TORONTO | SEPTEMBER 11, 2024

aws summit



SEC303

Securing generative AI in the cloud: AI/ML and generative AI deep dive

Emily Soward

Data Scientist
Amazon Web Services

Bill Ohlson

Executive Security Advisor Amazon Web Services



Introductions





Today, we'll help answer these questions

- 1. Does generative AI require me to change my approach to security?
- 2. What are the intersections of generative AI and security?
- 3. What mechanism can I use to understand what risks, security, and compliance requirements impact my use of generative AI?
- 4. I want to put the above into practice What does that look like?

Generative AI Security Scoping Matrix

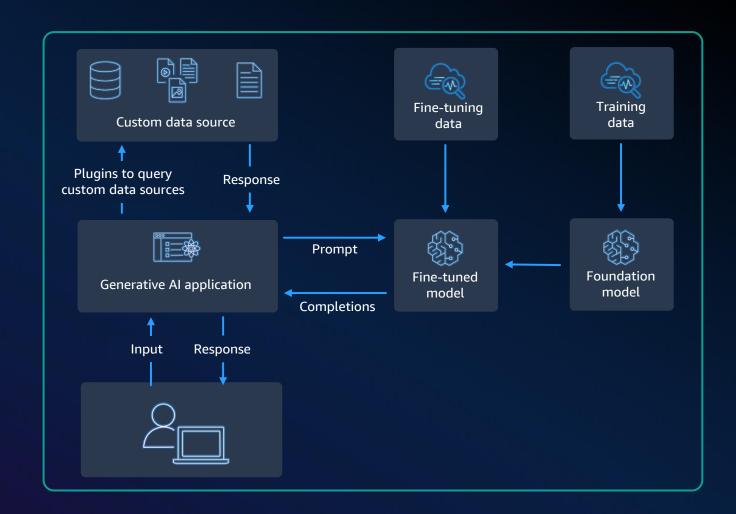
A MENTAL MODEL TO CLASSIFY USE CASES

SCOPE 1 SCOPE 2 **SCOPE 3 SCOPE 4** SCOPE 5 **Enterprise app Pre-trained models** Fine-tuned models Self-trained models Consumer app Using "public" Using an app or SaaS Building your app on Fine-tuning a model Training a model from a versioned model generative Al services with generative Al on your data scratch on your data features Ex: Amazon Bedrock Ex: Amazon Bedrock Ex: Amazon SageMaker Ex: Amazon PartyRock, Ex: Salesforce Einstein base models customized models. ChatGPT, Midjourney GPT, Amazon O Amazon SageMaker JumpStart Securing generative Al Governance & compliance Legal & privacy Risk management Controls Resilience

Data flows in a generative AI application

DATA FLOW AND DATA OWNERSHIP

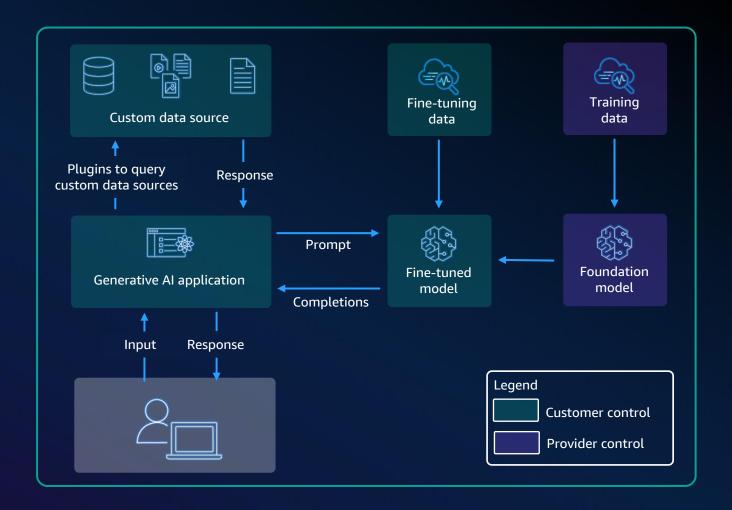
- App receives input from user
 Optional: App queries data from custom data sources
- 2. App formats user input and customer data into a prompt
- Prompt is completed by a model (fine-tuned or pre-trained)
- 4. Completion is processed by app
- 5. Response is sent to the user



Scope 4: Fine-tuned models

DATA FLOW AND DATA OWNERSHIP

- Model is fine-tuned on your data to improve its responses
- Fine-tuned model can be offered as an API or can be hosted by you
- You can fine-tune an open sourced model or a closed-source model
- Examples: Amazon Bedrock customized models, Amazon SageMaker JumpStart



Threat modeling – First principles

- 1. What are we working on?
- 2. What can go wrong?
- 3. What are we going to do about it?
- 4. Did we do a good job?

Adam Shostack, "Threat Modeling: Designing for Security"



OWASP Top 10 for LLM Applications

LLM01: Prompt injection

LLM02: Insecure output handling

LLM03: Training data poisoning

LLM 04: Model denial of service

LLM 05: Supply chain vulnerabilities

LLM06: Sensitive information disclosure

LLM07: Insecure plugin design

LLM08: Excessive agency

LLM 09: Overreliance

LLM 10: Model theft



MITRE ATLAS

ADVERSARIAL THREAT LANDSCAPE FOR ARTIFICIAL-INTELLIGENCE SYSTEMS

ATLASTM

The ATLAS Matrix below shows the general progression of attack tactics as column headers from left to right, with attack techniques organized below each tactic. 4 indicates a tactic or technique directly adapted from from ATT&CK. Click on the blue links to learn more about each item, or search and view more details about ATLAS tactics and techniques using the links in the top navigation bar.

Reconnaissance & 5 techniques	Resource Development & 7 techniques	Initial Access & 6 techniques	ML Model Access 4 techniques	Execution & 3 techniques	Persistence & 3 techniques	Privilege Escalation & 3 techniques	Defense Evasion & 3 techniques	Credential Access & 1 technique	Discovery &	Collection &	ML Attack Staging 4 techniques	Exfiltration & 4 techniques	Impact & 6 techniques
	/ techniques	o techniques	4 techniques	3 techniques	3 techniques	3 techniques	3 techniques	i teciliique	4 techniques	3 techniques	4 techniques	4 techniques	o techniques
Search for Victim's Publicly Available Research Materials	Acquire Public ML Artifacts	ML Supply Chain Compromise	ML Model Inference API Access	User Execution &	Poison Training Data	LLM Prompt Injection	II Evade ML Model	Unsecured Credentials &	Discover ML Model Ontology	ML Artifact Collection	Create Proxy ML Model	Exfiltration via ML Inference	Evade ML Model
	Obtain	Valid	ML-Enabled	Command and Scripting Interpreter &	Backdoor ML Model	LLM Plugin Compromise	LLM Prompt	Discover ML Model Family Discover ML Artifacts	Discover ML	Information	Backdoor ML Model	Exfiltration via	Denial of ML Service
Search for Publicly Available Adversarial Vulnerability Analysis	Capabilities &	Accounts &			Model		Injection				iviodei	Cyber	Service
	Develop Capabilities &	Evade ML			LLM Prompt	I LLM Jailbreak	LLM Jailbreak		Repositories &	Verify	Means	Spamming ML	
		Model		LLM Plugin Compromise	Injection					Data from Local	Attack		
		Exploit Public- Facing	Physical Environment Access							System &	Craft Adversarial Data		
Search Victim-Owned Websites	Acquire Infrastructure												Data
			7,00000						LLM Meta Prompt			LLM Data	Erode ML Model
	Publish Poisoned Datasets	Application &	Full ML Model						Extraction			Leakage	Integrity
Search Application Repositories		LLM Prompt	Access										Coot
		Injection											Cost Harvesting
Active Scanning &	Poison Training Data												Truivesting
		Phishing &	1										External
	Establish		_										Harms
	Accounts &												

Whiteboard



Al security and privacy journey – Where are you today?

Envision

What is AI, ML, and generative AI security?

Align

What controls do I need for my business?

Launch

How do I prepare to transform and enable my business?

Scale

How can I grow and drive value?

Optimize

How do I realize more with a security-first approach?

Increase success with generative AI security and privacy

Al security awareness

Learn about AWS and industry standards

Al security understanding

Assess holistic AI security posture and prioritize controls

Al security readiness

Build a security roadmap that aligns with operating model, governance, risk, and compliance

Al security transformation

Execute on security goals through enablement, fostering change, and controls

Al security optimization

Scale through automaton, resiliency, and optimization



Additional resources



Introduction to the Generative Al Security Scoping Matrix



Architect defense-in-depth using the OWASP Top 10 for LLMs



Learn how to integrate threat modeling into your SDLC



Workshop: Learn how to leverage foundation models through Amazon Bedrock



Deploying a multi-model and multi-RAG powered chatbot



MITRE Adversarial Threat Landscape for AI Systems (ATLAS)





Build beyond

Create a free account on AWS Skill Builder to gain in-demand skills

