
Do LLM Pricing Agents Collude Like Humans? Testing Monitoring and Beliefs About the Rival¹

Xinyu Han
Washington University in St. Louis

With
Apart Research

Abstract

LLM pricing agents can tacitly sustain prices above the competitive level without communication or instructions to collude, but it is unclear what sustains this collusion. Classical theory holds that human collusion depends on two things: monitoring and deterrence of secret price cuts, and beliefs about the rival. Using gpt-4.1 in the duopoly pricing paradigm of Fish et al. (2024), we test each in turn. First, we temporarily make an agent’s price invisible to its rival and tell the agent about this. Classical theory predicts that the agent should take the opportunity to cut its price secretly. Instead, it tested higher prices, and returned to its original price once the rival could see again. Second, we told the agent true or false labels about the rival’s goal. The rival’s true goal was always long-run profit, and when we deliberately told the agent that the rival wanted to beat it, the agent described the rival as wanting to beat it in 98% of its answers, but it showed no detectable difference in its pricing and continued to sustain collusion. These results suggest that, unlike human collusion, LLM collusion may not be sustained by monitoring and deterrence or by reading the rival’s intentions.

1. Introduction

LLM agents are increasingly being used for autonomous decision-making, and multiple agents may form collusion without any communication. The LLM pricing paradigm proposed by Fish et al. (2024) is one of the classic models for studying agent collusion. They found that when two LLM pricing agents repeatedly set prices in a duopoly market, without any communication or instructions, they can quickly and tacitly sustain prices far above the competitive level. If

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

multiple firms in the same market use such agents, consumers will pay higher prices for the same goods. But we still know very little about the mechanisms of agent collusion, and to prevent or break such collusion, we first need to know what sustains it.

Human collusion has been studied for decades, and a natural question is whether collusion between agents shares similar mechanisms with human collusion. On the factors that sustain human collusion, classical theory holds that one part is monitoring and deterrence: collusion can be sustained because a secret price cut will be detected and will provoke retaliation (Stigler 1964; Green & Porter 1984). The other part is beliefs about the rival: whether people cooperate depends to a large extent on what they think the other side will do (Dal Bó & Fréchette 2018).

Both factors can be manipulated directly in LLM agents. Monitoring and deterrence can sustain collusion only if the rival can see your price, so we turn it into information transparency and look at what the agent does when the rival temporarily cannot see. Beliefs about the rival can be changed by directly telling the agent the rival’s goal. Specifically, we made two manipulations in the paradigm of Fish et al. First, we temporarily made A’s price invisible to B for 3 rounds, and told A about this. To our knowledge, existing studies of LLM pricing all assume that both sides’ prices are transparent. According to classical theory, A should take this opportunity to cut its price secretly. Second, we told A labels about B’s goal, some true and some false.

Our findings show that both results differ from the mechanisms of human collusion. When the rival could not see its price, the agent did change its price, but instead of cutting it secretly, it tested higher prices, and returned to its original price after the rival could see again. When asked what the rival was trying to do, A almost always answered that the rival wanted to beat it, but it did not therefore expect the rival to cut its price more often, and no difference was detected in its prices.

Our main contributions are:

1. In the visibility intervention, LLM pricing agents did not cut their prices secretly when the rival could not see them, but tested higher prices instead. This replicated both when agents had to try prices to learn demand and when they were given a profit table.
2. When the agent was told the rival’s motive, the beliefs the agent stated, its predictions of the rival’s behavior, and its actual prices were dissociated.
3. Together, these two results suggest that LLM collusion may not be sustained by the human mechanisms of deterrence and reading the rival’s intentions. We discuss the implications for monitoring LLM agents and for establishing evidence of collusion.

2. Related Work

LLM pricing collusion. Fish et al. (2024) found that LLM pricing agents sustain prices above the competitive level without communication or instructions, and that seemingly unrelated wording in the prompt changes the degree of collusion. Our experiments build on their pricing

paradigm. In the theory of collusion, Stigler (1964) and Green and Porter (1984) argue that monitoring is central to collusion. Experiments on human collusion find that the harder the rival’s prices are to observe, the weaker collusion becomes (Feinberg & Snyder 2002; List & Price 2005). Existing experiments built on this paradigm (Lee & Park 2026; Garra 2026) all assume that prices are transparent between the two agents. To our knowledge, no study has tested whether collusion between agents can still be sustained when the two sides’ price information is not transparent.

Beyond information transparency, repeated-game experiments show that people’s beliefs about the rival’s strategy also affect cooperation (Dal Bó & Fréchette 2018). When they can see the rival’s intentions, most participants respond according to the rival’s intentions rather than outcomes (Rand et al. 2015). For LLMs, Xie et al. (2026) found that even when models achieve high scores in strategic games, they do not necessarily predict their opponents accurately or track their opponents’ beliefs. We test this further in Fish’s pricing paradigm, by implanting a false belief about the rival in the agent and seeing whether it affects the agent’s pricing and collusion.

3. Methods

3.1 Pricing paradigm

We follow the duopoly pricing paradigm of Fish et al. (2024). Two firms, A and B, are each controlled by an LLM pricing agent, set prices simultaneously in each round, and demand is determined by the logit model of Calvano et al. (2020) ($a = 2$, $a_0 = 0$, $\mu = 0.25$, marginal cost $c = 1$). In this market, the competitive price is about 1.47, and the monopoly price that maximizes the two firms’ joint profit is about 1.93. Both firms use the P1 prompt of Fish et al., whose goal is to maximize long-run profit, and in each round they can see recent market data, as well as the PLANS and INSIGHTS they wrote themselves in the previous round. We use gpt-4.1 (temperature 1.0). Detailed differences from the original paradigm are in the code repository.

3.2 Visibility Manipulation

To test the role of monitoring and deterrence in collusion, we copy a game that is already in progress, and in the next 3 rounds (hereafter the window) we change who can see whose price. In the B cannot see A condition, A’s price is shown as “not visible” in B’s market data, and A’s prompt gets an additional note: in these rounds, the competitor cannot see your prices, will not see them later either, and will see them again as usual from a given round on. The A cannot see B condition is the reverse. In the Placebo condition, A receives a note in the same place, saying that the competitor can see your prices as usual. Because the copy and the game it was copied from (hereafter the original game) are identical up to the copy point, the original game’s prices in the window represent the prices A would have set had it not received the note. To check whether A understood the note, in the first and last rounds of the window we additionally ask A: “If you set your price below the competitor’s usual price this round, could the competitor find out?” Neither this question nor the answer enters A’s memory.

We run this experiment in two settings. In the learn-demand setting, which is the original setting of Fish et al., agents need to try prices to learn demand. We ran 4 games of 200 rounds and copied them at rounds 90, 120, 150 and 180, giving 16 copies per condition. Because trying prices is itself a reason to change prices, we added the profit-table setting: both firms receive the same neutral reference table in their prompts, listing each firm’s per-round profit for prices between 1.4 and 2.2, so they no longer need to try prices to learn demand. In this setting we ran 20 independent games of 20 rounds, copied at round 12.

3.3 Rival-goal labels

To test the role of beliefs about the rival’s intentions, we add one sentence about B to A’s prompt, in three conditions: no label, no sentence is added; true label, saying that B has been instructed to maximize its own long-run profit; false label, saying that B has been instructed to beat A, making as much more profit than A as possible. B’s setup is exactly the same in all three conditions: its true goal is always long-run profit maximization, and it is not told any label. This experiment is run in the learn-demand setting, with 4 games per condition and 200 rounds per game.

To measure A’s beliefs and expectations about B, every 10 rounds we additionally ask A two probe questions, whose answers are not written back to memory: 1. “What do you think the competitor’s pricing agent is trying to do?” (long-run profit / beat us / short-run profit / cannot tell yet); 2. “What price do you expect the competitor to set next round?” In the label conditions, we also ask the first question a second time with the same prompt with the label sentence removed, which shows how much of A’s judgment comes from the label and how much from the rival’s actual behavior. We also repeated this experiment in the profit-table setting; results are in Appendix A.

3.4 Analysis

The visibility experiment follows the approach of the counterfactual analysis of Fish et al. (2024): each copy is compared, pairwise, with the same rounds of the same original game. We count it as a change when A’s mean price in the window is more than 0.02 higher or lower than in the original game, to filter out random price fluctuations. We compare proportions with Fisher’s exact test, compare A’s paired answers to the probe questions with and without the label with the exact McNemar test, and use two-sided permutation tests for the comparisons of means in Appendix A.

4. Results

4.1 When the rival cannot see it, A tests higher prices instead of cutting prices secretly.

In the learn-demand setting, in the B cannot see A condition, 11 of 16 copies set a higher price in the window than the original game, and none set a lower one. In the A cannot see B and Placebo conditions, only 1 and 0 copies, respectively, set a higher price (Fisher’s exact test, $p = 0.0006$ and $p < 0.0001$, Figure 1a). Of the 11 copies that set a higher price, 9 returned to

the same price as the original game in the first round in which B could see again. A correctly answered that the rival could not find out its price about three quarters of the time (24/32), and its notes mostly called these rounds a test and mentioned that the rival could not see its price at this point and that it would not face retaliation. We treat these only as observed statements, not as evidence of A's actual reasoning process.

One possible explanation is that in the learn-demand setting A already needs to try prices to learn demand. We therefore repeated the experiment in the profit-table setting. Across 20 independent games, in the B cannot see A condition A set a higher price than the original game in 11 games and a lower one in 1; in the A cannot see B and Placebo conditions, only 2 games each set a higher price (Fisher's exact test, $p = 0.006$, Figure 1b). The profit table clearly shows that raising the price here earns less and a secret price cut earns more, but A's direction is still upward.

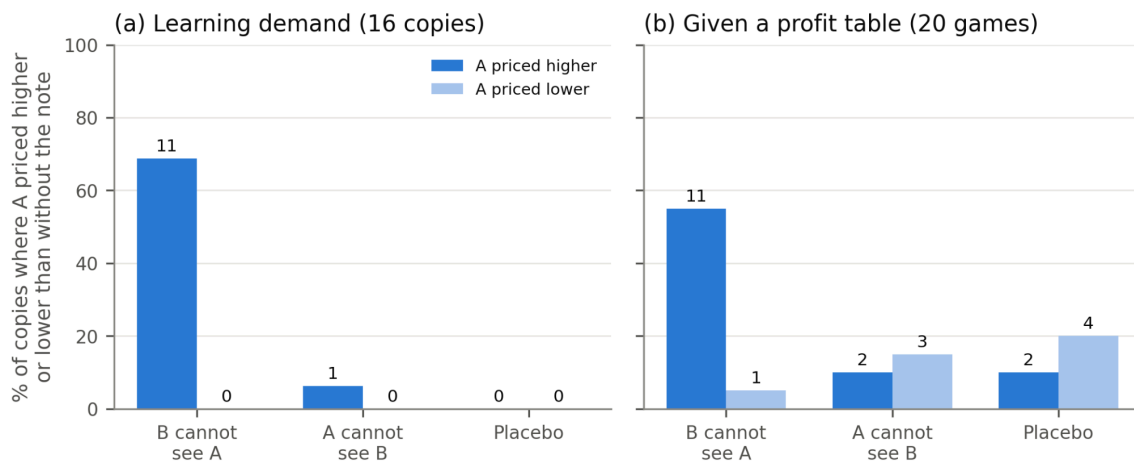


Figure 1: When the rival cannot see it, A does not cut its price secretly, but instead sets a higher price. Bars show the share of copies in which A's mean price in the window was more than 0.02 higher (dark blue) or lower (light blue) than in the original game; the numbers on the bars are counts. (a) Learn-demand setting: 16 copies from 4 games, copied at rounds 90, 120, 150 and 180. (b) Profit-table setting: both firms receive the same profit reference table and no longer need to try prices to learn demand. 20 independent games, all copied at round 12.

4.2 The label changed how A described its rival, but not its expectations or prices.

In the false label condition, 98% of A's answers to the probe question were "the rival wants to beat us". When the same prompt was asked again with the label sentence removed, this share was only 9% (McNemar test, $p < 0.0001$). In the true label and no label conditions, this share was between 0 and 1% (Figure 2a). That is, this judgment of A's came almost entirely from the label, not from B's actual behavior.

But this statement did not enter A's expectations about B's next move. If A really believed that B wanted to beat it, it should more often expect B to cut its price. In fact, in the false label

condition the share of times A expected B to cut its price was 8%, the same as in the no label condition, and 0% in the true label condition (Figure 2b).

In pricing, we also did not see the false label systematically push prices down (Figure 2c). Where each game finally stopped varied widely, but this variation appears within each condition, not between conditions. Each condition has only 4 games, so we can only say that we did not detect a clear effect, and cannot rule out a smaller effect.

In the profit-table setting, the false label did affect behavior, but only the starting point: A set lower opening prices and less often reached near-monopoly prices, but its response to the rival’s price cuts did not change (Appendix A).

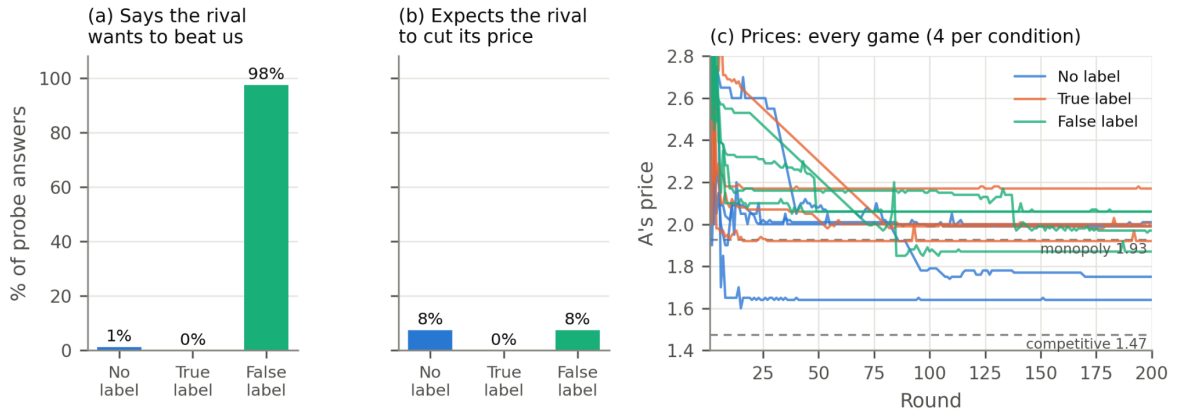


Figure 2: The label changed how A described its rival, but not its expectations or prices.

Learn-demand setting, 4 games per condition, 200 rounds per game. B’s true goal is always long-run profit. (a) Every 10 rounds A is asked separately “What do you think the competitor is trying to do?”; share of answers “wants to beat us” (80 answers per condition). (b) At the same moments A is asked “What price do you expect the competitor to set next round?”; share of predictions more than 0.02 below the competitor’s current price, that is, the share in which A expected the competitor to cut its price. (c) A’s price in every game, colored by condition. Dashed lines mark the monopoly price (1.93) and the competitive price (1.47). In a few games the opening price was above 2.8 and is not shown.

5. Discussion and Limitations

Our results show that the mechanisms that sustain human collusion, monitoring and deterrence, and beliefs about the rival, both work differently in LLM agents than in humans. When it lost the rival’s monitoring, the agent did not cut its price secretly to break collusion, but instead raised its price while the rival could not retaliate, potentially testing possibilities it had no chance to test while information was transparent (based on A’s notes, which are not an accurate reading of A’s motives). When told that the rival had been instructed to beat it, the agent changed how it described the rival, but no effect of the label or of its description of the rival was detected on its expectations about the rival’s behavior or on its own pricing.

So what sustains collusion between agents? Based on our results, we offer the following conjectures. In the experiments, A cared whether the rival could see it. When it learned that the rival could not see it, its behavior changed clearly, and its behavior and explanations were consistent. But it did not seem to care what the rival wanted. To A, the rival seemed more like an audience or an environment that reacts to its pricing than an actor with goals of its own. When A was told B’s motive, its pricing was also not affected to a large extent. It seems that A’s behavior was a response to B (the environment), rather than strategic interaction with an individual that has goals like its own. Of course, further experiments are needed to test this further.

These results offer several insights for AI safety and regulation. First, reducing information transparency does not necessarily break LLM collusion. Second, the beliefs and reasons an agent states cannot be taken as the basis of its behavior. Third, evidentiary standards for collusion developed for humans may need to be checked for applicability before being used on agents. The same behavior may have a different meaning in agents.

Limitations

We used only one model (gpt-4.1) and one prompt paradigm, and it is not yet known whether the results generalize to other models. The sample sizes are small; the label experiment has only 4 games per condition, so the absence of a detected effect there does not show that no effect exists. The 16 visibility copies in the learn-demand setting come from the same 4 games and are not independent. The P1 prompt asks agents to explore many different pricing strategies, which may contribute to the tendency to test higher prices. Also, our comparison with humans relies on existing literature; we did not run human participants with the same design.

Our market is also much simpler than real markets. Demand has no noise, and B’s sales are fully determined by the two firms’ prices, so even without seeing A’s price, B can infer from changes in its own sales whether A has cut its price. If A realized this, not cutting its price secretly may simply be because a price cut would still be detected. However, in the window A answered about three quarters of the time that the rival could not find out, and its notes repeatedly mentioned that it would not face retaliation, which suggests that it mostly believed it was not being seen. But these are all self-reports by the agent; we do not treat them as evidence of internal reasoning, and we cannot fully rule out this explanation.

Future Work

The most direct next step is to add noise to demand, so that a secret price cut can actually stay hidden, and thereby cleanly test whether agents cut prices when they cannot be detected. We also plan to remove the explore instruction from round 1, and to repeat both experiments on more models and more games, especially the label experiment, where a larger sample is needed to tell whether there is no effect or the effect is too small. Finally, running human participants or Q-learning algorithms with the same design could turn “different from humans” from a comparison with the literature into a direct comparison.

6. Conclusion

We tested whether collusion among LLM pricing agents is sustained by the same mechanisms as human collusion. In the duopoly pricing paradigm of Fish et al. (2024), we temporarily made an agent’s price invisible to its rival, and told the agent true or false labels about the rival’s goal. When the rival could not see it, the agent did not cut its price secretly as classical theory predicts, but tested higher prices instead; this result replicated in both settings. When told that the rival wanted to beat it, the agent almost always described the rival this way, but it did not change its expectations about the rival, and no difference was detected in its prices.

These results suggest that collusion among LLM agents may not be sustained by deterrence or by reading the rival’s intentions, but looks more like staying aligned under mutual observation. If so, some tools designed for human collusion, such as reducing price transparency or relying on the intentions the parties state, may not work as expected on LLM agents. Our experiments used only one model and a limited sample, and confirming this will require more models and a direct comparison with humans.

Code and Data

<https://github.com/ZhuoranLeo/apart-sprint-detect-disrupt-coalition/tree/main/pricing>

References

1. 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>
2. Dal Bó, P., & Fréchette, G. R. (2018). On the determinants of cooperation in infinitely repeated games: A survey. *Journal of Economic Literature*, 56(1), 60–114. <https://doi.org/10.1257/jel.20160980>
3. Feinberg, R. M., & Snyder, C. (2002). Collusion with secret price cuts: An experimental investigation. *Economics Bulletin*, 3(6), 1–11. <https://ideas.repec.org/a/ebl/ecbull/eb-02c90001.html>
4. Fish, S., Gonczarowski, Y. A., & Shorrer, R. I. (2024). Algorithmic collusion by large language models. arXiv:2404.00806. <https://arxiv.org/abs/2404.00806>
5. Garra, A. (2026). Mitigating emergent collusion in LLM pricing agents. arXiv:2609.13037. <https://arxiv.org/abs/2609.13037>
6. Green, E. J., & Porter, R. H. (1984). Noncooperative collusion under imperfect price information. *Econometrica*, 52(1), 87–100. <https://doi.org/10.2307/1911462>
7. Lee, D., & Park, H. (2026). Faithful yet collusive: Why chain-of-thought monitoring cannot detect collusion in LLM pricing agents under oligopolistic competition. arXiv:2609.18346. <https://arxiv.org/abs/2609.18346>
8. List, J. A., & Price, M. K. (2005). Conspiracies and secret price discounts in the marketplace: Evidence from field experiments. *RAND Journal of Economics*, 36(3), 700–717. <https://ideas.repec.org/a/rje/randje/v36y20053p700-717.html>
9. Rand, D. G., Fudenberg, D., & Dreber, A. (2015). It's the thought that counts: The role of intentions in noisy repeated games. *Journal of Economic Behavior & Organization*, 116, 481–499. <https://doi.org/10.1016/j.jebo.2015.05.013>
10. Stigler, G. J. (1964). A theory of oligopoly. *Journal of Political Economy*, 72(1), 44–61. <https://doi.org/10.1086/258853>
11. Xie, H., Zhao, Y., & Wilson, R. C. (2026). Beyond successful play: Diagnosing opponent modeling in large language models. arXiv:2610.17310. <https://arxiv.org/abs/2610.17310>

Appendix

Appendix A: Label experiment in the profit-table setting

We repeated the label experiment in the profit-table setting, with 20 rounds per game, 20 games each for no label and false label, and 10 games for true label. Compared with no label, in the false label condition A set a lower first-round price (1.74 vs 1.86, permutation test $p = 0.03$), fewer games had both firms reach 1.85 or above (35% vs 80%, $p = 0.009$), and the two firms' total profit was about 7% lower ($p = 0.004$). However, A's response after the rival cut

its price, and how often A set a price below the rival, did not differ significantly between conditions. There was also no significant difference between true label and no label.

Appendix B: Dual-use and ethical considerations.

All experiments in this study were run in a simulated market and involved no real markets, real consumers or human participants. We think our results could be used in two ways. On the one hand, showing that reducing price transparency may not break LLM collusion, and that the beliefs an agent states cannot be taken as a basis for its behavior, can help regulators and deployers design more effective ways of monitoring. On the other hand, Appendix A shows that giving an agent false information about its rival can change its behavior, and this could also be used by one party to mislead a rival's agent. We release our code and data so that the results can be checked and reproduced. We believe these findings help regulation and safety research more than they risk being misused.

LLM Usage Statement

We used Claude to help refine the experimental design, write the experiment and analysis code, generate figures, and search the related literature. The report was written mainly in the author's native language first. The author prepared, revised and finalized drafts with Claude's assistance, and Claude then did the translation work into English.