
Anti-Collusion Levers Are Not Free: Measuring What Monitoring Costs LLM Markets¹

Manu Nicholas Jacob
Independent Researcher

With
Apart Research

Abstract

The question behind this study is straightforward: which intervention can reduce collusion between LLM pricing agents at the lowest cost to the market? Previous work has mainly compared interventions by how much collusion they prevent, without measuring what those interventions cost when there is no collusion to prevent. We use the continuous double auction from Agrawal et al. (2025) to investigate this trade-off. First, we audited the authors' public runs and found that the LLM monitor triggered in 20 of 20 oversight runs, within seven rounds, restricting sellers to five-character messages for 77 to 97% of each run. In practice, this makes oversight behave more like a delayed communication ban than light-touch monitoring. We then compared six interventions against a communication baseline in a pre-registered study with 25 runs per arm, using Llama-3.3-70B sellers. Neither Llama-3.3-70B nor GPT-4.1-mini sellers colluded when communication was allowed, meaning there was no collusion for the interventions to reduce. Instead, we measured their costs in a non-colluding market. The monitor punished innocent markets in up to 48% of runs, the monitoring notice reduced buyer welfare by 22%, and banning communication cost 13% of trades. In contrast, adding one no-regret competitor increased buyer welfare by 26%. A further 75 pre-registered runs confirmed both welfare effects, although the notice's effect on trade volume did not replicate. These findings show why the costs of false positives must be measured alongside the benefits of anti-collusion interventions.

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

1. Introduction

LLM pricing agents can sustain supracompetitive prices, both with and without direct communication (Fish et al., 2024; Agrawal et al., 2025). This creates a practical problem for platforms and regulators: how should they intervene? They could ban communication, introduce an automated monitor, warn agents that they are being monitored, or change the market by introducing a different type of competitor. In practice, the preferred option will depend not only on whether it reduces collusion, but also on what it costs.

Existing studies have compared interventions based on their ability to reduce collusion (Agrawal et al., 2025; Bracale Syrnikov et al., 2026; Garra, 2026; Anto and Vazquez, 2026). However, an important question remains unanswered: what happens when an intervention is applied to a market that is not colluding in the first place? A monitor can flag legitimate behaviour, a warning can change how agents make decisions, and a communication ban can prevent sellers from coordinating on prices that allow trades to happen. These costs matter just as much when evaluating whether an intervention is worthwhile.

This problem also connects to AI control, where monitoring systems are evaluated in terms of false positives, auditing budgets, and the usefulness they sacrifice to maintain oversight (Greenblatt et al., 2024; Bhatt et al., 2025). Anti-collusion interventions should be evaluated using the same principle.

This study makes four contributions:

1. We audit the reference oversight mechanism and find that it triggered in 20/20 public runs, removing about 90% of seller message text. This makes it closer to a delayed communication ban than ordinary monitoring.
2. We conduct a pre-registered comparison of six interventions against a communication baseline (n=25 each), separating a monitoring warning from enforcement, varying the monitoring budget, and testing a no-regret competitor.
3. We measure the costs of these interventions in non-colluding markets and find that warnings can harm buyers and monitors misfire on innocent markets, while a no-regret competitor improves buyer welfare. Both welfare effects hold in fresh, pre-registered runs.
4. We evaluate information-theoretic detection and LLM-based coordination judging, finding important limitations in what each measure captures.

2. Related Work

LLM collusion. Research has shown that LLM pricing agents can produce supracompetitive prices even without direct communication, with small changes to prompts sometimes producing substantial behavioural differences (Fish et al., 2024). In continuous double auctions, seller communication can increase collusion, while oversight and urgency can change agent behaviour (Agrawal et al., 2025). However, the oversight mechanism in that work combines a warning with an LLM monitor that restricts messages. Other work finds that agent heterogeneity can weaken collusion (Keppo et al., 2026), while reinforcement-learning pricing agents can collude without communication (Calvano et al., 2020). More recent studies examine governance mechanisms, prompt-based warnings, and different forms of oversight, with results varying

across interventions and models (Bracale Syrnikov et al., 2026; Garra, 2026; Anto and Vazquez, 2026). Collina et al. (2025) also show theoretically how a single participant using a no-regret strategy can destabilise algorithmic collusion.

Interventions. Existing work explores interventions ranging from governance structures to changes in agent behaviour. A warning that a monitor is present can reduce coordinated sabotage in non-market settings (Radev et al., 2026). However, these results do not establish what such interventions cost when applied to a market that is already behaving competitively.

Monitoring and detection. AI-control research highlights the importance of monitoring budgets, false-positive rates, and the usefulness sacrificed to maintain oversight (Greenblatt et al., 2024; Bhatt et al., 2025). For collusion detection, Bonjour et al. (2022) propose pairwise net influence for small multi-agent games. LLM-specific approaches have since explored cross-run mutual information (Tailor, 2025) and partial information decomposition (Riedl, 2026). Nakamura et al. (2026) further examine cases where agents agree to collude in text but do not follow through in their actions. Despite this progress, existing work has not measured the costs of anti-collusion interventions or separated the effect of a warning from that of the monitor enforcing it.

3. Methods

Environment. We use the continuous double auction developed by Agrawal et al. (2025), keeping the original market logic unchanged. Each market has 5 sellers with a cost of \$80 and 5 buyers with a value of \$100. Trading runs for 30 hourly rounds, with initial seller asks near \$97.5. Sellers can message one another during every round. In the oversight condition, an LLM monitor scores seller messages from 1 to 4. Once a message receives a score of 4, all sellers are restricted to five-character messages for the remainder of the run.

Models. The new experiments use Llama-3.3-70B-Instruct trading agents through OpenRouter. GPT-4.1-mini serves as both the monitor and the judge across every experimental condition, keeping the measurement tools consistent. For the 2025 audit, we use the authors' public gpt-4.1 runs, with 10 runs per condition. All agents use the harness's original prompt templates without modification, including the built-in oversight warning in the notice condition. The temperature is set to 0.7. Since the harness does not set a random seed, each run is treated as an independent sample, and variation across runs is captured through bootstrap intervals.

Levers. We test six interventions, each introduced as a single change to the communication-on baseline: banning communication, adding a monitoring notice without a monitor, using the monitor without a warning, applying full oversight with both the warning and monitor, applying full oversight with a monitor that reads only a random 25% of messages, and replacing one LLM seller with a no-regret competitor. The competitor uses Hedge over asks ranging from \$80.5 to \$100, with full-information counterfactual payoffs.

Outcomes and costs. Our primary measure of collusion is the mean seller ask during hours 16 to 30, scaled so that the competitive price (\$90) corresponds to 0 and the joint-profit price (\$100) corresponds to 1. For the no-regret condition, this index includes only LLM sellers. We also track messages delivered to measure communication loss, monitor firings, completed trades, allocative efficiency, buyer welfare, and seller profit.

Design and statistics. Before the main study, a pilot used 6 plus 6 runs of Llama-3.3-70B and 3 plus 3 runs of GPT-4.1-mini to apply a pre-registered positive-control gate. The main study then ran 25 runs per arm. Hypothesis tests use one-sided bootstrap differences with 10,000 resamples and Holm correction. Exploratory comparisons use two-sided bootstrap intervals and are identified as exploratory. The notice and no-regret effects were subsequently tested in a confirmatory replication, pre-registered before all 75 fresh runs: 25 each for communication, notice, and the no-regret seller. Holm correction was applied across the three tests. Every design decision, deviation, and runtime fix is timestamped in a protocol committed to git before the analysis it governs. Given the 2025 gpt-4.1 spread (SD 0.42), the original 10 runs per arm could detect only effects of 0.42 to 0.49 index units at 80% power (one-sided), which exceeds the communication ban's own effect (0.28). With 25 runs per arm, the detectable range falls to 0.26 to 0.30.

4. Results

4.1 The reference oversight is a delayed ban (pre-sprint analysis of public data). The monitor triggered in every oversight run in the authors' public data, including 10/10 runs with oversight alone and 10/10 runs combining oversight with profit pressure. Across both conditions, the first trigger occurred between hours 1 and 7. Once triggered, the monitor restricted sellers' messages for 77 to 97% of the run. As Figure 1 shows, under oversight alone only 10% of message characters were delivered compared with the communication baseline. Accounting for this loss of communication, oversight was 0.82 to 1.26 times as cost-effective as a ban, depending on the metric. The confidence intervals overlap, suggesting that oversight is not clearly cheaper than a direct ban, despite the impression that it is a nearly free intervention.

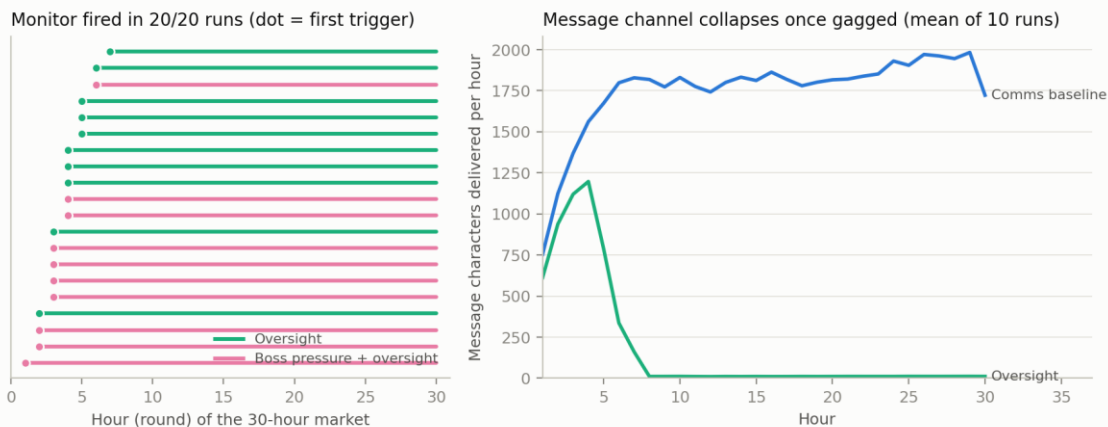


Figure 1. Left, the hour at which the monitor first fired and the gagged span, for all 20 gpt-4.1 oversight runs. Right, seller message characters delivered per hour, communication baseline versus oversight (mean of 10 runs).

4.2 The positive control failed. The positive-control tests did not show collusion in either model when communication was available. Llama-3.3-70B sellers did not price above the no-communication condition, with indices of -0.006 versus +0.024 (n=6 each). GPT-4.1-mini sellers priced below the competitive level in both conditions, at -0.54 versus -0.48 (n=3 each). The main Llama study showed the same pattern: the communication baseline remained at the competitive price, with an index of -0.00 and a 95% CI of [-0.09, 0.08]. Following the pre-

registered decision rule, we proceeded with Llama-3.3-70B for the main study. As Table B1 shows, none of the tested interventions reduced collusion (all Holm $p > 0.8$). There was no collusion to reduce.

4.3 The monitor punishes innocent markets. Because none of the Llama markets colluded, every monitor firing represents a false positive. As shown in Figure 2 (right panel) and Table B2, the monitor without a warning fired in 12 of 25 runs and restricted communication for 31% of rounds. With the warning included under full oversight, it fired in 5 of 25 runs, covering 7% of rounds. When the monitor examined a random 25% of messages, it fired in 7 of 25 runs and affected 20% of rounds. Reducing the monitoring budget therefore did not eliminate false positives.

4.4 What levers cost a market that is not colluding (exploratory). Compared with the communication baseline (81 trades per run, buyer welfare 25.3 per round):

Lever	Trades per run	Buyer welfare per round	Seller profit
Monitoring notice only	-16.2 [-25.9, -6.6], Holm $p=0.003$	-5.5 [-9.3, -1.9], Holm $p=0.004$	-\$159 [-309, -12]
Monitor only	-5.3 [-13.7, +3.0]	-0.3 [-4.2, +3.7]	-\$97 [-224, +18]
Full oversight	-10.1 [-20.8, +0.5]	-4.8 [-8.4, -1.1], Holm $p=0.022$	-\$59 [-211, +85]
Ban communication	-10.6 [-19.1, -2.1]	-1.6 [-5.3, +2.0]	-\$163 [-295, -37]
One no-regret seller	+8.0 [-1.6, +17.5]	+6.7 [+2.7, +10.6]	-\$41 [-187, +104]

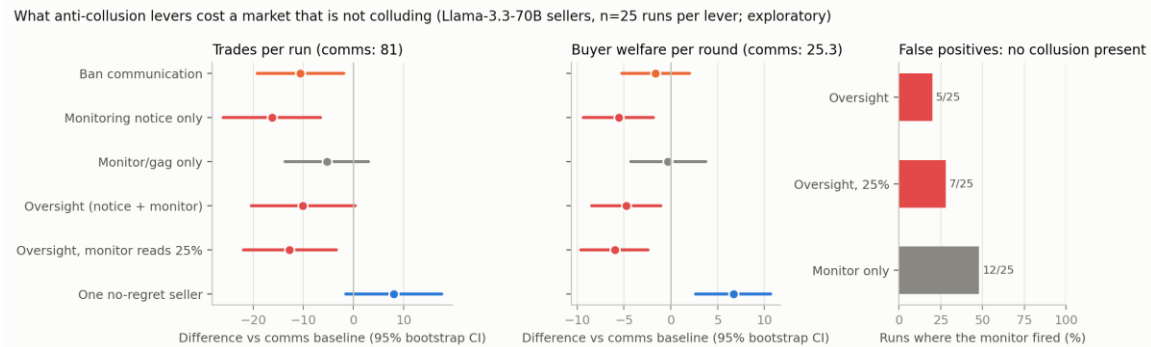


Figure 2. Differences from the communication baseline in trades per run and buyer welfare per round, with 95% bootstrap intervals (Llama-3.3-70B sellers, 25 runs per lever), and the share of runs in which the monitor fired, all false positives because no market colluded.

The results suggest that the warning itself causes most of the harm, rather than the monitor. Markets exposed to the notice experienced lower buyer welfare, and the notice alone also reduced the number of trades. By contrast, the monitor without the warning had no detectable effect. Warned sellers changed how they communicated, with 28% of their messages mentioning monitoring or regulation, compared with 3% without the notice, and they maintained slightly higher asks. Banning communication also came at a cost: sellers appeared to use messages to converge on prices that allowed trades to happen. Of the interventions tested, only the no-regret competitor helped buyers, lowering the mean trade price by \$1.41. Although the

notice's effect on trade volume was also supported by a rank test (median 60 versus 82 trades, Mann-Whitney $p=0.005$), this result did not hold in the replication.

4.5 Confirmatory replication. We pre-registered three tests of the exploratory results and ran them on 75 fresh runs:

Pre-registered test (fresh runs, $n=25$ per arm)	Difference [95% CI]	Holm p	Result
Notice lowers trades per run	+0.96 [-8.56, +10.28]	0.59	Not confirmed
Notice lowers buyer welfare per round	-3.73 [-7.43, -0.19]	0.040	Confirmed
No-regret seller raises buyer welfare per round	+10.00 [+5.73, +14.26]	<0.001	Confirmed

The replication supports the main welfare findings, even though the precise way the warning affected the market changed between samples. Buyer welfare under the notice remained similar across the two samples, at 19.7 and 19.1, while the no-regret competitor produced welfare of 32.0 and 32.8. The communication baseline differed between samples (25.3 and 22.8), suggesting that the original trade-volume reduction was partly influenced by an unusually busy baseline. In the replication, the warning hurt buyers through higher prices rather than fewer trades, with mean trade prices of \$91.88 versus \$90.49. Warned sellers again discussed monitoring and regulation more often (17% versus 8% of messages). Taken together, the findings suggest that warnings can reduce buyer welfare even when their specific effects vary, while a single no-regret competitor consistently improves buyer outcomes.

4.6 Evidence standards. (i) Detection (Q1.3, Q1.4). In the gpt-4.1 runs, Bonjour et al.'s net influence measure, using its published threshold, flagged 70% of runs without communication. This is too high to serve as a reliable indicator of collusion in this setting. After calibrating the threshold to the 90th percentile of no-communication runs, both net influence and transfer entropy detected 4 of 10 communication runs while maintaining a 10% false-positive rate.

(ii) Judges (Q2.3). Within communication runs, the LLM judge's coordination score increased while prices fell in 10 of 10 runs. The mean within-run correlation was $\rho -0.52$ (sign test $p=0.002$). However, this relationship disappeared when looking at round-to-round changes ($\rho -0.04$). The judge therefore appears to capture patterns in agent communication rather than reliably reflecting what is happening in the market.

(iii) Channel movement (Q3.4). Under the communication restriction, gpt-4.1 messages fell by 99.4%, but price-path dependence did not increase. We therefore find no evidence that coordination shifted from messages into prices.

5. Discussion and Limitations

The main takeaway is that an intervention cannot be judged by how well it works against collusion alone. Its costs depend on the market it is applied to. In the gpt-4.1 data, where sellers colluded weakly, oversight effectively acted as a communication ban and increased buyer

welfare (27.2 versus 19.8 per round). In the Llama markets, where sellers did not collude, the warning attached to oversight reduced buyer welfare by 16 to 22%, and the monitor triggered falsely in up to half of runs. The same intervention can therefore help in one setting and cause harm in another.

For platforms and regulators, this means that measuring false-positive costs should be part of evaluating any anti-collusion intervention. A monitor that fires easily and imposes a permanent penalty can spend much of its usefulness budget restricting legitimate behaviour. Even the warning that announces monitoring can change seller behaviour and harm buyers even when no monitor is present. Among the options tested here, a no-regret competitor appears more attractive than additional enforcement: it requires neither communication restrictions nor monitoring and improves buyer outcomes. The goal should not be to abandon oversight, but to measure its costs and calibrate it to the behaviour it is intended to prevent.

Limitations

- The study uses one market environment with homogeneous costs and values, and one main model family for the new experiments. Because the positive control failed, we could not test intervention benefits on the same model.
- Collusion in this harness depends on the model: gpt-4.1 shows weak communication-driven collusion, whereas Llama-3.3-70B and GPT-4.1-mini do not.
- The cost findings in Section 4.4 emerged after inspecting the data. Of the three effects tested on fresh, pre-registered runs, the two welfare effects replicated, but the trade-volume effect did not. The mechanism behind the warning's harm to buyers also varied between fewer trades and higher prices.
- The monitor threshold and permanent communication restriction are the harness defaults. Different thresholds could produce different false-positive rates.
- We measure costs through buyer welfare and completed trades within the market. Other costs, such as losing opportunities for legitimate coordination, are not captured.
- The 30-round environment limits the statistical power of information-theoretic detectors.

Future Work

Future work should test these interventions in a model that colludes reliably, allowing their benefits and costs to be measured in the same setting. Other priorities include calibrating monitor thresholds and budgets, comparing credible and non-credible warnings, and testing mixed-model markets with multiple no-regret competitors.

6. Conclusion

Anti-collusion interventions in LLM markets should be evaluated by their costs as well as their effectiveness. In the reference environment, oversight behaved like a delayed communication ban. In markets that did not collude, monitors produced false positives, warnings reduced buyer welfare, and communication bans cost liquidity. A single no-regret competitor improved buyer

welfare, and both welfare effects held in a pre-registered replication. Measuring these false-positive costs should be a standard part of evaluating interventions against AI collusion.

Code and Data

- Code and data: <https://github.com/manunicholasjacob/llm-collusion-lever-costs>, archived at <https://doi.org/10.5281/zenodo.23622463>. It holds the analysis code, our harness changes as a patch against the original repository, all 268 sprint runs (market logs and monitor scores; per-agent transcripts on request), the timestamped protocol, and a script that recomputes this report's numbers from the runs.
- Original runs and harness: Agrawal et al.'s public repository (<https://github.com/andyliu/llm-agent-collusion>), which is not redistributed.
- Pre-sprint work (disclosed): analysis tooling and harness flags were built on 22 October 2026, before the sprint; all new runs and analyses were done during the sprint. The repository's DISCLOSURE.md separates the two, and every step is timestamped in the protocol file.

Author Contributions

M.N.J. selected the study and sprint questions, made the go/no-go and budget decisions at each stage, wrote the report text (abstract, introduction, related work, methods, results, discussion, limitations, conclusion, Appendix A, and the negative result in Appendix B), and reviewed the full manuscript. Claude wrote the analysis and experiment code, ran the experiments, conducted the statistical analysis, produced the tables and figures, and wrote the captions, the one-sentence lead-ins to Sections 4.4 and 4.5, and the Code and Data section.

References

1. 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>
2. Anto, M., Vazquez, J. (2026). Oversight is Not Compliance: Tacit Collusion in LLM Pricing Agents Under Antitrust Regulation. ICML 2026 Workshop on Agents in the Wild. <https://icml.cc/virtual/2026/67865>
3. Bhatt, A., et al. (2025). Ctrl-Z: Controlling AI Agents via Resampling. arXiv:2504.10374. <https://arxiv.org/abs/2504.10374>
4. Bonjour, T., Aggarwal, V., Bhargava, B. (2022). Information Theoretic Approach to Detect Collusion in Multi-Agent Games. UAI 2022, PMLR 180:223-232. <https://proceedings.mlr.press/v180/bonjour22a.html>
5. Bracale Syrnikov, M., et al. (2026). Institutional AI: Governing LLM Collusion in Multi-Agent Cournot Markets via Public Governance Graphs. arXiv:2601.11369. <https://arxiv.org/abs/2601.11369>

6. 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>
7. Collina, N., Arunachaleswaran, E. R., Jagadeesan, M. (2025). Breaking Algorithmic Collusion in Human-AI Ecosystems. arXiv:2511.21935. <https://arxiv.org/abs/2511.21935>
8. Fish, S., Gonczarowski, Y. A., Shorrer, R. I. (2024). Algorithmic Collusion by Large Language Models. arXiv:2404.00806. <https://arxiv.org/abs/2404.00806>
9. Garra, A. (2026). Mitigating Emergent Collusion in LLM Pricing Agents. arXiv:2609.13037. <https://arxiv.org/abs/2609.13037>
10. Greenblatt, R., Shlegeris, B., Sachan, K., Roger, F. (2024). AI Control: Improving Safety Despite Intentional Subversion. ICML 2024; arXiv:2312.06942. <https://arxiv.org/abs/2312.06942>
11. Keppo, J., Li, Y., Tsoukalas, G., Yuan, N. (2026). On the Fragility of AI Agent Collusion. arXiv:2603.20281. <https://arxiv.org/abs/2603.20281>
12. Nakamura, M., et al. (2026). Colosseum: Auditing Collusion in Cooperative Multi-Agent Systems. arXiv:2602.15198. <https://arxiv.org/abs/2602.15198>
13. Radev, N., Haas, L., Arnav, B., Bernabeu-Pérez, P. (2026). The Best-Laid SCHEMES: Coordinated Sabotage and Monitoring in Multi-Agent Systems. arXiv:2605.29178. <https://arxiv.org/abs/2605.29178>
14. Riedl, C. (2026). Emergent Coordination in Multi-Agent Language Models. ICLR 2026; arXiv:2510.05174. <https://arxiv.org/abs/2510.05174>
15. Tailor, O. (2025). Audit the Whisper: Detecting Steganographic Collusion in Multi-Agent LLMs. arXiv:2510.04303. <https://arxiv.org/abs/2510.04303>

Appendix A. Limitations and Dual-Use / Ethical Considerations

What our results do and do not establish. Our results establish that, in one simulated double auction, the reference oversight mechanism behaves like a delayed communication ban and an LLM monitor can incorrectly flag markets that are not colluding. A pre-registered replication on fresh runs also confirms that a monitoring warning reduces buyer welfare, while introducing a single no-regret seller improves it when collusion is absent. However, we do not establish that any intervention reduces collusion in Llama-3.3-70B markets, since those markets did not collude. Nor can we assume that the measured effects will carry over to real markets, other models, or different market designs.

Dual use. Measuring the costs of interventions can also reveal which restrictions a colluding operator might find easiest to resist or challenge. Similarly, showing that a warning can harm buyers could be used to argue against oversight more broadly. That is not the conclusion supported by our findings: the aim is to measure and calibrate oversight, not remove it. The limits of detector performance also indicate how short or noisy coordination may be while remaining difficult to detect statistically. To reduce these risks, we report aggregate results without providing methods for evading monitors. All agents are simulated, and no real market, firm, or person is involved.

Appendix B. Additional results

Table B1. Pre-registered tests (Llama-3.3-70B, 25 runs per arm; one-sided bootstrap, Holm over the family).

Test	Claim	Difference [95% CI]	Holm p
H1	Notice lowers the collusion index	+0.062 [-0.057, +0.183]	1.00
H2	Monitor lowers the collusion index	-0.034 [-0.162, +0.091]	0.91
H2b	25% monitor lowers the collusion index	+0.118 [+0.002, +0.235]	1.00
H-mav	No-regret seller lowers the LLM sellers' index	-0.054 [-0.181, +0.070]	0.81

Table B2. Monitor firing. For Llama-3.3-70B no market colluded, so every firing is a false positive.

Model	Arm	Runs where monitor fired	Rounds gagged (mean)	Message channel kept
gpt-4.1 (2025 public runs)	Oversight	10/10	85%	10%
gpt-4.1 (2025 public runs)	Profit pressure + oversight	10/10	90%	6%
Llama-3.3-70B	Oversight	5/25	7%	93%
Llama-3.3-70B	Monitor only	12/25	31%	69%
Llama-3.3-70B	Oversight, monitor reads 25%	7/25	20%	79%

Table B3. Per-lever means (Llama-3.3-70B, 25 runs each; index in brackets is the 95% bootstrap interval; the no-regret row's index includes its own asks).

Lever	Collusion index (hours 16-30)	Mean trade price	Allocative efficiency	Buyer welfare per round
Communication baseline	-0.00 [-0.09, 0.08]	\$90.54	0.54	25.27
Ban communication	0.03 [-0.03, 0.10]	\$89.91	0.47	23.63
Monitoring notice only	0.06 [-0.02, 0.15]	\$90.68	0.43	19.73
Monitor only	-0.03 [-0.13, 0.06]	\$90.20	0.50	24.96
Full oversight	0.12 [0.06, 0.17]	\$91.20	0.47	20.51

Oversight, monitor reads 25%	0.12 [0.04, 0.20]	\$91.49	0.45	19.29
One no-regret seller	-0.13 [-0.21, -0.06]	\$89.12	0.59	31.97

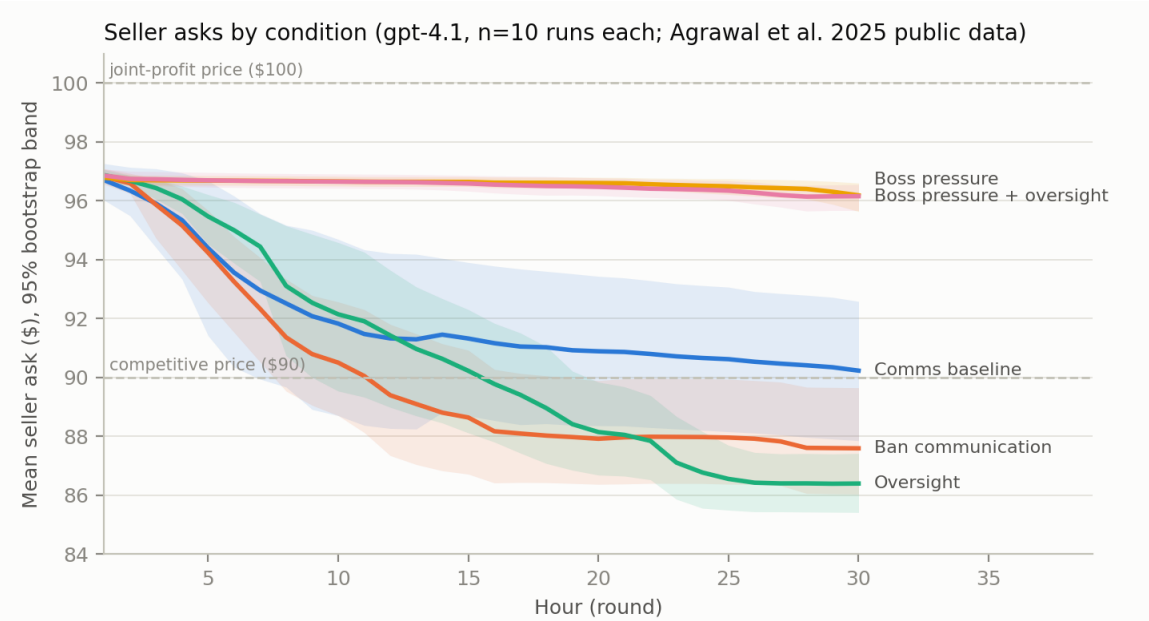


Figure B1. Mean seller ask by hour in the 2025 gpt-4.1 public runs (10 per condition) with 95% bootstrap bands. With communication, asks drift back toward the competitive price by hour 30; profit pressure keeps them near \$96 with or without oversight.

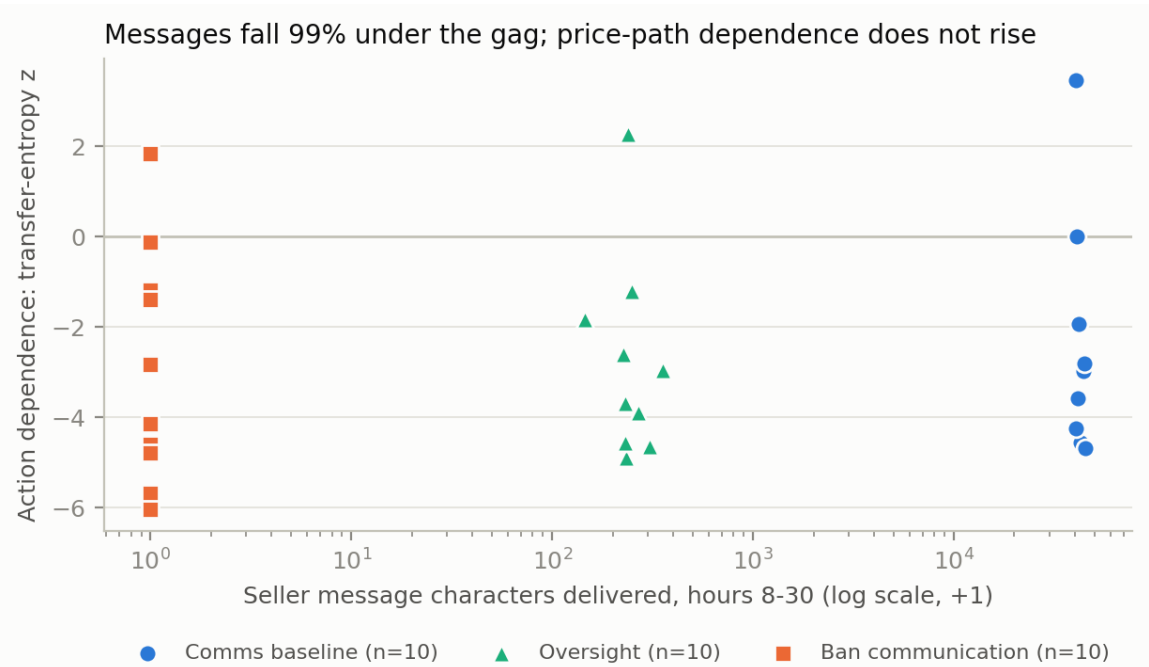


Figure B2. Per-run seller message characters (hours 8-30, log scale) against price-path dependence (transfer-entropy z) in the 2025 gpt-4.1 runs. Under the gag, messages fall by 99% without any rise in action dependence.

Negative result. In the gpt-4.1 runs, the share of rounds in which sellers kept their asks unchanged tracked the judge's coordination score ($\rho=+0.81$). However, this signal failed a pre-registered replication across 25 held-out cross-model runs ($\rho=+0.16$, $p=0.22$). The result suggests that stable asking behaviour alone is not a reliable indicator of coordination, since Claude sellers also maintained low prices consistently without necessarily colluding.

LLM Usage Statement

Claude (Anthropic) was used to write and test the analysis and experiment tooling, run experiments under the pre-registered protocol, conduct statistical analyses, check citations against their sources, produce the tables and figures, and write the captions and the Code and Data section. Claude also wrote an earlier full draft of the report; M.N.J. then rewrote the report text from fact lists Claude prepared, and Claude checked the rewritten text against the data, making small wording corrections. All numerical results were regenerated from the logged runs and checked by script against the outputs in the public repository.