rasotec

FREE CHECKLIST

# AI Agent and LLM Security Checklist

For chatbots, RAG assistants and AI agents with tool access, based on the OWASP Top 10 for LLM Applications (2025)

An LLM does not reliably distinguish between instructions and data. Every piece of text it processes (a user message, an uploaded document, an email, a web page or a tool result) can change its behavior. Use this checklist to review your AI application before release and after major changes. The references in brackets point to the matching OWASP category.

## 1 Inventory and architecture

- [ ] **All AI features documented:** models and providers, system prompts, data sources, connected tools, APIs and MCP servers.
- [ ] **Trust boundaries mapped:** which inputs come from users or external content, and which actions the model can trigger on whose behalf.
- [ ] **Models, libraries and MCP servers from trusted sources,** with pinned versions and a process for updates (LLM03 Supply Chain).

## 2 Prompt injection and system prompts

- [ ] **All content treated as untrusted:** user input, documents, emails, web pages and tool output can carry injected instructions (LLM01 Prompt Injection).
- [ ] **Security does not depend on the system prompt.** Access control and business rules are enforced in code, outside the model.
- [ ] **No secrets in the system prompt:** no credentials, internal URLs or information users must not see (LLM07 System Prompt Leakage).
- [ ] **Direct and indirect injection tested,** including jailbreaks and instructions hidden in documents or web pages the model reads.

## 3 Data, RAG and knowledge bases

- [ ] **Permissions enforced at retrieval:** users only retrieve documents they may see, and tenants are separated in the vector store (LLM08 Vector and Embedding Weaknesses).
- [ ] **Sensitive data minimized** in prompts, context, fine-tuning data and logs (LLM02 Sensitive Information Disclosure).
- [ ] **Ingestion pipeline protected** against poisoned or manipulated documents and training data (LLM04 Data and Model Poisoning).
- [ ] **Conversation history and logs** protected, access-controlled and deleted after a defined retention period.

## 4 Agents, tools and MCP integrations

- [ ] **Least privilege for every tool:** scoped API tokens, read-only where possible, no shared admin credentials (LLM06 Excessive Agency).

- [ ] **Human approval for consequential actions** such as payments, deleting data, sending emails or changing permissions.
- [ ] **Actions run with the user's permissions,** not with a privileged service account that any user can steer.
- [ ] **Tool results treated as untrusted input,** just like user messages.

## 5 Output handling

- [ ] **Model output encoded** before it is rendered in a browser, to prevent cross-site scripting (LLM05 Improper Output Handling).
- [ ] **Output never passed unchecked** to shells, SQL queries, code execution or internal HTTP requests (command injection, SQL injection, SSRF).
- [ ] **Links and images in output restricted,** so that data cannot be exfiltrated through attacker-controlled URLs.
- [ ] **Structured output validated** against a schema before it is processed.
- [ ] **Users informed** that answers can be wrong, and critical answers checked against sources (LLM09 Misinformation).

## 6 Infrastructure and abuse

- [ ] **Authentication and authorization** on every AI endpoint and API.
- [ ] **Rate limits, token limits and budget alerts** to prevent abuse and unexpectedly high costs (LLM10 Unbounded Consumption).
- [ ] **Provider API keys kept server-side** in a secrets manager, never in frontend code or mobile apps.
- [ ] **Logging and monitoring** that can detect injection attempts and unusual tool usage.
- [ ] **Security tests before release and after major changes,** following the OWASP Top 10 for LLM Applications and MITRE ATLAS.


### Test your AI application before attackers do

Our AI pentests cover prompt injection, data leakage, RAG pipelines, AI agents and MCP integrations. Details: **www.rasotec.com/tests/ai**

**Book a free 30-minute consultation** · **contact@rasotec.com** · **www.rasotec.com**