An Adversarial, Multi-Round Novelty Audit

INDEPENDENT RESEARCHER · 2026

The task

Before committing to a research contribution, establish honestly whether it is actually new. The setting makes this unusually hard. Work on agent safety, abstention, runtime enforcement, and agent-failure analysis is appearing almost weekly, much of it as preprints, and the closest neighbours are often only weeks old and not yet peer reviewed. The failure mode to avoid is the one that survives until a reviewer checks it: an overstated novelty claim, or a citation that turns out not to say what it was cited for. I treated the audit as adversarial by design, meaning that its job was to try to defeat my own claim rather than to confirm it.

Method

I ran the audit across three kinds of search surface, the open web, arXiv, and scholarly library-discovery databases, over seven rounds spread through the project rather than as a single pass. Spreading it out mattered, because the field kept moving, and a claim that looked safe in an early round had to survive neighbours that only appeared later. I read the substantive sources in full, cover to cover, rather than at abstract level, more than eighty papers, plus several proceedings volumes held as reference collections for the formal-methods, runtime-verification, multi-agent

Throughout, I held a strict evidence discipline that governs every substantive claim. Each one carries one of three tags: traced to a publication actually retrieved, my own reasoning, or plausible but unconfirmed. No citation, author, venue, year, or statistic is stated without being traced to a real source, and abstract-level knowledge is treated explicitly as insufficient for a citation, so anything read only at abstract level stays unconfirmed until it is read in full. Where a search returned nothing, I recorded that it returned nothing rather than quietly moving on. Source quality was part of the audit, not an afterthought: where a venue showed the markers of a predatory or self-asserted peer review, an in-house identifier prefix, a self-citation cluster, a future-dated issue, I recorded that and treated the work as grey literature to be cited only as an adjacent pointer, never as evidence about the world.

The hard part, done honestly

The audit repeatedly narrowed the claim, and I let it. Round by round, as strong neighbours surfaced, I conceded the parts of the idea that were already occupied and re-based the contribution onto the part that genuinely was not, rather than defending the original framing. It is worth setting out the three lines of prior work that did most of that narrowing, because the concessions are the substance of the result.

The first line is deterministic gates placed before an agent action. These exist, and several are strong, but they assess a different object. Pre-action authorisation evaluates capability and identity and is explicit that it authorises actions, not content (**Open Agent Passport, 2026**). A rule-based enforcement language performs deterministic checks at execution checkpoints over state and action (**AgentSpec, ICSE 2026**). An execution-authority layer generalises the classic reference monitor with non-compensatory aggregation and single-use permit tokens (**L-DREA, 2026**). A pre-action legitimacy boundary allows an action only when every required predicate holds (**Lavi, 2026**), and a bounded-autonomy controller adds an escalation ladder over allow, deny and approve (**Sherry et al., 2026**). Others gate on learned or safety-policy compliance rather than pure determinism (GuardAgent, 2025; **ShieldAgent, ICML 2025**; AGrail, 2025; TemporalGuard, 2026). Reading these in full established that the bare idea of a deterministic pre-execution gate over a probabilistic agent is not novel, so the contribution could not rest on it. What they share is an object that is not evidence sufficiency: authority, policy, legitimacy, or safety compliance.

The second line names evidence sufficiency and provenance directly, which was the more uncomfortable discovery, because it touches the vocabulary of the contribution. A review of auditable au-

ability layer with a six-dimensional taxonomy (**From Agent Traces to Trust, 2026**). A benchmark puts governance-evidence sufficiency in its title and scores whether emitted evidence reconstructs a past decision (**DEMM-Bench, 2026**). A misalignment detector formalises whether a proposed tool call is supported by traceable evidence, though it resolves to a binary aligned or misaligned via a model (**ProvenanceGuard, ASE 2026**). These conceded the term and the framing. What none of them do is index failures by the condition of the evidence and couple that index to a deterministic act, defer or abstain gate, which is where the residual claim had to live.

The third line derives failure taxonomies from execution traces, which forced me to give up the idea of a first trace-derived agent-failure taxonomy outright. A grounded-theory taxonomy is built over a large annotated multi-agent trace corpus and indexed to system design, coordination and verification (**MAST, NeurIPS 2025**). A trajectory-diagnosis method localises the first unrecoverable step and assigns a root-cause category (**AgentRx, 2026**). An agent-environment taxonomy annotates failed traces by exploration, exploitation and resource use (**Aegis, 2025**). A prospective-reflection framework distils a planning-error taxonomy from historical trajectories (**PreFlect, 2026**). A survey synthesises fifty-five such works into a meta-taxonomy (**LLM-agent trajectory analysis survey, 2026**). Trace derivation is thoroughly occupied. In the decisive rounds I retrieved and read the closest of these in full rather than trusting abstracts, and two later works pre-empted a sub-claim I had been treating as novel: a production-runtime taxonomy that tags a sandbox-denied observation as first-class, which partly anticipates the not-observed versus observed-absent idea as forensic tagging (**silent-failures taxonomy, 2026**), and the governance benchmark above, which types absence as a scoring label. I corrected the positioning and recorded each concession in the open.

Why reading in full, not abstracts, was load-bearing

The rule that abstract-level knowledge is insufficient for a citation is easy to state and expensive to keep, and this audit is where it earned its place, because on more than one occasion the full text of a paper said something different from its abstract, and the difference changed a conclusion.

A pre-execution refinement method for tool-use agents looked, at abstract level, like a deterministic checker and therefore a direct competitor on the determinism axis. Reading it in full reversed that: its rubric checks are judged by a language model, not by a deterministic algorithm, and the paper's own ablation shows that a purely deterministic abstract-syntax-and-signature checker per-

A second case was quantitative. An instrument-control study reports a headline success rate that its abstract attaches to the multi-device collaborative setting, while the body makes clear that the headline figure is the single-instrument rate and the collaborative rate is lower (**Zhang et al., 2026**). Because the evidence discipline forbids repeating a statistic that has not been verified against the source, I did not carry the abstract's number, and I recorded the discrepancy rather than smoothing over it. The general lesson is that a citation formed from an abstract is a liability waiting to be found by a reviewer, and the only defence is to read the thing.

The audit also had to identify, and then defuse, the single closest work on the object itself. A misalignment detector formalises exactly the relation my mechanism cares about, whether a proposed tool call is supported by traceable evidence, and it was the biggest positioning risk of the whole set. Reading it in full is what made the differentiation precise rather than defensive: it resolves to a binary aligned or misaligned verdict reached by invoking a language model to judge the relation, so it is neither three-way nor deterministic, and it neither indexes failures by evidence condition nor couples to a remediable-defer outcome (**ProvenanceGuard, ASE 2026**). Naming that contrast exactly, rather than hoping a reviewer would not notice the neighbour, is the difference between a claim that survives review and one that collapses on contact with it.

Two further works sit on the boundary this project must not cross, and they are cited precisely because they show where a neighbour steps over the line. A deterministic layer around GUI-testing agents issues pass, fail and inconclusive verdicts about the system under test (**Salva, 2026**), and the instrument-control system above emits a physical steady-state readiness verdict for the device (**Zhang et al., 2026**). Both are legitimate in their own settings, and both produce the external-state verdict that my boundary rule forbids, which makes them the cleanest possible illustration of what this project deliberately does not do.

What survived, and its residual risk

What remains, stated with its limits named, is the joint of two things that the neighbours have separately but not together: indexing agent failures by the condition of the evidence, and coupling that index to a deterministic gate that resolves to act, defer or abstain, with two refinements that survived, a deterministic split between a remediable defer and a terminal abstain, and the not-observed versus observed-absent primitive on the surface. To keep that defensible in review, I pinned four specific terminology contrasts against the nearest works, so a reader cannot mistake a shared word for a shared contribution: readiness and guardrail verdicts about a system under

2026). The residual risk is stated plainly: several of the closest works are recent preprints, the field is active, and a reviewer could still surface something not yet read. That is a defensible position, and it is worth more than an inflated one that does not survive review.

What it demonstrates

Literature synthesis at depth across a fast-moving field, intellectual honesty under the standing pressure to claim more than the evidence supports, source-quality judgement rather than uncritical citation, and a repeatable, auditable process that another person could follow and check against the same 97-item evidence base.

Selected references (a subset of the audited set)

- Cemri, M., Pan, M.Z., et al. (2025). Why Do Multi-Agent LLM Systems Fail? NeurIPS 2025 Datasets and Benchmarks. arXiv:2503.13657.
- Barke, S., et al. (2026). AgentRx: Diagnosing AI Agent Failures from Execution Trajectories. arXiv:2602.02475.
- Song, K., et al. (2025). Aegis: Taxonomy and Optimizations for Overcoming Agent-Environment Failures in LLM Agents. arXiv:2508.19504.
- Wang, Y., Cao, Y., Lin, L., Chen, J. (2026). PreFlect: From Retrospective to Prospective Reflection in Large Language Model Agents. arXiv:2602.07187.
- Wang, S., et al. (2026). A Survey for LLM Agent Trajectory Analysis: From Failure Attribution to Enhancement. IEEE Transactions on Software Engineering, early access. DOI 10.1109/TSE.2026.3717765.
- Wu (2026). When Errors Become Narratives: A Longitudinal Taxonomy of Silent Failures in a Production LLM Agent Runtime. arXiv:2606.14589.
- Solozobov (2026). DEMM-Bench: A Cross-Regime Benchmark for Agent-Runtime Governance-Evidence Sufficiency. arXiv:2606.20634.
- Wang, Y., et al. (2026). From Agent Traces to Trust: A Survey of Evidence Tracing and Execution Provenance in LLM Agents. arXiv:2606.04990.
- Theodorakopoulos, L., Theodoropoulou, A. (2026). Auditable LLM Autonomy for Operational Decision-Making: Big Data Evidence and Decision Traces. Computers, Materials and Continua. DOI 10.32604/cmc.2026.082270.
- She, Liang, Kang (2026). Safeguarding LLM Agents from Misalignment through Provenance

- Wang, H., Poskitt, C., Sun, J. (2026). AgentSpec: Customizable Runtime Enforcement for Safe and Reliable LLM Agents. ICSE 2026. arXiv:2503.18666.
- Gill-Lakhowal (2026). Deterministic Runtime Enforcement: The Execution Authority for Autonomous AI Agents (L-DREA). IEEE Access, 14. DOI 10.1109/ACCESS.2026.3719838.
- Lavi, G. (2026). The Pre-Action Legitimacy Gap in AI Systems. arXiv:2604.24153.
- Chen, Z., Kang, M., Li, B. (2025). ShieldAgent: Shielding Agents via Verifiable Safety Policy Reasoning. ICML 2025 (PMLR 267). arXiv:2503.22738.
- LeVine, Evers, Saltwick, Venkatesh (2026). RubricRefine: Improving Tool-Use Agent Reliability with Training-Free Pre-Execution Refinement. arXiv:2605.09730.
- Salva, S. (2026). Reliable execution of natural language test cases for GUI applications using LLM agents. Software Quality Journal, 34, 33. DOI 10.1007/s11219-026-09767-2.
- Ahmed (2026). Database-Native Reasoning: Treating the Database as the Cognitive Substrate for AI Systems. IEEE Access, 14, 54912-54921. DOI 10.1109/ACCESS.2026.3675808.
- Zhang, Z., et al. (2026). LLM-Enabled Multi-Agent Collaborative Instrument Control for Automated Microelectronic Testing. Advanced Devices and Instrumentation, 7, 0202. DOI 10.34133/adi.0202.

Download PDF ↓