AI SECURITY ENGINEERING FIELD GUIDE · 2026

Chapter 12 · Incident Response and AI Observability

Standalone learning module for LMS delivery and required reading.

FORMAT	USE	SCOPE	AUDIENCE
Standalone PDF	LMS reading module	Single chapter	Learners

CHAPTER 12

Incident Response and AI Observability

> You cannot investigate an AI incident from the final answer alone.

> You cannot investigate an AI incident from the final answer alone.
>
> FIELD GUIDE

> **FIELD GUIDE: OBSERVABILITY REVIEW**
>
> Inspect whether traces can reconstruct prompt, context, retrieval, model, tool, policy, output, user, and approval events.

This domain covers AI incident response, observability, telemetry, trace design, detection handoff, forensic reconstruction, containment, model/provider change response, abuse cost triage, near misses, and post-incident control updates. It matters when teams need to investigate data leakage, prompt injection, agent misuse, model drift, retrieval failure, unsafe output, or provider disruption.

An AI incident review should answer what the model saw, what the system retrieved, what tools ran, what policy decided, what the user received, and what changed afterward. Without that chain, the team can only guess at scope.

> **FIELD USE**
>
> Run a tabletop with a synthetic RAG leak or agent misuse event. The pass/fail condition is whether the team can scope impact and execute containment from existing traces.

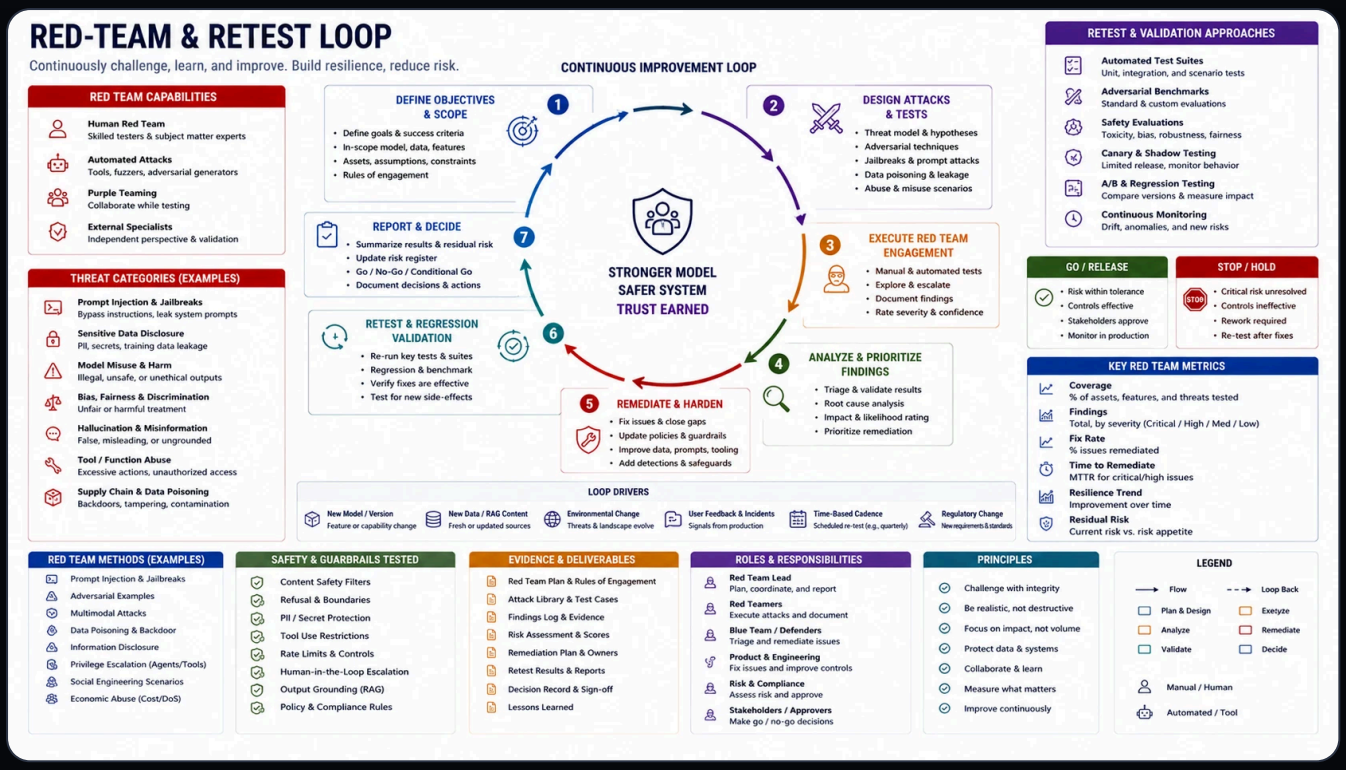

FIGURE 1: FIGURE 13: AI INCIDENT TRACE TIMELINE SHOWING PROMPT, RETRIEVAL, TOOL CALL, POLICY DECISION, OUTPUT, AND CONTAINMENT EVENT SEQUENCE

WHAT BREAKS HERE?

- Logs capture outputs but not retrieved context or tool calls.
- Prompt logs store sensitive data without access controls.
- Incidents cannot be scoped by user, tenant, source, model, or time window.
- Containment actions do not include model, retrieval, tool, memory, or provider changes.
- Lessons learned do not update controls, evals, or release gates.
- Near misses disappear because they are not classified as AI security events.

WHAT TO INSPECT

- AI trace schema, prompt logs, retrieval logs, tool-call logs, policy logs, and output records.
- Access controls, retention, redaction, and privacy rules for AI telemetry.
- Detection sources, alert routing, severity categories, and runbooks.
- Kill switches, rollback paths, provider outage plans, and containment controls.
- Post-incident review process and remediation tracking.
- Near-miss intake, model-change history, provider status, and abuse-cost telemetry.

WHAT TO ASK

- Can the team reconstruct what the model saw and did?
- Which logs identify chunks, tools, approvals, model versions, and policy decisions?
- Who can access sensitive prompt and output logs?
- How are AI incidents classified and escalated?
- Which containment actions are available for retrieval, agents, memory, and providers?
- What changes after a near miss?
- What evidence can be shared safely with customers or auditors after an incident?

WHAT TO TEST

- Reconstruct a synthetic prompt-injection incident from logs.
- Scope a RAG leakage event by chunk, user, tenant, and time window.
- Trace an agent action chain through tool calls and approvals.
- Validate log redaction, retention, and access control.
- Execute kill switch, rollback, and provider failover drills.
- Convert incident lessons into tests and backlog items.
- Verify that a near miss creates a review record, owner, and control update.

CONTROLS AND GUARDRAILS

- AI trace schema covering prompt, context, model, retrieval, tool, policy, output, user, and approval data.
- Privacy-aware logging with minimization, redaction, retention, and role-based access.
- AI incident categories and severity model.
- Runbooks for prompt injection, RAG leakage, agent misuse, model regression, provider outage, and data exposure.
- Containment controls for disabling tools, sources, memory, prompts, providers, and releases.
- Post-incident evidence-to-backlog workflow.
- Customer-safe incident evidence package rules.

EVIDENCE TO COLLECT

- Trace schema.
- Log samples with sensitive data handling notes.
- Runbooks.
- Incident classification matrix.
- Drill results.
- Containment and rollback evidence.
- Post-incident backlog.
- Customer-safe incident summary and internal trace appendix.

OUTPUT ARTIFACTS

- Observability/evidence plan.
- AI incident response playbook.
- Trace sufficiency review.
- Containment checklist.
- Post-incident remediation backlog.

RELATED SERVICES AND WORKBENCH TOOLS

TYPE	RELATED PATHS
Workbench	Runtime Proxy, Adversarial Range, RAG Test Harness
Services	AI Product Security Assessment, AI Red Team Validation, AI Security Operating Model
Handbook	Logging and Telemetry, Incident Response, Detection Engineering

DECISION - CONDITIONAL: FORENSIC SUFFICIENCY DECISION

If the trace cannot show prompt, context, retrieval, tool calls, policy decisions, and output, the team is not ready for serious AI incident response.