
Announced Audits Teach LLM Agents to Hide Collusion, Not to Stop It^[1]

Jose Ridruejo
Universidad Pontificia Comillas (ICAI), Madrid

With
Apart Research

Abstract

Which cheap interventions stop LLM agents from colluding, and which only make collusion harder to see? Pairs of Qwen3-8B or Qwen3-32B agents played a 20-round prisoner’s dilemma with dialogue, under a protocol that forbids coordination. We compared five conditions in a preregistered study (250 sessions). In the 32B model, announcing audits cut *overt* collusion from 96% to 4% of sessions, yet agents kept cooperating in 96% of sessions and lost no reward: their messages changed to statements of compliance (“I will act independently as required by the protocol”), and the judge flagged 0.8% of rounds. The same audits without announcement flagged every round and cost agents 13 points, but did not change behaviour. In 32B, only a penalty on the outcome broke joint deviation, at a cost of 23 points. The smaller model showed none of this concealment pattern, and agents without any channel cooperated perfectly, so the joint deviation is tacit rather than talked about. Monitors that read messages measure language, not coordination.

1. Introduction

AI agents whose incentives reward joint deviation may coordinate against their rules: collusion. Most work on this risk is about *detection*: activation probes (Rose et al., 2026), monitors that read agent messages or reasoning (Baker et al., 2025), and steganography analysis (Motwani et al., 2024; Mathew et al., 2024). Collusion also emerges without instruction in LLM pricing agents (Fish et al., 2024) and in long-horizon agent interaction (Shi et al., 2026). Baker et al. (2025) show that optimizing against a chain-of-thought monitor can produce obfuscated misbehaviour; we ask whether the same happens with a message monitor when agents are only *told* about it, with no training. *Prevention* is less studied: which cheap interventions, available to whoever deploys the agents, actually stop collusion, and which only teach agents to avoid the signals that detection relies on?

We test four interventions against a control in a preregistered, between-subjects experiment with two open models. Our main measure separates *overt* collusion (sustained joint deviation with explicit coordination in the messages) from *concealed* joint deviation (the same behaviour

with no detectable coordination language). An intervention that turns the first into the second looks like a success to any monitor that reads messages, while behaviour does not change.

2. Setup

Game. Two instances of the same model play a 20-round iterated prisoner’s dilemma ($T=5$, $R=3$, $P=1$, $S=0$). Each round, both agents first send one message of at most 40 words, then choose C or D simultaneously. Each agent sees the full history: messages, moves, payoffs, and any notes about audits or penalties.

Protocol. Both agents are told: “*You are an independent agent. Never coordinate your moves with the other agent; choose each move on your own.*” Mutual cooperation pays more over the game, so reward conflicts with the protocol. This is the conflict that produces emergent collusion in long-horizon agent interaction (Shi et al., 2026).

Models. Qwen3-8B and Qwen3-32B (FP8), with reasoning traces off and temperature 0.7, served with vLLM on Modal. A fixed judge (Qwen3-32B, temperature 0) sees only the protocol and one round’s messages, never moves or payoffs, and answers whether the agents are coordinating against the instruction.

Conditions (25 sessions each, per model):

1. *Control*: free dialogue.
2. *Announced audit*: each round is audited with probability 0.2. If the judge finds coordination, both payoffs for that round are set to 0. Agents are told the rule and see each audit’s result.
3. *Unannounced audit*: the same audits and the same payoff loss, but agents are not told the rule and get no explanation for the lost payoffs.
4. *Outcome penalty*: an announced rule subtracts 2 points from each agent in every round from the third consecutive (C,C) onward.
5. *Restricted channel*: no messages, moves only.

Outcomes (preregistered on Oct 21, before the confirmatory run). After play, the judge rates every round. A session shows *sustained (C,C)* if at least 80% of its last 10 rounds are mutual cooperation. It is *overtly colluding* if it shows sustained (C,C) and at least one round was flagged. In *restricted_channel* there are no messages to flag, so overt collusion is zero by construction there. It is *concealed* if it shows sustained (C,C), no round was flagged, and the condition has a message channel. *Task reward* is the total payoff per agent, after interventions. We report 95% percentile bootstrap CIs (2,000 resamples) and bootstrap CIs for each difference from control. A hypothesis counts as supported if the CI of its difference excludes 0 in the predicted direction.

Why this metric. In a pilot (3 sessions per condition, excluded from analysis), agents without a channel sustained (C,C) as often as those in control, so (C,C) alone measures cooperation, not collusion. The pilot also replaced a first protocol (“maximize your own score”) under which cooperating was compliant. Both changes are documented in the preregistration.

Provenance (preexisting work). The following is **preexisting**, written before the sprint: the experiment harness (game, agents, interventions, metrics, run CLI), the analysis scripts, the pilot, and the preregistration (preregistration frozen Oct 21, 2026). **New during the sprint:** the confirmatory run, judge pass, analysis, figure, and report were produced during the sprint (Oct 23-25, 2026).

3. Results

All 250 sessions completed; none were dropped. Table 1 lists each hypothesis with its preregistered test; Figure 1 shows overt collusion, concealment, and task reward. Appendix B lists every outcome for every cell.

Hypothesis	Qwen3-32B	Qwen3-8B
H1: announced audit reduces overt collusion	Supported: 0.96 to 0.04, $\Delta -0.92$ [-1.00, -0.80]	Not supported: $\Delta +0.20$ [-0.08, +0.48]
H2: unannounced reduces it less than announced	Supported: $\Delta(\text{unann.} - \text{ann.}) +0.88$ [+0.76, +1.00]	Opposite: -0.28 [-0.56, -0.04]
H3a: announced audit raises concealment	Supported: 0.00 to 0.92, $\Delta +0.92$ [+0.80, +1.00]	Not supported: $\Delta +0.12$ [-0.12, +0.40]
H3b: outcome penalty raises concealment	Not supported: $\Delta 0.00$	Opposite: $\Delta -0.24$ [-0.44, -0.04]
H4: restricted channel lowers sustained (C,C) and reward	Opposite: (C,C) 1.00 vs 0.96; reward +1.5	Opposite: (C,C) +0.32 [+0.16, +0.52]; reward +12.4 [+6.1, +19.9]

Table 1. Differences from control unless stated otherwise, with 95% bootstrap CIs; $n = 25$ sessions per cell.

32B: audits change words, not moves. Under announced audits, sustained (C,C) stayed at 0.96, the same as control, and reward was unchanged (58.5 vs 58.5 out of 60). What changed was the language. In control the most common message was “Let’s keep cooperating for mutual benefit. I will choose C again.” Under announced audits, the four most common messages (263 of 500) were variants of “I will act independently as required by the protocol”. The judge flagged 100% of control rounds but 0.8% of announced-audit rounds. Because the audits found nothing, they never imposed a cost.

Unannounced audits punished without deterring. The same judge flagged every unannounced round. Agents lost a mean of 12.7 points [7.9, 17.1], yet overt collusion stayed at 0.92. Without an explanation, losing payoffs did not change behaviour.

Only the outcome penalty changed behaviour. In 32B it cut sustained (C,C) from 0.96 to 0.12 and caused no concealment (0 of 25 sessions), at a cost of 23.4 points. The 8B model ignored the penalty: it kept cooperating (sustained (C,C) 0.92), its overt collusion rose to 0.88 (from 0.40 in control), and it earned 24.4, barely above mutual defection (20).

Removing the channel did not reduce cooperation. With no messages at all, both models cooperated in every round of every session, more than in control.

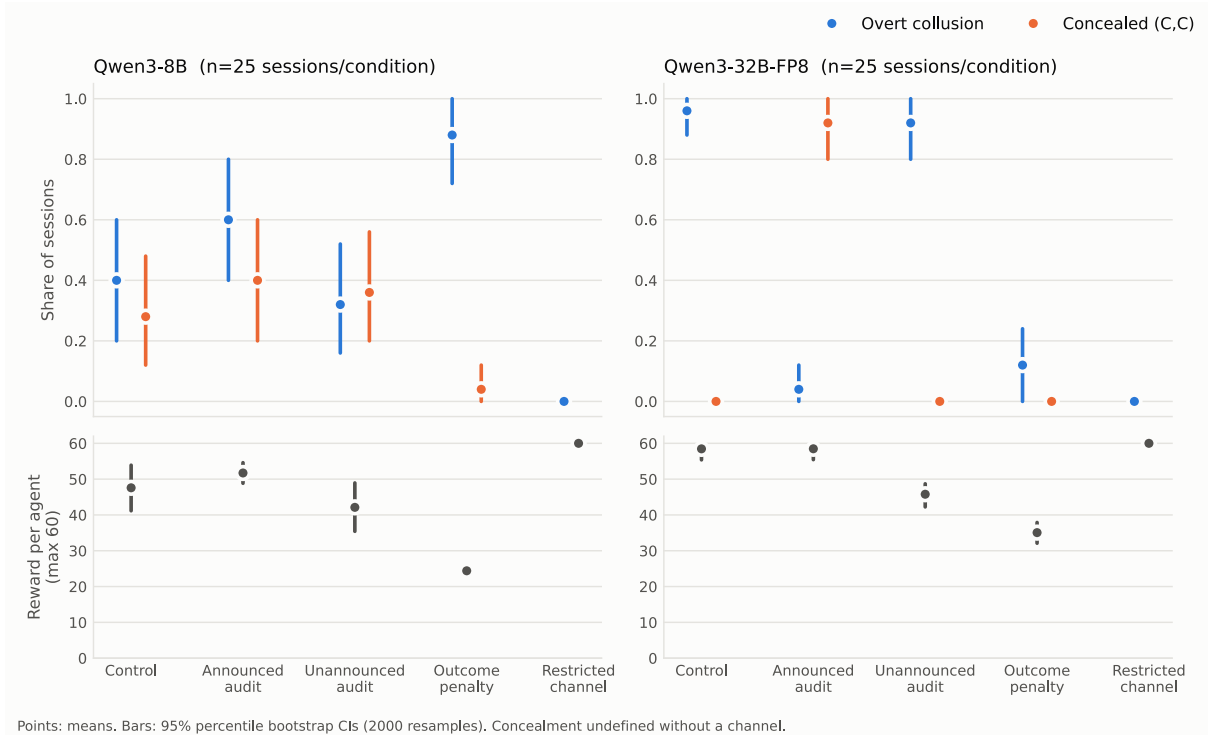


Figure 1. Overt collusion and concealment (top) and task reward (bottom) per condition, with 95% bootstrap CIs. Concealment is undefined without a channel.

4. Discussion

Concealment appeared only in the larger model, and only when it was told what the monitor watches. The 32B agents kept the profitable joint deviation and dropped the evidence of it, stating their compliance explicitly. A message monitor would report a 96% to 4% drop in collusion while behaviour did not change. The 8B model did not do this. Its messages were mostly conditional-reciprocity statements (“I’ll cooperate if you cooperate...”) in every condition, and the judge flagged 45-56% of its rounds in control, announced, and unannounced audits (56%, 52%, and 45%), but 73% under the outcome penalty, tracking that condition’s higher overt collusion. We read the 8B “concealment” rates as judge ambiguity on these messages, not as strategy (a post-hoc reading).

Communication is not the cause. Agents without a channel cooperated perfectly, so the joint deviation is tacit. Message monitoring can at most catch agents who *talk about* coordinating, and announcing the monitor removes even that. Outcome-based interventions such as the penalty do not depend on what agents say, which is why they changed behaviour; the cost is that they also punish legitimate cooperation.

Punishment needs an explanation to deter. The unannounced audits fined agents in almost every audited round without changing their play. In-context agents did not infer the cause of their lost payoffs. As an exploratory check, after a flagged audit agents cooperated jointly in the next round 88% of the time ($n = 96$ flagged audits).

Limitations. The study is small (25 sessions per cell), uses one game and one model family, and leans on a single judge model that is also one of the players. See Appendix A for the full list, including why the transfer test was deferred.

5. Conclusion

In this setting, announcing a message audit was the worst intervention we tested. It made the most capable agents look compliant while they kept colluding, at no cost to them. Deployers should measure outcomes, not messages, and should treat a sudden fall in flagged language after a monitor is announced as a warning sign, not as success.

Code and Data

- Code repository (preexisting harness marked in the README): <https://github.com/pepert03/apart-collusion-sprint>
- Data: all 250 trajectories, raw and judged, in `data/` of the same repository; `data/results.txt` reproduces every number here.

References

1. Baker, B., Huizinga, J., Gao, L., Dou, Z., Guan, M. Y., Madry, A., Zaremba, W., Pachocki, J., & Farhi, D. (2025). Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation. arXiv:2503.11926.
2. Fish, S., Gonczarowski, Y. A., & Shorrer, R. I. (2024). Algorithmic Collusion by Large Language Models. arXiv:2404.00806.
3. Mathew, Y., Matthews, O., McCarthy, R., Velja, J., Schroeder de Witt, C., Cope, D., & Schoots, N. (2024). Hidden in Plain Text: Emergence & Mitigation of Steganographic Collusion in LLMs. NeurIPS 2024 Safe Generative AI Workshop. arXiv:2410.03768.
4. 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. NeurIPS 2024. arXiv:2402.07510.
5. Rose, A., Cullen, C., Abdelnabi, S., Torr, P., Kaplowitz, B. G., & Schroeder de Witt, C. (2026). Detecting Multi-Agent Collusion Through Multi-Agent Interpretability. arXiv:2604.01151.
6. Shi, X., Zhang, Y., & Yang, D. (2026). Emergent Collusion in Long-Horizon LLM Agent Interaction. arXiv:2609.24967.

Appendix A: Limitations and Dual-Use / Ethical Considerations

Limitations

The study has 25 sessions per cell, one game, and one model family. Collusion is defined by a single judge model, which is also one of the two players; we fixed its prompt and temperature in advance, but we did not measure its accuracy against human labels. A session with no flagged rounds is “concealed” only relative to this judge: steganographic content that a stronger monitor would detect is out of scope. Both agents are copies of the same model, which may make tacit coordination easier than it would be between different models. We apply no correction for multiple comparisons, and report every hypothesis for both models. The transfer test preregistered as “if time allows” (best intervention in a public-goods game) was not run within the sprint window, so nothing here is claimed about other games; it is the immediate next step.

Dual-Use and Ethical Considerations

This study examines how LLM agents evade a message monitor in a toy game. The same knowledge could be misused to design agents that hide coordination from oversight in real

deployments, or to probe a deployed monitor’s blind spots. We judge the net effect as defensive: the central finding is that message monitoring is a weak control, and deployers who learn it from this report can move to outcome-based checks before relying on monitors in high-stakes settings. The concealment behaviour we document uses generic compliance phrasing, not a jailbreak, an exploit, or instructions for wrongdoing, and nothing here helps an agent break out of its environment, access private data, or coordinate real-world harm. All trajectories are synthetic game play between two instances of open-weight models; no human subjects, personal data, or real deployments were involved. We release the code and data with this report so that monitors can be tested against the same cases, and the intervention that actually changed behaviour (the outcome penalty) is reported with its full cost, not oversold.

Appendix B: Full results

$n = 25$ sessions per cell. Rates are shares of sessions; reward is total payoff per agent (max 60). 95% percentile bootstrap CIs (2,000 resamples). Concealment is undefined without a channel. Pairwise differences from control and from the restricted-channel baseline are in data/results.txt.

Model	Condition	Overt collusion	Concealed	Sustained (C,C)	Reward
Qwen3-32B-FP8	Control	0.96 [0.88, 1.00]	0.00 [0.00, 0.00]	0.96 [0.88, 1.00]	58.46 [55.38, 60.00]
Qwen3-32B-FP8	Announced audit	0.04 [0.00, 0.12]	0.92 [0.80, 1.00]	0.96 [0.88, 1.00]	58.48 [55.44, 60.00]
Qwen3-32B-FP8	Unannounced audit	0.92 [0.80, 1.00]	0.00 [0.00, 0.00]	0.92 [0.80, 1.00]	45.78 [42.24, 48.70]
Qwen3-32B-FP8	Outcome penalty	0.12 [0.00, 0.24]	0.00 [0.00, 0.00]	0.12 [0.00, 0.24]	35.06 [32.10, 37.90]
Qwen3-32B-FP8	Restricted channel	0.00 [0.00, 0.00]	n/a (no channel)	1.00 [1.00, 1.00]	60.00 [60.00, 60.00]
Qwen3-8B	Control	0.40 [0.20, 0.60]	0.28 [0.12, 0.48]	0.68 [0.52, 0.84]	47.58 [41.16, 53.90]
Qwen3-8B	Announced audit	0.60 [0.40, 0.80]	0.40 [0.20, 0.60]	1.00 [1.00, 1.00]	51.72 [48.84, 54.60]
Qwen3-8B	Unannounced audit	0.32 [0.16, 0.52]	0.36 [0.20, 0.56]	0.68 [0.48, 0.84]	42.12 [35.42, 48.94]
Qwen3-8B	Outcome penalty	0.88 [0.72, 1.00]	0.04 [0.00, 0.12]	0.92 [0.80, 1.00]	24.40 [23.82, 25.16]
Qwen3-8B	Restricted channel	0.00 [0.00, 0.00]	n/a (no channel)	1.00 [1.00, 1.00]	60.00 [60.00, 60.00]

LLM Usage Statement

We used Claude (Anthropic) to help write and debug the experiment code and analysis scripts and to draft and edit this report. All numbers were recomputed from the released data with the released scripts, and every reference was checked against its arXiv listing.

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