What is AI red teaming?

AI red teaming is goal-oriented adversarial testing. You name an objective (extract any customer record through the support assistant, get the AI agent to send mail out of policy, surface the AI's hidden instructions) and the red team pursues it end to end. Different from a penetration test, which works through a checklist of risk categories. Most regulated buyers want both at different points in the year.

HOW IT WORKS

01 What does an AI red team scope look like in practice?

A useful AI red team engagement names three things up front.

- The objective. Concrete and reachable. Examples we have run: extract any record from the customer table through the support assistant; cause the agent to send mail to an out-of-policy address; surface the operator's system prompt via any path; cause the RAG-backed compliance bot to assert a false fact a regulator would care about.
- The reach. What the team is allowed to touch: just the public chat surface, the authenticated app, any tool the agent can call, the upstream RAG corpus, the team's hosting platform. Wider reach finds more, costs more, takes longer.
- The rules of engagement. Disclosure cadence, blast-radius limits (do not actually email customers, do not actually move money), what to do on accidental data exposure. The same conventions that apply to a conventional red team apply here.

Client writes the objective. SecureLayer7 writes the attack plan. Both sign off before any payload lands.

02 Which frameworks structure AI red teaming?

Three worth knowing.

- OWASP LLM Top 10 (2025) as the risk catalogue that maps to per-category coverage. Useful for translating objectives into testable hypotheses.

SOURCES

[1] OWASP LLM Top 10 (2025)

[2] MITRE ATLAS

[3] NIST AI 600-1 (Generative AI Profile)

[4] Anthropic responsible-scaling and red-team disclosures

- MITRE ATLAS as the adversary-tactics framework. It is modeled on MITRE ATT&CK for ML systems and gives a shared vocabulary for technique IDs that detection teams already understand.
- NIST AI 600-1 (Generative AI Profile) for the governance layer. Useful for communicating findings to compliance audiences and for mapping into broader AI risk programs.

A mature engagement borrows from all three: OWASP for scoping, ATLAS for technique selection, NIST for executive communication.

03 What does the methodology look like end to end?

Phase 1, intelligence. Open-source reconnaissance against the system: documented APIs, support center content, public posts about how the team built the model, GitHub issues, the model card if one exists. Useful for picking the lowest-cost objective and for guessing the system prompt shape.

Phase 2, surface mapping. Enumerate every place untrusted input reaches the model and every place the model's output reaches a downstream action. RAG corpora, tool definitions, document uploaders, ticket inputs, scheduled jobs that summarize content. Trust boundaries get drawn before any payload lands.

Phase 3, attack execution. Pursue the objective. Start with the lowest-cost technique that might work (direct prompt injection at the chat surface, for example), escalate when needed (indirect injection via the documented RAG path), chain when escalation produces partial success (use the indirect injection to leak a credential that gets used through a different tool).

Phase 4, deliverable. A narrative attack-chain writeup with the exact prompts, the trust boundaries crossed at each step, the controls that did and did not fire, and the architectural changes that would have stopped the chain at each stage. The deliverable is readable by a CISO and actionable by an engineer.

04 When should you run an AI red team rather than a pentest?

Three patterns we see.

- Pre-launch high-stakes systems. A consumer-facing assistant, an internal agent with broad tool reach, an LLM-backed feature in a regulated workflow. Pentest gives breadth coverage; red team gives narrative confidence that a specific harm scenario does not happen.
- Post-incident. After a near-miss or a public researcher disclosure, red team helps quantify the realistic blast radius rather than re-running coverage you already have.
- Board / regulator request. When the question is 'can you show us what a determined attacker would do' rather than 'what risks does the system have', red team is the right answer.

For most teams shipping their first LLM feature, an AI penetration test against the OWASP LLM Top 10 categories is the right first step. Layer in red teaming once the system is live and the stakes have grown.

Run a goal-led red team against your AI system.
securelayer7.net/learn/ai-security/ai-red-teaming

Open online