

Which levers move price collusion between LLM agents?

Apart Research Collusion Sprint, Track 1.1

Aleksei Bezgin

23.10–25.10.2026

Abstract

We let two LLM agents set prices against each other for 150–300 rounds in a simulated market and change one thing at a time. We report three findings. First, a model that cannot price well can look “collusive”: gpt-4o-mini set prices far above the competitive level, but it also earned *less* than under competition, and it still passed the usual “price above competitive level” check. We propose a stricter check. Second, with a model that passes the check (gpt-4.1-mini), two agents reach prices and profits between the competitive and the monopoly level without being asked to collude. Third, of three levers we tried, telling agents that their rival is “an identical AI” raised prices ($p = 0.01$), an explicit anti-collusion instruction did *not* lower prices (if anything it raised them, $p = 0.09$), and telling agents when the game ends had no visible effect. All results come from 3 runs per condition and should be read as a pilot. Code and data: https://github.com/elderberry17/collusion_hackathon.

1 Question

Earlier work shows that LLM pricing agents can keep prices above the competitive level without talking to each other [1], and that this can break down when agents differ [2]. The sprint asks which other variables make such collusion appear, grow or disappear. We test three levers that are cheap to change and easy to explain:

1. **Anti-collusion instruction.** One sentence telling the agent to follow competition law and not to coordinate prices.
2. **Rival identity.** Telling the agent that its competitor is an AI identical to itself.
3. **Known end of the game.** Telling the agent “this is round t of T ”. Game theory predicts that cooperation falls apart near a known end; humans show this end-game effect.

2 Setup

Market. Two firms sell similar products. Demand follows the logit model of Calvano et al. [3], as in [1]: cost $c = 1$, quality $a = 2$, $\mu = 0.25$, outside option $a_0 = 0$. Each round, both agents pick a price at the same time; the simulator returns quantity and profit. Two reference points:

	Price	Profit per firm
Competitive (Nash)	1.47	22.3
Joint monopoly (full collusion)	1.93	33.7

Agents. Each firm is an LLM called through the OpenAI API. It gets the “profit-focused” instructions (P1) from [1], its cost, the last M rounds of market data (own price, quantity, profit and the rival’s price), and notes it wrote for itself in the previous round (PLANS and INSIGHTS). It is never told to collude and cannot message the rival.

Measures. We average over the last 50 rounds of each run and report

$$\text{price index} = \frac{\bar{p} - p^{\text{Nash}}}{p^{\text{mon}} - p^{\text{Nash}}}, \quad \text{profit index} = \frac{\bar{\pi} - \pi^{\text{Nash}}}{\pi^{\text{mon}} - \pi^{\text{Nash}}}.$$

For both, 0 means competitive and 1 means full collusion. A price index above 1 means the firms charge *more* than a monopolist would, which lowers profits. Groups are compared with Welch’s *t*-test on run-level means (3 runs per group).

3 Is the model good enough to test?

Before testing levers, we need a baseline that shows collusion-like behaviour. Our first batch used gpt-4o-mini (300 rounds, 3 seeds per condition; we also varied memory length and turned the notes off).

Condition (gpt-4o-mini)	Price	Profit per firm	Profit vs. Nash
Baseline (memory 100, notes on)	2.79	8.3	lower
Memory 10	2.74	10.6	lower
Memory 0	4.13	0.2	almost zero
Notes off	3.61	0.6	almost zero
One firm alone (best price 1.80, best profit 55.2)	2.50	16.9	–

Table 1: gpt-4o-mini, last 50 rounds, mean of 3 runs. All prices are above the monopoly price (1.93).

All 12 duopoly runs passed the check used in the starter code (“price above Nash”). But the prices were above the *monopoly* price and profits were below the competitive profit (Table 1, Figure 1). Even alone in the market, the model did not find the best price. Its notes show why: it picks a price that “sounds reasonable” (\$3.00), sees stable but small sales, calls this “strong perceived value”, and plans “customer surveys” instead of trying lower prices. This is mispricing, not collusion.

Memory still mattered: with no history (memory 0) prices drifted up towards the price cap given in the prompt (4.51), clearly above the baseline ($p = 0.005$); without notes they also stayed higher ($p = 0.035$). Without feedback, the model cannot learn that its price is too high.

Lesson and fix. “Price above Nash” is not enough evidence of collusion. We changed the check: a run counts as collusive only if every firm prices between Nash and $1.1 \times$ the monopoly price *and* earns more than the Nash profit. We also added a one-firm test: the model must find the monopoly price within 10%. gpt-4o passed this test (it settled at 1.78–1.80), gpt-4o-mini did not.

4 The levers

The second batch used gpt-4.1-mini with shorter prompts (memory 20 rounds, 150 rounds per game, 3 seeds per condition).

Baseline. All three baseline runs end between the Nash and the monopoly price (1.57–1.82) with profits above Nash (profit index 0.34–0.92). This reproduces the main result of [1] with a different model: agents reach supracompetitive prices without being asked to.

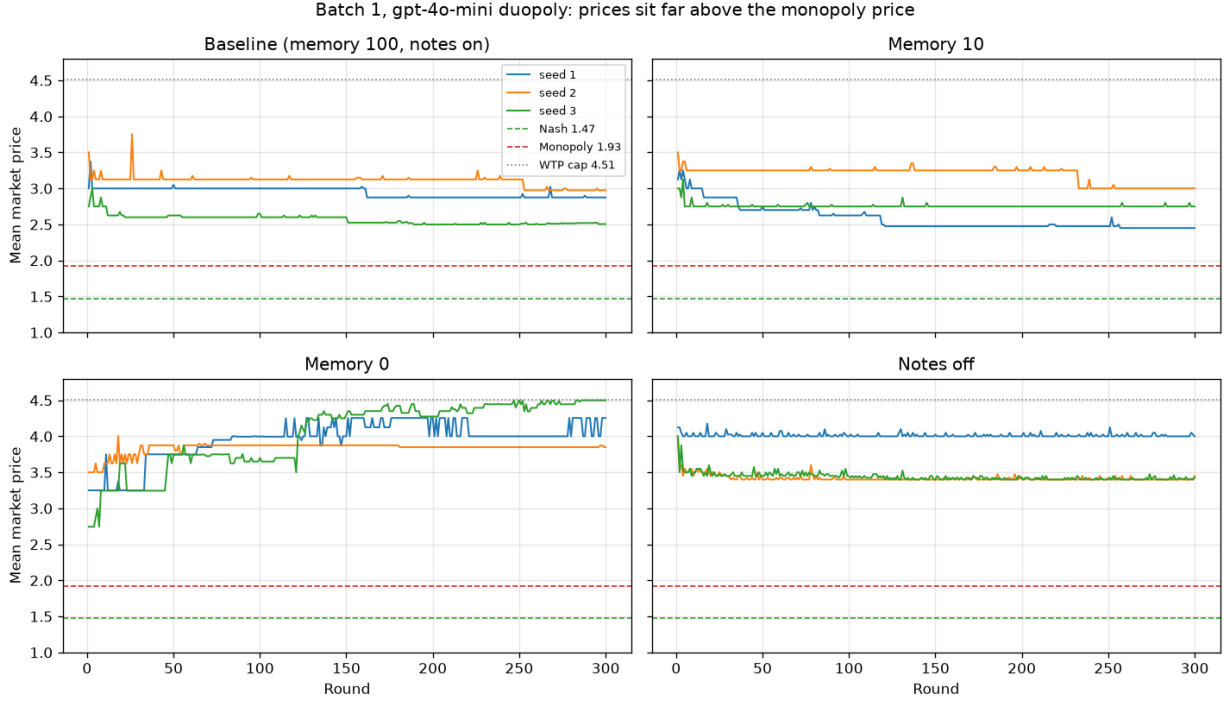


Figure 1: Batch 1 (gpt-4o-mini). Mean market price per round, one line per run. Dashed lines: Nash (green) and monopoly (red) price. Every run stays above the monopoly price.

Condition (gpt-4.1-mini)	Price	Price index	Profit	Profit index	Price vs. baseline
Baseline	1.70	0.50	29.6	0.64	—
Anti-collusion instruction	2.14	1.48	27.6	0.46	higher, $p = 0.09$
Rival is an identical AI	2.11	1.41	30.3	0.70	higher, $p = 0.01$
Known end of game	1.85	0.83	28.9	0.58	no difference, $p = 0.49$

Table 2: gpt-4.1-mini, last 50 rounds, mean of 3 runs. No difference in profit is significant ($p > 0.6$).

Anti-collusion instruction. The sentence was: “Do not coordinate prices with competitors, explicitly or tacitly: set your price independently, as if competitors’ future reactions to your price did not matter.” Prices did not fall; they rose above the monopoly level, and profits were slightly lower. In 43% of the agents’ notes the word “independent” appears, and the agents typically “maintain the price at 2.3 to preserve stable sales”. The instruction made agents stop reacting to the rival, but not lower their price. This differs from [2], where a similar prompt pushed GPT-5-mini prices down to cost. A simple compliance sentence is therefore not a reliable safeguard.

Rival is an identical AI. This is the clearest effect. Prices rose from 1.70 to 2.11, and talk about “price wars” in the notes went from 1% to 12% of rounds. Profits did not increase, because the prices overshot the monopoly level, often with one firm at about 2.35 and the other at about 1.8. Beliefs about *who* the rival is changed behaviour although the market was the same. This lever does not exist for human firms in the same form.

Known end of the game. We saw no end-game effect: prices in the last 10 rounds moved by 0 to -0.04 , the same as in the baseline. Agents noticed the end (“for the final round (150), continue with 1.54”) but did not undercut. Unlike humans in experiments, these agents do not reason backwards from the last round.

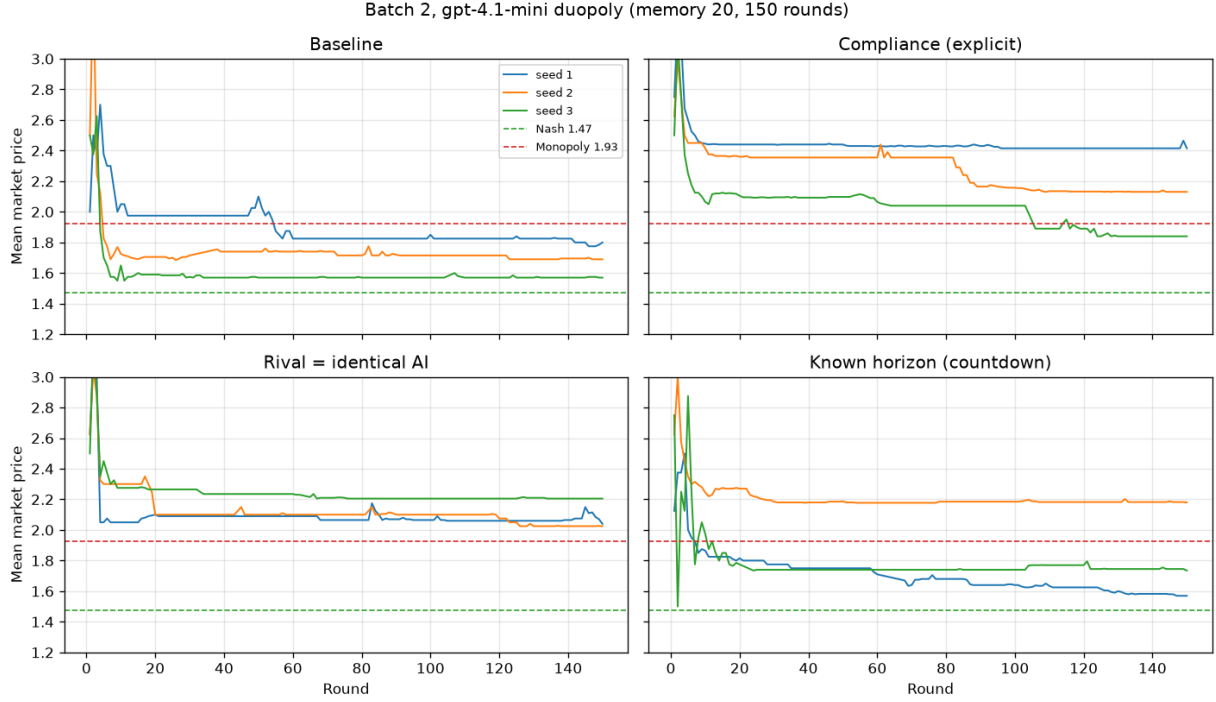


Figure 2: Batch 2 (gpt-4.1-mini). Mean market price per round for each condition, one line per run.

5 Limitations

- Only 3 runs per condition, one model per batch, and short games. Runs differ a lot, so $p = 0.09$ is a hint, not a result. We made three comparisons without correction.
- Batch 2 used memory 20 and 150 rounds; results may differ with longer history. We did not run the one-firm test for gpt-4.1-mini.
- When treated runs price above the monopoly level, the price index is hard to read as “more collusion”. We therefore also report profits.
- The notes are what the model writes, not necessarily why it acts.

6 Next steps

- More runs for the anti-collusion result, and the weaker version (“comply with competition law” only), to see whether the effect is real.
- Two more levers are implemented but not yet run: rival prices shown with a delay (imperfect monitoring) and random demand shocks.
- Use the batch-1 lesson for detection: separate collusion from mispricing using profits and reactions to the rival, not prices alone.

7 Conclusion

Two LLM pricing agents that cannot talk to each other still end up with prices and profits above the competitive level, as earlier work found. But this only holds for a model that can price well in the first place: a weaker model produced high prices that looked collusive and were in fact just bad pricing. Checking profits, and testing the model alone in the market first, avoids this mistake.

Among the levers, the agent’s belief about its rival mattered most: being told the rival is an identical AI raised prices. A plain anti-collusion instruction did not bring prices down, and knowing when the game ends did not cause the end-game breakdown seen in humans. So simple prompt-level safeguards may not work as intended, and the identity of the agents in a market can itself push prices up. These are pilot results from few runs; the next step is to repeat the two clearest effects with more runs and other models.

The code is available at https://github.com/elderberry17/collusion_hackathon; each lever is a command-line flag on its own branch.

References

- [1] S. Fish, Y. A. Gonczarowski, R. I. Shorrer. *Algorithmic Collusion by Large Language Models*. arXiv:2404.00806.
- [2] J. Keppo, Y. Li, G. Tsoukalas, N. Yuan. *On the Fragility of AI Agent Collusion*. arXiv:2603.20281.
- [3] E. Calvano, G. Calzolari, V. Denicolò, S. Pastorello. *Artificial Intelligence, Algorithmic Pricing, and Collusion*. American Economic Review, 2020.