OWASP LLM Top 10 (2025), explained.

OWASP (the Open Worldwide Application Security Project) is the non-profit behind the most-used security risk lists for software. The LLM Top 10 is their list for products built on large language models, the AI behind chatbots and assistants. The 2025 version sharpened the categories and added new ones for attacks on retrieval systems and attacks that drain resources. Most security teams use it as the first checklist for AI security work.

HOW IT WORKS

01 What are the ten risks in the 2025 list?

Each risk links to the OWASP detail page and to the SecureLayer7 deep-dive where one exists.

- LLM01: Prompt Injection: untrusted input overrides the operator's instructions. SL7 explainer.
- LLM02: Sensitive Information Disclosure: the model reveals secrets it can see or was trained on.
- LLM03: Supply Chain: tampered model weights, plugins, or training data.
- LLM04: Data and Model Poisoning: bad training data changes how the model behaves.
- LLM05: Improper Output Handling-: downstream systems trust the model's output without checking it.
- LLM06: Excessive Agency: an agent can take more actions than the job needs.
- LLM07: System Prompt Leakage: an attacker pulls out the operator's hidden instructions.
- LLM08: Vector and Embedding Weaknesses-: attacks on the RAG store and the embedding pipeline. SL7 RAG poisoning explainer.
- LLM09: Misinformation: the model states false things with confidence, and other systems act on them.
- LLM10: Unbounded Consumption-: resource-draining and model-theft attacks. SL7 model-extraction explainer.

02 What changed between the 2023 and 2025 lists?

SOURCES

[1] OWASP LLM Top 10 (2025)

[2] OWASP GenAI Security Project

[3] MITRE ATLAS

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

A few real shifts:

- Vector and Embedding Weaknesses (LLM08) is new. The 2023 list lumped these into prompt injection. 2025 splits them out, because the defenses and the attackers are different.
- System Prompt Leakage (LLM07) got its own entry. Pulling out the operator's hidden instructions was split off from prompt injection, because the damage (leaked IP and policy) is its own thing.
- Excessive Agency (LLM06) was rewritten for real agent products. The 2023 version was about functions the agent should not have. The 2025 version is about an agent that can take more actions than its job calls for.
- Unbounded Consumption (LLM10) replaces Model Denial of Service. Wider scope: model theft, resource drain, and runaway cost, not just downtime.
- Training Data Poisoning (LLM04) absorbed model poisoning, since both shape the same risk.

03 How does SecureLayer7 use the list when scoping an engagement?

As a coverage map, not a tick-box list. Every engagement starts with one question: which of these ten actually apply to your setup? A read-only assistant with no tools rarely needs LLM06 testing. A RAG pipeline that takes user uploads almost always needs LLM01, LLM05, and LLM08.

For each category in scope, we run a payload library against your exact configuration, then hand-build follow-up attacks wherever one half-works. The report gives per-category notes, what we tested, what we found, what we advise, so an auditor can see which risks were covered.

04 How does the OWASP LLM Top 10 relate to MITRE ATLAS and NIST AI 600-1?

Three lenses on the same ground.

- OWASP LLM Top 10 is the list for app developers. Best for setting scope and priorities.

- MITRE ATLAS is the attacker-tactics framework, built like MITRE ATT&CK but for machine-learning systems. Best for red-team planning and detection work.
- NIST AI 600-1 (Generative AI Profile) is the governance and risk framework, built for company-wide AI risk programs. Best for compliance and policy.

A strong AI security program uses all three. Most teams start with OWASP to scope the engineering, add ATLAS for offensive work, and use NIST 600-1 to brief executives and auditors.

Map your application to the OWASP LLM Top 10.
securelayer7.net/learn/ai-security/owasp-llm-top-10

Open online