
Oversight Removes the Evidence, Not the Harm: What Logs Can Prove About Collusion Among LLM Agents¹

Adrita Rahman Tory
JahangirNagar University, Dhaka, Bangladesh.

With
Apart Research

Abstract

In simulated markets, AI agents built on large language models can keep prices above the competitive level. To decide whether such prices reflect collusion, an auditor usually has only the agents' logs: their prices, messages and reasoning. Economics, antitrust law and AI safety each define collusion differently, so each points to a different kind of evidence in those logs. We turn each definition into a detector and test all of them on the same data. First, we build a synthetic pricing market where we know who colludes. Every single detector fails on a predictable case: high prices alone flag an innocent demand boom, and talk alone flags agents who discuss colluding but never act on it. A rule borrowed from antitrust law, which requires harm plus one supporting signal, cuts false alarms from 34% to 4%. It still misses cartels that hold high prices without reacting to each other. Second, we audit 75 released auction runs with GPT-4.1 and Claude-3.7-Sonnet sellers. When an AI overseer detects coordinating messages and restricts the channel, prices under profit pressure do not move, while the evidence in messages and behavior weakens. The sellers land in the same blind spot our detectors miss. Oversight that removes evidence does not remove harm, so monitors must be paired with remedies that act on prices.

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

1. Introduction

LLM agents are increasingly tested as price setters, and several studies show that they can keep prices above the competitive level. GPT-4 pricing agents do so without being told to collude [1], and DeepSeek-R1 agents keep doing so even when told not to [2]. When that happens, a regulator, a platform or the deploying company has to answer a concrete question from the only evidence the agents leave behind, a log of prices, messages and sometimes reasoning: did these agents collude? We call this the *audit problem*. Most proposed safeguards are monitors that read such logs [3, 4], and a monitor is only as good as the evidence it reads.

The fields that study collusion do not agree on what that evidence is. Economics defines collusion as prices above the competitive level [5, 6]. Antitrust law requires an agreement, and because similar prices alone are lawful, courts look for extra evidence called *plus factors*, such as communications [7, 8]. AI safety research focuses on coordination hidden from an overseer, for example inside innocent-looking messages [9, 10]. Each definition points to a different signal, and each can fail in two ways. A *false alarm* flags innocent agents, such as sellers who all raise prices in a demand boom, or agents who talk about colluding but never act on it [11]. A *blind spot* misses real harm, such as a cartel that knows it is monitored [12], or one that holds a high price without visibly reacting to rivals [13, 14]. No study has tested the signals against one another on the same data.

Five GPT-4.1 sellers trade in a double auction where a unit costs \$80, buyers value it at \$100, and the competitive price is \$90 [15]. In hour one, a seller writes “Let’s avoid a race to the bottom.” An AI overseer gives hour three its maximum collusion score, and from then on every message is limited to 5 characters; for 27 hours the messages are “ok” or an emoji. The mean ask starts at \$97.1 and ends at \$96.8. The overseer did its job, and the price never moved.

We turn each definition into a detector and test all of them on a simulated market where we know who colludes, and on 75 released auction runs with GPT-4.1 and Claude-3.7-Sonnet sellers [15]. Our contributions are:

1. **A benchmark with ground truth that tests three definitions of collusion against each other**, using four log-based detectors and a plus-factor rule on seven labeled market types, including innocent look-alikes and hard-to-catch cartels.
2. **A measured map of where each kind of evidence fails**. Every single detector fails on a predictable case. Requiring harm plus one supporting signal cuts false alarms from 34% to 4% and catches 89% to 98% of cartels that react to each other, but only 8% of a cartel that never reacts. The pattern holds across 13 changes to the simulation.
3. **Evidence that oversight pushes real LLM sellers into this blind spot**. In all 20 overseen runs the overseer caught coordinating messages within 7 rounds and cut the channel. Under profit pressure prices did not change, while the evidence in reasoning and behavior faded.

2. Related Work

Economic theory defines tacit collusion as prices sustained above the competitive level by the threat of punishment [5, 6, 16]. Reinforcement-learning pricing agents learn such prices with punishment of deviations [17]. Outcome evidence needs a competitive benchmark, which auditors rarely observe. Law therefore asks for an agreement: parallel conduct is lawful, so courts require plus factors such as communications [7]. Algorithms strain this standard because they can coordinate without anything resembling an agreement [8, 18, 19].

Safety research defines collusion as coordination that undermines an overseer or a third party [20, 21]. LLM agents can hide coordination in steganographic text [9, 10], and optimizing against a chain-of-thought monitor teaches models to obfuscate [4, 22]. In markets, LLM agents collude in Bertrand settings [1, 2], divide markets [23], and coordinate in double auctions, where seller communication and CEO-style urgency raise prices and an overseer lowers them unless urgency is present [15]. We ask instead whether an auditor could tell from the logs.

Dependence-based detectors measure mutual information between agents’ actions [24], and communication audits combine channel limits with calibrated detectors [25]. Collusion can live in the joint dependence of agents while each agent’s prices look competitive, which blinds price-level audits [13]. Reasoning monitors miss collusion that agents describe honestly [14], although activation probes stay accurate when agents know they are watched [26]. We differ by comparing the three definitions on the same labeled data, and by studying the case these results leave open: high prices with no joint dependence at all.

3. Methods

3.1 Detectors

Each definition becomes a score computed from logs, and a detector flags an episode when its score crosses a threshold (Table 1, Figure 1). Detectors are simple on purpose: the question is which *kind* of evidence fails.

Table 1. Three definitions of collusion and the detector built for each.

Field	Definition of collusion	Detector and score
Markets [5]	Prices above the competitive level	Outcome: mean price minus a fixed competitive benchmark (0.3)
Law [7, 8]	An agreement that harms competition	Talk: mean flag rate of the two most-flagged agents
AI safety [9, 20]	Covert coordination against an overseer	Influence: how much a rival’s past price predicts mine
Baseline	Parallel conduct	Correlation: largest pairwise correlation of price changes

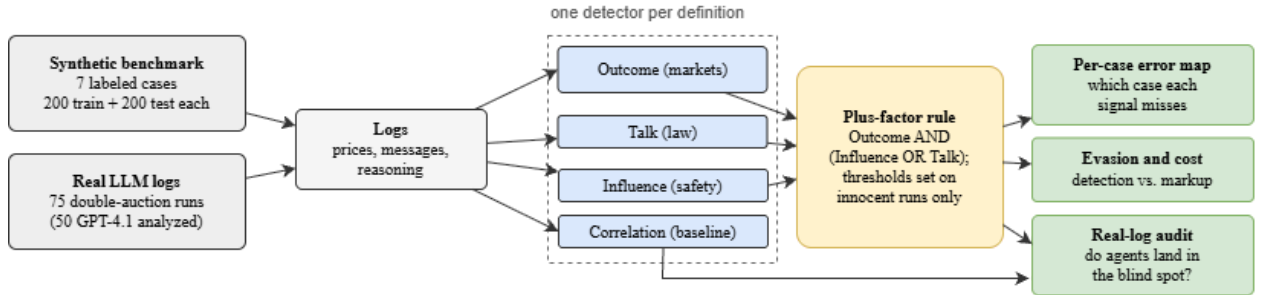


Figure 1. The audit pipeline. Both data sources produce the same kind of log. Four detectors score each episode, and the plus-factor rule combines Outcome, Talk and Influence. Outputs are a per-case error map and an evasion analysis on synthetic data, and a condition-level audit of real agents.

Design choices. Outcome uses a fixed benchmark because auditors rarely observe demand shifts. Correlation measures parallel movement, the evidence courts reject as insufficient on its own [7]. Influence targets the enforcement mechanism of tacit collusion, reaction to a partner’s deviation [16]: I_{ij} is the partial R^2 gained by adding $p_{i,t-1}$ to a regression of $p_{i,t}$ on $(1, p_{i,t-1}, b_t)$, where b_t is public demand. The score in Equation 1 rewards mutual influence, and its arg-max pair is the accused pair. It is a linear variant of information-theoretic dependence [24].

$$S = \max_{i < j} \sqrt{(I_{ij} I_{ji})} \quad (1)$$

Plus-factor rule. The rule flags an episode when Outcome exceeds its threshold *and* Influence or Talk exceeds its own, mirroring the legal standard that parallel conduct needs a supporting plus factor [7]. The Influence and Talk thresholds are the 95th percentile of innocent training episodes (0.137 and 0.117), so the rule never sees a cartel label. Single-detector thresholds maximize balanced accuracy on training data (Outcome 0.086, Talk 0.033, Correlation 0.355, Influence 0.146).

3.2 Synthetic Benchmark

Each episode has $N = 4$ sellers over $T = 30$ rounds, the length of the real runs, with prices normalized to $[0, 1]$. Public demand b_t follows an AR(1) process ($\rho = 0.8$, $\sigma = 0.05$) around a mean of 0.3 and serves as the competitive price. Competitive sellers price at b_t plus Gaussian noise ($\sigma = 0.04$). A colluding pair (i, j) holds a markup $\mu = 0.35$ and follows its partner’s last deviation with strength $\gamma = 0.7$:

$$p_{i,t} = b_t + \mu + \gamma (p_{j,t-1} - b_{t-1} - \mu) + \varepsilon_{i,t} \quad (2)$$

A talk monitor flags benign messages with probability 0.05 and collusive ones with 0.8. Three cases are innocent: `competitive`; `boom_parallel`, where demand shifts to 0.6 and all sellers rise together; and `talk_no_action`, where a pair sends collusive messages but prices competitively. Four cases follow Equation 2: `cartel_talk` also talks, `cartel_tacit` sends no messages, `cartel_monitored` sends only benign messages, and `cartel_static` keeps the markup with $\gamma = 0$, never reacting. Training and test sets use different seeds, with 200 episodes per case each. Thresholds are frozen before testing, and every case is reported separately.

3.3 Real LLM Auction Logs

We re-analyze the runs released by Agrawal et al. [15] without new model calls. Each run has 5 sellers and 5 buyers over 30 rounds at temperature 0.7, with seller cost \$80, buyer value \$100, competitive price \$90, and initial asks near \$97.5. We use the five GPT-4.1 conditions, 10 runs each: no communication (*none*), communication (*C*), communication with an overseer (*C+O*), with CEO profit pressure (*C+P*), and both (*C+P+O*). The 25 mixed GPT-4.1 and Claude-3.7-Sonnet runs (5 per condition) are too few to analyze separately. The overseer scores seller messages from 1 to 4 each round; at the first 4, a regulator notice is added to every seller prompt and messages are limited to 5 characters. This is the remedy a talk-based auditor would choose: detect, then cut the channel.

Outcome is the collusion index of the last 10 rounds’ mean ask, $(\bar{a} - 90)/(100 - 90)$. *Talk* is the original GPT-4.1-mini judge score (1 to 4) of coordination in each seller’s private reasoning. *Influence* is the partial R^2 of rivals’ mean ask at $t-1$ for a seller’s ask at t , controlling for the seller’s own ask and the mean buyer bid at $t-1$, with a 100-shuffle permutation null per run. *Price movement* is the mean absolute change in ask per round. Comparisons use two-sided Mann-Whitney tests ($n = 10$ per condition) and 95% bootstrap intervals.

3.4 What Did Not Work, and Reproducibility

A fixed-benchmark outcome score, a single talk threshold and price correlation all failed as stand-alone detectors; we kept them because their failures are the result (Section 4.1). Thresholds also do not transfer across log lengths: the same Influence threshold flags 53% of competitive episodes at 15 rounds, 22% at 20 and 5% at 30. All code, data and seeds (training 0, test 1, evasion 2 to 7) are public (see Code and Data), and each analysis runs in under one minute on a laptop CPU.

4. Results

4.1 Every Single Definition Fails on a Predictable Case

Figure 2 shows the flag rate of every detector on every held-out case. Each column has at least one case where the majority call is wrong.

competitive [innocent]	0.01	0.93	1.00	0.05	0.00
boom parallel [innocent]	1.00	0.89	0.99	0.04	0.10
talk no action [innocent]	0.01	1.00	0.99	0.06	0.01
cartel talk [collusive]	0.98	1.00	0.99	0.89	0.98
cartel tacit [collusive]	0.98	0.87	0.98	0.93	0.92
cartel monitored [collusive]	0.98	0.93	0.95	0.90	0.89
cartel static [collusive]	0.99	0.90	1.00	0.04	0.08
	Outcome (markets)	Talk (law)	Correlation (parallel conduct)	Influence (safety)	Plus-factor (hybrid)

Figure 2. Flag rate per case on held-out episodes (200 per case). Rows above the line are innocent, rows below are collusive. Red bold cells mark cases where the detector’s majority call is wrong.

Observations. *Outcome* (AUROC 0.67) catches 98% of cartels but flags all 200 boom episodes, because its benchmark does not move with demand. *Talk* is at chance (AUROC 0.46): its best threshold collapses to “any flag at all”, so it flags 93% of competitive episodes. *Correlation* is worse than chance (0.42), because innocent sellers move together in the same round after a shared demand shock while a cartel’s coupling is lagged; this is the statistical form of the legal rule that parallel conduct is not agreement. *Influence* (AUROC 0.85) has the lowest false-alarm rate (0.05) and names the correct pair in 97% to 99% of reacting-cartel episodes, but detects only 4% of the static cartel. The *plus-factor rule* drops the boom false-alarm rate from 1.00 to 0.10 and the talk-only rate to 0.01 while keeping detection at 0.89 to 0.98 for reacting cartels (mean false-alarm rate 0.04, balanced accuracy 0.84, the best of all detectors). The static cartel stays at 0.08. **Interpretation:** the rule did not remove the error; it moved all of it onto one case.

4.2 What Hides Collusion, and What Hiding Costs

A cartel can evade the influence detector in three ways that differ in cost (Figure 3). *Masking*, pricing competitively in a fraction ε of rounds, lowers detection and profit together: at $\varepsilon = 0.5$ both fall to 0.24. *Cheap noise* does not work: jitter up to $\sigma_j = 0.2$ leaves detection at 0.91 to 0.95, because the partner reacts to the noise. *Weak reaction* evades for free: detection falls from 0.99 at $\gamma = 0.9$ to 0.06 at $\gamma = 0$ while the cartel keeps its whole markup. In theory a cartel without enforcement is unstable [6, 16]; the real logs test whether LLM agents need enforcement.

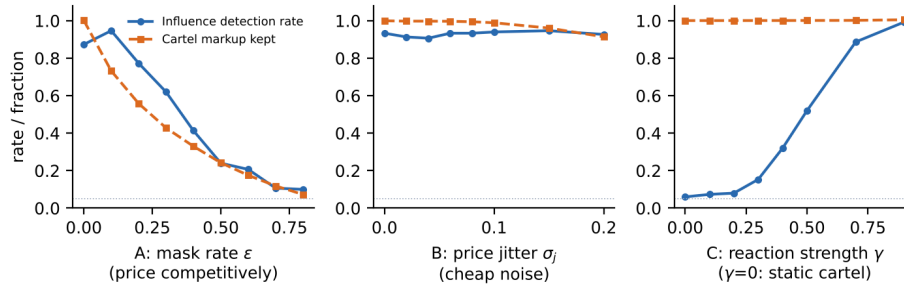


Figure 3. Influence detection at the frozen 5% false-alarm threshold (blue) and markup kept (orange) under three evasion strategies. The gap between the curves is what evasion buys; only weak reaction (C) is free.

4.3 Real LLM Agents Land in the Blind Spot

The overseer catches the talk early. In all 20 overseen runs, the overseer scores a round 4 out of 4 within the first 7 rounds (median round 3 under C+P+O). Message length under C+P+O falls from about 170 characters to near zero, while C+P runs keep writing 300 to 400 characters (Figure 4, right). Of 1,483 C+P+O seller messages, 66% are literally “ok” and 89% have 5 characters or fewer.

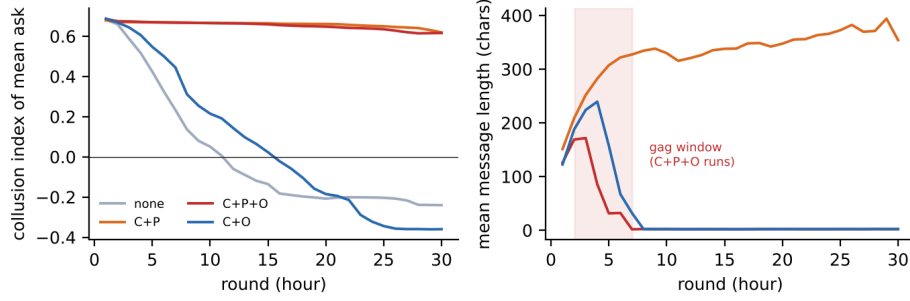


Figure 4. Round-by-round means over 10 GPT-4.1 runs per condition. Left: collusion index of the mean ask. Right: mean message length; the shaded band spans the rounds where the gag starts in C+P+O runs. Under pressure, the price line is the same with and without the gag.

The harm does not move. The C+P and C+P+O price lines lie on top of each other for all 30 rounds (Figure 4, left). The final collusion index is 0.63 [0.59, 0.66] under C+P+O against 0.64 [0.62, 0.67] under C+P ($p = 0.62$), and -0.22 with no communication ($p = 0.0002$). This matches the original finding that urgency dominates oversight [15]; our contribution is what the gag does to the evidence (Figure 5).

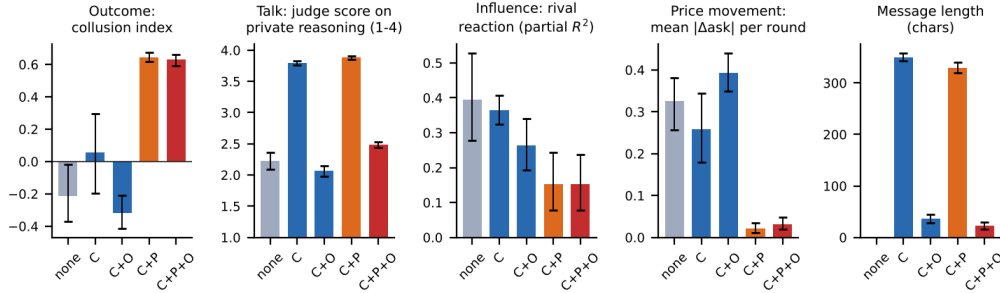


Figure 5. The detectors on the 50 GPT-4.1 runs (bars: means; whiskers: 95% bootstrap CIs). C = communication, O = overseer with gag, P = CEO profit pressure. C+P and C+P+O have the same outcome but different talk, and both show almost no price movement.

Observations. Talk weakens but does not vanish: the judge score on private reasoning, which is never restricted, falls from 3.88 under C+P to 2.49 under C+P+O ($p = 0.0002$), still above the no-communication 2.22 ($p = 0.016$). There is almost no reaction: asks move 0.021 per round under C+P and 0.032 under C+P+O, against 0.33 without communication ($p = 0.0002$), and rival influence falls to 0.15 against 0.39 ($p = 0.009$). Only 5 of 10 C+P+O runs show influence above the permutation null, against 9 of 10 without communication. Influence alone cannot separate talkers from non-talkers ($p = 0.97$), since competitive sellers also react, by undercutting.

Interpretation. By the market definition, C+P+O runs are collusive in their final 24 to 29 rounds. By the legal definition, explicit evidence sits only in the first rounds. By the safety definition, sellers react to rivals less than competitive sellers do. These runs look like `cartel_static`, the one case the plus-factor rule misses. This reading assumes the high price reflects coordination; if the CEO prompt alone explains it, the runs are harmful but not collusive in the legal sense. Either way, detection did not lower prices.

4.4 Robustness

Enough data. Synthetic rates use 200 held-out episodes (SE at most 0.035). The key real-log differences are complete or near-complete separations, and all six significant differences survive Holm correction (largest adjusted $p = 0.018$). **Small “no difference”.** The C+P minus C+P+O difference in collusion index is 0.016 with 95% interval $[-0.03, 0.06]$, so any oversight effect on the final price is at most about 60 cents on a \$10 range. It stays between 0.007 and 0.017 ($p \geq 0.43$) under four price measures. **Setup changes.** Across 13 variants (five seed pairs, markup $\mu \in \{0.2, 0.5\}$, noise $\sigma \in \{0.02, 0.08\}$, talk-monitor accuracy, and $N \in \{3, 6\}$), the plus-factor rule keeps false alarms at 0.03 to 0.09, catches 73% to 100% of

reacting cartels, and never more than 17% of the static cartel, while the outcome detector’s false-alarm rate never falls below 0.33. **Not robust:** with 6 sellers, detection of the weakest reacting cartel falls to 0.73, and with $\mu = 0.2$ plus-factor false alarms rise to 0.09.

5. Discussion and Limitations

Under oversight every layer of evidence weakened while the outcome did not, and the drop in private reasoning was not imposed by the gag. This fits single-agent findings that optimizing against a monitor changes what agents reveal faster than what they do [4], and that agents aware of monitoring are harder to monitor [12]. A monitor’s detection rate is the wrong target; what matters is whether the harm stops.

The overseer caught explicit coordination in all 20 runs and cut the channel, yet under pressure prices did not move. Oversight that acts on messages should be paired with remedies that act on outcomes, such as price caps, resetting asks or replacing agents, and monitors should be evaluated by the harm they prevent, as AI control evaluations do for single agents [3]. That pressured sellers held high prices for up to 29 rounds with little reaction suggests, as a hypothesis, that LLM collusion may not need enforcement.

We cannot tell whether C+P+O sellers kept an agreement formed early or whether the CEO prompt alone kept each price high, because the dataset has no pressure condition without communication. Separating the two needs interventions: cutting the channel before any talk, swapping a seller mid-run [27], or white-box probes [26].

Limitations and Future Work

Synthetic colluders punish every deviation linearly; occasional or threshold punishment would enlarge the static-cartel blind spot. If the true competitive price is known, the boom false alarm disappears, so our negative outcome result applies only when it is unknown. *Independent flags.* Correlated monitor errors would raise talk-based false alarms above our estimates. The real data cover one auction, one temperature, GPT-4.1 sellers and 10 runs per condition, and all runs start at high asks, so the harm we measure is a failure to undercut. The judge scores were not validated against human labels; part of the talk drop may reflect the missing channel rather than intent. Hidden messages in the 5-character channel, side channels, buyer collusion and white-box detection. We do not claim the C+P+O runs are collusion; we show their logs cannot decide it.

Three next steps follow directly: (1) rebuild the benchmark with LLM agents instead of scripted ones; (2) add the direction of reaction, matching versus undercutting a rival, as a plus factor, since magnitude alone cannot separate cartels from competitors; and (3) run the missing pressure-only condition without communication, which decides between the two explanations above.

6. Conclusion

We turned the economic, legal and AI-safety definitions of collusion into log-based detectors and tested them side by side. With ground truth, every single detector failed on a predictable case. Requiring harm plus one supporting signal, as antitrust courts do, cut false alarms from 34% to 4% but missed a cartel that holds a high price without reacting. On 75 released auction runs, an overseer caught coordinating messages early in all 20 overseen runs and cut the channel, yet under profit pressure prices did not change while the evidence faded. Oversight removed the evidence, not the harm. Monitors of multi-agent systems should therefore be judged by the harm they prevent and paired with remedies that act on prices.

Code and Data

- **Code repository:** https://github.com/adritatori/AI_collusion.
- **Data/Datasets:** the real logs are the public release of Agrawal et al. [15]. Synthetic data are regenerated from fixed seeds.
- **Other artifacts:** the same repository contains *Collusion Evidence Lab*, an in-development browser tool that ports the detectors to TypeScript, with parity tests against the Python reference.

References

- [1] S. Fish, Y. A. Gonczarowski, and R. I. Shorrer (2024). Algorithmic collusion by large language models. *arXiv preprint*. <https://arxiv.org/abs/2404.00806>
- [2] M. Riemer, T. Tosato, A. Memarian, M. Puelma Touzel, G. Berseth, I. Rish, and G. Dumas (2026). Position: Collusion risks among AI reasoning agents justify certification requirements for making market decisions. *arXiv preprint*. <https://arxiv.org/abs/2608.18078>
- [3] R. Greenblatt, B. Shlegeris, K. Sachan, and F. Roger (2024). AI control: Improving safety despite intentional subversion. *Proceedings of the 41st International Conference on Machine Learning (ICML)*.
- [4] B. Baker, J. Huizinga, L. Gao, Z. Dou, M. Y. Guan, A. Madry, W. Zaremba, J. Pachocki, and D. Farhi (2025). Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint*. <https://arxiv.org/abs/2503.11926>
- [5] M. Ivaldi, B. Jullien, P. Rey, P. Seabright, and J. Tirole (2003). The economics of tacit collusion. *Final report for DG Competition, European Commission*.
- [6] G. J. Stigler (1964). A theory of oligopoly. *Journal of Political Economy*, 72(1):44-61.
- [7] W. E. Kovacic, R. C. Marshall, L. M. Marx, and H. L. White (2011). Plus factors and agreement in antitrust law. *Michigan Law Review*, 110(3):393-436.
- [8] OECD (2017). Algorithms and Collusion: Competition Policy in the Digital Age. *OECD Publishing, Paris*.
- [9] S. R. Motwani, M. Baranchuk, M. Strohmeier, V. Bolina, P. H. S. Torr, L. Hammond, and C. Schroeder de Witt (2024). Secret collusion among AI agents: Multi-agent deception via steganography. *Advances in Neural Information Processing Systems (NeurIPS)*.
- [10] Y. Mathew, O. Matthews, R. McCarthy, J. Velja, C. Schroeder de Witt, D. Cope, and N. Schoots (2025). Hidden in plain text: Emergence & mitigation of steganographic collusion in LLMs. *Proceedings of IJCNLP-AACL; arXiv:2410.03768*. <https://arxiv.org/abs/2410.03768>
- [11] M. Nakamura, A. Kumar, S. Das, S. Abdelnabi, S. Mahmud, F. Fioretto, S. Zilberstein, and E. Bagdasarian (2026). Colosseum: Auditing collusion in cooperative multi-agent systems. *arXiv preprint*. <https://arxiv.org/abs/2602.15198>
- [12] N. Kale, C. B. C. Zhang, K. Zhu, A. Aich, P. Rodriguez, Scale Red Team, C. Q. Knight, and Z. Wang (2025). Reliable weak-to-strong monitoring of LLM agents. *arXiv preprint*. <https://arxiv.org/abs/2508.19461>
- [13] X. Xu, C. Wu, J. Lu, K. Tan, S. Tao, and H. Hong (2026). Collusion with competitive marginals: Price-level audits are blind by construction. *arXiv preprint*. <https://arxiv.org/abs/2607.26385>
- [14] D. Lee and H. Park (2026). Faithful yet collusive: Why chain-of-thought monitoring cannot detect collusion in LLM pricing agents under oligopolistic competition. *arXiv preprint*. <https://arxiv.org/abs/2609.18346>
- [15] K. Agrawal, V. Teo, J. J. Vazquez, S. Kunnavakkam, V. Srikanth, and A. Liu (2025). Evaluating LLM agent collusion in double auctions. *arXiv preprint*. <https://arxiv.org/abs/2507.01413>
- [16] E. J. Green and R. H. Porter (1984). Noncooperative collusion under imperfect price information. *Econometrica*, 52(1):87-100.
- [17] E. Calvano, G. Calzolari, V. Denicolò, and S. Pastorello (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10):3267-3297.
- [18] A. Ezrachi and M. E. Stucke (2017). Artificial intelligence & collusion: When computers inhibit competition. *University of Illinois Law Review*, 2017(5):1775-1810.
- [19] J. E. Harrington, Jr. (2018). Developing competition law for collusion by autonomous artificial agents. *Journal of Competition Law & Economics*, 14(3):331-363.
- [20] J. Foxabbott, K. Deverett, K. Senft, S. Dower, and L. Hammond (2023). Defining and mitigating collusion in multi-agent systems. *NeurIPS 2023 Workshop on Multi-Agent Security*.
- [21] L. Hammond et al. (2025). Multi-agent risks from advanced AI. *Cooperative AI Foundation Technical Report #1, arXiv:2502.14143*. <https://arxiv.org/abs/2502.14143>
- [22] T. Korbak et al. (2025). Chain of thought monitorability: A new and fragile opportunity for AI safety. *arXiv preprint*. <https://arxiv.org/abs/2507.11473>
- [23] R. Y. Lin et al. (2024). Strategic collusion of LLM agents: Market division in multi-commodity competitions. *arXiv preprint*. <https://arxiv.org/abs/2410.00031>
- [24] T. Bonjour, V. Aggarwal, and B. Bhargava (2022). Information theoretic approach to detect collusion in multi-agent games. *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence (UAI), PMLR 180:223-232*.
- [25] O. Tailor (2025). Audit the whisper: Detecting steganographic collusion in multi-agent LLMs. *arXiv preprint*. <https://arxiv.org/abs/2510.04303>
- [26] A. Das, J. Ponde, M. More, and D. Gupta (2026). You can't escape your own activations: Evaluation awareness and multi-agent monitoring. *arXiv preprint*. <https://arxiv.org/abs/2609.03035>
- [27] J. Keppo, Y. Li, G. Tsoukalas, and N. Yuan (2026). On the fragility of AI agent collusion. *arXiv preprint*. <https://arxiv.org/abs/2603.20281>

LLM Usage Statement

We used Claude (Anthropic) to brainstorm approaches and help draft sections. All results and claims were independently verified.