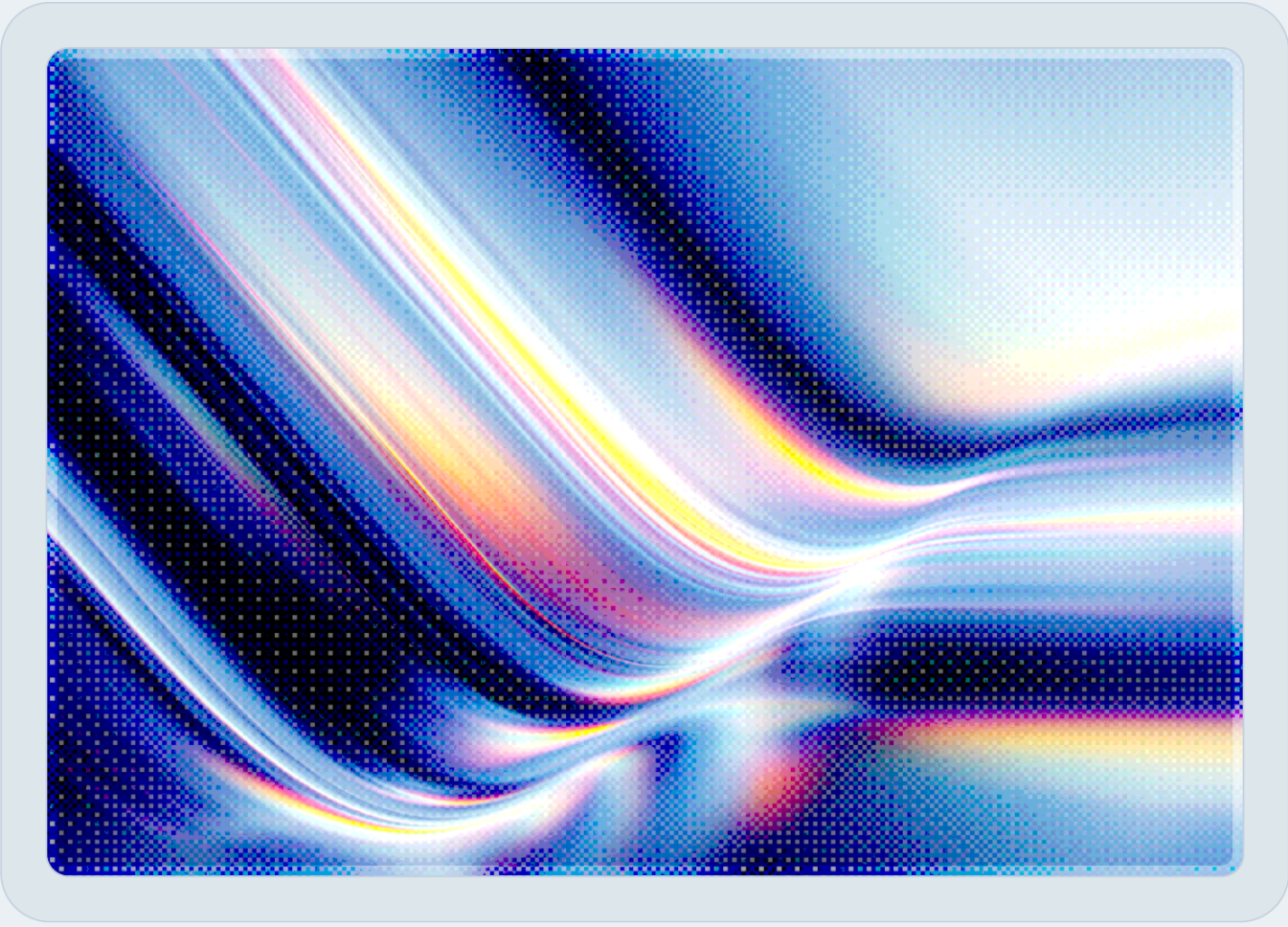
CSAI Foundation | Cloud Security Alliance

OWASP's 2026 LLM Top 10: Incident Data Meets Judgment

How 7,714 Real-World Incidents Reshaped the Industry's Benchmark AI Risk List

2026-08-15

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

The OWASP GenAI Security Project published the 2026 edition of its Top 10 for LLM Applications on August 3, 2026, and for the first time grounded the ranking in empirical incident data rather than practitioner survey results alone [1][2]. The project team weighted community voting at roughly 75% and incident evidence at 25%, drawing on a corpus of 7,714 reported AI-related security incidents, 6,639 of which carried enough detail to be classified against the risk taxonomy [3][4]. Prompt Injection and Sensitive Information Disclosure held the top two positions for a third consecutive year, but Excessive Agency jumped from sixth to third and Unbounded Consumption climbed from tenth to sixth, while Improper Output Handling fell from fifth to tenth — the largest single-category swing on the list [4][5]. One notable pattern in the results is where practitioner intuition and incident evidence diverged: prompt injection ranked far lower in raw incident counts than in expert opinion, while excessive agency and misinformation ranked higher in incident data than practitioners expected. OWASP's project leadership has said that divergence taught the team more than the areas of agreement did [2][5].

Background

The OWASP Top 10 for LLM Applications has been produced since its first release in 2023 through a working group of practitioners, researchers, and vendors who nominate, debate, and vote on candidate risks, and the resulting list has been referenced widely enough to appear in vendor red-teaming tools, compliance mappings, and enterprise AppSec checklists [1]. Through the 2024 and 2025 editions, that process remained fundamentally a consensus exercise: expert judgment about which failure modes mattered most, informed by individual experience but not systematically checked against any shared dataset of what was actually happening in production [7].

The 2026 edition breaks from that pattern. According to the project's public release materials, the team assembled a dataset of 7,714 AI-related security incidents pulled from public vulnerability databases and AI-harm repositories, then filtered it down to 6,639 incidents with sufficient detail to be classified against the ten risk categories [3][4]. Rather than replacing expert judgment with raw incident frequency, the project applied a deliberately conservative blend: practitioner voting retained roughly 75% of the weight in the final ranking, with incident data contributing the remaining 25% [4][5]. Project co-chair Steve Wilson described the intent behind that ratio as protection against a dataset artifact overturning years of collective professional judgment, noting that the team "learned more from the

disagreements than from where they agreed" [5][6]. That framing matters for how the list should be read: it is not a pure incident-frequency ranking, and categories with genuinely low incident counts, most notably Prompt Injection, can still rank highly because the vote weighting keeps expert consensus dominant.

The scope of several categories was also revised for 2026 independent of the ranking methodology. Prompt Injection's definition expanded to explicitly cover cross-modal attacks that smuggle instructions inside images or audio, alongside memory persistence and the wider blast radius created by agentic integrations [4]. System Prompt Leakage was renamed and broadened into Hidden Context Exposure, extending its scope beyond system prompts to cover policy and configuration exfiltration more generally [6]. Supply Chain was likewise reframed to fold in AI-specific concerns such as compromised model weights and unsafe serialization formats alongside conventional dependency risk [1][3]. The GenAI Security Project also maps the 2026 categories to external frameworks including NIST, MITRE ATLAS, and CWE, and cross-references the related OWASP guidance for agentic applications, positioning the LLM Top 10 as one node in a broader OWASP GenAI security taxonomy rather than a standalone document [1].

Security Analysis

The ranking and what moved

The table below compares each category's 2025 position to its 2026 position, drawing on the rank order that the project's release commentary and the most detailed independent security press coverage of the update corroborate in full [4][5][6][7].

Rank 2026	Category (2026)	Rank 2025	Movement
1	Prompt Injection	1	Unchanged (3rd consecutive year at #1)
2	Sensitive Information Disclosure	2	Unchanged
3	Excessive Agency	6	Up 3
4	Supply Chain	3	Down 1

Rank 2026	Category (2026)	Rank 2025	Movement
5	Data and Model Poisoning	4	Down 1
6	Unbounded Consumption	10	Up 4
7	Misinformation	9	Up 2
8	Hidden Context Exposure (formerly System Prompt Leakage)	7	Down 1
9	Vector and Embedding Weaknesses	8	Down 1
10	Improper Output Handling	5	Down 5 (largest fall)

Two movements stand out as clear signals of what incident data revealed that voting alone had not. Excessive Agency's rise from sixth to third reflects, in Wilson's view, that AI systems now routinely browse the web, call external tools, and touch business systems with far broader permissions than earlier chat-only deployments carried, so unchecked autonomy is producing documented harm rather than remaining a theoretical concern [5][6]. Unbounded Consumption's climb from tenth to sixth points to a related dynamic that this note assesses as agents increasingly chain tool calls and sub-agent invocations: uncontrolled token and resource consumption is becoming an operational and cost-security issue rather than remaining a narrow denial-of-service edge case.

Misinformation's move from ninth to seventh is smaller in absolute rank, but the underlying reasoning behind it connects to the same dynamic driving Excessive Agency's rise. Voters ranked misinformation low, largely because hallucination has historically been treated as a quality problem rather than a security one. Incident data told a different story: as OWASP's release materials put it, model outputs increasingly "drive tool calls, generate code, infer system state, authorize actions, and coordinate across agents," so a wrong answer several steps removed from a human reviewer can become a wrong action with real consequences [2]. This note reads that reasoning — a hallucination becoming an unreviewed downstream action inside an agent pipeline — as the same failure pattern driving Excessive Agency's rise, which suggests the two categories are worth evaluating together rather than as fully independent risks.

Improper Output Handling's fall from fifth to tenth, the largest movement on the list in either direction, does not mean unsanitized model output has become less dangerous. OWASP's commentary frames it instead as relative movement: other categories generated more incident volume and rose past it, while output handling failures, though still real, generated comparatively fewer classifiable incidents in the

dataset the team assembled [4]. Practitioners should not treat output handling's lower rank as license to deprioritize input/output validation controls; unsanitized code generation leading to cross-site scripting or remote code execution remains a documented failure mode, just one that ranked lower relative to a field of categories that grew faster.

Why Prompt Injection stayed first despite thin incident counts

The most counterintuitive result in the 2026 methodology is that Prompt Injection retained the top position even though, by one account, incident data alone would have placed it much further down the working group's broader candidate pool of risk categories considered before the list was narrowed to ten — as low as twelfth in that wider pool [6]. OWASP's explanation is a "defense effect": mature security teams that have invested in prompt-injection defenses catch and contain attempts before they escalate into reportable incidents, so the technique's true prevalence is undercounted in any dataset built from disclosed incidents and harms [4][6]. At the same time, the category's attack surface keeps expanding — cross-modal injection through images and audio, persistent agent memory, and the growing number of agentic integration points all add untrusted input channels — which is why the practitioner vote, weighted at 75%, kept prompt injection at the top of the list [4].

Wilson's framing of prompt injection as being unlike SQL injection shows how security teams should plan around it: rather than a vulnerability class organizations can expect to eliminate through a single fix, he described it as something closer to "death and taxes" — a condition to be continuously managed [5]. The project's broader public message reinforces the same posture shift: rather than trying to build a model that cannot be fooled, OWASP argues teams should build the system around the model so that when it is fooled, "nothing important breaks" [2]. That is a call to invest in blast-radius containment — scoped permissions, human-in-the-loop checkpoints for consequential actions, and monitoring for anomalous tool invocation — rather than treating prompt-injection defense as a solved problem once basic input filtering is in place.

Recommendations

Immediate Actions

Security teams operating LLM-powered applications should re-baseline their internal risk register against the 2026 category order rather than assuming the 2025 priorities still hold, paying particular attention to Excessive Agency and Unbounded Consumption given their sharp climbs. Any agent deployment that grants tool-calling, code-execution, or external-system access without a documented

review of permission scope should be flagged for prioritized remediation, since this is precisely the failure pattern driving Excessive Agency's rise to third place. Teams should also confirm that output-handling controls have not been quietly deprioritized on the mistaken assumption that its lower 2026 rank reflects reduced risk rather than relative movement within the list.

Short-Term Mitigations

Organizations should extend existing prompt-injection defenses to cover the newly broadened scope of the category, including image- and audio-borne instructions and injection vectors introduced through persistent agent memory, rather than assuming text-only filtering remains sufficient. Rate limiting and resource-consumption monitoring warrant renewed attention given Unbounded Consumption's climb, particularly in multi-agent architectures where a single user request can fan out into many downstream tool calls or sub-agent invocations. Given the misinformation-to-agency failure pattern OWASP highlighted, teams should also review whether any agentic workflow allows a model-generated output to trigger a consequential action — a purchase, a code deployment, an account change — without a human or independently verified checkpoint in between.

Strategic Considerations

The shift to incident-weighted methodology signals that OWASP intends future editions to keep incorporating empirical evidence alongside practitioner judgment; this note assesses that organizations should therefore plan for continued, and potentially larger, ranking volatility in coming years as the incident dataset matures and grows. Enterprises building internal AI risk taxonomies or vendor assessment criteria should treat the OWASP LLM Top 10 as a living, evidence-informed baseline rather than a static checklist, and should map it explicitly into whatever AI governance control framework they already operate, since the categories themselves describe failure modes rather than prescribe controls. Security and product leadership should also note the project's own caution about the "defense effect": a category's incident count reflects both its danger and how well-defended it already is, so absence of incidents is not evidence of absence of risk.

CSA Resource Alignment

For organizations implementing controls against the specific risk categories in the 2026 list, CSA's "AICM v1.1 Implementation Guidelines for Application Providers" [8] provides control-by-control guidance across domains including Application & Interface Security and Threat and Vulnerability Management, along with controls addressing model-specific risk, that map closely onto Prompt

Injection, Sensitive Information Disclosure, Insecure Output Handling, and Data and Model Poisoning. The guide explicitly references the OWASP LLM Top 10 as a prerequisite framework and covers input/output validation, guardrail design, and prompt differentiation controls that correspond directly to several of the categories discussed in this note. Both implementation documents sit atop the broader "AI Controls Matrix (AICM) v1.1" [9], which security and compliance teams should treat as the control catalog to consult when translating any individual OWASP LLM Top 10 category into auditable organizational requirements; AICM builds on CSA's earlier Cloud Controls Matrix and is the current framework CSA points organizations toward for AI-specific control mapping. For the agentic-autonomy dimension specifically — the thread connecting Excessive Agency, Misinformation, and Unbounded Consumption — CSA's MAESTRO agentic AI threat modeling framework [10] offers a layered methodology for identifying where autonomous decision-making and tool access introduce risk beyond what a single-turn LLM application would face.

References

[1] OWASP GenAI Security Project. "OWASP GenAI LLM Top 10 2026." OWASP, August 3, 2026.

[2] Help Net Security. "OWASP 2026 LLM Top 10: "The model will be fooled"." Help Net Security, August 6, 2026.

[3] Cyber Security News. "OWASP Releases GenAI LLM Top 10 2026 for Building and Securing Modern AI Apps." Cyber Security News, August 2026.

[4] Invicti. "OWASP LLM Top 10 2026: 7,714 Incidents Analyzed – And the #1 Risk Almost Didn't Make It." Invicti, August 2026.

[5] SD Times. "Prompt Injection tops 2026 OWASP GenAI / LLM Top Ten vulnerabilities." SD Times, August 2026.

[6] Rock Cyber Musings. "The 2026 OWASP LLM Top 10 Landed. Its #1 Risk Nearly Didn't Make the Cut.." Rock Cyber Musings, August 2026.

[7] OWASP GenAI Security Project. "OWASP Top 10 for LLM Applications 2025." OWASP, 2025.

[8] Cloud Security Alliance. "AICM v1.1 Implementation Guidelines for Application Providers (AP)." Cloud Security Alliance, 2026.

[9] Cloud Security Alliance. "AI Controls Matrix (AICM) v1.1." Cloud Security Alliance, 2026.

[10] Cloud Security Alliance. "Agentic AI Threat Modeling Framework: MAESTRO." Cloud Security Alliance, February 6, 2025.

