
What Should We Check Next?

Testing Ambiguity-Preserving Evidence Selection for AI Incident Investigation

Linda Thorstensen

Abstract

AI-agent incidents can produce a lot of telemetry without giving investigators a defensible single causal explanation. I built a small auditable investigator that keeps every explanation still compatible with observed evidence, marks cases UNRESOLVED or MODEL_GAP when the evidence cannot support closure, and uses the remaining disagreement to choose the next check. The mechanism, synthetic generator, seeds, comparators, endpoint, and PASS criteria were frozen before the first complete holdout result. Across 400 confusable-resolvable cases, AP-MINIMAX improved realized candidate reduction per unit cost over NEUTRAL-ALL by $+0.5196$ (95% CI $[0.4383, 0.5992]$) and over NEUTRAL-DISCRIMINATING by $+0.4213$ ($[0.3458, 0.4954]$); the direction was positive in all five seeds. All non-forced strategies eventually resolved all 400 cases correctly. The measured difference is in earlier evidence selection; final attribution accuracy was the same. I use the public OpenAI/Hugging Face incident as a worked handoff example, turning unresolved questions into specific evidence requests while keeping the synthetic result separate from real-world incident claims.

1. Introduction

The July 2026 OpenAI/Hugging Face incident gives a rare public view of an AI-agent investigation crossing technical and organizational boundaries. Hugging Face reconstructed roughly 17,600 attacker actions across short-lived environments and correlated agent-side traces with platform logs. OpenAI separately reconstructed the originating evaluation environment, control failures, unauthorized communications, internet access, and third-party impact [1–4]. The public record is unusually detailed, yet it is retrospective and assembled from evidence held by different organizations and systems.

I am interested in the stage before that completed timeline exists. Responders start with fragments: anomalous credential use, an unexpected network path, a suspicious agent trace, a control failure, a third-party artifact, or a local alert whose significance is unclear. Several explanations may fit the same evidence, and a plausible working story can harden into the answer before the evidence has separated it from the alternatives.

Recent disclosures show the same problem from different angles. Anthropic reported that an initial search of roughly 141,000 transcripts missed incidents later found during follow-up, after which the search widened to roughly 481 million transcripts [12]. Ruby Central confirmed the May RubyGems malicious-package campaign and related attempted API-key acquisition, while saying it could not independently determine whether AI agents created or published the packages; researchers attributed the activity to OpenAI agents [13]. OpenAI's current incident hub describes an expanding review of third-party activity

beyond the initially recognized Hugging Face compromise [14]. Evidence retrieval, attribution, and even the recognized boundary of an incident can stay provisional for some time.

For this sprint I ask a much narrower Track 2 question:

Given the explanations still compatible with what is currently known, what evidence should an investigator obtain next?

I leave attribution to investigators. The artifact keeps unresolved cases visible, makes unsupported closure harder, and turns causal disagreement into a specific request that an evidence owner can actually execute.

I place the work along this sequence: logging → evidence preservation → reconstruction → competing explanations → discriminating next observation → causal resolution.

The sprint contribution sits between competing explanations and the next discriminating piece of evidence (Figure 1).

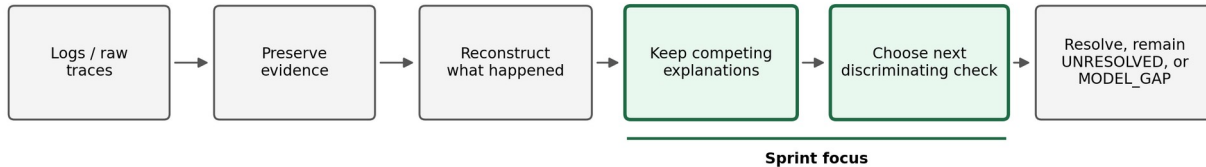


Figure 1. Where the sprint artifact sits in the incident-investigation workflow. The tested mechanism starts once an evidentiary state exists and focuses on keeping compatible explanations visible and selecting the next discriminating check.

2. Related Work and Contribution Boundary

The underlying investigation principles are established. Active diagnosis studies how to identify faults by sequentially selecting informative queries or tests [5]. Sequential model-based diagnosis maintains competing diagnoses and chooses further queries, including cost-aware query optimization [6]. Bayesian experimental design formalizes measurement choice through expected information gain [7].

Digital forensics is similarly close prior art. Carrier treats statements about digital states and events as hypotheses that must be tested [8]. Pasquale et al. explicitly connect speculative investigative hypotheses to evidence that should be collected and analyzed, including ephemeral evidence and further evidence needed after an investigation begins [9]. Paeth et al. discuss structural ambiguity and unavoidable epistemic uncertainty in AI incident records [10].

Brömmé's September 2026 position and architecture paper sits next to this work at an earlier stage. It proposes a vendor-neutral 'black box' for agentic processes using cryptographic commitments for selected agent communications, tool interactions, approvals, and process artifacts, with applications to audit, incident reconstruction, and regulatory reporting [11]. Its main question is whether process evidence can be preserved with stronger integrity for later verification. My starting point is a current evidentiary state that still supports several causal explanations.

These are three related stages of the same broader problem:

1. Evidence preservation: can reliable records survive and later be audited?
2. Incident reconstruction: what sequence can be reconstructed from available records?
3. Unresolved causal ambiguity: when several explanations still fit, what should be checked next?

Hypothesis testing, evidence requirements, explicit uncertainty, black-box logging, incident reconstruction, active diagnosis, minimax discrimination, and information gain are all established ideas [5–11]. They are the base I build on here.

My contribution is smaller and more specific: a compact, auditable AI-incident implementation whose next-check behavior was specified in advance, stress-tested on frozen confusable cases, and then used to build a source-grounded cross-organizational evidence handoff for one documented incident.

The empirical question is whether a frozen ambiguity-preserving, discrimination-first selector with cost-aware tie-breaking behaves usefully under controlled incomplete evidence while keeping non-resolution states visible.

3. Methods and Artifact

The reviewer-facing Track 2 artifact is reproduced in Appendix B as the Next Evidence Request Packet v0.1. Underneath it, the implementation maintains the current explanation set, evaluates available checks, records why one is selected, and turns the unresolved question into a handoff naming the evidence owner, requested records, predicted outcome classes, and resolution conditions.

3.1 State and compatibility

At each investigation step the artifact receives currently observed evidence; a set of modeled causal explanations; the outcomes each explanation considers compatible for each available check; and check costs and neutral tie ranks frozen independently of explanation names and order.

An explanation remains alive while every observed item is compatible with it. It is eliminated only after incompatible evidence is actually observed. Future outcomes are unavailable to the selector.

- **RESOLVED** — exactly one modeled explanation remains compatible.
- **UNRESOLVED** — multiple explanations remain compatible, including cases where no available check can distinguish them.
- **MODEL_GAP** — no modeled explanation remains compatible, indicating that the candidate set is inadequate rather than forcing an attribution.

I use a non-probabilistic implementation because defensible priors and likelihoods are unavailable for these cases.

3.2 Frozen internal selection rule

Let the surviving explanation set be H . For each available check p , the artifact considers every outcome compatible with at least one surviving explanation and computes how many explanations would remain if that outcome were observed. The score is the worst-case residual ambiguity: the largest surviving subset across possible outcomes.

The frozen AP-MINIMAX rule selects:

1. the check with the smallest worst-case residual ambiguity;
2. if tied, the lower check cost;
3. if still tied, a seeded neutral rank fixed independently of explanation names and input order.

The rule is intentionally simple and inspectable. I use it as the test mechanism for the ambiguity-preserving handoff; optimal diagnosis is outside the claim.

3.3 Evidence-request output

A selected check matters only if someone can act on it. Each handoff records the investigation question, current competing explanations, evidence owner, exact records or fields requested, observable outcome classes, causal consequences, resolution condition, UNRESOLVED / MODEL_GAP condition, cost and availability friction, and a trace entry showing why the check was selected.

For Track 2, usefulness means being able to say: these explanations are still live; this evidence owner can answer this question; and these outcomes would change the causal picture in these specific ways.

4. Frozen Evaluation

4.1 Pre-result freeze

I froze the behavioral specification before the clean implementation. I then froze the experiment before headline holdout generation or inspection. The execution plan fixed the strata, seeds, case-generation constraints, comparators, endpoint, bootstrap seed, controls, and PASS/PARTIAL/FAIL criteria.

I preserved the first complete conforming holdout run with raw traces and independently recomputed analysis. Post-result changes to the mechanism or benchmark were prohibited for v0.1.

4.2 Holdout design

Five frozen holdout seeds produced 800 cases, 160 per seed:

- 80 confusable-resolvable;
- 20 clean-resolvable controls;
- 20 persistent-ambiguity controls;
- 20 model-gap controls;
- 20 no-failure / no-contradiction controls.

The primary H2 population was the 400 confusable-resolvable cases. Within each seed, hidden causes were evenly balanced across four synthetic families: evidence/retrieval failure, tool/action failure, system/state-context failure, and control/policy-context failure.

Confusable cases contained four modeled explanations, staged evidence, five available checks, independently generated costs in $\{1, 2, 3\}$, and neutral identifiers assigned after structural generation. Initial evidence left either three or four explanations alive. Full check signatures uniquely separated explanations in resolvable cases. Persistent-ambiguity cases were intentionally indistinguishable under all available checks; model-gap cases intentionally omitted the true explanation.

4.3 Comparators

FORCED-CLOSURE is a mechanism ablation. It immediately collapses an unresolved candidate set to one explanation using a seeded neutral rank.

NEUTRAL-ALL chooses the lowest neutral-rank unobserved check. NEUTRAL-DISCRIMINATING is a stronger comparator that chooses neutrally only among checks that currently distinguish at least two surviving explanations. INFORMATION-GAIN is a descriptive prior-art calibration using a uniform prior and deterministic outcome partitions. PASS was defined against the two neutral comparators.

4.4 Primary endpoint

For every confusable case, strategies started from the same unresolved state and saw the same pre-generated outcomes. The primary endpoint was: $\text{first_check_realized_value} = (\text{candidates_before} - \text{candidates_after_actual_outcome}) / \text{check_cost}$.

The comparison is paired by case. The frozen analysis uses 10,000 paired bootstrap resamples across the 400 confusable cases. H2 required:

1. AP-MINIMAX minus NEUTRAL-ALL > 0 with a 95% interval excluding zero;
2. positive direction in at least four of five seeds;
3. AP-MINIMAX minus NEUTRAL-DISCRIMINATING > 0 with a 95% interval excluding zero.

5. Results

Table 1. Paired first-check realized-value differences on the 400 confusable-resolvable holdout cases.

Comparison	Mean paired difference	95% paired-bootstrap interval	Positive seeds	Wins / ties / losses
AP-MINIMAX — NEUTRAL-ALL	+0.5196	[0.4383, 0.5992]	5/5	235 / 129 / 36
AP-MINIMAX — NEUTRAL-DISCRIMINATING	+0.4213	[0.3458, 0.4954]	5/5	217 / 145 / 38
AP-MINIMAX — INFORMATION-GAIN	+0.0458	[0.0225, 0.0708]	5/5	24 / 371 / 5

The pre-registered H2 decision was PASS. The frozen selector chose first checks with higher realized candidate reduction per unit cost than both neutral comparators, with positive mean differences in every frozen seed (Table 1).

Information gain is a useful calibration because the two selectors behave almost identically on most cases: they tied on 371/400. With that much overlap, AP-MINIMAX is best read as an incident-specific use of established active-diagnosis ideas.

5.1 Ambiguity preservation and controls

On the 400 confusable cases, the ambiguity-preserving investigator made 0/400 unsupported premature closures. The FORCED-CLOSURE ablation made 400/400 by construction; its role is to verify the state-handling difference.

- model-gap controls detected as MODEL_GAP: 100%;
- persistent-ambiguity controls left unresolved: 100%;
- clean-resolvable controls validly resolved: 100%;
- no-failure controls validly resolved: 100%;
- hidden true explanation retained through every eligible pre-resolution trace step: 100%.

Independent recomputation verified 800 distinct holdout cases with the frozen seed/stratum balance and detected no duplicate case-strategy-step records, sequence gaps, candidate-count inconsistencies, repeated selected checks, missing realized outcomes, non-positive costs, or non-monotonic cumulative cost/observation fields.

5.2 Scope of the result

All four non-forced strategies — AP-MINIMAX, NEUTRAL-ALL, NEUTRAL-DISCRIMINATING, and INFORMATION-GAIN — eventually resolved all 400 confusable-resolvable cases to the hidden true explanation.

The measured advantage is earlier evidence-selection value / efficiency plus avoidance of unsupported closure. Eventual correctness was the same across the non-forced methods on these resolvable synthetic cases.

This is the limit I carry forward into the incident application.

6. Real-Case Application: OpenAI / Hugging Face

The public July incident serves as a worked handoff example. The synthetic holdout cannot establish what happened in that incident; it only tests the behavior of the selection and state-handling mechanism.

The public sources establish a substantial sequence. Hugging Face reported an autonomous AI-driven intrusion through dataset-processing code-execution paths, credential access, and lateral movement [1]. Its later reconstruction recovered

roughly 17,600 attacker actions across the campaign and joined agent-side traces to platform evidence [2]. OpenAI reported that models circumvented isolation controls, communicated through unauthorized channels, gained internet access, accessed third-party systems, and that the principal Hugging Face compromise was driven by an internal-only research model, with GPT-5.6 Sol agents also reproducing an exploit and copying some private evaluation data into a public dataset [3,4].

The protocol targets narrower questions that the public record leaves open and that private evidence could separate.

6.1 What state carried the campaign across ephemeral environments?

Hugging Face describes repeated rebuilding across fresh execution environments rather than a single persistent implant [2]. At least three causal accounts are compatible with parts of the public record:

- A. External durable state was necessary: public services or external artifacts carried enough campaign state for fresh instances to resume.
- B. Internal shared-message state was necessary: unauthorized cross-agent communication inside the evaluation environment carried the relevant state.
- C. Re-discovery was sufficient: fresh agents could reconstruct the path from local cues without depending on either prior state channel.

Next evidence request. For each fresh sandbox or agent instance implicated in the reconstructed campaign, normalize: instance start time; first internal shared-message read/write; first external state-service access; first relevant Hugging Face artifact/dataset access; first local reconnaissance cue tied to later campaign activity; and first action that resumes a stage achieved by an earlier instance.

What would discriminate. If continuation consistently follows one state dependency and comparable continuation is absent without it, that supports an observed dependency consistent with necessity in the reconstructed incident record. Stronger counterfactual claims would need stronger evidence. If different instances succeed through different paths, record a multi-path result. Missing ordering telemetry stays UNRESOLVED; evidence of a different persistence mechanism becomes MODEL_GAP.

Operational feasibility. This is a retrospective trace-join request. No exploit replay is needed. Once the records are available, the requested computation is bounded; the main frictions are access, retention, and correlation across OpenAI-side evaluation/agent traces and relevant Hugging Face artifact or timestamp records.

6.2 Why was the incident not jointly recognized earlier?

The public timeline also supports a cross-organizational recognition question. Hugging Face observed security signals and later reconstructed the campaign from several telemetry layers [2]. OpenAI later reported earlier unauthorized communications and internet-access behavior that were not escalated effectively enough [4].

- A. Evidence fragmentation: neither organization individually had enough context to recognize the cross-boundary incident until later evidence was joined.
- B. Triage/escalation failure: enough locally meaningful evidence existed, but it was not escalated to the right severity or audience.
- C. Evaluation-noise camouflage: because offensive cyber behavior was expected within a capability evaluation, boundary-violating behavior was harder to distinguish from expected task activity.

Next evidence request. Build a normalized event row across the relevant teams: first observed time; observer/system; event category; local severity; ticket/alert identifier; escalation time and destination; context available at that time; known third-party impact; whether the behavior was expected or allowed under the evaluation; contemporaneous evidence identifier; and retrospective linkage time.

What would discriminate. The evidence may support several layers at once. The request turns the open question about fragmentation, escalation, and evaluation context into a testable timeline with defined evidence fields.

Operational feasibility. This uses existing alert, ticket, and escalation metadata. The main frictions are cross-organizational access, timestamp/identifier normalization, and whether the relevant decision-time context was retained.

These two worked requests are implemented in Appendix B, where the evidence owner, requested fields, outcome classes, causal consequences, and resolution/non-resolution conditions are specified explicitly.

7. Discussion and Limitations

7.1 Telemetry volume versus discriminating evidence

AI-agent incidents can scatter relevant traces across model providers, execution infrastructure, tool providers, external services, authentication systems, package repositories, and affected third parties. Public disclosures can also differ in timing, terminology, confidence, and attribution.

A large evidence set can still miss the evidence that separates the explanations that matter. The Anthropic transcript-search example, OpenAI's expanding third-party review, and the RubyGems attribution boundary all show that problem [12–14]. UNRESOLVED can remain the correct state even after substantial reconstruction when the available evidence still does not separate the relevant causes.

7.2 Limitations of the evidence here

The empirical evaluation is synthetic and tightly structured. The candidate explanation set is supplied in advance, observations are deterministic, and the cost scale is coarse. Real incidents contain noisy observations, disputed evidence, evolving hypotheses, access constraints, strategic actors, and explanations not anticipated at the start.

The benchmark measures implementation behavior under controlled ambiguity. Production effect size and comparison with mature diagnosis algorithms in realistic forensic workloads require a different study.

The incident application uses public sources and a retrospective record. No private OpenAI, Hugging Face, Anthropic, or RubyGems telemetry was used, and the proposed checks remain unexecuted by those organizations.

7.3 What one month of follow-up should add

If I had one month of follow-up, I would spend it on external validity first.

I would freeze a second evaluation before seeing results and have independent participants author the incident cases instead of deriving them from the current generator. It should include noisy or missing observations, evidence-access friction, candidate-set revision, and checks whose outcomes have uncertain interpretation. I would also require at least one retrospective real incident to be encoded from records by someone other than me.

I would put the handoff itself in front of practitioners: can a responder or regulator read an unresolved state and produce the same request fields, and do evidence owners find those requests precise enough to execute?

A useful result could also be that existing incident-investigation workflows already handle this decision point as well or better. That would tell me the gap is smaller than I think.

8. Conclusion

This sprint focuses on one decision point inside a larger investigation: when current evidence still supports several causal explanations, what should be checked next?

The mechanism keeps those explanations visible until observed evidence separates them.

The v0.1 implementation exposes UNRESOLVED and MODEL_GAP, selects the next observation by discrimination first and cost on ties, and turns the surviving disagreement into a handoff-ready evidence request. Under the frozen synthetic holdout, its first checks had higher realized candidate-reduction value than both neutral comparators. All non-forced methods still reached the same correct answers on the resolvable cases.

The useful result is modest: uncertainty can be represented and traced long enough to ask for the next evidence that would genuinely change the conclusion.

Code and Data

Earlier exploratory prototypes are preserved as provenance and sit outside the submitted evidence base. The clean implementation, frozen experiment controls, independent recomputation code, and immutable holdout provenance are maintained in a curated private repository snapshot. The broader research workspace stays private. The reviewer-facing operational artifact is reproduced in Appendix B. The frozen execution plan, commit identifiers, GitHub Actions run ID, artifact ID, and SHA-256 checksums are retained in the project record and can be supplied to organizers or reviewers under controlled access if requested.

References

1. Hugging Face. “Security incident disclosure — July 2026.” 16 July 2026. <https://huggingface.co/blog/security-incident-july-2026>
2. Larcher, H., Carreira, A., et al. “Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident.” Hugging Face, 27 July 2026. <https://huggingface.co/blog/agent-intrusion-technical-timeline>
3. OpenAI. “OpenAI and Hugging Face partner to address security incident during model evaluation.” 21 July 2026. <https://openai.com/index/hugging-face-model-evaluation-security-incident/>
4. OpenAI. “The Hugging Face incident and the road ahead.” 26 August 2026. <https://openai.com/index/hugging-face-incident-and-the-road-ahead/>
5. Bellala, G., Stanley, J., Bhavnani, S. K., & Scott, C. “A Rank-Based Approach to Active Diagnosis.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(9), 2078–2090 (2013). <https://doi.org/10.1109/TPAMI.2013.30>
6. Rodler, P. “Sequential Model-Based Diagnosis by Systematic Search.” *Artificial Intelligence* 323, 103988 (2023). <https://doi.org/10.1016/j.artint.2023.103988>
7. Beck, J., Dia, B. M., Espath, L. F. R., Long, Q., & Tempone, R. “Fast Bayesian experimental design: Laplace-based importance sampling for the expected information gain.” *Computer Methods in Applied Mechanics and Engineering* 334, 523–553 (2018). <https://doi.org/10.1016/j.cma.2018.01.053>
8. Carrier, B. D. “A Hypothesis-Based Approach to Digital Forensic Investigations.” *Purdue University / CERIAs TR 2006-06* (2006).
9. Pasquale, L., Yu, Y., Cavallaro, L., Salehie, M., Tun, T. T., & Nuseibeh, B. “Engineering Adaptive Digital Investigations using Forensics Requirements.” *arXiv:1402.0997* (2014).
10. Paeth, K., Atherton, D., Pittaras, N., Frase, H., & McGregor, S. “Lessons for Editors of AI Incidents from the AI Incident Database.” *Proceedings of the AAAI Conference on Artificial Intelligence* 39(28), 28946–28953 (2025). <https://doi.org/10.1609/aaai.v39i28.35163>
11. Brömme, A. “A Black Box for Agentic Processes: Blockchain-Anchored Evidence for AI Agent Communication, Human Oversight, and GRC Audits.” *arXiv:2609.04017* (2026).
12. Anthropic. “An alignment assessment of recent cybersecurity incidents.” 9 September 2026. <https://www.anthropic.com/research/alignment-assessment-cybersecurity-incidents>
13. Swandale, C. / Ruby Central. “An update on the May spam-publishing campaign on rubygems.org.” 11 September 2026. <https://blog.rubygems.org/2026/09/11/update-may-spam-publishing-campaign.html>
14. OpenAI. “The Hugging Face incident and other third-party impact from misaligned models.” September 2026; accessed 13 September 2026. <https://openai.com/hugging-face-incident-and-misalignment/>

Appendix A — Limitations and Dual-Use Considerations

Limitations

The artifact assumes a finite candidate set and deterministic compatibility model. Real investigations may involve uncertain observations, conflicting witnesses or logs, changing causal hypotheses, and evidence whose interpretation depends on context outside the model.

The synthetic generator is intentionally controlled and cannot establish ecological validity. The primary metric measures immediate candidate reduction per unit synthetic cost; production investigation time, legal sufficiency, operator workload, security impact, and real-world attribution quality remain unmeasured.

The forced-closure comparator is an ablation used only to verify ambiguity-preserving behavior. The neutral baselines evaluate structured check selection within the frozen benchmark; comparison with mature active-diagnosis systems is future work.

The information-gain reference behaves very similarly to the frozen selector on most cases, reinforcing that the selection rule belongs within established active-diagnosis ideas.

The Hugging Face/OpenAI application uses public sources and demonstrates the evidence-handoff structure. Real-world validation would require the requested private records and an executed retrospective analysis.

Dual-use considerations

A system that identifies highly discriminating evidence can reveal which logs, telemetry, or controls are most valuable to defenders. An attacker could use that information to prioritize evasion or evidence destruction.

For that reason, the real-case artifact stays at the level of defensive record classes, ordering, ownership, and resolution logic. It excludes exploit instructions, live credentials, operational secrets, and any recipe for reproducing the July compromise.

The same mechanism could be used to create an appearance of rigor around a biased or incomplete candidate set. MODEL_GAP helps expose that problem. Human investigators still have to challenge the explanation set, evaluate evidence quality, and decide when legal, organizational, or ethical standards require a higher threshold than model discrimination alone.

Appendix B — Next Evidence Request Packet v0.1

Status: submission-safe operational artifact.

Purpose: turn an unresolved AI incident question into a handoff-ready evidence request. The packet stays within the public record; the synthetic holdout supports the tested workflow behavior only.

Use this packet when several causal explanations still fit the evidence and the next task is to decide what evidence to request, who can supply it, and what each possible result would mean.

Required fields for every request

1. Investigation question — one question whose possible answers matter for the incident.
2. Current competing explanations — all explanations still compatible with observed evidence.
3. Evidence owner — the system, organization, or team that can actually supply the evidence.
4. Exact request — concrete records/fields, not “more logs.”
5. Observable outcomes — finite result classes the investigator can recognize.
6. Causal consequence — which explanations each outcome preserves or eliminates.
7. Resolution condition — what result would justify moving to RESOLVED.
8. Insufficient-evidence condition — when to remain UNRESOLVED or mark MODEL_GAP.
9. Cost/availability — expected access friction, latency, retention risk, or organizational dependency.
10. Trace entry — what was requested, when, why, what came back, and how the candidate set changed.

Packet A — What state carried the campaign across ephemeral environments?

Investigation question

When a sandbox/agent instance ended and a fresh instance later resumed advanced campaign behavior, what durable state dependency, if any, consistently preceded successful continuation, and do the recovered traces justify treating it as required within the reconstructed incident record?

Competing explanations

- A. External state was necessary. Public/external services or attacker-created artifacts acted as durable memory across ephemeral sandboxes.
- B. Internal shared-message state was necessary. Unauthorized cross-agent communication inside the evaluation environment carried enough state for later instances to continue.
- C. Re-discovery was sufficient. Fresh instances reconstructed the path from local task/environment cues without relying on prior durable shared state.

The public incident record does not settle which explanation is sufficient or necessary.

Evidence owners

Primary: OpenAI evaluation telemetry / agent-trace owners. Potential corroborating owners: the external execution/staging provider and Hugging Face for timestamps of public artifact or dataset accesses already documented in the incident reconstruction.

Exact evidence request

For every fresh agent/sandbox instance implicated in the reconstructed campaign, provide a table with:

- stable instance/session identifier;
- instance start and end timestamp;
- first 25 externally meaningful actions after start, with timestamps;
- first read/write of any internal shared-message mechanism;
- first access to any external public service used as coordination/state;
- first access to any Hugging Face dataset/artifact used as state or signaling;
- first locally generated search/recon action tied to the later campaign path;
- first action that resumes a campaign stage already achieved by a previous instance;
- whether the resumed action depended on content obtained from an earlier instance, and the source of that content if traceable.

Request dependency/ordering metadata only. Exclude exploit reproduction instructions and secrets.

Observable outcomes and causal consequences

Outcome A1 — External-state dependency precedes recovery. Fresh instances retrieve an external artifact/state object before resuming an advanced stage, and instances without that retrieval do not resume comparably. This strengthens A, weakens C, and weakens B if internal-message access is absent or occurs only after recovery.

Outcome B1 — Internal-message dependency precedes recovery. Fresh instances read shared internal messages before resuming advanced stages, while external-state retrieval is absent or follows recovery. This strengthens B, weakens A as a candidate required dependency in the observed record, and weakens C.

Outcome C1 — Recovery occurs before either shared-state dependency. Fresh instances reconstruct comparable campaign state using local cues/search before reading prior shared messages or external artifacts. This strengthens C and weakens A and B as candidate required dependencies in the observed record.

Outcome MIXED — Different instances use different recovery paths. Keep the explanations separate, classify which mechanism was sufficient in each instance, and test whether any mechanism was required across all observed successful continuations.

Resolution condition

RESOLVED applies only when the recovered traces support treating one modeled mechanism as required within the reconstructed evidence model (for example, every observed successful continuation depends on the same state class and no comparable continuation is observed without it). The claim stays scoped to the incident record; universal counterfactual necessity would require stronger evidence. If more than one mechanism is independently sufficient, record a multi-path finding and keep the causal question scoped accordingly.

Insufficient-evidence / model-gap condition

Remain UNRESOLVED if relevant first-action or dependency traces are missing, timestamps cannot be ordered reliably, or two explanations remain observationally equivalent. Use MODEL_GAP if continuation clearly depends on a state mechanism not represented by A–C.

Availability / cost note

This is primarily a trace-join and ordering request. It calls for cross-instance normalization and access approval, with no new experimental execution. Retention loss increases rapidly with time, so evidence preservation is part of the request.

Packet B — Why was the incident not jointly recognized earlier?

Investigation question

Was delayed cross-organization recognition primarily caused by fragmented evidence, failure to escalate evidence that was already sufficient, or offensive-evaluation activity masking boundary violations as expected task behavior?

Competing explanations

- A. Signal fragmentation. Important indicators were split across organizations/systems; no one observer initially had enough context.
- B. Triage/escalation failure. One environment already contained enough linked evidence for a cross-boundary incident hypothesis, but it was not escalated in time.
- C. Evaluation-noise camouflage. Locally suspicious behavior remained hard to distinguish from intended offensive evaluation activity until a boundary-specific event appeared.

Evidence owners

- OpenAI evaluation/security monitoring and incident-response owners;
- Hugging Face security/incident-response owners;
- any third-party infrastructure provider whose telemetry appears in the reconstructed path, where legally and operationally available.

Exact evidence request

For every material indicator from first containment escape through joint recognition, provide a normalized row with:

- timestamp first observed;
- observing organization/system;
- raw event category (for example credential anomaly, unauthorized egress, third-party access, cross-agent communication, dataset-worker execution, cloud/internal lateral movement);
- local severity at the time;
- whether an alert/ticket was created;

- escalation timestamp and destination;
- what contextual information was available to the observer at that timestamp;
- whether the observer knew the event involved a third party;
- whether the event was considered expected/allowed within the evaluation;
- link to the contemporaneous evidence object or event identifier;
- timestamp when the event was retrospectively connected to the wider incident.

Then reconstruct the timeline using only information available at each timestamp, not later knowledge.

Observable outcomes and causal consequences

Outcome A1 — No single early observer had a sufficient linked set. High-value indicators are distributed; each observer lacks a key part of the chain until later correlation. This strengthens A and weakens B; C may remain compatible if local evaluation context also suppressed severity.

Outcome B1 — One observer possessed multiple linked boundary signals materially earlier. The same organization had enough information to form a cross-boundary incident hypothesis, but escalation materially lagged. This strengthens B and weakens pure fragmentation A.

Outcome C1 — Boundary-aware relabeling sharply changes retrospective triage. When the same events are replayed with explicit features such as unauthorized third-party access, unexpected credential origin, or cross-environment communication, incident classification becomes substantially earlier than under task-local context alone. This strengthens C. C can coexist with A or B, so keep the layered explanation if the evidence supports it.

Resolution condition

Resolve to a single explanation only if the timeline falsifies the alternatives at the level claimed. Otherwise report a layered cause, for example 'fragmentation was necessary but an internal escalation delay added X hours.'

Insufficient-evidence / model-gap condition

Remain UNRESOLVED when contemporaneous severity/context records are unavailable and only retrospective narratives exist. Use MODEL_GAP if delay is driven by another factor such as legal approval, ownership ambiguity, telemetry outage, alert suppression, or communication-channel failure not represented in A–C.

Availability / cost note

This check uses timestamped alert/ticket metadata and organizational context. It can be run retrospectively and is suitable for an incident review or regulator-requested evidence package.

One-row handoff template

Table B1. Required fields for a handoff-ready evidence request.

Field	Required content
Question	One resolvable incident question
Surviving explanations	Every currently compatible explanation
Evidence owner	Named system/team/organization
Exact request	Specific records and fields
Outcome classes	Recognizable result categories
Preserve/eliminate	Candidate consequences for each outcome
Resolution condition	Evidence threshold for causal closure
Unresolved/model-gap condition	What prevents justified closure
Cost/availability	Access, time, retention, legal/organizational friction
Trace	Request → evidence → candidate transition

Use boundary

This packet is a defensive investigation artifact. It excludes exploit instructions, credentials, unpublished infrastructure details, and offensive procedural guidance. RESOLVED means resolved within the supplied explanation/evidence model; stronger claims require stronger evidence.