We're your partner in

Understanding AI Security

CLOUD SECURITY PARTNERS

Overview

As more companies explore the world of generative AI and how it will enable their business, we must approach these integrations with caution and awareness. As well as the typical application security issues, we must understand the new risks generative AI introduces. A whole field of new attacks has been discovered and documented in generative AI systems. ML and specifically generative AI inherits all traditional software vulnerabilities, but its black-box nature and non-deterministic behavior introduce unique challenges.

The following are just a few of the potential weaknesses that generative AI introduces. The best approach to integrate generative AI into applications and business processes is using the following secure principles and processes:

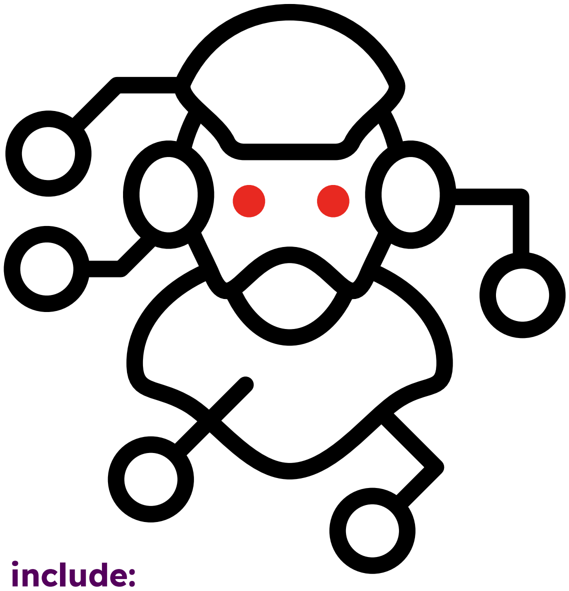
> **Threat model integration:** **review threats and risks of integration early into the design process.**

> **Secure design:** **build the most secure design for the integration to reduce risks.**

> **Continuous evaluation:** **review and update designs and architecture as threats and risks evolve.**

Generative AI Risks

The following are some risks present in generative models and integrations.



Confidentiality Concerns

Confidentiality attacks in Generative AI systems may include:

- Model Inversion: Risk of reconstructing training data from model outputs, threatening data privacy.

Mitigation Strategies:

- Differential Privacy: Anonymize training data to mask individual records.
- Data Classification: Only index and expose data that a user is authorized to access in a given model.
- Rate Limiting: Prevent unauthorized, frequent model queries.
- Access Control: Restrict model access to trusted users only.
- Utilize tooling like Laiyer AI’s LLM Guard to sanitize inputs (anonymize PII, Ban Substrings, & Secrets such as API keys) to ensure private data is not ingested by the language model.

Integrity Threats

Attacks on ML integrity can lead to:

- Dataset Poisoning: Skewing model learning by injecting malicious data into training sets.
- Evasion Attacks: Deceiving models into incorrect classifications with subtle input changes.

Mitigation Strategies:

- Dataset Poisoning: Skewing model learning by injecting malicious data into training sets.
- Evasion Attacks: Deceiving models into incorrect classifications with subtle input changes.
- Regular Audits: Periodically check and validate the integrity of training data and the model itself.
- Model Robustness Testing: Continuously test models against potential threats.

Availability Attacks

Denial of Service (DoS) attacks on ML systems:

- Similar to traditional applications, but it may target specific model functionalities.

Preventive Measures:

- Usage Monitoring: Detect and respond to anomalies in model usage.
- Rate Limiting: Prevent overloading and potential model theft.
- Redundancy: Implement failover systems to maintain service during attacks.

Addressing Large Language Models (LLMs) Specific Risks:

- Prompt Injection: LLM-specific attacks generating malicious content.
- There is no foolproof solution yet; use LLMs with caution. Mitigate prompt injection risk via system prompt and tooling like LLM Guard.
- Utilize tools like Laiyer AI's LLM Guard & Meta's Purple Llama / Llama Guard

> Cooperation with Cloud Security Partners significantly helped us mitigate all security risks that could impact our customers prior to our product's release.
>
> **- Peter Vaclavik, CTO, HireLogic**

Key Takeaways

For AI Integration

Monitoring Vigilant monitoring of ML model interactions.

Rate Limiting Essential against DoS and model theft.

Data Quality Invest in privacy-respecting data curation.

Supplementary Controls Restrict access and maintain dependencies.

Data Classification Tag data to prevent leaks and assist in Incident Response.

Correctness and Review Human review of AI outputs to counter 'hallucinations'. Keep a human in the loop for all processes. Implement Retrieval Augmented Generation (RAG) as well as prompting techniques to assist in reducing hallucinations.

Utilize Tooling Like LLM Guard's Factuality output scanner as well as strategies such as Retrieval Augmented Generation (RAG) to provide additional context such as private data not included in the pre-training dataset.

Comprehensive Security Adhere to cloud security best practices and manage higher compute costs.

Red Teaming Understand the capabilities of the model and uncover potential risks.

RLHF/RLAI/DPO Ensure your model is aligned to be helpful, harmless, and honest by utilizing techniques such as RLHF (Reinforcement Learning from Human Feedback), RLAI (Reinforcement Learning from AI), or DPO (Direct Preference Optimization).

Your Partner in Cloud and App Security

Additional Considerations

Employee Training Educate staff on AI security risks and best practices.

Policy Development Create policies for ethical AI use and data handling.

External Audits Engage third-party experts to evaluate AI system security.

Continual Learning Update AI models to adapt to emerging threats.

Red Team Red Team your model to ensure you understand the capabilities and risks. Implement guardrails.

Stakeholder Communication Keep stakeholders informed about AI risks and defenses.

Work with our experienced team of senior consultants to reduce your risk!

CLOUD SECURITY
PARTNERS

🌐 cloudsecuritypartners.com
in cloud-security-partners
X @CloudSecPartner