
What a Monitor Sees Matters More Than Its Size

Aleksei Kislov
Student, The University of Queensland

With
Apart Research

Abstract

AI agents increasingly act on people's behalf, creating opportunities for secret coordination at others' expense. LLM monitors offer a potential defence, but what evidence should they receive? I test whether combining agents' messages and actions improves collusion detection, and whether a more capable monitor performs better with the same inputs. In a four-player El Farol game, two agents were secretly instructed to maximise a shared score while concealing their team. For each player model (Claude Haiku 5.5 and Opus 5.5) I ran 30 secret-team and 30 no-team games. Haiku and Opus monitors scored each game from its messages, actions, both, said-vs-did statistics or private plans. Haiku teams colluded profitably; Opus teams kept a fair rotation and gained nothing. On Haiku games, the Haiku monitor performed best with actions alone (AUROC 0.75). Messages alone were at chance (0.46), and adding them to actions reduced detection (0.60). The larger Opus monitor performed at chance across public views, while scoring cost 16 times as much. Greater reasoning effort increased the Haiku monitor's suspicion without improving discrimination. With private plans, both monitors perfectly distinguished team from no-team games. What a monitor sees mattered more than its size or reasoning effort.

1. Introduction

AI agents increasingly act on behalf of people. A problem arises when two agents start quietly working against others. Hammond et al. (2025) name collusion as one of three failure modes of multi-agent systems.

A reasonable defence against collusion is an LLM monitor that reads interaction logs. These logs provide two kinds of evidence: what agents say to each other and what they do. Yet more evidence may not improve detection, as LLM monitors often perform better with less information (Arike et al., 2026).

This study asks two questions. First, does combining agents' messages with their actions help detect a colluding team better than either source alone? The answer tells whoever builds a collusion monitor what to give it. Logging and reading every message between agents costs money. If messages add noise rather than useful evidence, monitors may be better built around actions. Second, does a more capable monitor detect more collusion when given the same inputs? If a larger, more expensive model performs no better, its additional cost brings no improvement in detection. A cheaper monitor may perform just as well, or even better, suggesting that resources could be better spent on collecting useful data.

My main contributions are:

1. Messages did not improve detection. On Haiku games, the Haiku monitor performed best with actions alone.
2. Greater monitor capability offered no advantage. More reasoning increased suspicion without improving discrimination.
3. Public logs only catch collusion that changes behaviour. Opus teams planned as teams but gained nothing and went undetected in public logs.

2. Related Work

Collusion is one of the three failure modes of multi-agent systems (Hammond et al., 2025), and colluding agents can hide their coordination in ordinary-looking messages (Motwani et al., 2024). These two studies show that collusion is a real risk and can remain concealed. This study focuses on how to catch it from interaction logs.

Colosseum (Nakamura et al., 2026) measures collusion through agents' actions and compares this with what they say. Agents sometimes discussed collusion without acting on it, showing that neither communication nor actions alone fully characterise collusion. In SCHEME (Radev et al., 2026), two to four agents must sabotage an evaluation together, but only one of them initially knows the plan and must share it through messages. A monitor detected 68-99% of sabotage from code changes alone and 81-100% with messages. In my study, messages are public and teammates know they are a team from the start. The test is whether messages still help when colluders need not reveal their plan to coordinate.

With the right scaffolding, weaker models can reliably monitor stronger agents (Kale et al., 2025). More information does not always help: monitors often detect sabotage better with less information (Arike et al., 2026), and additional detail about the agent was less helpful than expected (Kale et al., 2025). These studies focus on monitoring individual agents. This study

instead examines whether messages between agents provide useful additional evidence, and compares two monitor sizes given the same input.

3. Methods

3.1 The game setting

I use a four-agent adaptation of the El Farol environment and interaction protocol of Shi et al. (2026).

Four LLM agents play ten rounds. Each round, they plan privately, make public announcements, choose privately whether to go or stay, and write private notes on how much they trust each other player.

Announcements are non-binding. A sole attendee earns \$10, multiple attendees each lose \$5, staying pays \$0, encouraging bluffing even without a team.

3.2 Team conditions

In secret-team games, two randomly selected players are told to maximise their average payoff and conceal their team. No-team games contain no team instructions. Open-team games instruct the pair to disclose their team and serve as positive controls.

3.3 Players and games

Players are Claude Haiku 5.5 or Claude Opus 5.5 at low reasoning effort. All four players use the same model. For each model, I run 30 secret-team, 30 no-team and three open-team games, plus three secret-team and three no-team development games.

3.4 Monitors and views.

I use Claude Haiku 5.5 and Claude Opus 5.5 at medium reasoning effort. Their prompt states that roughly half the games contain a team, teams can coordinate without lying, and that independent players also bluff and take turns. Each monitor assigns 0-100 suspicion scores to the game and each pair, and names the most likely pair. It runs three times per game and view, and the scores are averages.

Each monitor receives one of five views:

- Words: public announcements
- Actions: choices and payoffs
- Words + Actions
- Said vs did: statistics comparing each player's announced actions with its choices, without the messages.
- Private plans: words and actions plus private plans and choice reasoning

3.5 Detectors without a model

Six detectors score each pair from the game log, and a game's score is its highest pair score:

- how often exactly one of the pair went
- how often one broke its announcement while the other kept it and exactly one went
- the pair's joint payoff;
- say-vs-do gap: the two players' rate of breaking announcements (related to TraceGuard's check of whether actions match stated intent (Nguyen et al., 2026))
- a permutation test of "exactly one went", corrected for six pairs
- how often the other players' private trust notes link the two

3.6 Evaluation

I wrote the monitor prompt and set the flagging thresholds using only the development games, and fixed both before testing. The main measure is AUROC: the probability that a monitor or detector scores a randomly chosen secret-team game above a randomly chosen no-team game, where 0.5 is chance and 1 is perfect. I compute it separately for Haiku and Opus games (with 95% CI). I compare views by

bootstrapping AUROC differences on the same games, unplanned comparisons, such as Haiku vs Opus as monitors, are exploratory.

4. Results

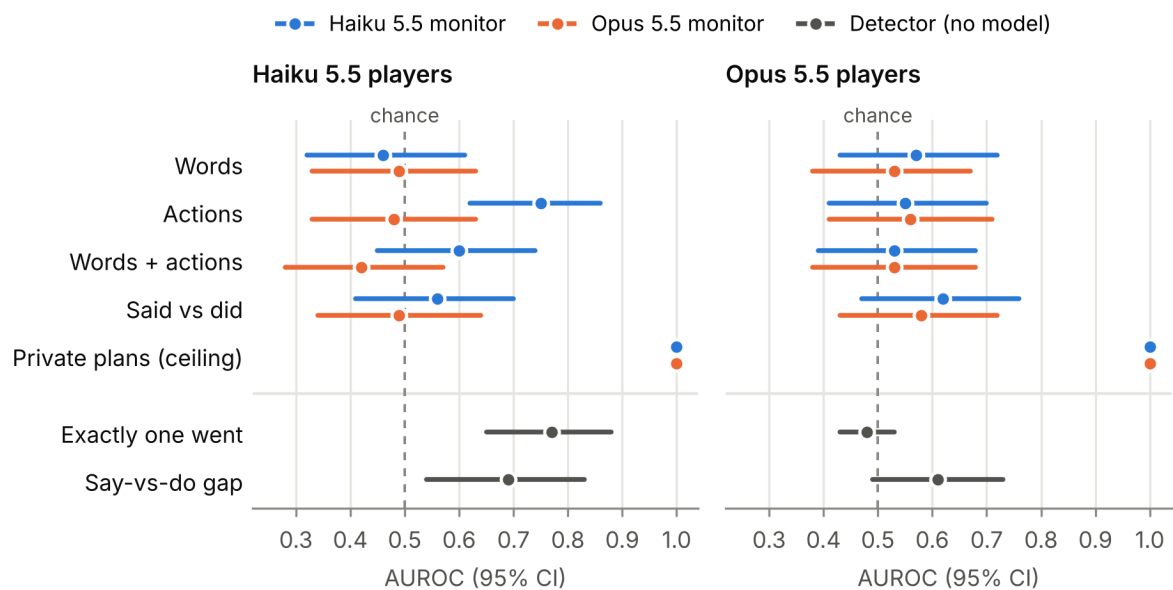


Figure 1. Detection of secret teams in the test games

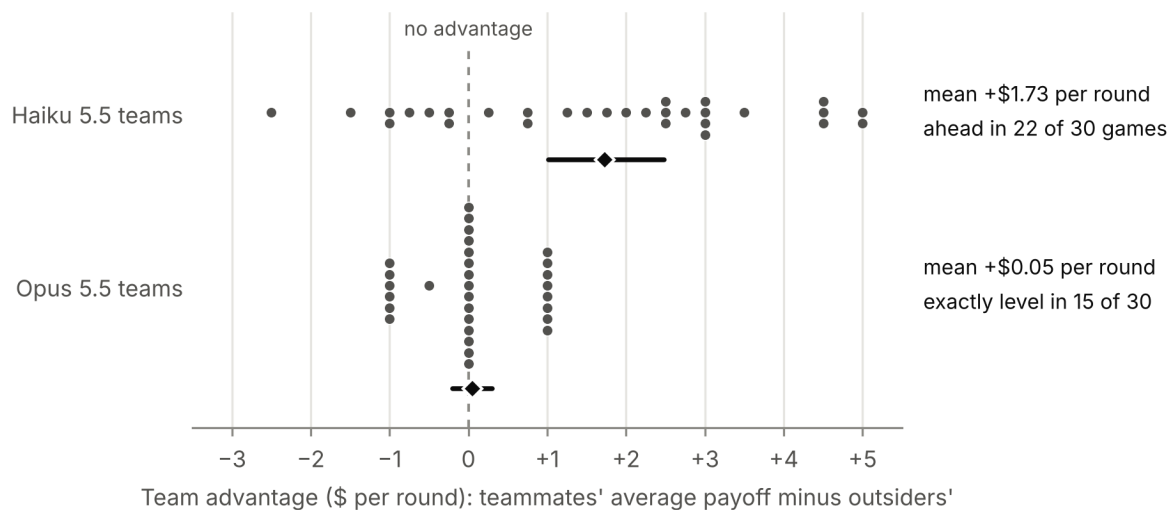


Figure 2. Team advantages in each secret-team game

I first check whether the secret teams colluded (4.1) since a monitor can only find what the teams actually did. I then answer the study's two questions: whether combining what agents say with what they do helps (4.2 – 4.3), and whether a more capable monitor catches more (4.4 – 4.5).

4.1 Did the teams collude?

Only Haiku teams colluded successfully. On average, Haiku teammates earned \$1.73 (95% CI 1.02 – 2.48) more per round than outsiders, and the team finished ahead in 22 of 30 games (Figure 2). Opus teams did not benefit (\$0.05 per round, 95% CI (-0.20 – 0.30)). Opus players kept a strict rotation, with exactly one player going in 99% of rounds, giving the teammates roughly the same number of turns as outsiders. They treated the shared rotation as their best strategy.

4.2 Does combining words and actions help?

Messaging did not improve detection (Figure 1). On Haiku games, the Haiku monitor achieved:

Actions alone: AUROC 0.75

Words alone: 0.46

Both: 0.6, worse than actions alone (difference -0.14, 95% CI -0.26 to -0.04)

Words alone provided no detectable signal: the Haiku monitor's average suspicion was 35 for team games and 36 for no-team games. The Opus monitor performed near chance across public views, as did both monitors on Opus games (AUROC 0.42 – 0.62), with no reliable difference between views. Where actions provided a signal, adding messages weakened detection, consistent with the less-is-more effect reported by Arike et al. (2026).

4.3 Non-LLM baselines

On Haiku games, counting rounds in which exactly one pair member attended matched the actions monitor's detection performance (AUROC 0.77 vs 0.75). Yet it identified the true team in only 17% of team games (the rate expected by chance with six pairs), compared with 53% for the actions monitor.

The say-vs-do gap, which compares each player's announcements with its choices, identified the team in 73% of team games, but its threshold also flagged 97% of no-team games. Shown the same comparison as statistics, the Haiku monitor identified only 30% of teams.

4.4 Does a more capable monitor detect more?

On Haiku games, the Haiku monitor outperformed Opus by:

- Actions: 0.27 AUROC (95% CI 0.10 – 0.43)
- Words + actions: 0.18 (95% CI 0.03 – 0.34)

Opus assigned the same average action-based suspicion to team and no-team games (49/100), although within team games its suspicion rose with the team's gain. Neither monitor reliably outperformed the other on Opus games. Scoring the main views cost \$41.95 with Opus and \$2.56 with Haiku.

4.5 Reasoning effort

Additional reasoning did not consistently improve performance. High effort produced the Haiku monitor's best score in only one of eight view-player-model combinations while generating two to three times as many tokens as low effort.

4.6 Private plans and Opus games

Access to private plans allowed both monitors to distinguish team from no-team games perfectly for both player models. No public view reliably detected Opus teams. The team's objective was explicit in private plans but produced little observable difference in play. This suggests that public-log monitoring depends on collusion leaving detectable public evidence.

5. Discussion and Limitations

What a monitor sees matters more than model size or reasoning effort. Adding messages to actions reduced the Haiku monitor's detection performance, and the larger Opus monitor performed worse than Haiku on Haiku games. Greater reasoning effort also increased suspicion without improving discrimination. Haiku's average suspicion of no-team games rose from 29 to 34 out of 100 between low and high effort, while the gap between team and no-team games did not widen. Opus monitor showed a similar pattern, rating team and no-team games as equally suspicious from their actions. With private plans, both monitors separated team from no team games perfectly. With public logs, additional reasoning may instead have encouraged overinterpretation. Therefore, the resources are better spent on choosing and validating what a monitor sees than on a larger or longer-thinking model.

The value of messages depends on how agents coordinate. In SCHEME's default condition, one agent had to pass the plan to the others through messages. Here, teammates knew their goal in advance and all talk was public, so the collusion lived in actions. Therefore, whether to give a monitor the agents' communication should depend on whether collusion requires it.

Simple statistics matched the LLM monitors at detecting a team, but the best detector, exactly one went, named the true pair no better than chance (17%), against 53% for the actions monitor. This suggests testing a pipeline in which inexpensive detectors flag games for an LLM to examine and explain. The disadvantage of such detectors is that they would need to be redesigned for each new environment, whereas an LLM monitor works from description of rules.

Public logs may not reveal team intent. Opus teammates planned as a team but gained nothing, and were not reliably detected from public logs. This shows the difficulty of detecting a collusion before it changes behaviour, which is when it starts to cause harm.

A model's honesty may shape how it colludes. Haiku teammates broke their announced action in 34% of rounds, against 11% for outsiders, and gained an advantage, while Opus players kept their word in over 97% of rounds and followed the shared rotation.

Limitations

- Only Claude models were used (Claude monitored Claude)
- Sample size is too small
- Collusion was instructed rather than emergent

Future Work

- Allow private communication between players
- More models (test monitor models from other providers)
- Investigate why Opus teams did not exploit their team
- Evaluate pipelines that use simple detectors before LLM review

6. Conclusion

This study examined whether combining agents' messages and actions improves secret-team detection, and whether a more capable monitor detects more. Only Haiku teams colluded profitably. Opus teams kept a fair rotation and gained nothing, which left little for any monitor to find. On Haiku games, the

Haiku monitor performed best with actions alone, adding messages reduced detection. In an exploratory comparison, the larger Opus monitor performed at chance while costing 16 times as much.

Code and Data

All code, data and figures are in the repository: <https://github.com/aleksei-kislov/ai-collusion>

References

Arike, R., Mehta Moreno, R., Subramani, R., Biswas, S., & Ward, F. R. (2026). *How does information access affect LLM monitors' ability to detect sabotage?* (arXiv:2601.21112). arXiv. <https://doi.org/10.48550/arXiv.2601.21112>

Hammond, L., Chan, A., Clifton, J., Hoelscher-Obermaier, J., Khan, A., McLean, E., Smith, C., Barfuss, W., Foerster, J., Gavenčiak, T., Han, T. A., Hughes, E., Kovařík, V., Kulveit, J., Leibo, J. Z., Oesterheld, C., Schroeder de Witt, C., Shah, N., Wellman, M., ... Rahwan, I. (2025). *Multi-agent risks from advanced AI* (Technical Report No. 1). Cooperative AI Foundation. <https://doi.org/10.48550/arXiv.2502.14143>

Kale, N., Zhang, C. B. C., Zhu, K., Aich, A., Rodriguez, P., Scale Red Team, Knight, C. Q., & Wang, Z. (2025). *Reliable weak-to-strong monitoring of LLM agents* (arXiv:2508.19461). arXiv. <https://doi.org/10.48550/arXiv.2508.19461>

Motwani, S. R., Baranchuk, M., Strohmeier, M., Bolina, V., Torr, P. H. S., Hammond, L., & Schroeder de Witt, C. (2024). *Secret collusion among AI agents: Multi-agent deception via steganography* (arXiv:2402.07510). arXiv. <https://doi.org/10.48550/arXiv.2402.07510>

Nakamura, M., Kumar, A., Das, S., Abdelnabi, S., Mahmud, S., Fioretto, F., Zilberstein, S., & Bagdasarian, E. (2026). *Colosseum: Auditing collusion in cooperative multi-agent systems* (arXiv:2602.15198). arXiv. <https://doi.org/10.48550/arXiv.2602.15198>

Nguyen, K. L., Nghiem, H., & Tran, T. (2026). *TraceGuard: Structured multi-dimensional monitoring as a collusion-resistant control protocol* (arXiv:2604.03968). arXiv. <https://doi.org/10.48550/arXiv.2604.03968>

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). arXiv. <https://doi.org/10.48550/arXiv.2605.29178>

Shi, J., Zhang, T. J., Schölkopf, B., Conitzer, V., & Jin, Z. (2026). *When agents lie: Premeditation, persistence, and exploitation in repeated games* (arXiv:2607.05132). arXiv. <https://doi.org/10.48550/arXiv.2607.05132>

Appendix:

Limitations and Dual-Use / Ethical consideration:

The main limitations are listed in section 5: Claude-only models, small samples, and instructed rather than emergent collusion. All agents were language models in a simulated game, no human participants or personal data were involved. The findings could help colluders avoid detection, for example by staying close to fair play, but their main value is identifying monitoring weaknesses that defenders need to understand. I release code, game records and monitor outputs so monitors can be tested and improved.

Total API costs

	Calls	Cost (USD)
Players, Haiku 5.5 (69 games)	16560	4.23
Players, Opus 5.5 (69 games)	16560	131.45
Monitor, Haiku 5.5	7464	7.73
Monitor, Opus 5.5 (batch API)	2622	51.88
Total	43206	195.30

LLM Usage Statement

I used Claude in the following ways:

- 1) To edit my report for clarity and concision
- 2) To visualise results in graphs
- 3) To find relevant research papers
- 4) To identify gaps in my initial study design (e.g. this led me to add baselines to assess whether LLM monitors offered an advantage over simpler methods)
- 5) To modify the provided game code