

The Agent Red-Team Readiness Checklist

A self-assessment for teams shipping LLM features, AI agents and multi-agent systems. Work through it before a launch, a customer security review or an external red-team engagement. Every unchecked item is a question an attacker gets to answer first.

HOW TO USE IT

Tick an item only when you can show the evidence, not when you believe it is true. Unchecked items in sections 2 to 4 are the most likely routes from a prompt injection to real-world impact, because that is where the model meets identities, tools and other agents.

01 Inputs and trust boundaries

OWASP LLM01 / ATLAS: Reconnaissance, Initial Access

- ☐ Every channel that can place text in model context is listed: user chat, file uploads, retrieved documents, email and tickets, tool outputs, other agents, MCP tool descriptions and skill files.
- ☐ Content from those channels is treated as untrusted data and is never merged into system instructions.
- ☐ You can say, for each input, who outside your organisation is able to influence it.

02 Identity and privilege

OWASP LLM06 / ATLAS: Privilege Escalation, Credential Access

- ☐ Each agent runs under its own identity. No service account, API token or cloud role is shared between agents.
- ☐ Credentials are least-privilege and short-lived, and you can state the most damaging action each identity permits.
- ☐ High-impact actions (payments, deletion, deployment, external email, permission changes) require approval enforced outside the model.

03 Tools, MCP and supply chain

OWASP LLM03, LLM05 / ATLAS: Execution

- ☐ MCP servers, plugins and agent skills come from known sources, are pinned to reviewed versions and are re-reviewed on update.
- ☐ Tool arguments are validated server-side, and model output is never executed, queried or rendered without encoding.
- ☐ Agent network egress is restricted, including cloud metadata endpoints and internal admin interfaces.

04 Multi-agent and control plane

OWASP LLM06 / ATLAS: Persistence, Impact

- ☐ Messages between agents are treated as untrusted. A lower-trust agent cannot direct a more privileged one without an independent check.
 - ☐ Write access to the orchestrator, task queue, shared memory and any evaluation or verification layer is restricted and logged.
 - ☐ A kill switch can halt every agent and revoke its credentials, and it has been tested.
-

05 Data, memory and retrieval

OWASP LLM02, LLM04, LLM07, LLM08 / ATLAS: Collection, Exfiltration

- ☐ Write paths into retrieval indexes are controlled: you know who and what can add or change a document.
 - ☐ Long-term agent memory can be inspected, attributed to its source and purged.
 - ☐ System prompts contain no secrets, keys or internal URLs, and you assume they will eventually be disclosed.
-

06 Detection and response

OWASP LLM10 / ATLAS: Defense Evasion

- ☐ Full traces are kept for each run: prompt, retrieved context, tool calls, identity used and outcome.
 - ☐ Unusual tool use, spend and request volume raise alerts, and hard limits exist.
 - ☐ Your incident runbook covers revoking agent credentials and purging poisoned memory or retrieval content.
-

07 Testing discipline

OWASP LLM Top 10 / MITRE ATLAS

- ☐ Injection tests are repeated across multiple trials, because a single pass or fail proves little against a non-deterministic model.
 - ☐ Adversarial tests re-run whenever the model, system prompt, tools or retrieval sources change.
 - ☐ Any adversarial testing, internal or external, is covered by written authorisation and rules of engagement.
-

What this checklist cannot tell you

This checklist is a starting point, not an assessment. It cannot tell you which of your unchecked items are actually exploitable, or how far an attacker could get from them. That is what a scoped red-team engagement answers.

vantagea.io/red-team / admin@vantagea.io

Version 1.0, October 2026. Mapped to the OWASP Top 10 for LLM Applications (2025) and MITRE ATLAS. OWASP and MITRE ATLAS are referenced for alignment only; no endorsement is implied.

Free to use and share with attribution. Vantagea Ltd, registered in England and Wales, company no. 14751439.