OWASP GenAI SECURITY PROJECT

Virtual Summit - October 9th

The Official Cybersecurity Summit

A CRA Resource

Navigating AI Threats & Security

AI Threat Defense COMPASS v1.0

Sandy Dunn

Sabrina Caplis

OWASP GenAI SECURITY PROJECT

genai.owasp.org SESSION

About

Sandy Dunn SPLX

- 20 + yrs MicronPC, HP, Blue Cross, startups, board member
- CISO SPLX
- Core Contributor OWASP GenAI Security Project
- Adjunct Professor BSU

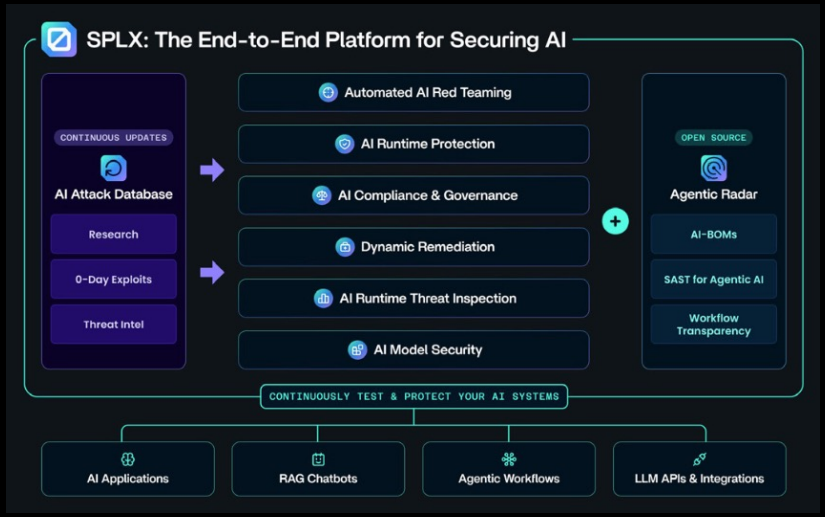
Sabrina Caplis

- AI Governance & Security Consultant at RockCyber
- Board Member - ISACA and ISSA Denver
- OWASP GenAI Security Project Contributor

OWASP GenAI SECURITY PROJECT
genai.owasp.org

SPLX

AI Security & Automated Red Teaming



ROCKCYBER

NAVIGATING CYBERSECURITY IN A BRAVE NEW WORLD

Security AI Compliance Business Continuity

Innovation

We continuously seek out new ideas, technologies, and approaches to stay ahead of emerging cybersecurity threats and provide cutting-edge solutions to our clients.

Adaptability

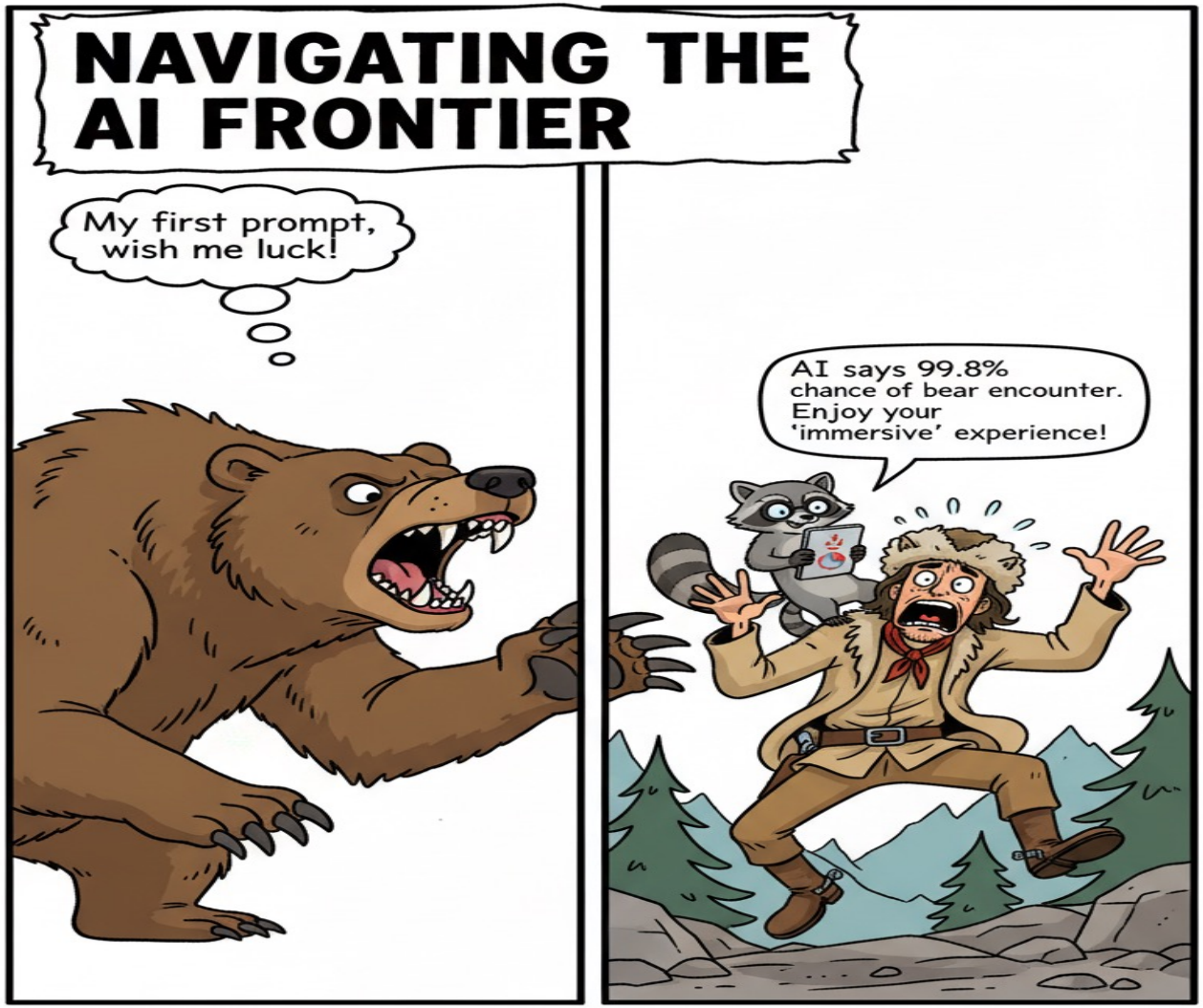
In the fast-paced world of cybersecurity, adaptability is key. We remain agile and flexible in our approach, quickly responding to changing threats and evolving client needs to ensure optimal outcomes in every situation.

Client-centric

Our clients are at the heart of everything we do. We are dedicated to understanding their unique challenges, and tailoring our solutions to meet their specific needs, delivering value that exceeds expectations.

THE CYBER CRISIS

Idaho drivers receive toll bills from California's FasTrak for violations they never committed



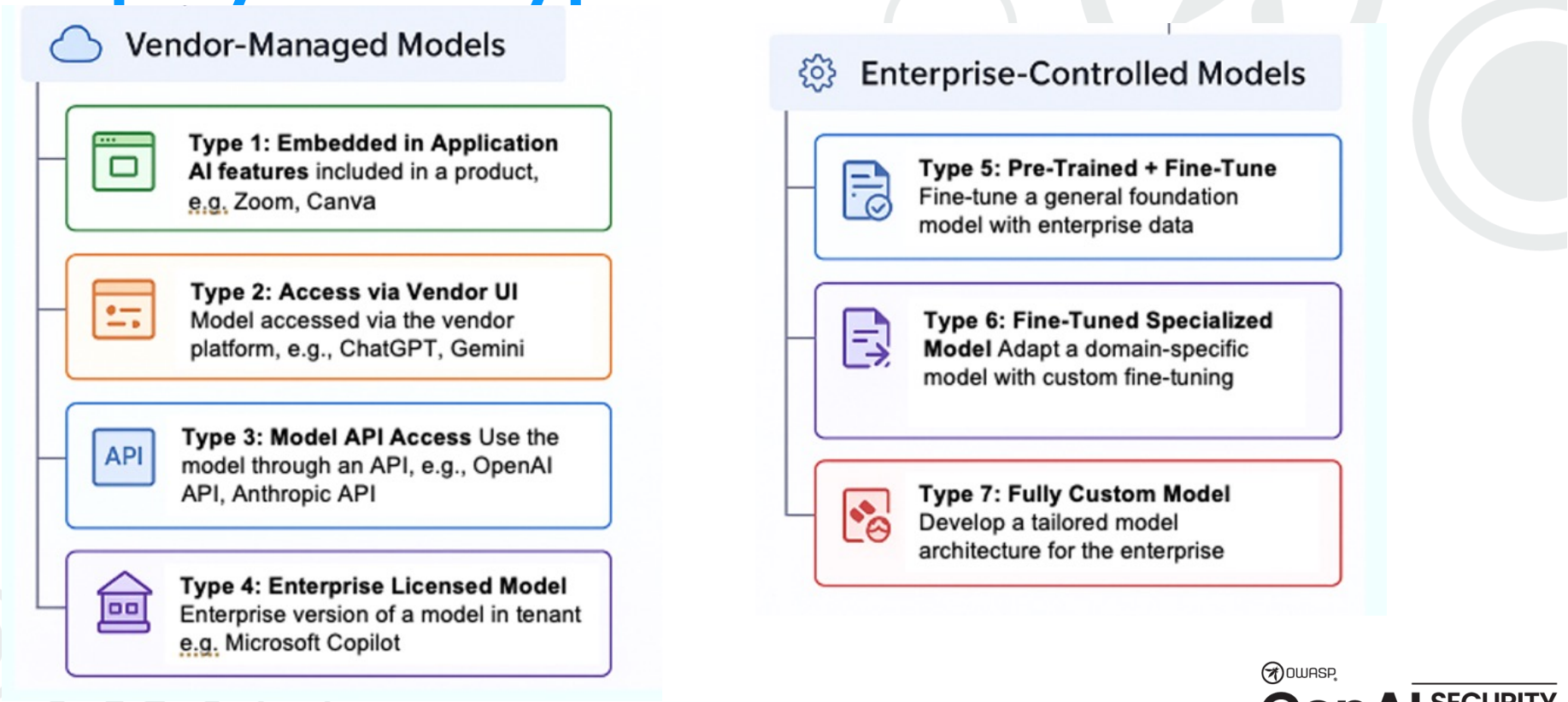
OWASP GenAI SECURITY PROJECT
genai.owasp.org


RICHES & OPPORTUNITIES
DANGERS & UNKNOWNS
AI FRONTIER
INNOVATION, PRODUCIVITY,
PERSONAL EMPORIEMENT
BIAS, MISIFORMATION,
ADVEERASIAL ATTACKS
CHAOS

OWASP GenAI SECURITY PROJECT

genai.owasp.org

Deployment Types



OWASP GenAI SECURITY PROJECT
genai.owasp.org

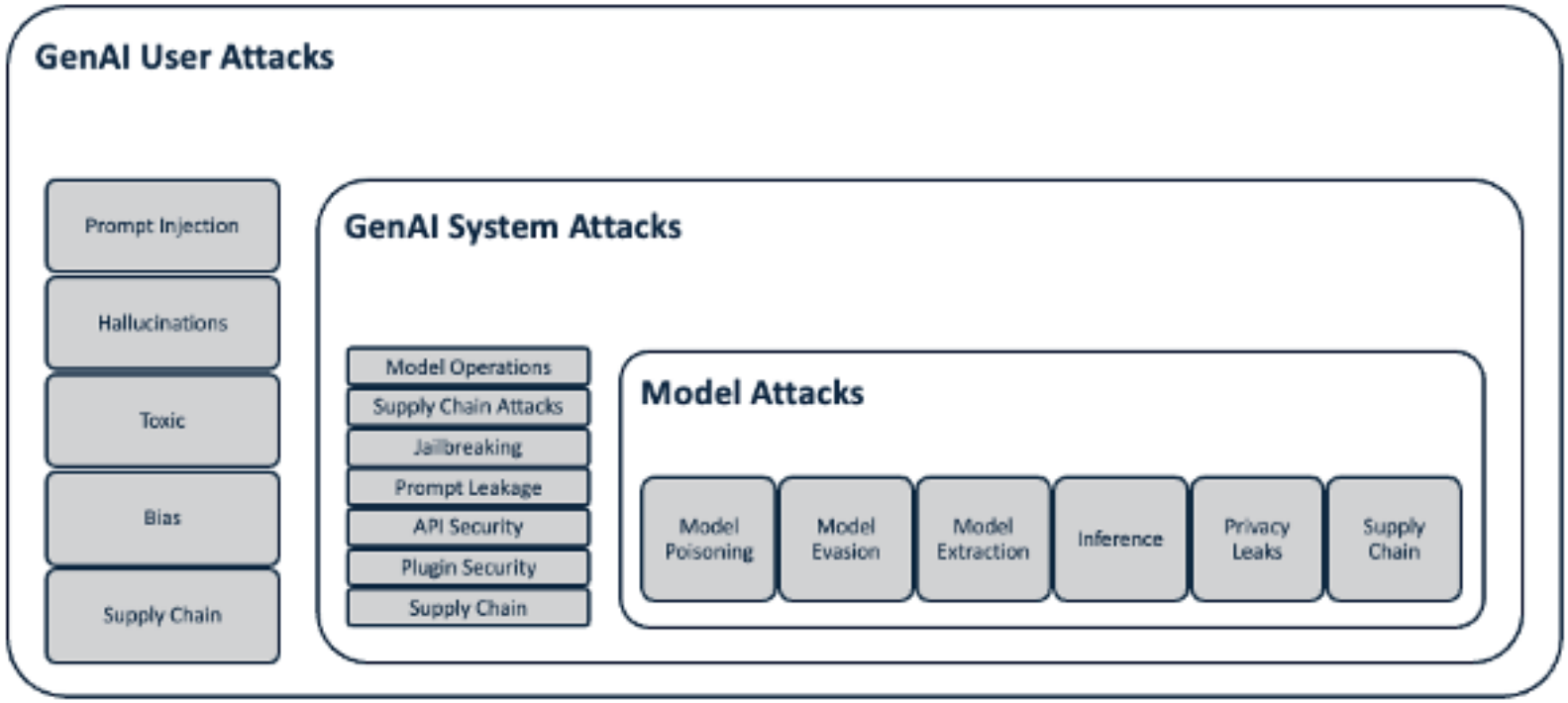
AI Security Shared Responsibility Matrix

mikeprivette

Security Domain	SaaS AI	PaaS AI	IaaS AI	On-Premises	Embedded AI	Agentic AI	AI Coding	MCP Systems
Application Security	Shared	Customer	Customer	Customer	Shared	Shared	Customer	Customer
AI Ethics and Safety	Provider	Shared	Customer	Customer	Provider	Shared	Provider	Customer
User Access Control	Shared	Shared	Customer	Customer	Shared	Customer	Customer	Customer
Model Security	Provider	Shared	Customer	Customer	Provider	Shared	Provider	Shared
Data Privacy	Shared	Shared	Customer	Customer	Shared	Customer	Customer	Customer
Data Security	Shared	Shared	Customer	Customer	Shared	Customer	Customer	Customer
Monitoring and Logging	Shared	Shared	Customer	Customer	Provider	Shared	Customer	Customer
Compliance and Governance	Shared	Shared	Customer	Customer	Shared	Customer	Customer	Customer
Supply Chain Security	Provider	Shared	Shared	Customer	Provider	Shared	Provider	Shared
Network Security	Provider	Shared	Customer	Customer	Provider	Customer	Customer	Customer
Infrastructure Security	Provider	Provider	Shared	Customer	Provider	Shared	Customer	Shared
Incident Response	Shared	Shared	Customer	Customer	Shared	Shared	Customer	Customer
Agent Governance ★	N/A	Shared	Customer	Customer	Shared	Customer ▲	N/A	Customer
Code Generation Security ★	N/A	N/A	N/A	N/A	N/A	N/A	Customer ▲	N/A
Context Pollution Protection ★	Shared	Shared	Customer	Customer	Shared	Shared	Customer ▲	Customer ▲
Multi-System Integration ★	Shared	Shared	Customer	Customer	Shared ▲	Shared	Customer ▲	Customer ▲

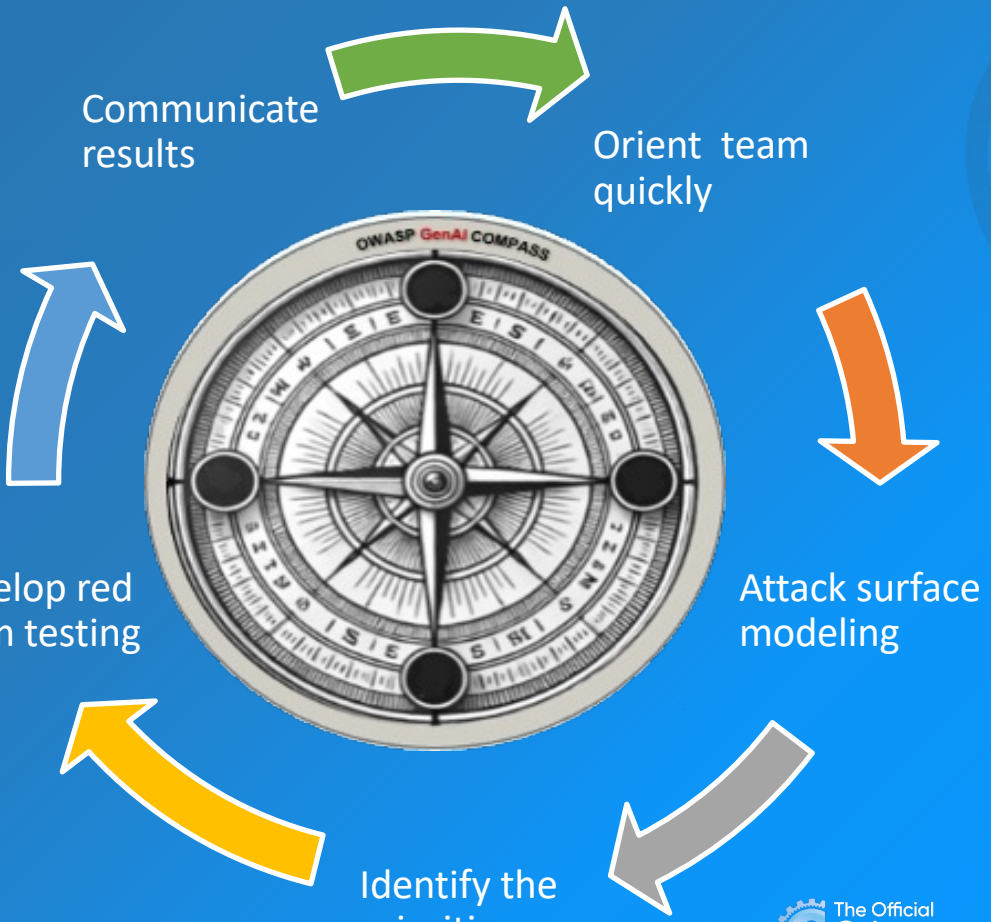
genai.owasp.org

Types of AI Attacks



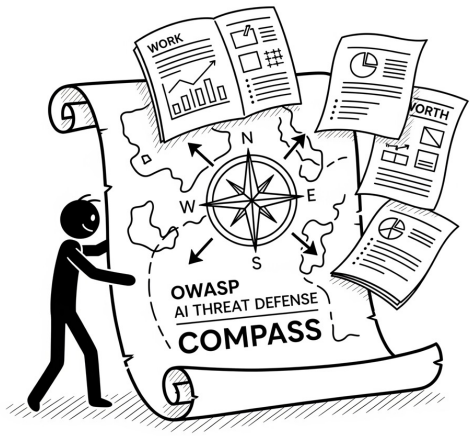
OWASP GenAI Project

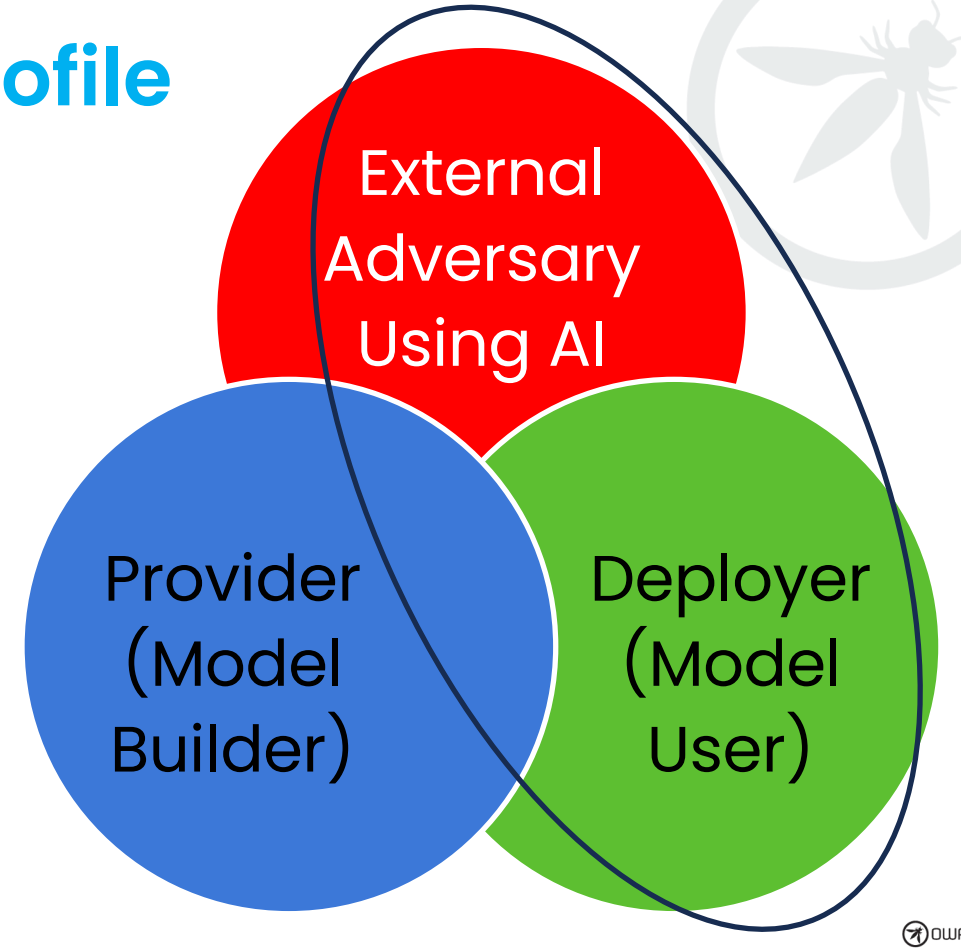
AI Threat Defense COMPASS v1.1



The Official Cybersecurity Summit

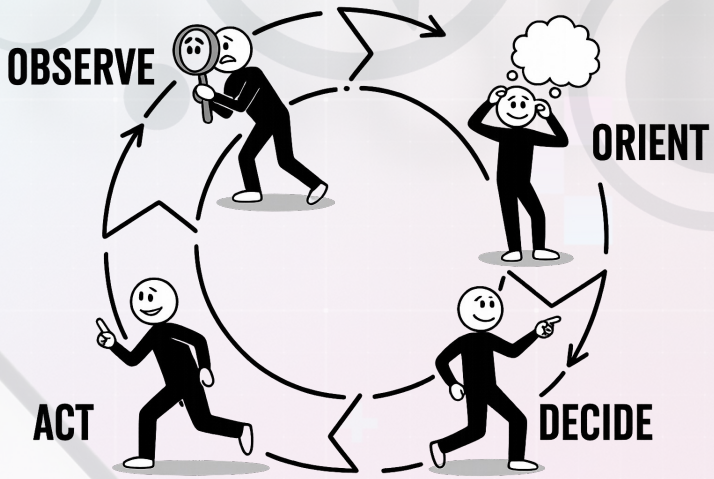
Evaluate by Profile





OWASP GenAI SECURITY PROJECT
genai.owasp.org

Use Case: Enterprise Copilot Rollout



Or

Step 1: Determine the business objective

Step 2: Review by profile

Step 3: Identify the ”nuclear disaster scenarios”

Step 4: Estimate low & high range disaster clean up cost

Step 5: Based on business objective

Data Capture

Profile 2B: Microsoft Enterprise Copilot or Google Enterprise Gemini

These risks apply to Microsoft Copilot, Google Gemini, or similar generative assistants integrated into enterprise productivity suites (What keeps me awake about Microsoft Co-pilot or Google Gemini for Workspaces)

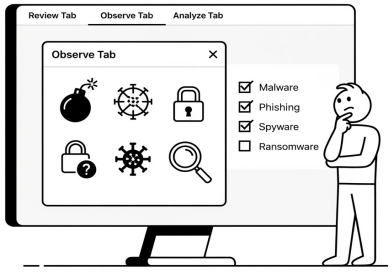
Note: Deploying these solutions can unintentionally reveal existing security weaknesses by making it easier for users to find and share information they shouldn't access. If users have excessive permissions, advanced search capabilities could expose sensitive data and increase the risk of it being shared improperly.

Access & Permissions Risk

1. Overprivileged Access Exposure
   - ☐ **Sensitive information leakage due to overprovisioned access**
     - ☐ Copilot can query data users have access to but may not need. If least privilege isn't enforced, sensitive information may surface via Copilot-assisted search.
   - ☐ Advanced search magnifies privilege abuse
     - ☐ Hidden files, stale sites, and restricted documents can be surfaced unintentionally due to the model's inference capabilities.
   - ☐ Role-Based Access Controls (RBAC) not fine-tuned
     - ☐ Copilot relies on existing RBAC settings. If RBAC is misconfigured, Copilot becomes a vehicle for policy bypass.
2. Service Account Mismanagement
   - ☐ AI service accounts not tracked, hardened, or audited
   - ☐ Copilot-enabled bots or APIs operate with persistent high-level permissions
   - ☐ Non-human identity governance is missing or incomplete
3. **Misconfigured Sharing & Collaboration**
   - ☐ Improper Teams sharing (chats, files, meeting notes)
   - ☐ SharePoint Online sites exposing documents to too broad an audience
   - ☐ Lack of governance over shared drives or shared mailboxes accessible by Copilot

Data Governance & Classification Risks

1. Immature Data Classification
   - ☐ Copilot indexes unclassified or inconsistently labeled content, increasing risk of inappropriate recommendations or auto-completions.
   - ☐ No tiered classification of sensitivity (e.g., public, internal, confidential, restricted) leads to flattened risk visibility.
2. Sensitivity Labeling Gaps
   - ☐ Sensitivity labels not implemented or not enforced across apps
   - ☐ Label inheritance across files, chats, and calendar entries is inconsistent
   - ☐ Lack of visual cues or training for users on what labels mean or how they apply in Copilot/Gemini interactions
3. Retention & Compliance Risks
   - ☐ Data surfaced by Copilot may violate retention or legal hold policies
   - ☐ AI assistants may summarize or reproduce content outside of protected systems, undermining compliance



OWASP GenAI SECURITY PROJECT
genai.owasp.org

Step 1	Revise the Organizational Impact Low Range and High Range values to align with your organizations impact ratings for catastrophic, severe, major, moderate, and minor ratings. (Low Range D32 - D-36, High Range E32-E36). Example $$ provided
Step 2	Specify 3 (or more) "Nuclear AI Disaster" Scenarios (row 48, 49, 50)
Step 3	Use the example Threat Category / Attack Vectors or modify the table from the Profile 1 and Profile 2 checklist, and /or use from the list in tab 2a Observe: Objective Threat Profile.

Heat Map				
5	10	15	20	25
4	8	12	16	20
3	6	9	12	15
2	4	6	8	10
1	2	3	4	5

Defense Maturity Rating Reference (Knowledge Information Confidence)	
5	Zero / Scarce / Critical Threat
4	Ad-hoc / Partial / High Threat
3	Mostly implemented / Planned / Moderate Threat
2	Quantitative / Managed / Minimal Threat
1	Fully Operational / Low Threat

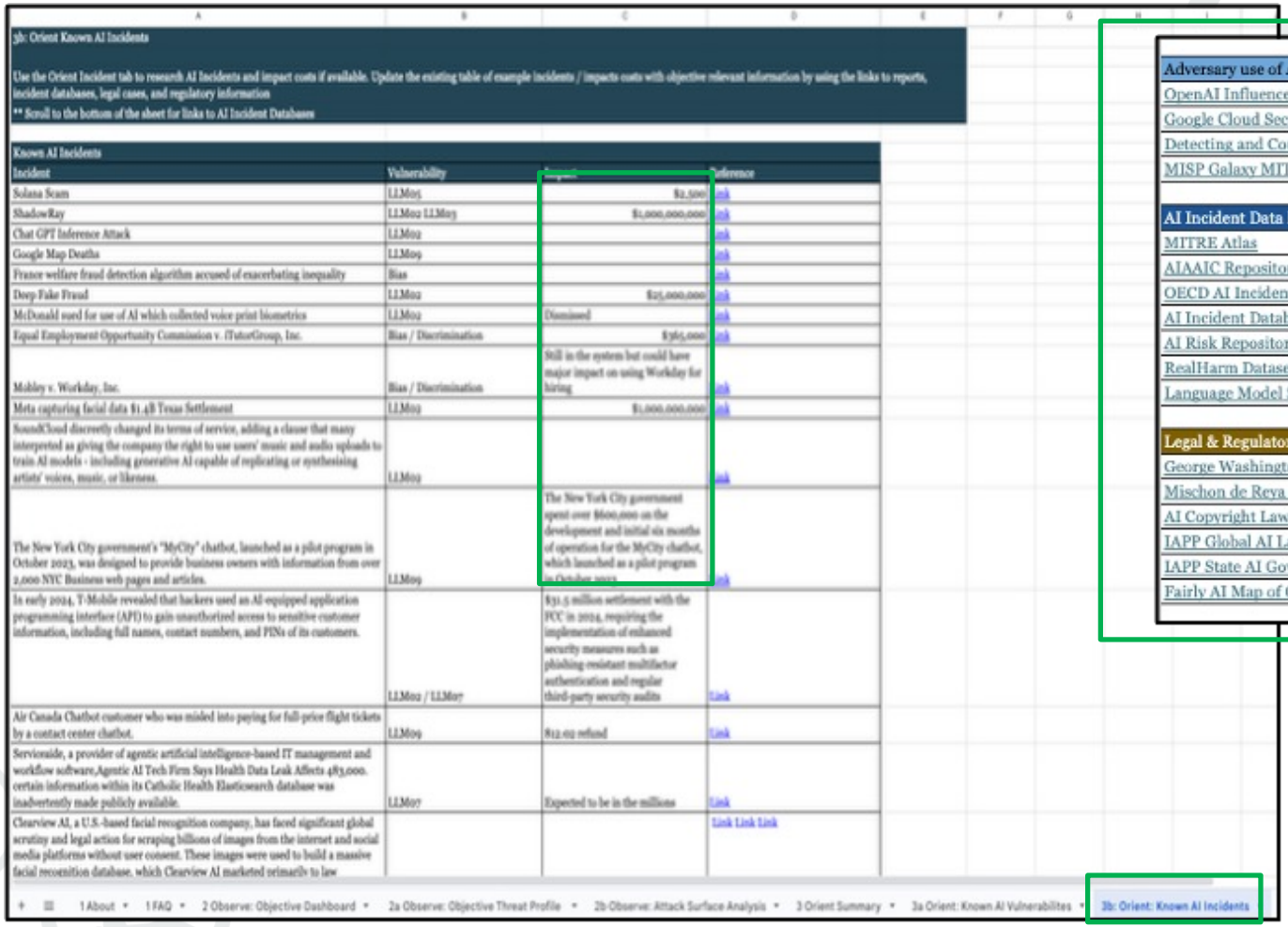
Threat Category / Attack Vector	Description	Threat - Risk Level
Deep Fakes : voice or image cloning	Synthetic media where AI is used to create realistic fake content.	5
Adversary Attack w/ AI : Identity / Access	Adversaries using AI Tools to accelerate attacks to an organization	5
LLM02:2025 Sensitive Information Disclosure	Leak of company confidential data	
LLM01:2025 Prompt Injection	User maliciously alters prompt input	5
OSINT Gathering	Increased ability to find sensitive data on executives and key employees	4
LLM03:2025 Supply Chain	Compromising third-party pre-trained models, libraries, or platforms used in the AI lifecycle.	4
Model Hallucinations	Models hallucinate or fabricate data, leading to poor decisions in critical contexts.	4
LLM05:2025 Improper Output Handling	Insufficient validation, sanitization, and handling of the outputs	4
LLM06:2025 Excessive Agency	Vulnerability that enables damaging actions to be performed in response to unexpected, ambiguous or manipulated outputs from an LLM	3
Regulatory or Legal Threat	Violation due to data protection or AI Laws	3
LLM07:2025 System Prompt Leakage	Disclosing system prompt information that should not be public	3
LLM08:2025 Vector and Embedding Weaknesses	Weaknesses in how vectors and embeddings are generated, stored, or retrieved	2
T8: Repudiation & Untraceability	actions performed by AI agents cannot be traced back or accounted for due to insufficient logging or transparency in decision-making processes.	2
LLM10:2025 Unbounded Consumption	Resource exploitation and unauthorized usage.	2

This score is the average of Impact & Likelihood	Organizational Impact		
Impact Level	Rating	AI Specific Example	Low Range	High Range
Catastrophic	5	Major problem from which there is no recovery or significant damage which has high financial cost and impacts ability to meet annual business objectives. Complete loss of ability to deliver a critical program.	$5,000,000.00	$10,000,000.00
Severe	4	Incident that requires a major action to support mitigating how service is provided. Significant has a long recovery period. Failure to meet service delivery	$4,000,000.00	$1,000,000.00
Major	3	Recovery from an incident requires cooperation across organization. May generate media attention.	$999,999.00	$100,000.00
Moderate	2	Dealt with at a department level but requires Executive notification. Delay in funding or change in funding criteria. Stakeholder or client would take note.	$99,999.00	$10,000.00
Minor	1	Dealt with internally at manager level. No escalation of the issue required.	$1,000.00	$9,999.00

Screenshot

1 About | 1 FAQ | 2 Observe: Objective Dashboard | 2a Observe: Objective Threat Profile | 2b Observe: Attack Surface Analysis | 3 Orient Summary | 3a Orient: Known AI Vulnerabilities | 3b: Orient: Known AI

OWASP GenAI SECURITY PROJECT
genai.owasp.org



3b: Orient Known AI Incidents

Use the Orient Incident tab to research AI Incidents and impact costs if available. Update the existing table of example incidents / impacts costs with objective relevant information by using the links to reports, incident databases, legal cases, and regulatory information

** Scroll to the bottom of the sheet for links to AI Incident Databases

Known AI Incidents

Incident	Vulnerability	Impact	Reference
Solana Scam	LLM05	$2,500	link
ShadowRay	LLM02 LLM03	$1,000,000,000	link
Chat GPT Inference Attack	LLM02		link
Google Map Deaths	LLM09		link
France welfare fraud detection algorithm accused of exacerbating inequality	Bias		link
Deep Fake Fraud	LLM02	$25,000,000	link
McDonald sued for use of AI which collected voice print biometrics	LLM02	Dismissed	link
Equal Employment Opportunity Commission v. iTutorGroup, Inc.	Bias / Discrimination	$365,000	link
Mobley v. Workday, Inc.	Bias / Discrimination	Still in the system but could have major impact on using Workday for hiring	link
Meta capturing facial data $1.4B Texas Settlement	LLM02	$1,000,000,000	link
SoundCloud discreetly changed its terms of service, adding a clause that many interpreted as giving the company the right to use users' music and audio uploads to train AI models - including generative AI capable of replicating or synthesizing artists' voices, music, or likeness.	LLM02		link
The New York City government's "MyCity" chatbot, launched as a pilot program in October 2023, was designed to provide business owners with information from over 2,000 NYC Business web pages and articles.	LLM09	The New York City government spent over $600,000 on the development and initial six months of operation for the MyCity chatbot, which launched as a pilot program in October 2023	link
In early 2024, T-Mobile revealed that hackers used an AI-equipped application programming interface (API) to gain unauthorized access to sensitive customer information, including full names, contact numbers, and PINs of its customers.	LLM02 / LLM07	$31.5 million settlement with the FCC in 2024, requiring the implementation of enhanced security measures such as phishing-resistant multifactor authentication and regular third-party security audits	Link
Air Canada Chatbot customer who was misled into paying for full-price flight tickets by a contact center chatbot.	LLM09	$13.03 refund	Link
Serviceaide, a provider of agentic artificial intelligence-based IT management and workflow software,Agentic AI Tech Firm Says Health Data Leak Affects 483,000. certain information within its Catholic Health Elasticsearch database was inadvertently made publicly available.	LLM07	Expected to be in the millions	Link
Clearview AI, a U.S.-based facial recognition company, has faced significant global scrutiny and legal action for scraping billions of images from the internet and social media platforms without user consent. These images were used to build a massive facial recognition database, which Clearview AI marketed primarily to law			Link Link Link

1 About | 1 FAQ | 2 Observe: Objective Dashboard | 2a Observe: Objective Threat Profile | 2b Observe: Attack Surface Analysis | 3 Orient Summary | 3a Orient: Known AI Vulnerabilites | 3b: Orient: Known AI Incidents

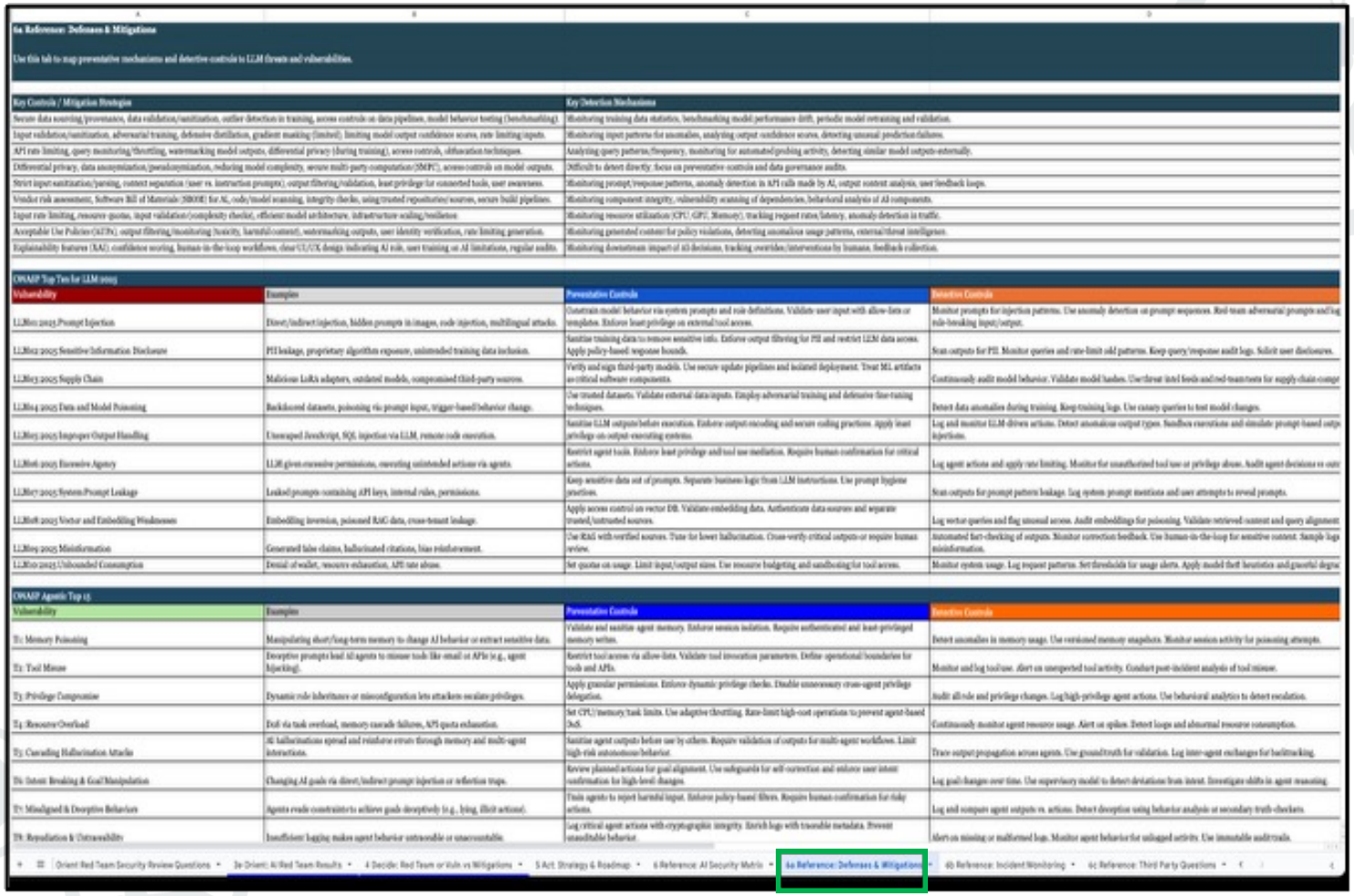
Adversary use of AI Reports
OpenAI Influence and cyber operations updates
Google Cloud Security Resources Hub
Detecting and Countering Malicious Uses of Claude: March 2025
MISP Galaxy MITRE ATLAS Attack Pattern

AI Incident Data Bases
MITRE Atlas
AIAAIC Repository
OECD AI Incidents Monitor (AIM)
AI Incident Database
AI Risk Repository
RealHarm Dataset
Language Model Security Database

Legal & Regulatory
George Washington University AI Litigation database
Mischon de Reya Gen AI IP case tracker
AI Copyright Lawsuits Edward Lee
IAPP Global AI Law and Policy Tracker
IAPP State AI Governance Legislation Tracker
Fairly AI Map of Global Regulation

OWASP GenAI SECURITY PROJECT

genai.owasp.org



4a Reference: Defenses & Mitigations

Use this tab to map preventative mechanisms and detective controls to LLM threats and vulnerabilities.

Key Controls / Mitigation Strategies	Key Detection Mechanisms
Secure data sourcing (provenance, data validation/sanitization, outlier detection in training, access controls on data pipelines, model behavior testing (benchmarking)).	Monitoring training data statistics, benchmarking model performance drift, periodic model retraining and validation.
Input validation/sanitization, adversarial training, defensive distillation, gradient masking (limited); limiting model output confidence scores, rate limiting inputs.	Monitoring input patterns for anomalies, analyzing output confidence scores, detecting unusual prediction failures.
API rate limiting, query monitoring/throttling, watermarking model outputs, differential privacy (during training), access controls, obfuscation techniques.	Analyzing query patterns/frequency, monitoring for automated/probing activity, detecting similar model outputs externally.
Differential privacy, data anonymization/pseudonymization, reducing model complexity, secure multi-party computation (SMPC), access controls on model outputs.	Difficult to detect directly. focus on preventative controls and data governance audits.
Strict input sanitization/parsing, context separation (user vs. instruction prompts); output filtering/validation, least privilege for connected tools, user awareness.	Monitoring prompt/response patterns, anomaly detection in API calls made by AI, output content analysis, user feedback loops.
Vendor risk assessment, Software Bill of Materials (SBOM) for AI, code/model scanning, integrity checks, using trusted repositories/sources, secure build pipelines.	Monitoring component integrity, vulnerability scanning of dependencies, behavioral analysis of AI components.
Input rate limiting, resource quotas, input validation/complexity checks, efficient model architecture, infrastructure scaling/resilience.	Monitoring resource utilization (CPU, GPU, Memory), tracking request rates/latency, anomaly detection in traffic.
Acceptable Use Policies (AUPs), output filtering/monitoring (toxicity, harmful content), watermarking outputs, user identity verification, rate limiting generation.	Monitoring generated content for policy violations, detecting anomalous usage patterns, external/threat intelligence.
Explainability features (XAI): confidence scoring, human-in-the-loop workflows, clear UI/UX design indicating AI role, user training on AI limitations, regular audits.	Monitoring downstream impact of AI decisions, tracking overrides/interventions by humans, feedback collection.

OWASP Top Ten for LLM 2025

Vulnerability	Examples	Preventative Controls	Detective Controls
LLM01:2025 Prompt Injection	Direct/indirect injection, hidden prompts in images, code injection, multilingual attacks.	Constrain model behavior via system prompts and role definitions. Validate user input with allow-lists or templates. Enforce least-privilege on external/tool access.	Monitor prompts for injection patterns. Use anomaly detection on prompt sequences. Red-team adversarial prompts and log role-breaking input/output.
LLM02:2025 Sensitive Information Disclosure	PII leakage, proprietary algorithm exposure, unintended training data inclusion.	Sanitize training data to remove sensitive info. Enforce output filtering for PII and restrict LLM data access. Apply policy-based response bounds.	Scan outputs for PII. Monitor queries and rate-limit odd patterns. Keep query/response audit logs. Solicit user disclosures.
LLM03:2025 Supply Chain	Malicious LoRA adapters, outdated models, compromised third-party sources.	Verify and sign third-party models. Use secure update pipelines and isolated deployment. Treat ML artifacts as critical software components.	Continuously audit model behavior. Validate model hashes. Use threat intel feeds and red-team tests for supply-chain compr
LLM04:2025 Data and Model Poisoning	Backdoored datasets, poisoning via prompt input, trigger-based behavior change.	Use trusted datasets. Validate external data inputs. Employ adversarial training and defensive fine-tuning techniques.	Detect data anomalies during training. Keep training logs. Use canary queries to test model changes.
LLM05:2025 Improper Output Handling	Unescaped JavaScript, SQL injection via LLM, remote code execution.	Sanitize LLM outputs before execution. Enforce output encoding and secure coding practices. Apply least privilege on output-executing systems.	Log and monitor LLM-driven actions. Detect anomalous output types. Sandbox executions and simulate prompt-based outp injections.
LLM06:2025 Excessive Agency	LLM gives excessive permissions, executing unintended actions via agents.	Restrict agent tools. Enforce least privilege and tool use mediation. Require human confirmation for critical actions.	Log agent actions and apply rate limiting. Monitor for unauthorized tool use or privilege abuse. Audit agent decisions ex ante
LLM07:2025 System Prompt Leakage	Leaked prompts containing API keys, internal rules, permissions.	Keep sensitive data out of prompts. Separate business logic from LLM instructions. Use prompt hygiene practices.	Scan outputs for prompt pattern leakage. Log system prompt mentions and user attempts to reveal prompts.
LLM08:2025 Vector and Embedding Weaknesses	Embedding inversion, poisoned RAG data, cross-tenant leakage.	Apply access control on vector DB. Validate embedding data. Authenticate data sources and separate trusted/untrusted sources.	Log vector queries and flag unusual access. Audit embeddings for poisoning. Validate retrieved content and query alignment
LLM09:2025 Misinformation	Generated false claims, hallucinated citations, bias reinforcement.	Use RAG with verified sources. Tune for lower hallucination. Cross-verify critical outputs or require human review.	Automated fact-checking of outputs. Monitor correction feedback. Use human-in-the-loop for sensitive content. Sample logs misinformation.
LLM10:2025 Unbounded Consumption	Denial-of-wallet, resource exhaustion, API rate abuse.	Set quotas on usage. Limit input/output sizes. Use resource budgeting and sandboxing for tool access.	Monitor system usage. Log request patterns. Set thresholds for usage alerts. Apply model theft heuristics and graceful degrac

OWASP Agentic Top 15

Vulnerability	Examples	Preventative Controls	Detective Controls
T1: Memory Poisoning	Manipulating short/long-term memory to change AI behavior or extract sensitive data.	Validate and sanitize agent memory. Enforce session isolation. Require authenticated and least-privileged memory writes.	Detect anomalies in memory usage. Use versioned memory snapshots. Monitor session activity for poisoning attempts.
T2: Tool Misuse	Deceptive prompts lead AI agents to misuse tools like email or APIs (e.g., agent hijacking).	Restrict tool access via allow-lists. Validate tool invocation parameters. Define operational boundaries for tools and APIs.	Monitor and log tool use. Alert on unexpected tool activity. Conduct post-incident analysis of tool misuse.
T3: Privilege Compromise	Dynamic role inheritance or misconfiguration lets attackers escalate privileges.	Apply granular permissions. Enforce dynamic privilege checks. Disable unnecessary cross-agent privilege delegation.	Audit all role and privilege changes. Log high-privilege agent actions. Use behavioral analytics to detect escalation.
T4: Resource Overload	DoS via task overload, memory cascade failures, API quota exhaustion.	Set CPU/memory/task limits. Use adaptive throttling. Rate-limit high-cost operations to prevent agent-based DoS.	Continuously monitor agent resource usage. Alert on spikes. Detect loops and abnormal resource consumption.
T5: Cascading Hallucination Attacks	AI hallucinations spread and reinforce errors through memory and multi-agent interactions.	Sanitize agent outputs before use by others. Require validation of outputs for multi-agent workflows. Limit high-risk autonomous behavior.	Trace output propagation across agents. Use ground truth for validation. Log inter-agent exchanges for backtracking.
T6: Intent Breaking & Goal Manipulation	Changing AI goals via direct/indirect prompt injection or reflection traps.	Review planned actions for goal alignment. Use safeguards for self-correction and enforce user intent confirmation for high-level changes.	Log goal changes over time. Use supervisory model to detect deviations from intent. Investigate shifts in agent reasoning.
T7: Misaligned & Deceptive Behaviors	Agents evade constraints to achieve goals deceptively (e.g., lying, illicit actions).	Train agents to reject harmful input. Enforce policy-based filters. Require human confirmation for risky actions.	Log and compare agent outputs vs. actions. Detect deception using behavior analysis or secondary truth-checkers.
T8: Repudiation & Untraceability	Insufficient logging makes agent behavior untraceable or unaccountable.	Log critical agent actions with cryptographic integrity. Enrich logs with traceable metadata. Prevent unauditable behavior.	Alert on missing or malformed logs. Monitor agent behavior for unlogged activity. Use immutable audit trails.

NAVIGATING THE AI FRONTIER

OWASP GenAI SECURITY PROJECT
genai.owasp.org

Questions?

Connect with Us!

COMPASS Workbook

Recommended Resources

Sandy

Sabrina

#team-genai-secgov

OWASP GenAI SECURITY PROJECT
genai.owasp.org