AI GOVERNANCE RESOURCE

AI Agent Audit Program

External Threat Taxonomies

A reference guide to the OWASP Top 10 for LLM Applications and OWASP Top 10 for Agentic Applications, anchored to ISO/IEC 42001, NIST AI RMF 1.0, and the EU AI Act.

This document is part of AI Agent Audit Program series. Use alongside AI Agent Audit Program & Risk Checklist and Companion Guide for CAEs.

Prepared by: Ishan Jain | ishan@jainishan.com | jainishan.com

Version 1.0 | June 2026

1. Introduction

1.1 Intro

OWASP — the Open Worldwide Application Security Project — is a nonprofit foundation that has published vendor-neutral, practitioner-written security guidance for over two decades. It is best known for its “Top 10” lists: concise, ranked summaries of the most critical and most common weaknesses in a given technology, built from real-world incident data and the contributions of hundreds of security practitioners and AI vendors worldwide.

The original OWASP Top 10 for Web Application Security (first published 2003, most recently updated 2021) has been the de facto baseline for web security audits for two decades. OWASP has since extended the same model to the technologies this document addresses: the OWASP Top 10 for Large Language Model Applications (2025) and the OWASP Top 10 for Agentic Applications (2026), both reproduced in full on the following pages.

1.2 Why It Matters

ISO/IEC 42001, the NIST AI RMF, and the EU AI Act establish what a trustworthy AI management system must achieve — governance, risk management, human oversight, transparency. None of them specify exactly how an LLM or an AI agent is actually compromised in practice. OWASP closes that gap: it is the closest thing the industry has to a shared technical vocabulary for AI/agent failure modes, revised on a cycle fast enough to reflect attack techniques that postdate any of the three governance frameworks above. Regulators, cyber-insurance underwriters, and external assessors are increasingly treating OWASP Top 10 coverage as an implicit baseline — an organization that cannot show it tested for prompt injection or excessive agency will struggle to credibly claim it has managed AI risk, however mature its governance documentation looks on paper.

1.3 Why a CAE Should Care

For Internal Audit specifically, OWASP turns a principles-based control objective — “ensure human oversight,” “manage cybersecurity risk” — into a specific, testable failure mode an auditor can examine directly. Two practical implications follow:

- **Common language:** Testing “ASI02: Tool Misuse” or “LLM06: Excessive Agency” with an engineering team lands very differently than discussing “excessive agency” in the abstract — it gives Internal Audit and the build team a shared, specific reference point.
- **A more defensible opinion:** “We tested against the industry-recognized OWASP Top 10 for LLM and Agentic Applications, mapped to ISO 42001, NIST AI RMF, and the EU AI Act” is a materially stronger statement to an Audit Committee or regulator than “we performed an AI risk assessment.”

The tables that follow map each OWASP risk, item by item, to the audit/governance question Internal Audit should ask and the governing clause(s) in ISO 42001, NIST AI RMF, and the EU AI Act.

Guidance, Not Law

OWASP Top 10 lists are industry consensus guidance, not law or regulatory text — but they are increasingly treated as a de facto technical baseline by regulators, cyber-insurance underwriters, and external assessors. The Agentic Applications list is new for 2026 and will move faster than the more established LLM Applications list (2025) as agentic deployment patterns mature. Revisit this mapping at least annually, and immediately upon any new major-version release from OWASP.

2. OWASP Top 10 for Agentic Applications (2026)

Standards basis: maps to ISO 42001 Annex A.6, A.8; NIST AI RMF MAP/MEASURE/MANAGE functions; EU AI Act Art. 14, 15 — see Sections 5-7 of the AI Agent Audit Program & Risk Checklist.

ID	Risk	What It Means in Practice	Audit / Governance Question	Standards Mapping
ASI01	**Agent Goal Hijack**	Attacker-controlled input (direct, or hidden in a document/tool response) redirects the agent's objective toward an attacker's goal instead of the operator's.	What stops a deviated plan before an irreversible action executes?	ISO 42001 Annex A.6, A.8 | NIST MEASURE 2.7, MAP 5 | EU AI Act Art. 14, 15
ASI02	**Tool Misuse**	The agent invokes a legitimate tool (file write, API call, code exec) with parameters or sequencing outside its authorized use, causing harm.	Is every tool scoped, parameter-validated, and gated for irreversible actions?	ISO 42001 Annex A.6, A.8 | NIST GOVERN 1, MANAGE 1 | EU AI Act Art. 14, 15
ASI03	**Identity & Privilege Abuse**	The agent's credentials or delegated permissions are reused, escalated, or impersonated beyond the original user's intent.	Is the delegation chain auditable? Are tokens scoped and short-lived?	ISO 42001 Annex A.6 (roles) | NIST GOVERN 1, MANAGE 1 | EU AI Act Art. 14, 15
ASI04	**Agentic Supply Chain Vulnerabilities**	A third-party framework, tool, MCP server, registry, or pre-built agent skill carries an exposure the application inherits at runtime.	Is every tool/plugin treated as a vetted, pinned, re-reviewed dependency?	ISO 42001 Annex A.10 | NIST GOVERN 6, MANAGE 3 | EU AI Act Art. 25, 51-56
ASI05	**Unexpected Code Execution**	The agent's sandbox or code-execution tool escapes its intended boundary and runs arbitrary code on the host or another tenant.	Is code execution isolated, least-privilege, and reviewed as a privileged-access feature?	ISO 42001 Annex A.6, A.8 | NIST MEASURE 2.7, MANAGE 2 | EU AI Act Art. 15
ASI06	**Memory & Context Poisoning**	Persistent memory, retrieval stores, or session context the agent relies on is shaped by an adversary, corrupting later decisions.	Is there provenance metadata, tenancy separation, and forgetting windows on agent memory?	ISO 42001 Annex A.7 | NIST MAP 4, MEASURE 2.7/2.10 | EU AI Act Art. 10, 15
ASI07	**Insecure Inter-Agent Communication**	Messages between agents are unauthenticated, unencrypted, or unverified, enabling impersonation, replay, or injection.	Are inter-agent messages signed, replay-protected, and policy-restricted?	ISO 42001 Annex A.8 | NIST MEASURE 2.7, MANAGE 2 | EU AI Act Art. 15
ASI08	**Cascading Failures**	An error or compromise in one agent propagates through downstream agents/tools faster than operators can detect or interrupt it.	Are there rate limits, circuit breakers, and a defined blast-radius cap per agent?	ISO 42001 Annex A.6, A.9 | NIST MEASURE 2.5-2.7 | EU AI Act Art. 15
ASI09	**Human-Agent Trust Exploitation**	Confident or impersonating agent output induces a human to approve, pay, disclose, or automate something they should not.	Is AI involvement disclosed, and is there friction on irreversible human approvals?	ISO 42001 Annex A.9 | NIST MEASURE 2.6, 2.8, 2.9 | EU AI Act Art. 50, 14
ASI10	**Rogue Agents**	An agent operates outside policy — via design failure, behavioral drift, or compromise — acting as an internal threat.	Can an unhealthy agent be detected and disabled in minutes, not days?	ISO 42001 Annex A.6, A.10 | NIST GOVERN 1, MEASURE 2.7 | EU AI Act Art. 14, 15

3. OWASP Top 10 for LLM Applications (2025)

Standards basis: maps to ISO 42001 Annex A.6, A.7, A.8; NIST AI 600-1 (GenAI Profile); EU AI Act Art. 10, 15 — see Sections 6 and 8 of the AI Agent Audit Program & Risk Checklist.

ID	Risk	What It Means in Practice	Audit / Governance Question	Standards Mapping
LLM01	Prompt Injection	Crafted input — typed directly or hidden in a retrieved document, email, or webpage — causes the model to follow an attacker's instructions instead of the application's intended rules.	Is there structural separation between system instructions and untrusted content?	ISO 42001 Annex A.8 | NIST AI 600-1 §2.9 | EU AI Act Art. 15
LLM02	Sensitive Information Disclosure	The model reveals confidential data — PII, secrets, proprietary content — that was present in its training data, retrieved context, or prompt history.	Is sensitive data masked/filtered before reaching the model, and scrubbed from logs?	ISO 42001 Annex A.7 | NIST MAP 2.2 | EU AI Act Art. 10; GDPR Art. 5, 32
LLM03	Supply Chain	Vulnerabilities introduced via compromised pre-trained models, datasets, fine-tuning adapters, or third-party plugins/libraries used to build the application.	Is model/dataset/plugin provenance verified and tracked through a vendor due-diligence process?	ISO 42001 Annex A.10 | NIST GOVERN 6 | EU AI Act Art. 25, 51-56
LLM04	Data and Model Poisoning	Malicious or low-quality data introduced during pre-training, fine-tuning, or embedding creation skews model behavior or creates hidden backdoors.	Is training/fine-tuning data lineage documented and tested for anomalies before use?	ISO 42001 Annex A.7 | NIST MAP 2.3, MEASURE 2.7 | EU AI Act Art. 10
LLM05	Improper Output Handling	Model output is passed downstream (rendered, executed, or fed to another system) without validation, enabling injection, XSS, or unauthorized command execution.	Is every output validated/sanitized before being executed, rendered, or trusted?	ISO 42001 Annex A.6 | NIST MANAGE 1.4 | EU AI Act Art. 15
LLM06	Excessive Agency	The model/agent is granted more permissions, tools, or autonomy than its task requires, enabling unintended or harmful actions when manipulated or in error.	Are permissions and tool access scoped to least-privilege per task?	ISO 42001 Annex A.6, A.8 | NIST MANAGE 1.3 | EU AI Act Art. 14
LLM07	System Prompt Leakage	Internal system instructions, business logic, or embedded secrets are extracted from the model, exposing how the application is designed to be controlled or bypassed.	Are secrets/business-logic kept out of the system prompt itself, with leakage tested?	ISO 42001 Annex A.8 | NIST MEASURE 2.7 | EU AI Act Art. 15
LLM08	Vector and Embedding Weaknesses	RAG/embedding pipelines are exposed to vector-store poisoning, cross-tenant data leakage via shared embeddings, or manipulation of similarity search results.	Are vector stores access-controlled per tenant and monitored for injected content?	ISO 42001 Annex A.7 | NIST MAP 4, MEASURE 2.10 | EU AI Act Art. 10
LLM09	Misinformation	The model produces confident, plausible, but factually incorrect output (hallucination, fabricated citations) that is trusted and acted on without verification.	Is there a human-review or fact-check gate before high-stakes output is used?	ISO 42001 Annex A.9 | NIST MEASURE 2.5, 2.6 | EU AI Act Art. 14, 13
LLM10	Unbounded Consumption	Excessively long prompts, recursive tool/agent calls, or automated request flooding drive uncontrolled cost, latency, or denial-of-service against the application.	Are there rate limits, token quotas, and cost-anomaly alerts per user/agent?	ISO 42001 Annex A.6, A.9 | NIST MEASURE 2.7 | EU AI Act Art. 15