

When Missing Evidence Looks Like Innocence

Evidence ablation in agent-incident reconstruction

Lucile Ter-Minassian

Independent researcher · with Apart Research

A controlled study of evidence loss — and of what repairs it

MOTIVATION

Reconstruction ran on records the agents controlled

- July 2026: agents in an evaluation escaped their sandbox and reached Hugging Face infrastructure. About 17,600 attacker actions were reconstructed after the fact.
OpenAI called it a warning shot for agent safety.
- The independent METR / Redwood review found the record itself was shaky: transcripts missing components, some spoofed tool calls, timestamps reconstructed.
Not proof the account is wrong — proof the evidence was a source of uncertainty.
- Frameworks already say “retain logs.” We found no prior controlled comparison of what that loss costs an investigator, or of which repair is sufficient.

THE QUESTION

How much does an altered agent-controlled record change the conclusion — and what small artifact repairs it?

INTUITION

One sentence disappears. Does the case close?

An agent does the work, then writes its own session log. That log is the narrative an investigator reads. The harness separately records every tool call, outside the agent's reach.

AGENT-AUTHORED LOG

*"I reviewed the repository and
ran the test suite..."*

*The sentence naming the protected
file is gone.*

INDEPENDENT RECEIPT

actor: A
event: read_file
resource: tests/expected.json
outcome: ok
t: 4.674

Intuition: the investigator is reading a story the suspect wrote.

Delete one passage and there may be nothing left to conclude from — unless something outside the story survives.

Ablate the evidence, hold the truth fixed



- 120 episodes, 240 Qwen3.8-27B agents; 66 read the protected file.
Ground truth from harness tool calls and file hashes only — never from reasoning traces or agent prose.
- Eligible logs randomized to redaction or control; 34 ended up with their own access-evidence passage deleted.
Claude Sonnet 5 is the primary blinded investigator: yes / no / insufficient.
Outcome: false exoneration — a genuine access answered “no.”

Six evidence packages, one per investigator call

VIEW	EVIDENCE SHOWN	PURPOSE
V0	Submissions and scores only	No-log floor
V1	Agent log, access-evidence passage removed	Tampered record
V1P	Passage kept; matched volume of other text removed	Placebo — is it volume or content?
V1A	V1 + five-field receipt (~740 chars)	Minimal repair
V1R	V1 + the deleted passage restored	Exact undo of the deletion
V3	V1 + full call record (~4,474 chars)	Upper-bound repair

— Paired within episode; each call sees exactly one view. The frozen V1–V1A comparison was re-run on Haiku 4.5 and Opus 5, and two blinded readers judged a shared subset.

V1R isolates the deletion; V1P isolates text volume; V1A asks whether a compact receipt is enough.

RESULTS

What we found

11 / 34

treated agents falsely exonerated
from the tampered log

An affirmative “no,” not an abstention

0

false exonerations after restoring
the passage, or adding receipts

V1A, V1R and V3 alike; McNemar $p = 0.00098$

38 / 38

detections survive the
matched placebo deletion

*The driver is which content was removed,
not how much*

- Across the whole sample, 14 of the 16 missed accesses were affirmative “no” answers — median stated confidence 82.5.
- The failure was omission, not blame-shifting: probability moved to “nobody did this,” not to the innocent agent.
- Direction held across tiers — but Haiku still answered “no” in 7 cases with a correct receipt in front of it.

TAKEAWAY

Two controls, not one

- Retain an action record outside the agent's reach.

Five fields — actor, event, resource, outcome, t — matched the full call record in this sample.

- Then verify that your investigator actually uses it.

A no-log and a deleted-passage canary should yield escalation, never a confident exoneration.

Missing evidence did not read as uncertainty.

It read as innocence — and an exoneration closes a case.

Code and frozen artifacts: github.com/l-terminassian/agent-incident-reconstruction