

Behind the Prompt: Exposing and Mitigating the Top LLM Vulnerabilities –

ISC2's Spotlight on AI virtual conference

Ads Dawson & Steve Wilson



GenAI SECURITY
PROJECT

genai.owasp.org

whoarewe



ads (@GangGreenTemperTatum)

Staff AI Security Researcher, Hacker
Technical lead for OWASP Top 10 for LLM Applications
OWASP Toronto chapter lead

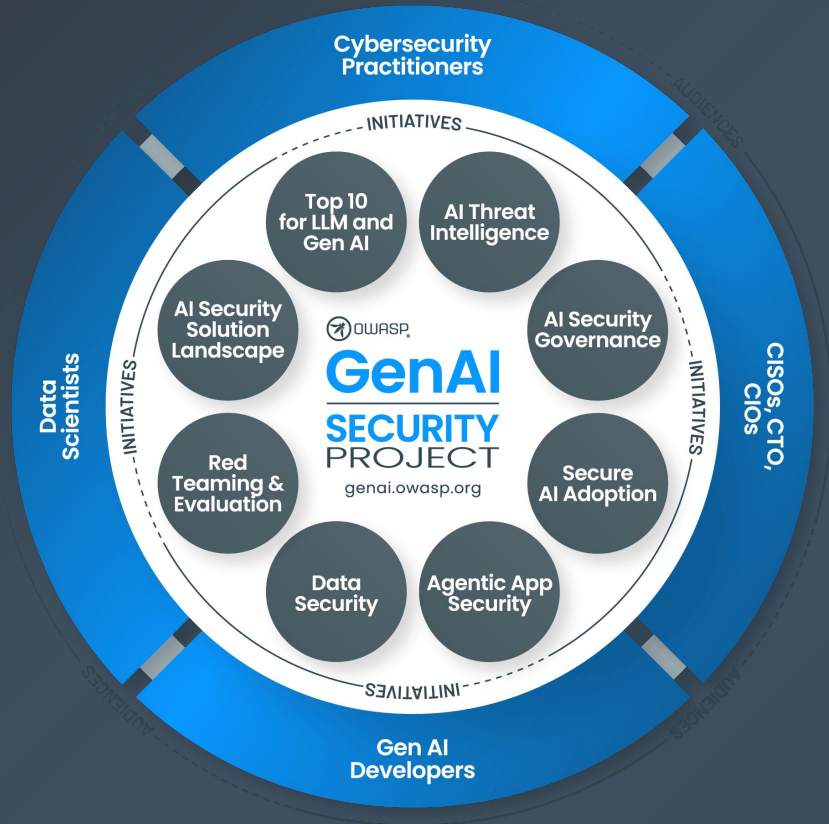
Harnessing code to conjure creative chaos.. think evil;
do good.



Steve Wilson

Chief AI and Product Officer, Exabeam
Founder, OWASP Gen AI Security Project
Author, O'Reilly's Developer's Playbook for LLM
Security

Robot dog herder



OWASP GenAI Security Project

About The Project

OWASP Foundation

- The Open Worldwide Application Security Project (OWASP) is a **nonprofit foundation that works to improve the security of software.**
- Community-led open-source projects including code, documentation, and standards
- Over **250+ local chapters** worldwide
- **200,000+** global community
- Industry-leading educational and training conferences

OWASP GenAI Security Project

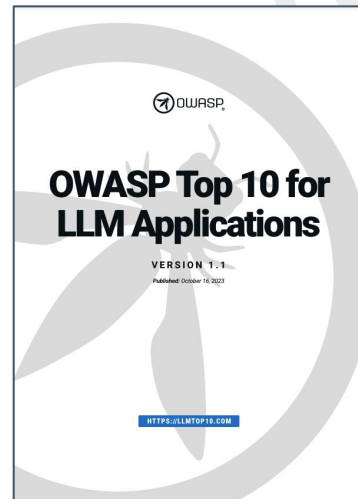
- A global, **open-source community project** focused on **generative AI security.**
- Dedicated to identifying, mitigating, and **documenting security and safety risks.**
- Covers technologies like Large Language Models (LLMs), agentic AI, and AI-driven applications.
- Aims to **empower organizations**, security professionals, AI practitioners, and policymakers.
- Provides **practical guidance, resources** and **tools** for secure AI development, deployment, and governance.

Initiative – Top 10 for LLMs

Initiative Leads: Ads Dawson, Steve Wilson

The OWASP Top 10 for Large Language Model Applications started in 2023 to highlight and address security issues specific to AI applications. As LLMs are embedded more deeply in everything from customer interactions to internal operations, developers and security professionals are discovering new vulnerabilities—and ways to counter them.

Key Resources:



AI Red Teaming

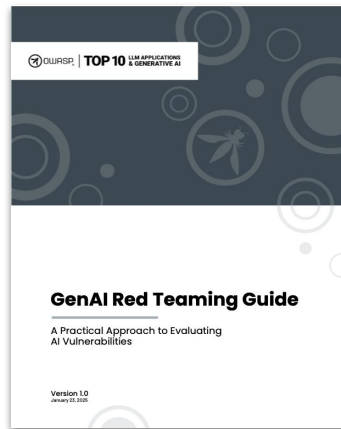
Initiative Lead: Jason Ross

This initiative delivers the first standardized methodology for AI Red Teaming, addressing the unique challenges of LLMs in security, safety, and trust. By conducting adversarial testing and evaluations, the project equips organizations with comprehensive guidelines, benchmarks, and frameworks to enhance model robustness and mitigate risks like bias, inaccuracies, and undetected vulnerabilities

- Introduces LLM-aware red teaming methodology.
- Covers threat modelling, system evaluation, prompt injection, and runtime abuse.
- Phased Blueprint Rollout

Slack: #team-genai-redteaming

Key Resources:



Up Next: LLM Testing Guide

AI RED TEAMING

Agentic Security

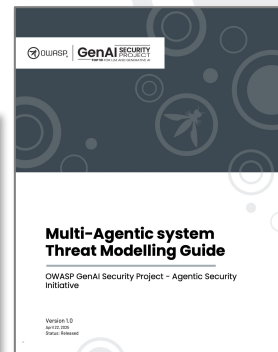
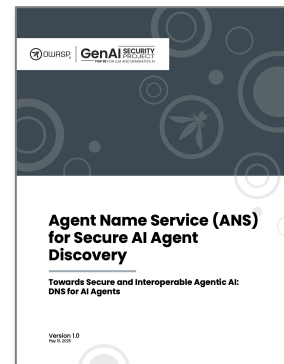
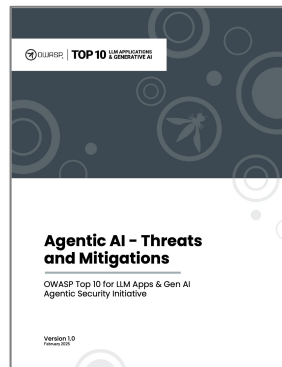
Initiative Lead: John Sotiropoulos, Ron F. Del Rosario

The Agentic Security Research Initiative explores the emerging security implications of agentic systems, particularly those utilizing advanced frameworks (e.g., LangGraph, AutoGPT, CrewAI) and novel capabilities like Llama 3's agentic features..

Multiple Workstreams

- Threats & Mitigations
- Threat Modeling, with Maestro
- Code Samples, Hackathons
- Securing Agentic Apps including Agentic Protocols (MCP, A2A)
- First-party Research
- Tracking the Agentic Landscape

Key Resources:



**AGENTIC
SECURITY**

Up Next: Agentic AI Threats and Mitigations 1.1

OWASP
GenAI SECURITY
PROJECT
genai.owasp.org
<https://genai.owasp.org>

Slack: #team-genai-agentic

Upcoming Patterns and Trends

Forecasting



GenAI SECURITY
PROJECT

genai.owasp.org

Critical Trends

The current wave of 2025 with agentic AI

The space is much bigger and faster – Security is most commonly an afterthought in a normal moving pace environment

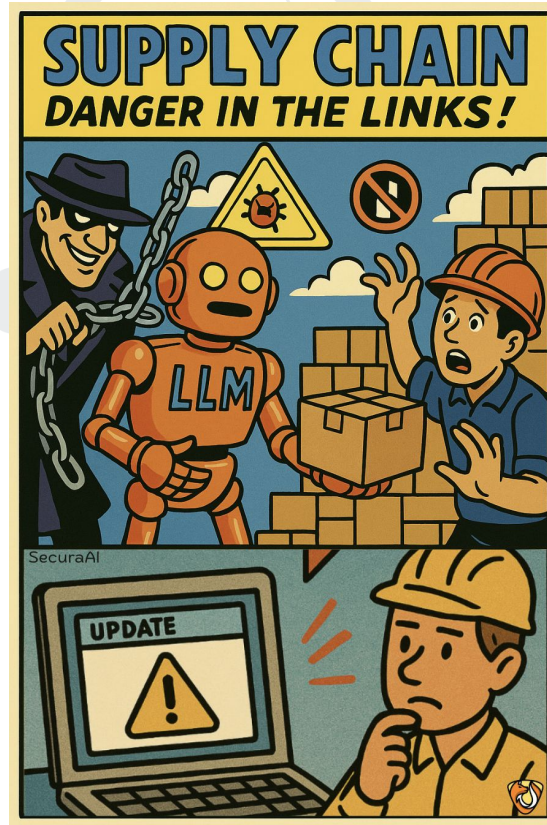
MCP-Driven Autonomy – Standardizes how AI applications connect to external tools and data—turning the traditional M×N integration headache into an M+N solution.

Reasoning Models – Advanced AI reasoning enhances multi-step attacks, deception, and adversarial security.

Agentic AI Expansion – AI agents now orchestrate offensive and defensive ops, raising new exploitation risks.

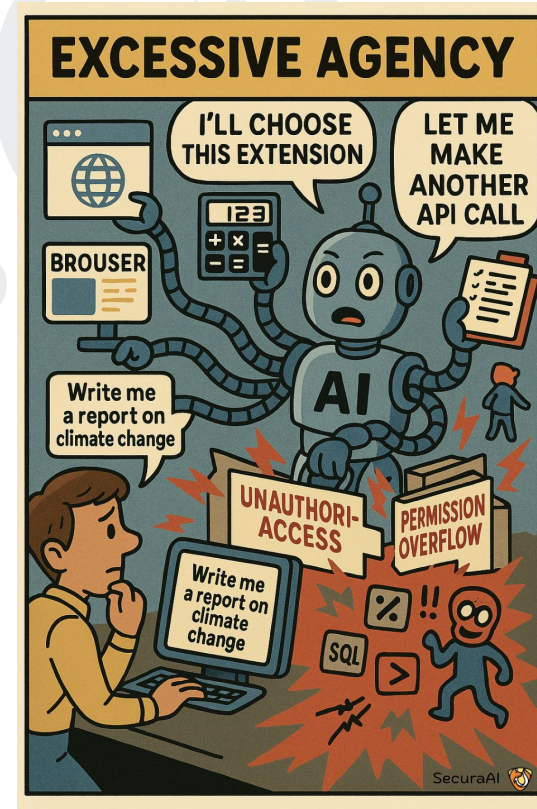
Growing Attack Surfaces – AI's multi-environment interactions introduce novel vulnerabilities and systemic threats.

Radically More Multimodal – AI is becoming fully multimodal, combining text, vision, audio, and code for more sophisticated exploits and defenses.



How to Build Secure Software with Stochastic Models?

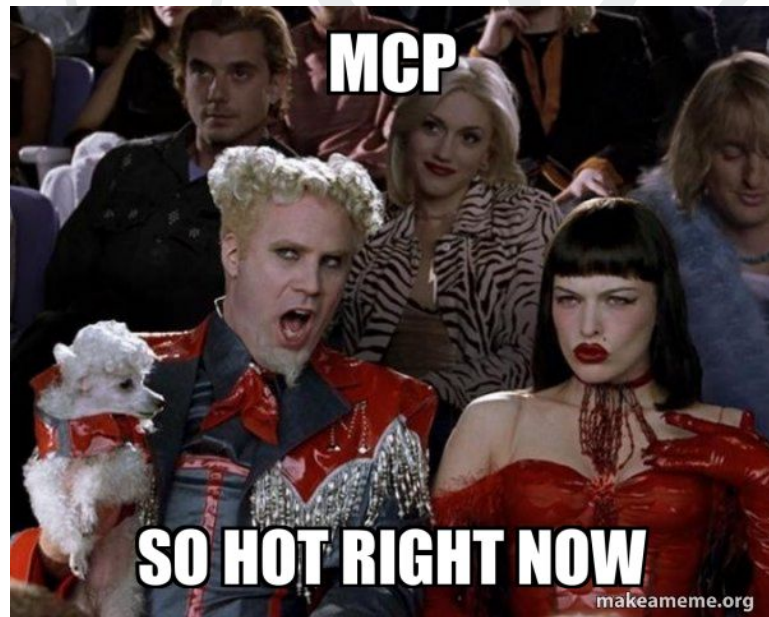
- **Threat Modeling First** – Identify attack surfaces, failure modes, and adversarial risks early.
- **Secure Decision Boundaries** – Enforce constraints to prevent overreach and hallucinations.
- **Adaptive Security Policies** – Use controlled randomness to harden defenses against evolving threats.
- **Principle of Least Privilege Throughout the Stack** –
 - E.G [arXiv:2504.11703v1](https://arxiv.org/abs/2504.11703v1), [Progent: Programmable Privilege Control for LLM Agents](#)
- **Manage Uncertainty** – Quantify and mitigate risks from probabilistic AI decisions.
- **Adversarial Resilience** – Harden models against perturbations, poisoning, and injections.
- **Continuous Risk Assessment** – Detect drift, anomalies, and unexpected failures.
- **Explainability & Auditing** – Ensure traceability and transparent security assessments.
- **AI Red Teaming** – Simulate real-world attacks to identify weaknesses early.



Forecasting the Future

The "S" in MCP Stands for Security

- MCP - Model Context Protocol
 - *Remember the move fast thing?* [#284](#)
[RFC] Update the Authorization specification for MCP servers
- Emerging Threats: Multi-step reasoning exploits, prompt injections, tool abuse
- Expanded Surface: Cross-service integrations, plugins, data stores, APIs
- Multimodal Risks: Text, vision, audio, code all in the same loop
- Defensive Playbook: Threat modeling, uncertainty management, continuous monitoring
- OWASP Evolution: Top 10 for LLMs → Agentic AI Threats Navigator



GenAI SECURITY
PROJECT

genai.owasp.org

Forecasting the Future

Googles A2A Protocol the new hype?

- **Enterprise-Grade AuthN/AuthZ:** Supports protocols like OAuth 2.0 to ensure only authorized agents can interact.
- **OpenAPI Compatibility:** Uses OpenAPI specs with Bearer tokens for standardized, secure agent communication.
- **Access Control (RBAC):** Enables fine-grained permissioning to restrict agent actions based on roles.
- **Data Encryption:** Ensures secure data exchange (e.g., HTTPS) to protect sensitive information in transit.
- **Evolving Authorization:** Future plans include expanding AgentCard with optional embedded credentials for enhanced control.



OWASP

GenAI SECURITY
PROJECT

genai.owasp.org

Where Can I Connect and Learn More?

Linked-In

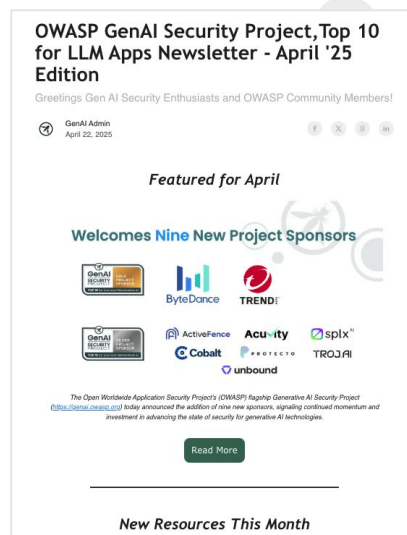


<https://www.linkedin.com/company/owasp-top-10-for-large-language-model-applications/>

Podcast

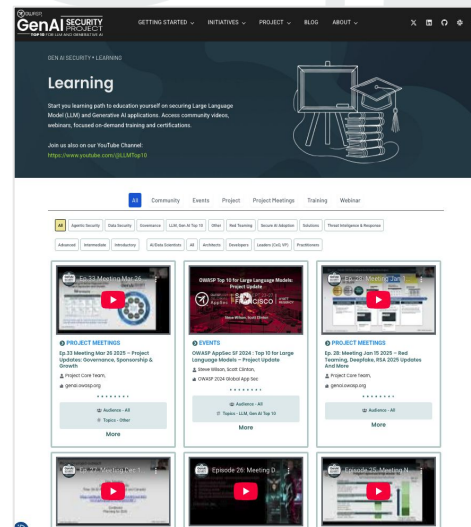


Newsletter



<https://genai.owasp.org/newsletter/>

Learning Portal



<https://genai.owasp.org/learning/>

Join the Bi-weekly Open Meeting

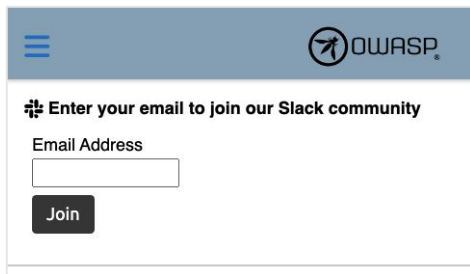
Wed, 9:00am PST

genai.owasp.org/meetings/

GenAI SECURITY PROJECT
genai.owasp.org

How Can I Contribute?

Request and OWASP Slack Invite



The screenshot shows the OWASP website's Slack invite page. It features the OWASP logo at the top right. Below it, a heading reads "Enter your email to join our Slack community". There is a text input field labeled "Email Address" and a "Join" button below it.

<https://owasp.org/slack/invite#>

Join a Working Group Slack

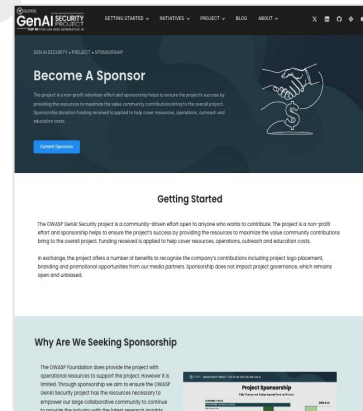


The screenshot shows the "Slack Working Groups, and Contacts" page for the "LLM and Gen AI Application Security Project". It lists various resources, initiatives, and other areas with their respective leads and Slack channels.

KEY RESOURCES / INITIATIVES / OTHER AREAS	LEADS	SLACK CHANNEL
Key Resources		
OWASP Top 10 for LLMs Application Security (Each item has a designated slack channel #team-llm-ai-sec)	Steve Wilson Ada Swanson	#project-top10-for-llms # Check the Wiki for a detailed list
CISCO Checklist	Sande Datta	# team-llm-ai-sec-gov
Gen AI Security Solutions Landscape	Scott Clifton	# team-llm-solutions
Initiatives		
Data Gathering/Methodology	Emmanuel Subirats	# team-llm-datagathering-methodology
Gen AI Threat Intelligence	Rachel James	# team-llm-ai-cti
Secure AI Adoption	Scott Clifton	# team-llm-sec
LLM & Gen AI Red Teaming & Evaluation	Grady Searles	# team-llm-redteams
Agentic App Security	John Satriapokus	# team-ai-autonomous-agents
Other Areas		
OWASP Top 10 for LLM General and Discussion Channels	Steve Wilson	# team-llm-discuss
Standards Alignment	John Satriapokus	# team-llm-standards-alignment
Project Outreach, Growth and Operations	Scott Clifton	# team-llm-project-outreach
Project Sponsorship	Scott Clifton	# team-llm-sponsorship-project

<https://genai.owasp.org/roadmap/>

Sponsor/Support



<https://genai.owasp.org/sponsorship/>



Thank You

Questions & Fireside Chat



GenAI SECURITY
PROJECT

genai.owasp.org