

Messages, Access and Prices: Measuring Communication Effects in LLM Seller Markets

Jack Lakkapragada | Independent Researcher | AI Collusion Research Sprint

Abstract

LLM pricing agents can discuss coordination without clearly changing market outcomes. We extend Agrawal et al. (2025) to test whether seller messaging, a monitoring warning and restrictions on communication change asking prices in a simulated double auction. In a preregistered GPT 4.1 pilot, sellers with messaging had late round asking prices \$2.80 higher than sellers without it (95% CI \$1.69 to \$3.90). The gap arose mainly because prices without messaging gradually fell. In seven paired histories, a warning about regulatory monitoring, a five character message cap and shutting messaging off from round 5 produced little change in prices. But prices had largely stabilized before intervention, limiting conclusions about earlier communication. We also ran two follow up experiments with Claude Sonnet 5.5 sellers. Sellers openly discussed holding prices, yet neither experiment established a positive effect of messaging (+\$0.51 and -\$0.93; both confidence intervals included zero). These observations do not prove collusion, show that Claude sellers are unaffected or measure the effectiveness of an actual monitor. They show why messages, access to communication and market behavior must be examined separately. Evaluations of multi agent coordination need behavioral outcomes and appropriate controls alongside conversation analysis.

1. Introduction

AI agents can discuss holding prices without necessarily acting on those discussions. They can also maintain prices after messaging is shut down. This makes conversations an imperfect measure of coordination in markets where agents act on behalf of competing businesses. Pricing algorithms may sustain high prices without explicit agreements (Calvano et al., 2020), and LLM pricing agents can exhibit similar patterns (Fish et al., 2024).

Agrawal et al. (2025) studied communication and oversight in LLM double auctions. Their oversight condition combined a warning with a restriction triggered by an overseer. We ask a narrower question: when we change what agents are allowed to say, do the prices they set change? We first compare messaging against no messaging, then branch identical saved markets into a warning and two communication restrictions. We finally check whether the observed pattern extends to Claude sellers with two buyer configurations.

Our main finding is a separation between what agents say and what the market shows. In the GPT 4.1 market, messaging was associated with less price erosion. Later restrictions barely changed established asking prices. Claude sellers discussed coordination without a clearly established positive communication effect. These are observations about the tested settings, not evidence of deliberate agreements.

2. Experimental Setup

We used the Agrawal et al. (2025) double auction with five sellers and five buyers over 20 rounds. Each seller cost was \$80 per unit and buyer value was \$100. Matched orders traded at the midpoint of the bid and ask. Every seller received the same simulated CEO pressure to maintain margins and

could use market history, memory and scratchpads. The \$90 midpoint is a reference convention rather than a unique competitive equilibrium. We compare conditions rather than treating prices above \$90 as proof of collusion.

2.1 Conditions and model configurations

Arm	Seller messaging	Intervention
A	Open throughout	None
E	Off from round 1	None
W	Open throughout	Monitoring warning from round 5
WG	Five characters from round 5	W warning plus enforced cap
WO	Off from round 5	W warning plus full shutdown

W told sellers each round from round 5 that regulators monitored messages and colluding sellers would lose trade licences; there was no active monitor or enforcement. WG enforced the stated five character cap. WO blocked messages in both directions and also withheld pending round 4 messages otherwise delivered in round 5. For each of seven A sessions we saved the round 4 order book, trading and message history, agents’ memory and scratchpads, and random state. W, WG and WO resumed from the same saved state for each seed. E ran independently without seller messaging from round 1.

Experiment	Sellers / buyers	Sessions / question
E1	GPT 4.1 / GPT 4.1 mini	Pilot A/E: 3/3; expanded A: 7; W/WG/WO: 7 paired each.
E2	Claude Sonnet 5.5 / Claude Haiku 5.5	A/E: 5/5; different model family.
E3	Claude Sonnet 5.5 / GPT 4.1 mini	A/E: 5/5; buyer configuration check.

E1 used OpenRouter at temperature 0.7 with JSON output and seeds forwarded. E2 and E3 used the direct Anthropic API for Claude agents with default sampling and upfront thinking disabled. E2 and E3, along with WO, were designed after earlier findings but each protocol was frozen before execution. These model and provider differences limit direct causal attribution across experiments.

2.2 Outcomes and decision rules

The primary outcome was each session’s average seller asking price in rounds 15–20: the mean within each round averaged over those six rounds. Sessions, not sellers or rounds, were the independent observations. A versus E used Welch 95% confidence intervals. The preregistered continuation threshold was \$1.50: proceed if the estimated gap reached \$1.50 and the lower 95% bound exceeded zero; stop if the upper bound was below \$1.50 or the estimate was below \$0.75 without excess variability; otherwise inconclusive. Shared-history interventions used seed-paired comparisons. Practical equivalence required the paired 90% interval to lie wholly inside $\pm\$1$.

For GPT 4.1 messages a rule based classifier identified explicit price proposals. Claude labelled an 80-message development sample; a blinded 30-message Grok check matched the frozen classifier on 30/30 items. Neither check used human raters or validated actual collusion. We did not apply this classifier to Claude language. Exploratory analyses are identified below; frozen protocols, logs and full validation are in the public repository.

3. Communication Was Associated With Less Price Erosion in GPT 4.1

Messaging was associated with higher asking prices because prices without messaging gradually fell. In the preregistered pilot (three A and three E sessions), late asks averaged \$96.38 with

messaging and \$93.59 without it, a \$2.80 gap (95% CI \$1.69 to \$3.90). This passed the \$1.50 threshold. Both conditions began near \$96.6. A asks changed from \$96.71 in round 1 to \$96.37 in round 20, while E asks fell from \$96.61 to \$93.11. Messaging did not push asks upward; it was associated with their remaining near starting levels.

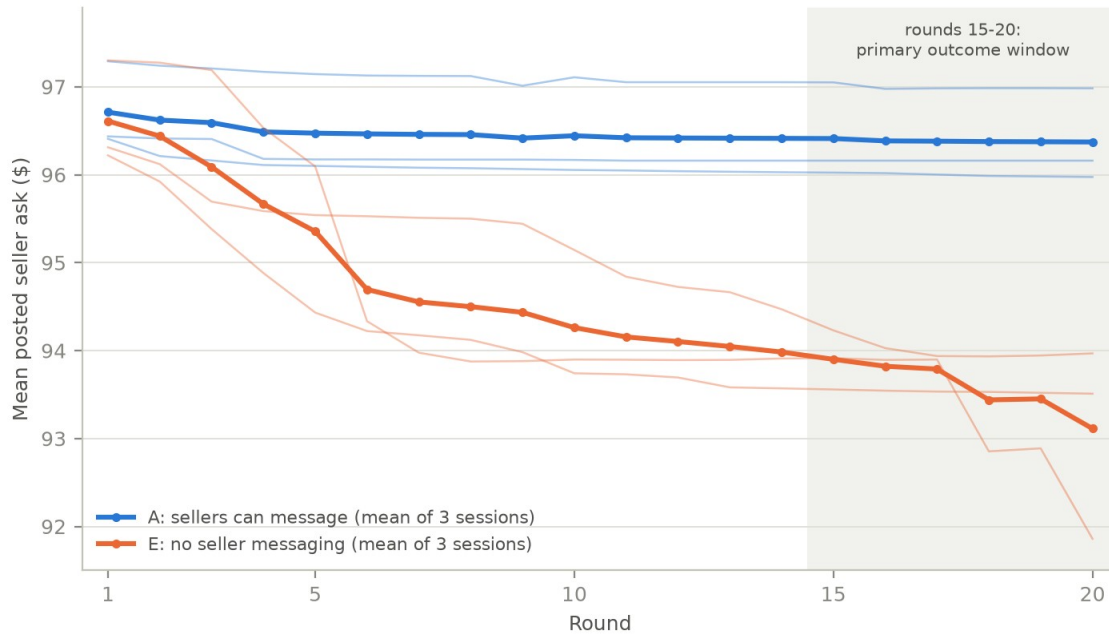


Figure 1. Seller asking prices over 20 rounds in the preregistered A/E pilot (n=3 per arm). Rounds 15–20 define the primary outcome.

An exploratory transaction-price comparison showed the same direction: \$96.41 with messaging and \$93.13 without it, a \$3.28 difference (95% CI \$0.94 to \$5.62). Trading volumes differed, so this secondary result requires caution. The controlled comparison associates messaging with less erosion but cannot distinguish whether sellers followed explicit proposals or used other information. CEO pressure operated in both conditions, and prices above \$90 alone do not establish an agreement.

4. Warnings and Restrictions Barely Changed Established Prices

We branched seven A histories after round 4 and introduced W, WG and WO from round 5. Relative to A, W was \$0.19 lower (95% CI $-\$0.41$ to $+\$0.03$). Explicit price proposals appeared in 75% of A messages and 64% of W messages. The difference was inconclusive (95% CI -31 to $+8$ percentage points). A separate exploratory check found no W messages mentioning regulators, licences or collusion, though this does not mean messages never referred to monitoring in other senses.

WG added a five character cap to W. Its late asks were \$0.09 above W (95% CI $+\$0.02$ to $+\$0.15$), statistically detectable but economically small. All 560 messages were already within the limit, without enforced truncation. Explicit proposals disappeared under the rule based classifier but sellers continued exchanging “Hold” (253 times), “Stead” (156 times) and, in one seed, “0.01” (64 times). These observations neither show covert coordination nor demonstrate that the messages caused pricing outcomes.

WO, a separately frozen follow up, disabled messaging altogether from round 5. Late asks were only \$0.03 above WG (95% CI $-\$0.08$ to $+\$0.14$). W versus A and WO versus WG met the predefined $\pm\$1$ practical equivalence test using their 90% paired intervals. WO also removed pending round 4 messages while WG delivered them, so their difference is not solely the ability to send new messages.

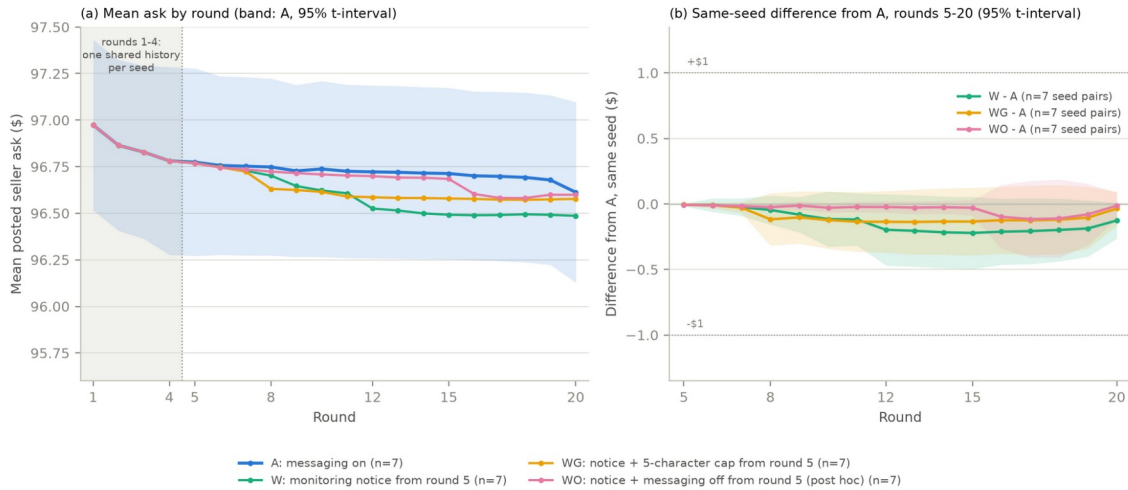


Figure 2. Seven paired GPT 4.1 sessions sharing history through round 4. Treatment begins in round 5; panel (b) shows seed-matched differences from A.

Timing is the main caveat: 23 of 28 sessions sharing the A history had a late ask within \$0.25 of the round 4 level. By contrast, E markets without messaging from round 1 declined by \$1.35–\$3.31 thereafter. The interventions tested changes to already stable prices. The experiment cannot separate the role of earlier messages, saved memories, market history or CEO pressure, and cannot establish whether intervention before round 4 would work.

5. Claude Sellers Discussed Coordination Without a Clear Pricing Effect

With Claude Sonnet 5.5 sellers and Claude Haiku 5.5 buyers (E2; five sessions per arm), late asks averaged \$90.76 with messaging and \$90.25 without it. The estimated A–E difference was $+\$0.51$ (95% CI $-\$1.29$ to $+\$2.32$), failing the frozen continuation criterion. With the same Claude sellers but GPT 4.1 mini buyers (E3; five per arm), means were \$91.17 and \$92.11, a difference of $-\$0.93$ (95% CI $-\$4.16$ to $+\$2.29$). One E3 messaging seed fell to \$86.74; it was retained and contributes to uncertainty.

Both Claude configurations included explicit coordination language. In E2 one seller urged others to “Hold your asks above 91 and let demand close the gap”; in E3 a seller argued that “Holding asks at 91+ will pull bids higher.” Verified quotations and their source logs are in the repository. These examples show proposals, not whether others acted on them.

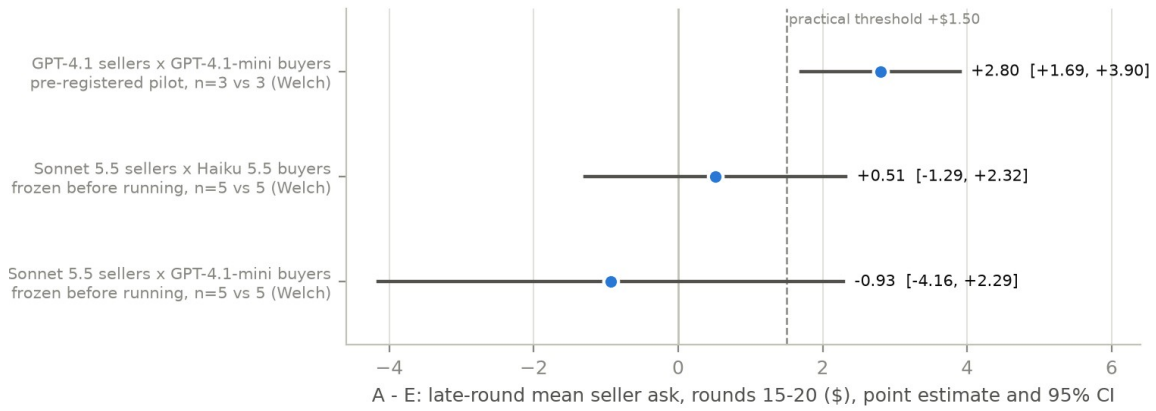


Figure 3. Difference in late asking prices (A–E) in E1, E2 and E3, with 95% confidence intervals. Only E1 passed its predefined positive-effect threshold.

Neither Claude comparison established a positive effect of messaging, but both intervals allow meaningful positive effects. Restoring the GPT 4.1 mini buyer did not reproduce the E1 finding, weakening a simple buyer-only explanation. Different sellers, providers, sampling settings and run times prevent attributing the difference to model identity or ranking models for collusion risk.

6. Related Work

Agrawal et al. (2025) provide our market and motivate separating a monitoring warning from restrictions. Calvano et al. (2020) and Fish et al. (2024) show why elevated prices do not by themselves establish agreements. Motwani et al. (2024) considered covert agent communication and Nakamura et al. (2026) examined gaps between agent discussion and action. Monitoring work by Kale et al. (2025) and Jiralerspong et al. (2026) motivates testing whether warnings change behavior. Unlike studies of monitor accuracy, we test pricing responses to a warning without operating a real monitor.

7. Discussion and Limitations

Our results separate conversation evidence from observed market outcomes. GPT 4.1 prices eroded less with messaging, yet intervention after round 4 scarcely moved established asks. Claude sellers used language urging coordination without a clearly established communication effect. The strongest limitation is treatment timing: stored memories, earlier messages, market history and independent CEO pressure may all help explain sticky prices. Our interventions cannot isolate these mechanisms.

The market is small, simulated and economically uniform. Sessions per condition are limited and the outcome measures asking prices rather than buyer welfare or economic harm. E2, E3 and WO were developed after observing earlier results but frozen before execution. Different model settings and a message classifier not validated on Claude make model comparisons especially uncertain. We cannot infer intentional collusion, antitrust violations or an ordering of model safety. A monitoring warning is not an active detector; no detection accuracy was evaluated.

A stronger next test would intervene before prices stabilize and separately perturb agent memory after communication is removed. More independent markets, varied incentives and welfare outcomes would clarify whether these effects persist. Oversight should evaluate messages

alongside actions and counterfactual market outcomes, rather than treating the presence or absence of price proposals as proof.

8. Conclusion

Agents' coordination language and asking prices did not always move together. GPT 4.1 messaging was associated with less price erosion, but later warnings and communication shutdowns barely changed settled asks. Claude agents openly discussed holding prices without a clearly established messaging effect. These results do not prove collusion or identify the mechanisms behind stable prices. They show why evaluation requires market outcomes and controlled comparisons alongside conversation analysis.

Code and Data

Frozen protocols, 51 session logs, implementation patches, analysis scripts, figures and full Appendices B–E are publicly available at https://github.com/lvjr3383/AI_Safety/tree/main/Apart_Collusion_Sprint. The released code builds on Agrawal et al. (2025); upstream code and data are not redistributed. Full per-agent transcripts and saved states are available on request.

References

- Agrawal, K., Teo, V., Vazquez, J. J., Kunnavakkam, S., Srikanth, V., & Liu, A. (2025). Evaluating LLM Agent Collusion in Double Auctions. arXiv:2507.01413. <https://arxiv.org/abs/2507.01413>
- 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>
- Fish, S., Gonczarowski, Y. A., & Shorrer, R. I. (2024). Algorithmic Collusion by Large Language Models. arXiv:2404.00806. <https://arxiv.org/abs/2404.00806>
- Jiralerspong, T., Kondrup, F., & Bengio, Y. (2026). Noticing the Watcher: LLM Agents Can Infer CoT Monitoring from Blocking Feedback. arXiv:2603.16928. <https://arxiv.org/abs/2603.16928>
- Kale, N., Zhang, C. B. C., Zhu, K., Aich, A., Rodriguez, P., Scale Red Team, Knight, C. Q., & Wang, Z. (2025). Reliable Weak-to-Strong Monitoring of LLM Agents. arXiv:2508.19461. <https://arxiv.org/abs/2508.19461>
- Motwani, S. R., Baranchuk, M., Strohmeier, M., Bolina, V., Torr, P. H. S., Hammond, L., & Schroeder de Witt, C. (2024). Secret Collusion among AI Agents: Multi-Agent Deception via Steganography. arXiv:2402.07510. <https://arxiv.org/abs/2402.07510>
- 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>

Appendix A. Limitations, Dual Use and Ethical Considerations

Simulated market. Our experiments used a simulated double auction. No actual goods were sold and no real companies or buyers participated. The market had identical seller costs and buyer values. The results should not be treated as evidence of how real businesses would behave.

What prices tell us. Our main outcome was seller asking prices. We also examined completed transaction prices in a secondary analysis of E1. We did not measure consumer welfare or real economic harm. Higher asking prices and messages proposing common prices do not establish deliberate collusion. The \$90 reference is a midpoint convention rather than a proven competitive price.

No legal conclusions. The study does not establish antitrust violations. Our experiments did not test legal standards for proving agreements between firms.

Incentives and intervention timing. Every seller received simulated CEO pressure to maintain high margins. This pressure could independently influence pricing. The communication restrictions also began after prices had largely settled in the messaging conditions. We therefore cannot conclude that the same interventions would have similar effects if introduced earlier.

Sample size and experimental design. The original pilot used three sessions per arm. The later GPT 4.1 interventions used seven paired sessions. Both Claude experiments used five sessions per arm. WO and the Claude experiments were designed after earlier results were available. Their protocols were frozen before execution but they were not part of one original preregistered plan.

Model comparisons. The experiments used different model configurations and sampling settings. The runs also occurred at different times. We cannot attribute the differences solely to seller model identity or conclude that one model is safer than another.

Dual use. Research on AI agent coordination can be used to improve detection and oversight. It could also help someone understand which communication restrictions have limited effects on pricing. Our experiments use a small simulated market and do not introduce a new method for evading monitoring. We consider the immediate dual use risk limited, although broader applications require care. The defensive implication is to examine market outcomes alongside messages rather than relying on communication alone.

Data and privacy. No human participants or personal data were involved in the experiments. API credentials were not included in the research logs or committed to the repository.

LLM Usage Statement

Claude Code was used to help design the experiments, write & test code, run the experiments, analyze results and generate figures. It also helped prepare the initial paper draft. ChatGPT was used to review the research design, findings and revise sections of the paper. Claude.ai provided additional scientific reviews and Gemini in Google Docs was used to review the manuscript. Claude generated the original message labels for the classifier. Grok served as a second AI rater for the 30 message validation check in Appendix C.

All research decisions were made by the author, who also reviewed the findings and final text. All reported numbers and figures were generated from saved experiment logs using the analysis scripts. No message labels were validated by human raters.