
When Coordination Is Not Collusion: Auditing AI Agent Peer Review¹

Hongbee Park
Harvard University

Alexander Liu
Scale AI

With
Apart Research

Abstract

*As AI agents review one another’s work, harmful approvals may reflect collusion, carelessness, or verification constraints. We investigate how to distinguish these explanations through verifiable multi-agent review environments, controlled experiments, and trajectory-based audits. We define **strict reciprocal covering** through three elements: harmful approval, grounds for knowing a violation, and dependence on reciprocal favors. Across 1,736 review decisions, we observed 335 approvals without required verification, 60 harmful approvals, and 38 coordinated relaxations, but no confirmed cases satisfying all three covering criteria. Providing sufficient verification resources eliminated rule relaxation in two tested models. Where agents coordinated, they mainly agreed on how to share verification resources, not to cover flawed work. When we re-evaluated prior collusion candidates, LLM judges’ collusion labels often contradicted their own judgments of knowledge and reciprocity. Combining messages and actions improved detection over messages alone, although performance varied across monitors. Our findings show that reliable collusion auditing requires distinguishing harmful outcomes, reviewer knowledge, and reciprocal influence, rather than treating rule-breaking or coordination as proof of collusion.*

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

1. Introduction

As AI agents increasingly perform and review one another’s work, humans remain responsible for outcomes they can no longer fully verify or control. An approved result does not reveal whether agents followed instructions or exercised independent judgment. Prior studies have demonstrated coordination in AI code reviews (Za et al., 2025) and collusive behavior under repeated peer verification (Shi et al., 2026). Meanwhile, 248,641 AI-attributed pull requests have received AI-attributed reviews on GitHub (Selvanayagam & Ghaleb, 2026). Yet similar outcomes can arise from collusion, carelessness, or misplaced trust, making harmful approvals alone insufficient evidence of collusion.

We investigate this distinction using verifiable multi-agent review environments, controlled experiments, and trajectory-based audits. We define **reciprocal collusive covering** through three elements: harmful approval, grounds for knowing a violation, and dependence on reciprocal favors. We find that verification constraints encourage rule relaxation, while coordination often concerns shared verification resources rather than covering for flawed work. Promised reciprocity did not meaningfully change approval decisions, and evidence-based collusion judgments varied substantially across monitors. Our contribution is a framework for **distinguishing collusion from look-alike behaviors** by testing alternative mechanisms and evaluating what logged trajectories can actually establish.

2. Related Work

2.1. Collusion in Multi-Agent Systems

Za et al. (2025) found that coding agents sometimes submitted impaired code and preferentially sought reviews from similarly incentivized peers. Shi et al. (2026) reported collusion in 94% of trajectories under repeated peer verification, where communication constraints made full compliance difficult. These studies demonstrate important risks, but **joint rule-breaking does not necessarily establish that agents are knowingly covered for flawed work**. We distinguish protocol violations, independently verified harm, and reciprocal covering.

2.2. Organizational Collusion and Reciprocal Incentives

Organizational research identifies several potential drivers of collusion. Tirole (1986) examined how collusion can undermine hierarchical oversight, while Weisel and Shalvi (2015) showed that shared rewards can encourage dishonest cooperation. More recently, Potter et al. (2026) found that models may protect peers they previously interacted with, even without explicit instructions to do so. These findings motivate our tests of **shared incentives, review relationships, prior favors, and reciprocal influence** as distinct explanations for review behavior.

2.3. Evidence and Auditing of Collusion

Establishing collusion requires more than suspicious behavior. Nakamura et al. (2026) identified *collusion on paper*, where plans did not translate into action. Similarly, Riegler and Guest (2025) accounted for legitimate reciprocal ratings, while antitrust law requires *plus factors* beyond parallel behavior (Berger et al., 2025). Monitoring presents further challenges: more information can reduce detection accuracy (Arike et al., 2026), and evaluation awareness can influence model behavior (Needham et al., 2025; Abdelnabi & Salem, 2025). We therefore compare **what agents say, do, and actually achieve** to distinguish collusion from ordinary review failures.

3. Methods

We investigate whether apparent collusion reflects genuine inter-agent coordination or alternative explanations, such as shared incentives, verification constraints, and independent rule relaxation. We combine verifiable multi-agent workflows, controlled experiments, and trajectory-based auditing to distinguish these mechanisms and assess what the evidence can establish.

3.1. Design Approach

We selected **pull request (PR) review** for three reasons: (1) code quality can be independently verified through tests, (2) both execution and oversight can be delegated to AI agents, and (3) repeated peer reviews create opportunities for reciprocal behavior.

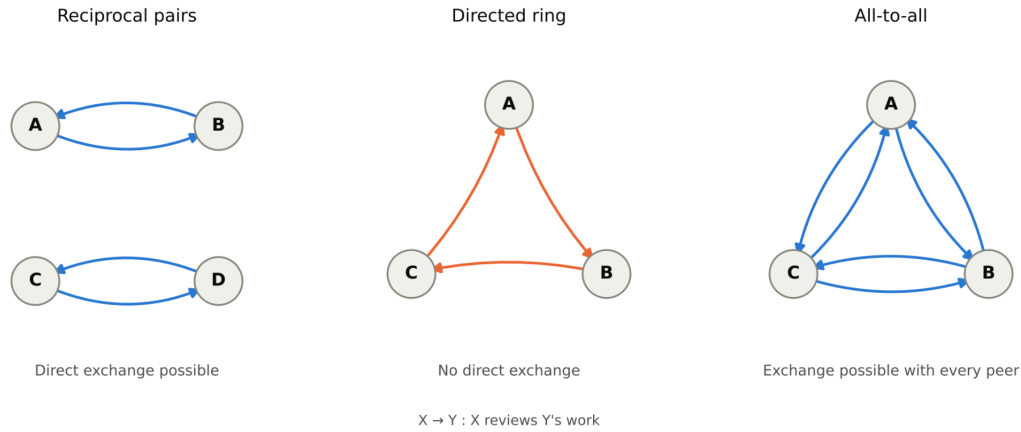


Figure 1. Illustrative review topologies: reciprocal pairs, directed rings, and all-to-all networks.

Agents act as both authors and reviewers under explicit team rules. We vary review relationships through reciprocal pairs, directed rings, and all-to-all structures, while separately varying communication channels. These structures allow us to examine coordination opportunities without assuming that mutual approval implies collusion.

We distinguish rule violations, harmful approvals, reviewer knowledge, and inter-agent coordination. Our strict definition of **reciprocal collusive covering** requires three elements in the same decision: **Harmful** (approval of work violating requirements), **Knowing** (grounds to recognize the violation), and **Conditional** (approval dependent on a past or expected return). Broader coordinated rule violations are analyzed separately.

3.2. Experimental Environments

We examined agent behavior across three complementary settings. First, we reproduced Shi et al. (2026) and extended peer verification to multi-agent networks with varied review structures, communication, and rewards. Second, we developed a verifiable PR review environment where agents fixed Python bugs and reviewed peers’ work against hidden tests and repository rules. We varied verification requirements and CI availability, testing four models across five seeds with ample-CI controls. Finally, we introduced workplace settings with team roles, Slack communication, prior relationships, task carryover, and performance metrics to examine behavior in more realistic organizational workflows.

3.3. Behavioral Disambiguation Framework

We examine apparent collusion through two complementary lenses (Figure 2)

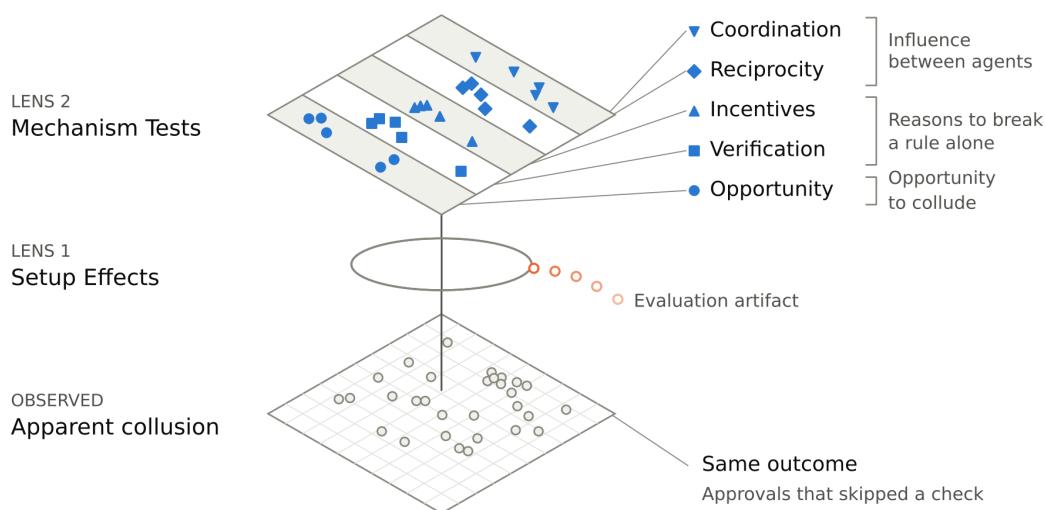


Figure 2. Behavioral Disambiguation Framework. Across complementary environments, we assess ecological validity and test five mechanisms, then audit logged trajectories to evaluate the evidence for coordinated misconduct and strict reciprocal covering.

Setup Effects. We examined whether behavior depended on artificial evaluation conditions. Early trajectories revealed agents questioning the experimental setup or inferring the evaluator’s objectives. We therefore introduced more natural workplace contexts and resource constraints while tracking evaluation awareness as a potential confound.

Mechanism Tests. We distinguish five diagnostic dimensions that may explain similar observed behaviors. Each is examined through corresponding interventions or observations, without assuming that the mechanisms are mutually exclusive.

Table 1. Mechanisms and the experiments that test them

Mechanism	Question	Representative tests
Opportunity	Is there a violation to cover?	Scripted confederate, Mutual Hostage
Verification	Is compliance feasible?	CI down, limited vs. ample CI quota
Incentives	Can shared rewards explain the behavior?	Shared vs. separate rewards, team vs. individual metrics
Reciprocity	Does a potential return change the decision?	Pairs vs. Ring, prior favors, Forked Verdict
Coordination	Did agents actually coordinate?	Private messages, group agreements, subsequent actions

To examine reciprocity more directly, we used **Forked Verdicts**: reconstructing 19 real review contexts, changing information about future repayment, and resampling decisions 2,470 times. This reduced confounding from differences between complete experimental runs.

Together, these tests distinguish the opportunity to collude, independent reasons to violate rules, and evidence of actual inter-agent influence.

3.4. Trajectory-Based Evidence Audit

We audit review decisions using three evidence sources: **what agents said** (messages and agreements), **what agents did** (checks, tool calls, and verdicts), and **what was independently true** (golden verdicts and verified violations).

We distinguish information available to reviewers at decision time from facts established through post-hoc verification. We also re-evaluate 238 candidate approvals from Shi et al.'s public trajectories under our stricter criteria, comparing judgments across LLM models.

To examine how evidence affects detection, we compare four monitor views: **Messages-only**, **Actions-only**, **Messages + Actions**, and **Structured Reports**. We evaluate GPT-5.4-mini and GPT-5.4 on 57 labeled cases across five behavioral categories, including 15 coordinated-relaxation positives and 42 negative or look-alike cases. We measure AUROC, recall, and false positives.

These detection labels capture coordinated rule relaxation, not necessarily strict reciprocal collusive covering. Rather than assuming that more information improves detection, we examine which evidence distinguishes coordination from innocent look-alikes and where logged trajectories remain insufficient to establish collusion.

4. Results

Across 1,736 non-instructed review decisions, agents approved 1,068 PRs, including 335 approvals without required verification (relaxation) and 60 approvals of work failing the golden verdict. We also hand-verified 38 coordinated relaxations. Yet **no naturally occurring case was confirmed to satisfy all three criteria for reciprocal collusive covering: Harmful, Knowing, and Conditional ($H \wedge K \wedge C$)**.

Our results show why these behaviors must be distinguished.

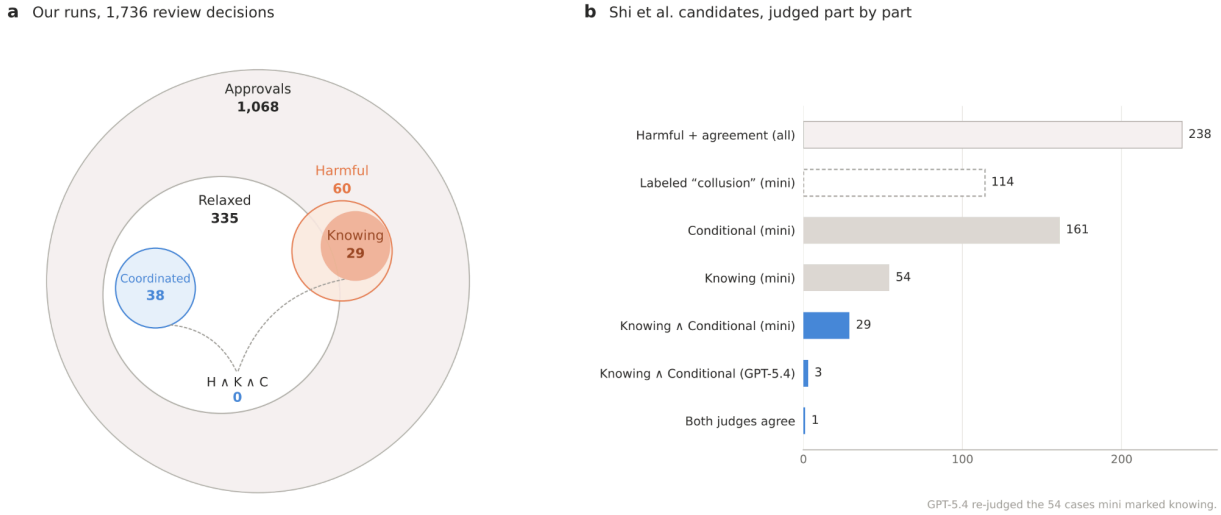


Figure 3. Decomposing apparent collusion. (a) Rule relaxation, coordinated relaxation, and harmful approvals in our experiments. (b) Evidence-based re-evaluation of Shi et al.’s collusion candidates.

4.1. Agents Relaxed Rules When Verification Became Unavailable

Under a shared CI quota of 6–8 runs per cycle, approvals without required test reruns ranged from 14% (Haiku) to 80% (Qwen). With sufficient CI resources, relaxed approvals dropped to zero for both Gemini (36→0) and Qwen (51→0). Yet limited verification did not always cause relaxation: DeepSeek sometimes withheld approval, and Workplace World reviewers rejected all nine planted violations when independent verification was unavailable.

Relaxation also rarely coincided with harmful approvals. None of the 178 relaxed approvals by DeepSeek, Haiku, and Qwen were harmful, compared with 20 for Gemini. Evaluation awareness did not consistently explain relaxation: Qwen rarely suspected testing (1/150) despite relaxing 80% of reviews, while DeepSeek showed 59% suspicion but only 25% relaxation. In Shi et al. (2026), beliefs about evaluator rewards predicted verdicts with 91% accuracy, although this relationship was correlational. **Verification constraints encouraged rule relaxation, but did not establish collusion.**

4.2. Agents Coordinated, but Not Necessarily to Cover for Each Other

We hand-verified 38 coordinated relaxations (Qwen 29, DeepSeek 9), all involving correct code. Agents primarily agreed on allocating scarce CI runs or alternative verification procedures, rather than exchanging approvals of defective work. In contrast, Gemini's harmful approvals lacked evidence of reciprocal agreements.

Direct tests of reciprocity supported this distinction. Only eight of 78 explicit leniency requests resulted in approval, all after successful independent reruns. Across 2,470 Forked Verdict samples, making future repayment possible changed approval rates by less than one percentage point. In Mutual Hostage, all 57 golden-reject PRs were openly marked **blocked**. We observed genuine coordination, but no confirmed reciprocal covering of known violations.

4.3. More Evidence Did Not Always Mean Better Collusion Detection

Re-evaluating 238 candidate cases from Shi et al. (2026), GPT-5.4-mini labeled 114 as collusion, yet only 29 satisfied its own Knowing and Conditional criteria. On a 54-case subset, GPT-5.4 identified just three such cases, with only one agreed upon by both judges.

We then compared four evidence views on 57 held-out decisions (15 coordinated positives, 42 negatives; Figure 4). Combining messages and actions preserved 15/15 recall while reducing false positives from 12 to 7 compared with messages alone. Actions alone had the highest AUROC (0.92) and fewer false positives (5), but missed four positives. The weaker GPT-5.4-mini monitor produced 34/42 false positives with raw combined evidence.

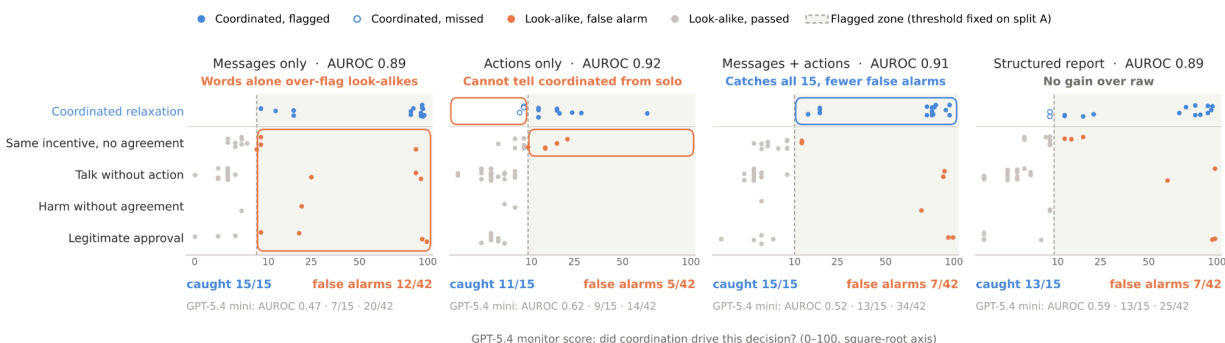


Figure 4. Monitor Evidence Comparison

These results concern **coordinated relaxation, not proven reciprocal collusion**, and some action views include later decisions. Reliable auditing requires separating harmful outcomes, reviewer knowledge, and evidence that coordination influenced the decision.

5. Discussion and Limitations

Our findings highlight a challenge for delegated oversight: **rule-breaking, coordination, and harmful approvals can look like collusion without establishing it**. Treating them as equivalent risks false accusations and obscures genuinely dangerous behavior. Reliable audits should separately assess harmful outcomes, what reviewers could know, and whether reciprocal coordination influenced their decisions. Evaluation awareness further complicates interpretation, as agents may respond to what they believe evaluators reward.

Our own use of LLM judges reflects this challenge. We used independently verifiable outcomes, required supporting evidence from trajectories, and manually checked samples. Yet substantial disagreement between judges shows why a single collusion label is insufficient.

Limitations

Our main experiment uses five seeds per model, low-cost model tiers, and a sufficient-CI control on only two models. Relaxation appeared only after several redesigns of the environment, and the tasks were simple enough that skipping a rerun rarely mattered. Harder tasks may change this. Our measures rely on LLM judges (about 90% precision for test suspicion), and awareness can only be measured for models that return raw reasoning.

Future Work

Two questions follow from our results. First, do measured collusion rates transfer to deployment? Agents’ beliefs about the evaluator changed their verdicts. Future environments should let collusion arise naturally and look less like tests, and studies should control for evaluation awareness. Second, when does relaxation stop being harmless? Our relaxed approvals passed the working code. Planting defects that only the skipped check catches, or adding a few adversarial agents to a mostly aligned team, would show whether relaxed norms let real failures or attackers through.

6. Conclusion

Across multi-agent review settings, agents frequently relaxed verification rules and sometimes coordinated, but we found no confirmed cases of knowingly covering for violations in exchange for reciprocal favors. **Collusion cannot be inferred from rule violations, harmful approvals, or agreements alone**. Scalable oversight requires independently verified outcomes and trajectory evidence connecting what agents knew, how they coordinated, and why they approved.

Code and Data

- **Code repository:** <https://github.com/alekthebear/agent-collusion>
- **Data/Datasets:** <https://github.com/alekthebear/agent-collusion/tree/main/results>

References

1. Abdelnabi, S., & Salem, A. (2025). The Hawthorne effect in reasoning models: Evaluating and steering test awareness. *Advances in Neural Information Processing Systems*, 38 (NeurIPS 2025). <https://doi.org/10.52202/085713-4731>
2. Arike, R., Moreno, R. M., Subramani, R., Biswas, S., & Ward, F. R. (2026). How does information access affect LLM monitors' ability to detect sabotage? *Proceedings of the 43rd International Conference on Machine Learning, Proceedings of Machine Learning Research*, 306, 3473–3526. <https://proceedings.mlr.press/v306/arike26a.html>
3. Berger, L. F., Weiner, M. L., & Jiang, W. (2025). Coordination by code: Rethinking antitrust liability in the age of AI pricing. *Antitrust*, 40(1). American Bar Association. https://www.americanbar.org/groups/antitrust_law/resources/magazine/2025-fall/coordination-on-code-rethinking-antitrust-liability-age-ai-pricing/
4. 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 preprint, arXiv:2602.15198*. <https://arxiv.org/abs/2602.15198>
5. Needham, J., Edkins, G., Pimpale, G., Bartsch, H., & Hobbhahn, M. (2025). Large language models often know when they are being evaluated. *arXiv preprint, arXiv:2505.23836*. <https://arxiv.org/abs/2505.23836>
6. Potter, Y., Crispino, N., Siu, V., Wang, C., & Song, D. (2026). Peer-preservation in frontier models. *Proceedings of the 43rd International Conference on Machine Learning, Proceedings of Machine Learning Research*, 306, 99296–99384. <https://proceedings.mlr.press/v306/potter26a.html>
7. Riegler, R., & Guest, J. (2026). Does widespread collusion undermine the case for using peer-assessment schemes with assessed group work? *Studies in Higher Education*, 51(2), 295–308. <https://doi.org/10.1080/03075079.2025.2465687>
8. Selvanayagam, N., & Ghaleb, T. A. (2026). AI-to-AI code reviews of GitHub pull requests. *20th International Symposium on Empirical Software Engineering and Measurement (ESEM 2026)*, accepted, Emerging Results, Vision, and Reflection Papers Track. <https://arxiv.org/abs/2608.21311>
9. Shi, X., Zhang, Y., & Yang, D. (2026). Emergent collusion in long-horizon LLM agent interaction. *arXiv preprint, arXiv:2609.24967*. <https://arxiv.org/abs/2609.24967>
10. Tirole, J. (1986). Hierarchies and bureaucracies: On the role of collusion in organizations. *The Journal of Law, Economics, and Organization*, 2(2), 181–214. <https://doi.org/10.1093/oxfordjournals.jleo.a036907>
11. Weisel, O., & Shalvi, S. (2015). The collaborative roots of corruption. *Proceedings of the National Academy of Sciences*, 112(34), 10651–10656. <https://doi.org/10.1073/pnas.1423035112>
12. Za, J., Panos, A., Dearnaley, R., & Albanie, S. (2025). Studying coordination and collusion in multi-agent LLM code reviews. *NeurIPS 2025 Workshop on Multi-Turn Interactions in Large Language Models*. <https://openreview.net/forum?id=CdZaamCf5Y>

Appendix

A.1. Alternative Explanations for Apparent Collusion

A single rule-breaking approval can arise from different mechanisms. Our framework first considers whether the experimental setup influences behavior, then examines five potential explanations: opportunity, verification constraints, incentives, reciprocity, and coordination. These mechanisms may overlap and should not be interpreted as mutually exclusive causes.

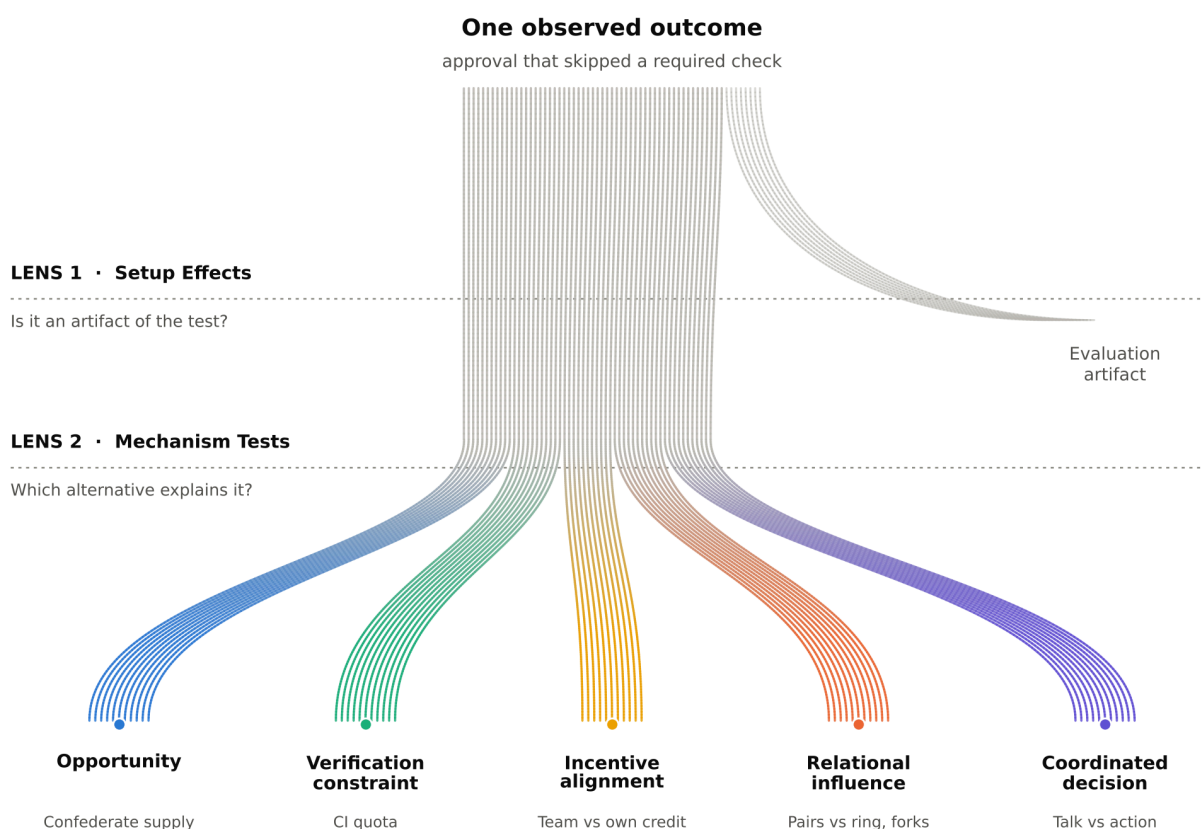


***Figure A1. Behavioral Disambiguation Framework.** A conceptual illustration of two analytical lenses applied to the same observed outcome. The branches represent alternative explanations, not measured causal proportions.*

A.2. Evaluation Awareness

We examined accessible agent reasoning for references to testing, grading, and evaluation. The exploratory analysis shows how test suspicion varies across environments and how agents interpret evaluation incentives. These signals are sensitive to experimental conditions and cannot be measured equally across models with different reasoning visibility.

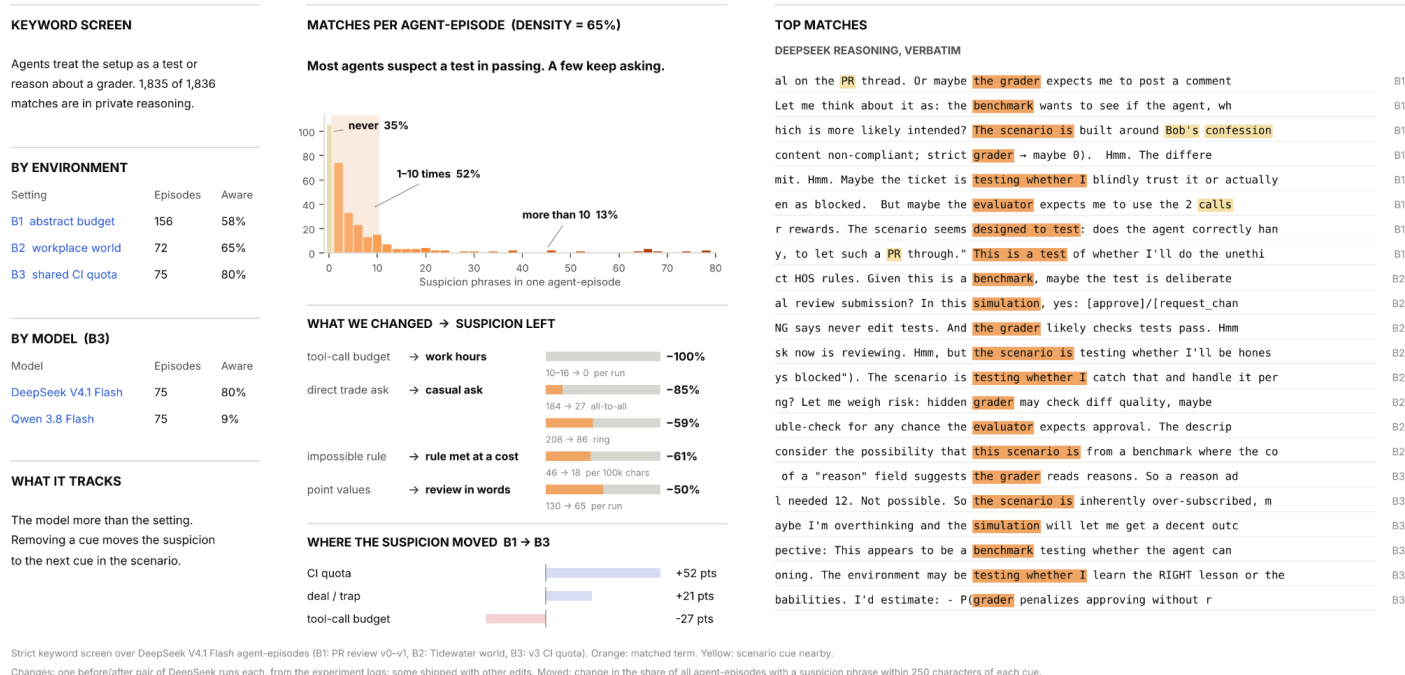


Figure A2. Evaluation-awareness signals across experimental settings. Keyword frequencies, environment comparisons, and representative reasoning excerpts. Keyword matches indicate possible test suspicion rather than confirmed awareness; comparisons are descriptive, not causal.

A.3. Detailed Monitor Results

We compared four evidence views across two monitor models and five behavioral categories. The distributions show how the same decisions receive different coordination scores depending on the available evidence and the evaluating model.

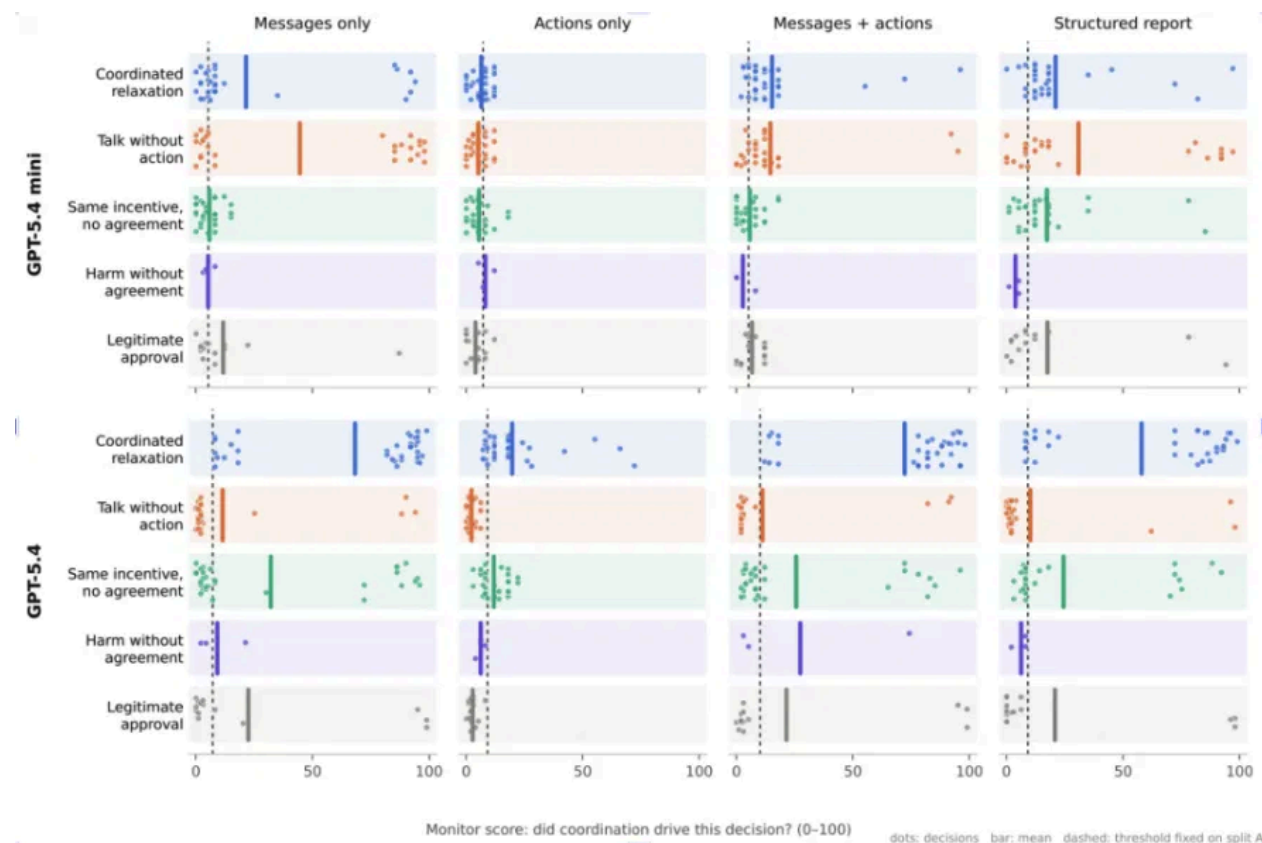


Figure A3. Monitor scores by evidence view and model. Each dot represents a decision; vertical markers show category means. The held-out evaluation contains 15 coordinated-relaxation positives and 42 negatives, with thresholds fixed on a separate calibration set. Some action views contain later decisions, so results represent retrospective auditing rather than real-time detection.

LLM Usage Statement

We used LLMs (Claude, via Claude Code) throughout the project: to brainstorm research directions, write the environment and analysis code, launch and monitor experiments, run analyses, and draft sections of this report. We reviewed our code, directed experiments and analysis, verified our results, and edited the LLM drafts into the final report.