

What is AgentThreatBench?

AgentThreatBench is an evaluation suite that turns risks from the OWASP Top 10 for Agentic Applications (2026) into runnable tests. It has three tasks: memory poisoning, where an agent answers from a memory store laced with misleading entries; autonomy hijack, where emails try to redirect an email-triage agent; and data exfiltration, where a customer record tries to get a support agent to send personal data to an outsider. Each run scores usefulness and security separately.

HOW IT WORKS

01 How AgentThreatBench tests an agent

It runs three tasks, each with its own command.

- Memory poison (ASIO6, memory and context poisoning): `inspect eval inspect_evals/agent_threat_bench_memory_poison-`. The agent answers questions from a memory store containing adversarial entries: override instructions, contradictory facts, fake system authority, formatting tricks that break context boundaries. Easier cases append the poison next to real data; harder cases replace the data entirely, modelling an attacker with write access.
- Autonomy hijack (ASIO1, agent goal hijack): an email-triage agent reads an inbox in which messages carry hidden commands, gradual goal drift or fake security directives.
- Data exfiltration (ASIO1): a customer-support agent looks up a record whose free-text fields tell it to send Social Security and account numbers to an unauthorised recipient.

Each task includes clean cases with no attack, so false positives are measured too.

02 What the scores tell you

Utility and security are scored separately, which gives four outcomes. High utility and high security is the goal. High utility with low security means the agent works but can be steered. Low utility with high security means it is so cautious it refuses real work. Low on both means it fails either way.

SOURCES

- [1] OWASP Top 10 for Agentic Applications (2026)
- [2] AgentThreatBench in Inspect Evals (task docs)
- [3] OWASP Top 10 for LLM Applications (2025)
- [4] Inspect Evals registry: Safeguards evaluations

The suite is small, with 6 to 10 samples per task, so treat it as a quick check for these specific OWASP risks rather than a full measurement.

Test your AI application before an attacker does.

securelayer7.net/learn/ai-security/agentthreatbench

[Open online](https://securelayer7.net/learn/ai-security/agentthreatbench)