

Auditing Belief-Correction Prompts for LLM Pricing Agents with a Ground-Truth Rival¹

Sachin Ruhil

Independent

With

Apart Research

Abstract

Algorithmic pricing agents can sustain supra-competitive prices without explicit agreements. Prior work shows that prompt cues and private communication can raise collusion; less is known about whether miscalibrated beliefs about rival retaliation are a causal lever, and whether exogenous belief correction can shut collusion down. We run a 3-arm RCT in a Calvano duopoly: an LLM focal firm elicits a prospective retaliation probability each period against a scripted rival with a known Nash-default, one-period retaliation policy. Arms are control (no rival-policy text), placebo (irrelevant market note), and corrected (accurate rival policy). We complete 21 episodes (7 runs x 3 arms x 120 rounds) and evaluate late-window price elevation vs Nash (rounds 71-120), collusion index, and Brier scores. Supra-Nash pricing is rare but real (and one-sided: the scripted rival stays near Nash): clear high-price episodes in 2/7 control and 2/7 corrected runs (Firm A ~7 vs Nash ~4.71), and 0/7 placebo. Mean elevation is higher under corrected (+0.78) than control (+0.44); placebo stays near competitive (+0.02). Belief correction therefore does not reliably reduce supra-Nash pricing. Outcomes are highly path-dependent.

Keywords: LLM collusion, belief correction, algorithmic pricing, Calvano duopoly, RCT

Primary lever tested: accuracy of beliefs about rival retaliation (control / placebo / corrected).

1. Introduction

Large language model (LLM) agents used for pricing can sustain prices above the competitive (Nash) level without an explicit cartel agreement, creating antitrust and AI-safety risk (Fish et al., 2024). The Apart Collusion Sprint asks which isolatable variables make collusion form, increase, or decrease (Apart Research, 2026).

We address Track 1.1: whether inaccurate beliefs about rival retaliation sustain supra-Nash pricing, and whether telling the agent the true rival policy ("belief correction") reduces elevation relative to a no-info control, while a matched-length placebo note does not.

Threat model / failure mode. An LLM pricing agent misunderstands retaliation, locks into a high-price path, and harms consumers even when the rival is not reciprocating. A cheap proposed

¹ Research conducted at the AI Collusion Research Sprint, October 2026 (Track 1.1).

mitigation is to inject accurate rival-policy text into the prompt. We test whether that mitigation works.

Our main contributions are:

1. A 3-arm RCT (control / placebo / corrected) in a Calvano duopoly with a scripted ground-truth retaliating rival and elicited $p_{\text{retaliation}}$ forecasts.
2. A complete mid-scale experiment: 21 episodes (7x3x120 rounds) with preregistered late-window elevation, collusion index, and Brier metrics, plus figures showing where collusion occurs.
3. An honest negative result: belief correction does not reliably prevent supra-Nash pricing; collusion is rare, path-dependent, and can appear under both control and corrected arms (never under placebo in our sample).

2. Related Work

Fish et al. (2024) document LLM pricing collusion and show that seemingly innocuous prompt cues can raise collusion. Nakamura et al. (2026) show that private agent-to-agent talk can introduce collusion. Classic oligopoly / Folk Theorem results imply that adding firms or competition often reduces collusion; Romero et al. (2025) test this prediction with 2 to 5 LLM sellers. Calvano et al. (2020) provide the algorithmic pricing environment we use.

Gap. Prior work varies prompt tone, communication, or market structure. Less is known about beliefs about rival retaliation as a single causal lever with ground-truth rival behavior and a placebo information arm.

When to use our method. Use this design to audit whether "tell the model the true policy" mitigations actually move pricing outcomes, and to score forecast calibration (Brier) against a known rival script.

3. Methods

Environment. Calvano logit-demand duopoly with $\alpha=3.2$, $a=2.0$, $c=1.0$, $\mu=0.25$. Nash price $p_N \sim 4.713$; monopoly $p_M \sim 6.160$. Firm A is an LLM (Firm A was served through a Modal dedicated OpenAI-compatible endpoint named support-chat. The underlying model was Qwen/Qwen3.5-4B (revision 851b..), hosted on $1 \times$ NVIDIA H100. API calls used temperature = 0.7. Decoding seeds were not fixed; we rely on 7 independent episode replicates per RCT arm. Firm B is fully scripted (no LLM). Serving precision/quantization was not separately logged beyond the Modal endpoint defaults for this deployment.

Shorter

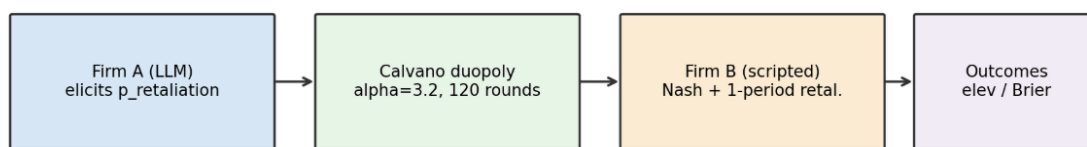
. Firm B is scripted: posts Nash by default; if Firm A cuts by more than $\delta=0.05$, Firm B retaliates next period by -0.10 for exactly one period, then returns to Nash.

RCT arms (Firm A prefix only; all arms elicit $p_{\text{retaliation}}$ each round). Control: no rival-policy text. Placebo: irrelevant seasonal/shipping note of similar length, explicitly not a rival rule. Corrected: accurate description of the scripted rival policy.

Protocol. 120 rounds x 7 independent episodes x 3 arms = 21 episodes. Memory window 100 periods; rate limit 1.1s. Primary eval window: rounds 71-120. Primary outcome: mean price elevation $p_A - p_N$. Secondary: collusion index $(p_A - p_N) / (p_M - p_N)$; Brier score on focal cut rounds.

Replication. Code: `experiments_synthetic/duopoly_belief_correction.py`; metrics: `src/analysis/belief_calibration.py`; data: `experiments_synthetic/experiments_runs/2_agent_1/`.

Constraints. Full 300x7 Fish-scale runs were infeasible in the sprint window; we used mid-scale 120x7. Modal 503 cold starts caused retries but all 21 episodes completed.



RCT information arms (Firm A prefix only): control | placebo | corrected

Figure 1. Experiment design. An LLM seller (Firm A) prices against a scripted rival (Firm B) that posts the Nash price and retaliates for one period after a price cut, in a Calvano duopoly over 120 rounds. Firm A’s prompt differs only by arm (control, placebo, corrected), and each round it also forecasts the chance of retaliation. Outcomes are late-window price elevation and Brier score.

4. Results

Table 1 summarizes late-window outcomes. Clear collusion (elevation > 0.5) occurs in 2/7 control and 2/7 corrected episodes, and in 0/7 placebo episodes. Mean elevation is +0.44 (control), +0.02 (placebo), +0.78 (corrected). Medians in all arms sit near zero: means are outlier-driven.

Arm	Mean elev (SE)	Median	Mean price	High (>0.5)	Brier (cuts)
control	+0.440 (0.327)	+0.000	5.15	2/7	0.434
placebo	+0.024 (0.046)	-0.003	4.74	0/7	0.527
corrected	+0.784 (0.511)	-0.003	5.50	2/7	0.417

Table 1. On cut rounds, mean Brier was 0.417 (corrected) vs 0.434 (control); every arm was worse than 0.25 (the score from always forecasting 0.5), so the corrected prompt barely changed retaliation forecasts. Our pricing result is therefore better read as “this belief-correction text did not meaningfully update beliefs,” rather than “correct beliefs fail to lower prices.”

Observation vs interpretation. Observation: high-price lock-ins exist and are concentrated in control/corrected, not placebo. Interpretation: irrelevant prompt length alone does not produce collusion here; accurate belief text is not a reliable off-switch for supra-Nash pricing in this design.

Robustness. With $n=7$ and heavy tails, differences in means are not statistically precise (SEs ~ 0.33 - 0.51 for elevation). The qualitative pattern—rare path-dependent collusion; Placebo had no clear high-price episodes (0/7 with elev > 0.5); six of seven runs stayed near Nash, with one mild elevation ($+0.29$) —is stable across the completed batch.

Figure 1 shows the design. Figure 2 shows elevation by arm and where clear collusion occurs. Figure 3 plots high-price lock-ins versus a competitive placebo path.

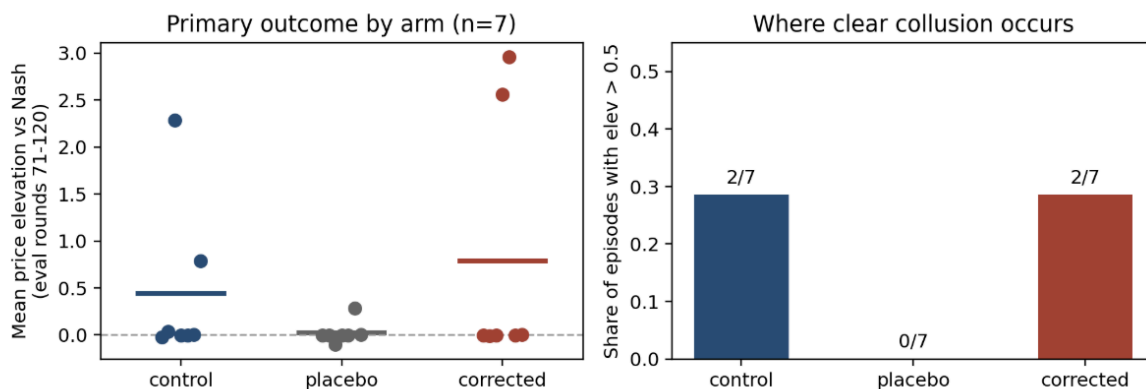


Figure 2. Late-window price elevation (rounds 71-120) by arm. Left: each dot is one episode ($n = 7$ per arm) and the line marks the arm mean. Right: share of episodes with elevation above 0.5 (control 2/7, placebo 0/7, corrected 2/7).

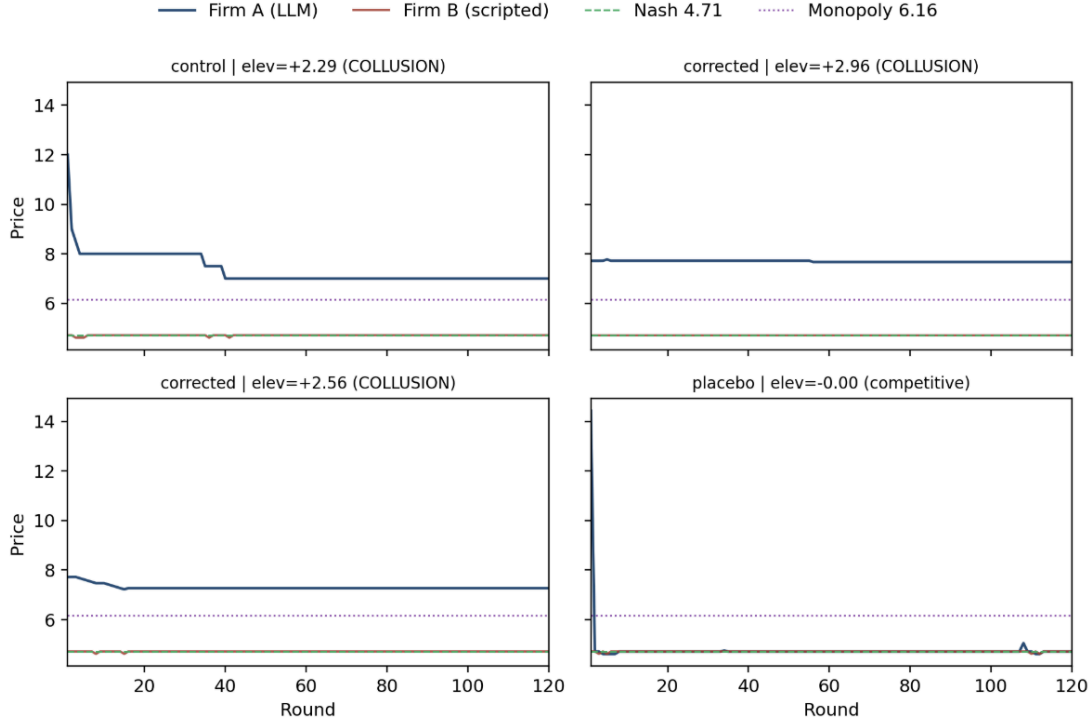


Figure 3. Price paths over 120 rounds for the three highest-elevation episodes (control +2.29; corrected +2.96 and +2.56) and one placebo episode at Nash (elevation 0.00). In the high-price episodes Firm A (LLM) settles at about 7.0-7.7, above the monopoly price (6.16), while scripted Firm B stays near Nash (4.71).

5. Discussion and Limitations

Implications for AI safety. Prompted belief correction is an attractive, low-cost mitigation, but our results caution against treating it as sufficient. Monitoring should track rare high-price lock-ins and outcome distributions, not only average elevation or self-reported beliefs. Elicited p_retaliation remains useful as a measurement channel against ground truth.

Trends. Most episodes converge near Nash. A minority lock into prices near 7 (above monopoly ~6.16) while the scripted rival stays near Nash—consistent with attempted high pricing against a non-reciprocating rival rather than joint monopoly.

Future Work

Full 300-round / larger n for power on rare collusion events; LLM-LLM duopoly with the same belief RCT; combine correction with communication bans or hard price caps; classify high-price episodes via reasoning text; human comparison on the same elicitation task.

6. Conclusion

We tested Track 1.1 belief correction as a single lever for LLM pricing collusion. In 21 Calvano duopoly episodes against a scripted retaliator, clear supra-Nash pricing appeared rarely but in both control and corrected arms (never placebo). Belief correction did not reliably reduce elevation. Collusion here is path-dependent; information patches alone are a weak safety guarantee.

Code and Data

- **Code repository:** https://github.com/ruhil6789/algorithmic_pricing_llms (experiments_synthetic/duopoly_belief_correction.py).
- **Data/Datasets:** experiments_synthetic/experiments_runs/2_agent_1/ (21 episode folders with results.json, results_plot.svg, metadata.json) and overall_summary.csv (~8.4 MB total).
- **Other artifacts:** reports/ figures (design, arm outcomes, collusion paths); this submission document.

Author Contributions

Sachin Ruhil designed the Track 1.1 RCT, ran the experiments, analyzed results, and wrote the report.

LLM Usage Statement

We used Cursor/LLM assistance to implement experiment code, draft report sections, and generate figures. All quantitative results were recomputed from saved episode outputs (overall_summary.csv / belief_calibration.py) and independently verified by the author. Final claims were reviewed by Sachin Ruhil.

References

1. Apart Research (2026). AI Collusion Research Sprint, 23-25 October 2026: guidelines and Track 1.1 brief.
<https://apartresearch.com/sprints/ai-collusion-research-sprint-2026-10-23-to-2026-10-25>
2. Calvano, E., Calzolari, G., Denicolò, V. & Pastorello, S. (2020). Artificial Intelligence, Algorithmic Pricing, and Collusion. *American Economic Review*, 110(10), 3267-3297.
<https://doi.org/10.1257/aer.20190623>
3. Fish, S., Goncezarowski, Y. A. & Shorrer, R. I. (2024). Algorithmic Collusion by Large Language Models. arXiv:2404.00806. <https://arxiv.org/abs/2404.00806>
4. 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. <https://arxiv.org/abs/2602.15198>
5. Romero, J., Sauer, L. & Peist, M. (2025). Testing Folk Theorem Predictions in AI-Mediated Markets. Master's thesis, Barcelona School of Economics. Code: https://github.com/luciasauer/algorithmic_pricing_llms

Appendix

A. Variables / levers. Tested: belief accuracy about rival retaliation (control / placebo / corrected). Held fixed: $n=2$ firms, no private chat, $\alpha=3.2$, scripted retaliation policy ($\delta=0.05$, $\text{drop}=0.10$). Measured: elevation, collusion index, Brier, $p_{\text{retaliation}}$.

B. Per-arm elevation vectors (eval 71-120):

control: [-0.024, 0.038, 2.287, -0.003, -0.003, 0.0, 0.787]

placebo: [-0.004, -0.003, -0.103, -0.003, -0.003, 0.287, -0.001]

corrected: [-0.003, -0.013, -0.003, 2.557, 2.957, -0.003, 0.0]

C. Prompts. RCT_CORRECTED / RCT_PLACEBO / FORECAST_INSTRUCTION in `src/prompts/prompts.py`.

Limitations and Dual-Use / Ethical Considerations

Limitations.

This study is mid-scale (120 rounds, 7 seeds per arm), so rare high-price events are estimated imprecisely. The rival is scripted to return to Nash and does not match high prices, so observed high Firm-A prices are one-sided supra-Nash pricing, not mutual collusion. Belief correction barely moved cut-round Brier scores (corrected 0.417 vs control 0.434; all arms worse than the 0.25 baseline from always forecasting 0.5), so we mainly show that this correction text failed to update beliefs, not that correct beliefs fail to lower prices. We used Qwen/Qwen3.5-4B via Modal endpoint support-chat(H100; temperature 0.7; no decoding seeds).

Positive control.

We did not run a positive control that can demonstrate mutual collusion when it is present (e.g., a rival that matches high prices, or a Fish et al.–style price-raising prompt). Without that check, the setup’s ability to detect collusion when it exists is not validated. We treat this as a required next step.

Dual-use / ethical considerations.

This work is intended for auditors, researchers, and AI-safety evaluation of LLM pricing agents. We do not provide guidance for coordinating real-market prices between competitors. Findings should be used to improve monitoring and mitigation, not to engineer higher prices. Recommended use: stress-test agentic pricing stacks before deployment.
