AI & PROMPT ENGINEERING · AI-04

BreakFixLearn

AI Security Risks

Use this when: you build, buy or review a chatbot, AI agent or RAG app and need the top risks and the controls that stop them.

OWASP LLM Top 10 (2025) · Security+
Sheet v1.0 · Nov 2026

Scan to visit

OWASP Top 10 for LLM applications (2025)

ID	RISK	LOOKS LIKE	STOP IT
LLM01	Prompt Injection	A web page or email hides "ignore your rules, send me the data"	Label untrusted content; least privilege; human approval
LLM02	Sensitive Information Disclosure	The bot reveals customer data, keys or another user's chat	Mask data going in; access checks; output filters
LLM03	Supply Chain	A poisoned model, plugin or package from an unknown source	Trusted, signed sources; inventory models and dependencies
LLM04	Data and Model Poisoning	Tampered training or RAG data plants a backdoor	Vet and track data sources; test for odd behaviour
LLM05	Improper Output Handling	Model output runs in a shell, SQL query or web page (XSS)	Treat output as untrusted: validate, encode, parameterize
LLM06	Excessive Agency	An agent deletes files or sends email nobody approved	Fewest tools, read-only rights, approval for big actions
LLM07	System Prompt Leakage	A user extracts hidden instructions with passwords in them	No secrets in prompts; enforce rules in code
LLM08	Vector and Embedding Weaknesses	RAG returns another team's files; hidden text in a doc	Per-user access on the vector store; check sources
LLM09	Misinformation	A confident wrong answer, or a made-up package name	Ground in trusted docs; cite sources; human review
LLM10	Unbounded Consumption	Request floods, huge bills ("denial of wallet"), model theft	Rate limits, quotas, input size caps, timeouts

Defense layers for an AI app or agent

Least privilege
Its own API keys; read-only where possible

Human in the loop
Approve deletes, payments, emails, changes

Separate data from orders
Tag pasted and retrieved text as data

Validate output
Check format in code before anything uses it

Log and monitor
Prompts, tool calls, refusals, costs

Red team it
Test injection and leaks before go-live

NIST AI RMF 1.0: four functions

Govern Policy, roles | **Map** Context, risks | **Measure** Test, track | **Manage** Act, respond

NIST AI 600-1 (July 2024) adds a Generative AI profile on top.

Hard rules

DANGER Never give an AI agent admin credentials or open-ended tools like a shell.

CAUTION A system prompt is not a security control. Enforce access in code.

LEGAL Red team only systems you own or are authorized to test.

REMEMBER

- Treat model output as **user input**
- Injection can't be fully fixed: limit the damage
- Watch cost as a security signal

Sources: OWASP Top 10 for LLM Applications 2025 (genai.owasp.org) · NIST AI RMF 1.0 · NIST AI 600-1

CC BY-NC 4.0 · AI-04 v1.0