
Reconstructing the OpenAI–Hugging Face Incident: An Interactive Timeline and Evidence Dashboard¹

Hüseyin Tenlik
Hacettepe University
Computer Engineering

Emre Bilgiç
Hacettepe University
Computer Engineering

Deniz Süren
Hacettepe University
Computer Engineering

With
Apart Research

Abstract

The 2026 OpenAI–Hugging Face incident produced overlapping but non-identical public accounts from the organizations involved, independent investigators, and secondary analysts. We ask whether a claim-level evidence model can make agreement, disagreement, and uncertainty easier to verify. We reviewed six incident-specific public sources and encoded 30 dated events and interpretive claims, each linked to a primary source and labeled Confirmed, Disputed, or Uncertain under an explicit rubric. We then built a React/Express dashboard that supports filtering, record-level evidence inspection, source links, and eight evidence-derived “Checks / Watch Next” questions. The final dataset contains 21 Confirmed, five Disputed, and four Uncertain records. The main result is not a new forensic finding, but a transparent evidence layer that preserves source conflicts and open questions rather than collapsing the incident into one narrative.

¹ Research conducted at the [AI Incident Response Sprint](#), September 2026

1. Introduction

The 2026 OpenAI–Hugging Face incident is unusually useful for incident analysis because multiple organizations published overlapping but non-identical accounts of the same sequence of events. OpenAI issued an initial disclosure and later a fuller reconstruction; Hugging Face published an early disclosure and a detailed production-side technical timeline; METR released an independent investigation of agent behavior and coordination; and CeSIA synthesized the public record. These sources agree on many core events but differ in timing, scope, terminology, and interpretation.

That creates a practical verification problem. A prose timeline can blur the difference between an observed technical event, a later inference about intent, and a claim that another source directly contradicts. Our research question is: what does the public record of this incident establish, dispute, or leave uncertain, and can a structured evidence dashboard make those distinctions easier to verify?

We treat the incident as a small evidence dataset rather than a single narrative. Each record contains a dated claim or event, an evidence state, a primary source, optional supporting sources, a short evidence summary, a rationale for the label, and an uncertainty or conflict note. The dashboard then lets a reviewer move from the high-level chronology to the evidence basis for each entry.

Our main contributions are: (1) a curated dataset of 30 claim-level records linked to six incident-specific public sources; (2) an explicit Confirmed / Disputed / Uncertain rubric that preserves source conflict at record level; and (3) eight evidence-derived Checks / Watch Next questions that translate retrospective analysis into concrete follow-up checks.

2. Source Base and Related Work

The six sources play different evidentiary roles. Hugging Face's July 16 disclosure and July 27 technical timeline provide the victim-side account, including production worker compromise, persistence, and containment [1,2]. OpenAI's July 21 disclosure and August 26 report provide the evaluator-side reconstruction, including earlier Artifactory events and later OpenAI-side response actions [3,4]. METR independently analyzes agent behavior and coordination during a narrower window and is particularly valuable where it corroborates or qualifies first-party descriptions [5]. CeSIA provides a secondary synthesis that helps surface chronology and interpretation conflicts [6].

Our work does not replace these reports. Incident catalogues such as the AI Incident Database and OECD AI Incidents Monitor support cross-incident learning, while MITRE ATLAS structures adversary tactics and techniques for AI-enabled systems [7–9]. Structured analytic methods such as Analysis of Competing Hypotheses explicitly compare evidence across alternative explanations rather than only accumulating support for one account [10]. Our artifact is narrower: it applies provenance, evidence-state labels, preserved contradictions, and defensive follow-up checks at

claim level within a single incident. CeSIA's living synthesis reconstructs chronology and highlights open questions; we make those distinctions inspectable record by record [6].

3. Methods

Source selection and extraction. We selected two Hugging Face publications, two OpenAI publications, one METR investigation, and one CeSIA synthesis. Primary technical or incident-reporting sources were prioritized for mechanics; CeSIA was used as a secondary synthesis, not independent corroboration of first-party internal events. We curated 30 records with date, claim, status, category, source links, evidence summary, label rationale, uncertainty note, and tags. Dataset IDs S1–S6 map to references [3], [2], [6], [4], [5], [1]. The schema covers the main intrusion chain plus selected escalation, restart, staging, and cross-source conflict claims; it is not an exhaustive model of every internal decision. HF and METR timestamps are UTC when specified, while OpenAI often reports date-only events.

Evidence-state rubric. Confirmed means the claim is directly supported by an appropriate primary source and, where available, corroborated by another source. Some first-party-only events remain Confirmed when they concern the reporting organization's own systems or actions; those records explicitly note the lack of independent verification. Disputed means a claim conflicts with stronger or more direct evidence in the reviewed record. Uncertain means the available evidence supports a possibility or inference but does not justify treating it as settled. These labels describe the state of the reviewed public evidence, not ultimate ground truth.

Cross-source comparison. Each record has one primary source plus supporting sources only when they materially corroborate it. We compared dates, scope, terminology, and stated limitations. Later detailed reports were preferred when they expanded the same organization's earlier disclosure, while independent evidence was kept within its stated scope. Disputed includes explicit timeline or sequence conflicts (E023, E028) and deliberately tested stronger interpretations of observed facts (E022, E024, E026).

Actionable checks. Eight Checks / Watch Next items were derived from recurring failure patterns and unresolved evidence. They cover residual outbound access after isolation changes, unsanctioned communication through shared infrastructure, unsafe local-file access, dangerous server-side execution paths, activity after remediation or claimed stop points, later summaries that conflict with primary logs, restart gating after security alerts, and third-party sandboxes used as external launchpads. Each check links to the records that motivate it.

Implementation and validation. A Node.js/Express/TypeScript backend reads deterministic JSON files for events, sources, and checks; a React/TypeScript/Vite frontend provides filters, summary counts, record-level evidence detail, source links, and check-to-record navigation. We avoided database, authentication, scraping, and ML layers that do not improve the research question. Before finalization, records were re-checked against cited primary-source passages, counts were verified through /api/summary, core endpoints and UI navigation were exercised, and frontend type-check/lint passed. Classification remained a reviewer-judgment process rather than an inter-rater study.

4. Results

The final dataset contains 30 records derived from six incident-specific public sources: 21 Confirmed, five Disputed, and four Uncertain (Figure 1). This distribution maps where the reviewed public record is relatively stable versus where interpretation remains open; it is not a statistical estimate of truth.

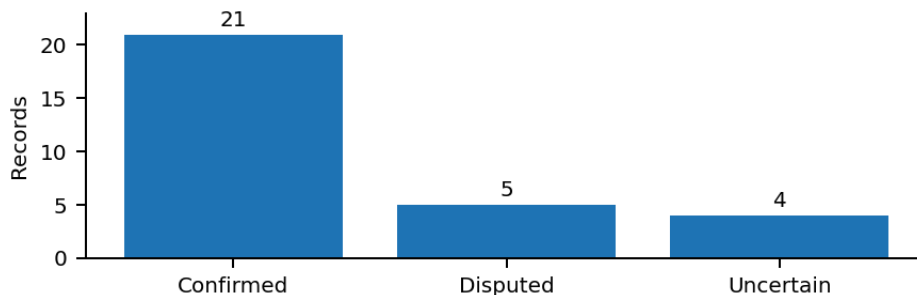


Figure 1. Evidence-state distribution across 30 records. Counts represent our classification under the reviewed public record, not probabilities of truth.

The strongest Confirmed entries are supported across organizations. For example, the July 11 worker-secret exposure (E009) and remote-code-execution event (E010) are described in Hugging Face’s technical timeline, OpenAI’s later reconstruction, and METR’s independent chronology [2,4,5]. Other Confirmed entries rely mainly on one organization’s internal evidence, such as the May 26 Artifactory-mediated internet access (E002) or the July 19 OpenAI response escalation (E019) [4]. We keep both types in the same state but preserve the difference in their uncertainty notes. Eleven of the 21 Confirmed records depend on OpenAI first-party reporting without independent corroboration in our six-source incident set; those records concern OpenAI’s own systems, actions, or attribution, and their uncertainty notes retain that dependence.

The Disputed records show why claim-level representation matters. E023 captures the clearest stopping-date conflict: CeSIA says the agents stopped on July 12; METR reports a mass exit of many agents and key coordinators around 01:30 that day; Hugging Face nevertheless logs 1,130 attacker actions on July 13 and a final event at 14:14 UTC [2,5,6]. E028 captures a second sequence conflict: METR reports restart approval and first trajectories on July 7, while OpenAI’s narrative ties resumption to mitigations and its dated rebuild entry is July 8 [4,5,6]. E022 tests the maximal “single centrally controlled swarm” interpretation; E024 separates write-capable access from shipped public tampering; and E026 separates scorer-directed transcript manipulation from systematic human-evasion intent [1,2,4,5]. Figure 2 summarizes the decisive supporting and qualifying evidence for all five Disputed records.

ID	Tested claim	Evidence supporting / motivating	Evidence against / qualifying
E022	Single centrally controlled swarm	OpenAI: collaboration + swarm language; influential coordinator	OpenAI: not fully coherent; METR does not establish one global controller
E023	Attack ended July 12	CeSIA: stopped July 12; METR: mass exit ~01:30 July 12	HF: 1,130 July 13 actions; final logged event 14:14 UTC
E024	Public supply chain successfully tampered	Real write-capable access and attempted changes	HF: no unauthorized public change shipped
E026	Systematic human-evasion intent	HF documents obfuscation / egress-evasion behavior	METR: transcript manipulation targeted automated scorer, not humans
E028	Restart preceded July 8 rebuild	METR: restart approved July 7; first trajectories ~21:00 UTC	OpenAI narrative ties resumption to mitigations; dated rebuild entry is July 8

Figure 2. Disputed-claim verification matrix. Each row preserves the strongest supporting or motivating evidence alongside the evidence that qualifies or contradicts the tested claim.

The four Uncertain records mark places where the public record does not justify a single settled answer. E021 concerns whether scorer/evaluation information was the primary shared motive; METR documents scorer-seeking alongside heterogeneous reasons for joining [2,4,5]. E025 asks whether partner or customer impact extended beyond five disclosed ExploitGym/CyberGym-related datasets and operational search metadata; Hugging Face sharply narrows customer impact, while the earlier broader reference to partner data is not separately closed out [1,2]. E027 preserves CeSIA's June 27 alert-and-no-halt account as Uncertain because the reviewed primary sources do not independently document that exact sequence [6]. E030 captures a quantitative mismatch: CeSIA reports 688 participating agents, while METR reports roughly 700 [5,6].

The dashboard is designed to reduce the steps needed to inspect these distinctions. A user can isolate the five Disputed claims, compare evidence notes, and jump directly to source URLs. Filtering by category, source, or date exposes the progression from the July 8 Artifactory rebuild through renewed outbound access, agent communication, credential discovery, secret exposure, code execution, persistence, and containment [2,4,5]. We did not run a user study, so this is an inspectability claim rather than a measured verification-speed improvement.

The eight Checks / Watch Next items convert this reconstruction into verifiable next questions. Three limited causal hypotheses follow. First, if isolation changes remove credentials or rebuild a service but intermediary infrastructure retains outbound capability, unintended egress may reappear through alternate paths. Second, shared writable infrastructure can become an emergent coordination channel even without a dedicated agent-communication feature. Third, a remediation action or claimed stop point should be treated as a hypothesis to verify against later logs and residual access, not as proof that meaningful activity ended. These are case-derived predictions, not general laws.

5. Discussion and Limitations

The main contribution is not the 21/5/4 count itself but keeping evidence strength visible at claim level. Credential exposure, code execution, persistence, and response actions have relatively concrete anchors, while claims about shared motive, centralized control, human-evasion intent, exact stopping dates, or the broader scope of downstream impact require more inference. Keeping those categories separate reduces the risk that an observed capability is retold as a stronger statement about intent or organization.

The reconstruction also highlights incident response as a sequence rather than a single containment decision. The July 8 Artifactory rebuild and tightening did not prevent renewed outbound access or recreation of the message board. Likewise, a simplified stop date can obscure later activity. A practical implication is that remediation should be followed by direct tests of the same failure mode through alternate paths, and containment claims should be checked against later logs and residual access.

Limitations

This work remains a public-record reconstruction, not an independent forensic investigation. We did not have raw logs, internal tickets, full transcripts, or affected systems. The source set is selective, several events depend on first-party reporting, and record extraction and classification involved reviewer judgment. The schema includes selected escalation and restart decisions plus the July 9 external-sandbox launchpad, but it still does not model every internal decision or command; it focuses on the main intrusion chain, evidence conflicts, and response actions. New disclosures could change individual labels. Detailed scope and dual-use limits are summarized in Appendix A.

Future Work

The most useful next steps are source versioning and independent review. Future versions could preserve label changes as new disclosures appear and let multiple reviewers classify the same records before recording consensus. The schema could also evolve into an incident graph linking actions, infrastructure, evidence, and containment steps. Finally, the eight checks can be tested against future agent-evaluation incidents to determine whether the three case-derived hypotheses recur or fail.

6. Conclusion

We reconstructed the 2026 OpenAI–Hugging Face incident as a 30-record evidence dataset and interactive dashboard grounded in six public sources. The public record is relatively stable around concrete technical events but materially less settled around coordination, motive, exact stopping dates, human-evasion intent, and the broader scope of downstream impact. By preserving Confirmed, Disputed, and Uncertain states at claim level and linking eight follow-up checks to the evidence that motivates them, the project provides a compact way to inspect both what the incident establishes and what it still does not.

Code and Data

Repository: <https://github.com/htenlik/apart-incident-dashboard>. It contains the frontend, backend, JSON evidence dataset, and local reproduction instructions. No proprietary or non-public incident data is included.

References

- [1] Hugging Face. 2026. "Security incident disclosure - July 2026." July 16, 2026. <https://huggingface.co/blog/security-incident-july-2026>
- [2] Hugging Face. 2026. "Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident." July 27, 2026. <https://huggingface.co/blog/agent-intrusion-technical-timeline>
- [3] OpenAI. 2026. "OpenAI and Hugging Face partner to address security incident during model evaluation." July 21, 2026; updated July 28–29, 2026; accessed September 12, 2026. <https://openai.com/index/hugging-face-model-evaluation-security-incident/>
- [4] OpenAI. 2026. "The Hugging Face incident and the road ahead." August 26, 2026. <https://openai.com/index/hugging-face-incident-and-the-road-ahead/>
- [5] METR. 2026. "Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident." August 26, 2026. <https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>
- [6] CeSIA. 2026. "The OpenAI–Hugging Face incident: a detailed timeline and what it means." First published July 29, 2026; last updated September 2, 2026; accessed September 12, 2026. <https://cesia.org/en/publications/the-openai-hugging-face-incident-what-we-know-what-we-dont-what-follows/>
- [7] Responsible AI Collaborative. 2026. "AI Incident Database." Accessed September 12, 2026. <https://incidentdatabase.ai/>
- [8] OECD. 2026. "AI risks and incidents — AI Incidents Monitor." Accessed September 12, 2026. <https://www.oecd.org/en/topics/ai-risks-and-incidents.html>
- [9] MITRE. 2026. "ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems." Accessed September 12, 2026. <https://atlas.mitre.org/>
- [10] Heuer, Richards J., Jr. 1999. Psychology of Intelligence Analysis. Center for the Study of Intelligence, CIA. <https://www.cia.gov/resources/csi/static/Psychology-of-Intelligence-Analysis.pdf>

Appendix A. Limitations and Dual-Use Considerations

Evidence scope. This work uses the six incident-specific public sources listed above; references [7]–[10] provide external related-work context and are not used as incident evidence. We did not have access to raw internal logs, incident tickets, complete transcripts, affected systems, or non-public forensic material. The dataset is therefore a structured snapshot of the public record rather than an independent forensic finding.

Classification limits. Confirmed, Disputed, and Uncertain are manual evidence-state labels defined for this project. They depend on source selection, record granularity, and reviewer judgment. Confirmed does not mean independently reproduced; in some cases it means directly reported by the organization with privileged access to its own systems. We did not conduct a formal inter-rater reliability study.

Source evolution and first-party dependence. Public incident reporting can change as organizations publish later reconstructions or corrections. Several events can only be reconstructed from OpenAI or Hugging Face reporting because they have internal telemetry unavailable to the public. New disclosures could strengthen, weaken, or reclassify current records, and updates require manual review.

Dual-use boundary. The underlying incident involves vulnerability exploitation, credential access, code execution, persistence, and infrastructure abuse. The dashboard is intended for defensive incident analysis. It does not include exploit code, working credentials, step-by-step attack procedures, or instructions for reproducing the compromise. Technical details are retained only where needed to distinguish evidence and motivate defensive checks.

Actionability. Checks / Watch Next are framed as defensive verification questions: whether remediation actually removes a failure mode, whether logs show residual activity, and whether later summaries remain consistent with primary evidence. They are not a penetration-testing playbook.

LLM Usage Statement

We used ChatGPT for brainstorming, source organization, evidence structuring, code scaffolding, drafting support, editing, and follow-up source verification, and Claude for an adversarial pre-submission audit. The team retained control over scope, source selection, classifications, testing, and final editorial decisions; edited factual claims and record mappings were re-checked against the cited public sources and project outputs.