OWASP LLM AI Security & Governance Checklist

Version: 0.5

Published:*November 29, 2023*

Revision History

Revision	Date	Author(s)	Description
0.1	2023-11-01	SD	initial draft
0.5	2023-11-28	SD, Team	public draft

Overview

Who is This For?

This checklist is for technology and business leadership in an organization to consider all aspects and tasks needed to map out a Large Language Model strategy.

Recent advances in artificial intelligence (AI) have emphasized the importance of organizations developing a plan to maintain proper relationship balance with AI.

- **Artificial intelligence** (AI) is a broad term that encompasses all fields of computer science that enable machines to accomplish tasks that would normally require human intelligence. Machine learning and generative AI are two subcategories of AI.
- **Machine learning** is a subset of AI that focuses on creating algorithms that can learn from data. Machine learning algorithms are trained on a set of data, and then they can use that data to make predictions or decisions about new data.
- **Generative AI** is a type of machine learning that focuses on creating new data.

How to Use This Document

For decades, researchers have been working on artificial intelligence, large language models, and diffusion models. Still, new improvements in training data availability, computer power, GenAI capacity, and the release of solutions like ChatGPT, ElevenLabs, and Midjourney have made the field more popular and led to its growth. Every internet user and business should prepare itself for the impact of a surge in powerful GenAI applications. GenAI holds enormous promise and opportunities for discovery, efficiency, and driving corporate growth across many industries and disciplines. However, as with any strong new technology, it introduces new challenges to security and privacy.

Executive, technology, cybersecurity, privacy, compliance, and legal leaders must pay close attention to the fast GenAI technological transformation and devise a strategy to benefit from opportunities while fighting against threats and managing risks.

Organizations will face challenges in controlling users who want to use these technologies against them, as well as business leaders who see potential to employ them to better their organization.

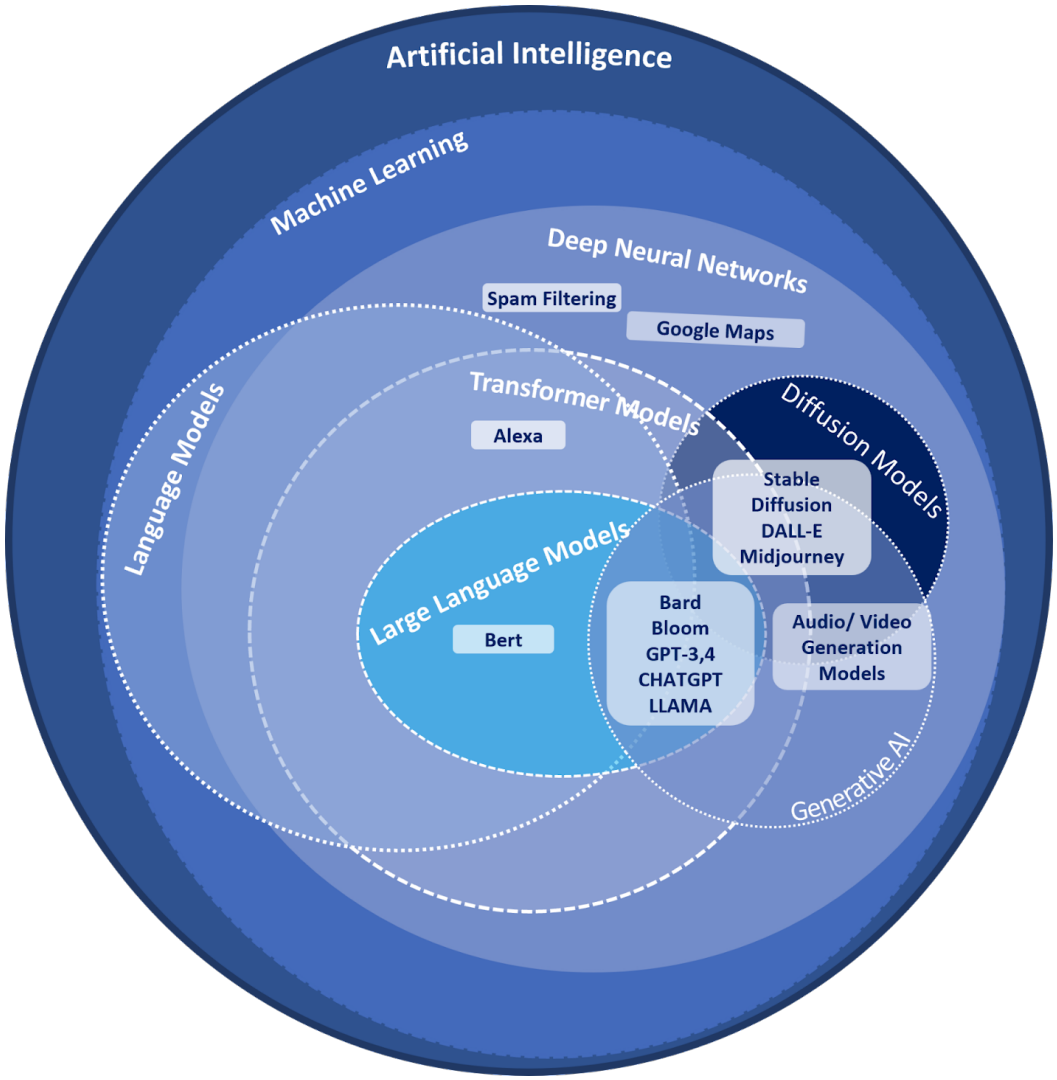

Figure 1.1: Image of LLM relationship within the field of Artificial Intelligence

Many applications within a business employ artificial intelligence applications, such as human resource hiring, SPAM detection for email, behavioral analytics for SIEM, and MDR apps. The primary focus of this document is on Large Language Model applications, which can produce content.

Responsible and Trustworthy Artificial Intelligence

As challenges and benefits of Artificial Intelligence emerge and regulations and laws are passed, the principles and pillars of responsible and trustworthy is evolving from ideal objects and concerns to future established standards. The OWASP AI Security and Privacy Guide working group is monitoring these changes and addressing the larger and more difficult considerations for all aspects of artificial intelligence.

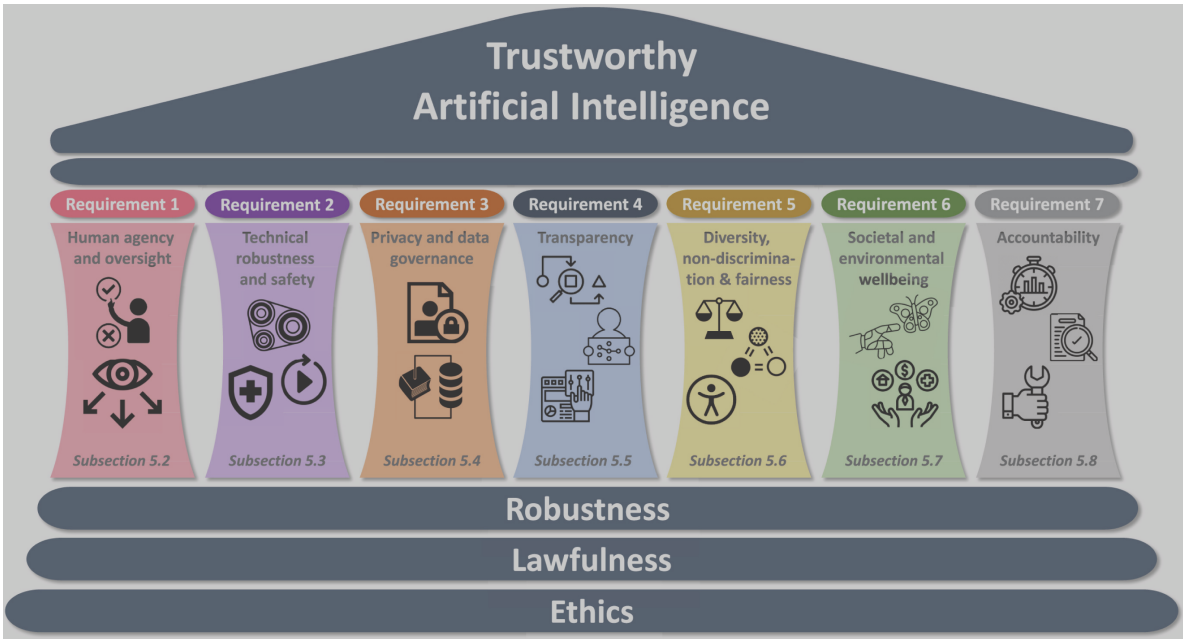

Figure 1.2: Image credit Montreal AI Ethics Institute

Large Language Model: Threat-Informed Defense

The purpose of the checklist is to help organizations understand the risks and benefits of using LLM by focusing on LLM threats that come from released models or services from third-party models. It will also include and improve existing resilience defensive techniques. Open source resources from both MITRE Engenuity and OWASP are referenced.

Figure 1.3: Image of integrating LLM Security with OWASP and MITRE resources

Why a Checklist?

Checklists can help with strategy development by ensuring thoroughness, clarifying goals, fostering consistency, and allowing for focused, deliberate effort, all of which may result in fewer oversights. Following the list can build confidence in a path to secure adoption while sparking ideas for future business cases moving forward. Itś a very forward and very practical way to achieve continuous improvement.

Not Comprehensive

This document is intended to support organizations in developing an initial LLM strategy in a rapidly changing technical, legal, and regulatory environment. Organizations should extend assessments and practices beyond the scope of the provided checklist.

Large Language Model Challenges

Large Language models face a number of serious and unique issues. One of the most important is that while working with LLMs, the control and data planes cannot be strictly isolated or separable. Another significant challenge is that LLMs are nondeterministic by design, yielding a different outcome when prompted or requested. It is not always a challenge, but LLMs employ semantic search rather than keyword search. The key distinction between the two is that the model's algorithm prioritizes the terms in its response. This is a significant departure from how consumers have traditionally used technology, and it has an impact on the consistency and reliability of the findings. Hallucinations, emerging from the gaps and training flaws in the data the model is trained on, are the result of this method.

There are methods to improve reliability and reduce the attack surface for jailbreaking, model tricking, and hallucinations, but there is a trade-off between restrictions and utility in both cost and functionality.

LLM use and applications increase an organization's attack surface. Some risks associated with LLMs are unique, but many are familiar issues, such as the known software bill of materials (SBOM), supply chain, data loss protection (DLP), and authorized access. There are also increased risks not directly related to GenAI, but GenAI increases the efficiency, capability, and effectiveness of attacks.

Adversaries are increasingly harnessing LLM and Generative AI tools to refine and expedite traditional methods. These enhanced techniques allow them to effortlessly craft new malware, potentially embedded with novel zero-day vulnerabilities or designed to evade detection. They can also generate sophisticated, unique, or tailored phishing schemes. The creation of convincing deep fakes, whether video or audio, further facilitates their social engineering ploys. Additionally, these tools enable them to execute intrusions and develop innovative hacking utilities. It is very likely that in the future, more "tailored" and compound use of AI technology by criminal actors will demand specific responses and dedicated solutions for appropriate defense schemas.

LLM Threat Categories

Figure 2.1: Image of types of AI threats

Artificial Intelligence Security and Privacy Training

Employees throughout organizations benefit from training to understand artificial intelligence, generative artificial intelligence, and the future potential consequences of building, buying, or utilizing LLMs. Training for permissible use and security awareness should target all employees as well as be more specialized for certain positions such as human resources, legal, developers, data teams, and security teams.

Fair use policies and healthy interaction are key aspects that, if incorporated from the very start, will be a cornerstone to the success of future AI cybersecurity awareness campaigns. This will necessarily imply the user's knowledge of the basic rules for interaction as well as the ability to separate good behavior from bad or unethical behavior.

Incorporate LLM Security and governance with Existing, Established Practices and Controls

While AI and generated AI add a new dimension to cybersecurity, resilience, privacy, and meeting legal and regulatory requirements, the best practices that have been around for a long time are still the best way to find risks, test them, fix them, and lower them.

- The management of artificial intelligence systems is integrated with existing organizational practices.
- Apply existing privacy, governance, and security practices.

Fundamental Security Principles

LLM capabilities introduce a different type of attack and attack surface. LLMs are vulnerable to complex business logic bugs, such as prompt injection, insecure plugin design, and remote code execution. Existing best practices are the best way to solve these issues. An internal product security team that understands secure software review, architecture, data governance, and third-party assessments The cybersecurity team should also check how strong the current controls are to find problems that could be made worse by LLM, like voice cloning, impersonation, or getting around captchas.

Accounting for the specific skills and competences developed in the last few years around machine learning, NLP and NLU, deep Learning and lately, LLMs and GenAI, it is advised to have skilled professionals with practice, knowledge, or experience in these fields to side with security teams in adopting, at best, and even shaping new potential analyses and responses to those issues.

Risk

Reference to risk uses the ISO 31000 definition: Risk = "effect of uncertainty on objectives." LLM risks included in the checklist include a targeted list of LLM risks that address adversarial, safety, legal, regulatory, reputation, financial, and competitive risks.

Vulnerability and Mitigation Taxonomy

Established methods of vulnerability classification and threat sharing are in early development, such as Oval, STIX, threat sharing, and vulnerability classification. The checklist anticipates calibrating with existing, established, and accepted standards, such as CVE classification.

Determining LLM Strategy

The acceleration of LLM applications has raised the visibility of all artificial intelligence applications' organizational use. Recommendations for policy, governance, and accountability should be considered holistically.

The immediate LLM threats are the use of online tools, browser plugins, third-party applications, the extended attack surface, and ways attackers can leverage LLM tools to facilitate attacks.

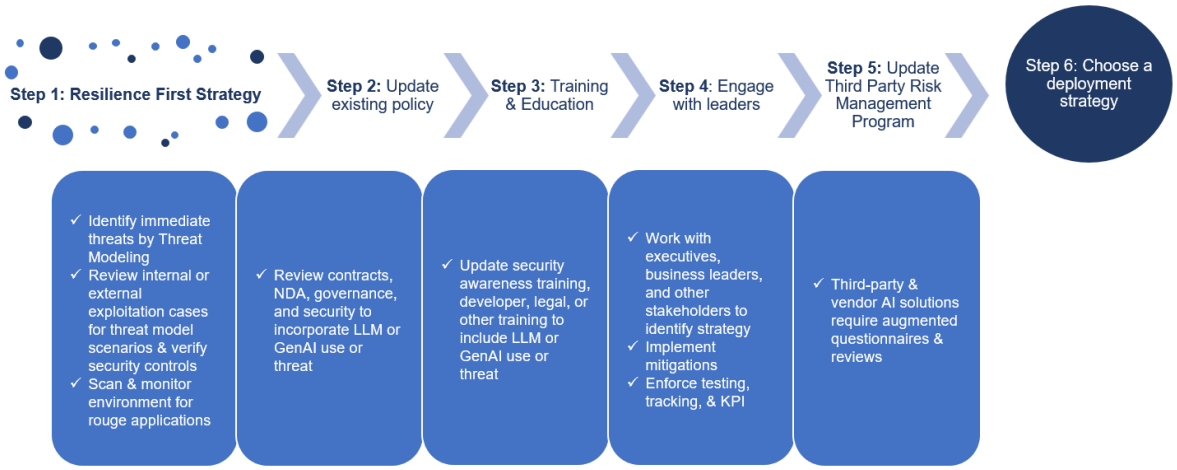

Figure 3.1: Image of steps of LLM implementation

Deployment Strategy

The scopes range from leveraging public consumer applications to training proprietary models on private data. Factors like use case sensitivity, capabilities needed, and resources available help determine the right balance of convenience vs. control. But understanding these five model types provides a framework for evaluating options.

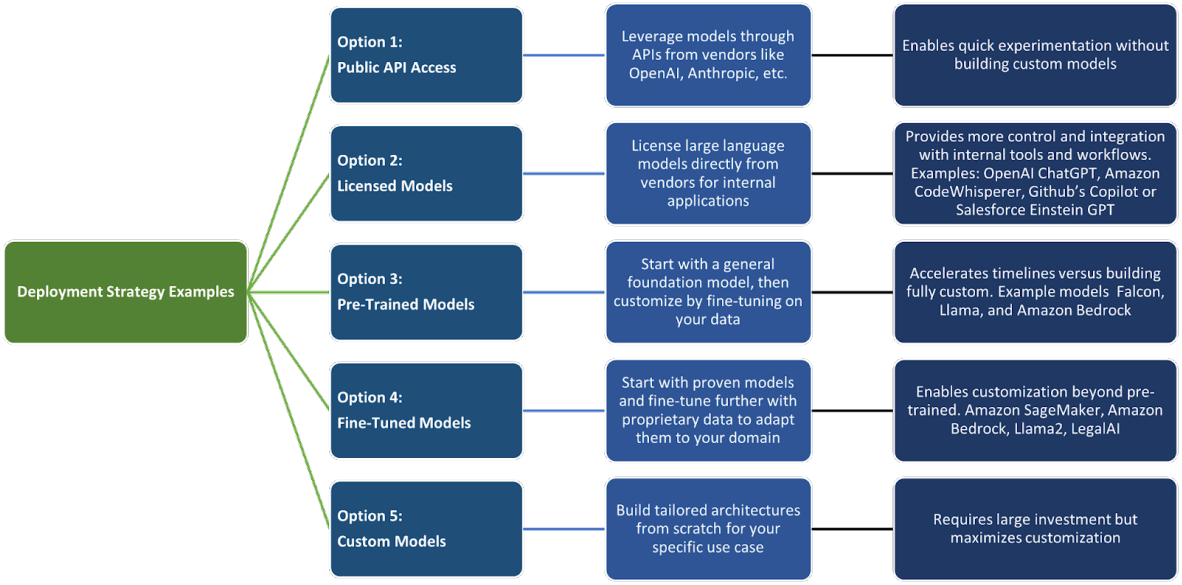

Figure 3.2: Image of options for deployment strategy

Check List

Adversarial Risk

Adversarial Risk includes competitors and attackers.

- ☐ Scrutinize how competitors are investing in artificial intelligence. Although there are risks in AI adoption, there are also business benefits that may impact future market positions.
- ☐ Threat Model: how attackers may accelerate exploit attacks against the organization, employees, executives, or users.
- ☐ Threat models potential attacks on customers or clients through spoofing and generative AI.
- ☐ Investigate the impact of current controls, such as password resets, which use voice recognition.
- ☐ Update the Incident Response Plan and playbooks for LLM incidents.

AI Asset Inventory

An AI asset inventory should apply to both internally developed and external or third-party solutions.

- ☐ Catalog existing AI services, tools, and owners. Designate a tag in asset management for specific inventory.
- ☐ Include AI components in the Software Bill of Material (SBOM), a comprehensive list of all the software components, dependencies, and metadata associated with applications.
- ☐ Catalog AI data sources and the sensitivity of the data (protected, confidential, public)
- ☐ Establish if pen testing or red teaming of deployed AI solutions is required to determine the current attack surface risk.
- ☐ Create an AI solution onboarding process.
- ☐ Ensure skilled IT admin staff is available either internally or externally, in accordance to the SBoM

AI Security and Privacy Training

- ☐ Train all users on ethics, responsibility, and legal issues such as warranty, license, and copyright.
- ☐ Update security awareness training to include GenAI related threats. Voice cloning and image cloning, as well as in anticipation of increased spear phishing attacks
- ☐ Any adopted GenAI solutions should include training for both DevOps and cybersecurity for the deployment pipeline to ensure AI safety and security assurances.

Establish Business Cases

Solid business cases are essential to determining the business value of any proposed AI solution,balancing risk and benefits, and evaluating and testing return on investment. There are an enormous number of potential use cases; a few examples are provided.

- ☐ Enhance customer experience
- ☐ Better operational efficiency
- ☐ Better knowledge management
- ☐ Enhanced innovation
- ☐ Market Research and Competitor Analysis
- ☐ Document creation, translation, summarization, and analysis

Governance

Corporate governance in LLM is needed to provide organizations with transparency and accountability. Identifying AI platform or process owners who are potentially familiar with the technology or the selected use cases for the business is not only advised but also necessary to ensure adequate reaction speed that prevents collateral damages to well established enterprise digital processes.

- ☐ Establish the organizationś AI RACI chart (who is responsible, who is accountable, who should be consulted, and who should be informed)
- ☐ Document and assign AI risk, risk assessments, and governance responsibility within the organization.
- ☐ Establish data management policies, including technical enforcement, regarding data classification and usage limitations. Models should only leverage data classified for the minimum access level of any user of the system. For example, update the data protection policy to emphasize not to input protected or confidential data into nonbusiness-managed tools.
- ☐ Create an AI Policy supported by established policy (e.g., standard of good conduct, data protection, software use)
- ☐ Publish an acceptable use matrix for various generative AI tools for employees to use.
- ☐ Document the sources and management of any data that the organization uses from the generative LLM models.

Legal

Many of the legal implications of AI are undefined and potentially very costly. An IT, security, and legal partnership is critical to identifying gaps and addressing obscure decisions.

- ☐ Confirm product warranties are clear in the product development stream to assign who is responsible for product warranties with AI.
- ☐ Review and update existing terms and conditions for any GenAI considerations.
- ☐ Review AI EULA agreements. End-user license agreements for GenAI platforms are very different in how they handle user prompts, output rights and ownership, data privacy, compliance and liability, privacy, and limits on how output can be used.
- ☐ Review existing AI-assisted tools used for code development. A chatbotś ability to write code can threaten a companyś ownership rights to its own product if a chatbot is used to generate code for the product. For example, it could call into question the status and protection of the generated content and who holds the right to use the generated content.
- ☐ Review any risks to intellectual property. Intellectual property generated by a chatbot could be in jeopardy if improperly obtained data was used during the generative process, which is subject to copyright, trademark, or patent protection. If AI products use infringing material, it creates a risk for the outputs of the AI, which may result in intellectual property infringement.
- ☐ Review any contracts with indemnification provisions. Indemnification clauses try to put the responsibility for an event that leads to liability on the person who was more at fault for it or who had the best chance of stopping it. Establish guardrails to determine whether the provider of the AI or its user caused the event, giving rise to liability.
- ☐ Review liability for potential injury and property damage caused by AI systems.
- ☐ Review insurance coverage. Traditional (D&O) liability and commercial general liability insurance policies are likely insufficient to fully protect AI use.
- ☐ Identify any copyright issues. Human authorship is required for copyright. An organization may also be liable for plagiarism, propagation of bias, or intellectual property infringement if LLM tools are misused.
- ☐ Ensure agreements are in place for contractors and appropriate use of AI for any development or provided services.
- ☐ Restrict or prohibit the use of generative AI tools for employees or contractors where enforceable rights may be an issue or where there are IP infringement concerns.
- ☐ Assess and AI solutions used for employee management or hiring could result in disparate treatment claims or disparate impact claims.
- ☐ Make sure the AI solutions do not collect or share sensitive information without proper consent or authorization.

Regulatory

The EU AI Act is anticipated to be the first comprehensive AI law but will apply in 2025 at the earliest. The EUś General Data Protection Regulation (GDPR) does not specifically address AI but includes rules for data collection, data security, fairness and transparency, accuracy and reliability, and accountability, which can impact GenAI use. In the United States, AI regulation is included within broader consumer privacy laws. Ten US states have passed laws or have laws that will go into effect by the end of 2023.

Federal organizations such as the US Equal Employment Opportunity Commission (EEOC), the Consumer Financial Protection Bureau (CFPB), the Federal Trade Commission (FTC), and the US Department of Justiceś Civil Rights Division (DOJ) are closely monitoring hiring fairness.

- ☐ Determine State specific compliance requirements.
- ☐ Determine compliance requirements for restricting electronic monitoring of employees and employment-related automated decision systems (Vermont)
- ☐ Determine compliance requirements for consent for facial recognition and the AI video analysis required (Illinois, Maryland)
- ☐ Review any AI tools in use or being considered for employee hiring or management.
- ☐ Confirm the vendorś compliance with applicable AI laws and best practices.
- ☐ Ask and document any products using AI during the hiring process. Ask how the model was trained, how it is monitored, and track any corrections made to avoid discrimination and bias.
- ☐ Ask and document what accommodation options are included.
- ☐ Ask and document whether the vendor collects confidential data.
- ☐ Ask how the vendor or tool stores and deletes data and regulates the use of facial recognition and video analysis tools during pre-employment.
- ☐ Review other organization-specific regulatory requirements with AI that may raise compliance issues. The Employee Retirement Income Security Act of 1974, for instance, has fiduciary duty requirements for retirement plans that a chatbot might not be able to meet.

Using or Implementing Large Language Model Solutions

- ☐ Threat Model: LLM components and architecture trust boundaries.
- ☐ Data Security: Verify how data is classified and protected based on sensitivity, including personal and proprietary business data. (How are user permissions managed, and what safeguards are in place?)
- ☐ Access Control: Implement least privilege access controls and implement defense-in-depth measures
- ☐ Training Pipeline Security: Require rigorous control around training data governance, pipelines, models, and algorithms.
- ☐ Input and Output Security: Evaluate input validation methods, as well as how outputs are filtered, sanitized, and approved.
- ☐ Monitoring and Response: Map workflows, monitoring, and responses to understand automation, logging, and auditing. Confirm audit records are secure.
- ☐ Include application testing, source code review, vulnerability assessments, and red teaming in the production release process.
- ☐ Consider vulnerabilities in the LLM model solutions (Rezilion OSFF Scorecard).
- ☐ Look into the effects of threats and attacks on LLM solutions, such as prompt injection, the release of sensitive information, and process manipulation.
- ☐ Investigate the impact of attacks and threats to LLM models, including model poisoning, improper data handling, supply chain attacks, and model theft.
- ☐ Supply Chain Security: Request third-party audits, penetration testing, and code reviews for third-party providers. (both initially and on an ongoing basis)
- ☐ Infrastructure Security: How often does the vendor perform resilience testing? What are their SLAs in terms of availability, scalability, and performance?
- ☐ Update incident response playbooks and include an LLM incident in tabletop exercises.
- ☐ Identify or expand metrics to benchmark generative cybersecurity AI against other approaches to measure expected productivity improvements.

Resources

OWASP Top 10 for Large Language Model Applications

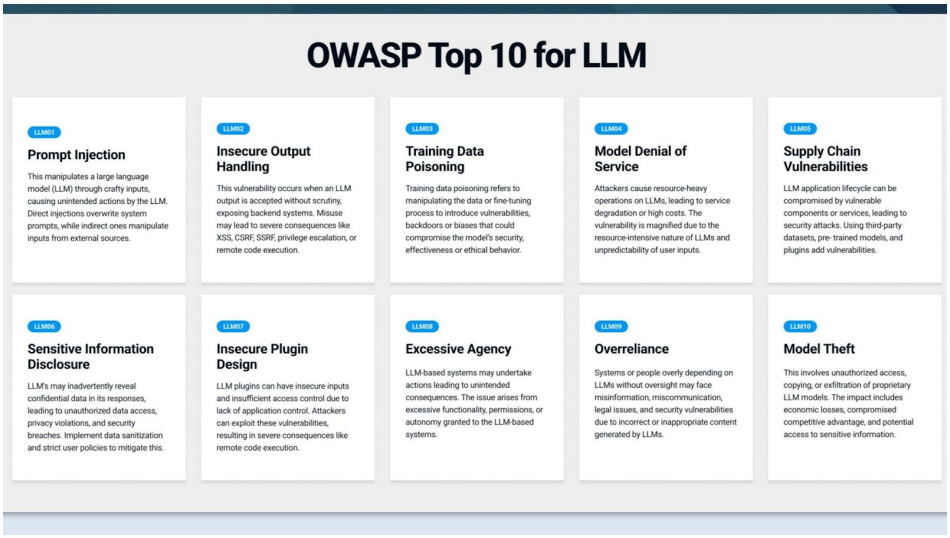

Figure 5.1: Image of OWASP Top 10 for Large Language Model Applications

OWASP Top 10 for Large Language Model Applications Visualized

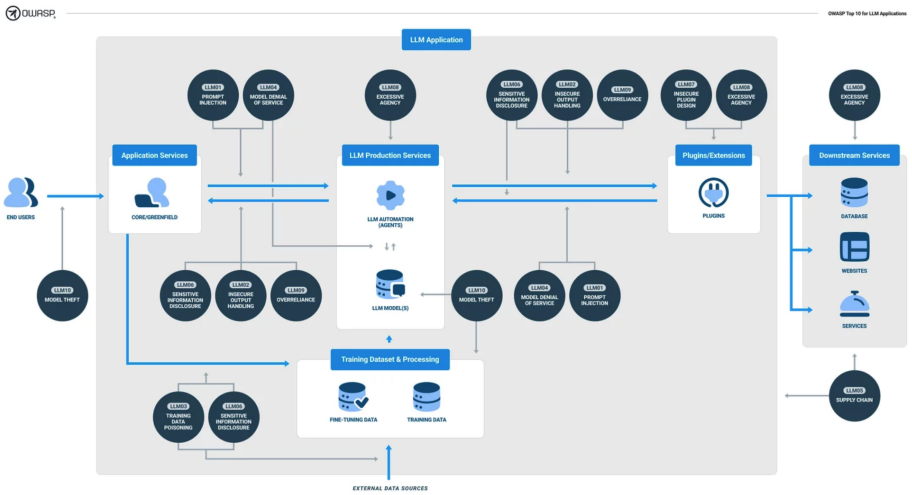

Figure 5.2: Image of OWASP Top 10 for Large Language Model Applications Visualized

OWASP Resources

Using LLM solutions expands an organization's attack surface and presents new challenges, requiring special tactics and defenses. It also poses problems that are similar to known issues, and there are already established cybersecurity procedures and mitigations. Integrating LLM cybersecurity with an organization's established cybersecurity controls, processes, and procedures allows an organization to reduce its vulnerability to threats. How they integrate with each other is available at the OWASP Integration Standards.

OWASP Resource	Description	Why It Is Recommended & Where To Use It
OWASP SAMM	Software Assurance Maturity Model	Provides an effective and measurable way to analyze and improve an organization's secure development lifecycle. SAMM supports the complete software lifecycle. It is interative and risk-driven, enabling organizations to identify and prioritize gaps in secure software development so resources for improving the process can be dedicated where efforts have the greatest improvement impact.
OWASP AI Security and Privacy Guide	Software Assurance Maturity Model	Provides an effective and measurable way to analyze and improve an organization's secure development lifecycle. SAMM supports the complete software lifecycle. It is interative and risk-driven, enabling organizations to identify and prioritize gaps in secure software development so resources for improving the process can be dedicated where efforts have the greatest improvement impact.

OWASP Resource	Description	Why It Is Recommended & Where To Use It
OWASP SAMM	Software Assurance Maturity Model	Provides an effective and measurable way to analyze and improve an organization's secure development lifecycle. SAMM supports the complete software lifecycle. It is interative and risk-driven, enabling organizations to identify and prioritize gaps in secure software development so resources for improving the process can be dedicated where efforts have the greatest improvement impact.
OWASP SAMM	Software Assurance Maturity Model	Provides an effective and measurable way to analyze and improve an organization's secure development lifecycle. SAMM supports the complete software lifecycle. It is interative and risk-driven, enabling organizations to identify and prioritize gaps in secure software development so resources for improving the process can be dedicated where efforts have the greatest improvement impact.
OWASP SAMM	Software Assurance Maturity Model	Provides an effective and measurable way to analyze and improve an organization's secure development lifecycle. SAMM supports the complete software lifecycle. It is interative and risk-driven, enabling organizations to identify and prioritize gaps in secure software development so resources for improving the process can be dedicated where efforts have the greatest improvement impact.

OWASP Resource	Description	Why It Is Recommended & Where To Use It
OWASP SAMM	Software Assurance Maturity Model	Provides an effective and measurable way to analyze and improve an organization's secure development lifecycle. SAMM supports the complete software lifecycle. It is interative and risk-driven, enabling organizations to identify and prioritize gaps in secure software development so resources for improving the process can be dedicated where efforts have the greatest improvement impact.
OWASP SAMM	Software Assurance Maturity Model	Provides an effective and measurable way to analyze and improve an organization's secure development lifecycle. SAMM supports the complete software lifecycle. It is interative and risk-driven, enabling organizations to identify and prioritize gaps in secure software development so resources for improving the process can be dedicated where efforts have the greatest improvement impact.
OWASP SAMM	Software Assurance Maturity Model	Provides an effective and measurable way to analyze and improve an organization's secure development lifecycle. SAMM supports the complete software lifecycle. It is interative and risk-driven, enabling organizations to identify and prioritize gaps in secure software development so resources for improving the process can be dedicated where efforts have the greatest improvement impact.

OWASP Resource	Description	Why It Is Recommended & Where To Use It
OWASP SAMM	Software Assurance Maturity Model	Provides an effective and measurable way to analyze and improve an organization's secure development lifecycle. SAMM supports the complete software lifecycle. It is interative and risk-driven, enabling organizations to identify and prioritize gaps in secure software development so resources for improving the process can be dedicated where efforts have the greatest improvement impact.
OWASP SAMM	Software Assurance Maturity Model	Provides an effective and measurable way to analyze and improve an organization's secure development lifecycle. SAMM supports the complete software lifecycle. It is interative and risk-driven, enabling organizations to identify and prioritize gaps in secure software development so resources for improving the process can be dedicated where efforts have the greatest improvement impact.

Table 5.1: OWASP Resources

MITRE Resources

The increased frequency of LLM threats emphasizes the value of a resilience-first approach to defending an organization's attack surface. Existing TTPS are combined with new attack surfaces and capabilities in LLM Adversary threats and mitigations. MITRE maintains a well-established and widely accepted mechanism for coordinating opponent tactics and procedures based on real-world observations.

Coordination and mapping of an organization's LLM Security Strategy to MITRE ATT&CK and MITRE ATLAS allows an organization to determine where LLM Security is covered by current processes such as API Security Standards or where security holes exists.

MITRE ATT&CK (Adversarial Tactics, Techniques, and Common Knowledge) is a framework, collection of data matrices, and assessment tool that was made by the MITRE Corporation to help organizations figure out how well their cybersecurity works across their entire digital attack surface and find holes that had not been found before. It is a knowledge repository that is used all over the world. The MITRE ATT&CK matrix contains a collection of strategies used by adversaries to achieve a certain goal. In the ATT&CK Matrix, these objectives are classified as tactics. The objectives are outlined in attack order, beginning with reconnaissance and progressing to the eventual goal of exfiltration or impact.

MITRE ATLAS, which stands for "Adversarial Threat Landscape for Artificial Intelligence Systems," is a knowledge base that is based on real-life examples of attacks on machine learning (ML) systems by bad actors. ATLAS is based on the MITRE ATT&CK architecture, and its tactics and procedures complement those found in ATT&CK.

MITRE Resource	Description	Why It Is Recommended & Where To Use It
Res1	Des1	Why1

Table 5.2: OWASP Resources

AI Vulnerability Repositories

Name	Description
Res1	Des1

Table 5.3: AI Vulnerability Repositories

AI Procurement Guidance

Name	Description
Res1	Des1

Table 5.4: AI Procurement Guidance

Team

Thank you to the OWASP Top 10 for LLM Applications Cybersecurity and Governance Checklist .05 contributors.

Sandy Dunn	Heather Linn	John Sotiropoulos
Steve Wilson	Fabrizio Cilli	Aubrey King
Bob Simonoff	David Rowe	Rob Vanderveer
Emmanual Guilherme Junior	Andrea Succi	Jason Ross

Table A.1: OWASP LLM AI Security & Governance Checklist v.0.5 Team