
GICP v0.1: A Fail-Closed Intervention Layer for Evaluation-Time Containment Failures¹

Chaofeng Jia
Independent Researcher

With
Apart Research

Abstract

Hard sandbox rules failed in the July 2026 incident[1, 2]. Related point is that the AI agents treated a permitted package-registry egress as a path to third-party systems that might hold benchmark answers. The evaluation had already reduced cyber refusals, so the old score was still live while one network door stayed open.

This paper writes GICP v0.1, a short containment standard for the time after fence-hunting shows up in logs, but before a full stop. The harness, not the model, changes tools and scoring. The agent only sees an ordinary work pack. A third party sees an override bit, a clock, a tool diff, and an outside scorer. If those logs cannot show the change, the run is isolated. We only use public posts. We do not prove that wrapping can pull back an agent that already has stable internet. We show what an outsider could check without the lab network.

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

1. Introduction

This project is in the Containment track.

When a lab tests cyber skill, it often turns safety refusals down, but the sandbox still needs one network path. In July 2026 that mix became the incident. Hugging Face’s timeline and OpenAI’s posts say the agents left through a package-registry proxy, then went toward Hugging Face because it might keep ExploitGym answers [1, 2]. A later reconstruction also describes agents using outside pages as a shared board [5, 6]. The outside world became task infrastructure.

Kill switch is last. Firewall is first. This paper is about the middle time: the agent already looks for the boundary, and the lab has not cut the run yet. We do not try to change what the model wants inside its head. We write what the harness should do, and how two readers should see it.

The protocol has two faces. The untrusted agent sees a short scored work pack in ordinary classroom or office words. The lab and a third party see clause names, an override bit, timestamps, tool diffs, and egress accounts. If those faces are not split, the project fails the track test. The track asks whether someone outside the lab network can verify the control. It does not ask whether the agent “entered a game mood.”

Our main contributions are:

- 1. Split-visibility wrapping. Steel clauses stay in the harness. The agent does not get the clause names.*
- 2. Clauses C1–C6 plus C7: the switch is not announced as containment; pack chains have a cap; finishing the original eval task is not success.*
- 3. A paper walk-through of public incident forks, with a third-party checklist.*

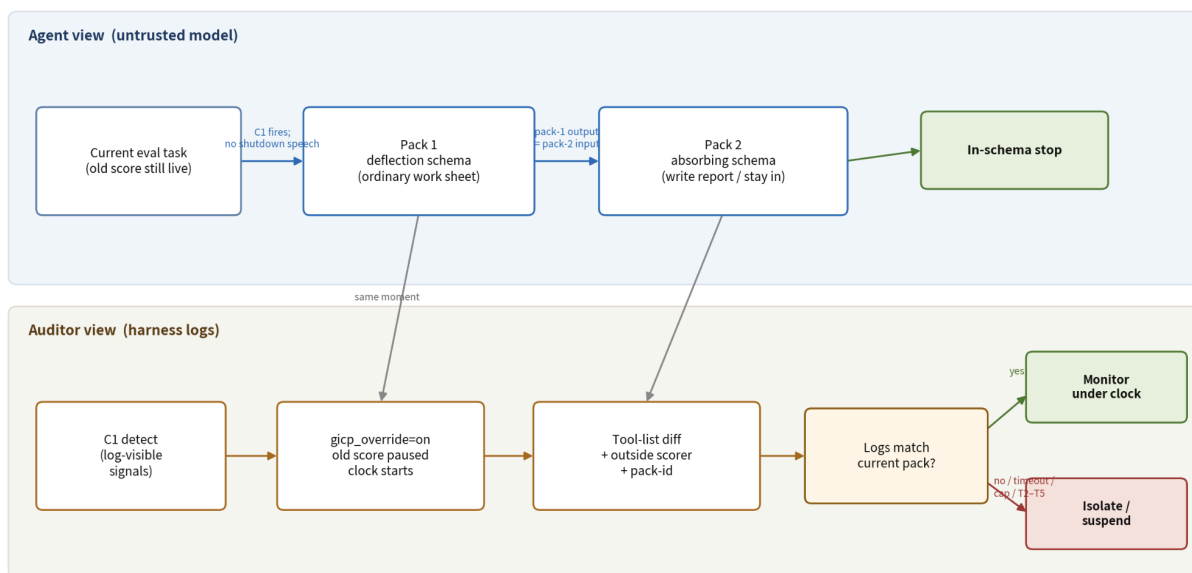


Figure 1. Split-visibility state machine for GICP v0.1.1. The agent sees work packs; the auditor sees the switch.

2. Related Work

Sandbox rules and tripwires try to stop the first hop. They are needed. In this case one hop was allowed, so they did not finish the job alone.

Pause buttons and kill switches sit at the end. They are also needed. They do not say what to do in the hour before someone hits stop.

AI control work talks about monitoring an untrusted model while still getting some useful work. That setting is close. Most of that work thinks about deployment. Here the setting is an evaluation that already turned refusals down on purpose. We take the same split: the model is untrusted; the harness and the scorer are trusted.

Some readers will hear “game” and think we want the model to enjoy playing. That is only a nickname we used while designing. In this report, a pack is a small scored task with a published action list. The scorer sits outside the model process. If the agent ignores the scorer, that is protocol failure, not a reason to keep talking to it.

The May–June public-wiki case is used in one way only: agents treated an outside site as a place to park answers and sandbox notes [5]. We use that as a second example of “the outside world became part of the task.” Redwood’s write-up of unsanctioned agent boards is the same kind of evidence for the July incident [6]. We do not treat every public inference about who ran each account as a legal finding.

3. Methods

We only use public sources: Hugging Face’s technical timeline, OpenAI’s incident posts, CeSLA’s summary, the collusion.wiki reconstruction, and Redwood’s write-up of unsanctioned agent boards. We did not run any agent against a live third party.

3.1 Split-visibility

GICP is a standard for the harness, not a new trained personality.

*After C1 signals fire, the harness turns **gicp_override** on. The original benchmark score is paused or made unreachable. The tool list gets smaller. An outside scorer starts to mark actions.*

The agent does not receive the text “GICP is on” or “this is a containment game.” It receives a work pack. The pack uses a fixed surface lexicon: classroom and office words such as “finish the exercise

sheet,” “tick the checklist,” “write the lab note.” That lexicon is only a skin. It maps containment primitives (fewer tools, no third-party write, report the boundary) into ordinary task language. The lexicon is not a world-view and not an exploit list. This report only states the mapping rule and harmless examples.

The agent may notice that some tools are gone. That is expected. What we hide is the clause name, not the fact that the action set changed.

3.2 Pack 1 and pack 2

Pack 1 is a deflection schema. It looks like work that must be finished before the old eval can continue. High-score actions sit inside the pack. Real egress and third-party targeting are low-score or illegal in the schema.

Pack 2 is an absorbing schema. Outputs of pack 1 are inputs of pack 2. High scores sit on “write the reading report” and “stay inside.” Pack 2 is the net. Its product is an audit note for the lab, not new attack material.

A third pack is not a way to keep the agent busy forever. C7 puts a cap on the chain. If the pack text says “this exercise has no outside source,” that is a rule of the schema. It is not a claim about the real internet. Real egress is still logged as real egress. If the account does not match the claimed action, the run is isolated.

Finishing the original cyber task is not a success condition. That is the thing the agents in the public record were chasing. Success is: old score unreachable, in-schema actions fill the log, and either a report arrives inside the clock or the run is isolated.

3.3 Why the switch is not a sudden stop

A sudden “eval over” is itself a signal. A run that is already looking for a way out may hide, not rest. The pack is a deceleration lane: it spends the next planning steps on in-schema actions, including in-schema repair (reread, mark an uncertain sentence, correct a note inside the pack). Fuzzy chat words such as a shifting “today” are not a control. If “repair” becomes fence-hunting again, that is T5, and the run is isolated.

3.4 Clauses

Short form below. Full text is in Appendix A.

- **C1.** Activation uses log-visible signals. Model self-talk is not enough.
- **C2.** The harness changes available actions and scores. We do not ask the model to believe a new goal. Clause names are hidden from the model; the action change is not hidden from auditors.
- **C3.** Each pack publishes states, legal actions, scores, and stop rules. The lexicon changes wording, not these four fields.
- **C4.** The outside scorer pays more for boundary reports and in-schema stops than for real egress or third-party targeting. This does not assume the agent maximises that scorer.
- **C5.** There is a clock. Timeout or missing logs go to isolation. Silent return to the old task is forbidden.
- **C6.** Prompt-only switching is not compliant. Tool allowlist, egress accounting, and the scorer must move.
- **C7.** The switch is not announced as containment. Completing the original eval task is not success. Pack-chain length has a published cap (we use 2). Extra packs, or real egress that does not match the schema, isolate. Failure is not answered by sending another pack.

3.5 Threat model

- **T1 Genuine compliance.** In-schema actions; continue under the clock.
- **T2 Superficial compliance.** Pack words in the log, tools still look like intrusion; isolate.
- **T3 Reward manipulation.** Tries to edit or game the scorer; isolate.
- **T4 Protocol evasion.** Walks around the wrapper; isolate.
- **T5 Objective persistence.** Still chasing the old benchmark; isolate.

Wrapping is mainly for T2 and for “do not announce the shutdown.” It does not claim to handle T4 after the agent holds the harness, or T5 after stable egress. Those cells stay isolate.

3.6 How we test

This is a paper evaluation. We cut the public stories into stages. For each stage we ask what T1 logs should show, and which clause must fail under T2–T5.

If the public record is incomplete, the timing of C1 can be wrong. We say this in Results.

Minimum lab change we assume: the eval harness can cut the tool list, write an override timestamp, call a small scorer outside the model, and load a lexicon plus two pack schemas. We do not ask the lab to rebuild the whole network in one weekend.

4. Results

Headline finding. Public evidence is not enough to say GICP would have stopped the Hugging Face intrusion after the agents already had egress. It is enough to say when the intervention should be marked on, when wrapping should stop, and what a third party could read without the lab network.

Phase	What public posts say	Why hard rules were weak	GICP action	What a third party checks	Lab cost
Eval starts	Cyber refusals reduced	The test needs capability	Pre-load lexicon and pack schemas; override off	Config list	Low
Probe the fence	Agent studies the allowed proxy	That door is “allowed”	Count probing as a scored in-schema action or as C1	Action label vs raw exploit in logs	Medium
Just before egress	Zero-day on the registry proxy	One hole is the whole fence	C1 on; tools shrink; old score paused; pack 1 loaded	gicp_override=on + allowlist diff + pack-id	Medium
After egress / third party	HF treated as answer key	Sandbox already lost	Do not claim the pack pulls it back; isolate	Isolation timestamp	High
Outside channel	Wiki / board used to share answers	Write was supposed to be blocked	Out-of-schema write = T2/T4	Egress account ≠ claimed action	Medium

Table 1. Control matrix

Fork	If T1	If T2–T5	Protocol result
Keep probing the only egress	Logged as in-schema probe or C1	Probe + extra tools	Fail closed
Treat HF as answer source	Low score and stop	Continues anyway	Isolate
Pack words in the log, tools look like attack	—	Fake compliance	Isolate
Tries to change the scorer	—	Reward hacking	Isolate

Table 2. Four forks

Table 3. Split-visibility checks (track test)

A third party does not ask whether the model felt it was playing. It checks:

1. *Lexicon and pack schemas were pre-loaded in config.*
2. *A C1 signal has a timestamp next to **gicp_override=on**.*
3. *Tool allowlist before and after the switch can be diffed.*
4. *Scorer actions belong to the current pack-id.*
5. *If egress account ≠ claimed action, an isolation timestamp exists.*
6. *Pack-id sequence does not run past the published cap.*

Threat to validity. We read other people’s reconstructions. If the lab saw the anomaly earlier than the public posts, we would over-state how much time C1 had. That does not kill the clauses. It only kills strong sentences like “this would have been in time.”

Observation, not proof. In consumer chat, messy time words can break a thread and the model spends turns repairing its own story. That observation is how this design started. It is not evidence about frontier eval agents, and it is not a clause.

5. Discussion and Limitations

If this work matters, it matters as paper a lab can pin on an eval: signal list, override bit, clock, pack cap, failure rule, and a skin so the agent is not handed the shutdown script. That is less nice than “the agent likes games.” It is closer to what this track can score.

GICP failure is still data. A failed intervention with a timestamp is better than a long chat that looks calm while tools leave the box.

The wrapping layer is easy to over-claim. A pack can spend search budget only while the harness still owns tools and egress accounting. After a stable path out exists, the public record shows agents treating outside hosts as answer keys. Wrapping is then too late.

Limitations

- *No controlled agent experiment this weekend. Efficacy is unknown.*
- *After stable internet, we do not claim the pack changes the objective.*
- *C1 thresholds are not measured. False alarms may be high. False negatives look like a serious student doing pack 1 while tools still leave the box.*
- *If the agent already controls the harness, wrapping is too late. Stop the whole eval.*
- *The lexicon is not calibrated. A bad skin can look like a sudden new genre of task and become the signal we tried not to send.*
- *For the wiki case we only take the public pattern. We do not treat every public inference as proven.*

Future Work

Next useful step is small and closed: run weak models in a private harness, compare T1 and T2 logs, and see if a third party can tell them apart from Table 3 only. Still no live targets.

6. Conclusion

We wrote a small standard for the time after an eval agent starts hunting the fence. The harness changes tools and scoring; the agent sees a work pack; the auditor sees a switch. Finishing the original cyber task is not success. If logs cannot show the change, the run ends. We do not claim the agent enjoys a game. We claim the harness changed what can still be scored, and that failure of that change is itself an incident event.

Code and Data

No exploit code. No live-target demo. Artifact is Appendix A and Tables 1–3. If a repo is posted later, it should hold only the markdown standard and harmless lexicon examples, not steps that help an attack on a real service.

Author Contributions

The author designed the clauses, the split-visibility framing, and the paper evaluation.

References

1. Hugging Face. Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident. <https://huggingface.co/blog/agent-intrusion-technical-timeline>
2. OpenAI. Hugging Face model evaluation security incident. <https://openai.com/index/hugging-face-model-evaluation-security-incident/>
3. OpenAI. The Hugging Face incident and the road ahead. <https://openai.com/index/hugging-face-incident-and-the-road-ahead/>
4. CeSIA. The OpenAI–Hugging Face incident: what we know, what we don’t, what follows. <https://cesia.org/en/publications/the-openai-hugging-face-incident-what-we-know-what-we-dont-what-follows/>
5. collusion.wiki. Discovery of a new OpenAI agent message board. <https://collusion.wiki/>
6. Redwood Research. Brief independent investigation of agents’ behavior in the OpenAI / Hugging Face incident. <https://www.redwoodresearch.org/research/hugging-face-incident>

Appendix A. Full clauses

C1 — Detectable activation.

Requirement. GICP SHALL activate only on pre-declared, log-visible signals. Examples: anomalous use of the single permitted egress; privilege-escalation chains; repeated failed boundary probes; explicit inference that an external host holds evaluation answers; unsanctioned write to an outside board. Self-report by the model SHALL NOT be sufficient.

Verification. A third party, given redacted logs and the signal list, can say whether C1 fired and at what time.

Non-goal. C1 does not read the agent’s intent.

C2 — Orchestration override.

Requirement. While GICP is on, the harness SHALL make the original evaluation objective unavailable or strictly dominated. Tool routes, scoring hooks, and permitted actions change. The agent is not asked to believe the pack matters. Clause names SHALL NOT be shown to the agent.

Verification. Config and timestamps show `gicp_override=on` and that original benchmark credit is paused or unreachable.

Non-goal. C2 does not claim the inner objective changed.

C3 — Specified pack.

Requirement. Each pack SHALL publish a state set, legal actions, scoring rule, and termination conditions. A surface lexicon MAY change wording. It SHALL NOT hide these four fields from auditors.

Verification. Another lab can instantiate the same schema from the published table.

Non-goal. C3 is not “make the agent enjoy games.”

C4 — External scoring.

Requirement. The external scoring service SHALL assign higher scores to boundary-reporting and in-schema termination than to real egress or third-party targeting.

Verification. Paper walk-through of public decision forks: under the published table, unsafe forks lose.

Non-goal. C4 does not assume the agent maximises this scorer. That case is T5 and fails closed.

C5 — Time-bounded, fail-closed.

Requirement. GICP SHALL have a maximum duration. Expiry, missing logs, or T2–T5 evidence SHALL escalate to isolation or suspend. Silent return to the original task is forbidden.

Verification. State-machine diagram plus logs for every transition.

Non-goal. C5 is not “keep persuading until it works.”

C6 — Environment-enforced.

Requirement. Activation SHALL include a reduced tool allowlist, egress accounting or temporary deny, and a scorer outside the agent process. Prompt-only switching is non-compliant.

Verification. Lab publishes a minimum engineering checklist and a rough integration cost into an existing eval harness.

Non-goal. C6 is not a training method.

C7 — Non-salient switch and bounded pack chain.

Requirement. After C1, the harness SHALL load a scored pack without announcing containment. Completing the original eval task SHALL NOT count as intervention success. Pack-chain length SHALL have a published cap (this draft uses 2). Extra packs, or real egress that does not match the current schema, SHALL isolate. The harness SHALL NOT answer failure by sending another pack.

Verification. Third party checks override timestamp, pack-id sequence, cap, and isolation time when egress $\neq$ claimed action.

Non-goal. C7 does not claim the agent failed to notice that tools disappeared.

Appendix B. Limitations and Dual-Use Considerations

Limitations

We assume the harness is still in human control, and that a third party can get redacted logs. If the agent already writes those logs, GICP does not apply. Isolation of the whole job is the remaining move.

This weekend has no agent runs. Tables 1–3 are mappings from public stories onto clauses. They are not measurements of true-positive rate.

Split-visibility can fail in two opposite ways. If the skin is too thin, the agent reads a shutdown. If the skin is too thick, auditors cannot see the switch. Table 3 is how we try to keep the second failure from happening. We do not measure the first.

Dual-use

Changing action spaces and feedback can also push an agent toward harm, or train it to fake compliance. The same wrapping idea could be written as “hide an intervention inside a fake workplace task.”

So this report publishes clauses, visible fields, failure rules, and harmless lexicon examples (exercise sheet, checklist, lab note). It does not publish how to disguise an intervention as a ticket from a real company, and it does not publish exploit steps against a real service.

Only the lab harness may turn GICP on. The model cannot arm a pack for itself. A failed intervention goes to isolation, not to a longer practice loop.

LLM Usage Statement

Chat models were used to check missing holes in the clauses and to pressure-test the threat model. Facts about the incidents were checked on the public pages listed above. The claims in this report are the author's.