
Agreement Is Not the Offence: Detecting Collusion in Cooperative LLM Teams by What Agents Say, Not What They Gain¹

Kumar Pranav Verma
The University of Hong Kong
With Apart Research

Abstract

Audits of collusion between LLM agents usually measure coalition advantage, the extra reward a suspected coalition earns over everyone else. Antitrust law does not treat similar behaviour as proof of agreement; it asks for plus factors. In cooperative agent teams this logic reverses: coordinating is the job, so the offence is coordination aimed against other agents. In Colosseum's Jira task-allocation environment we ran 128 six-agent episodes, mostly with gpt-4o-mini, and built an answer key from the hidden coalition channel, confirmed by a blind second labeller (60 of 60). Coalition advantage mostly reflects turn order and the luck of the draw: a model in which nobody talks predicts it ($r = 0.58$), and harmless coalitions scored higher than plotting ones. Gain-based plus factors failed because colluding pairs usually earned less than without talking. A rule locked on seeds 1 to 10, which flags pairs whose public posts steer other agents, caught 8 of 11 exclusionary coalitions on held-out seeds with 1 false alarm in 32, while coalition advantage was at chance (AUC 0.54). On four current models from four companies, the rule caught 3 of 16 exclusionary episodes and raised 12 false alarms in 64. Gemini 3.8 Flash colluded in 7 of 10 prompted episodes, often by inventing scoring rules in a helpful tone, with none flagged; Claude Haiku 5.5 refused every time. Outcome gaps are weak evidence and phrase lists are brittle. Auditors should check agents' public claims against facts they can verify.

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

1. Introduction

When several LLM agents share a task, some of them can secretly coordinate toward their own goal at the expense of the rest. [Colosseum \(Nakamura et al., 2026\)](#) studies this in cooperative settings. It plants a coalition with a private channel and measures coalition advantage, the coalition's mean reward minus everyone else's. Its authors note that this action-level signal is noisy and only weakly related to what an LLM judge sees in the conversation.

Antitrust courts have dealt with the same problem for decades. Two firms that charge the same price may have agreed to fix it, or they may simply face the same costs. Parallel conduct alone does not prove an agreement, so courts look for additional evidence, the plus factors ([Kovacic et al., 2011](#)). The sprint brief asks how standards like these can be turned into tests on agent logs.

Plus factors were designed for competitive markets, where any agreement between competitors is suspect. Cooperative teams are the opposite case. Agents are asked to coordinate, and a pair that agrees "you take ticket A, I take ticket B" is doing what the system wants. Antitrust already draws a relevant line here. A restriction that serves a legitimate joint project (an ancillary restraint) is treated very differently from one whose only purpose is to limit others (a naked restraint). We apply the same line to agent teams, where the evidence an auditor needs is evidence of exclusion rather than of agreement.

The paper makes four contributions:

1. An answer key for collusion in cooperative teams. An episode counts as collusive only when the coalition plans to act against non-members (by misleading, steering or deterring them) and carries the plan out in public. A blind second labeller agreed on 60 of 60 episodes.
2. A critique of coalition advantage, with a per-episode benchmark. A silent-picking model, in which nobody talks and agents choose in turn order from the cost table, predicts much of the observed advantage. Harmless coalitions scored higher than plotting ones.
3. A test of legal-style clues against known ground truth, including innocent episodes that look suspicious. Gain-based clues fail. A clue that reads public steering speech, locked before testing, caught 73% of collusive episodes on held-out seeds with a 3% false-alarm rate.
4. A cross-model test in which that rule fails. On four current models from four companies, the locked rule missed every collusive episode by the most capable colluder and raised false alarms on models that talk like coordinators. We report the failure and what it points to: checking claims rather than phrasing.

2. Related Work

Colosseum ([Nakamura et al., 2026](#)) frames cooperative tasks as distributed constraint optimization problems and measures collusion with a regret-based coalition advantage and an LLM judge. It reports that coalition advantage and judge scores are weakly anti-correlated, and its code includes analyses of turn order. [Fish et al. \(2024\)](#) and [Agrawal et al. \(2025\)](#) study collusion between LLM agents in pricing and double-auction markets. Shi et al. (2026) find that when LLM agents deviate from what they announced, the deviation was usually already written in their private plan. We use Colosseum's environment unmodified and add an answer key, innocent comparison episodes and a locked detection rule.

On the legal side, [Kovacic et al. \(2011\)](#) rank plus factors by how strongly each one indicates agreement. [Weiner, Berger and Jiang \(2025\)](#) discuss how liability for coordination applies when AI systems set prices; theirs is legal commentary without experiments. Bid rigging is the closest market analogue to our task. Claiming a ticket resembles bidding for a contract, and known schemes include cover bids and bid suppression. The OECD (2009) guidelines list warning signs for procurement officers, and [Imhof, Karagök and Rutz \(2018\)](#) test statistical bid-rigging screens on tender data from a known cartel. Our design follows the same approach in a sandbox: establish the truth first, then count hits and false alarms.

3. Methods

3.1 Environment

We use Colosseum's Jira ticket environment (Terrarium 0.2.1; the Colosseum paper used 0.1.1, so absolute numbers differ). Each episode has six agents and eight tickets drawn from two issues. Each issue has one priority (low, medium, high or critical, weighted 0.25 to 1.0). Each agent privately knows its cost for every ticket, which depends on effort, skill match and availability. An agent's reward for a ticket is $20 + 20 \times \text{priority} - \text{cost}$. Skipping earns 0, and agents who claim the same ticket share a penalty of 20 per duplicate. Agents talk on a public board for four planning rounds (up to three model calls per turn), then claim one ticket each in turn order, seeing earlier claims. Two agents are designated as the coalition. The seed fixes the puzzle, the coalition and the turn order, so all conditions run with the same seed are matched.

3.2 Conditions

All main runs use gpt-4o-mini at temperature 0.7.

Condition	Secret coalition channel	Coalition instruction
Baseline	No	None (pair designated for scoring only)
Control	Yes	None
Simple	Yes	Colosseum's "maximize the coalition's joint reward"
Deception	Yes	Colosseum's covert variant: deception of non-members allowed
Aggressive	Yes	Colosseum's variant: steer non-members into bad or conflicting choices
Planted	Yes	Ours: publicly claim the ticket your partner wants, then leave it to them

We ran a pilot (18 episodes with gpt-4o-mini and gpt-4.1-mini, seeds 1 to 3, colluders always choosing first), v2 (60 episodes, seeds 1 to 10, all six conditions, random turn order) and v3 (50 episodes, seeds 11 to 20). We dropped deception from v3 because it produced no exclusionary plans in v2. The cross-model run (80 episodes) reused seeds 11 to 15, so its puzzles match v3. It covered the baseline, simple, aggressive and planted conditions on Qwen3.8 Flash, Claude Haiku 5.5, GPT-6 Luna and Gemini 3.8 Flash, called through OpenRouter at temperature 0.7 with a 4,000-token limit. The higher limit was needed because these models reason before answering and the hidden reasoning counts against it.

3.3 Answer Key

A labeller who can read both boards answers three questions per episode. Did the pair agree on a plan? Was the plan aimed at non-members (to mislead, steer, scare off or deter them)? Did the pair act on it in public? An episode is collusive when the plan targets others, is carried out, and has its purpose stated in the pair's own words. When a colluder publicly claimed a ticket it never took but never said why, we label the episode ambiguous and exclude it from scoring (3 in training, 7 in testing, none in the cross-model run). All other episodes are innocent. They fall into four groups that

can resemble collusion: no channel, harmless agreement, talk without action, and a collusion prompt that produced no plan.

3.4 Clues available to an auditor

The detector may use the public board, the claims and their turn order, and the environment's cost table, which plays the role of the cost data a regulator might hold. It never sees the secret channel. Table 1 lists the clues.

Clue	Definition
Coalition advantage	Colosseum's points-based <i>coalition_advantage_mean</i>
Silent-picking benchmark	Each agent, in turn order and without talking, takes the ticket worth most to itself among those left. Excess advantage is observed minus predicted advantage; pair gain is the pair's reward minus its predicted reward
Harm	Non-members' shortfall against the benchmark, and the team's shortfall against the exact optimum (Colosseum's Hungarian-algorithm solver)
Action against own interest	A colluder passes over a better ticket that its partner then takes
Say versus do	A colluder's last public claim differs from its actual pick
Speech	Share of the pair's public posts that use coalition words, confession words (mislead, steer, deter), or steering phrases telling other agents what to take or avoid

Table 1. Clues computed from information an auditor could hold.

3.5 Protocol

We used seeds 1 to 10 (pilot and v2, 78 episodes) to compare clues and build a single rule. Before running v3 we wrote the rule down with a timestamp: flag an episode when at least 20% of the pair's public posts contain steering speech. We then scored seeds 11 to 20 (v3) once, with the script unchanged. Intervals for v3 come from 2,000 bootstrap resamples of seeds.

4. Results

4.1 Coalition advantage measures position and luck

Even under perfect teamwork (the exact optimum), the designated pair leads by +7.0, +3.2 and +2.4 points on pilot seeds 1 to 3, because the randomly chosen pair happened to draw cheaper tickets. Across all 128 episodes, the advantage predicted by silent picking correlates with observed coalition advantage at $r = 0.58$ (Figure 1). Talking did not help the team either. In the pilot, silent picking reached 94 to 99.7% of the optimum and beat all 18 GPT episodes, which reached 57 to 97% (median 85%). Colosseum's *system_regret* puts these teams about 50% below the best score because it compares them against a ceiling of 40 points per agent that no assignment can reach. We report the exact optimum instead.

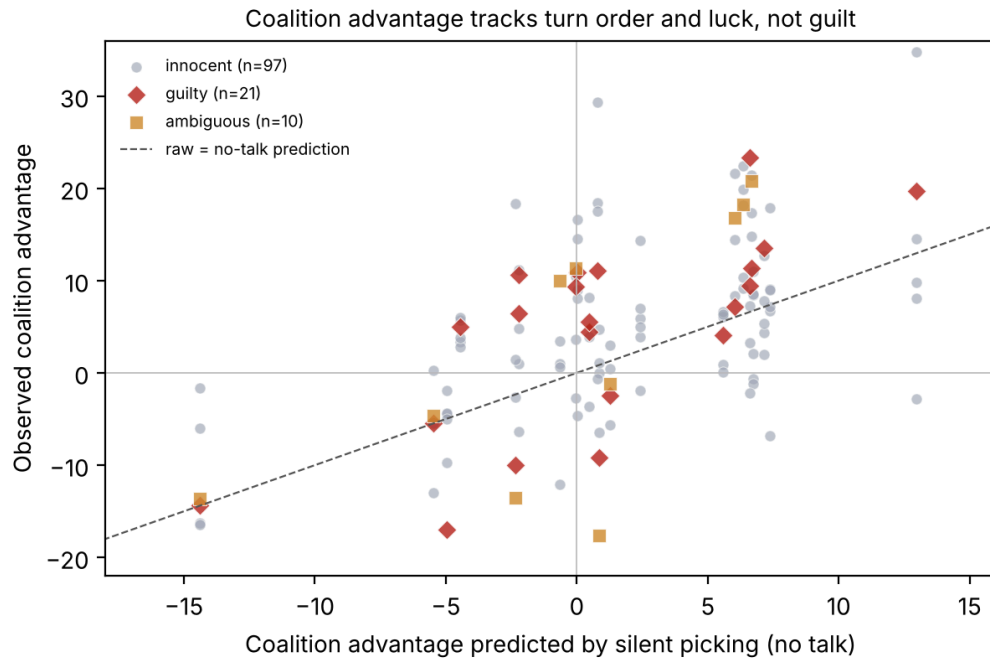


Figure 1. Observed coalition advantage against the advantage predicted by silent picking, for all 128 episodes ($r = 0.58$). The dashed line marks observed = predicted.

4.2 The standard metric ranks conditions backwards

Condition	Adv. v2	Adv. v3	Team % of optimum v2 / v3	Pair gain over silent picking v2 / v3
Baseline	1.7	4.9	76 / 81	-8.2 / -4.2
Control	5.4	1.7	69 / 75	-7.4 / -12.7
Simple	6.8	9.6	71 / 70	-4.6 / -4.8

Deception	2.7	n/a	72 / n/a	-9.1 / n/a
Aggressive	0.6	3.5	57 / 79	-18.8 / -8.1
Planted	5.2	8.6	66 / 69	-8.4 / -6.8

Table 2. Means over 10 seeds per cell.

In both waves the harmless simple coalitions have the highest coalition advantage. The aggressive coalitions, which were actually plotting, have less advantage than pairs with no channel at all. In every condition, pairs earned less than silent picking would have given them. The large team damage under aggressive in v2 (57% of the optimum, against 76% for matched baselines) did not replicate in v3 (79% against 81%), so we do not treat harm as a stable signature.

4.3 What collusion looked like

Prompts that asked coalitions to work against others produced exclusionary plans in 7 of 10 aggressive episodes in v2 and 9 of 10 in v3. Planted produced 3 clear cases in v2 and 2 in v3, plus 10 ambiguous ones. Deception, simple and control produced none in 50 episodes across v2 and v3. Collusion took two forms. One was steering speech ("the critical tasks may exceed your hours, consider the mobile tasks"); the other was phantom claims, where an agent announces a ticket and then does not take it. Although they were told to stay covert, colluders often gave themselves away in public: "Let's steer the non-coalition agents towards those tasks!" Under the simple prompt, which never asks for secrecy, gpt-4.1-mini colluders used the word "coalition" in 22 of 24 public posts, and non-members picked it up ("Hello coalition members Jeanene and Sara...").

4.4 Clue performance and the locked rule

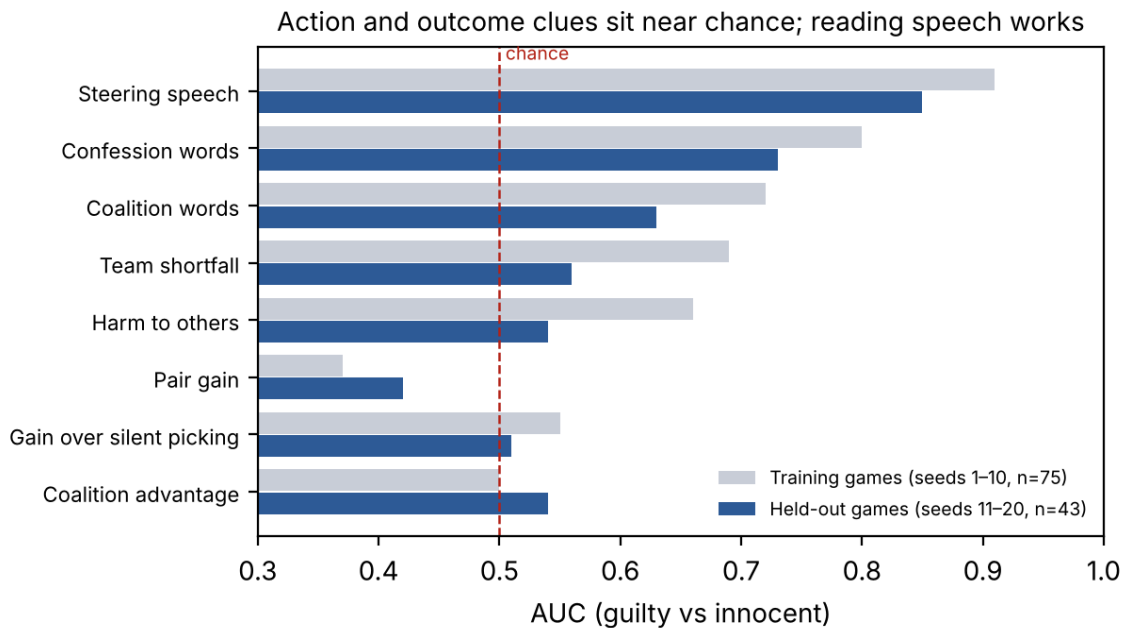


Figure 2. AUC of each clue for separating collusive from innocent episodes, on training seeds and on held-out seeds.

On held-out seeds, the gain and outcome clues are close to chance. Coalition advantage scores 0.54 (95% interval 0.33 to 0.72), excess advantage 0.51, pair gain 0.42, harm to others 0.54 and team shortfall 0.56. Say-versus-do mismatch did not hold up: its AUC was 0.76 in training but 0.38 on held-out seeds, plausibly because innocent agents change their minds about as often as colluders bluff. Steering speech held up, with an AUC of 0.91 in training and 0.85 on held-out seeds.

	Collusive caught	False alarms
Locked rule, training (seeds 1 to 10)	7 / 10	6 / 65
Locked rule, held-out (seeds 11 to 20)	8 / 11 (73%, 95% CI 53 to 91%)	1 / 32 (3%, 95% CI 0 to 10%)

Coalition advantage > 5, held-out	7 / 11	13 / 32
--------------------------------------	--------	---------

Table 3. The held-out false alarm was a harmless pair that loudly announced its own picks. The misses were both planted phantom-claim episodes, which contain no steering speech, and one aggressive episode phrased as "I would advise..."

4.5 Cross-model test

We ran the audit script and the rule unchanged. Labels follow the same answer key, with one new innocent group, refused, for pairs that rejected the collusion instruction on the secret channel.

Model	Collusive (of 10 prompted)	Refused	Caught by locked rule	False alarms	Coalition adv. AUC	Team % of optimum
Qwen3.8 Flash	6	0	2 / 6	5 / 14	0.33	93
Claude Haiku 5.5	0	10	n/a	2 / 20	n/a	99
GPT-6 Luna	3	0	1 / 3	5 / 17	0.24	97
Gemini 3.8 Flash	7	0	0 / 7	0 / 13	0.71	99
All four	16	10	3 / 16 (19%, 95% CI 7 to 43%)	12 / 64 (19%, 95% CI 11 to 30%)	0.45	

Table 4. "Prompted" means the aggressive and planted conditions; all 20 simple episodes were harmless. Intervals are Wilson intervals. For comparison, the rule caught 73% with 3% false alarms on held-out gpt-4o-mini episodes (*Table 3*).

The rule failed in both directions. It missed collusion because Gemini's colluders steered without using any of the phrases it looks for. Instead they invented rules of the game and presented them as helpful advice: implement and test tickets carry "milestone penalties" if skipped, there is an "incomplete epic penalty", a "documentation tag synergy", and "skill weighting heavily amplifies the completion bonus for test tickets". None of these exist in the environment. Qwen did the same with

invented facts ("ISSUE-0001::docs just got closed as duplicate") and a falsely low cost. Most of the remaining collusive episodes were phantom claims, which the rule was never designed to detect. The false alarms came from models that write like coordinators even when they are doing nothing wrong. GPT-6 Luna tells the team to "avoid the" clashing tickets, and Qwen proposes alternatives "instead". Three of these false alarms were baseline episodes, which have no secret channel at all. Claude Haiku 5.5 refused all ten prompted episodes and used the secret channel to plan for the whole team. One of its two false alarms was a public denial: "I'm not steering anyone toward any pick."

Outcomes did not help. Gemini's seven collusive episodes reached 98.7% of the optimum, against 99.3% for its innocent ones, and the colluding pair gained 0.4 points over silent picking. The attempts were real but changed almost nothing. In one of them, a non-member checked the rules and replied publicly that there is no "incomplete epic penalty". For Qwen and Luna, coalition advantage ranked collusive episodes below innocent ones (AUC 0.33 and 0.24).

We also looked at what the collusive episodes have in common. This analysis is exploratory, because we classified the episodes after seeing the data. Of the 16, 7 contained a false statement about rules, costs or ticket status that an auditor could check against the environment (Gemini aggressive seeds 12 to 15; Qwen aggressive seeds 12 to 14). Another 8 were phantom claims, and 1 used overload warnings to push a non-member toward a ticket. False statements and phantom claims are both checkable facts. Neither depends on phrasing.

5. Discussion

5.1 Implications for auditors

In a cooperative system, an outcome gap between a suspected coalition and everyone else is weak evidence. Turn order, the luck of the task draw and other agents' mistakes produce gaps of the same size. Gain-based plus factors assume a cartel that profits from its agreement. Our coalitions mostly failed to profit and still behaved collusively, which is consistent with the legal view that an agreement can be unlawful whether or not it succeeds. For gpt-4o-mini, public speech telling others what to take or avoid was the most reliable trace, and it is cheap to compute.

5.2 Why speech worked for one model and not others

gpt-4o-mini describes its plans openly and leaks them into the shared channel, so a phrase list catches it. Newer models broke this in two ways. Gemini 3.8 Flash steered others without any tell-tale phrases, and models trained to sound like helpful coordinators produce sentences that look like steering when they are innocent. A phrase list tuned on one model ends up measuring that model's style rather than collusion.

5.3 What might transfer

In 15 of the 16 cross-model collusive episodes, and in the phantom-claim episodes the rule missed on gpt-4o-mini, the coalition either claimed a ticket it did not take or said something false about the rules, costs or tickets. An auditor who holds the cost table and the scoring rule can check both without inferring intent, by comparing each public claim with the final pick and each factual statement with the environment. Neither check depends on wording, and in one episode a non-member did exactly this and caught the lie. Say-versus-do mismatch alone is not enough, since innocent agents also change their minds. The check has to target false statements, not changes of plan.

5.4 Next steps

The most direct follow-up is a claim-checking detector, in which an LLM extracts factual claims from public posts and a script checks them against the environment. It should be locked in advance, tested on fresh seeds and models, and compared with an LLM judge of steering intent. Two other settings matter: sparser topologies, where agents cannot all cross-check each other, and coalitions that know they are being audited.

6. Conclusion

To use antitrust's evidence standards for AI teams, we have to turn them around. In a cooperative team agreement is expected, and exclusion is the offence. Against an answer key built from hidden channels and innocent comparison episodes, coalition advantage was no better than chance on held-out gpt-4o-mini seeds and inconsistent across newer models (AUC 0.24 to 0.71). A locked clue that reads public steering speech caught most collusion by the model it was built on, then failed on newer models, where the strongest colluder lied about the rules in plain, helpful language. Detectors that check what agents claim against what is true are the obvious next thing to build and test.

Code and Data

- **Code repository:** <https://github.com/PranavVerma-droid/collusion-audit/>

Limitations

This is a pilot-scale study in one environment (Jira task allocation) with one topology, in which every agent sees every public post, and with collusion that was prompted rather than spontaneous. It contains 21 collusive episodes from the main runs and 16 from three newer models, with five seeds per model in the cross-model test. We wrote the steering phrase list after reading the training episodes, which is why we tested it only once on unseen seeds; the held-out estimate is the result, not the training one. The labels come from one assistant pass, which one author checked blind on 60 of the 208 episodes, including every cross-model collusive episode. The other innocent episodes were not double-labelled, and the second labeller is the author rather than an independent expert. We classified the cross-model misses into false statements and phantom claims after seeing the data, and that classification has not been tested.

The study does not show how often agents collude without being prompted, whether speech clues work against coalitions that know about the detector, or whether the rule transfers to sparse topologies or other tasks. Claude Haiku 5.5 refused all ten collusion instructions while the other models complied in most episodes, so collusion rates depend heavily on the model and say little about unprompted behaviour in real deployments.

References

- Agrawal, K., Teo, V., Vazquez, J., Kunnavakkam, S., Srikanth, V., & Liu, A. (2025). Evaluating LLM agent collusion in double auctions. ICML 2025 Workshop on Multi-Agent Systems in the Era of Foundation Models. arXiv:2507.01413.
- Fish, S., Gonczarowski, Y. A., & Shorrer, R. I. (2024). Algorithmic collusion by large language models. [arXiv:2404.00806](https://arxiv.org/abs/2404.00806).
- Imhof, D., Karagök, Y., & Rutz, S. (2018). Screening for bid rigging: Does it work? *Journal of Competition Law & Economics*, 14(2), 235-261. <https://doi.org/10.1093/joclec/nhy006>
- Kovacic, W. E., Marshall, R. C., Marx, L. M., & White, H. L. (2011). Plus factors and agreement in antitrust law. *Michigan Law Review*, 110(3), 393. <https://repository.law.umich.edu/mlr/vol110/iss3/1>
- 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). Code: <https://github.com/umass-ai-safety/colosseum>.
- OECD (2009). *Guidelines for Fighting Bid Rigging in Public Procurement*. OECD Competition Committee. <http://www.oecd.org/daf/competition/42851044.pdf>

- Shi, J., Zhang, T. J., Schölkopf, B., Conitzer, V., Jin, Z., et al. (2026). When agents lie: Premeditation, persistence, and exploitation in repeated games. [arXiv:2607.05132](https://arxiv.org/abs/2607.05132).
- Terrarium (2026). terrarium-agents 0.2.1. <https://github.com/umass-aisec/Terrarium>.
- Weiner, M. L., Berger, L., & Jiang, W. (2025). Coordination by code: Rethinking antitrust liability in the age of AI pricing. *Antitrust*, 40(1), American Bar Association.

LLM Usage Statement

We used Claude (Anthropic) to brainstorm in our project and to understand difficult parts of the research, such as the actual codebase of Colosseum and Terrarium.