Trend Research

TREND MICRO

SECURITY FOR AI BLUEPRINT

A step-by-step guide for introducing cybersecurity to your AI application innovations

FERNANDO CARDOSO
Director of Product Management

EXECUTIVE SUMMARY

In the rapidly evolving landscape of AI, cybersecurity professionals face the critical challenge of safeguarding business applications and employee use of AI technologies. As artificial intelligence becomes integrated across organizations, it is essential to develop robust security strategies that address the unique risks posed by AI usage, ensuring both protection and visibility at every level of the enterprise.

According to McKinsey, cybersecurity is one of the top three risks associated with the use of Generative AI. This highlights a key issue: many organizations are either restricting or blocking AI initiatives due to security concerns. A comprehensive cybersecurity strategy, however, can empower businesses to mitigate these risks, fostering innovation and enabling the safe deployment of AI technologies without unnecessary friction.

To address these challenges, we collaborated with organizations globally to identify common obstacles in protecting AI systems. Our findings have led to the creation of a strategic blueprint that outlines best practices for securing critical AI applications.

By understanding the AI attack surface and conducting threat modeling specifically for Large Language Models (LLMs), we have developed a six-layer cybersecurity framework designed to defend against the most prevalent threats targeting AI-driven systems.

This blueprint provides organizations with actionable insights to safeguard their AI applications, ensuring they can leverage AI's full potential while protecting both business operations and customer trust.

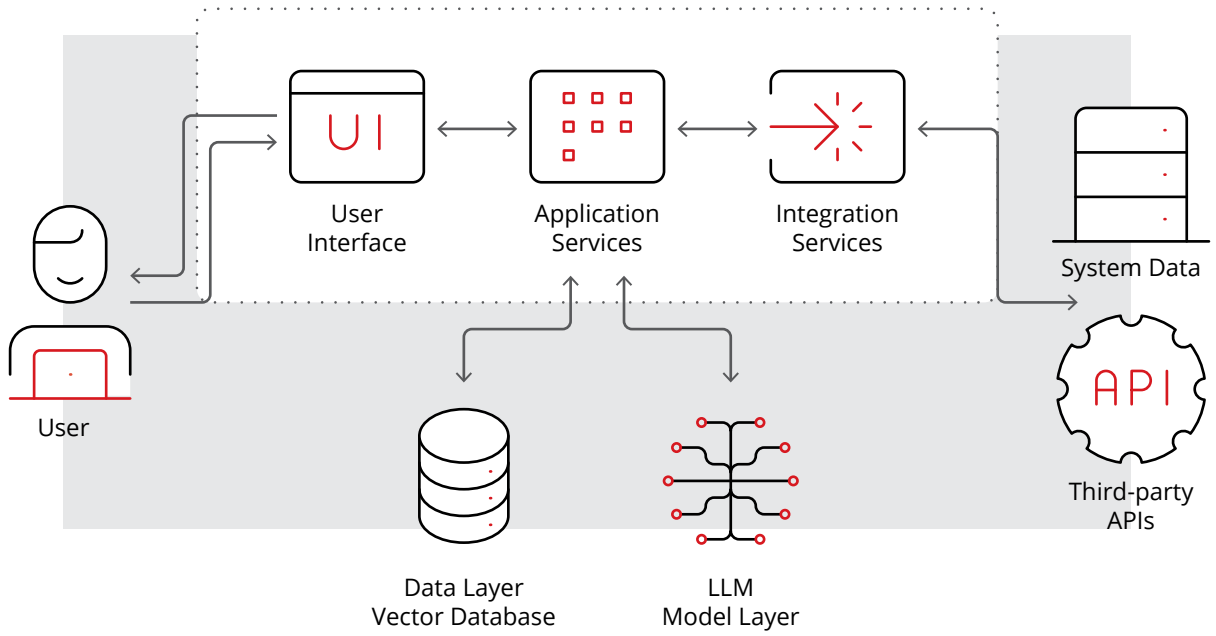
TABLE OF CONTENTS

WHAT IS AN LLM APPLICATION?

Large Language Model (LLM) applications are defined by one or more of the following capabilities:

General Knowledge and Thinking: Achieved through extensive model training on diverse datasets

- **Specific Domain Expertise:** Developed through fine-tuning on specialized datasets for specific tasks or domains
- **Specific Use Case Data:** Developed by integrating reference data from databases or external sources, often using Retrieval-Augmented Generation (RAG) techniques



AI Application Framework

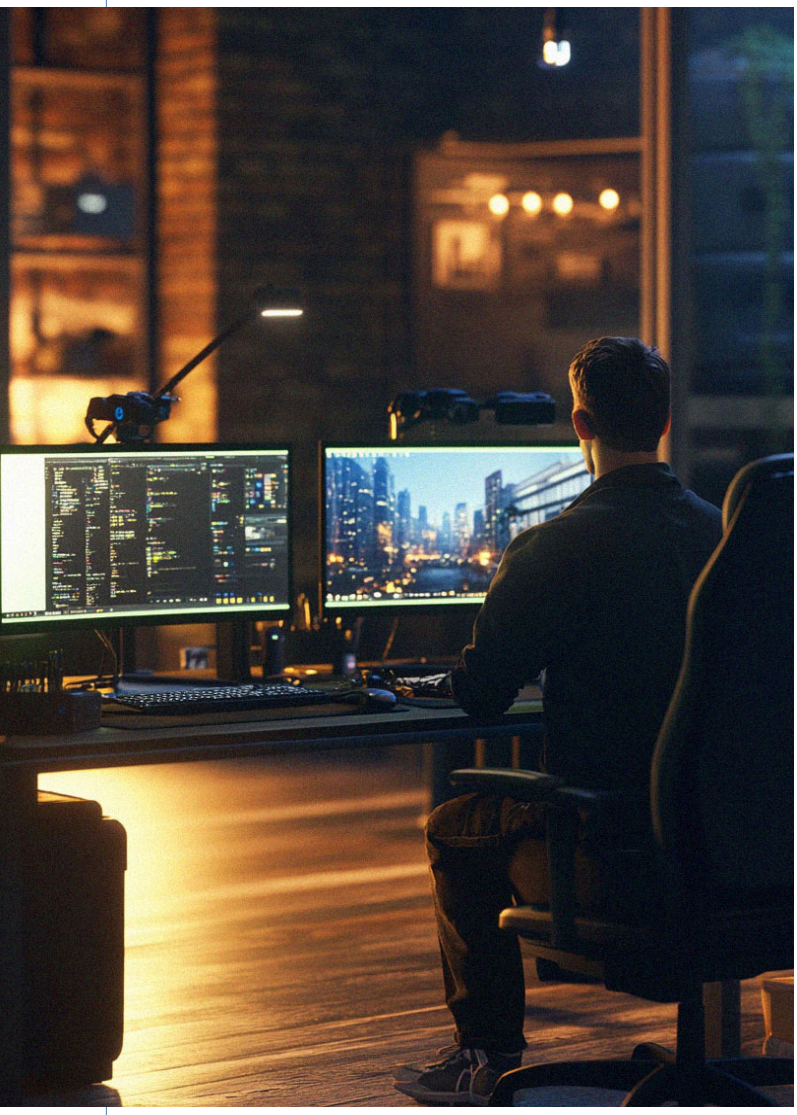
LLM applications also employ prompt templates to guide the LLM's responses with real input data and predefined instructions. Utilizing these capabilities and prompt templates, developers can create sophisticated, context-aware, and reliable LLM applications. LLMs are trained on vast quantities of text data and can understand and generate human-like text. Applications can range from chatbots and virtual assistants to content generation, language translation, sentiment analysis, and more. Developers and researchers use LLMs in various fields like education, healthcare, customer service and cybersecurity, among others.

Application Services

The Application Services component handles all common operations and specific application logic in an LLM application. It manages user authentication, session management, interaction logging and API integrations. This layer orchestrates tasks, routes requests and executes business logic tailored to the application's needs. By ensuring seamless interaction between LLM agents and other components, including databases and external data sources, Application Services provide a foundation for efficient, secure and flexible application operation.

LLM Agents

The LLM Agents component functions as a processor to handle input and output between the LLM and other components. It uses prompt templates to structure and format incoming data, ensuring the LLM processes it accurately. These agents also manage the output generated by the LLM, adapting it to fit the application's specific requirements and user context. By acting as an intermediary, LLM Agents streamline interactions, optimize the model's performance and ensure seamless communication within the application.



Large Language Model (LLM)

The LLM component is the core engine that drives the application's capabilities. It can be an external LLM provider, such as OpenAI's GPT series or Google's Gemini, or a self-hosted LLM customized to specific needs. Trained on extensive datasets, the LLM provides the fundamental ability to understand and generate human-like text. This component leverages general knowledge and specific domain expertise, acquired through model training and fine-tuning, to deliver accurate and contextually relevant responses. By processing structured inputs from the Prompt Template and interacting with various data sources, the LLM is essential for delivering intelligent and dynamic user interactions within the application.

HOW TO INTEGRATE SECURITY INTO AI APPLICATION ARCHITECTURES?

How do I integrate security into my AI application architecture? This is one of the most common questions from organizations and security teams—primarily because security functions are often only engaged by innovation and AI teams late in the development cycle. By the time this stage of the project is reached, AI training models have already been executed, the AI application is running in some internal and external tests, and in some cases, it's already gone to production. Implementing security at this stage can be complicated for security teams—who must understand all the phases in the AI pipeline process and AI applications to better recommend how to implement security layers and helping minimize any risks.

Security is often overlooked during development of AI applications and architecture, due to three key challenges:

The development team and AI engineers are united in their mission to build applications that empower businesses to innovate, grow and deliver maximum value to their customers.

1. Limited understanding of AI models, applications and architecture among security teams.
2. Pressure on business units to rapidly deliver innovative AI applications and fixes often leads development teams to overlook secure design principals into AI architectures and pipelines.
3. Security teams are still uncertain about which layers of protection are necessary for AI applications because of the missing risk exposure and impact visibility in their environments, creating confusion for some companies.

The main challenges to integrating security and how to overcome them

Today, many security teams have limited expertise in programming languages such as Ruby, Go, Node.js, Java, and Python. Additionally, some team members feel uncomfortable with concepts such as LLMs, RAG, and AI model training.

Historically, when development and innovation teams were less prevalent, security practitioners weren't responsible for verifying security processes within the development pipeline. Consequently, code security has largely depended on developers' backgrounds and the best practices they follow. In most cases, developers enjoy significant autonomy in their processes, focusing on delivering results that meet business expectations.

However, recent cultural shifts, such as the adoption of DevSecOps, are prompting organizations to prioritize cybersecurity across all areas, rather than focusing solely on isolated environments. This heightened awareness is essential for ensuring the security of AI innovations and applications under development.

> Embedding cybersecurity in the early stages of AI development is essential for safeguarding business integrity, reducing risk and driving future innovation.

These changes have ushered in a new suite of tools designed to enhance security maturity and incident response within modern AI infrastructures. It won't be long before the market and security teams fully embrace this new reality.

To effectively integrate cybersecurity into the AI application and model-building process, organizations should invest in cross-functional teams that bridge the gap between developers and security professionals. This collaboration is essential for cultivating a DevSecOps culture that prioritizes security at every stage of the AI development lifecycle.

As AI models become more integral to business operations, understanding potential risks and vulnerabilities is crucial. Security teams should be trained in the nuances of machine learning frameworks and data handling practices, while developers need insights into secure design principals or coding practices to mitigate risks during model training and deployment. By fostering this mutual understanding and responsible AI culture, organizations can better secure their AI applications against threats such as adversarial attacks and data breaches.

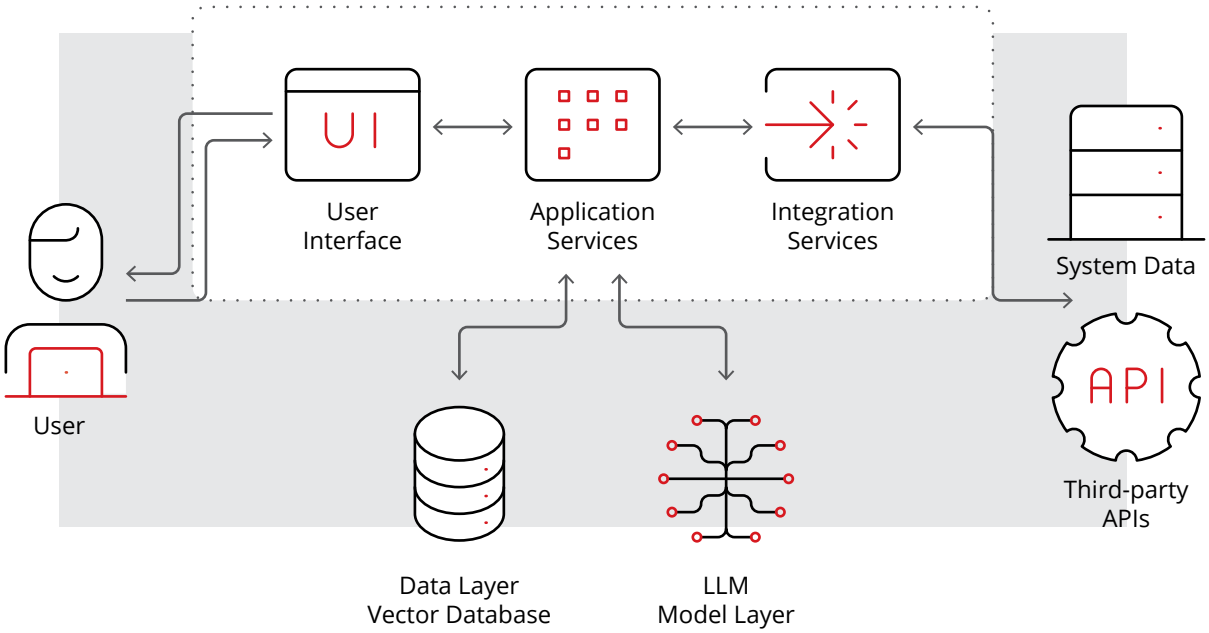
In today's landscape, where every organization is becoming software-driven, a robust approach to security is non-negotiable. Organizations that fail to adapt may face critical challenges, as adversaries grow more sophisticated and regulations become more widespread. Engaging with the open-source community—where many cutting-edge security practices are shared and developed collaboratively—can also provide valuable resources and insights.

In summary, integrating cybersecurity into the AI application lifecycle is not just a technical necessity, it's a strategic imperative.

THE AI ATTACK SURFACE IN ACTION

The attack surface of an AI system refers to the various points where an attacker can attempt to compromise the system. Here are some examples of potential attack surfaces in AI applications:

1. **Data Inputs:** AI models are heavily reliant on data. Attackers can manipulate input data to produce incorrect outputs—a technique known as adversarial attacks. For instance, altering pixels in an image could cause a model to misclassify the image entirely.
2. **Model Training:** During the training phase, if an attacker gains access to the training dataset, they could introduce biased or malicious data, leading to a compromised model. This is known as data poisoning.
3. **APIs and Interfaces:** If an AI application exposes APIs for interaction, these can be exploited. For example, attackers might try to send malformed requests or use injection techniques to manipulate how the AI processes data.



AI Application Framework

4. **Model Deployment:** Once the model is deployed, attackers can attempt to extract sensitive information from it through techniques like model extraction, where they try to recreate the model by querying it repeatedly to understand its behavior.

5. **User Interfaces:** The front-end applications that utilize AI models can also be vulnerable. If a user interface is not properly secured, attackers might exploit it to gain unauthorized access or inject malicious scripts.

6. **Third-Party Libraries:** Many AI applications rely on third-party libraries and frameworks. If these dependencies have vulnerabilities, they can be exploited to compromise the entire application.

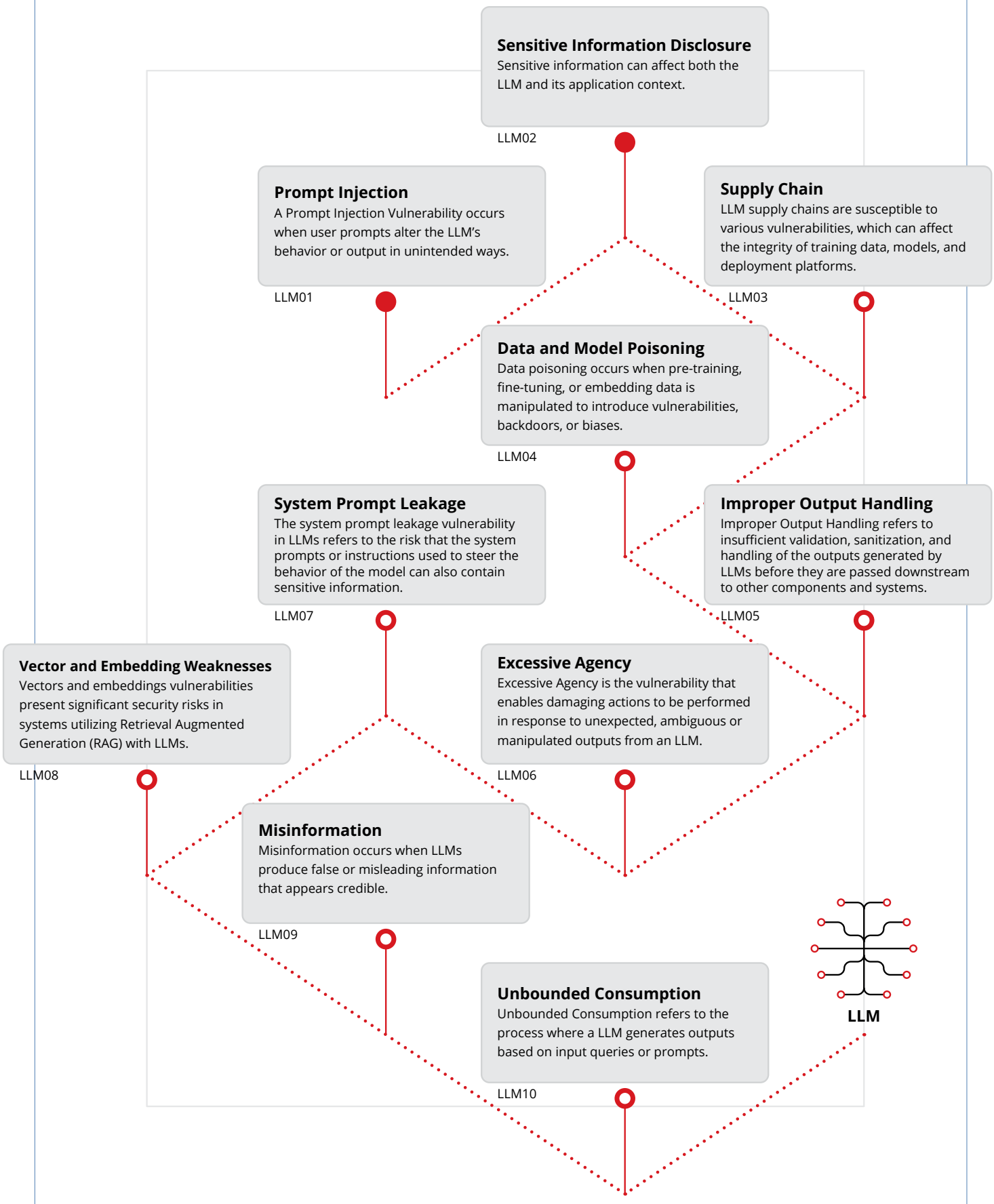
7. **Cloud Infrastructure:** If the AI model is hosted on cloud platforms, misconfigured cloud settings can expose it to unauthorized access or attacks, such as denial-of-service (DoS) attacks.

By understanding these attack surfaces, organizations can implement better security measures to protect their AI systems throughout the development and deployment lifecycle.

THREAT MODELLING FOR LLM

The **OWASP Top 10 for Large Language Model Applications** project seeks to educate and inform developers, organizations, and stakeholders about the unique security challenges associated with deploying and managing Large Language Models (LLMs) and Generative AI technologies. The project provides valuable resources, including a comprehensive list of the **Top 10 vulnerabilities** frequently encountered in LLM-based applications. These vulnerabilities are assessed based on their potential impact, ease of exploitation, and prevalence in real-world scenarios.

Key risks covered include issues such as **prompt injection, data leakage, inadequate sandboxing, and unauthorized code execution**, among others. The primary goal is to raise awareness of these threats, recommend practical mitigation strategies, and enhance the overall security resilience of LLM-driven systems. By addressing these vulnerabilities, the project aims to foster safer AI development practices and better protect both users and organizations from emerging risks in the rapidly evolving AI landscape.



OWASP TOP 10 for LLMs

A BLUEPRINT FOR SECURING AI

The first step is to secure your data. Data is the backbone of artificial intelligence, driving its ability to learn, adapt, and make informed decisions—so it is crucial to safeguard this data. Next, you must secure your AI models. This is essential to protect sensitive data, ensure reliable performance, and maintain trust in distributed systems by mitigating risks associated with data breach and vulnerability risks.

As part of the shared responsibility model, it is also your job to secure your AI infrastructure in the cloud—ensuring you have visibility of any misconfigurations that could make you vulnerable.

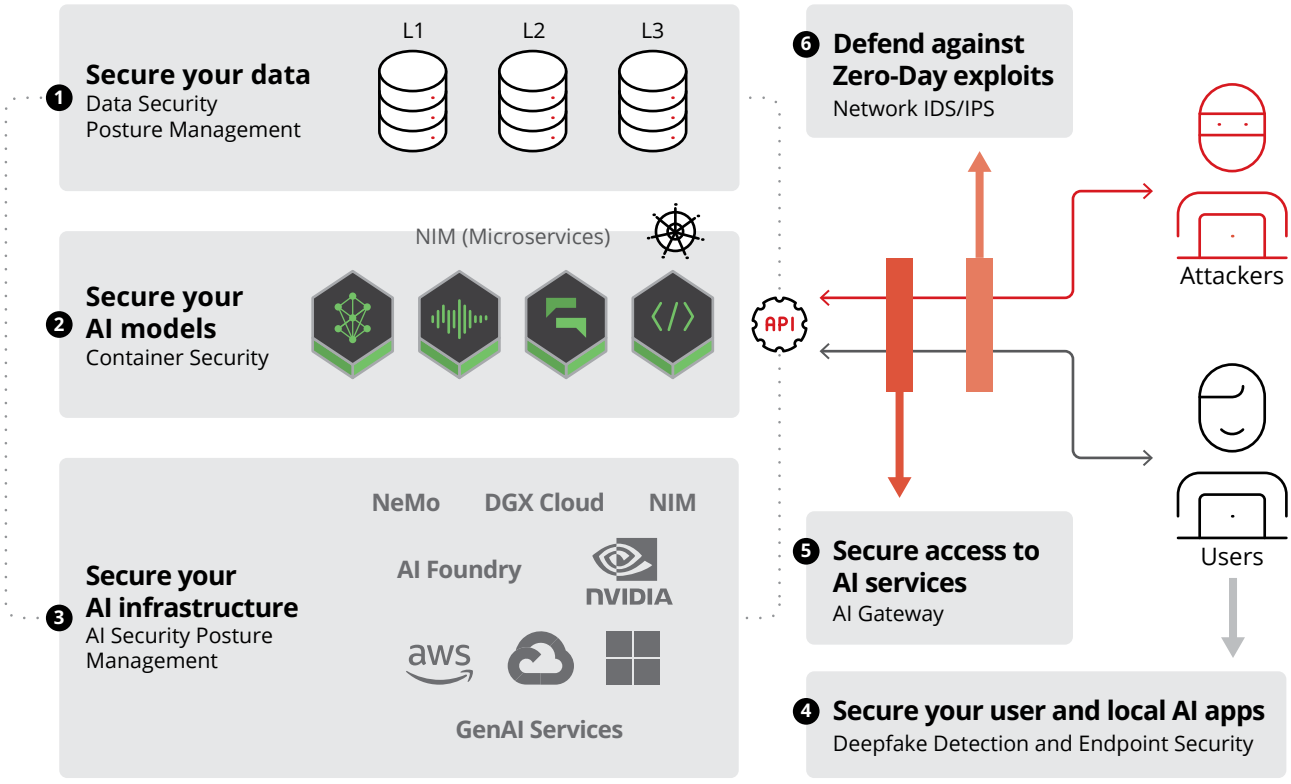
You'll also need to secure your users and AI applications from malicious uses of AI, such as deepfakes, as well as prevent suspicious programs, including malware and ransomware, from manipulating your AI applications.

Of course, managing employee access to AI services and filtering their prompts.

The 6 Steps to Securing AI Applications:

Innovate boldly but secure responsibly, ensuring that trust and risk management keep pace with AI's accelerating impact.

1. Secure your data
2. Secure your AI models
3. Secure your AI infrastructure
4. Secure your users and local AI apps
5. Secure access to AI services
6. Defend against 0-day exploits



Security for AI Blueprint

Let's take a closer look at how you can secure your AI innovations.

1 - Secure Your Data

Today's data landscape is complex and securing your data is crucial. One of the primary challenges organizations face in doing so is an absence of proper classification. Without clear classification, sensitive information is at risk of being mishandled or exposed. Data Security Posture Management (DSPM) helps by systematically identifying and categorizing data, ensuring that information is adequately protected based on its sensitivity and regulatory requirements. Here are some other reasons why DSPM is critical to data security:

Addressing Data Classification:

- Lack of proper data classification can lead to mishandling or exposure of sensitive information.
- DSPM systematically identifies and categorizes data, ensuring adequate protection based on sensitivity and regulatory requirements.

Managing Public AI Services:

- Public AI services can inadvertently expose sensitive data.
- DSPM enables the implementation of strict access controls and monitoring of data interactions, reducing the risk of unauthorized access.

Protecting Sensitive Data in AI Models:

- Sensitive data is often used in AI models, necessitating secure handling.
- DSPM uses methods to validate sensitive data before training AI models or generating database vectors from an organization's source knowledge, helping to mitigate data leakage risks.

Preventing Data Exfiltration:

- DSPM continuously monitors for sensitive data while providing visibility into potential attack paths. By correlating this information, you can better understand critical risks, enabling real-time identification of potential data breaches.
- Quick responses to detected anomalies guarantee that only authorized personnel can access sensitive information.

2 - Secure Your AI Models

Protecting AI models is crucial for several reasons:

Intellectual Property Protection

- AI models often represent significant investments in research and development. If compromised, attackers can steal proprietary algorithms or sensitive data.

Data Integrity and Trust

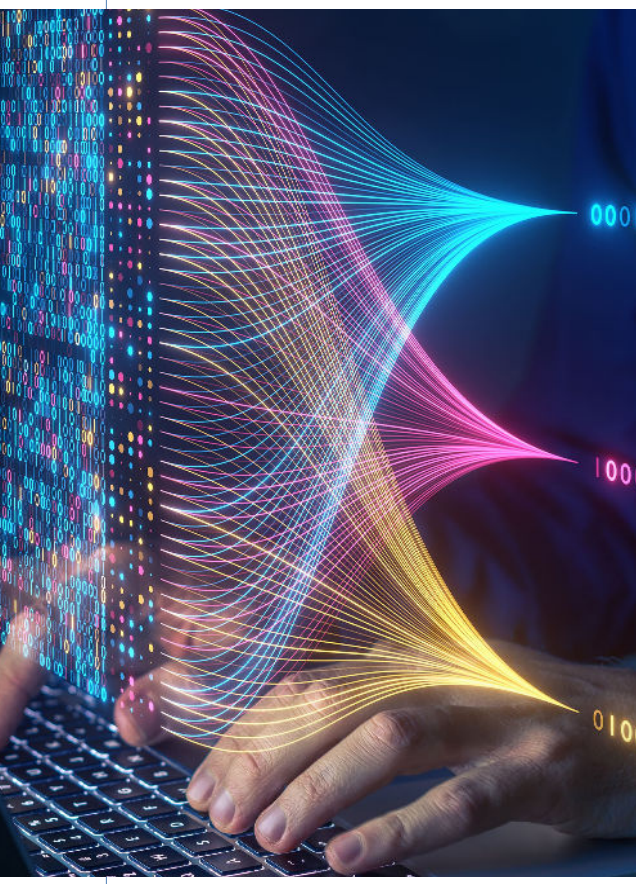
- Manipulated models can produce incorrect outputs, leading to faulty decisions. Ensuring the integrity of AI models is essential to maintain trust in AI-driven applications.

Regulatory Compliance

- Many industries are subject to data protection and privacy regulations. Safeguarding AI models helps organizations comply with these requirements.

Operational Continuity

- Attacks on AI models can disrupt services and lead to downtime, impacting business operations and customer experience.



AI models are increasingly deployed in containerized environments due to their scalability, flexibility and ease of integration. However, these containers can also become prime targets for attack. To protect AI models running in containers, implement robust container security measures such as:

Container Security Runtime

- **Real-time Monitoring:** Use container security tools that monitor runtime behavior to detect anomalies and potential threats in real-time. This helps identify suspicious activities that could indicate a breach.

Vulnerability Scanning:

- **Regular Scans:** Conduct vulnerability scans on your container images before deployment and during runtime. This identifies known vulnerabilities that could be exploited.

Network Segmentation

- **Isolate Containers:** Use network segmentation to isolate containers running AI models from other parts of your infrastructure. This limits the potential attack surface and contains breaches

Logging and Incident Response (through XDR)

- **Comprehensive Logging:** Enable detailed logging of container activity. This helps in forensic analysis in case of an incident, and improves your overall security posture.

3 - Secure Your AI Infrastructure

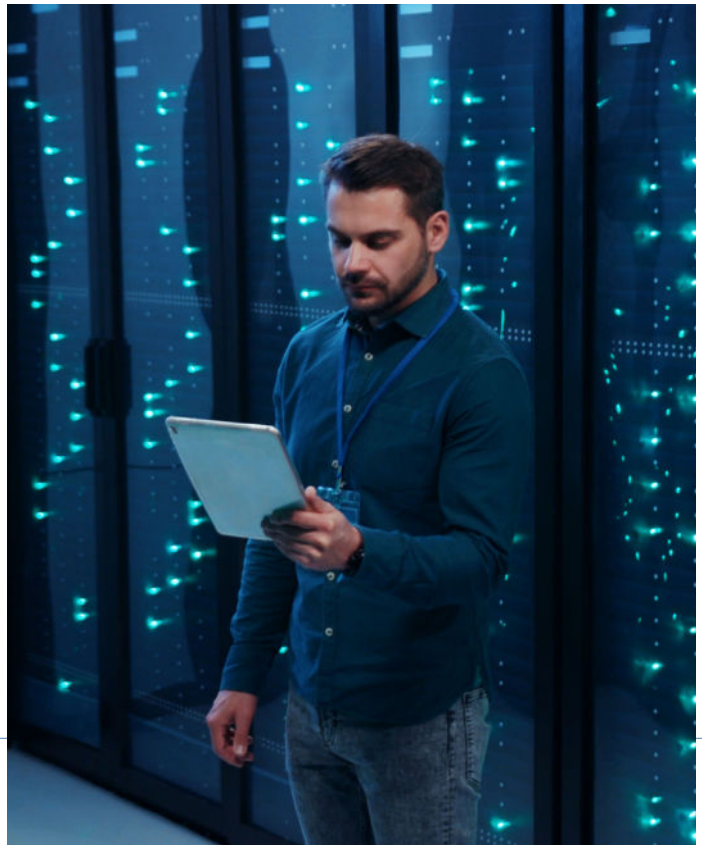
Leveraging AI security posture management (AI-SPM) is a great way to help maintain the security, compliance and integrity of AI infrastructure. It enables you to proactively gain visibility into and mitigate risks, protect your assets, and foster trust in your internal and external AI applications. Let's examine AI-SPM in more detail:

Visibility into Infrastructure

- It can provide comprehensive visibility into your AI infrastructure to understand the multiple components, configurations and data flows within your cloud environments. Enhanced visibility helps identify and manage risks effectively.

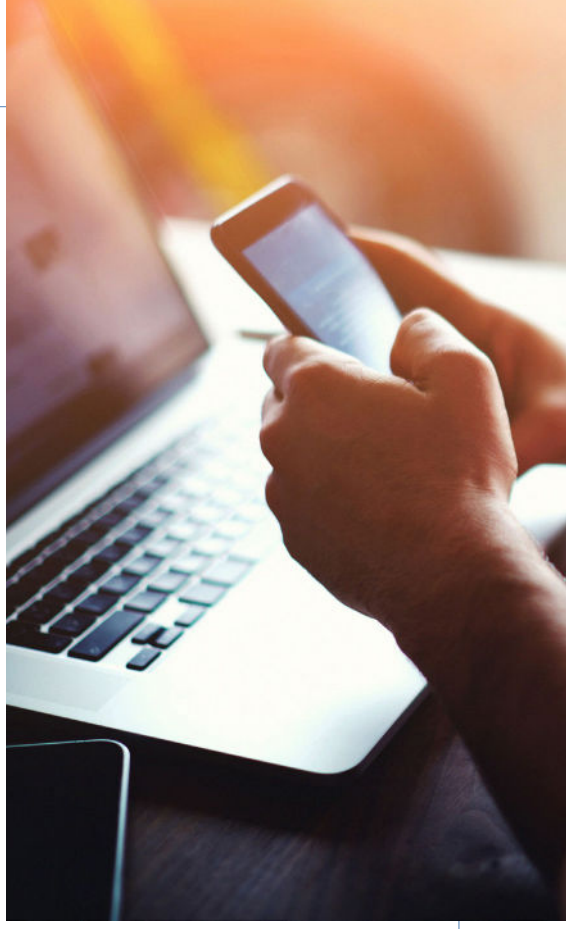
Proactive Risk Management

- By continuously monitoring and assessing the security posture of AI infrastructure, AI-SPM allows organizations to adopt a proactive approach to risk management, rather than a reactive strategy.



Identifying Misconfigurations

- Misconfigurations can lead to significant security vulnerabilities. AI-SPM combines with the Cloud Secure Posture Management (CSPM) to help detect these issues early, enabling teams to rectify them before they can be exploited by malicious actors.



Vulnerability Management

- AI applications often rely on complex software stacks, which can introduce vulnerabilities. AI-SPM continuously scans for known vulnerabilities and provides insights on how to address them. This ensures your applications are less prone to attacks exploiting potentially hidden vulnerabilities.

Compliance Monitoring

- Many organizations must adhere to industry standards and regulations (e.g., GDPR, HIPAA, PCI-DSS). AI-SPM combined with CSPM and Kubernetes Secure Posture Management (KSPM) helps monitor compliance violations, ensuring that AI applications meet necessary legal and regulatory requirements, thereby reducing the risk of fines and reputational damage.

4 - Secure Your Users and Local AI Apps

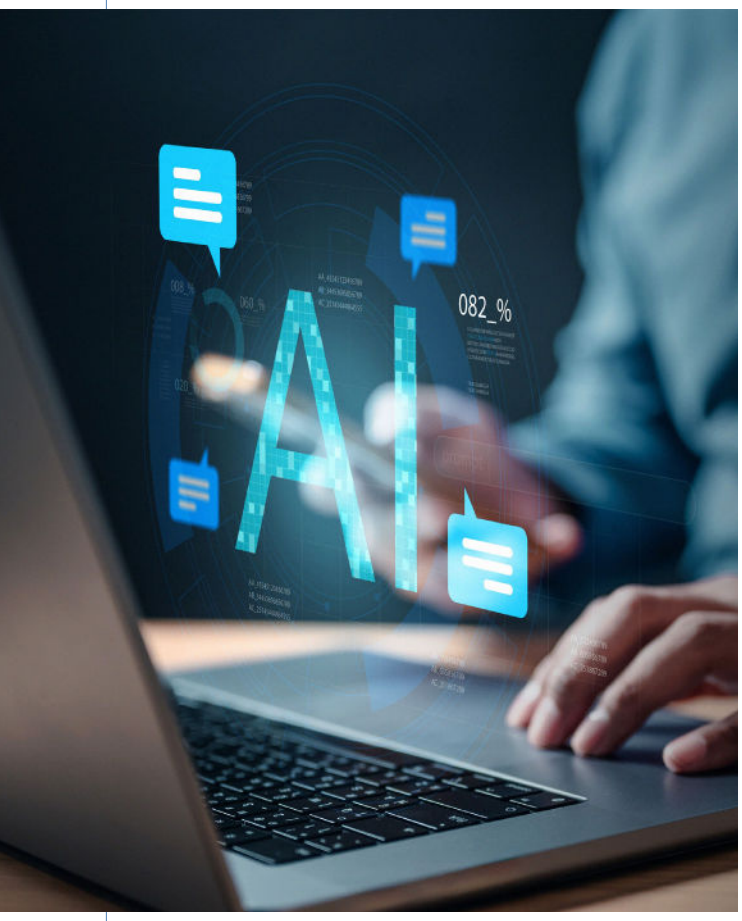
More and more software vendors, organizations and enterprises are leveraging AI into their applications and workflow. At first glance, these applications will inherently be treated as a source of truth because they are from a trusted source, but this imposes a risk for the consumer and ultimately the enterprise that are leveraging these applications. Data privacy remains a significant concern, raising the question: to what extent can an organization trust an AI cloud application that potentially leverages unrestricted data access?

What are the implications when employees inadvertently upload sensitive data to unapproved AI applications? The use of AI applications in workflows creates a new attack vector for threat actors to leverage. If a threat actor has access to the data used in these AI applications, manipulate that information to anything they want. Attackers can launch malicious code, redirect to known bad websites or launch social engineering attacks to extort data or release ransomware payloads.

Localization of AI Applications will not only minimize the exposure of sensitive data, but also reduce privacy risks compared to cloud services. Localizing AI applications can help reduce operational costs and improve overall efficiency, but there are risks in this. Data integrity of AI models and essential Retrieval-Augmented Generation (RAG) files is crucial to prevent tampering that can lead to incorrect outputs. Data poisoning which is an AI attack on the rise can compromise AI data models which then can result in serious consequences like incorrect results or security incidents. Constant monitoring and protection are needed for these files because they are treated as a source of truth in AI applications.

While AI has increased efficiency in various job functions for the better, we must also look at how AI has increase effectiveness of attacks that threat actors are conducting, deepfake attacks being a topic on the rise. These attacks have been idolized in movies for a long time but now many organizations are seeing these attacks due to the rise of AI innovations. Gone are the days of being able to distinguish a real authentic human being in a video conference call. Organizations need to be diligent on protection with this new attack vector.

A protection layer that can be used to circumvent these attacks is to utilize deepfake detection software on front line employee machines. Since these employees are utilizing video conference calls multiple times during the work week, deepfake detection software can effectively prevent security breaches, extortion, and coercion that may arise from manipulated videos. This technology not only helps to combat highly sophisticated attacks but also plays a crucial role in protecting your brand reputation.



5 - Secure Accesss to AI Services

As organizations adopt AI technologies, securing access is also critical. Related risks include unsanctioned usage, noncompliance with company policies, data exfiltration, and the potential for malicious prompt injection attacks—where bad actors manipulate AI models to produce unintended or harmful outputs. Additionally, unsecured GenAI responses can lead to unpredictable behavior, and denial of service attacks can disrupt operations, particularly for private models hosted by the organization.

To mitigate these risks, it is essential to implement robust security measures like Zero Trust Secure Access (ZTSA). This approach is particularly effective for managing user access to both private and public AI services.

Here are some ways to leverage ZTSA in order to mitigate AI access risks:

Central visibility of AI services

- Including shadow AI services that may be used by employees. Monitor for AI services usage across the organization.

Detect and prevent the AI-related threats

- Ensure you have the guardrails in place to prevent Prompt Injection and Data Leakage via AI services.

Enforce access control

- Use strict access controls to limit who can interact with GenAI resources.

Utilize prompt filtering

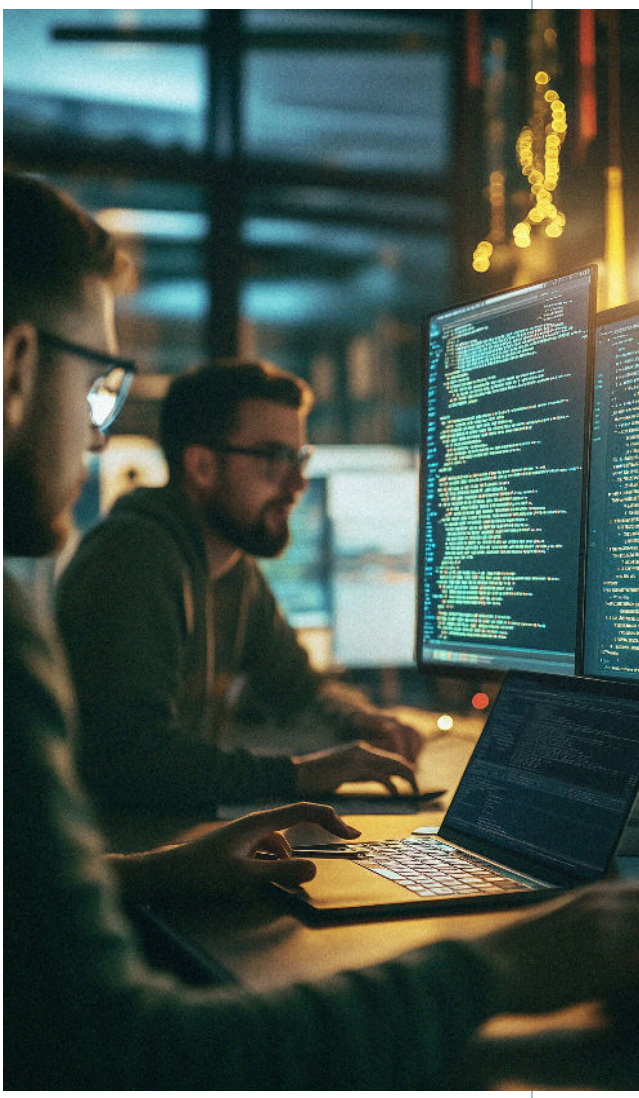
- Filter prompts to detect and block malicious inputs.

Risk-based access control

- **Zero Trust Access Control:** Security capability to always check the risk level of the user before granting the access to system or applications.
- **Rate Limiting:** Helps prevent abuse of services and spiraling costs, ensuring usage remains within safe parameters.
- **Response Filtering: Ensures** AI outputs are safe and appropriate before being delivered to users.
- **Reverse Proxy:** Acts as an intermediary between users and AI services adding an extra layer of protection.

By adopting a Zero Trust framework and employing varied security technologies, organizations can effectively manage risks associated with GenAI, ensuring safe and compliant usage of these powerful tools.

6 - Defend Against Zero Day Exploits



Zero-day exploits target unknown vulnerabilities, and can therefore lead to significant data breaches, operational disruption and reputational damage. The sensitive nature of data processed by AI systems makes them highly attractive targets for cybercriminals in this regard. Attackers can exploit these flaws before patches are available, causing substantial harm before mitigation strategies can be implemented.

Network IDS/IPS can help to mitigate this risk via key elements such as:

Real-Time Threat Detection

- Network intrusion detection and prevention systems (IDS/IPS) monitor traffic in real time to identify suspicious activities and potential zero-day exploits, providing immediate alerts and protection.

Virtual Patching

- A simplified, automated solution to shielding vulnerabilities from exploits, helping organization to quickly mitigate threats and support compliance efforts.

Behavioral Analysis

- Uses advanced analytics and machine learning to establish baselines of normal behavior, enabling it to detect anomalies that may indicate a zero-day attack.

Threat Intelligence

- Leverages a vast repository of threat intelligence to stay updated on emerging threats, allowing you to proactively protect against known and unknown vulnerabilities.
- **Automated Response:** The (Intrusion Prevention System) IPS capabilities allow for automatic blocking of identified threats, minimizing the risk of successful exploits and reducing the response time for incidents.
- **Comprehensive Coverage:** Provides protection across various protocols and applications, ensuring that all layers of the AI infrastructure are monitored and secured.

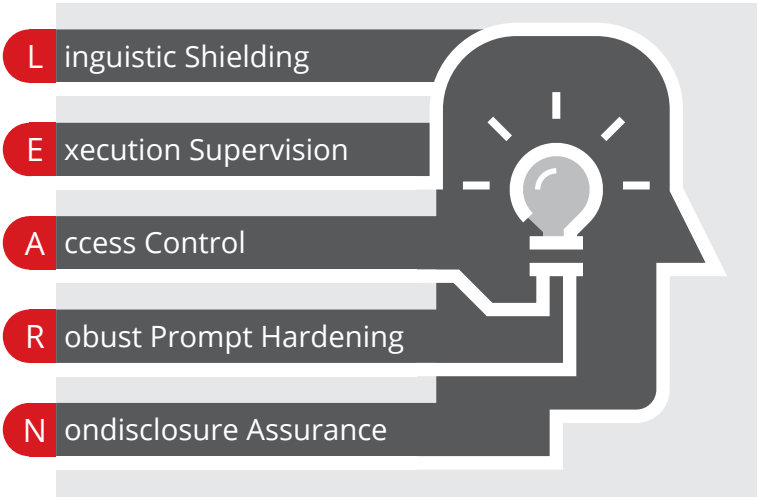
LEARN ARCHITECTURE OVERVIEW

This white paper aims to guide organizations in implementing robust security practices for Large Language Model (LLM) applications. With the rapid adoption of LLMs across various industries, addressing the unique security challenges these applications face is crucial for the future of many companies across the globe.

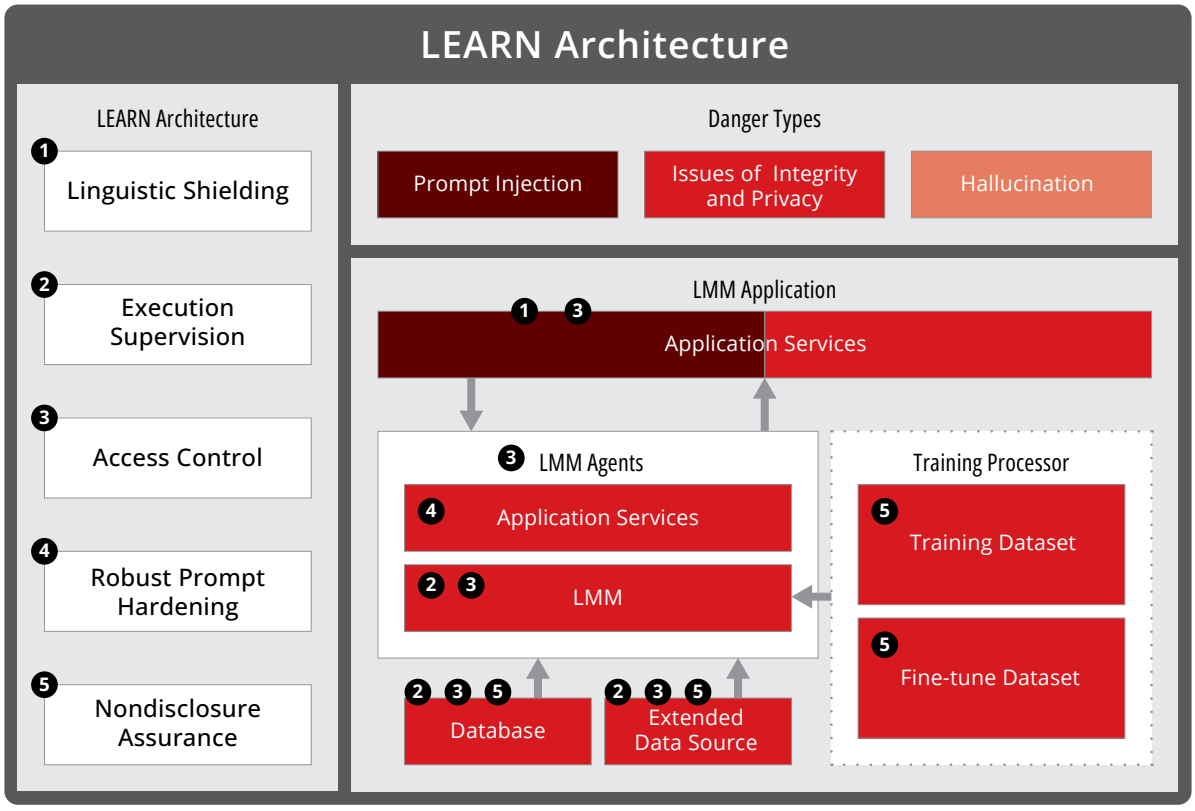
The LEARN Architecture is a comprehensive framework designed by Trend Research to mitigate risks such as prompt injection, as well as issues of integrity, privacy and hallucinations. By mapping these risks to specific best practices and aligning them to the OWASP Top 10 for LLM applications, this best practice guide provides insights and tools to help developers build secure, reliable, and efficient LLM applications.

About LEARN

A comprehensive framework designed to mitigate risks such as prompt injection, issues of integrity and privacy, and hallucinations.



LEARN Overview



LEARN Architecture, a security architecture for developing LLM applications

The LEARN Architecture is designed to enhance the security, integrity and performance of LLM applications. It consists of five best practices:

- **Linguistic Shielding:** Implementing measures to protect the language model from harmful inputs and ensure its outputs are appropriate and safe.
- **Execution Supervision:** Monitoring and controlling the interactions and processes involving the language model to maintain its integrity and reliability.
- **Access Control:** Restricting and managing access to the language model and its data to prevent unauthorized use and ensure that only appropriate entities can interact with it.
- **Robust Prompt Hardening:** Strengthening the queries and prompts used with the language model to prevent manipulation and ensure the model's responses are aligned with intended outcomes.
- **Nondisclosure Assurance:** Implementing techniques to protect sensitive information, ensuring the language model does not inadvertently disclose personal or confidential data.

This architecture aims to address the various security challenges associated with LLM applications by incorporating best practices and strategies across multiple dimensions of interaction with the model. The approach is applicable to on-premises datacenters or cloud architectures.

To learn more, check our LEARN Architecture whitepaper.

WRAP-UP

In conclusion, securing AI technologies requires a comprehensive approach that emphasizes several key areas: protecting data, securing AI models, reinforcing infrastructure, safeguarding users and local applications, controlling access to AI services, and defending against zero-day exploits. These layered strategies are essential for ensuring the integrity and reliability of AI systems.

AI innovation could transform your organization, enabling intelligent data analysis, managing complex supply chain systems, discovering novel drugs to cure disease, and making public services smarter. However, to fully realize these benefits, organizations must maintain visibility into risks, conduct thorough assessments, and implement effective monitoring to detect suspicious activities and vulnerabilities.

By addressing these security concerns, organizations can accelerate AI adoption, remove cybersecurity barriers, gain rapid insights, and ensure regulatory compliance. A proactive security strategy not only enhances trust in AI systems, but also empowers businesses to harness AI's transformative power effectively.

Want more insights like this?

TrendMicro.com/ai

SECURITY FOR AI BLUEPRINT

TREND MICRO™ | Global Leader in Cybersecurity

Trend Micro, a global cybersecurity leader, helps make the world safe for exchanging digital information. Fueled by decades of security expertise, global threat research, and continuous innovation, Trend Micro's AI-powered cybersecurity platform protects hundreds of thousands of organizations and millions of individuals across clouds, networks, devices, and endpoints. As a leader in cloud and enterprise cybersecurity, Trend's platform delivers a powerful range of advanced threat defense techniques optimized for environments like AWS, Microsoft, and Google, and central visibility for better, faster detection and response. With 7,000 employees across 70 countries, Trend Micro enables organizations to simplify and secure their connected world.

For more information visit www.TrendMicro.com.