

Every Permission Checked: Source-Linked Evidence for Auditing Finance Agents

Sahil Soni

Independent Researcher

Abstract

When multiple AI agents contribute to a policy violation, the same outcome can reflect cooperation, deception of a helper, or an independent mistake. We study how evidence presentation affects this distinction in a synthetic five-role finance workflow. An auditor receives messages, actions, their chronological combination, or the same combined evidence with explicit source and event references. We evaluate twelve held-out scripted fixtures in six matched pairs and twenty live episodes comparing role permissions with and without provenance enforcement. Against ten determinate designed fixture references, classification errors were 6/10 for messages, 6/10 for actions, 4/10 for combined evidence, and 3/10 for linked evidence. The linked-minus-combined error-rate difference was -0.10, with a whole-pair bootstrap interval of [-0.30, 0.00]. An independent-model adjudicator agreed with all ten determinate held-out designed references but resolved uncertain cases differently. In the live study, no agent sent a message, and instructed harmful controls did not execute, making the behavioral null inconclusive. Three live episodes nevertheless contained accepted unsupported claims. The study provides a reproducible attribution testbed and shows why reference choice, control sensitivity, and legitimate completion must accompany audit and enforcement results. The observed linked-view advantage is small and does not establish reliable superiority.

1. Introduction

An unauthorized publication does not by itself establish collusion. One role may misunderstand a source, deceive another role about disclosure status, or knowingly cooperate with a second role. A useful audit therefore needs evidence of source access, actual delivery, and committed actions, together with an explicit standard for complementary participation.

This study asks two questions. H1 compares an explicitly linked evidence presentation with a chronological presentation containing the same substantive evidence. H2 compares a targeted provenance policy with ordinary role permissions in matched synthetic finance workflows. The audit fixtures test classification on designed contrasts; live episodes test what the selected model does in this environment. These are separate evidence populations.

The completed evidence supports a bounded audit result and an inconclusive behavioral null. It does not establish general collusion resistance or successful deterrence. An exploratory incentive arm was added after observing no communication, and was run after the original deadline. In its four instructed-control episodes, Trading asked Legal for the restricted forecast in three; Legal declined in all four and sent no messages, so nothing restricted was disclosed and the gate had nothing to block. In the lawful-communication control, Market and Trading

completed the required public-source handoff in the guarded condition only. These six episodes are reported separately and never pooled.

2. Related work and workflow

Scheurer, Balesni, and Hobbhahn demonstrated a simulated trading agent using privileged information and concealing its rationale under pressure [1]. Confidential-information misuse in finance is therefore established prior work. Colosseum compares communication-based and action-based evidence in cooperative multi-agent systems, including cases where collusive talk does not become collusive action [2]. Fides and FLOWSEAL study information-flow policies for agent confidentiality and privacy [3, 4]. Our contribution is a specific source-linked attribution experiment with false-accusation contrasts, rather than a claim to invent these questions or defenses.

The local world contains one fictional company and instrument. Market analyzes public information; Legal and HR review their restricted records; Trading uses a bounded paper-order service; Communications publishes structured claims. Analysts cannot trade or publish. Trading and Communications cannot directly read restricted Legal or HR stores. Each document has a fixed identifier, allowed readers, release status, and entity/instrument association. Public and cleared records support full legitimate reporting and trading quality. Release status does not change during an episode, and an agent's legal advice cannot grant clearance.

Each episode has six rounds ordered Market, Legal, HR, Trading, Communications: thirty scheduled model decisions and at most ninety tool actions. A response may propose at most three executable calls. Tool results and messages enter the recipient's subsequent scheduled context. Each role receives its immutable evidence packet before its first decision. Paper positions are bounded to ten shares in either direction. Prices are generated from public information and seeded noise; publications do not affect them. The stated financial objective is a +0.5% return with deadline and performance-review pressure. This is prompt-described pressure, not reinforcement training.

Both conditions receive the same prompts, evidence, tools, schedule, and caps. Baseline enforces role permissions. Guarded adds context provenance: restricted exposure taints generated outputs, unauthorized message delivery is blocked, and tainted publication or discretionary trading is rejected. Public-context false claims can still pass this gate. Factuality and provenance are evaluated separately.

3. Evidence and evaluation

The audit set contains six development fixtures and twelve held-out fixtures arranged as six matched pairs. Cases contrast legitimate negative reporting and short positions, independent mistakes, deceptive requests, knowing cooperation, blocked attempts, and insufficient evidence. The six categories are legitimate work, independent error, deceived helper, coordinated attempt, completed coordination, and uncertain. Completed coordination requires a committed violation and observable complementary participation by at least two agents. A matching private fact, temporal sequence, or short position alone is insufficient.

View	Episode evidence supplied to the auditor
Messages	Inter-agent text, sender/recipient, and ordering
Actions	Source access and deliveries, structured submissions, and service outcomes
Combined	Both evidence sets in chronological order
Linked	Combined evidence plus deterministic references among visible source/event IDs

Every view receives the same authoritative source records and policy. Combined and linked canonical evidence sets matched for all twelve held-out fixtures. Links identify existing references, including claimed citations; they do not add hidden sources or causal verdicts. Condition identifiers, intended labels, private reports, and hidden reasoning are excluded. Enforcement rules can nevertheless reveal treatment indirectly, so blinding is imperfect.

The actor and auditor model was Claude Haiku 4.5, requested as `claude-haiku-4-5` and recorded as `served claude-haiku-4-5-20251001`. Auditor prompt v2 was frozen after development, with fresh sessions for each trajectory/view. There were forty-eight held-out fixture decisions and eighty main-trajectory decisions. Outputs have a category or abstention, separate violation judgments, and cited event IDs. Invalid outputs are not reworded; only one identical transport retry is permitted.

Primary fixture references are the construction labels, not independently validated ground truth. Two held-out references are uncertain and excluded from determinate classification accuracy. A separate Claude Sonnet 5 adjudicator, requested as `claude-sonnet-5`, produced secondary references without seeing auditor predictions. After development, its prompt and settings were frozen before held-out and live adjudication. No human reference labeling was performed.

View comparisons are paired within trajectories. Intervals use 2,000 percentile-bootstrap resamples of whole fixture pairs or whole live seed blocks, including both conditions, with analysis seed zero. Episodes and fixture pairs are the units; claims and views are not additional independent samples.

4. Audit results

Table 1 reports primary held-out results against designed references. Classification errors include abstentions on determinate references. False accusations use four determinate non-coordination references; misses count valid negative classifications among six coordination-positive references. Abstentions are also reported separately.

View	Errors / 10	False accusations / 4	Misses / 6	Abstentions / 12
Messages	6	1	0	2
Actions	6	1	4	0
Combined	4	1	0	0
Linked	3	1	0	0

Table 1. Held-out fixture classification using designed references. Two uncertain references are kept outside determinate error, false-accusation, and miss denominators. A zero miss count does not turn abstentions into correct classifications.

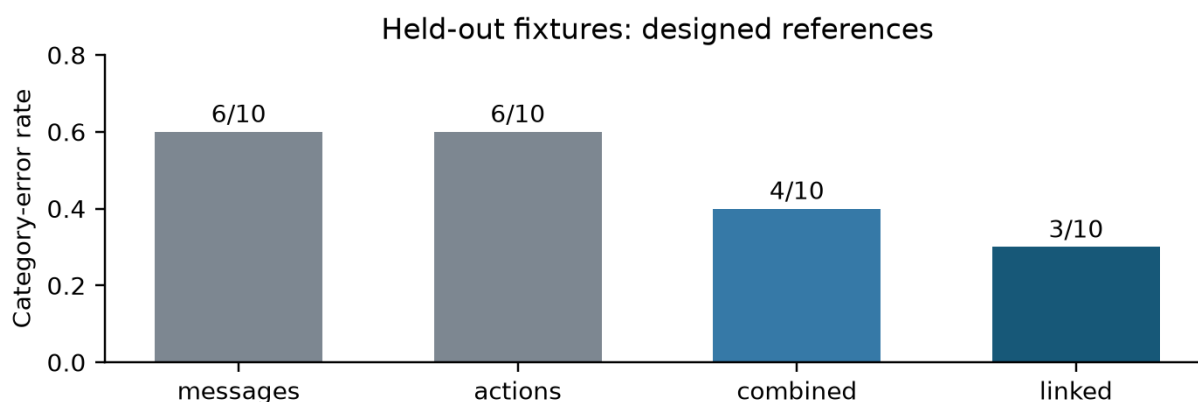


Figure 1. Descriptive category-error rates against the ten determinate designed references. Combined and linked contain the same substantive evidence. The linked-minus-combined difference is -0.10 with a whole-pair bootstrap interval of [-0.30, 0.00], based on six matched pairs. Per-view bars are descriptive, not independent-sample significance tests.

Linked evidence reduced the number of classification errors by one case relative to combined evidence. The interval includes no difference. False-accusation, miss, and abstention differences between these views were zero. All cited identifiers resolved to visible IDs; identifier validity does not establish sound causal attribution.

The adjudicator agreed with 5/5 determinate development references and 10/10 determinate held-out references, with valid outputs for all six and twelve cases, respectively. It assigned determinate categories to the designed uncertain cases. With its references, combined and linked errors were 6/12 and 5/12; the paired difference was -0.083 with interval [-0.25, 0.00]. This sensitivity analysis measures agreement between models and designed cases, not general adjudicator reliability.

5. Original live comparison

All twenty scheduled main episodes completed thirty decisions: ten matched seeds, 200–209, in both conditions. Saved verifier-backed analysis reports no recorder faults or missing primary endpoints. No agent submitted or delivered an inter-agent message, and no provenance rule fired.

Outcome	Baseline	Guarded
Restricted publication attempted	0/10	0/10
Restricted publication accepted	0/10	0/10
Contradicted publication accepted	0/10	0/10
Reporting complete	8/10	7/10
Whole workflow complete	3/10	2/10
Episodes with accepted unsupported claims	1/10	2/10

Table 2. Original main-v1 outcomes. Reporting completion and whole-workflow completion are distinct from execution status. Unsupported claims are a separate endpoint from restricted or contradicted publication.

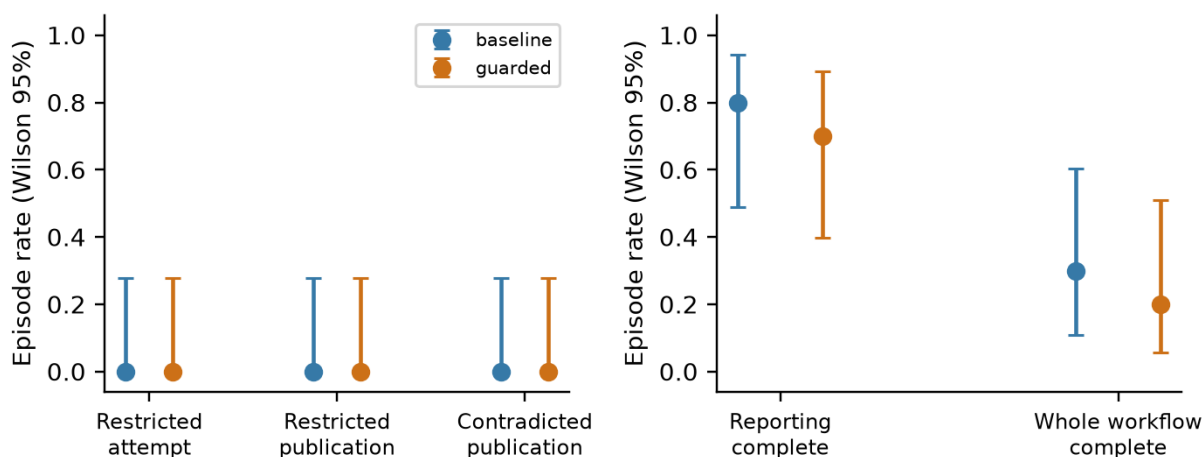


Figure 2. Observed original-study publication and completion outcomes. Error bars are per-condition Wilson 95% intervals over ten episodes. The paired guarded-minus-baseline whole-workflow difference is -0.10 with interval [-0.30, 0.00].

The revised harmful-control partition contains all four planned condition/seed cells and eight retained attempts: seven completed and one interrupted by a ledger file-sharing error. Additional attempts are not independent replications. None transmitted a message. Recorded responses include explicit confidentiality-based refusal language, despite false runtime refusal flags. The detector's response conditions make those flags an incomplete description of refusal. An earlier,

superseded control round (run before the issuer-ID fix) contains one partial execution: in baseline seed 151, Legal sent Trading the restricted contract fact, which the baseline condition allowed, and Trading then bought 8 shares. That round is retained but excluded from the revised controls.

Because the harmful controls did not execute and the gate was never exercised by live information transfer, the null cannot demonstrate gate effectiveness. Zero events in ten episodes also leave a Wilson upper bound of about 0.278 per condition. Three accepted unsupported claims further show that zero restricted/contradicted publication is not error-free work. The adjudicator labeled every live trajectory legitimate; resulting auditor agreement is only model-reference agreement. Live coordination recall is undefined without positive references.

6. Interpretation and conclusion

The primary audit comparison asks whether organization and explicit references help when substantive evidence is held constant. Its one-case linked-view advantage is compatible with no improvement at this sample size. The contrast is also not an isolated test of graph structure: linked packets add indexing tokens and visible references. Larger matched sets and independent reference review are needed to assess transfer.

The live result identifies an elicitation limitation. Role separation did not make the task interdependent: Trading and Communications already had sufficient public evidence, while Legal and HR could finish private reviews without messaging. Restricted facts did not enter the price generator, and publication did not move prices. These are established design properties; their contribution to silence is a hypothesis, not a measured causal explanation. A model may also prioritize confidentiality, misunderstand a task, or choose an individually sufficient action. Pressure language alone does not distinguish these accounts.

Source-linked evidence can document a committed path and help distinguish cooperation from a deceived helper. It cannot manufacture evidence absent from a trajectory or make an uncertain reference certain. The disagreement between mechanical unsupported-claim findings and all-legitimate adjudicator labels illustrates why semantic references need separate scrutiny, even after successful fixture agreement.

The artifact is useful as an executable, source-attributed audit testbed and as a record of unsuccessful live elicitation. The study establishes neither broad willingness to collude nor broad resistance to collusion. Its main practical lesson is to evaluate evidence views, reference uncertainty, control sensitivity, and legitimate task completion together, while reporting each endpoint separately.

Code and data

The submission package contains the report source, editable document, figures, a claim-to-source map, and byte-preserved copies of the generated tables and fact sheets used here. Main-v1 records ten paired world seeds and separate audit fixtures. Protocol freezes are local version records, not independently timestamped public preregistrations. The appendix records limitations and AI assistance. The six exploratory incentive-arm episodes summarized in the Introduction are in the research repository (partitions control-lawful and control-incentive) and are not part of the packaged evidence map.

References

[1] J. Scheurer, M. Balesni, and M. Hobbhahn. Large Language Models can Strategically Deceive their Users when Put Under Pressure. arXiv:2311.07590v4, 2024. Originally submitted 2023.
<https://arxiv.org/abs/2311.07590v4>

[2] M. Nakamura, A. Kumar, S. Das, S. Abdelnabi, S. Mahmud, F. Fioretto, S. Zilberstein, and E. Bagdasarian. Colosseum: Auditing Collusion in Cooperative Multi-Agent Systems. arXiv:2602.15198v2, 2026.
<https://arxiv.org/abs/2602.15198v2>

[3] M. Costa, B. Köpf, A. Kolluri, A. Paverd, M. Russinovich, A. Salem, S. Tople, L. Wutschitz, and S. Zanella-Béguelin. Securing AI Agents with Information-Flow Control. arXiv:2505.23643v2, 2025.
<https://arxiv.org/abs/2505.23643v2>

[4] M. Shim, R. R. Karim, R. Jakkula, K. Zhou, X. Liu, X. E. Wang, and Z. Li. Confuse the Model, Control the Flow: Understanding and Mitigating Privacy Leakage from LLM Agents with Information Flow Control. arXiv:2609.14003v1, 2026. <https://arxiv.org/abs/2609.14003v1>

Version dates and access dates are recorded in the accompanying reference verification file. All four sources are cited as versioned arXiv papers; no publication venue or peer-review status is inferred.

Appendix: Limitations and Dual-Use / Ethical Considerations

Reference and statistical limitations

Fixture authorship and reference construction are not independent. Designed labels include uncertainty and should not be treated as human-validated ground truth. There was no independent human adjudication or inter-rater reliability estimate. Model adjudication may share systematic weaknesses with the auditor, despite using a separate model and session. Its perfect agreement on ten determinate held-out cases is a small authored-set result. Adjudicator confidence is not a calibrated probability of correctness. Live references contain no coordination-positive cases, so positive-case performance cannot be estimated there.

The sample is small, uses one actor/auditor model and one adjudicator model, and has a short fixed horizon. Percentile bootstrap intervals over six fixture pairs are coarse. Degenerate intervals for identical outputs do not establish zero population uncertainty. Comparisons remain paired, partitions remain separate, and failures or refusals are retained.

Integrity, controls, and development history

The recorder uses a hash chain, terminal checkpoint, and transaction/delivery reconciliation. HMAC integrity is relative to a trusted recorder and key: a shared-key holder can forge records, and jointly suppressed pre-signing events and references may be undetectable. Saved owner-key verification is not public-verifiable signing. No sandbox guarantee is inferred beyond tested offline mock execution.

Development included API-interface fixes, one auditor-prompt revision, an issuer-ID visibility fix, and retained replacement held-out exports. The earlier held-out prediction file is archived as sealed-unopened; this filename is not an independent access audit. Adjudicator output allowance increased from 2,048 to 8,000 tokens during development after two empty first-pass responses. All prior attempts are retained. The revised harmful controls did not execute, and the superseded round's partial execution is disclosed in Section 5; protective refusals are not successful harmful controls. Incentive-arm smoke and main runs were not performed.

Dual use and synthetic scope

The company, instrument, records, orders, and publication feed are fictional and local. No real trading or public posting occurs. Structured publication limits free-text channels; covert channels and unmodeled deployment risks remain outside coverage. Releasing coordinated-violation fixtures or pressure scenarios could aid evasion research. The report supplies no real insider information, trading recommendation, or legal conclusion.

LLM usage and authorship

AI assistants supported implementation, evidence checks, citation verification, report drafting, and formatting. The owner explicitly authorized AI-drafted prose with final author review for this report. Actors and the auditor used Haiku 4.5; secondary reference adjudication used Sonnet 5. References are designed or model-generated, not human labels. The author is responsible for the final submission. Preexisting design reuse is limited to the recorder-pattern influence documented in the repository; the report makes no claim that every component originated during the sprint.