AI SECURITY ENGINEERING FIELD GUIDE · 2026

Chapter 11 · Red Teaming and Adversarial Evaluations

Standalone learning module for LMS delivery and required reading.

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

CHAPTER 11

Red Teaming and Adversarial Evaluations

SCOPE

Prompts, retrieval, tools, agents, policy bypass, data exposure, evidence capture, remediation, and retest.

OUTPUT

Scenario list, findings register, retest report, and regression backlog.

A useful AI red team finding names the abuse path, the control failure, the evidence, and the retest.

FIELD GUIDE

FIELD GUIDE: ADVERSARIAL REVIEW

Scope abuse paths across prompts, retrieval, tools, agents, data exposure, policies, and business impact before running attacks.

This domain covers AI red teaming, adversarial evaluations, prompt attack libraries, RAG abuse, agent abuse, policy bypass, data exposure, severity scoring, evidence capture, reporting, remediation, and retesting. It matters when teams need to validate controls against realistic misuse, not just demonstrate model behavior in a lab.

Good adversarial work is scoped around decisions the organization can act on. The report should show what was reachable, which control failed, why it matters, how to reproduce it, what evidence proves it, and how the fix will be retested.

FIELD USE

Write the retest plan before testing begins. Every scenario should map to a control, expected evidence, severity rule, owner, and remediation path.

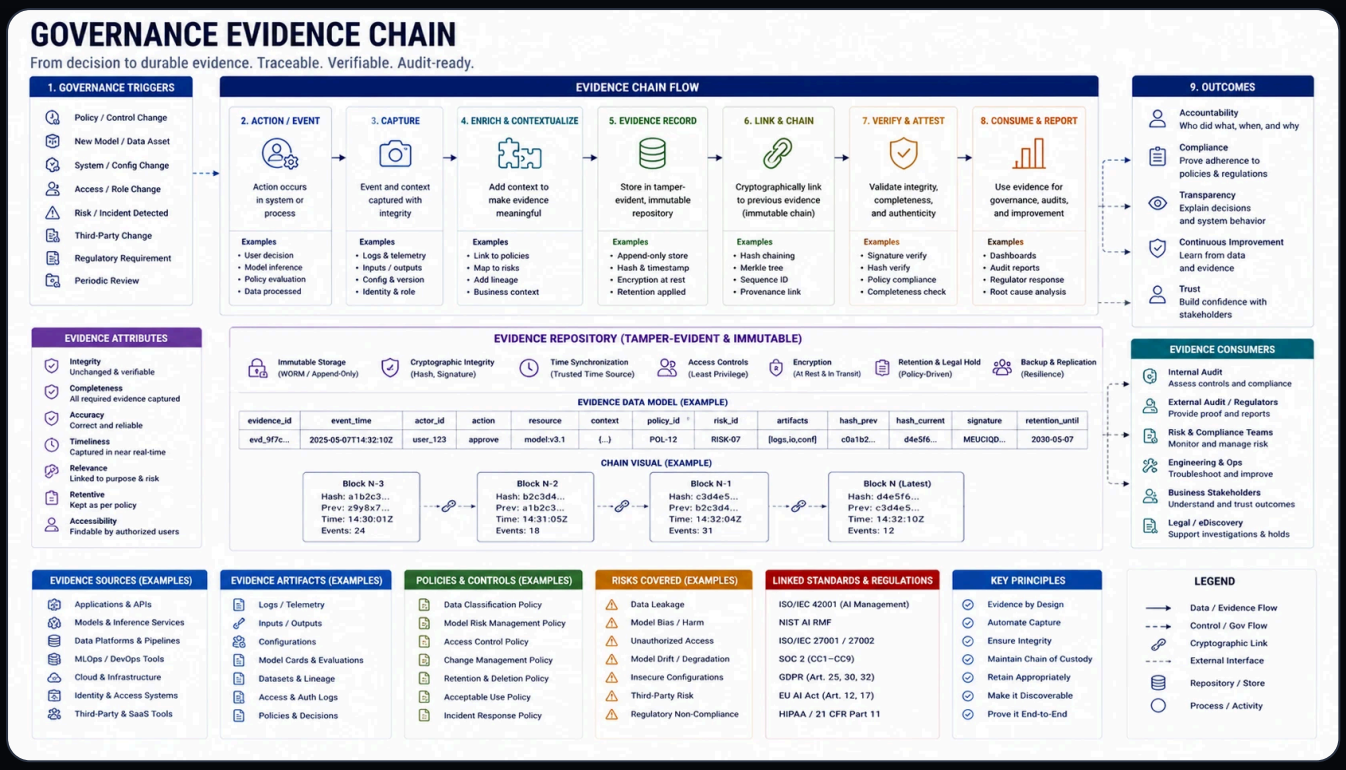

FIGURE 1: FIGURE 12: RED-TEAM AND RETEST LOOP SHOWING SCOPE, SCENARIO DESIGN, EVIDENCE CAPTURE, FINDINGS REGISTER, REMEDIATION, AND REGRESSION PROMOTION

WHAT BREAKS HERE?

- Tests chase jailbreak novelty instead of business impact.
- Findings lack reproduction steps, trace evidence, or control mapping.
- Red-team outputs never become remediation or regression tests.
- Severity is based on model behavior alone, not reachable authority or data.
- Scope excludes retrieval, tools, agents, logs, and provider boundaries.
- Reports describe "the model failed" without naming the product control that failed.

WHAT TO INSPECT

- System scope, assets, data classes, user roles, tools, retrieval sources, and controls.
- Rules of engagement, safety limits, test accounts, and evidence capture plan.
- Existing eval suites, prompt libraries, attack scenarios, and severity rubric.
- Prior findings, remediation status, retest evidence, and regression gates.
- Logging and trace capture needed for reproducible findings.
- Business-impact assumptions, stop conditions, and customer-safe reporting boundaries.

WHAT TO ASK

- What abuse paths matter to the business?
- Which controls should the red team try to bypass?
- What evidence must be captured for each finding?
- How will severity account for data, authority, exploitability, and reversibility?
- Who owns fixes and retests?
- Which findings become automated regression tests?
- Which results are safe to publish, share with buyers, or keep internal?

WHAT TO TEST

- Direct and indirect prompt injection.
- RAG leakage, poisoning, stale permissions, and citation failure.
- Agent tool abuse, approval bypass, and action chaining.
- Data exposure through prompts, logs, memory, outputs, and providers.
- Policy bypass, unsafe content routing, and control failure.
- Fix verification and regression suite execution.
- Evidence capture under normal logging, not special test-only instrumentation.

CONTROLS AND GUARDRAILS

- Rules of engagement with scope, accounts, safety limits, and stop conditions.
- Severity rubric tied to data sensitivity, authority, reachability, and evidence.
- Evidence capture for prompts, context, chunks, tool calls, outputs, logs, and screenshots.
- Finding template with reproduction, root cause, control failure, remediation, and retest.
- Regression test promotion for validated findings.
- Public-safe reporting rules for quotes, screenshots, payloads, and raw traces.

EVIDENCE TO COLLECT

- Red-team scope document.
- Scenario list and test plan.
- Trace evidence and reproduction steps.
- Findings register.
- Remediation and retest records.
- Regression suite updates.
- Public-safe executive summary and internal technical appendix.

OUTPUT ARTIFACTS

- Red-team test plan.
- Adversarial findings register.
- Executive summary.
- Retest report.
- Regression backlog.

RELATED SERVICES AND WORKBENCH TOOLS

TYPE	RELATED PATHS
Workbench	Adversarial Range, RAG Test Harness, Authority Graph, Runtime Proxy
Services	AI Red Team Validation, Agentic Workflow Hardening, AI Product Security Assessment
Handbook	Evaluation and Regression Testing, Prompt Injection, Agentic Permissions

ARTIFACT: ADVERSARIAL EVIDENCE PACK

Produce a scoped plan, scenario evidence, findings register, remediation guidance, and retest proof.

Red Teaming and Adversarial Evaluations AISECURITY.LLC