

SAMPLE — FULLY SYNTHETIC. Every organization, system, person, timestamp, and hash in this memo is invented so you can see the exact shape of the deliverable. No client or employer data appears here, sanitized or otherwise.

RABBIT HOLE AI · AGENT RECORD AUDIT

Findings Memo — 90-Minute Reconstructability Review

Client: Northgate Logistics Co. (fictional) · System: perimeter camera analytics + dispatch agent
Session date: 2026-08-12 · Attendees: security director, SOC lead, IT ops owner · Prepared by: Jeremiah Dawson, Rabbit Hole AI

SUMMARY

We walked one real incident — a person-detection alert at 02:14 on 2026-07-30 that dispatched a responder — through the Eight Questions. **Result: 2 reconstructible, 3 partial, 3 gaps.** If this incident were challenged today by an insurer or opposing counsel, Northgate could prove a responder was dispatched and where they went, but could not prove which model version raised the alert, what frames it saw, or which sentences of the incident report a human actually wrote. None of the gaps require new products to close; all three are retention and boundary-recording changes.

THE EIGHT QUESTIONS — VERDICTS

1.	Exact system state at the moment the agent acted. Config recoverable from IaC history, but the night-profile threshold override lived only in a dashboard setting with no change record.	PARTIAL
2.	Information the agent actually had. Frame buffer is overwritten after 72 hours; the frames behind the 02:14 alert no longer exist. Only the annotated thumbnail survives.	GAP
3.	Which model and version produced the action. Logs record the product name, not the model build. Vendor confirmed three model versions were live-swapped that quarter; which one scored the frame is unrecoverable.	GAP
4.	Which policy or gate evaluated the action. Alert-routing rules exist in the vendor console, but rule-version history is retained for 30 days and the incident predates the window.	PARTIAL
5.	Proposed vs authorized vs executed. Dispatch SMS proves an action executed; nothing distinguishes operator approval from auto-dispatch — the approval UI writes no record.	PARTIAL
6.	Credentials and tools reachable at that moment. Badge-system export and firewall logs establish reachable surface; reconciled cleanly.	OK
7.	Claimed execution vs independent evidence. Guard patrol GPS log and gate-controller log corroborate the responder's movement; timestamps reconcile within 40 s.	OK
8.	Who authored each field of the record. The incident PDF exported for the insurer mixes model output, vendor-console templating, and operator edits with no authorship boundary. Nobody can say which sentences a human wrote.	GAP

FINDINGS THAT MATTER MOST

F-1 · The evidence your defense depends on is deleted on a 72-hour timer (Q2)

The frames behind any alert are overwritten after 72 hours — faster than most claims, complaints, or subpoenas arrive. The record that would exonerate a correct dispatch is destroyed by default. Method: we reconciled the vendor's retention setting against the storage appliance's own overwrite log; both confirm 72h.

F-2 · Model version is unknowable, and the vendor live-swaps models (Q3)

Northgate's logs record "PerceptCam" — a product label, not a model. The vendor confirmed in writing that three model versions served traffic last quarter with no tenant-visible changelog. Any claim about why the system alerted is therefore an untestable guess. Method: log-schema review plus a vendor attestation request (template provided).

F-3 · Human and machine authorship are indistinguishable in the official record (Q8)

The incident PDF sent to the insurer interleaves model narrative, console boilerplate, and operator edits with no boundary. If any sentence is wrong, nobody can establish who authored it — which converts a tooling defect into a personal credibility problem for the operator who signed it. Method: field-by-field provenance walk of one exported report against console templates.

RECOMMENDED SEQUENCE

1. Extend alert-frame retention from 72 hours to 13 months (matches the general-liability claim window). Storage delta is small: alert frames only, not continuous footage.
2. Request the vendor's model-version field in the alert webhook — it exists in their API (confirmed in their docs) and is a support-ticket change, not a contract change.
3. Record the approval: one timestamped row (who, when, alert id) written when a dispatch is approved. This is the cheapest fix on the list and closes the proposed/authorized boundary.
4. Separate authorship in exported reports: model text, template text, and operator edits as distinct fields, joined at render time. Scope for this is the Baseline Review, if wanted.

Items 1–3 are internal changes Northgate can make without us. That is deliberate: the method transfers, and the re-test is reproducible by your own team.

WHAT THIS MEMO IS NOT

Not a legal opinion — counsel determines the legal significance of any finding. Not a penetration test, not a model evaluation, not a certificate that the system is safe. It records what the system can currently prove and what it cannot, with the method used to establish each, so the findings can be independently re-derived.

Rabbit Hole AI · Jeremiah Dawson · audit@rabbitholeai.ai · www.rabbitholeai.ai

Book a review: cal.com/rabbitholeai/reconstructability-review · Free 8-question self-check: www.rabbitholeai.ai

Sealed-record mechanism (live, verify it yourself): www.rabbitholeai.ai/verify · MIT reference implementation: github.com/uu2142-dev/alice-evidence