



CALICO brought to you by  **TIGERA**

Generative AI Security: 6 Critical Risks and Defending Your Organization

Table of Contents

What Is Generative AI?	3
How do Generative AI Models Work?	3
Generative AI Security Risks	4
Enterprise Best Practices for Generative AI Cybersecurity	7
AI Security with Calico	9

What Is Generative AI?

Generative AI is an advanced field of artificial intelligence that focuses on creating new content autonomously, by learning patterns from extensive datasets. This technology powers various tools, which have recently entered the mainstream, including AI-driven chatbots like OpenAI ChatGPT and Google Gemini, and image generators such as DALL-E and Stable Diffusion.

While generative AI can perform many of the roles of traditional machine learning models, particularly in the field of natural language processing (NLP), generative AI introduces completely new applications for AI, such as generation of textual content, images, videos, and audio. At the basis of this field are so-called 'foundation models', competent at a broad variety of tasks, which require extensive research, huge datasets and massive computing power. Among the current state-of-the-art foundation models are OpenAI's GPT, Google's Gemini, and Meta's open source LLaMA series of models.

How do Generative AI Models Work?

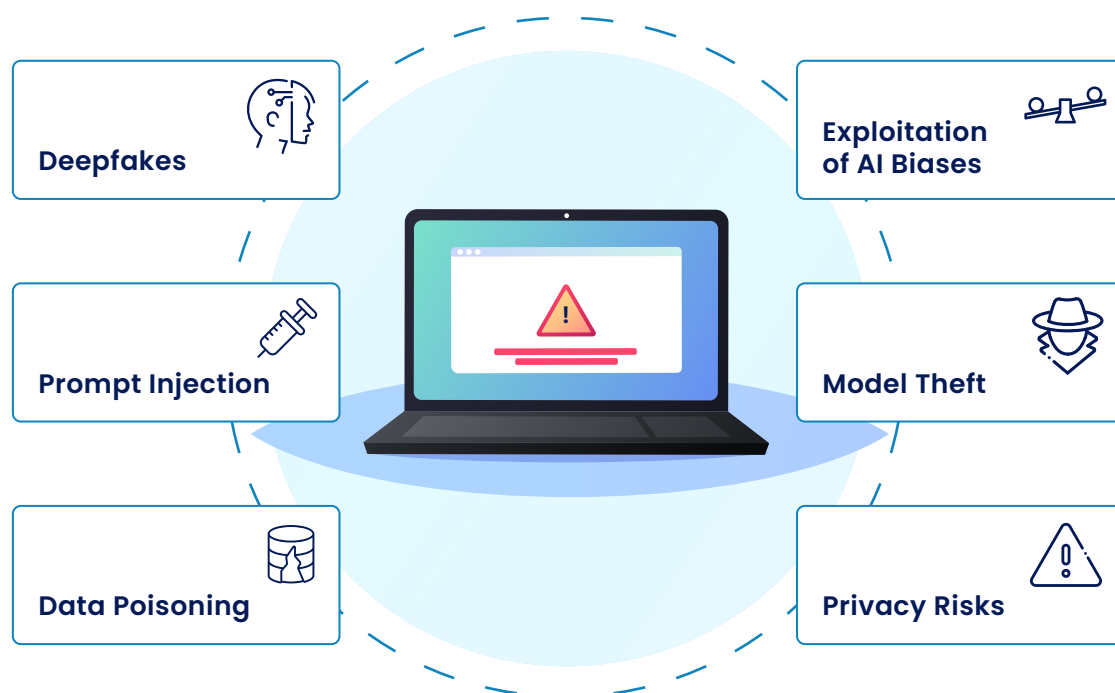
Generative AI models utilize several neural network architectures to achieve their capabilities. State-of-the-art image generation models use the diffusion technique, which involves training a neural network to denoise images blurred with Gaussian noise. Models are trained to reverse the process of adding noise to an image, thus generating new, photorealistic images based on natural language prompts.

Text generation models are typically based on the Transformer architecture, first introduced in a 2017 paper by Google, which underpins today's advanced large language models (LLMs). Transformers utilize an attention mechanism to manage and predict sequences in data, enhancing their ability to generate coherent and contextually relevant outputs based on the input data they process.

A newer innovation in generative AI is the introduction of multi-modal models, which are trained simultaneously on text, images, and videos, and are able to receive inputs and generate outputs combined of any type of information. The first fully multi-model generative AI model was Google Gemini, which is able to understand and reason about textual, image, or video content.

Generative AI Security Risks

While generative AI has tremendous potential in improving human productivity, it also raises new and severe security risks. Here are some of the important risks that must be considered by individuals or organizations developing, deploying, or using generative AI technology.



1 Deepfakes

Deepfakes represent a significant challenge in cybersecurity, stemming from generative AI's ability to create convincing fake images and videos. These can be used to impersonate individuals in phishing attacks or to spread misinformation. The realism of deepfakes makes it difficult for both individuals and traditional security systems to detect falsifications, increasing the risk of security breaches.

Moreover, the propagation of deepfakes can undermine trust and credibility in digital communications. As generative AI continues to improve, the sophistication and indistinguishability of deepfakes are likely to increase, posing elevated threats to personal and organizational security.

2 Prompt Injection

Prompt injection occurs when malicious inputs are provided to an AI system, leading it to generate outputs that facilitate a security breach. This form of attack can alter the behavior of AI-powered systems, causing them to malfunction or produce harmful outputs. Detecting such manipulation can be challenging, as it requires constant vigilance over the inputs and outputs of AI systems.

For organizations using generative AI, safeguarding against prompt injection involves rigorous input validation and continuous monitoring of AI activities. Early detection and response are crucial to mitigate potential damages and maintain system integrity.

3 Data Poisoning

Data poisoning is a tactic where attackers corrupt the training data of AI systems, leading to compromised model integrity. By inserting false or misleading data into the training set, attackers can skew the AI's output or cause it to fail under specific conditions. This manipulation can seriously affect decision-making processes and operational effectiveness.

To combat data poisoning, cybersecurity professionals must implement stringent data validation and cleansing processes. Ensuring the accuracy and reliability of training data is vital to maintaining the effectiveness and trustworthiness of generative AI applications.

4 Exploitation of AI Biases

Generative AI systems can inherit or develop biases based on the data they are trained on, which can be exploited by attackers. For example, if an AI system disproportionately associates certain behaviors or traits with specific demographics, it could lead to skewed security measures against those groups. Exploiting these biases can enable attackers to engineer specific scenarios that the AI fails to recognize or mishandles.

Addressing AI biases involves regular audits, diverse data sets for training, and the implementation of fairness algorithms to detect and mitigate biased outcomes. Maintaining ethical standards and fairness in AI operations is critical to its success and acceptance in cybersecurity roles.

5 Model Theft

Model theft is a growing concern in the context of generative AI in cybersecurity. Attackers may steal AI models to understand their structure and functioning, which can be exploited to bypass security mechanisms. Stolen models can also be replicated or sold illegally, undermining the competitive advantage of original developers.

Protecting AI models involves securing data pipelines and storage, implementing robust access controls, and continuously monitoring for any unauthorized access. As AI models become more integral to business operations, ensuring their security against theft is crucial to maintaining operational integrity and trust.

Model theft is not limited to developers of foundation models. Many organizations are fine-tuning and training generative AI models on their own data. These custom models are a new type of intellectual property, which represents a valuable target for attackers.

6 Privacy Risks

Privacy risks arise when generative AI handles sensitive personal or organizational data. There are concerns about how data is used, stored, and accessed within AI systems, especially considering that breaches could expose vast amounts of confidential information. Ensuring data privacy is paramount to maintaining trust and compliance with regulations like GDPR.

To secure privacy, organizations providing AI services or deploying them for their customers and employees must enforce strict access controls, data encryption, and anonymization techniques where possible. Comprehensive privacy policies that address the specifics of generative AI are essential to safeguard sensitive information against unauthorized access or leaks.

Learn more in our detailed guide to [generative AI security risks](#).

Enterprise Best Practices for Generative AI Cybersecurity

Here are a few steps organizations can take to prevent and mitigate threats when using generative AI technology.

1 Bias and Fairness Assessments

Regular bias and fairness assessments are crucial for organizations employing generative AI. These assessments help identify and correct any discriminatory behavior or unfair outcomes produced by AI systems. By routinely evaluating AI outputs and decisions, organizations can ensure that their AI implementations are ethical and unbiased, fostering trust and reliability.

Additionally, these assessments promote transparency in AI operations, crucial for regulatory compliance and public acceptance. By proactively managing AI biases, companies can prevent harm and discrimination, supporting a just and equitable digital environment.

2 Input Validation

Input validation is essential to safeguard generative AI systems from manipulative or harmful inputs that could compromise system integrity. By rigorously checking data before it is processed, organizations can prevent unwanted or malicious content from influencing AI behavior. This practice is particularly important in environments where AI interacts directly with users or other systems.

Implementing comprehensive input validation protocols enhances the resilience of AI systems against attacks such as data poisoning and [prompt injections](#). It contributes to the overall security architecture, ensuring that only clean, verified data influences AI decisions and functions.

3 Leverage Explainable AI

Explainable AI (XAI) refers to methods and techniques that make the outcomes of AI models understandable by humans. Leveraging XAI within cybersecurity helps stakeholders understand how security decisions are made, fostering greater trust and confidence in AI-driven systems. Transparency in AI processes is crucial for troubleshooting, optimizing operations, and complying with regulatory standards.

XAI also aids in identifying the reasons behind any anomalous or malicious activities detected by AI systems, enabling more effective mitigation strategies. As cybersecurity threats grow more sophisticated, employing explainable AI can provide clarity and improve response measures.

4 Maintain Clear Documentation of AI Model Development and Deployment

Maintaining clear documentation of AI model development and usage is fundamental for auditability and continual improvement. This documentation should include details of the data used, model training processes, versioning information, performance metrics, use cases, and information about model fine-tuning. Accurate records ensure that AI systems are transparent and their functions fully understood by developers, operators, and auditors alike.

Such rigorous documentation practices not only help in maintaining operational consistency but also assist in compliance with legal and regulatory requirements. It serves as a basis for refining AI models, troubleshooting issues, and enhancing system reliability.

5 AI-Specific Incident Response Plans

AI-specific incident response plans are tailored to address potential issues arising from the deployment of AI technologies. These plans include scenarios like data breaches, system misconfigurations, and AI model failures. Having a specialized response mechanism ensures that teams can quickly and efficiently handle AI-related incidents, minimizing negative impacts.

These plans should integrate with broader cybersecurity strategies, providing a cohesive response across all technological fronts. They play a crucial role in maintaining service continuity, safeguarding data, and protecting the integrity of AI systems.

6 Regular Updates of AI Systems

Regular updates are vital to ensuring that generative AI systems remain secure against newly discovered vulnerabilities. This is especially important in light of the rapid development of generative AI models. By keeping AI software and models up to date, organizations can protect against exploits that target outdated systems. This practice extends to all components of the AI infrastructure, including data processing pipelines and integration points.

Scheduled audits and updates help maintain the robustness of AI systems, adapting to both technological advancements and emerging security threats. Effective patch management reduces the risk of security breaches and enhances overall system resilience, supporting sustained operational performance.

AI Security 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