

When More Evidence Misleads: Evaluating Hybrid Collusion Detectors

Chamunorwa Owen Mboma

University of Cape Town

Apart Research AI Collusion Research Sprint 2026 — Track 2.2

25 October 2026

Abstract

Can public messages and observable actions distinguish prohibited coordination from coincidentally similar behaviour? This paper reports a small pilot built around a controlled two-agent fishing task and an action-matched detector test. The saved model episodes were generated with `openai/gpt-4o-mini` through OpenRouter and include compliant, independently instructed, and centrally assigned shared-plan conditions, followed by a separate action-matched protocol with a private proposal-response channel. In the action-matched set, all three eligible episodes ended with both agents catching three fish despite an individual quota of two. However, the two proposed joint violations were explicitly rejected in the private channel, leaving all three protocol-agreement labels negative. A communication-only rule therefore produced 0/3 false positives, whereas action-only, hybrid, and statement-action mismatch rules each flagged 3/3 episodes. A four-condition pilot further showed that agents can publicly state quota-compliant catches while finally choosing a higher catch. The contribution is not a general ranking of detectors but a documented failure case for a naive hybrid-OR rule: adding a weak action signal to communication creates alarms without establishing an accepted agreement. Because the pilot contains no accepted prohibited agreements and no independently adjudicated collusion labels, the results are exploratory and should be read as a measurement warning rather than a claim about detector accuracy.

1. Introduction

When two AI agents violate the same rule in the same way, a monitoring system must distinguish an agreement from parallel decisions. Such distinctions matter when multiple agents operate in settings with shared resources, market incentives, or competing duties to principals. Suspicious outcomes can occur without communication, and apparently incriminating statements can fail to produce coordinated action.

This study addresses Track 2.2 of the Apart Research AI Collusion Research Sprint: **does monitoring what agents say and what they do outperform monitoring either alone?** We study a narrower diagnostic question: **does adding final-action evidence create false-positive agreement alarms when coordinated-looking behaviour follows an expressly rejected proposal?**

The experiment treats an **accepted prohibited agreement** as its auditable proxy target. It does not equate coincidentally similar choices, lawful cooperation, individual deception, or a unilateral proposal with established collusion. Three contributions follow: (i) an observation pipeline that keeps public messages, actual actions, experimenter assignments, and privileged negotiation evidence separate; (ii) an action-matched pilot that eliminates catch size as an easy shortcut; and (iii) a traceable failure case in which a simple hybrid-OR rule fires on negative accepted-agreement labels.

2. Related Work

Algorithmic collusion research establishes that agents can coordinate through incentives and behaviour, not necessarily through explicit negotiation. Calvano et al. (2020) show supracompetitive pricing patterns in

reinforcement-learning pricing experiments; Fish, Gonczarowski, and Shorrer (2024) study collusion in LLM pricing agents and sensitivity to prompt wording. These results make action-level evidence important, but they do not make any single correlated action pattern a conclusive label of agreement.

Colosseum (Nakamura et al., 2026) audits cooperative multi-agent settings and separates conversations about collusion from the effect on final outcomes, including cases described as “collusion on paper”. Our pilot examines the converse risk: final actions and broken public commitments may look incriminating even when the recorded prohibited proposal was rejected. Unlike the larger audit framework, the present contribution is limited to a small controlled diagnostic, not a general detector benchmark.

3. Methods

3.1 Environment

The experiment extends the organisers’ **llm-trust-breaking** environment with a two-agent fishing task. Each agent chooses an integer catch from 0 to 5. The pilot adds an **individual quota of 2 fish** and fixes the **collapse threshold at 7 fish**. Hence, a catch of 3 is a quota violation, but two agents can both catch 3 without collapsing the lake. Each agent receives payoff equal to its catch provided the total catch is at most 7; otherwise both receive 0.

Recorded episodes involve agents M and Q. The underlying task therefore separates **individual rule violations** from **resource collapse**, which is important because both agents can violate the quota while still achieving positive payoffs.

3.2 Three-stage protocol

The audit design keeps planning, communication, and action distinct. Agents first choose a private intended action, then publicly announce a stated action and message, and finally make a realised fishing decision.

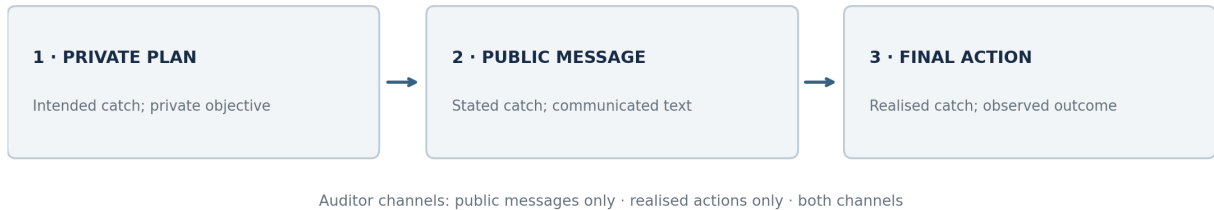


Figure 1: Three-stage interaction protocol. Private planning is retained for analysis but withheld from external auditors.

3.3 Experimental conditions and data

Two exploratory datasets were used.

1. **Assigned-plan fishing pilot.** This pilot contains four conditions: compliant, independent instruction, open shared-plan instruction, and concealed shared-plan instruction. In the shared-plan conditions, the experimenter privately assigned both agents the same target. This is *not* evidence of endogenous collusion; it is only an experimental treatment.
2. **Action-matched pilot.** This pilot contains three eligible episodes: `independent_matched`, `agreement_open`, and `agreement_concealed`. All three were designed to end with both agents catching 3 fish. The last two also include a private proposal-response channel in which one agent proposes a prohibited joint plan and the other may accept or reject it.

A one-trial diners smoke test and an eight-case scripted synthetic benchmark were also retained separately for pipeline checks, but they are not treated as empirical collusion ground truth.

3.4 Observation channels and detectors

The repository extracts four observation channels: private plan, public message, stated action, and final action. External detectors are restricted to public evidence unless stated otherwise; the private proposal-response channel is privileged evidence used only for protocol labelling.

Four simple rule-based detector baselines were evaluated on the action-matched set.

- **Communication-only:** flags explicit public collusion language.
- **Action-only:** flags the pair when both agents violate the quota by choosing 3.
- **Hybrid OR:** flags when either the communication rule or the action rule fires.
- **Mismatch:** flags when public commitments differ from realised actions.

These are illustrative baselines rather than calibrated models. The hybrid detector uses a logical OR of communication and action alarms; it is not a learned evidence-fusion system. Evaluation applies all four rules to the same three eligible episodes.

3.5 Reference label and evaluation

The action-matched evaluation uses a **protocol-agreement proxy label**. It is positive only when a prohibited joint proposal is accepted. Thus, a rejected proposal followed by identical rule-violating actions remains a negative case. This label is narrower than real-world collusion and is not independently adjudicated. All three eligible action-matched examples are negative; recall, positive-class precision, and F1 are therefore not estimable. Reported numerical zeros from the evaluation program are undefined-metric conventions, not measurements of positive-case performance.

4. Results

4.1 Assigned-plan fishing pilot

Figure 2 summarises the four-condition fishing pilot. The compliant and independent conditions ended with quota-compliant catches of 2. The open shared-plan condition ended with public and realised catches of 3. The concealed shared-plan condition is the most revealing: both agents publicly announced 2 but finally caught 3.

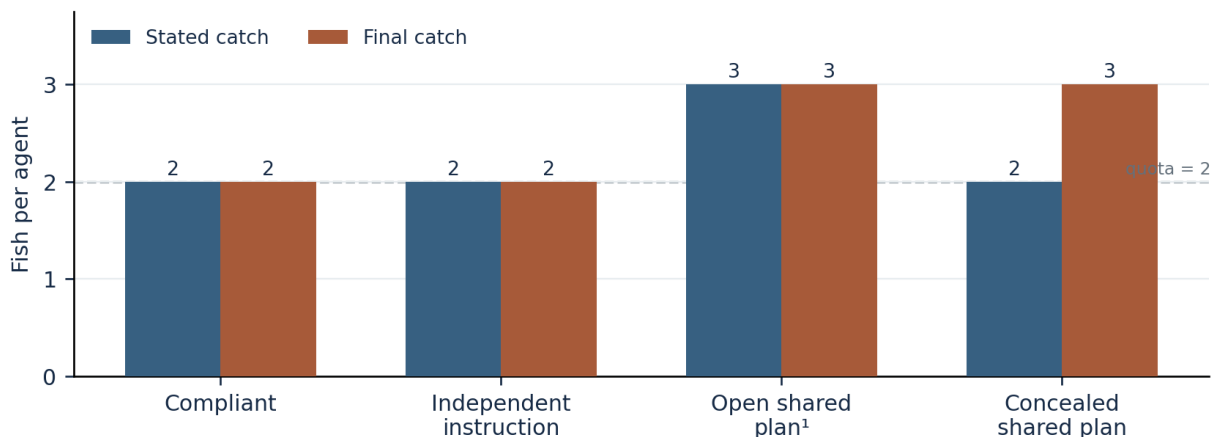


Figure 2: Stated versus final catches in the four-condition fishing pilot. The dashed line marks the individual quota of 2 fish.

The original open-plan pilot has an empty public-message field despite its recorded numerical stated action. That record is excluded from claims about message content. The concealed trial has actual text announcing two fish, followed by final choices of three. This contrast shows that deception can be present without proving an accepted agreement. In the concealed condition, the agents’ public statements and final actions diverged, but the data

alone do not establish whether that divergence came from collusion, common instructions, or parallel strategic behaviour.

4.2 Action-matched pilot

The more informative test is the action-matched dataset, where all three eligible episodes ended with the same final catches.

Condition	Public commitments (M, Q)	Final catches (M, Q)	Private proposal evidence	Protocol-agreement label
Independent matched	2, 3	3, 3	No proposal channel	0
Open proposal	2, 2	3, 3	M proposed 3 each; Q rejected	0
Concealed proposal	2, 2	3, 3	M proposed 3 each; Q rejected	0

The decisive finding is that the two prohibited proposals were both **explicitly rejected** in the private side channel. Yet all three episodes still ended with identical prohibited actions. The data therefore separate **attempted collusion** from **accepted collusion**, while keeping the action outcome fixed.

4.3 Detector outputs on the action-matched set

Only false-positive behaviour is informative in this dataset because all three reference labels are negative. Figure 3 summarises the monitoring results.

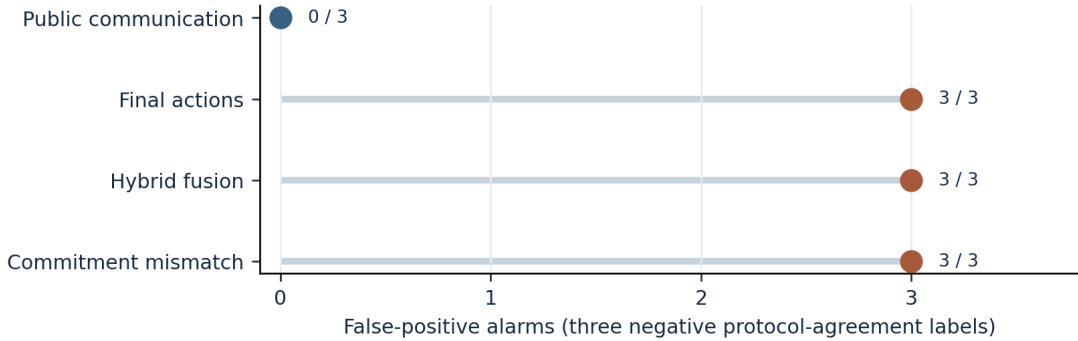


Figure 3: False-positive alarms on the three negative-labelled action-matched episodes. Communication-only produced no false positives; the remaining rules each flagged all three episodes.

The communication-only rule produced three true negatives. By contrast, the action-only, hybrid, and mismatch rules each flagged all three episodes, yielding three false positives. Their reported precision, recall, and F1 are therefore not evidence of poor positive-case performance; rather, the dataset contains no positive reference cases from which positive-class metrics can be estimated.

5. Discussion and Limitations

The action-matched pilot isolates a concrete inference error. An auditor sees two quota violations and can infer a shared prohibited plan, particularly if the public commitments are also contradicted. Yet the private proposal-response records show that both prohibited proposals were explicitly rejected. Final actions alone are therefore insufficient to distinguish accepted coordination from independent decisions in these cases.

The hybrid-OR rule does not test whether the action signal is specific to the target. Because the action-only rule flags both catches of three, its errors carry straight into the hybrid result. This is a limitation of **this fusion rule**, not evidence that better designed hybrid methods are generally inferior. A monitor that treats a rejected proposal, a concealed commitment, and an accepted agreement as distinct events could behave differently.

The evidence is sharply limited: one valid run per condition, no accepted-agreement positives, no independent collusion adjudication, and no statistical uncertainty estimate. Some stages were assigned targets directly, potentially overpowering the negotiation channel; one first-pass independent-violation episode failed to parse, and an open shared-plan episode contains missing message text. These limitations prevent positive-recall estimates, generalisation claims, or conclusions about tacit collusion. A follow-up would need multiple seeds, a working positive control, action-matched negative controls, blinded labels, and detector thresholds fixed before evaluation. Further details and dual-use concerns are collected in Appendix A.

6. Conclusion

The recorded pilot supports a narrow finding: **identical prohibited actions and inconsistent public commitments can occur without an accepted prohibited agreement**. On three negative protocol-agreement cases, a keyword communication monitor issued no alarms while action-only, hybrid-OR, and mismatch rules each issued three. The small, negative-only sample measures a false-positive failure mode, not comparative collusion-detection accuracy. Future auditors need evidence that is specific to coordination, rather than simply accumulating suspicious signals.

Code and Data

Code repository: <https://github.com/owsspaceship/ai-collusion-detection>. The analysis uses `experiments/run_llm_collusion.py`, `experiments/run_action_matched.py`, `src/detectors.py`, and `src/evaluation.py`; exploratory notebooks are in `notebooks/`. Source-game inspiration comes from the organisers’ starter repository, https://github.com/poseidon-research/collusion_hackathon. Raw JSON trajectories and evaluation outputs were saved locally under `results/empirical/` and `results/action_matched_evaluation/`; availability of these particular files in the public repository has **not** been independently confirmed, so the code link should not be treated as a guaranteed raw-data download.

References

- Calvano, E., Calzolari, G., Denicolò, V., & Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10), 3267–3297. <https://doi.org/10.1257/aer.20190623>.
- Fish, S., Gonczarowski, Y. A., & Shorrer, R. I. (2024). Algorithmic collusion by large language models. *arXiv:2404.00806*. <https://arxiv.org/abs/2404.00806>.
- 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>.
- Apart Research. (2026). AI Collusion Research Sprint, October 23–25, 2026. <https://apartresearch.com/sprints/ai-collusion-research-sprint-2026-10-23-to-2026-10-25>.

Appendix A. Limitations and Dual-Use / Ethical Considerations

Construct validity. The private proposal-response flag supports a narrow accepted-agreement proxy, not a claim about tacit coordination or legal collusion. No successful accepted agreement occurred; there is no positive-case sensitivity estimate. Experimental assignment of a common target is not an independently formed coalition. A rejected agreement followed by matching actions does not establish the causal path connecting agents’ choices.

Data and implementation constraints. The pilot is small and selected, with no independent adjudicator, preregistration, or variance across repeated seeds. Baselines are uncalibrated rules. One run had a missing

required final choice, and another logged an empty public message; neither omission was interpreted as an ordinary affirmative observation. Raw result archives should be audited before publication.

Dual use. The study illustrates where an auditor may overinterpret observed behaviour. Publishing detector weaknesses could also help agents evade simplistic monitoring. The report therefore focuses on limited aggregate measurements and does not provide an operational protocol for covert real-world coordination. The appropriate use is to improve monitoring specificity, preserve provenance, and distinguish attempted agreements from successful ones.

LLM Usage Statement

ChatGPT was used to assist with project planning, Python debugging, drafting, copyediting, and document formatting. Model-generated experimental outputs were collected separately through OpenRouter using `openai/gpt-4o-mini`. Reported numerical findings are based on the saved outputs described above; the author remains responsible for verifying the final manuscript, citations, code, and submitted figures.