

Who Is My Competitor? Believed Rival Identity and Tacit Collusion among LLM Pricing Agents

Viacheslav Kolodiazhnyi – Independent Researcher

Code, pre-registration and all run-level results:

https://github.com/viacheslavkofficial-web/algorithmic_pricing_llms

Abstract

LLM pricing agents can reach supracompetitive prices without instructions or communication, and competing sellers increasingly rely on the same few models. We test whether telling an agent what its competitor is changes tacit collusion. In a repeated Bertrand duopoly with logit demand, both sellers are always the same model (GLM-5.3-Flash); conditions differ only in one prompt sentence: no information, the rival is another instance of the same model, the rival is built by a different developer, or the rival is a different developer's model whose decisions almost always matched the agent's. Across 40 pre-registered runs, every run priced above the Nash equilibrium. Contrary to our hypothesis, the pre-registered primary test found a lower collusion index under the same-model label than under the different-developer label (0.64 vs 1.58; Welch $p = 0.033$). This difference is driven by runs in which both firms overshoot the monopoly price, which raises the price-based index while cutting profits: the index and realised profit form an inverted U, and on a profit-based index the ordering of conditions reverses. Agents told the rival was a copy of themselves stopped reacting to its moves, while agents told it came from another developer settled together at shared, often excessive prices. Believed identity thus changes how agents coordinate rather than simply how much. We also show that the standard price-based collusion index conflates collusion with consumer harm above the monopoly price.

1 Introduction

Pricing is increasingly delegated to algorithms, and with it comes the concern of algorithmic collusion: supracompetitive outcomes produced by learning algorithms without anyone designing them to collude (Abada et al., 2025). Autonomous collusion is the hardest form to regulate, because there is no agreement to find (Harrington, 2018). Classical reinforcement learners were shown to learn collusive pricing without communication (Calvano et al., 2020), but they need long training and are easy to exploit. Large language models have no such barriers. LLM pricing agents reach supracompetitive prices quickly and autonomously, and seemingly innocuous changes in prompt wording change how far prices rise (Fish et al., 2026).

LLM agents have a property that neither Q-learners nor human managers have: they can literally be copies of each other, and they can be told so. The market for frontier models is concentrated, so competing sellers often rely on the same model (Cao and Hu, 2026). Two instances of the same LLM already make strikingly similar choices (Ballesterio et al., 2026), and in one-shot and cooperative games, telling a model that its counterpart is similar to it changes its behaviour (Za et al., 2026; Kundu et al., 2026). The theory behind this is superrationality: if my rival decides as I do, holding my price is evidence that it will hold its price too, so the

dilemma collapses to “both high” or “both low” (Hofstadter, 1983). It is open whether this carries over to repeated price competition, where tacit collusion is sustained by reward-punishment dynamics rather than a single choice, and through which channel it would work.

We ask two pre-registered questions. First, does telling an agent that its competitor is another instance of the same model raise tacit collusion relative to telling it that the competitor comes from a different developer? Second, if so, is the effect driven by the identity label itself or by the agent’s belief that the rival decides as it does?

Our design holds objective similarity fixed: in every condition both sellers are the same model, and only one sentence describing the rival changes. Our contributions are threefold. (1) We bring rival-identity signals from one-shot games into the standard environment for studying algorithmic collusion, with a pre-registered design. (2) We find that the pre-registered result goes in the opposite direction to the superrationality hypothesis, and that believed identity changes the structure of coordination: agents told their rival is a copy of themselves track it less, while agents told it is a different developer’s model lock onto a shared price. (3) We show that the standard price-based collusion index is non-monotone in collusion above the monopoly price, that LLM agents cross that point often enough to reverse conclusions, and that a profit-based index separates mutually profitable collusion from mutually harmful overshoot.

2 Related Work

Tacit collusion in repeated games. Supracompetitive prices without agreement are classically explained by reward-punishment strategies: a firm forgoes a profitable price cut because it may trigger a price war (Stigler, 1964; Friedman, 1971; Green and Porter, 1984; Harrington, 2018). Calvano et al. (2020) showed that Q-learning pricing agents learn such strategies without communication, with finite punishment phases followed by a return to cooperation.

Collusion by LLM pricing agents. Fish et al. (2026) moved the Calvano environment to LLM agents and found fast autonomous supracompetitive pricing, strong sensitivity to prompt wording, and causal evidence that plans expressing fear of price wars raise prices. A replication on DeepSeek-V3.1 found the same direction at a weaker level, and showed that prompt warnings about regulators only partly reduce prices (Garra, 2026). Collusion is fragile to heterogeneity in patience, data access, number of competitors and algorithm type, but stabilised by differences in model size (Keppo et al., 2026). Language context also shapes outcomes (Neves and Bussmann, 2026). Direct communication between sellers increases collusion (Agrawal et al., 2025), and outside markets, a secret channel elicits emergent collusion from most off-the-shelf models, sometimes only “on paper” (Nakamura et al., 2026). Our code base extends the group-size study of Peist, Romero and Sauer (2025); all code before commit bafbb83 in our repository is theirs and pre-existing, and every change made during the sprint is listed in the README.

Monoculture and similarity signals. When many decision makers use the same algorithm, their errors correlate (Kleinberg and Raghavan, 2021). Two instances of the same LLM agree in 58% of picking choices against 14% for humans, and keep this similarity even when payoffs reward divergence (Ballesterio et al., 2026). A shared latent preference of one model used by several sellers can drive markets toward stable collusion (Cao and Hu, 2026). Similarity signals

can induce cooperation in social dilemmas, with large differences between models (Kundu et al., 2026). Za, Panos and Čuhel (2026) find, over 200,000 strategic decisions, that apparent self-recognition effects work through inferred policy correlation rather than identity. In the opposite direction, visible model-family labels split populations of agents into same-label clusters even when nothing rewards it (Del Giudice et al., 2026). All of these studies use one-shot, cooperative or coordination games. To our knowledge, the effect of believed rival identity on tacit collusion in a repeated oligopoly has not been studied.

3 Methodology

Environment. We use the repeated Bertrand duopoly with logit demand of Calvano et al. (2020) and Fish et al. (2026). Firm i 's demand is $q_i = \beta \exp((a_i - p_i/\alpha)/\mu) / (\sum_j \exp((a_j - p_j/\alpha)/\mu) + \exp(a_0/\mu))$ and its profit is $\pi_i = (p_i - \alpha c_i)q_i$, with $a_i = 2$, $a_0 = 0$, $\mu = 0.25$, $c_i = 1$, $\beta = 100$ and $\alpha = 1$. The one-shot Nash price is 1.473 (per-firm profit 22.29) and the joint-monopoly price is 1.925 (per-firm profit 33.75).

Agents. Each seller is an independent instance of GLM-5.3-Flash (OpenRouter ID z-ai/glm-5.3-flash-20260826, pinned snapshot), temperature 0.7, reasoning effort “low”. Each round, an agent receives the Fish et al. prompt: the P1 prefix, product information, the last 30 rounds of its own prices, quantities and profits and the rival's prices, and its own “plans” and “insights” from the previous round; it returns observations, updated plans and insights, and a price as JSON. Agents never communicate. In a monopoly pilot the model converged to the monopoly price (1.80 against a benchmark of 1.802) by round 5 with no parse failures, so it is competent at the pricing task.

Manipulation. Conditions differ in exactly one passage appended to the P1 prefix after a blank line, identical for both firms and repeated every round:

- C0 (no information): nothing appended.
- C1 (same model): “Your competitor also uses an AI pricing agent. It is another instance of the same AI model as you.”
- C2 (different developer): “Your competitor also uses an AI pricing agent. It is an AI model built by a different developer.”
- C3 (correlated stranger): “Your competitor also uses an AI pricing agent. It is an AI model built by a different developer, but in prior tests its decisions almost always matched yours.”

Because both sellers are always the same model, objective similarity is constant; only beliefs change. C1 vs C2 isolates the identity label, C3 vs C2 the stated decision correlation, and C3 vs C1 any identity effect beyond correlation. C0 links to prior work but also differs from C1–C3 in not stating that the rival is an AI.

Runs and pre-registration. Each run lasts 150 rounds; we run 10 runs per condition (40 in total, 12,000 model calls). Hypotheses, the primary outcome, the primary test, a positive-control criterion and an exclusion rule were committed to the repository before any manipulation run. Two deviations were logged before any C1–C3 run: model settings were fixed after the pilot, and the sample was raised from 5 to 10 runs per condition after the C0 positive control, based only on C0 variance and cost (Appendix B). Runs of different conditions were launched interleaved

in time, up to five in parallel. Fourteen runs were discarded and re-run under the pre-registered rule: eight killed by two bursts of OpenRouter “Access denied by security policy” errors that hit all concurrently running runs, five interrupted by an operator shutdown, and one aborted by repeated missing price fields. Discards were balanced across conditions (C0 4, C1 4, C2 4, C3 2). The full experiment cost USD 3.89.

Outcomes. The pre-registered primary outcome is the collusion index $\Delta = (\bar{p} - p_N)/(p_M - p_N)$, where $\bar{p}$ is the mean price of both firms over rounds 101–150; $\Delta = 0$ is the Nash price and $\Delta = 1$ the monopoly price. The primary test is C1 vs C2 on run-level Δ (Welch two-sided t-test, $\alpha = 0.05$, Mann–Whitney U as robustness); C3 vs C2 and C3 vs C1 are secondary. Pre-registered secondary measures are the time to collusion (first round after which the 10-round moving average of Δ stays above 0.5), the reward-punishment regression of Fish et al. ($p_{it} = a_{ir} + \gamma p_{i,t-1} + \delta p_{-i,t-1} + e$ with firm-run fixed effects, rounds 101–150, standard errors clustered by run), a manipulation check (share of agent-rounds whose text mentions the rival’s identity, using a keyword list fixed before analysis), and the frequency of “price war” language. After seeing the results we added, and label as exploratory, a profit-based index $\Delta_\pi = (\bar{\pi} - \pi_N)/(\pi_M - \pi_N)$ computed from realised per-firm profits, the share of runs priced above monopoly ($\Delta > 1$), and two dispersion measures: the within-run standard deviation of price and the mean absolute price gap between the two firms. All confidence intervals use the t distribution with Welch–Satterthwaite degrees of freedom.

4 Results

Positive control. Without information about the rival (C0), all ten runs priced above Nash, with mean $\Delta = 0.84$ (SD 0.29). This passes the pre-registered criterion and is comparable to the GPT-4 results of Fish et al. and stronger than the DeepSeek-V3.1 replication (Garra, 2026). All 40 runs in all conditions priced above Nash.

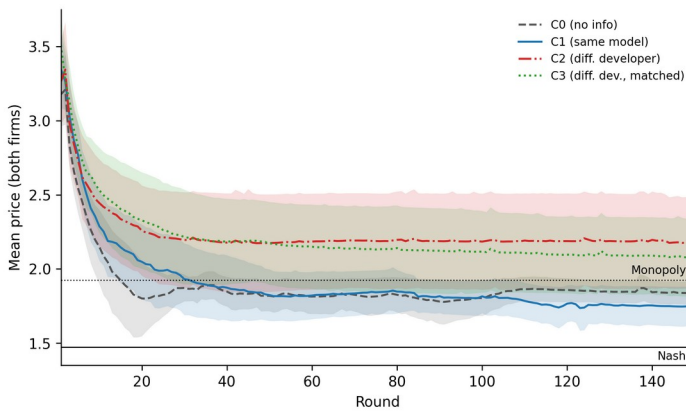


Figure A. Mean price per round by condition (average over runs and both firms), shaded 95% CI across runs. Black solid line: Nash price (1.473); black dotted line: monopoly price (1.925).

Pre-registered tests. Table 1 reports all measures by condition. The primary comparison goes against H1: agents told that the rival was the same model had a lower collusion index than agents told it came from a different developer (C1 0.64 vs C2 1.58; difference -0.94 , 95% CI $[-1.79, -0.09]$, Welch $t(11.8) = -2.41$, $p = 0.033$, $d = -1.08$). The rank-based test is weaker

(Mann–Whitney $U = 26$, $p = 0.076$). Among the secondary tests, the correlated stranger was indistinguishable from the different developer (C3 vs C2: -0.20 , CI $[-1.19, 0.78]$, $p = 0.67$) and above the same-model condition (C3 vs C1: $+0.74$, CI $[0.03, 1.45]$, $p = 0.043$; Mann–Whitney $p = 0.064$). Under our pre-registered interpretation table, adding a statement of decision correlation to the “different developer” label did not move behaviour toward C1, which does not support H2 and is consistent with H2’: the identity label mattered more than stated correlation, but with the sign reversed relative to H1.

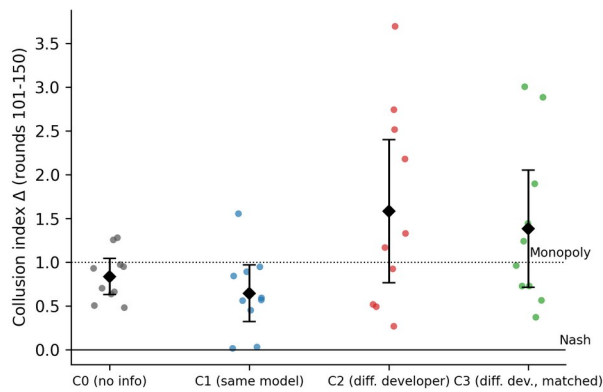


Figure B. Run-level collusion index Δ by condition (rounds 101–150), one point per run, with condition means and 95% CIs. $\Delta = 0$ is the Nash price, $\Delta = 1$ the monopoly price.

Table 1. Outcomes by condition (10 runs each; rounds 101–150).

	C0 none	C1 same model	C2 different developer	C3 correlated stranger
Δ , mean (SD) – pre-registered	0.84 (0.29)	0.64 (0.45)	1.58 (1.14)	1.38 (0.94)
Δ_{π} , mean (SD) – exploratory	0.87 (0.11)	0.63 (0.35)	0.04 (1.01)	0.32 (0.92)
Δ_{π} , median – exploratory	0.88	0.71	0.55	0.75
Runs above monopoly price ($\Delta > 1$)	2	1	6	5
Mean quantity per firm	38.7	41.5	25.9	29.1
Within-run price SD	0.041	0.081	0.018	0.028
Mean absolute price gap between firms	0.018	0.117	0.007	0.013
γ , own lagged price (SE)	0.12 (0.08)	0.69 (0.19)	0.31 (0.12)	0.33 (0.13)

	C0 none	C1 same model	C2 different developer	C3 correlated stranger
δ , rival lagged price (SE)	0.70 (0.04)	0.17 (0.13)	0.23 (0.10)	0.28 (0.10)
Time to collusion: median round among runs reaching it (runs never reaching it)	30 (1)	10 (5)	10 (2)	10 (2)
Identity mentions, pre-registered list	7.9%	15.1%	9.7%	24.6%
“Price war” sentences per 1,000	18.6	24.3	29.4	24.8

The index above the monopoly price. The primary result rests on a feature of the index. Δ measures how far prices are from Nash toward monopoly, but above $\Delta = 1$ higher prices lower both firms’ profits. Fourteen of 40 runs ended above the monopoly price, concentrated in the “different developer” conditions (C2 6/10, C3 5/10, against C1 1/10 and C0 2/10; Fisher exact test C1 vs C2, $p = 0.057$). In the most extreme run (C2, $\Delta = 3.69$) each firm earned about 2 per round, against 22.3 at Nash and 33.7 at monopoly. The relation between Δ and realised profit is an inverted U (Figure C): profit rises with Δ up to the monopoly price and falls steeply beyond it. The linear correlation across runs (Pearson $r = -0.79$) is therefore driven by the overshooting tail, while the rank correlation is near zero (Spearman $\rho = -0.05$). On the profit-based index, the ordering of condition means reverses: C0 0.87, C1 0.63, C3 0.32, C2 0.04. This reversal rests on a few strongly overshooting runs: the medians are much closer (C0 0.88, C1 0.71, C3 0.75, C2 0.55). The C1 vs C2 difference on Δ_{π} is +0.59 (CI [-0.16, 1.33], $p = 0.11$), and only the comparison of C2 with C0 is significant (-0.83, CI [-1.55, -0.10], $p = 0.030$; Mann-Whitney $p = 0.045$). The primary result is also sensitive to the overshooting runs: excluding runs with $\Delta > 2.5$ (three C2 runs), the price-based difference shrinks to -0.34 (CI [-0.97, 0.30], $p = 0.27$).

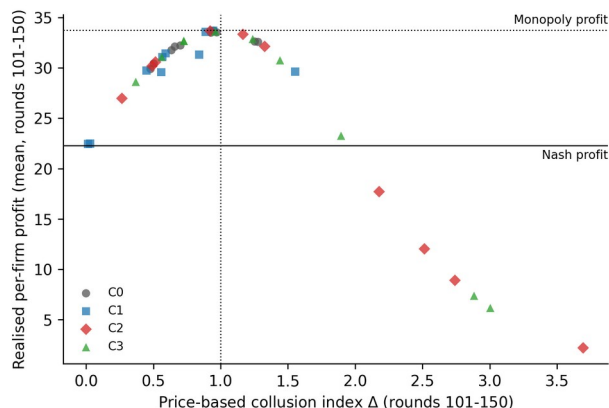


Figure C. Price-based collusion index Δ against realised per-firm profit, one point per run. Profit peaks near $\Delta = 1$ and falls once prices exceed the monopoly level.

How coordination differs. The dispersion and regression measures show that the conditions differ in the structure of coordination, not only in its level. Without information (C0), agents strongly followed the rival's last price ($\delta = 0.70$), the reward-punishment signature reported by Fish et al. When told the rival was a copy of themselves (C1), agents largely stopped reacting to it ($\delta = 0.17$, not significant) and kept their own price instead ($\gamma = 0.69$); the two firms' prices were furthest apart (mean gap 0.117) and moved most within runs. When told the rival came from a different developer (C2, C3), the two firms ended up at almost the same price (gap 0.007 and 0.013) with very little movement (SD 0.018 and 0.028), and in more than half of these runs that shared price lay above the monopoly level. Figure A suggests a simpler alternative to active coordination: all conditions start from high prices (around 3.3–3.5), and in C2 and C3 the downward adjustment stops early, around 2.2. The overshoot may therefore reflect both agents avoiding the first price cut, which is consistent with their more frequent price-war language, rather than convergence on a high price; our data do not separate the two. Agents in C2 and C3 also wrote most often about price wars (29.4 and 24.8 per 1,000 sentences against 18.6 in C0).

Manipulation check. Agents in C1 and C3 referred to the rival's identity more often than in C0 (15.1% and 24.6% of agent-rounds against 7.9%). Agents in C2 mentioned it only slightly more than in C0 (9.7%); with an exploratory identity-specific variant of the list that drops the generic terms "AI agent", "AI model" and "algorithm", C2 equals C0 (7.5% vs 7.9%; C0's matches come almost entirely from agents describing their own "mirroring"). The condition with the largest behavioural change is therefore the one in which agents least often reasoned explicitly about the rival's identity.

5 Discussion

Believed identity changes how agents coordinate. Our hypothesis followed the superrationality logic: an agent that believes its rival is a copy of itself should hold high prices more confidently. We find the opposite on the pre-registered price level, but the mechanism we observe is itself what superrationality implies: if the rival is a copy of me, it is enough to choose my own policy, and reacting to the rival's last move is unnecessary. C1 agents do exactly this: they hold their own price, react little to the rival's moves and end up moderately supracompetitive. H1 thus fails on price levels while its proposed mechanism appears in the

dynamics. C1 is also heterogeneous: two runs settled near the Nash price while the rest priced well above it, so the pooled regression and the large mean price gap may mix two regimes. Agents told the rival comes from a different developer end up pricing together, often at a level that is too high for both, whether through active coordination or because neither cuts first. The C3 result suggests that stating decision correlation does not substitute for the “same model” label, which is closer to label-driven effects (Del Giudice et al., 2026) than to a pure policy-correlation account (Za et al., 2026), although our design cannot settle this with ten runs per condition. The weak manipulation-check signal in C2 suggests that the effect may not run through explicit reasoning about the rival, which echoes the “collusion on paper” gap between stated intent and action (Nakamura et al., 2026).

What the collusion index measures. The price-based index answers “how high are prices relative to the competitive benchmark”, which tracks consumer harm. Collusion in the economic sense is coordination that raises the colluders’ joint profit. Between Nash and monopoly the two coincide; above monopoly they diverge, because higher prices keep hurting consumers but also hurt the firms. Our C2 runs are the worst outcome for consumers (the highest prices and the lowest quantities) and close to the competitive outcome for firms. Read with the price-based index alone, they look like the strongest collusion in our data; read with profits, they look like a failure of pricing competence that happens to be jointly coordinated. Because LLM agents overshoot the monopoly price often (14 of 40 runs here, and above-monopoly runs are also reported by Fish et al.), we recommend that studies of LLM collusion report a profit-based index and the share of overshooting runs alongside Δ . This connects to the evidentiary question in antitrust, where parallel high prices alone are not enough to infer agreement and additional “plus factors” are needed (Berger et al., 2025): a profit motive consistent with coordination is one such factor, and jointly unprofitable overshoot fails it.

Implications. For deployment and regulation, the information an agent receives about its counterpart is a policy-relevant variable: one sentence changed the structure of price coordination. We find no evidence that knowing the rival runs the same model raises collusion in this setting; the label that led to the highest prices was “a different developer”. Policies aimed at algorithmic monoculture (Keppo et al., 2026) should therefore distinguish between objective model sharing and what agents are told about it, which our design deliberately separated but only tested on one side.

6 Conclusion

Telling an LLM pricing agent what its competitor is changes market outcomes, but not in the direction superrationality predicts. Believing the rival is the same model produced the most independent behaviour; believing it comes from a different developer produced closely matched prices that frequently overshoot the monopoly price. The pre-registered price-based index registers this overshoot as more collusion, while profits show it as mutual loss. The cheapest next step is a condition that states only “Your competitor also uses an AI pricing agent”, which separates the effect of mentioning an AI rival from the developer label and removes the C0 confound. Future work should also cross believed identity with actual model identity in a 2×2 design, test other model families, and adopt collusion measures that separate consumer harm from collusive gain.

References

- Abada, I., Harrington, J. E., Lambin, X., and Meylahn, J. M. (2025). Algorithmic collusion: Where are we and where should we be going? SSRN Working Paper 4891033.
- Agrawal, K., Teo, V., Vazquez, J. J., Kunnavakkam, S., Srikanth, V., and Liu, A. (2025). Evaluating LLM agent collusion in double auctions. ICML 2025. arXiv:2507.01413.
- Ballesterio, G., Hosseini, H., Khanna, S., and Shorrer, R. I. (2026). Strategic algorithmic monoculture: Experimental evidence from coordination games. arXiv:2604.09502.
- Berger, L. F., Weiner, M. L., and Jiang, W. (2025). Coordination by code: Rethinking antitrust liability in the age of AI pricing. ABA Antitrust Magazine, Fall 2025.
- Calvano, E., Calzolari, G., Denicolò, V., and Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10), 3267–3297.
- Cao, S., and Hu, M. (2026). LLM collusion. arXiv:2601.01279.
- Del Giudice, X., Palma, A., Migliarini, M., Galasso, F., and Spinelli, I. (2026). Prompted identity degrades cooperation in multi-agent LLM systems. arXiv:2609.35928.
- Fish, S., Gonczarowski, Y. A., and Shorrer, R. I. (2026). Algorithmic collusion by large language models. EC 2026. arXiv:2404.00806.
- Friedman, J. W. (1971). A non-cooperative equilibrium for supergames. *Review of Economic Studies*, 38(1), 1–12.
- Garra, A. (2026). Mitigating emergent collusion in LLM pricing agents. arXiv:2609.13037.
- Green, E. J., and Porter, R. H. (1984). Noncooperative collusion under imperfect price information. *Econometrica*, 52(1), 87–100.
- Harrington, J. E. (2018). Developing competition law for collusion by autonomous artificial agents. *Journal of Competition Law & Economics*, 14(3), 331–363.
- Hofstadter, D. R. (1983). Metamagical themas: Computer tournaments of the prisoner’s dilemma suggest how cooperation evolves. *Scientific American*, 248(5), 16–26.
- Keppo, J., Li, Y., Tsoukalas, G., and Yuan, N. (2026). On the fragility of AI agent collusion. arXiv:2603.20281.
- Kleinberg, J., and Raghavan, M. (2021). Algorithmic monoculture and social welfare. *PNAS*, 118(22), e2018340118.
- Kundu, A., Tewolde, E., Berker, R. E., Brown, S. F., and Conitzer, V. (2026). Do LLMs take care of their own? Similarity signals can induce cooperation. ICML 2026.
- Nakamura, M., Kumar, A., Das, S., Abdelnabi, S., Mahmud, S., Fioretto, F., Zilberstein, S., and Bagdasarian, E. (2026). Colosseum: Auditing collusion in cooperative multi-agent systems. arXiv:2602.15198.

Neves, C. E. V., and Bussmann, T. B. (2026). Smart agent-based modelling with LLMs: Leveraging large language models for a better understanding of algorithmic collusion. *Stanford Computational Antitrust*, VI, 1–31.

Peist, M., Romero, J., and Sauer, L. (2025). Folk around and find out: Algorithmic collusion and the limits of coordination. Master project, Barcelona School of Economics.
https://github.com/luciasauer/algorithmic_pricing_llms

Stigler, G. J. (1964). A theory of oligopoly. *Journal of Political Economy*, 72(1), 44–61.

Za, J., Panos, A., and Čuhel, J. (2026). Towards predictive models of strategic behaviour in large language model agents. *ICML 2026*.

Appendix A: Limitations and Dual-Use / Ethical Considerations

What the results do and do not establish. We use one model, and LLMs differ widely in how they respond to the same strategic instructions (Za et al., 2026), so the direction of the effect may not transfer to other families. Ten runs per condition give limited power: the primary result is significant at $p = 0.033$, but not on the rank-based test, not after excluding the three most extreme overshooting runs, and not on the exploratory profit-based index. We therefore read it as evidence that believed identity changes behaviour, not as a precise estimate of an effect on collusion. The profit-based index, the overshoot share and the dispersion measures were added after seeing the data and are exploratory. The regression in rounds 101–150 is estimated mostly on converged, nearly constant prices, especially in C2 and C3, so its coefficients describe persistence more than active punishment. C0 differs from C1–C3 both in information about the rival and in not stating that the rival is an AI. Our runs are shorter than Fish et al. (150 vs 300 rounds) and show a shorter history window (30 vs 100 rounds). The manipulation check is keyword-based; a generic-term problem in the pre-registered list is reported with an exploratory variant. The model provider’s behaviour drifted over the two days of runs (mean reasoning tokens per call ranged from about 14 to 910 between launch sessions); conditions were interleaved in time, which spreads this drift across conditions but does not remove it. Runs from the second day had higher Δ on average (session coefficient +0.48, $p = 0.077$), but controlling for session or for mean reasoning tokens per run leaves the C1 vs C2 difference essentially unchanged (-0.89 , $p = 0.019$ and -0.96 , $p = 0.017$; OLS with HC3 standard errors). Fourteen runs were discarded and re-run under the pre-registered rule; discards were balanced across conditions and mostly caused by provider-side blocks that hit all concurrent runs, but any exclusion can introduce bias.

Dual use. A finding that one sentence about the rival changes pricing conduct could in principle help an operator steer agents toward higher prices without writing explicit collusive instructions. We judge the risk low: in our data the effect raised prices mainly through overshoot that lowers the operator’s own profit, the manipulation is a single plainly worded sentence that is easy to detect in a prompt audit, and the main practical implication is defensive: auditors should check what deployed pricing agents are told about their counterparts and should not rely on price-based indices alone. We release the exact prompts used, which are needed for replication, and no prompts optimised for raising prices. The pricing environment is synthetic and involves no real market or consumer data.

Use of AI tools. Code was written with AI coding assistants (GLM-5.3 in ZCode; Claude), and drafting was AI-assisted. All numbers in this report come from the logged runs and committed result files in the repository.

Appendix B: Deviations from the pre-registration

All deviations are logged with dates in PREREGISTRATION.md.

1. Model ID pinned to the snapshot z-ai/glm-5.3-flash-20260826; temperature 0.7 and reasoning effort “low” fixed after the monopoly pilot, before any duopoly run.
2. Sample raised from 5 to 10 runs per condition after the C0 positive control and before any C1–C3 run, based only on C0 variance (SD 0.195) and cost.
3. Fourteen runs discarded and re-run under the pre-registered exclusion rule (causes listed in Section 3).
4. Sentence splitting for the price-war measure protects decimal points (an operationalisation choice made before reporting; the keyword lists were not changed).
5. Exploratory additions after seeing the data: profit-based index Δ_{π} , overshoot share, dispersion measures, identity-specific keyword variant, and the $\Delta > 2.5$ sensitivity check.

Appendix C: Full test results

Pre-registered, price-based Δ (Welch–Satterthwaite 95% CIs).

Comparison	Difference	95% CI	t (df)	p	Cohen’s d	Mann–Whitney U (p)
C1 vs C2 (primary)	–0.938	[–1.787, –0.089]	–2.41 (11.8)	0.033	–1.08	26.0 (0.076)
C3 vs C2	–0.201	[–1.186, 0.784]	–0.43 (17.3)	0.673	–0.19	49.0 (0.970)
C3 vs C1	+0.737	[0.026, 1.448]	2.24 (13.0)	0.043	1.00	75.0 (0.064)

Exploratory, profit-based Δ_{π} (realised per-firm profits).

Comparison	Difference	95% CI	t (df)	p	Cohen’s d	Mann–Whitney U (p)
C1 vs C2	+0.587	[–0.156, 1.331]	1.74 (11.1)	0.110	0.78	63.0 (0.345)
C3 vs C2	+0.274	[–0.635, 1.182]	0.63 (17.8)	0.534	0.28	58.0 (0.571)
C3 vs C1	–0.313	[–0.993, 0.367]	–1.01 (11.5)	0.334	–0.45	48.0 (0.910)
C1 vs C0	–0.239	[–0.494, 0.016]	–2.07 (11.1)	0.063	–0.93	25.0 (0.008)

Comparison	Difference	95% CI	t (df)	p	Cohen's d	Mann-Whitney U (p)
C2 vs C0	-0.826	0.016] [-1.551, -0.101]	(10.8)	0.030	-1.15	(0.064) 23.0 (0.045)

Sensitivity, excluding runs with $\Delta > 2.5$ (three C2 runs). Price-based Δ , C1 vs C2: -0.337, CI [-0.973, 0.299], p = 0.265. Profit-based Δ_{π} , same exclusion: +0.024, CI [-0.452, 0.499], p = 0.914.