OWASP AI Summit

Introducing a Solutions Framework for LLMs, Gen AI Security

Scott Clinton

Core Team and Sub-project Lead

Scott Clinton

President, SCVentures, Ltd
Advisory Board, Zafran Security

Project Core Team Lead, OWASP LLM

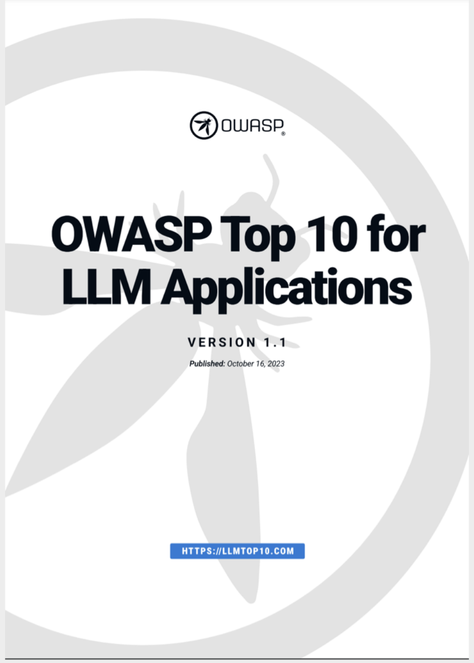
Formerly: Trend Micro, VMWare, MobileIron, Qualys, Red Hat, Sun Microsystems

15+ Years Leading Cyber Security Product and Solution Portfolios, Multiple Security Acquisitions, 12+ Years in Big Data and AI Products.

Co-lead Founding of the Project Liberty Alliance for Federated Identity (SAML)

Contributing Technical Author

Leads: Steve Wilson & Ads



OWASP

OWASP Top 10 for LLM Applications

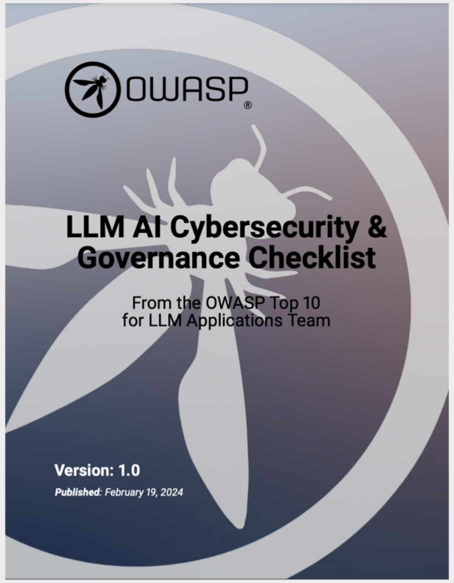
VERSION 1.1

Published: October 16, 2023

HTTPS://LLMTOP10.COM

Top 10 List:

- Developers
- AppSec Teams



OWASP

LLM AI Cybersecurity & Governance Checklist

From the OWASP Top 10 for LLM Applications Team

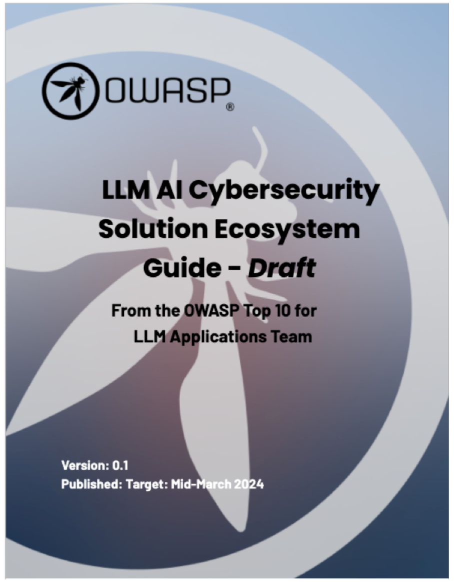
Version: 1.0

Published: *February 19, 2024*

Checklist:

- CISOs
- Compliance Officers

Lead: Sandy Dunn



OWASP

LLM AI Cybersecurity Solution Ecosystem Guide - *Draft*

From the OWASP Top 10 for LLM Applications Team

Version: 0.1

Published: Target: Mid-March 2024

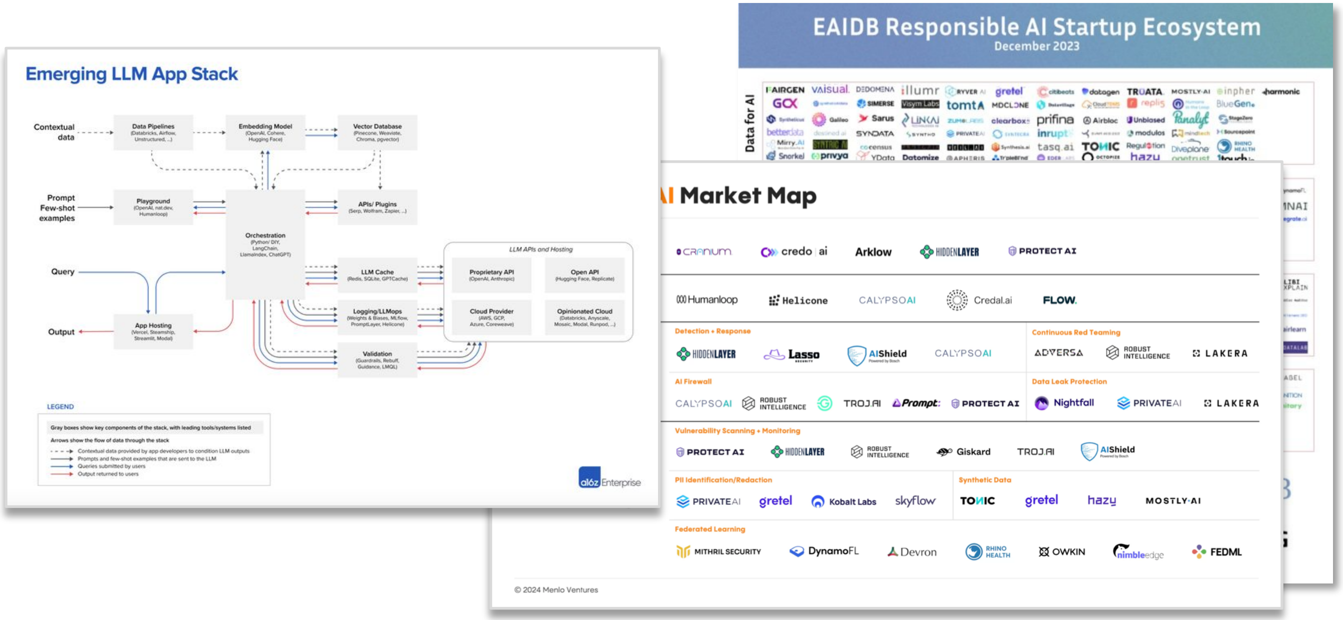
Solutions Guide:

- Development Leaders
- Security Operations

Lead: Scott Clinton

Why a Solution Framework

- Generative AI App Stack is rapidly evolving
- New security solutions and technologies emerging
- Organizations need as much guidance to keep up.
- Many POVs are biased (vendor)


Emerging LLM App Stack
Contextual data
Prompt Few-shot examples
Query
Output
Data Pipelines (Databricks, Airflow, Unstructured, ...)
Embedding Model (OpenAI, Cohere, Hugging Face)
Vector Database (Pinecone, Weaviate, Chroma, pgvector)
Playground (OpenAI, nat.dev, Humanloop)
Orchestration (Python/ DIY, LangChain, LlamaIndex, ChatGPT)
APIs/ Plugins (Serp, Wolfram, Zapier, ...)
App Hosting (Vercel, Steamship, Streamlit, Modal)
LLM Cache (Redis, SQLite, GPTCache)
Logging/LLMops (Weights & Biases, MLflow, PromptLayer, Helicone)
Validation (Guardrails, Rebuff, Guidance, LMQL)
LLM APIs and Hosting
Proprietary API (OpenAI, Anthropic)
Open API (Hugging Face, Replicate)
Cloud Provider (AWS, GCP, Azure, Coreweave)
Opinionated Cloud (Databricks, Anyscale, Mosaic, Modal, Runpod, ...)
LEGEND
a16z Enterprise
EAIDB Responsible AI Startup Ecosystem
December 2023
Data for AI
I Market Map
Detection + Response
Continuous Red Teaming
AI Firewall
Data Leak Protection
Vulnerability Scanning + Monitoring
PII Identification/Redaction
Synthetic Data
Federated Learning
© 2024 Menlo Ventures

Gray boxes show key components of the stack, with leading tools/systems listed

Arrows show the flow of data through the stack

- Contextual data provided by app developers to condition LLM outputs
- Prompts and few-shot examples that are sent to the LLM
- Queries submitted by users
- Output returned to users

OWASP LLM, AI Solution Framework

GenAI Security Solutions Project Charter

Provide a reference resource for organizations seeking to address the identified risks and enhance their security programs, making their research easier by creating a solutions reference guide to categorize, define, and align applicable technology solution areas with the emerging LLM and Generative AI Application threat landscape.

Link to charter

Target Audience

Developers, AppSec, DevSecOps, MLSecOps, Data Engineers, Data Scientists, CISOs, and security leaders looking to develop strategies to secure LLMs and Generative AI applications seek reference guidance POV on the areas to consider or track in enhancing their security programs.

Documenting Consistent LLM and Generative AI Application Lifecycle Stages

Model Lifecycle Stages	Security Solution Categories
Ethical and Compliance Governance	Model governance and compliance ethics encompass the frameworks, policies, and processes that ensure machine learning models are developed, deployed, and maintained responsibly, transparently, and in accordance with ethical standards, legal regulations, and industry best practices. This includes considerations around data privacy, model fairness, accountability, transparency, and the mitigation of biases, aiming to foster trust and reliability in AI systems while protecting individuals and society from potential harms.
Consumption	Model consumption and usage refer to the process where deployed machine learning models are utilized by end-users or applications to make predictions, analyze data, or generate insights. This final step in the model lifecycle involves interacting with the model through APIs, applications, or other interfaces to apply the model's capabilities to real-world problems and data. Ensuring secure and efficient access to the model's functionalities is crucial for maintaining data integrity, user privacy, and system reliability.
Observability/Monitoring	The model monitoring and serving process is crucial for maintaining the performance and reliability of machine learning models once they are deployed in production. This process ensures that the models continue to operate as expected over time, providing accurate predictions or analyses even as input data changes. Below is a concise overview of this process, followed by specific security controls pertinent to model monitoring and serving:
Serving and Deployment	The model provisioning and serving process involves deploying trained machine learning models so they can be used to make predictions or analyses on new data in a production environment. This process ensures that the model is accessible and performs optimally under real-world operational conditions.
Building	The model building process for Large Language Models (LLMs) and generative AI involves several sequential steps, starting from the initial design to the final training and validation stages.
Data Preparation	The process of data model preparation for Large Language Models (LLMs) and generative AI involves several critical steps, ensuring the model is trained on high-quality, diverse, and relevant datasets. This preparation is vital for the model's ability to generate accurate, coherent, and contextually appropriate responses.

Aligning Solutions to Gen AI Application Lifecycle

Model Lifecycle Stages	Security Solution Categories	
Ethical and Compliance Governance	Data Protection Privacy Protection Access Control and Authentication Audit Trails and Activity Monitoring	Bias Detection and Mitigation Tools AI Compliance Audits Incident Response and Reporting Education and Training
Consumption	Continuous Red Teaming Detection and Response Data Leak Protection	
Observability/Monitoring	Secure Model Access Input and Output Data Encryption API Security Access and Activity Logging	Anomaly Detection Secure Configuration Management
Serving and Deployment	Authentication and Authorization Data Encryption Secure APIs Access Logging and Monitoring Model Tampering Protection	Network Security Data Privacy Compliance Regular Security Audits
Building	Secure Data Storage Secure Data Transmission Access Controls to Training Environments	Model Integrity Controls Secure Deployment Vulnerability Management
Data Preparation	Data Masking and Anonymization Access Control Data Loss Prevention (DLP) Software Secure Data Storage Data Audit and Logging	Data Privacy Data Integrity Controls Synthetic Data Generation Network Security Controls Compliance and Governance Reporting

Mapping Lifecyle Fit and Top 10 for LLM Coverage

Data supplied by vendors and OSS project leaders.

Company	Product/Solution	Lifecyle Stage(s)	Top 10 For LLM Coverage
Company	Product	Consumption	**LLM01, LLM04**
Open Source	Project	Data Preparation	**LLM05**
...			
...			

Provide an on-line web resource for organizations and OSS projects to list themselves and coverage

Why OWASP

Continues to deliver on the OWASP mission to provide actionable resources for the challenges development and security teams are facing but does not focus on the broader long term issues such as surveillance, fairness, and social impact.

This project is intended to serve as a companion to the OWASP Top 10 for Large Language Model Applications List and the CISO Cybersecurity & Governance Checklist

Roadmap

4.26.24 - Initial draft .05

4.26.24 - 5.3.24 - Initial comment and input period

5.5.24 - first draft

5.6.24 - 5.20.24 - second review and input period

5.31.24 - publication

Contribute – Join the OWASP SLACK channel #llm-team-solutions

THANK YOU