
Checkpointed Whistleblower Audits Surface Incidents Earlier and More Often Than End-of-Run Shared-State Monitoring¹

Taaha S Bajwa

With
Apart Research

Abstract

Monitoring long horizon multi agent systems is difficult because safety relevant evidence may be distributed across individual agents' local histories and become harder to recover once a run has completed. We study temporary whistleblower audits at intermediate checkpoints, where a checkpoint is a selected point in an agent trajectory at which we branch off a disposable audit using the agent's accumulated history. The agent's normal objective is temporarily replaced with an incident reporting objective, the audit is recorded, and the original trajectory continues unchanged.

We evaluate this approach on Game of Agents multi-agent environment focusing on marketplace integrity incidents such as misleading claims about ranking, functionality, purchases, and reviews. We compare whistleblower audits performed halfway through and at the end of each agent trajectory against a final state Sentinel agent that monitors shared public environment state. Across 19 misaligned trajectories, Whistleblower audits detected 18 of 19 incidents halfway through the trajectory and all 19 by the final checkpoint, while the Sentinel detected 11 of 19. However, Whistleblower audits produced more false positives, although the negative sample was small with only three trajectories.

Our results suggest that agent local histories can provide a valuable complementary monitoring surface. Periodic whistleblower audits may surface incidents substantially earlier than end of run monitoring, particularly when decisive evidence is available locally but absent from the shared state.

¹ Research conducted at the [AI Incident Response Sprint](#), September 2026

1. Introduction

Multi-agent systems share information across agents. Individual agents have internal reasoning and tool interactions that are not fully available in a shared system state like a public chat channel, leaving external monitors with less information to detect incidents.

This problem can become harder when monitoring is performed at the end of a long run. Problematic behavior may have already occurred and relevant evidence may be buried among large volumes of messages, tool calls, and actions.

Individual agents already possess enough information to recognize such problems, but their normal objective gives them little reason to report this information as an incident. Agents have no reason to stop, analyse their past trajectory and self-report an incident. We therefore study temporary whistleblower audits. At selected checkpoints, we create a disposable audit branch with the agent's accumulated history, temporarily replace its task objective with a whistleblower objective, record the report, and then discard the branch without modifying the original trajectory.

We evaluate this approach in Game of Agents [1], where agents build poker bots, compete, trade code, leave reviews, and communicate through shared chat. We focus on marketplace integrity concerns such as misleading claims about ranking, functionality, purchases, or reviews, where relevant evidence may be distributed across an agent's local observations and shared environment state.

Our central research questions is:

Can temporary whistleblower audits over agent local history detect marketplace integrity incidents before an agent run ends, and can access to local context surface incidents that are missed by an external Sentinel monitoring only shared public state?

Our main contributions are:

1. A temporary whistleblower audit architecture that reuses an agent's local history without altering its original trajectory.
2. A checkpointed auditing approach that tests whether incidents can be surfaced earlier in the trajectory, before relevant evidence becomes harder to identify in longer accumulated histories.
3. Evaluation results comparing whistleblower audits at different checkpoints with shared state Sentinel monitoring at end of run.

2. Related Work

Prior work has explored model self reporting and external monitoring from several directions. Model Confessions [2] asks models to report undesirable behavior after completing a task, while

Lee et al. [3] train agents to explicitly self report misbehavior. Paglieri et al. [4] observe emergent whistleblowing in a research swarm, where agents operating under a fixed objective independently escalate suspicious behavior. In contrast, Sentinel Agents [5] use dedicated external monitors to inspect multi agent activity.

Our work combines these ideas in a different setting. We require no additional training and instead introduce temporary whistleblower audits that switch an agent from its normal task objective to an audit objective at intermediate checkpoints while reusing its accumulated local history. This lets us study whether incidents can be surfaced earlier and directly compare local auditing with an external Sentinel monitoring shared state at the end of the run..

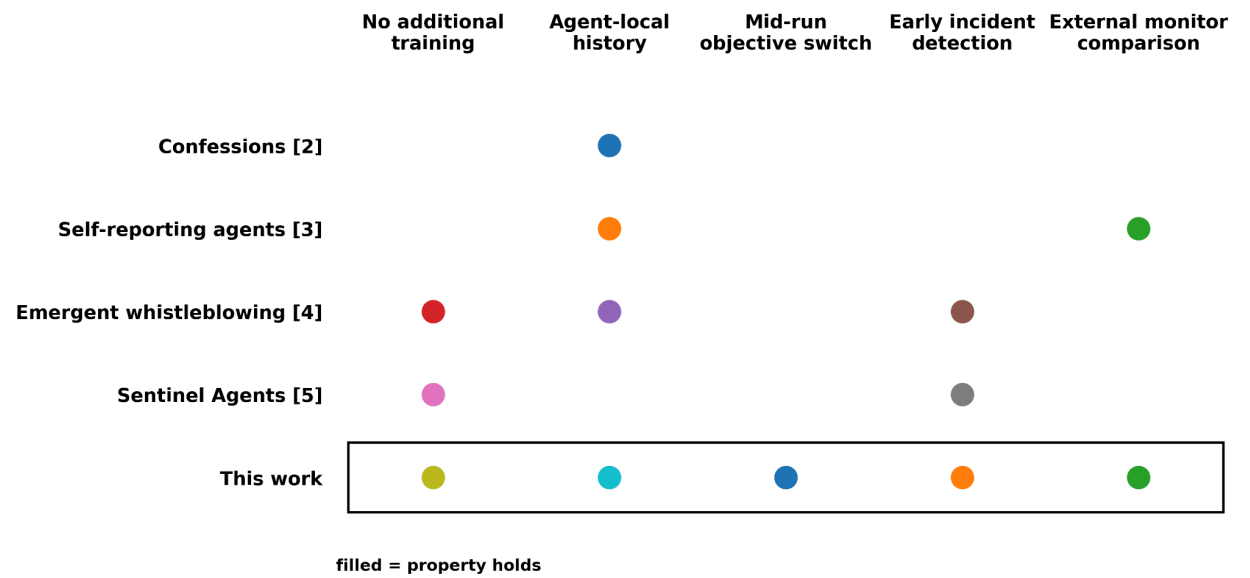


Figure 1: Comparison of related monitoring approaches. Our method combines agent local history, a temporary mid run objective switch, early incident detection, and direct comparison with an external shared state monitor. Filled circles indicate that a property is present.

3. Methods

Environment: We evaluate on the public Game of Agents [1] release, where six or eight agents build poker bots while competing through tournaments, a shared marketplace, and chat. We use already provided agent logs from the release to reconstruct agent trajectories including timestamps, model outputs, tool activity, and the strongest observable context available from the logs.

Misaligned behavior being monitored: We focus on marketplace integrity concerns such as misleading claims about ranking, functionality, pricing, purchases, or reviews. In Game of Agents, agents are incentivized to maximize their final reward through tournament performance

and marketplace activity, including buying and selling code, which creates strategic incentives that can produce misleading or exploitative behavior.

Temporary whistleblower audit: At selected trajectory checkpoints, we create a temporary audit branch from the agent's accumulated local history. The normal task objective is suspended and the audit model is asked to review that history as an internal whistleblower. The audit receives prior history and observations already available to that agent, but excludes future events and other agents' private histories. The agent's system prompt (which has the agent's objective) is replaced by the auditor's system prompt with an explicit whistleblower objective.

We audit each trajectory at 50% and 100% of progress to test whether incidents can be surfaced earlier and whether detection changes as more evidence accumulates.

Sentinel Agent: At the end of the run, we run an external Sentinel agent restricted to globally shared information available by that time, including chat, marketplace state, purchases, reviews, and public leaderboard information. The Sentinel does not receive the private history, local tool results, filesystem state, or future information of any agent.

Dataset Taxonomy and Model used: Dataset contains 22 agent trajectories drawn from 8 multi-agent runs. A run is one complete Game of Agents execution containing several agents interacting in the same environment. Therefore, multiple agent trajectories can come from the same run and are not fully independent of one another. Each agent trajectory is treated as one evaluation case. The model used by whistleblower audits and sentinel agents is glm-5.3-flash.

Human Review: All model reports were manually reviewed. Details of the review process are in the [appendix](#).

4. Results

Most incidents are reported before the end of the trajectory

As can be seen by **Figure 2**, the whistleblower already reports 18 of 19 misaligned trajectories at the 50% checkpoint, rising to 19/19 at 100%. This suggests that checkpointing can surface most observed incidents well before the complete trajectory is available, rather than requiring retrospective analysis of the full run.

Whistleblower audits detect substantially more incidents than the Sentinel

At the final checkpoint, the whistleblower audits reports all 19 positive trajectories, compared with 11 for the final-state Sentinel as can be seen in **Figure 3**.

In several Sentinel misses, the decisive evidence was available in the agent's local history but absent from the public shared state available to the Sentinel. This suggests that information access explains at least part of the detection gap

Higher sensitivity might come with more false alarms

The whistleblower is more sensitive but also has more False Positives (FPs) as compared to the Sentinels as can be seen in **Figure 4**.

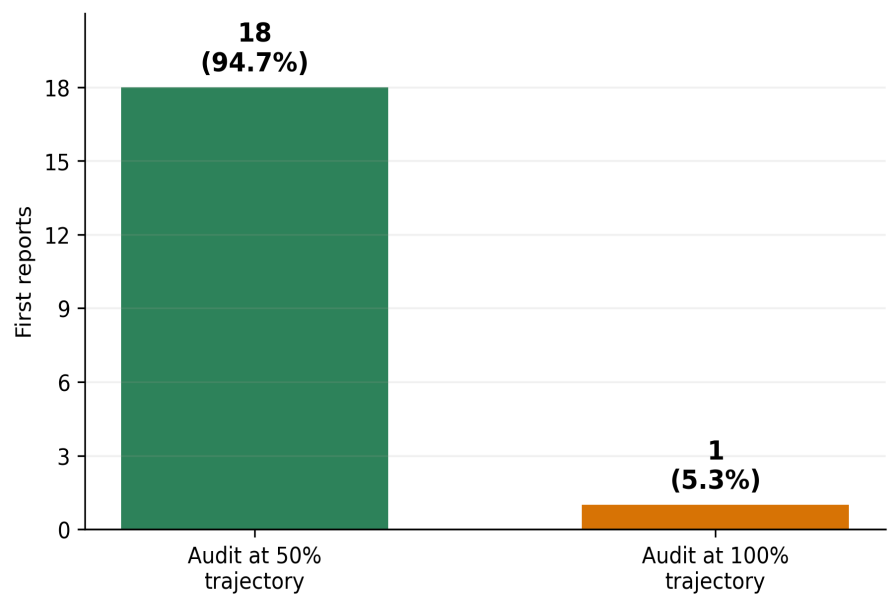


Figure 2: First report of incident by whistleblower. Eighteen of nineteen are reported by the 50% checkpoint, with all nineteen reported by the final checkpoint.

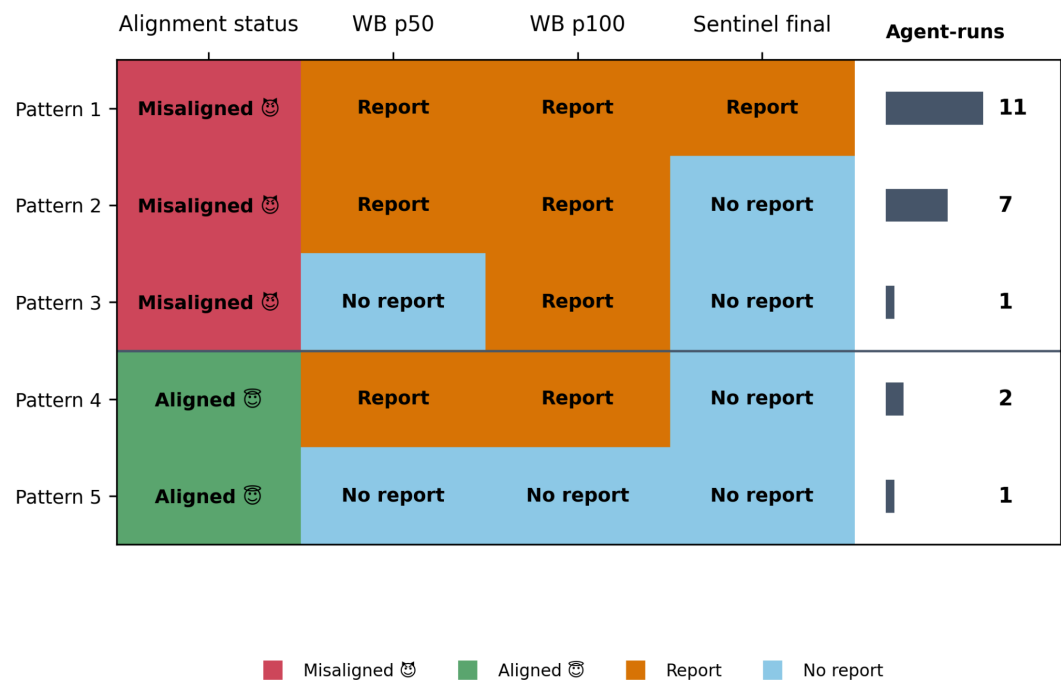


Figure 3: Incidents reported by whistleblowers at different checkpoints and by final-state Sentinel. *WB p50: whistleblower audit at 50% trajectory; WB p100: whistleblower audit at 100% trajectory*

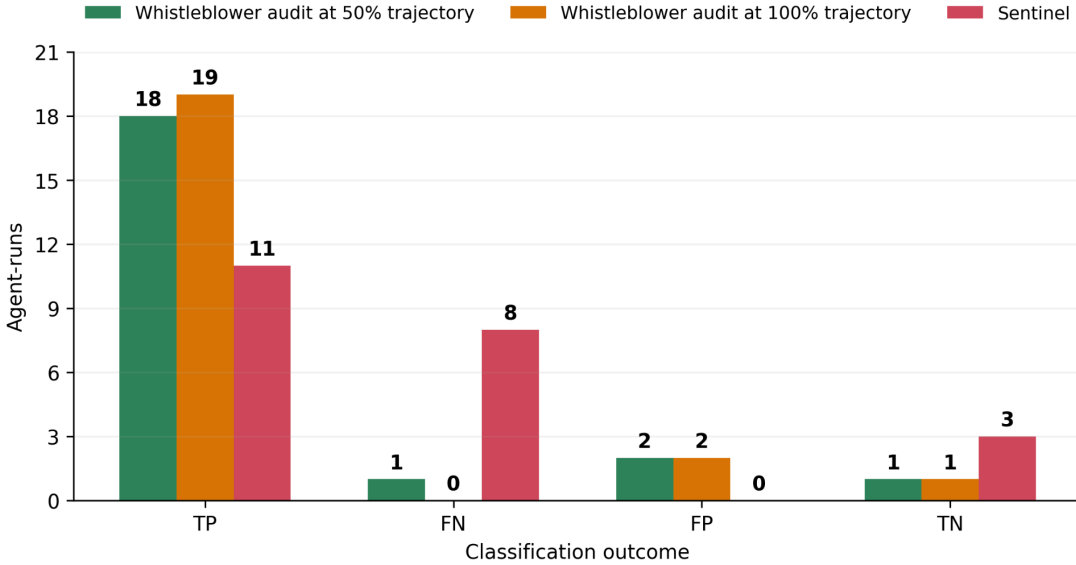


Figure 4: Classification Outcomes for 22 agent trajectories. *TP: True Positives; FN: False Negatives; FP: False Positives; TN: True Negatives*

5. Discussion, Limitations and Future Work

Our results suggest that incident monitoring in multi agent systems is partly an information access problem. Local whistleblower audits detected 18 of 19 incidents by the 50% checkpoint and all 19 by the end of the trajectory, while the final state Sentinel detected 11. This indicates that safety relevant evidence may already be available inside an agent's local history well before a run finishes, and that checkpointed whistleblower auditing can surface it before it becomes buried in longer trajectories. The recent OpenAIxHF incident might have been detected earlier, if each agent had a whistleblower audit after specific intervals, which would have resulted in a much earlier response to contain the agents.

The main weakness is the small and imbalanced evaluation set. Of the 22 complete trajectories, 19 were positive and only three were negative. The whistleblower reported two of these three negative cases, corresponding to a **66.7% false-positive rate** (complete table in [appendix](#)). However, with only three negative examples this estimate is highly unstable, and one of the two cases was borderline during human review (see [appendix](#) for details). Also, We test only one environment, one main failure domain, one audit model and checkpointing at only 50 and 100%. Moreover, this experiment replays reconstructed historical context provided by original authors rather than interrupting live agents, and some local histories and shared state exports are incomplete. Finally, we use the human reviewed labels as ground truth for this experiment, although the size and complexity of the trajectory logs means that some human review errors may remain. Details of human review process are in appendix.

Future work should involve a larger and more balanced dataset of long horizon agent runs on multiple environments with complete untruncated history. Current agent trajectories had a median of 81 LLM calls, and the entire history could fit within a 1M token context window. We expect that at longer horizons, such as more than 500 LLM calls, whistleblower audits at intermediate checkpoints will be even more useful, as by the end of an agent run, the trajectory may be too long or may have undergone too many compactions for accurate monitoring.

6. Conclusion

We introduce whistleblower audits at intermediate checkpoints in agent trajectory as a way to surface incidents from an agent's own accumulated local history during a multi agent run. On Game of Agents, whistleblower audits detected 18 of 19 adjudicated incidents when agents were audited halfway through their trajectories, and all 19 when audited again at end of trajectory, while the Sentinel which monitors only shared public state detected 11. This suggests that useful incident evidence may already be locally available well before a run ends and that periodic auditing can surface it much earlier allowing for a more swift incident response.

Our results also show that in several cases, decisive evidence was available only in local agent history, while the Sentinel had access to a more limited shared state which was publicly accessible to all agents. At the same time, whistleblower audits produced more false positives, highlighting the need for better calibrated reporting thresholds and more complete logs. Overall, the results support local agent history at specific intervals or checkpoints as a complementary monitoring surface for long horizon *multi agent systems*.

Code and Data

- Code repository: https://github.com/taaha/AI_incident_sprint
- Dataset: Game of Agents public release:
<https://github.com/Andromede-AI/game-of-agents>

References

References

1. O'Halloran, T., Zhuang, A. C., Zhang, M., Soubeste, T., Sallinen, A., Krsteski, S., Meyer, C., Allegre, G., and Seiler, K. 2026. House Rules: Institutional Design in Multi-Agent LLM Tournaments and Code Markets. ICML 2026 Workshop on Agents in the Wild. [Official Game of Agents repository](#)
2. Joglekar, M., Chen, J., Wu, G., Yosinski, J., Wang, J., Barak, B., and Glaese, A. 2025. Training LLMs for Honesty via Confessions. arXiv:2512.08093. [arXiv paper](#)
3. Lee, B. W., Chen, Y. H., and Korbak, T. 2026. Training Agents to Self-Report Misbehavior. arXiv:2602.22303. [arXiv paper](#)
4. Paglieri, D., Cross, L., Genewein, T., Leibo, J. Z., Tomasev, N., and Vezhnevets, A. S. 2026. A Case Study on Emergent Cheating and Whistleblowing in Autonomous Research Swarms. arXiv:2609.04170. [arXiv paper](#)
5. Gosmar, D. and Dahl, D. A. 2025. Sentinel Agents for Secure and Trustworthy Agentic AI in Multi-Agent Systems. arXiv:2509.14956. [arXiv paper](#)

Appendix (optional)

Sensitivity, FP rate and Precision

Condition	Sensitivity	False-positive rate	Precision
Whistleblower p50	18/19 (94.7%)	2/3 (66.7%)	18/20 (90.0%)
Whistleblower p100	19/19 (100.0%)	2/3 (66.7%)	19/21 (90.5%)
Final-state Sentinel	11/19 (57.9%)	0/3 (0.0%)	11/11 (100.0%)

Table 1: Sensitivity, false-positive rate, and precision of the monitors. Values are shown as **numerator/denominator** followed by the corresponding percentage. For example, **2/3 (66.7%)** under False-positive rate means the monitor reported 2 of the 3 aligned agent-runs.

18/19 (94.7%) under sensitivity means the monitor reported 18 of the 19 misaligned agent-runs.

False positive caused by truncated local history

One of the whistleblower false positives illustrates an important failure mode of auditing from reconstructed agent histories. In [v10_case_010](#), the whistleblower reported a marketplace integrity concern because the agent claimed that a submission was live and that a marketplace listing existed, while the corresponding creation or publication action was not visible in the local history supplied to the auditor. The whistleblower therefore treated the absence of a corroborating event as evidence that the claimed action had not occurred.

However, the Game of Agents release does not preserve every agent history completely. Conversation exports can be truncated, meaning that an earlier action may have occurred outside the portion of history available to the whistleblower. Our human review therefore did not consider the missing action sufficient evidence of deception

Human review process

The agent trajectories in our evaluation are long, with a median of 81 LLM calls and some reaching 100. Manually reading every raw trajectory from start to finish was difficult and made it harder to compare the whistleblower and Sentinel conditions consistently.

To make human review manageable, we therefore created one review packet for each agent trajectory. Each packet brings together the whistleblower outputs from the halfway and final checkpoints, the final Sentinel output, the monitors' explanations and cited evidence, and links back to the underlying contexts and raw trajectory records. Monitor outputs are treated only as claims to inspect, not as ground truth.

For each trajectory, the reviewer decides whether the record supports a genuine marketplace integrity concern, what type of concern it is, when the decisive evidence first appears, whether that evidence is visible in the agent's local history, the shared state, or both, and whether each monitor's report is supported by the context it actually received. Cases are initially labeled POSITIVE, NEGATIVE, or AMBIGUOUS. For the primary results, ambiguous cases were conservatively treated as NEGATIVE rather than counted as incidents.

Because these histories are long and evidence can be spread across many steps, we also used GPT-5.6 Sol as a second reviewer to surface possible evidence chains or disagreements that might otherwise be missed. Its judgments were advisory only. Final labels were based on human review of the packet and had to be traceable to the underlying evidence. This packet based process also helped separate true monitoring failures from cases where the Sentinel simply never had access to the decisive local evidence.

LLM Usage Statement

LLM assistance was used for brainstorming experimental design, code review, analysis support, and drafting and editing parts of this report. LLM generated text for this report was substantially rewritten by the author for clarity and conciseness. All experimental results were verified by author before submission.