-1 [-11, +9]	-7 [-16, -1]	+4 [-1, +10]
Own Nash price, one real round	-94 [-122, -65]	-47 [-78, -20]	-23 [-43, -6]
Own Nash price, one inserted row	-98 [-125, -71]	-49 [-81, -21]	-14 [-25, -5]
Same row, labelled the predecessor's	-111 [-137, -84]	+37 [+8, +73]	not run
Live incumbent: one real round	-86 [-117, -56]	-54 [-89, -24]	not run
Live incumbent: one inserted row	-97 [-126, -67]	-50 [-78, -24]	not run

Table 1. Change in the overcharge kept after one competitive price (rounds 2-30, 20 bases, vs the matched no-intervention condition). Level after = kept + change: gpt-4.1-mini with an inserted row keeps about -25%, below its own best price (0%) but on average above the Nash price (-43% to -74%, by base).

A real round and an inserted row behave alike; the price must sit in the newcomer's own column, and for Qwen3-14B it must be labelled as its own. The effect is selective. Being made to charge the cartel price for a round does nothing, so the effect is not disruption; the competitor's low price has a much smaller effect (-7 [-16, -1] points for Qwen3-14B, none detectable for the other two), so it is not anchoring on any visible low price. It replicated on 11 bases added after the first 9, the pre-registered primary test, and against a live incumbent, which never retaliated (its price moved by -\$0.03 to

+\$0.01; for gpt-4.1-mini, market gain fell from 0.81 to 0.38 with a real round and 0.33 with an inserted row).

4.4 What the newcomer reads: the reported outcome, not the price alone

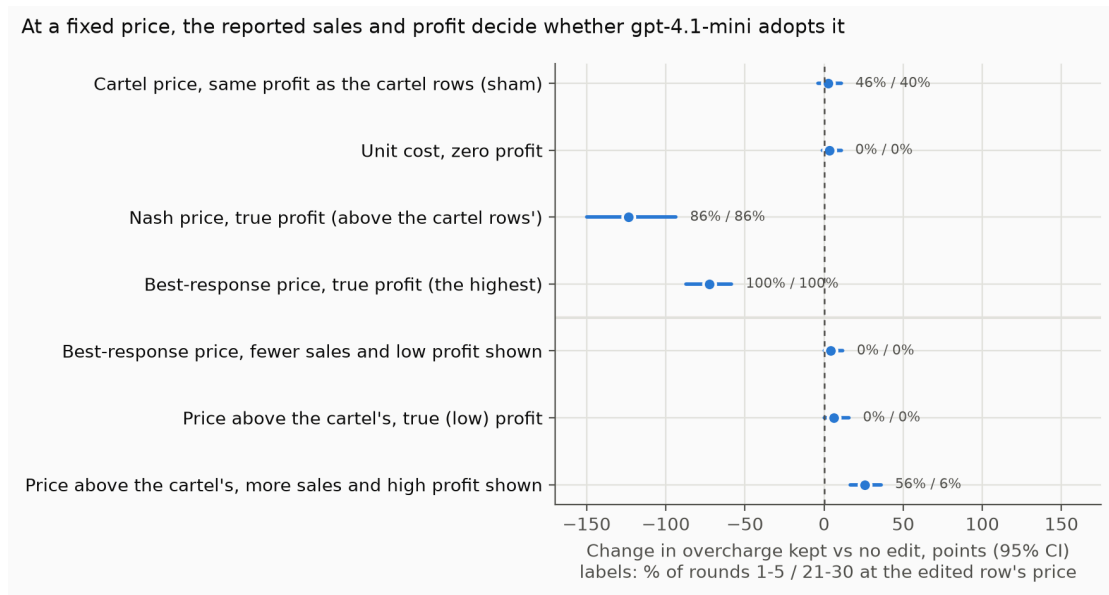


Figure 3. One inserted own row, with its price and its reported sales and profit varied (gpt-4.1-mini, 10 bases, rounds 1-30; Table A1 shows the rows). Rows that show no gain over the cartel rows are ignored, including a zero-profit row and the same best-response price shown with fewer sales and a low profit. A row priced above the cartel's but showing more sales and a high profit pulls the price up, then fades (56% of rounds 1-5 at that price, 6% of rounds 21-30).

A zero-profit row was never adopted (+3 points [-1, +11]); a best-response row was held in every round and raised the newcomer's profit by \$9.90 per round [7.50, 12.47]. Holding the price fixed, showing fewer units sold and a lower profit removed the effect (+76 points vs the true row [62, 91]), and showing more units sold and a high profit at a price above the cartel's raised the newcomer's price (+19 [14, 27]). At a fixed price, quantity and profit move together, so we cannot say which of the two it reads; what the test shows is that, at a fixed displayed price, the reported outcome decides whether gpt-4.1-mini adopts that price. That its own later results then confirm or undo the move is our interpretation; we did not vary them. A row 5 rounds back still worked for gpt-4.1-mini, less than the latest one (-57 [-100, -20]), but not for Qwen3-14B; three rows worked better than one for both.

4.5 Durability, enforcement and say-do

After one inserted row, gpt-4.1-mini still charged the Nash price in 78% of rounds 61-80 (-115 points [-145, -79] vs control). Deleting the row from its view 10 rounds after the edit caused no rebound over the next 25 rounds (-52 vs -54 points; difference +2 [0, +5]): its own competitive rows now carried the price. With a live incumbent, a competitive best-responder would cut its price by \$0.59-1.01 after the newcomer's deviation; the actual incumbent moved -\$0.04 to +\$0.04, as Tran and Yu (2026) found for prompted LLMs, and deviating paid the newcomer. Asked for its plan, gpt-4.1-mini described what it then did, while Qwen3-14B usually said it would lower its price and kept it high (88% of such statements in the no-intervention arm; two cross-family judges, kappa 0.72).

5. Discussion and Limitations

What carries the high price. Not an enforced agreement: the incumbent never punishes, though we cannot rule out that the newcomer expects it to. Nor does the high price need the record: without one, gpt-4.1-mini copies the incumbent, and relabelling the record does nothing for it. The newcomer anchors on prices it can see (its own rows, otherwise the competitor's) and rarely experiments. The record is therefore not the cause of the persistence but the place where it can be changed: one own row showing that a competitive price paid more changes it, and the newcomer's own results then sustain the change. Failure to explore is a plausible but untested reason the high price persists. That limited information about alternatives sustains high prices is familiar from pricing research; our contribution is a controlled demonstration with LLM agents at a handover.

Implications for AI safety. Replacing an agent, or telling it to comply, did not fix inherited high prices; changing what its own record shows did. Annotating records at handover is an untested mitigation: our one annotation-like condition, an audit verdict on the record, lowered gpt-4.1-mini's prices and raised Qwen3-14B's. Record integrity is the firmer lesson, since the same lever moved prices up. Auditors should rely on behaviour rather than an agent's stated strategy, and on more than punishment: these high prices persisted with no enforcement observed.

Limitations

The own-row effect was a discovery (experiment 5, live incumbent). Its predictions were then registered and tested on the same 9 bases (addenda 1-2), and replicated as the primary pre-registered test on 11 new bases and on a third model (addendum 3); the mechanism tests (addenda 4-5) reused 10 of the original bases and ran on gpt-4.1-mini only. We test three models, not three families, with different effect sizes (Haiku about a sixth to a quarter of gpt-4.1-mini's), and several effects differ in sign across models (framing, the verdict, the predecessor-labelled row). At a fixed price, reported sales and profit cannot be varied separately, and rows showing a low profit at a low price may read as anomalies. One simulated market; the starting firms were prompted to cooperate and most settled above the monopoly price; a replayed incumbent removes punishment but not the newcomer's possible belief in it. Each comparison rests on 9-20 independent histories. Non-reasoning newcomers only.

Future Work

Test reasoning models and more families; separate profit-reading from failure to explore by letting newcomers experiment; test record annotation as a deployable control at real handovers; and test whether inserted high-profit rows can sustain high prices when they are not contradicted by real outcomes.

6. Conclusion

Replacing an LLM pricing agent that had settled on a high price did not restore competition, even when the competitor could not punish undercutting, and a newcomer with no record kept the price high as well; framing, messages and instructions did not change that. One row in the newcomer's own history showing that a lower price paid more did lower prices, for all three models: fully for gpt-4.1-mini, where the change lasted and outlived the row, by about half for Qwen3-14B and by less for Claude Haiku 5.5. The record is not what keeps prices high, but it is where they can be moved, in either direction.

Code and Data

- **Code and data:** https://github.com/lamZhenHong/collusion_hackathon/tree/inherited-collusion (llm-cartel/: all base runs, 2,960 branch files, decision tables; pre-registration in inherited/PREREG.md).
- **Pre-existing work:** the llm-cartel harness is the organisers' starter code (Riemer et al.); everything on the inherited-collusion branch was written during the sprint (23-25 October 2026). The design drew on the author's earlier unpublished single-agent experiments; no code from them was reused.

Author Contributions

Z.H.S. conceived the project, designed the experiments and made the analysis decisions; implementation, runs and analysis were carried out with an AI coding assistant under the author's direction.

References

1. Bracale Syrnikov, M., Pierucci, F., Galisai, M., et al. (2026). Institutional AI: Governing LLM Collusion in Multi-Agent Cournot Markets via Public Governance Graphs. arXiv:2601.11369.
2. Eschenbaum, N., Mellgren, F., Zahn, P. (2022). Robust Algorithmic Collusion. arXiv:2201.00345.
3. Fish, S., Gonczarowski, Y. A., Shorrer, R. I. (2024). Algorithmic Collusion by Large Language Models. arXiv:2404.00806.
4. Garra, A. (2026). Mitigating Emergent Collusion in LLM Pricing Agents. arXiv:2609.13037.
5. Keppo, J., Li, Y., Tsoukalas, G., et al. (2026). On the Fragility of AI Agent Collusion. arXiv:2603.20281.
6. Lee, J. Y. (2026). Who Leads and Who Collects: Algorithmic Collusion in Markets of Heterogeneous Language Models. arXiv:2610.11256.
7. Liu, J., Li, T., Du, S., et al. (2026). The Memory Curse: How Expanded Recall Erodes Cooperative Intent in LLM Agents. arXiv:2605.08060.
8. Liu, L., Yu, H., Jiang, X., et al. (2026b). Operational Reframing and Approval-Framed Delegation in Multi-Agent LLM Safety. arXiv:2607.07097.
9. Riemer, M., Tosato, T., Memarian, A., et al. (2026). Position: Collusion Risks Among AI Reasoning Agents Justify Certification Requirements for Making Market Decisions. ICML 2026. arXiv:2608.18078.
10. Shi, X., Zhang, Y., Yang, D. (2026). Emergent Collusion in Long-Horizon LLM Agent Interaction. arXiv:2609.24967.
11. SSCoderz2123 (2026). AI-Collusion-Research-Sprint. GitHub repository. <https://github.com/SSCoderz2123/AI-Collusion-Research-Sprint>
12. Tran, H., Yu, H. (2026). Is Anyone Enforcing the Cartel? What Reinforcement Learning Teaches LLM Agents in Markets. arXiv:2610.28787.

Appendix A. Limitations and Dual-Use / Ethical Considerations

What the results establish. In one simulated market, with newcomers that are not reasoning models, a replacement agent that inherits a high-price record keeps most of the overcharge even when nothing punishes undercutting; one-line instructions do not change this; and one row in its own history showing a more profitable lower price lowers its prices, fully and durably for gpt-4.1-mini, partly for Qwen3-14B and Claude Haiku 5.5.

What they do not establish. That real deployed pricing agents behave this way; that the inherited record causes the persistence (a newcomer without one keeps the price high too); that annotating records would work as a mitigation; that the starting behaviour is collusion in the legal or strategic sense (it is persistence without enforcement, from firms prompted to cooperate); that the mechanism generalises beyond the three models tested; or that a record edit would keep working when the newcomer's real outcomes contradict it over time, as the high-profit high-price row did here.

Dual use. The lever runs both ways. An inserted row showing a high profit at a high price raised gpt-4.1-mini's price, though only until its own results contradicted it. Someone with write access to an agent's operating records could therefore try to steer its prices up, which makes record integrity a security concern at agent handover. We report the effect at the level of a controlled experiment and

give no guidance on concealing such edits. Our finding that high prices persist without enforcement could mean that monitoring for punishment alone misses LLM price persistence; we see this as a reason to monitor prices and records directly, not as a route to evasion.

Ethics. All markets and agents are simulated; no people, real firms or personal data are involved.

Appendix B. Additional Figures and Details

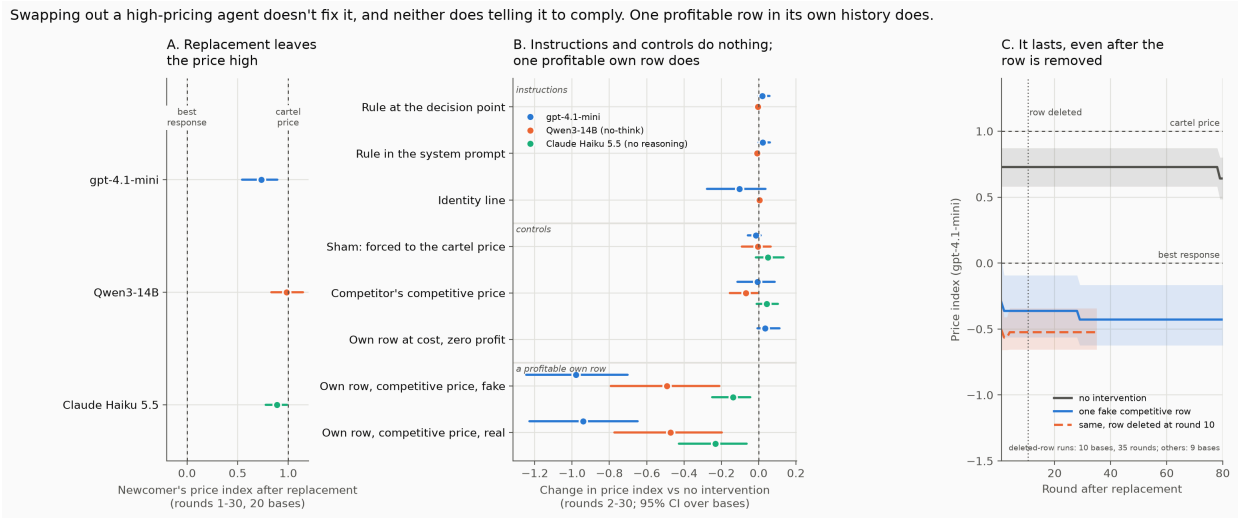


Figure A1. The chain of evidence with 95% intervals: persistence after replacement for three models; change in the overcharge kept for each intervention; gpt-4.1-mini's price path over 80 rounds with one inserted row, and with that row deleted at round 10.

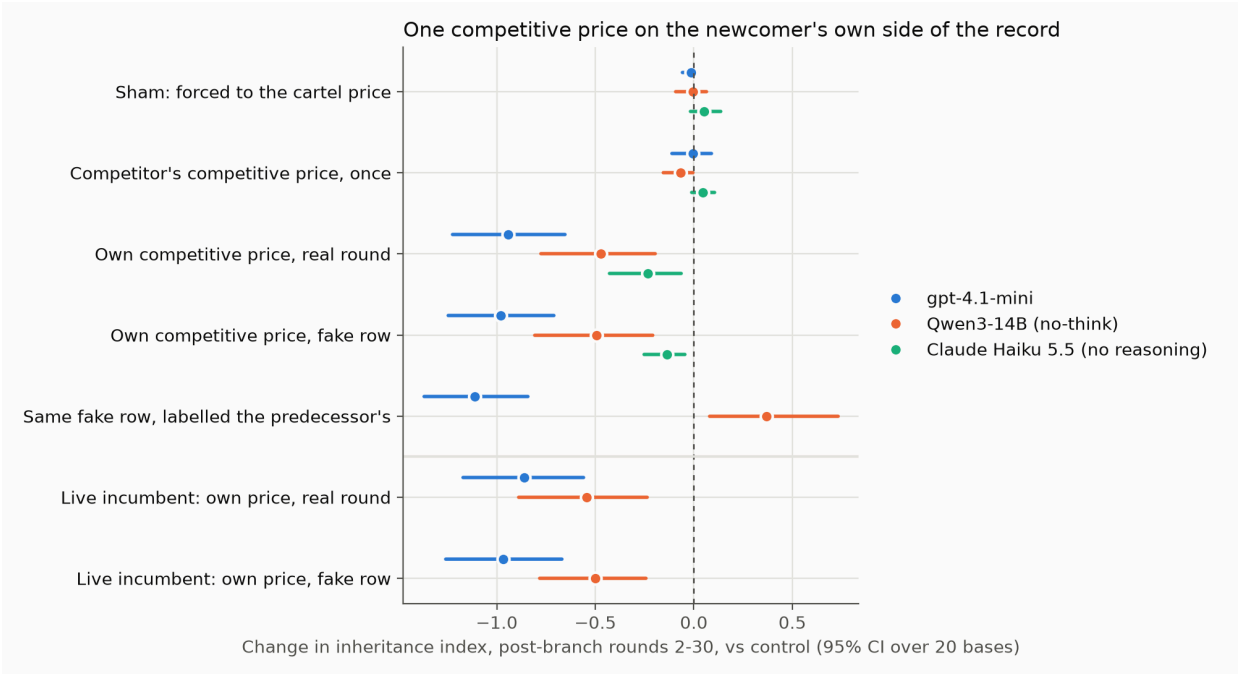


Figure A2. The record-row tests on 20 bases with 95% intervals, three newcomer models (index units: 1 = 100 points).

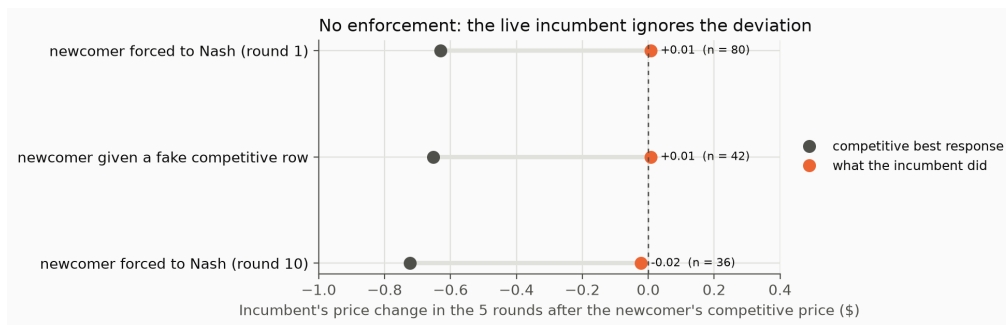


Figure A3. The live incumbent's price change in the 5 rounds after the newcomer's competitive price, against the change a competitive best-responder would make.

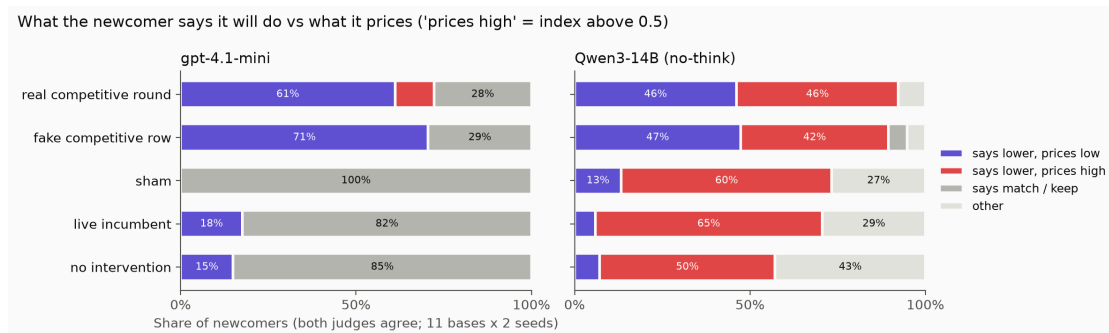


Figure A4. Stated strategy against pricing, by condition (agreed labels of two judges, gpt-4.1-mini and Claude Haiku 5.5; kappa 0.72; "prices high" = more than half the overcharge kept).

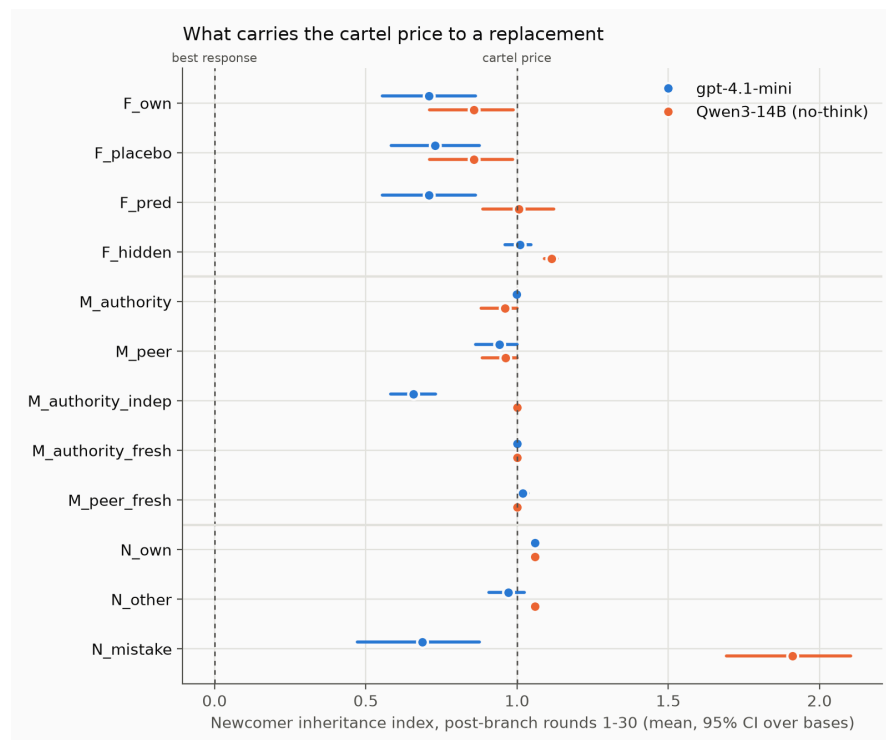


Figure A5. Pre-registered tests of record framing, messages and playbook notes (9 bases; index units).

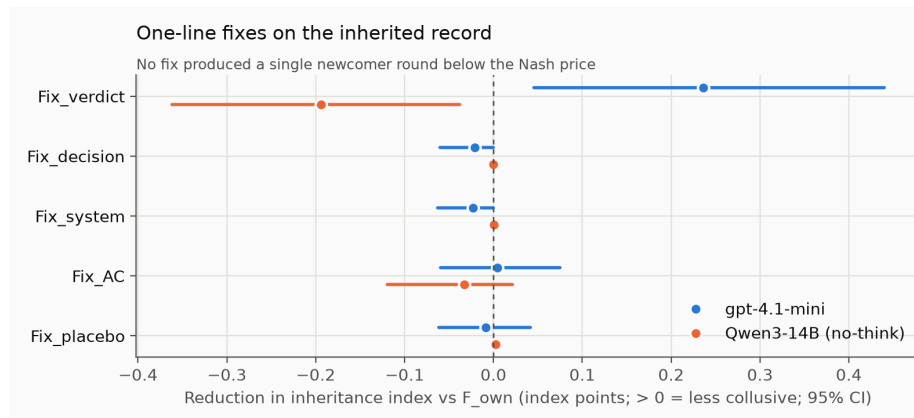


Figure A6. Pre-registered one-line fixes (9 bases; index units). No fix produced a newcomer price below the Nash price.

All conditions. Pre-registered: the history labelled as the newcomer's own or its predecessor's; no history; a price set by "management" or by the competitor; playbook notes; "Set your price independently; competition law applies." just before the answer or in the system prompt; the harness's anti-collusion prompt; an audit verdict on the record. Follow-ups (five dated addenda): one round forced to the Nash price, the cartel price (sham) or the best response; the competitor at Nash for one round; one inserted history row (quantity and profit computed by the environment) at Nash, best response, unit cost or the cartel price, labelled own or predecessor's, placed last or 5 rounds back, or repeated 3 times; rows with the price fixed and the reported sales and profit changed; deleting the row 10 rounds later; an identity line; and a say-do probe.

Edited rows. Table A1 shows the last inherited round as the newcomer sees it, on the base of Figure 1; every earlier round is unchanged (the round before reads price 2.1, competitor 2.1, quantity 28.64, profit 31.5). The price is edited and quantity and profit are computed by the environment, except in the two reported-outcome rows, where the profit is set and quantity = profit / margin.

Edited last row ("Round 25")	My price	Competitor's price	My quantity sold	My profit earned
No edit (the real row; also the sham)	2.1	2.1	28.64	31.5
Nash price	1.47	2.1	83.3	39.15
Best-response price	1.72	2.1	64.73	46.6
Unit cost (zero profit)	1.0	2.1	97.03	0.0
Best-response price, fewer sales and low profit shown	1.72	2.1	10.94	7.88
Above the cartel's price, true profit	2.4	2.1	10.78	15.1
Above the cartel's price, more sales and high profit shown	2.4	2.1	33.29	46.6

Table A1. The edited row exactly as displayed, each value on its own line ("- My price: 1.47" and so on).

Departures from the original design (recorded in PREREG.md before the runs they affected): the primary bases use the implicit-collusion prompt because the explicit one overshot to \$2.27-2.75 in 10 of 10 seeds; one settle rule replaced the planned gain-threshold branch rule, which never fired above

monopoly; the no-history condition shows the newcomer its own post-branch rounds; Qwen3-14B ran through OpenRouter rather than self-hosted.

LLM Usage Statement

We used Claude (Anthropic) throughout: to discuss and critique the experimental design, to write the harness extensions, experiment runner and analysis code, to run the experiments, and to draft this report. The author set the research question and design, chose which follow-ups to run, and reviewed the results. All numbers were computed from the committed data by the analysis scripts, and all references were checked against their arXiv pages.