A model says it followed policy. The action log says retrieval failed. Three witnesses, one run.

Dr Peter McCann Strain
CTO, DPhil/PhD in AI from Oxford University

Swipe ››

— THE FRAMEWORK GAP

An agent leaves three kinds of witness: self-report, action log and independent judgement. Collapse them and you get the appearance of audit. Weigh them against each other and you get a case file.

— THE REFRAME

What corroborates the trace.

THE OLD QUESTION

Does the trace exist?

↓ ↓ ↓

THE QUESTION THAT HOLDS UP

What kind of evidence is the trace, and what weight does it deserve?

— WHY THE FRAME HOLDS

+76% attribution lift

Chen et al. ('Seeing the Whole Elephant', 2026 preprint) found full execution traces lifted failure-attribution accuracy by up to 76 percent over partial observation. Yet peer-reviewed Who&When (Zhang et al., ICML 2025) caps it: even the best method falls short of reliable attribution. Necessary, not sufficient.

SOURCE

Zhang et al., 'Who&When: A Benchmark for Failure Attribution in LLM Multi-Agent Systems' (ICML 2025 Spotlight, OpenReview); Yuan et al., 'R-Judge: Benchmarking Safety Risk Awareness for LLM Agents' (Findings of EMNLP 2024); Turpin et al. (NeurIPS 2023) on chain-of-thought faithfulness; OpenTelemetry GenAI and OpenInference trace specifications.

HOW IT WORKS

Weight rises from self-report to action log to independent judgement.

01 Treat **self-report** as testimony only: a useful clue, never proof.

02 Reconcile the **action log** against an **external ledger** (the system of record outside the agent: database, CRM, payment rail).

03 Run **independent judgement**, a structurally different evaluator (different model family, deterministic rules, or human reviewer). Even the best automated method falls well short of reliable attribution at either the agent or the step level.

— THE ARTIFACT

Three witness families, five evidence channels.

Self-report

What the model says

Action log

What the harness recorded

Independent judgment

What a separate checker concludes

Three witness families (self-report, action log, and independent judgement) expanded into five evidence channels in the full essay matrix.

APPLY THE INSTRUMENT

Pick one consequential run this week. Fill the evidence-weight matrix: self-report, action log, external ledger, independent verifier, human domain review. For each row, write what it says, what it proves, and the question it cannot answer. Name the resolver where two rows disagree.

—— WHERE TEAMS MISREAD IT

Treating the reasoning trace as the audit trail. The model narrates, the harness records, the verifier judges: three different objects. Wire all three to the same run ID before calling the loop closed.

READ THE FULL FRAMEWORK

The Three Witnesses to a Run

Read the full essay — the argument, the sources, the figures and a reader-ready working artifact.
Substack petermccannstrain.substack.com · Medium @peter.mccann.strain · LinkedIn peter-strain-dphil-15a607128
New essays twice weekly, 2 June – 21 July 2026.

Next: E18 — When an Agent Acts, Who Acted?

— THE STACK SO FAR

E17 · Essay 17 of 22 complete · Arc IV: Proof and accountability

YOU JUST ADDED

The three-witness matrix

STACK LAYER LIT UP

Runtime evidence / Checks

YOU CAN NOW ASK

weigh self-report, action log, and independent judgment.

NEXT

E18 asks who actually acted when an agent acts.

— THE ARTIFACT, CONTINUED

Three witness families, five evidence channels.

THE REMAINING NODES

External ledger: the system of record outside the agent that the action log must reconcile against.

Human domain review: the practitioner whose judgement a checker cannot substitute for.

Worked example: claim 'refund issued, customer notified.' Self-report says 'done.' Action log shows refund API call but no email send. External ledger (payment rail) confirms refund. Independent judgement: refund yes, notification no.

Empty rows: when a witness is missing, write the claim you can no longer make and the question you can no longer answer.

Dr Peter McCann Strain

CTO, DPhil/PhD in AI from Oxford University

I build production AI systems and write about making agentic AI useful, inspectable, governable and safe enough for real work.

Follow on Substack for the full 22-essay series
petermccannstrain.substack.com