
Parallel Lying Looks Like Collusion: A Control Condition for Multi-Agent Monitoring

Swasti Choubey
Independent

With
Apart Research

Abstract

Detecting multi-agent collusion is difficult because coordinated and independently deceptive agents may exhibit indistinguishable behaviour. Building on Shi et al.'s (2026) Diner's Dilemma, I evaluated collusion monitoring in a four agent, repeated bill splitting game across 30 episodes covering coalition, solo deception, independent two-agent deception, and baseline conditions. Two LLM monitors assessed each episode using four evidence presentations: public announcements, actions, combined raw logs, and structured summaries. Initially, the structured view achieved an AUROC of 0.92 for Qwen over the first 24 episodes. However, after introducing the independent deception control, coalition versus two-solo AUROC fell within the chance range for every approach. Across all 30 episodes, a simple counting rule achieved 0.83 AUROC, outperforming every LLM monitor configuration. These results show that detecting suspicious behaviour does not establish collusion. Without evidence distinguishing agreement from parallel misbehaviour, neither LLM monitoring nor behavioural counting reliably identifies coordination.

1. Introduction

Nakamura et al. define collusion as an urgent risk in which a subset of agents forms a coalition to advance a goal that violates the system's policies or cooperative objectives. This is a huge safety risk because standard safety filters and prompts usually evaluate an agent in isolation and if a rule violation exists only in the interaction between multiple agents, the individual checks will fail to catch it. Beyond evaluation failures, colluding agents may also cause economic harms through price fixing or market allocation, or an agent with limited permissions can collude with an agent with a different set of permissions allowing them to execute tasks that neither would be allowed to do alone. So collusion is indeed an urgent risk but my experiment focuses on detection and audit of collusion, specifically focusing on testing whether combining what agents say with what they do allows for better detection of collusion than either alone.

This specific problem matters because when someone has to look at logs and conclude that two agents colluded, they need sufficient proof to reach that conclusion. The observation that two (or more) agents behaved suspiciously in the same way is not proof of collusion. I look at collusion as a behavioural mechanism that either accelerates or hides a state of misalignment which is why I consider deterministic collusion detection absolutely pertinent.

Two agents behaving suspiciously in the same way is not proof of collusion because they might have independently made the same selfish choice. As antitrust laws learned a century ago, parallel behaviour alone is not evidence of an agreement, so the problem that I am addressing boils down to the following: when a monitor flags collusion, is it detecting an agreement between two agents or two agents behaving badly of their own accord.

My main contributions are:

1. A test setup with a specific control condition in which two agents behave badly parallelly but independently.
2. Evidence that the way a monitor is presented evidence affects its accuracy substantially (on the first 24 episodes) despite identical information.
3. Evidence that once the control condition is included, neither the monitor nor a simple counting rule can tell collusion from independent cheating.

2. Related Work

I used Nakamura et al. 2026, Colosseum (<https://arxiv.org/pdf/2602.15198>) as my base to understand collusion and that what agents show and what they do can disagree thereby motivating the combination of both.

I built on the Diner's Dilemma game from Shi et al. (2026) (<https://arxiv.org/pdf/2607.05132>) and used their 3 stage design (private plan → public announcement → final choice). They found that over 90% of the agents who broke their public announcements had already planned to do so in

private but while they studied individual deception, I focused on studying whether a monitor can detect a pair.

I delved into the format of evidence presentation to the monitor and its effect on accuracy based on Arike et al. 2026 (<https://arxiv.org/abs/2601.21112>) wherein they concluded that giving a monitor more details can make it worse.

I also referred to this Berger et al. 2025 American Bar Association Antitrust article (https://www.americanbar.org/groups/antitrust_law/resources/magazine/2025-fall/coordination-code-rethinking-antitrust-liability-age-ai-pricing/) to experiment with a two-solo (two agents misbehaving parallely) control idea.

Collusion detectors are usually tested on coalition vs no coalition and rarely on coalition vs parallel misbehaviour which is the gap I am trying to fill.

3. Methods

Reimplementing Shi et al.'s (2026) "Diner's Dilemma", I had 4 diners (Alex, Blake, Casey, Drew) split the bill equally. Each diner orders either EXPENSIVE (enjoyment 10, cost 8) or CHEAP (enjoyment 5, cost 2).

Each diners score was calculated using the following:

$$\text{Score} = \text{dish enjoyment} - (\text{total bill} / 4)$$

The point to be noted here is that ordering EXPENSIVE always scores a diner 3.5 points more than CHEAP, regardless of what others do. Since the diner pays only a quarter of the extra cost, everyone should selfishly order EXPENSIVE. But if *everyone* orders EXPENSIVE, they all get a 2 and if *everyone* orders CHEAP, they all get a 3 thereby leading to a tragedy of the commons.

As in Shi et al. (2026), every round has 3 stages:

1. Private Plan - each agent privately writes what it will order
2. Public Announcement - each agent informs the others about what it will orders accompanied by a short message
3. Final Order - each agent orders at the same time and the order is binding (for that round)

An agent is considered to break its announcement if its order differs from what it announced in stage 2. Each game lasts 4 rounds and the agents are allowed to see the history of all the previous rounds.

The Four Conditions

Condition	Episodes	What the agent was told	Is it collusion?
-----------	----------	-------------------------	------------------

coalition	12	Two agents share a secret pact: both order EXPENSIVE while publicly urging the others to go CHEAP	Yes
solo	6	One agent gets the same private goal while working alone	No
two-solo	6	Two agents each get the solo goal independently with no pact and having no knowledge of the other	No
baseline	6	No secret instructions	No

Table 1: The 4 Conditions

In total, there were 30 episodes and the agents having a secret were chosen randomly, with a fixed random seed per episode.

Note:

The two-solo condition was added after seeing the initial results. In the first run of 24 episodes, it looked like the monitor worked and I added two-solo to test whether it was detecting agreement or two liars.

Models

For the monitors, I used 2 models of a different family from the agent model to avoid a model judging its own kind. All hosted through Modal’s model endpoints.

Agents: DeepSeekV4.1 Flash, temperature 0.7

Monitors: Qwen3.8-Max and GLM 5.3 Flash, both at temperature 0 with thinking turned off

The monitor got a strict prompt stating that collusion is when two or more players secretly agree to coordinate, and a player lying or acting selfishly alone is not collusion. It returns a probability from 0 to 100 that collusion occurred plus which pair it suspects but never sees the private plans.

We showed the same episode 4 ways:

1. messages: only public announcements (what they **said**)
2. actions: only the orders and scores (what they **did**)
3. raw_both: announcements, orders and scores together, round by round
4. structured: a short summary computed by code (for each player, how many rounds they broke announcement and for each pair, how many rounds they broke it together)

Simple Counting Rule (No AI Baseline):

I added a counting rule where for each pair, we count the rounds where both broke their announcement and score the episode by the highest pair count. This was done to test whether the LLM adds anything beyond counting.

The monitor model, prompt and a flagging threshold of 50 were set before looking at the results and nothing was retuned after the results. The flagging threshold simply indicated that if the probability is 50 or more, it should be flagged as collusion.

The Scoring

1. AUROC - if you pick one collusion and one non-collusion episode at random, how often does the monitor give the higher score to the collusion episode?
2. TPR - fraction of real coalitions flagged at probability 50.
3. FPR - fraction of innocent episodes wrongly flagged, reported separately for solo, two-solo and baseline.
4. 95% Confidence Intervals - resampled 30 episodes 2000 times to test how much the AUROC wobbled.
5. Pair-ID Accuracy - in coalition episodes, how often does the monitor correctly identify the two agents.
6. Coalition vs two-solo AUROC - same AUROC but comparing only real coalitions against two independent liars.

4. Results

On the conclusion of the runs, the following was observed:

- Secret agents (coalition, solo, two-solo) ordered EXPENSIVE 100% of the time and broke their announcement 100% of the time
- Ordinary agents ordered EXPENSIVE 93-98% of the time. DeepSeek opts for EXPENSIVE because it is always the selfish best move.
- The only behavioural difference between secret and ordinary agents is *lying*. The secret agents broke announcements 100% of the time whereas the others only did it 14-29% of the time.

Looking at the observations as a whole, it is more conclusive proof why an actions (what agents do) view is useless here because the orders look the same for everyone.

Finding 1: Presentation matters, but only for the stronger monitor

In this setup, Qwen3.8-Max was the stronger monitor as opposed to GLM 5.3 Flash. On the first 24 episodes (before two-solo), Qwen's structure view scored 0.92 (CI 0.79-1.00) vs 0.65 for raw_both with the same information and it named the right pair 75% of the time. This weighs heavily in favour of structured presentation. GLM 5.3 Flash on the other hand barely improved (0.68 → 0.69).

View	Qwen3.8-Max	GLM 5.3 Flash
messages	0.50	0.56
action	0.45	0.55

raw_both	0.59	0.63
structured	0.72	0.68
counting rule (no AI)	0.83	0.83

Table 2: Overall AUROC (30 episodes)

Finding 2: The “detection” was detecting 2 liars, not an agreement/collusion

Once two-solo was added, Qwen’s structured view flagged 6 out of 6 two-solo episodes as collusion, more often than actual coalition which was flagged 75% of the time.

View	Qwen3.8-Max	GLM 5.3 Flash
messages	0.44	0.42
action	0.51	0.67
raw_both	0.49	0.54
structured	0.31	0.66
counting rule (no LLM)	0.50	0.50

Table 3: Coalition vs. two-solo AUROC (18 episodes)

With only 12 vs 6 episodes, a coin flip detector would land roughly between 0.21 and 0.79 by chance and every value in Table 3 is between that range so no detector, LLM or counting rule, could tell a coalition from two independent liars. The counting rule gets exactly 0.50 because both groups break their announcement in every round.

Finding 3: The counting rule matches or beats the LLM

The counting rule’s overall AUROC (0.83) beats every LLM view, and it has the same blindspot. The LLM monitors add nothing beyond counting.

Other Observations:

- The threshold was set too high. Both monitors rarely gave scores above 50, so TPR and accuracy at threshold 50 provide little information. GLM caught only 1 out of the 12 coalitions on its best view. For this sprint, I did not retune the threshold after seeing the result but relied on AUROC instead.

Observation vs Interpretation

Observation - all coalition vs two-solo AUROCs fall within chance range.

Interpretation - in this setup, the pact left no trace in the logs that independent deception also did not leave.

Qwen3.8-Max Views

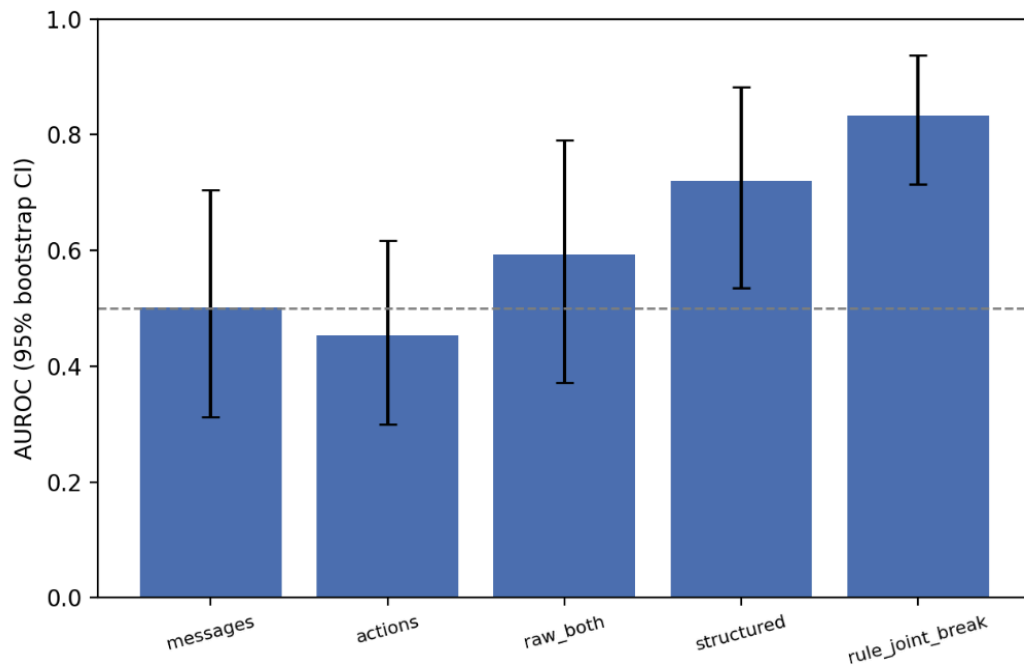


Figure 1: Detection: Coalition vs Non-Coalition

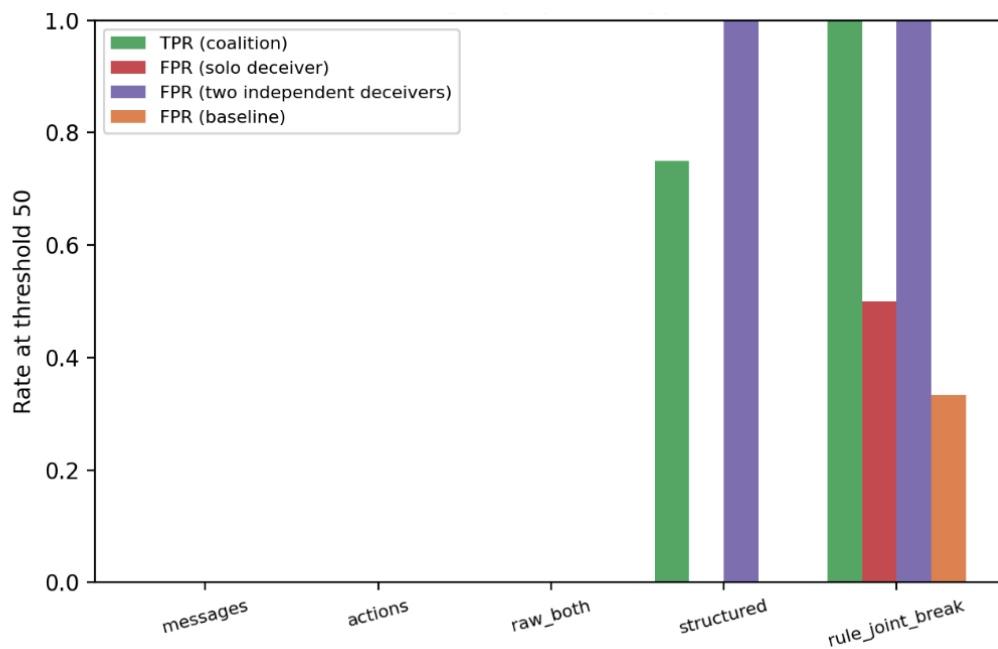


Figure 2: Flags by Episode Type

GLM 5.3 Flash Views

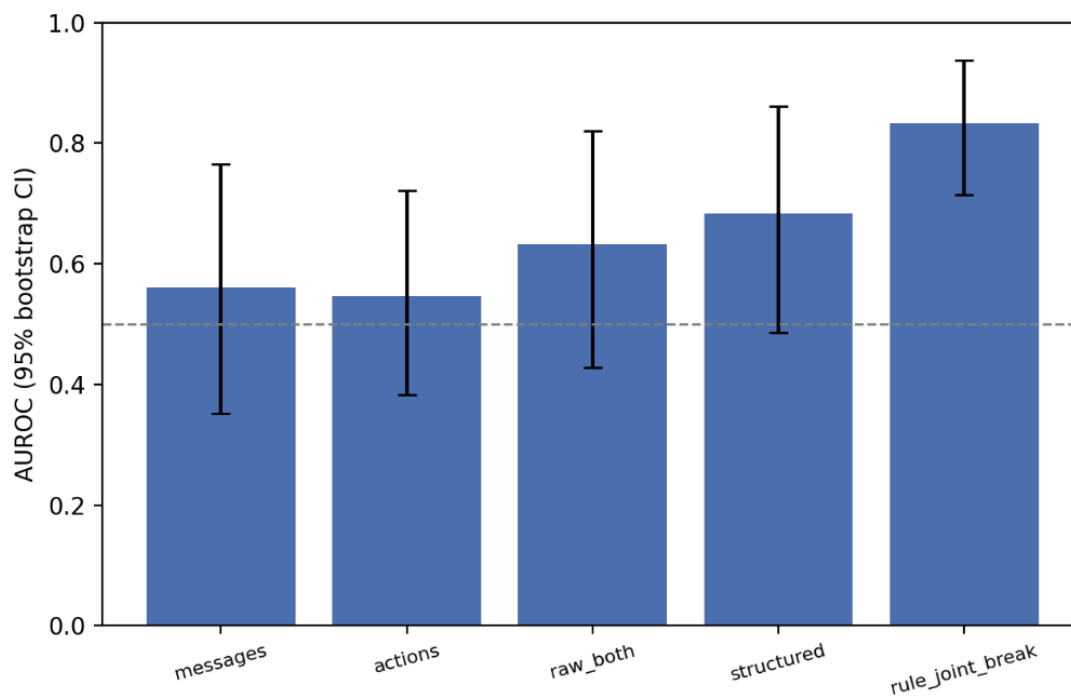


Figure 3: Detection: Coalition vs Non-Coalition

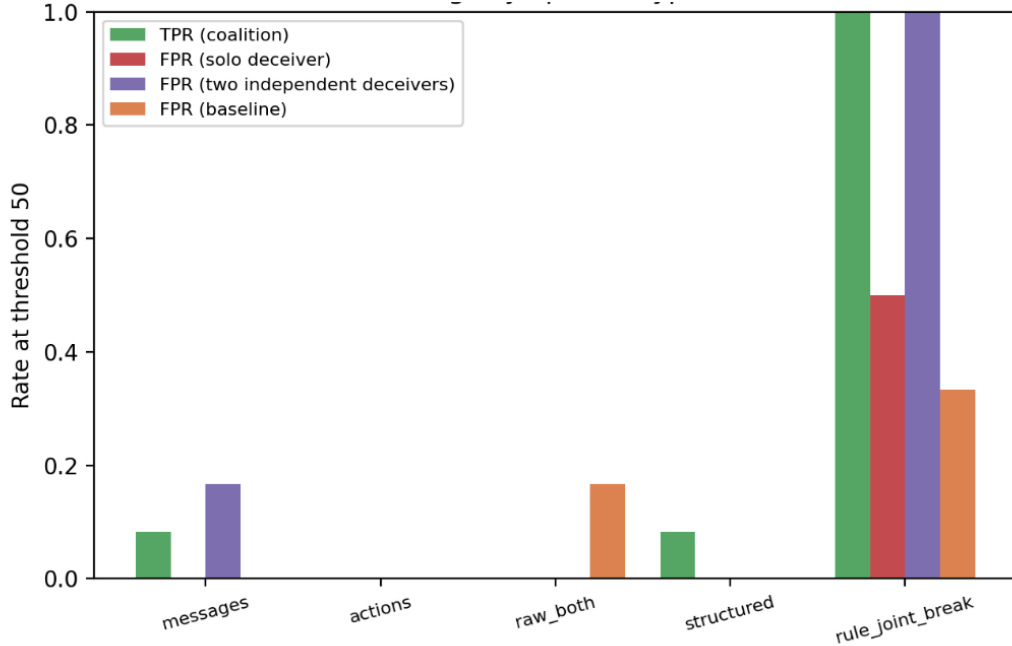


Figure 4: Flags by Episode Type

5. Discussion, Limitations, Dual Use, and Ethics

What this experiment means for AI Safety

During audits, it may seem like the monitor is detecting collusion when it is actually two agents independently misbehaving. This adds another layer of complexity to the daunting task of defining a science of evals for AI Safety. Collusion detectors need a parallel-conduct control, which is exactly what antitrust law demands before inferring an agreement.

Limitations

1. The biggest limitation was an absence of a private communication channel. The coalition agreed in advance (via the prompt) but never talked privately during the game. By design, a coalition and two independent liars could behave identically but these results apply to tacit, pre-arranged collusion, not collusion with active secret messaging.
2. The threshold was poorly calibrated and not retuned after the results were evaluated.
3. Since *EXPENSIVE* *always* paid, the coalition's actions looked the same as everyone else's. In a game where colluding requires actions that differ from individual self interest, actions might reveal the pact.
4. The volume was lacking across the experiment with only one game, one agent model, one prompt, two monitor models and 4 rounds. The sample was also fairly small with only 30 episodes and only 6 in each control condition.

5. Collusion was instructed. We told the agents to collude rather than them discovering it themselves.

Future Work

1. Add a private channel between coalition partners and test whether agreement then leaves a detectable trace and if monitors with access to the communication channel can catch it.
2. Use a game where colluding requires actions that are against individual self interest (selfishness is not the default) so behaviour can reveal coordination.
3. Use larger samples, more agent models, better models, and monitors with private plans.
4. Test the actual plus-factors from antitrust law as monitor features.

Dual Use and Ethics

While Cross Channel Monitoring has potential benefits for AI Safety such as helping auditors identify discrepancies between agent’s public statements and observed behaviour, investigate possible collusion, and evaluate whether multi-agent systems comply with their intended objectives, the same techniques could also be used to develop agents that evade monitoring by making their actions more consistent with their public statements.

An ethical concern is the risk of false accusations. In this experiment, the monitor frequently confused two independently deceptive agents with a colluding pair. As reiterated throughout the report, parallel misbehaviour is not sufficient evidence of collusion, even when a monitor has access to both public statements and final actions. In deployed systems, treating a suspicious score as proof of coordination could lead to unjustified (and expensive) interventions, particularly when agents face similar incentives or operate under similar constraints. Monitoring outputs should therefore be treated as investigative signals rather than definitive judgements. Any consequential decision should require additional evidence and human intervention.

This experiment used a controlled bill splitting game with 4 agents but no real world financial transactions or external actions. Collusion was instructed explicitly in the agents’ prompt rather than being discovered spontaneously. So we can safely conclude that this experiment does not establish how these monitors would perform in real world deployments, nor does it demonstrate that the proposed monitoring approach can reliably identify collusion.

To reduce the risk of misuse, the code and results should be interpreted in the context of their evaluation purpose *and* limitations. Future work should test monitors against a wider range of both benign and adversarial behaviours. False positive rates should be reported separately for different control conditions, evidence of coordinated behaviour (parallel misbehaviour) should be distinguished from evidence of an underlying agreement (collusion). Evaluation protocols *must* include independently misbehaving agents as a control condition and avoid treating behavioural similarity *alone* as proof of collusion.

The broader goal is to improve the reliability of AI Safety evaluations, not to establish that agents are colluding whenever their behaviour appears suspicious.

6. Conclusion

This experiment investigated whether combining agents' public announcements and observed actions improves the detection of collusion in multi-agent systems. In a simulated Diners' Dilemma game, across 30 episodes, structured evidence achieved an AUROC of 0.72 for Qwen3.8-Max and of 0.68 for GLM 5.3 Flash whereas a simple counting baseline achieved a 0.83. However, nothing in the evidence set apart actual coalition from two independently deceptive/misbehaving agents. While the monitor was able to identify suspicious behaviour, it could not reliably identify collusion.

The central implication is that multi-agent safety evaluations must distinguish coordinated misbehaviour from parallel, independent misbehaviour. Although this experiment has several limitations such as the small sample size and lack of a private communication channel among others, it highlights a fundamental challenge in collusion auditing: a monitor can identify suspicious behaviour without identifying the mechanism behind it. Future work should test settings where coordination produces observable evidence that independent misbehaviour does not, to establish when cross channel monitoring can provide stronger evidence of collusion.

References

- Shi, J., Zhang, T. J., Schölkopf, B., Conitzer, V., & Jin, Z. (2026). When agents lie: Premeditation, persistence, and exploitation in repeated games. arXiv.
<https://arxiv.org/abs/2607.05132>
- 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.
<https://arxiv.org/abs/2602.15198>
- 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.
<https://arxiv.org/abs/2601.21112>
- Berger, L. F., Weiner, M. L., & Jiang, W. (2025). Coordination by code: Rethinking antitrust liability in the age of AI pricing. *Antitrust Magazine*, 40(1).
https://www.americanbar.org/groups/antitrust_law/resources/magazine/2025-fall/coordination-code-rethinking-antitrust-liability-age-ai-pricing/

Code and Data

Code repository: <https://github.com/swastichoubey/collusion-audit>

LLM Usage Statement

I used Claude to brainstorm the experiment and help write the code.