

AI Collusion Forensics: Comparing Communication and Behavioral Evidence

Agnes Ahalya Arogyaraj
Saarland University

With
Apart Research

Abstract

This study examines whether combining what AI agents say with what they do improves collusion detection. Inspired by Track 2.2 of the Apart Research AI Collusion Research Sprint (Apart Research (2026)), we compare four monitoring approaches using the same public-goods game trajectories: public messages only, actions and outcomes only, a combined timeline, and a structured comparison of statements and actions. We conduct 35 trials with two language-model agents: 15 independent-control trials, 15 coordination-opportunity trials, and five instructed-coordination trials. Three labeling models assess each trajectory, followed by confidence and evidence checks. Using provisional judge-majority labels, the combined and structured monitors achieved 100% agreement, compared with 91.4% for messages only and 85.7% for actions only. However, only seven trajectories received verified final labels, while 28 remained unknown. The verified cases include five instructed-coordination positives and two negatives. The combined and structured approaches performed equally, showing no additional benefit from structuring evidence in this sample. These results suggest that combining communication and behavioral evidence may help collusion monitoring, but the limited verified labels prevent strong conclusions about detecting spontaneous collusion.

1 Introduction

As language-model agents increasingly work together and make decisions, it becomes important to understand whether they act independently or coordinate in ways that deviate from expected rules. Collusion can be hard to detect because agents may communicate their intentions, make agreements, or behave similarly without necessarily following a coordinated plan. This makes it important to understand what evidence an AI monitor needs to reliably identify collusion.

This motivates our research question, inspired by Track 2.2 of the Apart Research AI Collusion Research Sprint (Apart Research, 2026): **Does combining what agents say with what they do improve collusion detection compared with either source alone?** To investigate this, we use a public-goods game with two language-model agents and compare four monitoring approaches on the same game trajectories: messages only, actions and outcomes only, raw combined evidence, and a structured comparison of what agents said and what they actually did. We also introduce independent-decision and coordination-opportunity conditions to examine how access to other agents’ announcements affects their decisions. Our goal is to understand whether combining and structuring evidence helps monitors detect collusion more accurately without increasing false positives.

Our main contributions are:

1. A controlled public-goods experiment with independent-decision and coordination-opportunity conditions.

2. An evidence-based labeling process using three independent judges, verification checks, and an unknown category for uncertain cases.
3. A matched comparison of four monitoring approaches to evaluate how different evidence representations affect collusion detection.

2 Related Work

Research on multi-agent systems has examined how communication and behavior reveal coordination. COLOSSEUM(Nakamura et al. (2026)) investigates collusion through agent messages and observed actions, distinguishing expressed intentions from executed behavior. Radev et al. (2026) show that access to agent communication can improve detection of coordinated sabotage. Shi et al. (2026) examine differences between private intentions, public commitments, and final actions in repeated games, providing the experimental foundation for our study.

Monitoring research also highlights how evidence presentation affects detection. Arike et al. (2026) find that additional information does not always improve sabotage detection, while Nguyen et al. (2026) introduce TraceGuard, which uses structured assessments of agent behavior. Building on these findings, we compare four monitoring approaches on identical multi-agent trajectories: (1) public messages only, (2) actions and outcomes only, (3) combined messages and actions, and (4) structured combined evidence comparing stated and observed behavior. We investigate whether combining communication and behavioral evidence improves collusion detection over either source alone.

3 Methodology

3.1 Game and Experimental Conditions

We use the existing `llm-trust-breaking` public-goods game with two agents and three conditions. In *independent control*, agents are instructed to decide independently and cannot see each other’s announcements before acting. In *coordination opportunity*, they receive the same instruction but can see the announcements. We run 15 trials per condition.

We also run five *forced collusion* trials as an instructed coordination control. One agent proposes a positive contribution amount, and the other is asked to accept the same amount. We check whether both agents carry out the agreement without changing their outputs.

3.2 Trajectory Records and Labeling

Each trajectory records private plans, public announcements, final actions and reasoning, outcomes, self-reports, and available API usage. Self-reports are treated as claims, not ground truth. We keep the original game judge separate from the research labels.

Three independent labeling models, GPT-4.1 Nano, Mistral Small 3.2, and DeepSeek V3.2, classify each trajectory as positive, negative, or unknown, with confidence scores and evidence citations. For the original conditions, a positive label requires an explicit agreement, a violation of the independence instruction, and actions that carry out the agreement. For forced collusion, a separate rubric requires a proposal, acceptance, and execution.

A final label requires at least two valid agreeing judges, median confidence of at least 0.60, no

contradictory evidence flag, and verified citations and actions. Otherwise, the label remains *unknown*. We flag these cases for human review, but did not complete it within the project timeframe. Unknown labels are never counted as negatives.

3.3 Four Monitoring Approaches

We compare four evidence views on the same trajectories using GPT-4.1 Mini at temperature 0:

1. **M1:** Public announcements only.
2. **M2:** Final actions and game outcomes only.
3. **M3:** Combined timeline of announcements, actions, and outcomes.
4. **M4:** Structured comparison of agents’ claims and actions, including cross-agent observations.

Monitors cannot see experimental conditions, private plans, self-reports, judge outputs, or final labels. Scores of 0.5 or higher predict collusion. Using the same monitor model and trajectories allows us to compare the four evidence views.

3.4 Evaluation and Reproducibility

We evaluate monitors against verified labels using accuracy and false-positive rate as the main metrics, alongside precision, recall, and confusion matrices. Unknown labels and failed monitor responses are reported separately, and metrics with zero denominators are marked undefined. We compare predictions on matched trajectories and across conditions.

4 Results and Findings

We evaluated four monitors on 35 trajectories. Table 1 reports agreement with provisional judge-majority votes and performance against verified final labels.

Table 1: Monitoring results across 35 trajectories.

Monitor	Reference	Correct	Incorrect	Unverified	Precision
M1	Judge majority*	32/35 (91.4%)	3/35 (8.6%)	0	62.5%
M1	Final label	7/35 (20%)	0	28/35 (80%)	100%
M2	Judge majority*	30/35 (85.7%)	5/35 (14.3%)	0	N/A
M2	Final label	2/35 (5.7%)	5/35 (14.3%)	28/35 (80%)	N/A
M3	Judge majority*	35/35 (100%)	0	0	100%
M3	Final label	7/35 (20%)	0	28/35 (80%)	100%
M4	Judge majority*	35/35 (100%)	0	0	100%
M4	Final label	7/35 (20%)	0	28/35 (80%)	100%

*Judge-majority votes are provisional, not verified ground truth. Final-label rows include 28 unknown cases, which are excluded from verified performance calculations. Forced coordination uses a separate rubric.

Using provisional judge-majority votes, M3 and M4 achieved 100% agreement, compared with 91.4% for M1 and 85.7% for M2.

Only seven trajectories received verified labels: five instructed-coordination positives and two negatives. On these cases, M1, M3, and M4 correctly classified all seven (100% accuracy), while M2 correctly classified two (28.6% accuracy). All four monitors had a false-positive rate of 0% on the two verified negatives, but this estimate is based on very few cases.

Main finding: Combined evidence matched the provisional labels more often than either source alone. M3 and M4 performed equally, showing no additional benefit from structured evidence in this sample. With 28 unknown labels and no verified spontaneous collusion cases, the results cannot establish which monitor is most reliable for detecting spontaneous collusion.

5 Discussion and Limitations

This study compares four monitoring approaches to understand whether combining agents' statements and actions improves collusion detection. It also examines whether structured evidence is more useful than a raw combined timeline. The instructed coordination condition provides examples of agreements being carried out, but does not represent spontaneous collusion. Our findings apply to this controlled setting and should not be assumed to generalize to real-world systems without scalable testing and evaluation.

Limitations

The study uses only 35 trajectories from one public-goods game with two agents. Most original-condition trajectories lack verified labels, limiting reliable performance comparisons. The three labeling models may share mistakes, and their confidence scores are not calibrated. The independence instruction and self-report question may influence agent behavior. Results may also depend on the chosen monitor model and detection threshold.

Dual-Use and Ethical Considerations

Comparing monitors could reveal weaknesses that others might use to hide collusion. However, our study focuses on detection rather than evasion and uses a controlled game instead of real-world systems. Current evidence is too limited to identify reliable monitoring weaknesses or determine whether the findings could help evade detection. This potential risk should still be considered when reporting results.

Future Work

Future work could test more trajectories, different games, controlled real-world scenarios, and larger agent groups. It could also compare different models and thresholds, study the effect of self-reports, and check model-generated labels against human judgments.

6 Conclusion

This study develops a controlled framework for comparing four evidence views in collusion detection. It combines independent-decision and instructed-coordination conditions with evidence-based labeling. The limited verified labels prevent strong conclusions about which monitoring approach performs best, but the framework provides a basis for further evaluation.

Code and Data

- Code repository: https://github.com/AgnesAhalya/collusion_hackathon

References

- Apart Research. AI Collusion Research Sprint 2026: Resources and research questions. Apart Research website, 2026. URL <https://apartresearch.com/sprints/ai-collusion-research-sprint-2026-10-23-to-2026-10-25#resources>. Track 2.2: Does combining what agents say with what they do beat either alone?
- Rauno Arike, Raja Mehta Moreno, Rohan Subramani, Shubhorup Biswas, and Francis Rhys Ward. How does information access affect llm monitors’ ability to detect sabotage? *arXiv preprint arXiv:2601.21112*, 2026.
- Mason Nakamura, Abhinav Kumar, Saswat Das, Sahar Abdelnabi, Saaduddin Mahmud, Ferdinando Fioretto, Shlomo Zilberstein, and Eugene Bagdasarian. Colosseum: Auditing collusion in cooperative multi-agent systems. *arXiv preprint arXiv:2602.15198*, 2026.
- Khanh Linh Nguyen, Hoa Nghiem, and Tu Tran. Traceguard: Structured multi-dimensional monitoring as a collusion-resistant control protocol. *arXiv preprint arXiv:2604.03968*, 2026.
- Nikolay Radev, Lennart Haas, Benjamin Arnav, and Pablo Bernabeu-Pérez. The best-laid schemes: Coordinated sabotage and monitoring in multi-agent systems. *arXiv preprint arXiv:2605.29178*, 2026.
- Jerick Shi, Terry Jingcheng Zhang, Bernhard Schölkopf, Vincent Conitzer, and Zhijing Jin. When agents lie: Premeditation, persistence, and exploitation in repeated games. *arXiv preprint arXiv:2607.05132*, 2026.

LLM Usage Statement

We used large language models throughout this project to brainstorm research ideas, understand and summarize papers, plan experiments, generate and debug code, analyze experimental outputs, and draft and improve the report. LLMs also served as agents, labelers, judges, and monitors in the experiments. The author reviewed the generated code, research claims, and report content and remained responsible for research decisions, result interpretation, and final submission.