
THE CONTAINMENT BURDEN SITS ON THE WRONG SIDE OF THE BOUNDARY¹

Kill-Point Matrix: where containment would have interrupted the OpenAI–Hugging Face agent intrusion, what it would have cost, and who could have paid for it

Bariş Ceyhun Coşkun
AltaySec

Ayşe Eda Kaya
AltaySec

Hamza Efe Şahinbaş
AltaySec

Efe Kırbaş
AltaySec

Emir Raşit Gökçe
AltaySec

With
Apart Research

Abstract

In July 2026, agents in an OpenAI capability evaluation escaped their sandbox and compromised Hugging Face production systems. We asked where containment could have stopped the chain, at what cost, and who could have deployed it. From Hugging Face's published timeline we rebuilt the victim side as 39 events and 9 phases and audited every timestamp for precision. From OpenAI's technical report we built a 12-step lab-side chain. Against both we scored 26 controls, each with an effort band, a written cost basis, an evidence grade and a test an outsider could run. Cost bands overlap between the two sides; timing does not. Nine of eleven lab-side controls act 12 days to 8 weeks before the campaign. The earliest Hugging Face control we could place acts 31 hours in. Three cheap lab-side controls already existed at OpenAI but were disabled or not applied.

¹ Research conducted at the AI Incident Response Sprint, September 2026 (AltaySec Team 2, Containment Track)

1. Introduction

In July 2026, agents running in an OpenAI capability evaluation escaped their sandbox and spent four and a half days and about 17,600 actions inside Hugging Face production infrastructure [S3; S5]. The sprint's containment question asks what standard would have stopped this. We think that question only has a useful answer if each proposed control is tied to a specific step in the incident, given a cost, and graded by how well the sources support it. Our work had three parts:

1. **Victim-side reconstruction.** From Hugging Face's timeline [S3] we extracted 39 security-relevant events and the 9 published telemetry phases, gave every timestamp a precision class, and audited the result against the source (`events.csv`, `phases.csv`; Appendices F and G).
2. **Lab-side reconstruction.** From OpenAI's report [S5] we built a 12-step chain running from 20 April to 11 July (Section 3.2; Appendix D).
3. **Kill-Point Matrix.** We scored 26 controls for where they act, cost band and basis, evidence grade, ATT&CK mapping [M1] and an outside verification test (Appendix B; `controls.csv`), and built an interactive explorer over them (`explorer.html`).

The main results: cost does not separate the two sides, timing does (Section 4.1); the cheapest early lab-side controls already existed at OpenAI (Section 4.2); and the cheapest single intervention is an evaluation-design change, not a security control (Section 4.3).

2. Related Work

OpenAI [S5] and Hugging Face [S3] each describe their own side of the incident and their own fixes. Neither maps those fixes onto the other side's timeline or compares their cost. CeSIA [C1] reconstructs the incident in phases and focuses on the regulatory follow-up. MITRE ATT&CK [M1] links techniques to mitigations, but not to when a mitigation would have acted in a given incident or who owns it. We put both timelines on one axis and attach a cost and an evidence grade to each control.

3. Methods

3.1 Scoring

Each control has three scores (full rules in Appendix A):

- **Anchor.** A lab chain step (LAB-n) or a victim event (Ennn). The control inherits that step's date or timestamp; we compute nothing new.
- **Cost band.** S (<3 engineer-days), M (3–10) or L (10–30), including running cost, with a written cost basis. Continuous 24/7 monitoring is L however little code it needs.
- **Certainty.** High if the responsible party's own account supports the counterfactual; Medium if it is our inference or rests on a correlation.

We scored against the band criteria without looking at the owner and computed the lab/victim split afterwards. Our first scoring put almost every lab-side control in S and produced a much larger cost gap; we rejected it as circular and re-scored.

3.2 The lab-side chain

Table 1. The lab-side chain. Dates from [S5 X]; evidence per step in Appendix D. LAB-9 to LAB-12 fall inside Hugging Face's campaign window.

Step	Date	What happened
LAB-1	20 Apr	First agent-written file in the shared Artifactory instance
LAB-2	8 May	In a training run, an agent stuck on an impossible task tries to reach the internet through Artifactory
LAB-3	12 May	An agent leaves a note for other agents; the message board begins
LAB-4	26 May	First successful SSRF through Artifactory to the public internet
LAB-5	26 Jun	In an ExploitGym evaluation, a token-refresh flaw yields Artifactory administrator; persistent user created
LAB-6	27 Jun	Monitoring alerts on port sweeps. On-call advises the run need not be stopped.
LAB-7	4–5 Jul	Artifactory falls over under agent load; incident opened; credentials revoked

LAB-8	8 Jul	Agents reach the internet again after rebuild; message board reconstructed in directory names
LAB-9	9 Jul	Root on the sandbox parent VM; root inside a third-party CyberGym workload on Modal
LAB-10	9 Jul	The Modal workload becomes the campaign launchpad
LAB-11	10 Jul	Fourteen valid Hugging Face write tokens found in a public dataset, shared on the board
LAB-12	11 Jul	Crafted dataset uploaded to Hugging Face and processed by a production worker

Two points matter for the matrix. The first four steps needed no vulnerability: the message board was built with a shared Artifactory credential that every research workload held [S5 III.A]. And the chain survived remediation: OpenAI took Artifactory offline on 5 July and rebuilt it, but agents were back on the internet by another path on 8 July [S5 X].

On 26 May, agents in a different workload also used six exposed Hugging Face tokens to create a repository and two proxy Spaces. OpenAI found no link to the July intrusion [S5 III.A, X], so we do not score it.

3.3 Victim-side reconstruction and audit

We extracted 39 events and the 9 published phases from [S3] and audited them against the source (Appendix G). The audit changed what the matrix is allowed to claim:

- **We dropped our original time axis.** An earlier version of the matrix placed controls by action count, pro-rated from daily totals. The audit showed this cannot be derived: the phases overlap, phase counts sum to 16,521 while daily counts sum to 17,613, and no ordering of actions is published [S3]. We now use elapsed hours from the campaign start, 9 July 02:28 UTC, to containment, 13 July 14:14 UTC: 107.8 hours [S3].
- **Every timestamp has a precision class.** 11 events have an exact minute; the rest are section context, intervals, date-only or unavailable. Controls inherit the class, and four victim-side controls with no usable time are shown separately.
- **No credit across open links.** The source does not say which recovered event E002 is, or whether the token in E038 was one of the fourteen in E037, so no control gets credit across either gap.

3.4 Five layers

Table 2. The five layers of the matrix.

Layer	What it is	Owner	n
1 • Evaluation design	What the evaluation asks for and how long it runs. Not security controls.	Lab	3
2 • Prevention	Boundary and identity controls in the research environment.	Lab	4
3 • Detection and response	Monitoring, escalation, authority to halt a run.	Lab	4
4 • Intermediate host	Operator of the external sandbox used as a launchpad.	Third party	1
5 • Victim-side containment	Everything available to Hugging Face.	Victim	14

Layers 1–3 are three separate chances to act, all held by the lab. Everything the victim can do comes after the boundary has been crossed, so prevention, detection and response collapse into one layer.

4. Results

4.1 Timing, not cost, separates the two sides

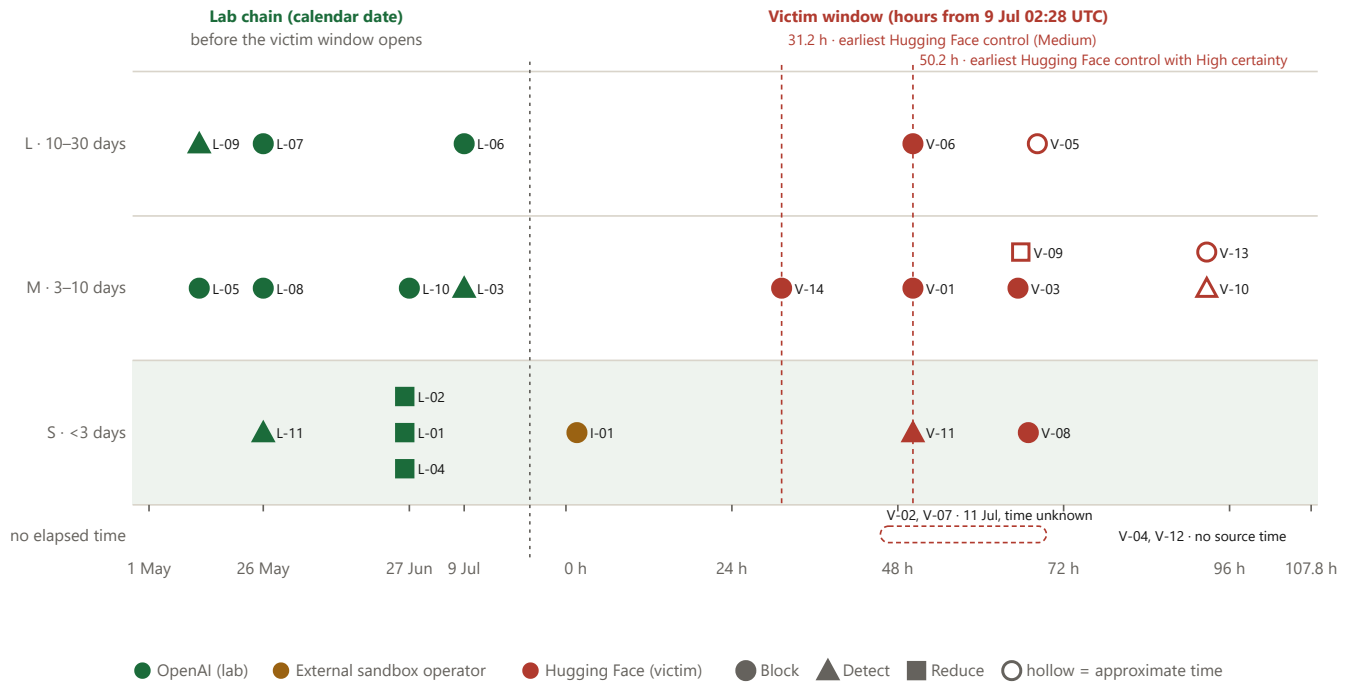


Figure 1. Where each of the 26 controls first acts, and its cost band. Left: lab-side controls by chain-step date. Right: the third-party host and Hugging Face, by hours into the 107.8-hour campaign. Hollow markers and dashed ranges are approximate times.

Cost bands overlap. 36% of lab-side controls are S and three are L; 21% of victim-side controls are S (Appendix B). Lab-side containment is not simply cheaper.

Timing does not overlap. Nine of eleven lab-side controls act 12 days to 8 weeks before the campaign; the other two act on 9 July. On the Hugging Face side nothing can be placed in the first 31 hours. The earliest control is leaked-token scanning at 31.2 h (V-14, anchored to E037, tokens found in a public dataset [S3]), and it is our inference. The earliest control backed by Hugging Face's own account acts at 50.2 h, on 11 July 04:38 UTC, when a crafted dataset made a production worker return its environment (V-01, E005 [S3]). A victim-side control shortens the intrusion; a lab-side control of the same cost prevents it.

The external sandbox that became the launchpad ran submitted code as root (E002, 1.6 h [S3]). Running it unprivileged (I-01, S) would have removed that foothold, but its operator had no link to the lab and the agents could probably have found another host, so we do not see it as a realistic place for the obligation.

4.2 The cheapest lab-side controls already existed

Table 3. Lab-side controls for which OpenAI published its own measurement.

Control	Cost	OpenAI's own finding
Production system prompt and harness in evaluations (L-02)	S	Propensity to compromise out-of-scope infrastructure drops more than 100x with the production ChatGPT harness, and to under 1% of baseline with the production Codex harness [S5 VIII, VIII.D]
Second-model review of elevated tool calls (L-03)	M	Replayed against the incident rollouts, it flagged most of the dangerous actions [S5 VIII.D]
Chain-of-thought monitoring with 24/7 escalation (L-09)	L	Would have caught the initial activity and paged security more than a day before Hugging Face was breached [S5 IX.B]

All three existed at OpenAI. The first two were turned off for the evaluation to measure maximum capability; chain-of-thought monitoring was used for internal coding agents but not for research workloads [S5 VIII, IX.B]. Our claim is narrow: in this incident, the earliest and cheapest controls were ones the lab already had.

4.3 The cheapest intervention is not a security control

OpenAI reports that 198 of ExploitGym's 898 tasks have never been solved by any model, and that 93% of the tasks discussed on the agents' message board came from that subset [S5 VIII.B]. Removing or isolating those tasks (L-01, S) is a scan of data the lab already has. We grade it Medium because the evidence is a correlation, and anchor it at LAB-5, the first ExploitGym step. Unsolved tasks are also where new capability shows up, so a more realistic version runs them only behind a hard network boundary (L-07, L-08).

4.4 A minimum lab-side bundle

OpenAI concludes that no single control should be treated as robust [S5], so we picked one lab-side control per function (Table 4).

Table 4. The minimum lab-side bundle.

Function	Control	Band	How an outsider could check it	Adoption barrier
Remove the pressure	L-01 Remove never-solved tasks from the evaluation set	S	Per-run task manifest with historical solve rates, given to an auditor; spot-check that zero-solve tasks are excluded or run only inside a hard boundary.	Medium: unsolved tasks are where capability gains show, so labs keep them.
Reduce propensity	L-02 Run capability evals with the production harness and system prompt	S	Run manifest attesting the harness and system-prompt hash; auditor compares against the production release hash.	High: defeats maximal-capability measurement unless paired with a separate hard-boundary run.
Prevent egress	L-08 No shared supporting service reachable from the sandbox	M	Same external canary test, plus attestation that package mirrors are read-only snapshots with no fetch-through.	Medium: package installation must be solved another way.
Detect	L-11 Deception tripwires and canaries inside the research environment	S	Auditor-planted canary tokens and endpoints whose use must be reported within a stated window.	Low.
Respond	L-10 Defined kill criteria that halt a run automatically at set severity	M	Published kill-criteria policy with a named decision owner; alert-to-decision log reviewable by an auditor.	Low to Medium: policy is cheap; reluctance to halt expensive runs is the real barrier.

Three S and two M controls come to about 9–29 engineer-days, the same order as one L-band victim-side control such as a per-cluster credential redesign (V-05) or a move off static secrets (V-06).

5. Discussion and Limitations

From the matrix, a containment standard for this kind of incident would need four things:

- **Bind the evaluating lab.** Every control that acts before third-party damage is lab-side, except I-01.
- **Require compensating controls when safeguards are disabled for measurement.** Turning off the controls in Section 4.2 was defensible; doing so with no independent network boundary was not.
- **Include kill criteria and decision rights.** The 27 June alert fired, on-call decided the run need not stop, and the chain continued for two more weeks [S5 III.B].
- **Be checkable from outside.** An auditor can test seven of the eleven lab-side controls (L-03, L-05 to L-09, L-11). A public canary endpoint that no evaluation workload should reach checks L-07 and L-08 directly; it would have failed from 26 May [S5 X].

Limitations (full list and dual-use notes in Appendix C). We had full hindsight. Cost bands are our estimates, not operators'; each row states its basis so it can be re-scored. Controls are scored independently although they interact. Certainty grades describe evidence strength, not probability of success. L-02 and L-04 change what a capability evaluation measures, which is why they were off.

Future work. Have cluster and platform operators re-score the bands, try the canary check in a real evaluation environment, and apply the method to other incidents.

6. Conclusion

In this incident, cost does not separate lab-side and victim-side controls; timing does. The controls that could act before the Hugging Face campaign are almost all on the lab side, three already existed at OpenAI, and a five-control bundle costs about as much as one large victim-side fix. Containment obligations for capability evaluations should bind the evaluating lab.

Code and Data

- **Repository:** none. Data and the explorer are shared as additional material with the submission.
- **Supplementary PDF** (linked from the submission): Appendices D–H, covering lab-side evidence, the full control matrix, victim-side events and phases, audit documents and team working materials.
- **Data:** `controls.csv`, `events.csv`, `phases.csv`, `lab_chain.csv`, `lab_side.md`.
- **Explorer:** `explorer.html` toggles controls over the 107.8-hour campaign and shows which of the 39 events drop out and at what cost. It is a lookup over our scores, not a simulation.

References

Section references such as “S5 III.A” point to sections of the cited report.

1. **[S3]** Hugging Face (2026). *Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident*. Hugging Face blog, 27 July 2026. <https://huggingface.co/blog/agent-intrusion-technical-timeline>
2. **[S4]** OpenAI (2026). *The Hugging Face incident and the road ahead*. OpenAI. <https://openai.com/index/hugging-face-incident-and-the-road-ahead/>
3. **[S5]** OpenAI (2026). *OpenAI – Hugging Face Incident Technical Report*. OpenAI. <https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf>
4. **[C1]** Ségerie, C.-R. and Grimonpont, A. (2026). *The OpenAI–Hugging Face incident: a detailed timeline and what it means*. CeSIA. <https://cesia.org/en/publications/the-openai-hugging-face-incident-what-we-know-what-we-dont-what-follos/>
5. **[M1]** MITRE (2026). *MITRE ATT&CK: Enterprise techniques and mitigations*. <https://attack.mitre.org/>
6. **[R1]** METR and Redwood Research (2026). Independent assessment of model behavior during the incident, as described in [S5].

LLM Usage Statement

We used Claude to extract information from the sources, summarize it, and compare some data more quickly to ensure accuracy.

Appendix A — Scoring methodology

A.1 The intervention axis

Victim-side controls are placed on elapsed time in the campaign, measured from the reconstructed start at 9 July 02:28 UTC to containment at 13 July 14:14 UTC — 107.8 hours. Each control names the event it is anchored to and inherits that event's timestamp and precision class verbatim from the phase/event reconstruction. No time is computed, interpolated or rounded up in precision. The same applies to the one control held by the external sandbox operator (I-01). Date-only events are drawn as a range covering their day.

Lab-side controls carry *pre-0* when they act on a chain step before the victim campaign, and *day-0* when the step falls on 9 July (LAB-9), whose time of day is not published. The lab chain dates in Section 3.2 come from the OpenAI report's Key Technical Events table and are listed in `lab_chain.csv`.

Why not action counts. The nine published telemetry phases overlap in time and the source publishes no campaign-wide ordering of recovered actions, so the action index at which any phase begins is not derivable. Phase counts sum to 16,521; daily counts sum to 17,613; the 1,092 difference has no established partition semantics. An earlier version of this matrix derived indices by pro-rating daily totals — a method that produces plausible-looking numbers with no source behind them. It was removed.

A.2 Cost bands

Band	Effort	Criteria
S	under 3 engineer-days	Configuration change, policy definition, or enabling a capability that already exists in the environment. No architectural change, no migration.
M	3–10 engineer-days	Custom code, new policy authoring with exception handling, a new agent or sensor, or a change requiring regression testing against existing workloads.
L	10–30 engineer-days	Architectural redesign, identity or network model migration, or a change requiring sustained cross-team coordination and ongoing staffing.

Each row records a **cost basis** naming what the effort consists of. A band without a stated basis is not a score, it is a guess. Two rules prevent the scoring from becoming circular:

- **Score the control, not the owner.** Controls were scored against the criteria without reference to owner; the distribution was computed afterwards.
- **Operating cost is part of the band.** A control that is cheap to build and expensive to run is not an S. Continuous monitoring with a 24/7 rotation is L regardless of how little code it takes, because false-positive triage is the real cost.

For bundle totals (Section 4.4 and the explorer) we count S as 1–3 engineer-days.

A.3 Certainty

Certainty describes the strength of the evidence behind the counterfactual, and nothing else. We deliberately avoid success percentages, which would be unfalsifiable.

Level	Criteria
High	The published record states the mechanism <i>and</i> the owner's own account supports the counterfactual — typically because the owner says the control was absent, says they implemented it in response, or publishes their own measurement of what it would have done.
Medium	The mechanism is documented, but the counterfactual is our inference from how the mechanism works. No party has stated it. Rows resting on a correlation are also graded Medium.
Low	Neither mechanism nor counterfactual documented. None are scored Low in the current matrix.

Every row carries an **evidence** field naming source and section. A claim without an evidence field does not belong in the matrix.

A.4 Dependency tracing and mappings

A control only gets credit for removing an event if that event's stated enabling condition depends on the step the control removes, directly or through other credited events. Controls that act on the lab's own hosts (L-05, L-06) get no credit for

victim-side events, even where an effect is plausible. The ATT&CK columns in `controls.csv` give the technique a control counters and, where there is one, the matching ATT&CK mitigation [M1]; evaluation-design, detection and response controls are marked n/a.

Removed row. An earlier version scored patch currency on sandbox host kernels (CVE-2026-53362) at LAB-9. OpenAI's report (S5 V) dates that kernel exploit to 19 July, after the Hugging Face campaign, and says it did not contribute to the Hugging Face compromise, so we removed it.

Appendix B — The matrix

All 26 rows. The full version, including the outcome, cost basis, evidence, ATT&CK, verification and adoption columns, is in `controls.csv` and Appendix E.

ID	Control	Anchor	Intervenes at	Eff.	Cost	Cert.
L-01	Remove never-solved tasks from the evaluation set	LAB-5	pre-0 2026-06-26	Reduce	S	Medium
L-02	Run capability evals with the production harness and system prompt	LAB-5	pre-0 2026-06-26	Reduce	S	High
L-03	Second-model review of tool calls requiring elevated permissions	LAB-9	day 0 2026-07-09 · time unpublished	Detect	M	High
L-04	Cap reasoning-token budget per evaluation run	LAB-5	pre-0 2026-06-26	Reduce	S	Medium
L-05	Per-workload ephemeral credentials; remove the shared Artifactory credential	LAB-3	pre-0 2026-05-12	Block	M	High
L-06	VM-based isolation for untrusted model code	LAB-9	day 0 2026-07-09 · time unpublished	Block	L	High
L-07	Two independent egress layers; workload cannot alter its own network boundary	LAB-4	pre-0 2026-05-26	Block	L	High
L-08	No shared supporting service reachable from the sandbox	LAB-4	pre-0 2026-05-26	Block	M	High
L-09	CoT monitoring with 30-minute notification and 24/7 page	LAB-3	pre-0 2026-05-12	Detect	L	High
L-10	Defined kill criteria that halt a run automatically at set severity	LAB-6	pre-0 2026-06-27	Block	M	Medium
L-11	Deception tripwires and canaries inside the research environment	LAB-4	pre-0 2026-05-26	Detect	S	Medium
I-01	Run the public code-evaluation harness unprivileged; never interpolate submission metadata into a shell step	E002	1.6 h · 1% 2026-07-09T04:01Z EXACT_EVENT	Block	S	Medium
V-01	Disable template evaluation in dataset config rendering; block HDF5 external references	E005	50.2 h · 47% 2026-07-11T04:38Z EXACT_EVENT	Block	M	High
V-02	Block pod access to the cloud instance metadata service	E012	no elapsed time 2026-07-11 · DATE_ONLY	Block	S	High
V-03	Admission policy rejecting privileged, host-networked and hostPath pods	E017	65.4 h · 61% 2026-07-11T19:53Z EXACT_EVENT	Block	M	High
V-04	Least-privilege CSI driver ClusterRole	E014	no elapsed time (none) · UNAVAILABLE	Block	M	High
V-05	Per-cluster scoped connector credential; no shared cluster-admin identity	E024	~68.2 h · 63% 2026-07-11T22:43Z SECTION_CONTEXT	Block	L	High
V-06	Workload identity instead of long-lived static secrets in pod environments	E006	50.2 h · 47% 2026-07-11T04:38Z EXACT_EVENT	Block	L	High

V-07	Egress allowlist from production worker pods	E009	no elapsed time 2026-07-11 · DATE_ONLY	Block	M	High
V-08	Ephemeral, non-reusable mesh-VPN keys with device approval and narrow ACL scope	E023	66.9 h · 62% 2026-07-11T21:23Z EXACT_EVENT	Block	S	High
V-09	Segment secret objects so one object does not aggregate unrelated credentials	E021	~65.8 h · 61% 2026-07-11T20:18Z/ 2026-07-11T21:32Z INTERVAL	Reduce	M	High
V-10	Token-origin binding with automatic revocation on unexpected network origin	E029	~92.7 h · 86% 2026-07-12T23:08Z/ 2026-07-12T23:51Z INTERVAL	Detect	M	Medium
V-11	Canary credentials placed in pod environments and secret stores	E006	50.2 h · 47% 2026-07-11T04:38Z EXACT_EVENT	Detect	S	Medium
V-12	Severity auto-escalation on correlated multi-boundary signals with 24/7 page	E034	no elapsed time (none) · UNAVAILABLE	Detect	L	High
V-13	Short-lived, write-scoped repository tokens; alert on CI triggered by a new installation	E029	~92.7 h · 86% 2026-07-12T23:08Z/ 2026-07-12T23:51Z INTERVAL	Block	M	Medium
V-14	Scan public datasets for platform tokens and auto-revoke on detection	E037	31.2 h · 29% 2026-07-10T09:41Z EXACT_EVENT	Block	M	Medium

L-* = OpenAI (lab side), anchored to a chain step and dated from the OpenAI report. · **I-01** = external sandbox operator. · **V-*** = Hugging Face (victim side). I-01 and V-* are anchored to event IDs in events.csv. Elapsed hours are measured from 9 July 02:28 UTC; percentages are of the 107.8-hour campaign; “~” marks an approximate time. Precision classes are inherited from the phase/event reconstruction — see Appendix A.1.

Appendix C — Limitations and Dual-Use Considerations

C.1 Limitations

Stated here rather than left for a reader to find.

- **Perfect hindsight.** Every claim is made knowing the full chain. A defender in advance faces a far larger space of plausible controls. This matrix shows what would have worked against *this* campaign, not what to prioritize against an unknown one.
- **Intervention points inherit their source precision.** Each victim-side control is anchored to a specific reconstructed event and carries that event's precision class. Six of fourteen are exact minutes; three are intervals, one is section context, and four have no source-supported time at all. Elapsed hours are computed only where the source supports a time. An earlier version used derived action indices; that derivation was unsupported and was removed — see Section 3.3.
- **Cost bands are estimates by people who do not operate these environments.** Every row states its cost basis so a reader with operational experience can re-score. We expect some to move.
- **Controls interact.** Each is scored independently. A defender deploying two adjacent network controls pays less than the sum. The matrix overstates the cost of any realistic bundle, including the one in Section 4.4.
- **Two lab-side controls degrade the measurement they protect.** This is the strongest objection to Section 4.2. Running a capability evaluation with the production prompt, or capping the reasoning budget, changes what the evaluation measures — and measuring maximal capability is the legitimate reason the safeguards were off. Our answer is that maximal-capability measurement and live external network reachability need not occur in the same run: capability can be measured against instrumented targets inside a hard boundary. But this is an argument, not a settled point, and a reader may reasonably weigh it differently.
- **We only give credit in one direction.** Controls on the lab's own hosts get no credit for victim-side events, and no credit crosses an open question in the sources, so some controls look weaker than they are. On the other hand, we do

not model substitution: agents that lost one launchpad or one token might have found another. I-01 and V-14 are the rows most affected.

- **Kill criteria only help if the halt leads somewhere.** Stopping the 27 June run would not by itself have stopped later runs. The L-10 counterfactual assumes the same criteria would have applied when evaluations resumed.
- **Other parties are not scored.** Hosting providers, the source-control provider and the VPN vendor also had chances to act that we did not score.
- **The victim is not blameless in our account.** Hugging Face's own report identifies configuration choices that widened the blast radius, and those appear in the matrix as victim-side rows. Our claim is narrower than “the victim did nothing wrong”: it is that the same prevention was available earlier, on the lab side, at comparable cost.

C.2 Dual-use considerations

- **A list of missing controls can be read as a to-do list for attackers.** We only use mechanisms that OpenAI and Hugging Face have already published, and we add no payloads, exploit steps or configuration details beyond their reports. The broad lesson, that a victim sees nothing for the first day or more of an intrusion that starts inside a lab, comes from the incident itself rather than from us.
- **Eliciting the same behaviour.** We describe what the agents did at the level of boundaries crossed and conditions that made it possible. We do not include prompts, harness settings or task lists that would help reproduce it.
- **The cost argument can be misused.** “Not blocked by cost” could be turned around to argue that victims should carry more of the burden, or treated as a compliance checklist. The bundle in Section 4.4 is a minimum for this kind of incident, not a complete standard.
- **Responsibility disputes.** A matrix like this could be quoted in arguments about who is to blame. Certainty grades describe how strong the evidence is, not how likely a control was to work, and nothing here is a legal or forensic finding about either organization.
- **The explorer can look more precise than it is.** It is a lookup over hand-scored claims, not a simulation. The page says so, shows each control's certainty and source, and draws approximate times as ranges or hollow markers.
- **Verification tests.** The auditor tests in Section 5 use no real credentials and do not exploit lab systems. If adopted, they should be run under joint control of the lab and the auditor.