

TIER

THE EIGHT TIERS · PART 7 · SECURITY AND THREAT TAXONOMIES · 8 MIN READ

7

OWASP LLM Top 10 (2026) and the AI Security Stack

Tier 7 of eight: where AI governance stops being paperwork, and the one ranking change that tells you where this is heading

3 Aug

2026 edition published

7,714

real-world incidents weighted in

6 → 3

Excessive Agency

3

legs of the lethal trifecta

The OWASP LLM Top 10 is the community-maintained ranking of the most critical security risks in large language model applications. The 2026 edition published on 3 August, and one movement in it tells you more about where AI risk is going than most of the research published this year.



Excessive Agency went from sixth place to third.

I want to spend a minute on why that matters, and then give you the rest of this tier — because this is the layer where governance stops being a document exercise and turns into something an engineer can actually test.

2026 EDITION

What changed in 2026, and why you should trust it more

The methodology changed, and that's as significant as the rankings.

This edition was the first weighted with real incident data — **7,714 real-world AI security incidents**, counted at 25% against a 75% community vote. Previous editions were expert opinion, which is fine but drifts toward whatever people are worried about rather than what's actually costing money.

25% incidents

75% community vote

The full 2026 list	
01	Prompt Injection
02	Sensitive Information Disclosure
03	Excessive Agency ↑ from 6
04	Supply Chain
05	Data and Model Poisoning
06	Unbounded Consumption
07	Misinformation
08	Hidden Context Exposure New
09	Vector and Embedding Weaknesses
10	Improper Output Handling

Prompt injection holds first place for a third consecutive edition, which at this point is less a finding than a standing indictment of the architecture.

Two movements are doing the real talking.

6 → 3 Excessive Agency, sixth to third. Loss is migrating from what models <i>say</i> to what agents <i>do</i>. A model that produces a wrong answer is a content problem. An agent with API credentials that takes a wrong action is an incident with a cost attached.	New → 8 Hidden Context Exposure, new at eighth. It replaces System Prompt Leakage, which was narrow — it only covered protecting the literal text of your system prompt. The new category widens the attack surface to retrieved documents, memory, user information, application state and tool responses.
--	--

Which is to say: everything we quietly stuff into the context window and don't think of as attack surface.

Read those two together and they're saying the same thing from different angles. The threat model moved while most governance programmes were still writing policies about model outputs.

SECOND LIST

The agentic companion

There's a second list. The **OWASP Top 10 for Agentic Applications** published on 9 December 2025, from the Agentic Security Initiative — over a hundred contributors, with a review board including people from NIST, Cisco, Microsoft and AWS.

Ten entries, ASI01 through ASI10, covering goal hijacking, tool misuse, privilege abuse, memory poisoning and rogue agents.

It extends the LLM list rather than replacing it, and each entry cross-references the corresponding LLM risks. The 2026 LLM edition also expanded its cross-references out to NIST, MITRE ATLAS and CWE — which, practically, makes both lists far easier to fold into a control library you already have.

OWASP LLM TOP 10

Governs what the model says.

OWASP AGENTIC TOP 10

Governs what the system does.

If you're running agents, you need both. The LLM list governs what the model says. The agentic list governs what the system does.

THE STACK

The rest of the stack

MITRE ATLAS

[v5.1.0](#) · [16 tactics](#) · [84 techniques](#)

MITRE ATLAS is an ATT&CK-style knowledge base of real adversary tactics and techniques against AI systems. Version 5.1.0, from November 2025, carries 16 tactics and 84 techniques with case studies drawn from actual incidents.

ATLAS is what you reach for when you need to show that your threat modelling is grounded in observed adversary behaviour rather than in a workshop where people imagined things. That distinction matters to auditors and it matters more to boards.

CSA AI Controls Matrix

[Cloud deployments](#)

The **Cloud Security Alliance AI Controls Matrix** does the same job for cloud deployments, built to interoperate with the Cloud Controls Matrix. If your estate is cloud-heavy and you already use the CCM, this is the path of least friction.

ISO/IEC 27090 & 27091

[Extends the 27001 family into AI](#)

ISO/IEC 27090 covers AI-specific security threats — data poisoning, model theft and extraction, membership inference, model inversion, evasion — with detection and mitigation guidance for each. It reached final publication stage in 2026. **ISO/IEC 27091** covers AI privacy protection; its DIS ballot closed in February 2026. Together they extend the 27001 family into AI, which is the easy road if your security programme already lives there.

Government & vendor guidance

[NCSC](#) · [CISA](#) · [Google](#) · [Databricks](#)

On the government side, the **UK NCSC and CISA Guidelines for Secure AI System Development** were endorsed by eighteen countries in November 2023 and cover secure design, development, deployment and operation. CISA has since added joint guidance on deploying AI systems securely and on AI data security. **Google's Secure AI Framework** and the **Databricks AI Security Framework** are the most substantial vendor contributions.

THE LETHAL TRIFECTA

The idea that beats all of it

The single most useful thing in this whole area isn't a framework. It's Simon Willison's formulation of the **lethal trifecta**.

An agent that combines three things is exploitable by construction:

1 access to private data	2 exposure to untrusted content	3 the ability to communicate externally
------------------------------------	---	---

Remove any one leg and the attack collapses.

That's it. That's the whole idea, and I'd argue it's worth more in an architecture review than a hundred pages of policy — because it turns a diffuse anxiety into a design constraint someone can actually check on a whiteboard in about ninety seconds.

If you take one thing from this piece into your next design review, take that.

THE TEST

Why this tier is the one that separates real programmes from theatre

Here's my honest test for whether an AI governance programme is real.

Look at it and ask whether anything in it came from this tier. If the answer is no — if it's all policies, registers, assessments and attestations, with nothing that produces a test result — then you've written a policy. You haven't built a control.

You've written a policy. You haven't built a control.

The EU AI Act's Article 15 requires accuracy, robustness and cybersecurity. That's an abstract obligation. This tier is how you actually satisfy it, and how you show someone you did.

FAQ

Things people ask me

What changed in the OWASP LLM Top 10 for 2026?

Excessive Agency moved from sixth to third, and Hidden Context Exposure came in new at eighth, replacing System Prompt Leakage. It was also the first edition weighted with real-world incident data — 7,714 incidents.

What is MITRE ATLAS?

A knowledge base of real adversary tactics and techniques against AI systems, modelled on MITRE ATT&CK. Version 5.1.0 carries 16 tactics and 84 techniques.

Are AI security frameworks legally required?

Not directly. But they're the practical means of satisfying requirements that are — EU AI Act Article 15 on accuracy, robustness and cybersecurity being the obvious one.

NEXT IN THE SERIES · PART 8

OECD AI Principles

Non-binding, and the source of the definition that decides who the EU AI Act applies to.

The Eight Tiers series — Part 0: The eight kinds · Part 1: Comprehensive statutes · Part 2: The treaty nobody is watching · Part 3: Where the obligations actually are · Part 4: The regulators who were already there · Part 5: ISO 42001 · Part 6: NIST AI RMF · **Part 7: Security taxonomies** · Part 8: Principles · Part 9: Putting it together