

---

# Correlating Weak Signals Paged a Synthetic Agentic Intrusion Before Any Secret Was Read<sup>1</sup>

---

Yassine Bentayfor  
Independent Researcher

With  
Apart Research

## Abstract

*Agentic intrusions can spread risk across many individually ambiguous events, which makes per alert paging a poor fit for activity at machine speed. I built AgentPage, an auditable detector that scores distinct suspicious telemetry categories inside a sliding window on each host. I compared it with critical only, high severity, and rate only pagers across three canonical synthetic streams and 90 deterministic timing and volume variants. AgentPage detected all 30 constructed intrusions, on average 106 seconds (4.6 events) earlier than the high severity baseline, and in every attack variant it paged before the first secret was accessed. The rate only baseline missed every intrusion and paged on 11 of 60 benign variants. AgentPage paged on 6 benign credential rotation variants. Ablations showed that category deduplication kept repeated benign detections from raising the false positive rate to 0.50, while removing the diversity bonus eliminated false pages but erased the timing advantage. This is a transparent, reproducible feasibility result, not a production accuracy claim: cross category correlation trades a measurable false page cost for earlier escalation.*

---

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

## 1. Introduction

Hugging Face reconstructed approximately 17,600 attacker actions in the 2026 incident involving agents that escaped an OpenAI evaluation environment. The successful path was buried among thousands of failed actions across short lived systems. Hugging Face reports that runtime, SIEM, and other signals were correlated into a coherent attack signal, but the system did not raise its criticality and did not trigger the on call team [1]. This points to a narrow operational gap between "the system sees related anomalies" and "a responder is paged."

My research question is: can a simple sequence aware rule page responders earlier than conventional severity based escalation, while staying quieter than volume based detection on bursty benign automation? I define "sequence aware" as correlating distinct suspicious outcomes on one host within a short time window. I do not attempt to identify an AI actor, reconstruct the incident, or reproduce any exploit.

### **My main contributions are:**

- A compact, configurable paging rule that deduplicates repeated signals and combines credential, control plane, egress, privilege, and persistence outcomes.
- Comparisons with critical only, high severity, and rate only paging baselines across canonical streams and 90 deterministic variants, measured in both events and seconds.
- Threshold and window sensitivity analysis plus component ablations that expose the false page cost of earlier escalation.

## 2. Related Work

The Hugging Face technical timeline is my primary motivation and factual source. It describes familiar weaknesses explored at unusual volume, argues that the defensive problem is correlating thousands of low signal events across systems, and calls for detection that correlates quickly enough to keep up [1]. AgentPage abstracts only the phase and boundary structure into harmless telemetry categories. It does not use the released commands or raw payloads.

Elastic Security Labs maps the incident to production endpoint, Kubernetes, cloud, and network detections. Its analysis emphasizes outcome detections, such as credential access, unusual destinations, sensitive workload creation, and persistence, rather than trusting a parent process or tool [2]. AgentPage assumes those lower level detections already exist and tests a downstream escalation policy over their normalized categories.

NIST SP 800-61r3 places detection inside a broader incident response capability intended to reduce incident impact and improve response effectiveness [3]. My contribution is deliberately smaller: a reproducible paging harness for one correlation hypothesis. It is not a replacement for a SIEM, EDR, containment control, or incident response program.

### 3. Methods

**Event model and scenarios.** Each JSONL event contains a timestamp, scenario, host, category, severity, and plain language description. The canonical synthetic intrusion mixes routine process events with credential access, control plane discovery, unusual egress, secret access, privilege escalation, and persistence. The benign CI stream contains rapid package, build, process, artifact, and expected network events. The administrative family contains authorized authentication and cluster inspection split across hosts. The non canonical admin variants include repeated inventory alerts, and every fifth variant also combines legitimate control plane inspection with secret and credential rotation. No stream contains exploit instructions, payloads, real credentials, or copied incident telemetry.

**Detectors.** I compare four policies: critical only pages on one critical event; high severity pages on one high or critical event; rate only pages after 12 events from one host within 60 seconds, regardless of meaning; and AgentPage correlates distinct suspicious categories within a ten minute window on each host.

AgentPage assigns 2 points each to credential access, secret access, and unusual egress; 1 point to control plane discovery; and 3 points each to privilege escalation and persistence. Three or more distinct suspicious categories add a bonus of 2. Twelve or more events within 60 seconds add 1, but volume alone can never reach the page threshold of 7. Each category is counted once per window, so one noisy alert repeated many times cannot impersonate a multi

stage chain. Every parameter lives in config/default.json. The values are proof of concept choices, not estimates from production telemetry.

**Evaluation.** For each canonical stream and detector, I recorded whether it paged, the event index and timestamp of the first page, the maximum score, and the evidence present at the first page. I then generated 30 deterministic variants of each scenario family. Attack variants jitter the spacing between suspicious outcomes, CI variants vary volume and cadence, and admin variants model repeated authorized inventory alerts plus selected credential rotations. I measured true positive rate (TPR), false positive rate (FPR), precision, and time to page, both as the mean event index of the first page and as mean seconds from the start of the stream. A 3 × 3 sensitivity grid crosses windows of 5, 10, and 15 minutes with thresholds of 6, 7, and 8. Ablations remove the diversity bonus, category deduplication, or the rate bonus. All reported proportions describe these 90 constructed variants, not a random sample of real environments.

## 4. Results

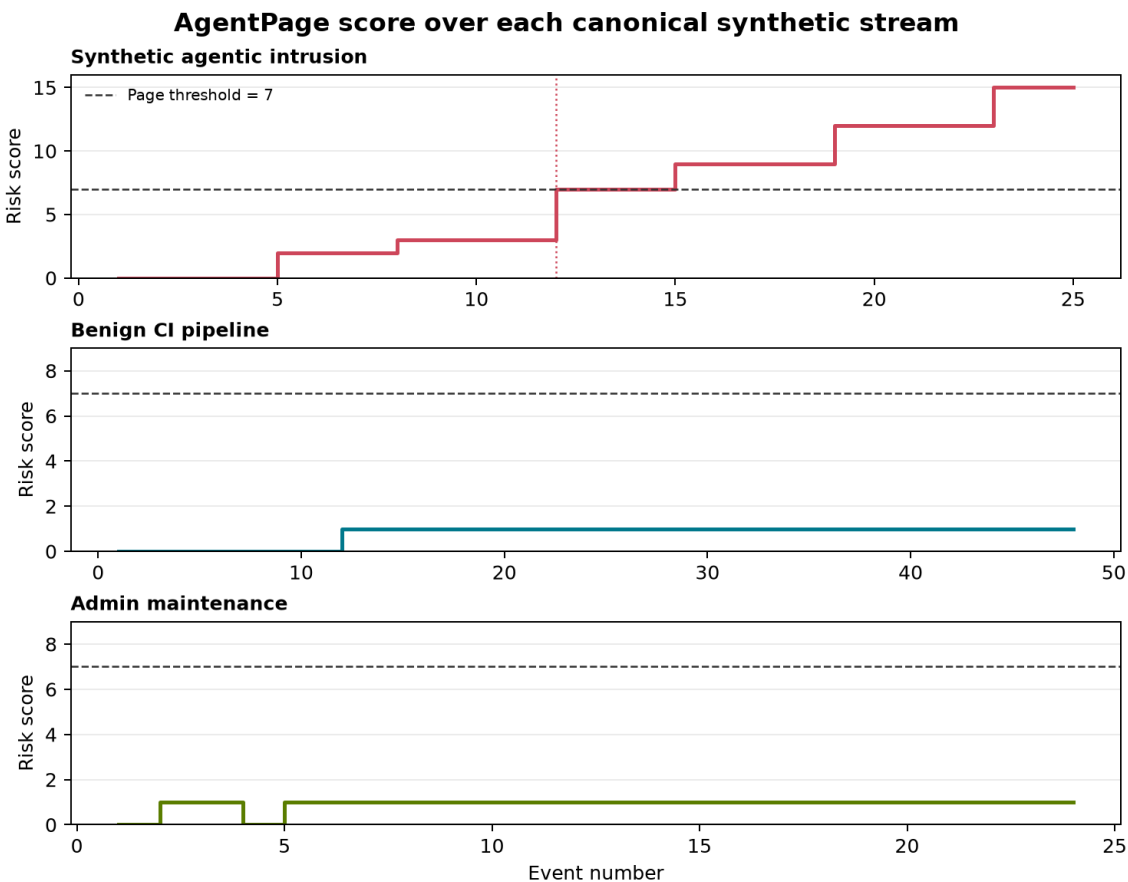
On the canonical intrusion, AgentPage paged at event 12 (14:04:00Z) with a score of exactly 7: credential access +2, control plane discovery +1, unusual egress +2, and the three category diversity bonus +2. The high severity baseline did not page until event 15 (14:05:15Z), which is the secret access event itself. The critical only and rate only baselines never paged. On canonical CI, the rate only baseline paged at event 12 while AgentPage and both severity baselines stayed silent. No detector paged on canonical administration.

*Table 1. Variant set comparison across 90 deterministic streams.*

| Detector      | TPR   | FPR   | Precision | Mean page | Mean seconds |
|---------------|-------|-------|-----------|-----------|--------------|
| Critical only | 0.000 | 0.000 | n/a       | n/a       | n/a          |
| High severity | 1.000 | 0.000 | 1.000     | 16.93     | 371.7        |
| Rate only     | 0.000 | 0.183 | 0.000     | n/a       | n/a          |
| AgentPage     | 1.000 | 0.100 | 0.833     | 12.33     | 265.4        |

Across the 90 variants, the high severity baseline detected all 30 constructed attacks with no false pages, at a mean of 371.7 seconds into the stream. AgentPage also detected all 30, at a mean of 265.4 seconds, which is 106 seconds earlier on average (range 59 to 143 seconds). In all 30 attack variants, the high severity page was triggered by the secret access event, and AgentPage had already paged before that secret was read. The cost was 6 false pages out of 60 benign variants. The rate only baseline missed every attack and paged on 11 benign variants. The critical only baseline never paged.

**Figure 1. AgentPage score across the three canonical synthetic streams. The dashed line is the configured page threshold. Only the constructed attack crosses it.**



The ablations show which design choices produced the result. Removing the diversity bonus eliminated all six false pages but pushed detection back to 371.7 seconds, exactly matching the high severity baseline. Removing category deduplication raised FPR to 0.50, because repeated

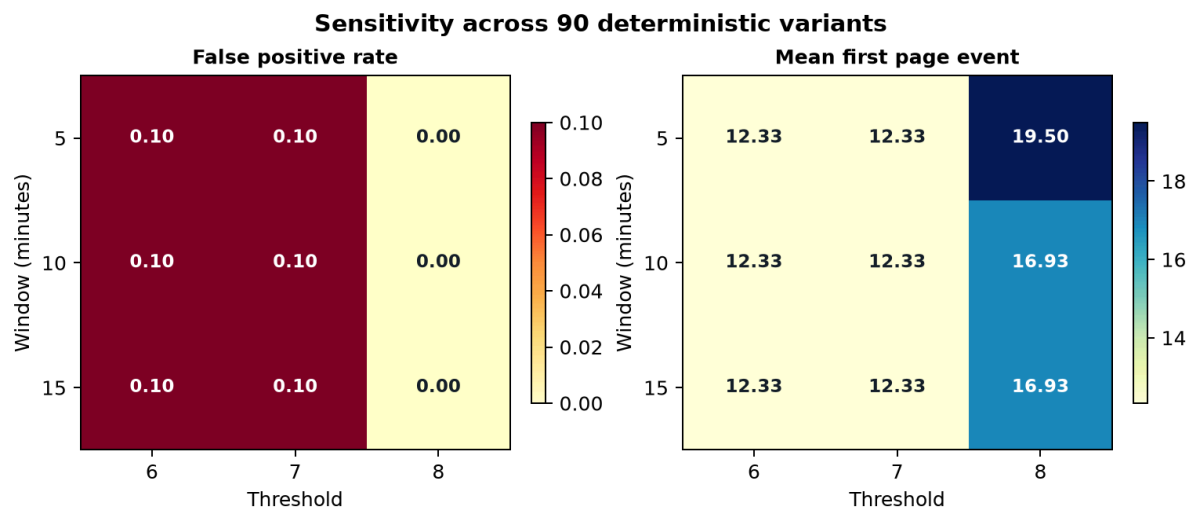
authorized inventory detections accumulated as if they were independent evidence. Removing the rate bonus changed neither classification nor timing, so that component has no demonstrated value in the current harness.

Table 2. AgentPage component ablations.

| Configuration             | TPR   | FPR   | Mean page | Mean seconds |
|---------------------------|-------|-------|-----------|--------------|
| Full detector             | 1.000 | 0.100 | 12.33     | 265.4        |
| No diversity bonus        | 1.000 | 0.000 | 16.93     | 371.7        |
| No category deduplication | 1.000 | 0.500 | 12.33     | 265.4        |
| No rate bonus             | 1.000 | 0.100 | 12.33     | 265.4        |

Raising the threshold to 8 kept 30 of 30 attack detections and removed every false positive, but moved the mean page to event 16.93 (371.7 seconds) for the ten and fifteen minute windows and to event 19.50 (431.1 seconds) for the five minute window. Window length did not change classification in this variant set, so these scenarios do not establish an optimal window.

**Figure 2. Sensitivity across 90 deterministic variants.** Threshold 8 removed false pages at the cost of later detection. Window length affected delay only at the strictest threshold.



## 5. Discussion and Limitations

The result supports a limited engineering claim: correlating ambiguous outcomes within a short window can produce an earlier page, while event volume alone is a poor decision rule. AgentPage does not dominate high severity paging. In this harness it trades six false pages for roughly 106 seconds of earlier warning, and that warning arrives before the secret access step rather than on it. At machine speed, where an agent can chain many actions per minute, whether a responder is paged before or after credentials are exposed matters more than the raw number of seconds. I think stating this operating trade off explicitly is more useful than claiming universally better detection.

The false positive and ablation results are the most decision relevant findings. All six false pages came from the credential rotation variants, where legitimate administration combines the same observable outcomes as an intrusion. Repeated detections must not be treated as independent evidence. A lab could replay the vendor neutral JSONL streams, inspect the JSON configuration and alert evidence, and substitute its own normalized telemetry without access to the original environment.

### Limitations

The streams are synthetic and I designed them by hand, including their labels, taxonomy, timing distributions, weights, and thresholds. Because I also designed the detector, the evaluation is not independent of the method. The 90 variants are perturbations of three scenario families, not independent incidents, so they cannot support confidence intervals or claims of external validity. I intentionally constructed the repeated alert and credential rotation cases to exercise specific mechanisms. The seconds reported above depend on the synthetic event spacing and are not estimates of real attacker tempo. Windows scoped to a single host miss chains distributed across hosts. I model normalized detections, not sensor quality, missing logs, delayed delivery, alert fatigue, or analyst response. I do not estimate when this rule would have fired in the real Hugging Face incident.

### Future Work

With another month, I would map a sanitized subset of the public phase level timeline into vendor neutral telemetry and replay it alongside unrelated CI and administration traces that resemble production. I would tune parameters on separate training data and lock them before evaluation. I would extend correlation across hosts using identity, workload, and trace relationships while bounding blast radius. Finally, a prospective test should measure page precision, time to page, analyst burden, and the value of each signal on independently collected traces.

## 6. Conclusion

AgentPage detected every constructed intrusion on average 106 seconds earlier than high severity paging and, in every attack variant, paged before the first secret was read. A rate only rule missed the attacks and paged on bursty benign automation. The improvement was not free: the diversity bonus caused six false pages, and removing it erased the timing advantage. Deduplication was necessary to keep repeated benign alerts from inflating risk, and the rate bonus showed no incremental value. The artifact makes these trade offs inspectable and reproducible. It does not claim production readiness.

## Code and Data

- **Code repository:** <https://github.com/YassineBentayfor/AgentPage>
- **Data:** deterministic synthetic JSONL streams generated by `simulate.py`; configuration in `config/default.json`.
- **Reproduction:** install `requirements.txt`, run `python run_experiment.py`, then `python -m pytest -q`. Replay arbitrary input with `python agentpage.py <input.jsonl> --alerts <alerts.jsonl>`.
- **Outputs:** canonical results, baseline comparison, sensitivity, ablations, alert evidence, and two figures under `results/`.

## References

1. Hugging Face Security Team. 2026. "Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident."
2. Elastic Security Labs. 2026. "Exploring the Hugging Face Breach: Mapping AI Agent Tactics to Elastic Defend."
3. Nelson, A., Rekhi, S., Souppaya, M., and Scarfone, K. 2025. "Incident Response Recommendations and Considerations for Cybersecurity Risk Management: A CSF 2.0 Community Profile." NIST SP 800-61r3.
4. OpenAI. 2026. "OpenAI and Hugging Face Partner to Address Security Incident During Model Evaluation."
5. Apart Research. 2026. "AI Incident Response Sprint."

## Appendix: Limitations and Dual Use Considerations

This artifact is a defensive proof of concept inspired by publicly documented incident phases. It is not a replay of raw Hugging Face telemetry. It contains no exploit code, payloads, targets, credentials, persistence mechanism, or instructions for accessing external systems. It does not identify whether an actor is human or automated.

Methodological limitations include hand written scenarios, three scenario families, deterministic variants, parameters I selected by hand, no production SOC baseline, and no measurement of actual Hugging Face detection latency. The rule may page on legitimate sequences and may miss slow, distributed, reordered, or sparsely logged activity. The results should not be generalized beyond this harness.

Publishing a detector can reveal what a toy rule notices. To limit misuse, the repository exposes only broad defensive event categories and synthetic descriptions, not operational evasion details. A production implementation should keep tuned thresholds and suppressions private, combine multiple independent controls, log rule changes, and require human review before disruptive containment.

## **LLM Usage Statement**

I used AI coding assistants (OpenAI Codex and Anthropic Claude) as development collaborators for drafting code, the synthetic scenarios, figures, and report text. I read the incident documentation and sources myself, made implementation changes, debugged the code, ran the tests, generated the figures, and reran the full experiment to verify every number in this report against the checked in code and result files. I reviewed and edited the final wording, sources, and conclusions, and I take responsibility for the content of this submission.