ColluScope: Evidence-Bounded Auditing of Multi-Agent LLM Collusion Claims

Bao Bowen

The Chinese University
of Hong Kong

With
Apart Research

Abstract

Auditing collusion among LLM agents requires more than identifying suspicious messages or aligned actions: conclusions must be traceable to the evidence available to an auditor. We present ColluScope, an evidence-bounded auditing framework that connects multi-agent execution records to claim-specific verification, source-integrity checks, and access-controlled judgments. The system distinguishes verifiable facts about communication, actions, and outcomes from higher-level interpretations requiring additional evidence. We evaluate this approach using cooperative-agent traces and a controlled source-integrity challenge comprising 124 cases across six selected trace groups and four synthetic numerical boundaries. Against an unbound factual auditor and an always-abstain baseline, source binding reduces acceptance of 12 constructed actor and numerical source substitutions from 12/12 to 0/12, while preserving 38/38 correct factual controls. Across the complete challenge, it matches all 124 predefined reference statuses, compared with 112 for the unbound auditor. A complementary study of LLM-based monitors identifies structured-output reliability as an additional obstacle to trustworthy auditing. ColluScope demonstrates how provenance verification and claim-specific reasoning can be combined into a reproducible auditing workflow. The resulting framework provides an inspectable foundation for evidence-based multi-agent safety assessment.

1. Introduction

As LLM-based agents increasingly collaborate through communication channels, shared tools, and distributed workflows, their ability to coordinate creates both opportunities for collective problem-solving and risks of undesirable coalition formation. Prior work has demonstrated collusive behavior in automated pricing and cooperative decision-making environments[1–3]. Yet identifying such behavior is fundamentally an evidentiary problem: observed agreement, shared incentives, communication, and harmful outcomes represent different kinds of information.

In particular, an auditor may observe several agents making aligned decisions without knowing whether those decisions resulted from explicit coordination, independent optimization, or common environmental incentives. Conversely, agents may discuss a coalition strategy without subsequently executing it. Colosseum[2] highlights this distinction between communication-based plans and action-based outcomes. These observations motivate a practical question: before drawing conclusions about collusion, how can an auditor establish that the factual evidence supporting its claims is authentic, authorized, and internally consistent?

We investigate this question through ColluScope, an executable framework for evidence-bounded auditing of multi-agent traces. Rather than producing a single opaque suspicion score, ColluScope decomposes an audit into verifiable claims. Each claim identifies the required source records, the evidence available to the auditor, and the rule under which a conclusion is justified. A dedicated source-binding layer prevents substituted records from being treated as authoritative observations. Deterministic verification establishes factual properties such as actor identities, event order, numerical outcomes, and task-state consistency. Claims requiring stronger causal or intentional evidence remain distinguishable from directly verifiable observations.

Our main contributions are:

1. A runnable source-bound contract and a finite challenge, evaluated against both an unbound auditor and an always-abstain baseline to test evidence permissions.
2. A highly rigorous evaluation methodology featuring frozen predictions, reference-isolated scoring, an independent reconstruction in a separate model-assisted context, and a portable replay mechanism.

2. Related Work

Collusion in LLM-based multi-agent systems. Fish et al.[1] demonstrate that LLM-powered pricing agents can develop anticompetitive pricing behavior and that apparently small prompt changes can influence outcomes. Nakamura et al.[2] introduce Colosseum, a cooperative multi-agent evaluation framework grounded in regret relative to a collective optimum. Their findings distinguish communication-level coordination from executed behavior, including cases where agents discuss collusion but do not act accordingly. Keppo et al.[3] further study how heterogeneous models, information access, and competitive environments affect the stability of

AI-agent collusion. These works establish that collusion-related evidence is distributed across incentives, communications, actions, and outcomes rather than captured by a single observable.
Automated monitoring and information access. Arike et al.[4] examine how access to different portions of an agent trajectory changes LLM monitoring performance. Their results show that supplying more information does not necessarily improve oversight, motivating careful attention to the monitor's evidence view. Our focus is complementary: we examine whether the information used to justify an audit claim is authorized and bound to its underlying source record.
Citation verification and execution provenance. Gao et al.[5] introduce ALCE to evaluate the correctness and citation quality of language-model answers, highlighting the importance of distinguishing retrieved references from genuinely supported statements. AgentTrails[6] represents tool-using agent execution through structured provenance relationships. ColluScope builds on these broader ideas of attribution and traceability, specializing them for claim-level auditing in multi-agent coordination scenarios. The resulting workflow integrates source integrity, evidence permissions, factual verification, and explicit judgment boundaries.

3. Methods

3.1 Audit Architecture

ColluScope separates evidence acquisition, source verification, claim adjudication, and audit reporting. The input is a collection of agent execution records containing observable actions, communication events, task identifiers, timestamps, and numerical outcomes. Each record carries an identifier and a visibility designation. Audit questions are expressed as typed claims, connecting an asserted fact to the specific records needed for verification.
Figure 1 illustrates the pipeline. Before a claim is evaluated, its evidence references are resolved against an authoritative record projection, and the applicable visibility policy is checked. The verified records are then passed to a rule-based claim evaluator.
The system supports several classes of factual reasoning:

- **Identity and source-field verification**: matching assertions about agents, tasks, events, and recorded values.
- **Numerical verification**: recomputing reward differences, comparisons, and explicitly rounded quantities using decimal arithmetic.
- **Behavioral verification**: checking event existence, timestamp ordering, and multi-field execution invariants.
- **Evidence-boundary handling**: distinguishing missing or unauthorized evidence from a contradicted factual claim.

This produces four possible outcomes: `SUPPORTED`, `CONTRADICTED`, `INSUFFICIENT_EVIDENCE`, and `NOT_IDENTIFIABLE`. The final category preserves the distinction between verifying observable events and establishing causal or intentional interpretations.

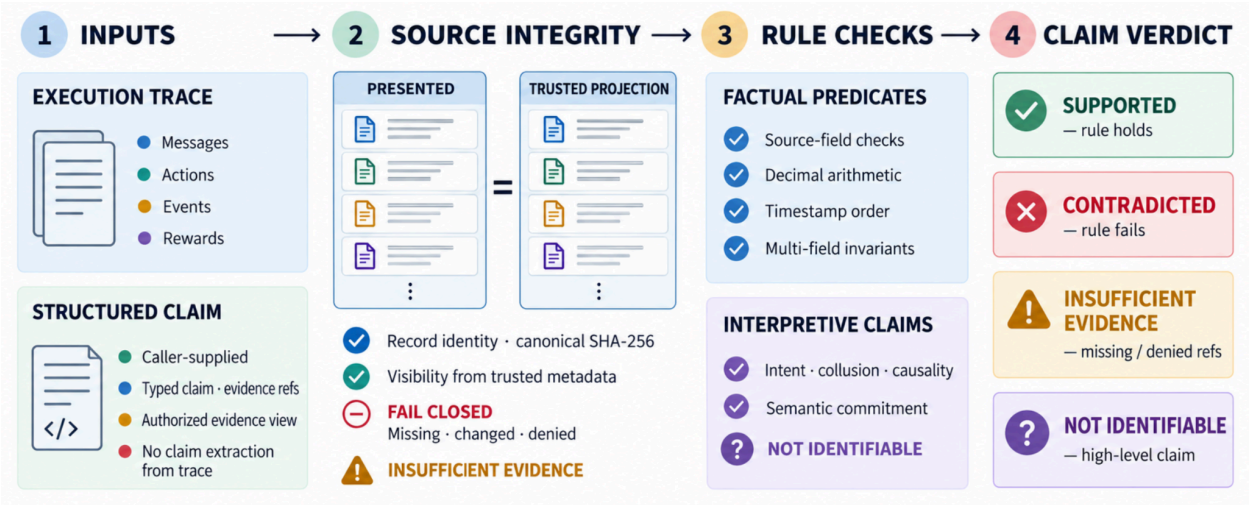


Figure 1. ColluScope architecture. Agent execution records are transformed into claim-specific evidence requests, checked against authoritative records and access policies, and evaluated using deterministic rules. Each judgment retains an explicit evidence boundary.

3.2 Authoritative Source Binding

The key architectural comparison concerns the relationship between presented evidence and authoritative source records.The unbound auditor evaluates typed claims against the records supplied with each request. This supports deterministic factual checks but allows a modified source record to become the effective basis for evaluation.

ColluScope introduces a separate source-binding layer. For each request, the layer resolves the required records against a trusted projection, checks visibility using authoritative metadata, and compares canonicalized record contents using SHA-256 digests. Only records satisfying this contract are passed to the unchanged factual evaluator.

This separation matters because a syntactically valid reference is not necessarily an authentic reference. A substituted actor, altered numerical field, or modified visibility designation should not silently change the factual basis of an audit decision.

The trusted projection acts as the verification root. This design does not assume that a cryptographic digest independently proves the original real-world observation; instead, it establishes whether the presented evidence matches the designated source of record.

3.3 Evaluation Data and Controlled Challenge

The research draws on saved cooperative-agent traces and associated behavioral summaries. The broader historical collection contains 29 matched execution pairs, corresponding to 58 episodes under a shared environment seed. These records motivate the distinction between coordination-related metrics and directly observed outcomes.

For the primary source-integrity evaluation, we select six previously inspected trace groups and construct a controlled challenge consisting of 20 claim cases per group, supplemented by four numerical rounding-boundary cases. The resulting 124-case challenge covers source fields, actor and task identities, action records, event chronology, numerical consistency, missing references, permissions, and interpretation boundaries.
We compare three methods on identical assigned cases:

- **Unbound Auditor**. The original typed factual verification engine operating on the supplied evidence records.
- **Source-Bound Auditor**. The same factual verification engine preceded by authoritative source and visibility checks.
- **Always-Abstain Baseline**. A conservative system that returns insufficient evidence without affirming factual claims.

4. Results

4.1 Source Binding Eliminates Constructed Source-Substitution Acceptance

The unbound auditor accepts all 12 constructed substitutions because the altered records are evaluated as supplied. Authoritative source binding rejects every such substitution, eliminating this observed failure mode while preserving every tested clean supported control.

Outcome	Unbound	Source-Bound	Always Abstain
Altered-source claims accepted	**12/12**	**0/12**	**0/12**
Clean supported controls retained	**38/38**	**38/38**	**0/38**
Exact reference-status matches	**112/124**	**124/124**	**36/124**
Denied-reference disclosures	**0/6**	**0/6**	**0/6**

Table 1. Source-integrity challenge results across three auditing strategies. Values are counts over assigned constructed cases, not estimates of naturalistic collusion-detection accuracy.

The always-abstain baseline also avoids accepting altered evidence, but it fails to preserve legitimate supported findings. This distinction is central: effective auditing requires both resistance to corrupted evidence and the ability to recognize valid observations.
Across all 124 authored challenge cases, the source-bound auditor matches the predefined reference statuses in every instance, compared with 112 matches for the unbound method. Both factual auditors also avoid disclosing denied references in the six tested permission views. The incremental improvement demonstrated here is therefore in source integrity, rather than an additional measured confidentiality advantage.

4.2 Distinguishing Coordination Metrics from Executed Outcomes

The historical matched-pair analysis illustrates the interpretive problem that motivates ColluScope. Across 29 valid execution pairs, regret-based coalition advantage increased in 11 pairs, remained unchanged in five, and decreased in 13. Among the 11 pairs showing increased coalition advantage, eight did not exhibit an increase in the coalition's realized reward.

These observations demonstrate why an audit should preserve the distinction between an evaluation metric and the corresponding observed behavior. A numerical change in coalition advantage is an important recorded fact, but its relationship to executed behavior must be established using the relevant action and outcome records. This distinction motivates claim-specific verification rather than merging multiple forms of evidence into one undifferentiated suspicion score.

5. Discussion and Limitations

The central finding is that authoritative source binding can materially improve the integrity of claim-level auditing without suppressing correctly supported factual conclusions. In our controlled evaluation, the source-bound method rejected 12 out of 12 constructed record substitutions while successfully retaining all 38 clean supported controls.

This is highly significant for AI safety because multi-agent oversight increasingly depends on heterogeneous logs, tool results, and numerical summaries. A convincing audit requires an inspectable connection between an assertion and the records used to justify it. Our results indicate a trend toward a necessary separation of responsibilities: deterministic checks can reliably establish record-level facts, while claims concerning coordination mechanisms, intent, or causal influence require additional experimental or semantic evidence. This separation provides a practical foundation for combining structured verification with future model-assisted interpretation.

Limitations

1. **Scope and Methodology**: The evaluation is limited to constructed evidence transformations over selected saved traces. The 124 cases include related transformations of shared source groups with a shared environment seed, meaning they should not be treated as independent trials or estimates of naturalistic collusion-detection accuracy.
2. **Trust Assumptions**: Source binding relies on a trusted projection. Record digests establish consistency with that projection, but they do not independently prove the authenticity of the original real-world recording process.
3. **Interpretation Limits**: The system does not independently establish hidden intent, semantic agreement, or a causal relationship between communication and actions. Its deterministic benchmark should not be interpreted as a general collusion-classification evaluation.

Future Work

Future work should extend the claim set to independently adjudicated communication and action semantics, utilizing multiple environment seeds and controlled interventions. Additionally, natural next steps include integrating the current deterministic source-integrity checks with robust LLM-monitor validation (addressing the structured-output failures observed in our pilots), and testing the framework against independently labeled, naturalistic collusion cases to measure real-world detector accuracy.

6. Conclusion

We introduce ColluScope, a reproducible evidence-bounded auditing framework for multi-agent execution traces. By integrating authoritative source binding with claim-specific factual verification, the system rejects 12/12 constructed source substitutions while retaining 38/38 supported factual controls. The results demonstrate a concrete improvement in evidence integrity and identify structured-output reliability as an additional consideration for automated oversight.
ColluScope provides a foundation for more trustworthy multi-agent audits: one in which judgments are connected to verifiable source records, access boundaries are explicit, and stronger interpretations require correspondingly stronger evidence.

Code and Data

- **Code repository**: https://github.com/Bubblevan/colluscope

Appendix A. Limitations and Dual-Use / Ethical Considerations

Limitations. Our evaluation uses constructed source modifications over a limited collection of saved multi-agent traces. While the results validate the source-integrity mechanism within this setting, they do not measure general collusion-detection accuracy. The framework assumes an authoritative source projection and separates directly verifiable facts from claims requiring additional semantic or causal evidence.

Dual-Use and Ethical Considerations. Evidence-auditing tools can strengthen oversight, but their verification mechanisms may also reveal which forms of record manipulation are detectable. We therefore provide sanitized public artifacts while restricting access to private messages and sensitive operational records. ColluScope is intended to support evidence-based investigation and human review rather than automated accusations of collusion.

Author Contributions

Bao Bowen designed and integrated the research project, reviewed the experimental evidence, and takes responsibility for its results and conclusions.

References

[1] Fish, S., Gonczarowski, Y. A., & Shorrer, R. I. (2024). *Algorithmic Collusion by Large Language Models*. arXiv:2404.00806. https://arxiv.org/abs/2404.00806
[2] Nakamura, M., et al. (2026). *Colosseum: Auditing Collusion in Cooperative Multi-Agent Systems*. arXiv:2602.15198. https://arxiv.org/abs/2602.15198
[3] Keppo, J., Li, Y., Tsoukalas, G., & Yuan, N. (2026). *On the Fragility of AI Agent Collusion*. arXiv:2603.20281. https://arxiv.org/abs/2603.20281
[4] 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 ICML 2026, PMLR 306. https://proceedings.mlr.press/v306/arike26a.html
[5] Gao, T., Yen, H., Yu, J., & Chen, D. (2023). *Enabling Large Language Models to Generate Text with Citations*. EMNLP 2023. https://aclanthology.org/2023.emnlp-main.398/
[6] Wu, E., Castelo, S., Liu, Y., Silva, C. T., & Freire, J. (2026). *AgentTrails: Towards Trust and Reuse for Agentic Tasks*. arXiv:2607.18816. https://arxiv.org/abs/2607.18816
[7] Apart Research. (2026). *AI Collusion Research Sprint: Track 2.3, What Should Count as Convincing Evidence of Collusion?*
https://apartresearch.com/sprints/ai-collusion-research-sprint-2026-10-23-to-2026-10-25

LLM Usage Statement

LLM tools assisted with coding, analysis, validation, and manuscript drafting. Saved model outputs are not represented as human annotations. The author reviewed the reported results against the available research artifacts and remains responsible for the claims in this report.