



CALICO brought to you by  **TIGERA**

LLM Security: Top 10 Risks and 5 Best Practices



Table of Contents

What Is Large Language Model (LLM) Security?	3
Importance of Security in LLM Usage	4
OWASP Top 10 Cyber Security Risks for LLM Applications	5
LLM Cybersecurity Best Practices	9
AI Safety with Calico	11

What Is Large Language Model (LLM) Security?

Large language models (LLM) are advanced AI systems trained on vast text-based datasets, able to create human-like text, code, and other interactions based on natural language prompts. LLMs entered mainstream use with the release of OpenAI's ChatGPT in late 2023, and there are now multiple popular LLM systems with hundreds of millions of users.

LLM security involves practices and technologies that protect LLMs and their associated infrastructure from unauthorized access, misuse, and other security threats.



This includes safeguarding the data they use, ensuring the integrity and confidentiality of their outputs, and preventing malicious exploitation. Effective LLM security measures address vulnerabilities inherent in the development, deployment, and operational phases of these models.

Security in LLMs is crucial due to their ability to process and generate large volumes of sensitive information, making them targets for various cyber threats. These models, which range from text generators to advanced decision-making systems, are integral to numerous applications across industries. Therefore, maintaining their security ensures they operate as intended and continue to be reliable assets in automated and assistive technologies.

Importance of Security in LLM Usage

1 Data Breaches

Large language models (LLMs) store and process massive amounts of data, making them prime targets for data breaches. Hackers who gain unauthorized access, or manipulate model inputs or outputs, can compromise both the model's integrity and the confidential data it holds. Effective security measures are crucial for preventing such incidents and ensuring the trustworthiness of LLM applications.

The implications of a data breach in LLM systems extend beyond simple data loss to include regulatory and reputational damage. Entities using LLMs must adopt rigorous data protection strategies, frequent security audits, and incident response plans to mitigate these risks.

2 Model Exploitation

Model exploitation occurs when attackers identify and leverage vulnerabilities within LLMs for malicious purposes. This can lead to incorrect or harmful outputs from the model, thereby compromising its effectiveness and safety.

Additionally, by manipulating inputs and exploiting model biases, attackers might generate or amplify false information. Such vulnerabilities necessitate ongoing monitoring of model behavior and the implementation of safeguards against these exploits.

3 Misinformation

LLMs can inadvertently generate and spread misinformation if not properly secured and monitored. This capability, if harnessed by malicious users, can lead to widespread dissemination of false information, potentially influencing public opinion. In extreme cases, models could cause financial or bodily harm, for example if they maliciously provide incorrect financial or health advice.

Therefore, it's vital for developers to implement moderation mechanisms and bias checks to minimize the risk of misinformation. Ensuring the model's output remains accurate and unbiased supports the reliability and credibility of LLM applications.

4 Ethical and Legal Risks

Misuse of LLM technology can lead to serious ethical and legal consequences. For instance, generating discriminatory or biased content can result in legal exposure and damage an organization's reputation. It's essential to integrate ethical guidelines and compliance checks in the deployment of LLMs.

Further legal risks involve compliance with international data protection regulations such as GDPR. Organizations must ensure that their use of LLMs adheres to all applicable laws and that they maintain transparency with users about data handling practices.

OWASP Top 10 Cyber Security Risks for LLM Applications

The Open Web Application Security Project (OWASP) has released a list of the most severe security threats facing LLM applications, the LLM Top 10. Below, we summarize the top 10 risks according to OWASP and how to mitigate them.



1 LLM01: Prompt Injection

Prompt injection is a security risk where attackers manipulate the input prompts to an LLM to elicit undesirable or harmful responses. This can compromise the model's integrity and output quality.

Safeguards against prompt manipulation involve validating and sanitizing all inputs to the model. Developers should also design LLM systems to detect anomalous input patterns that may indicate attempted prompt injections. These measures help in maintaining control over model interactions.

2 LLM02: Insecure Output Handling

Insecure output handling in LLMs refers to insufficient checks and balances that fail to prevent the dissemination of sensitive or harmful information. This can lead to breaches of privacy and exposure of sensitive data. Implementing robust output filtering and validation processes is essential to secure LLM applications.

Additionally, defining clear policies regarding output confidentiality and integrity helps in safeguarding important data. Continuous monitoring and regular updates also play crucial roles in maintaining output security.

3 LLM03: Training Data Poisoning

Training data poisoning involves tampering with the data used to train LLMs, aiming to corrupt the model's learning process. This can skew model outputs, leading to unreliable or biased results. Vigilance in data sourcing and validation is required to prevent poisoning.

Countermeasures include using verified and secure data sources, employing anomaly detection during training, and continuously monitoring model performance for signs of corruption.

4 LLM04: Model Denial of Service

A model Denial of Service (DoS) attack targets the availability of an LLM by overwhelming it with numerous or complex queries. This can render the model non-functional for legitimate users. Measures to prevent DoS attacks include rate limiting, robust user authentication, and deploying auto-scaling resources.

Furthermore, implementing fail-safes and recovery protocols ensures the model can maintain functionality or quickly recover from such attacks, minimizing downtime and service disruption.

5 **LLM05: Supply Chain Vulnerabilities**

Supply chain vulnerabilities in LLMs occur when the components or services the model relies on are compromised. This can lead to system-wide vulnerabilities. Ensuring all components are from trusted providers and regularly updated is key to securing the supply chain.

Additionally, running security assessments on all third-party services and integrating security at each stage of development and deployment can significantly reduce these risks.

6 **LLM06: Sensitive Information Disclosure**

Sensitive Information Disclosure risks arise when LLMs inadvertently reveal personal or proprietary information within their outputs. To counter this, strict data handling and output sanitization protocols must be enforced.

Data privacy measures, such as data anonymization and encryption, prevent the exposure of sensitive data. Regular audits and compliance checks ensure these protocols are followed and adapted to evolving data protection standards.

7 **LLM07: Insecure Plugin Design**

Insecure plugin design in LLMs can introduce vulnerabilities that compromise the entire system. Plugins should be designed with security as a priority, incorporating features like input validation and secure data handling.

Developer training and adherence to secure coding practices ensure that plugins remain robust against attacks. Periodic security assessments help identify and rectify potential vulnerabilities in plugin design.

8 LLM08: Excessive Agency

Excessive agency refers to LLMs making autonomous decisions without sufficient human oversight. This can lead to unexpected or undesired outcomes. Implementing clear guidelines and constraints on model autonomy can mitigate these risks.

Regular reviews and updates of the decision-making protocols keep the model's actions within desired boundaries. Human-in-the-loop architectures ensure ongoing oversight and intervene when necessary.

9 LLM09: Overreliance

Overreliance on LLMs can lead to dependency, where critical decision-making is left to the model without adequate verification. Encouraging practices like cross-verification and maintaining alternative decision-making processes reduces dependency risks.

Educating users on the limitations and appropriate use of LLM technology fosters a balanced approach, ensuring technology augments human decision-making without replacing it.

10 LLM10: Model Theft

Model theft involves unauthorized access and copying of proprietary LLM configurations. Protecting intellectual property through rigorous access controls, encryption, and legal measures is imperative.

Regular security updates and monitoring access logs help in detecting and responding to potential theft attempts. Additionally, legal frameworks ensure that perpetrators are held accountable, and damages are recovered when theft occurs.

LLM Cybersecurity Best Practices

1 Encrypt Data In Transit and at Rest

Encrypting data in transit involves using secure protocols such as HTTPS and SSL/TLS to safeguard data as it moves across networks. This prevents unauthorized intercepts and access to the data during transmission. Encryption of data at rest ensures that stored data is inaccessible without proper decryption keys, protecting it from theft or unauthorized exposure.

Implementing strong encryption not only secures information but also builds trust with users and stakeholders, and can support compliance with data protection regulations such as GDPR and HIPAA.

2 Implement Strict Access Controls

Access controls are critical in managing who can view or use the data within large language models. This can involve access controls implemented by the provider of a foundation model, such as OpenAI's GPT or Meta LLaMA, or access controls put in place by an organization deploying a model for use by employees or customers.

Using authentication mechanisms like multi-factor authentication (MFA) and employing role-based access control (RBAC) systems can limit access to sensitive data according to user roles and responsibilities. This minimizes the risk of unauthorized access and potential data breaches.

Additionally, maintaining stringent access logs and monitoring access patterns can help promptly detect and address abnormal behavior or potential security threats. Regular audits and reviews of access controls and permissions ensure that only authorized users have access to critical data, maintaining a secure environment for large language models.

3 When Training LLMs, Anonymize Data to Protect User Privacy

During the training of large language models, anonymizing data helps in minimizing privacy risks and protecting the identities of individuals represented in the data. Anonymization involves techniques like data masking or pseudonymization, which obscure or remove identifying information from datasets used in training LLMs.

Furthermore, ensuring that anonymized training datasets cannot be reverse-engineered to reveal personal information is paramount. Best practices include the use of advanced anonymization algorithms and regular checks to ensure the effectiveness of anonymization procedures.

4 Manage and Control Training Data Sources

Managing the sources of training data for LLMs involves ensuring the authenticity, accuracy, and security of the data. It is important to source data from reliable and secure sources to prevent the introduction of biases or malicious code into the model. Verification of data sources and continuous monitoring for inconsistencies is vital.

Controlling access to training data sources and maintaining secure backups of critical data sets also help mitigate risks associated with data corruption or loss. This requires strict guidelines and protocols for the handling and usage of training data, ensuring large language models are both robust and secure.

5 Develop and Maintain an Effective Incident Response Plan

An effective incident response plan is crucial for promptly addressing security breaches or other disruptions. This plan should include procedures for assessing the severity of an incident, containing the breach, and mitigating any damage. Communication strategies to inform stakeholders and users about the incident and the steps taken to resolve it are also vital.

Regular training of response teams and periodic drills to simulate security incidents ensure preparedness and functional responsiveness. Keeping incident response plans updated to reflect new cybersecurity threats and adapting to technological advancements in large language models helps maintain an effective defense against potential security threats.

AI Safety with Calico



Calico offers numerous features to address the many network and security challenges faced by cloud platform architects and engineers when deploying GenAI workloads in Kubernetes. Here are five Calico features for container networking and security for GenAI workloads:

1 Egress access controls

Calico provides granular, zero-trust workload access controls between individual pods in Kubernetes clusters to external resources such as LLMs. It provides fine-grained workload access controls using DNS egress policies and NetworkSets (using IPs/CIDRs in network policy).

2 Egress gateway

The Calico Egress Access Gateway assigns a fixed, routable IP to a Kubernetes namespace. All egress pod traffic from that namespace with an assigned routable IP address identifies the workload running within that namespace. This enables the cluster to securely scale while preserving the limited number of routable IPs and leveraging non-routable IPs for all other pod traffic within the cluster. The Calico Egress Gateway works with any firewall, enabling Kubernetes resources to access endpoints behind a firewall securely.

3 Identity-aware microsegmentation

Calico enforces microsegmentation to achieve workload isolation and secure lateral communication between pods, namespaces, and services. It enables teams to logically divide workloads into distinct security segments and then define granular security controls for each unique segment. Teams can isolate workloads based on environments, application tiers, compliance needs, user access, and individual workload requirements.

4 Observability and troubleshooting

Calico's Dynamic Service and Threat Graph provides a graph-based visualization of your Kubernetes deployments, including images, pods, namespaces, and services. It has built-in troubleshooting capabilities to identify and resolve security and compliance gaps, performance issues, connectivity breakdowns, anomalous behavior, and security policy violations.

5 Cluster mesh

Calico provides a centralized, multi-cluster management plane to enable security, observability, and advanced networking for workloads and services across multiple clusters in hybrid and multi-cloud environments. Calico provides unified security policy controls and federated endpoints and services.

Getting Started

Get hands-on experience with Calico for container security, networking, and observability based on your organization's needs.

[Start a Free Trial](#)



CALICO

brought to you by



TIGERA