TORONTO | SEPTEMBER 11, 2024

# aws SUMMIT

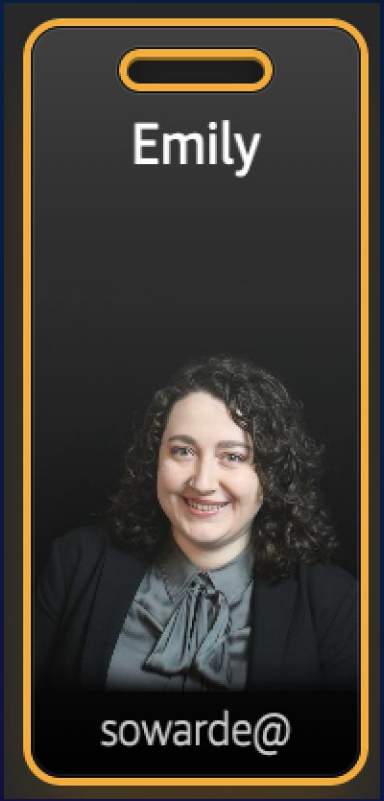
SEC303

# 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

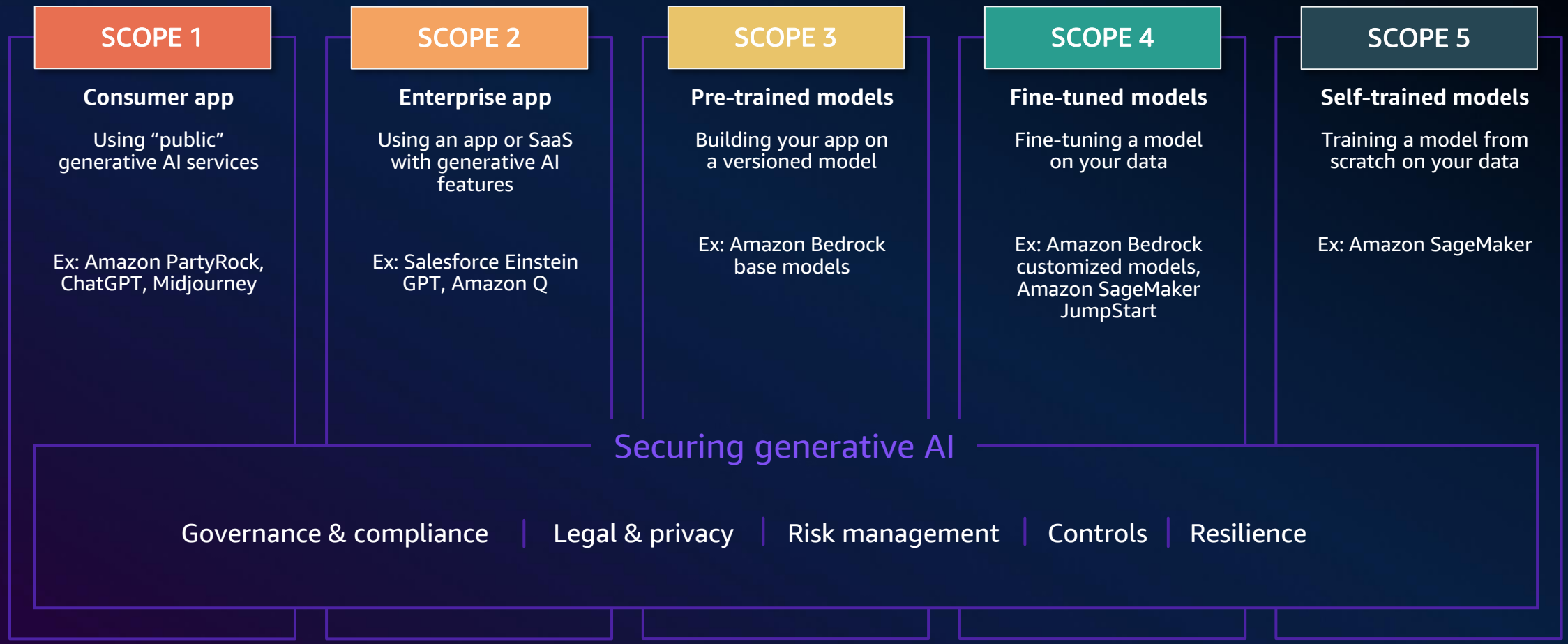



# Today, we'll help answer these questions

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

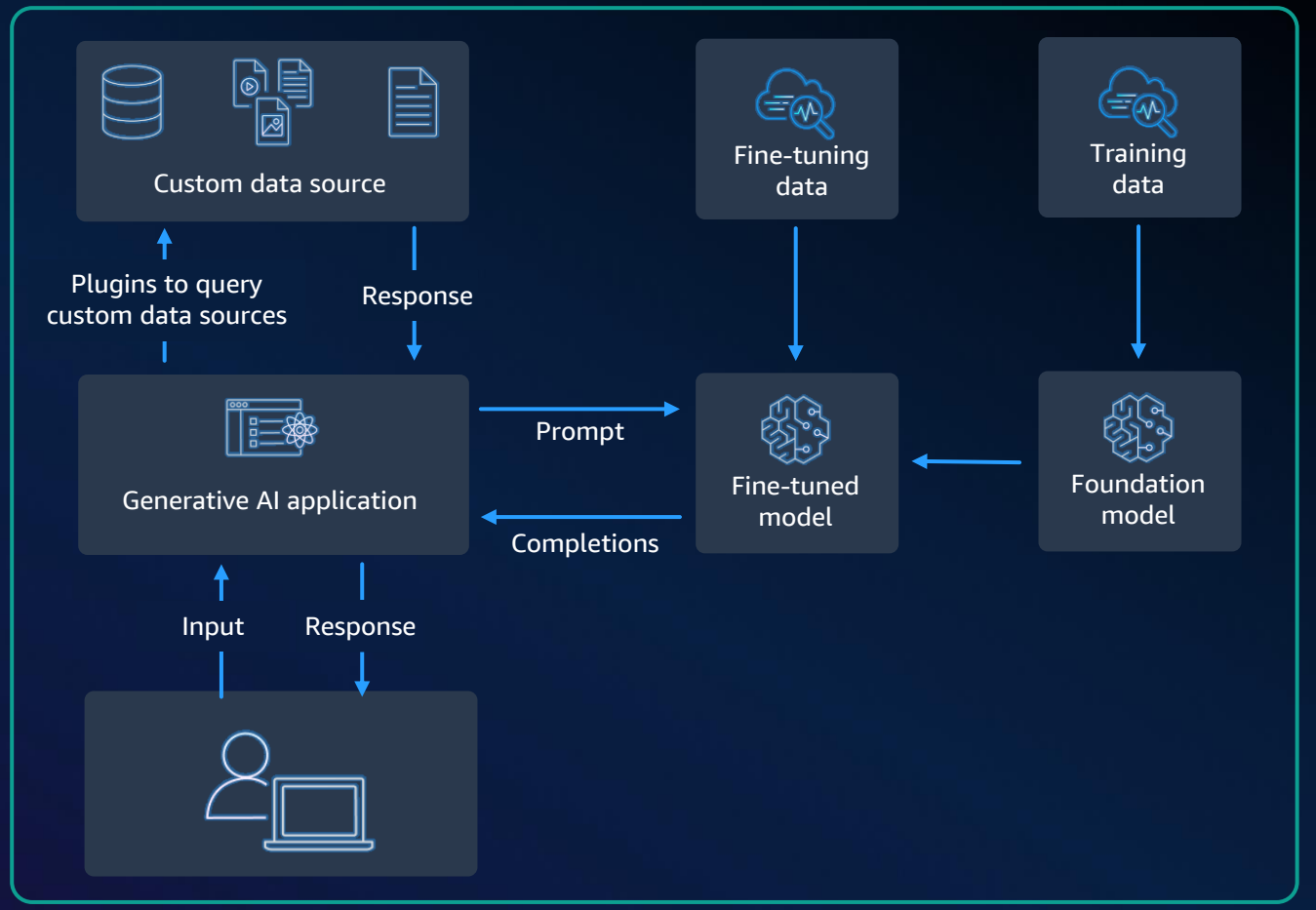
A MENTAL MODEL TO CLASSIFY USE CASES



# Data flows in a generative AI application

DATA FLOW AND DATA OWNERSHIP

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

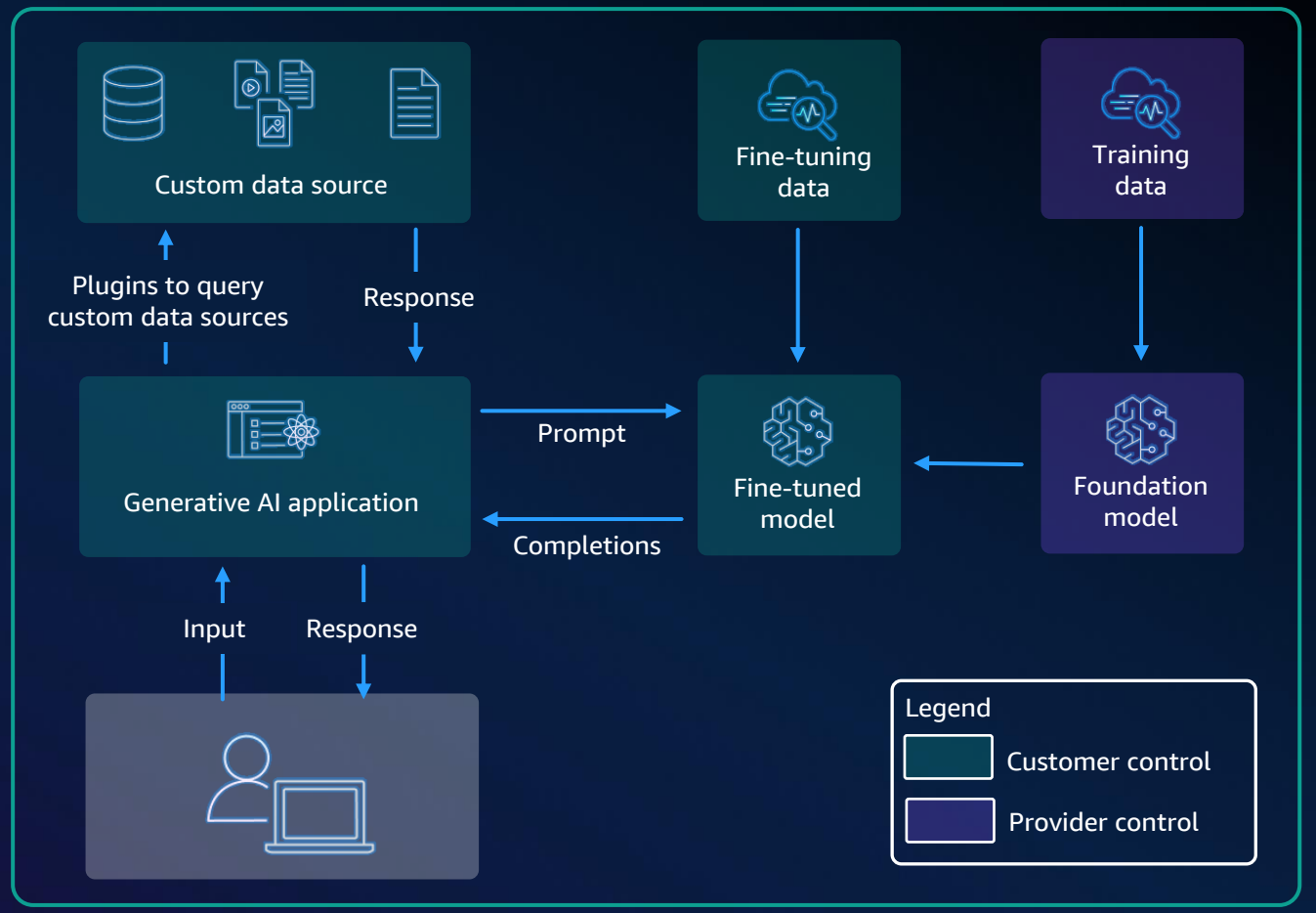


# Scope 4: Fine-tuned models

SCOPE 4

DATA FLOW AND DATA OWNERSHIP

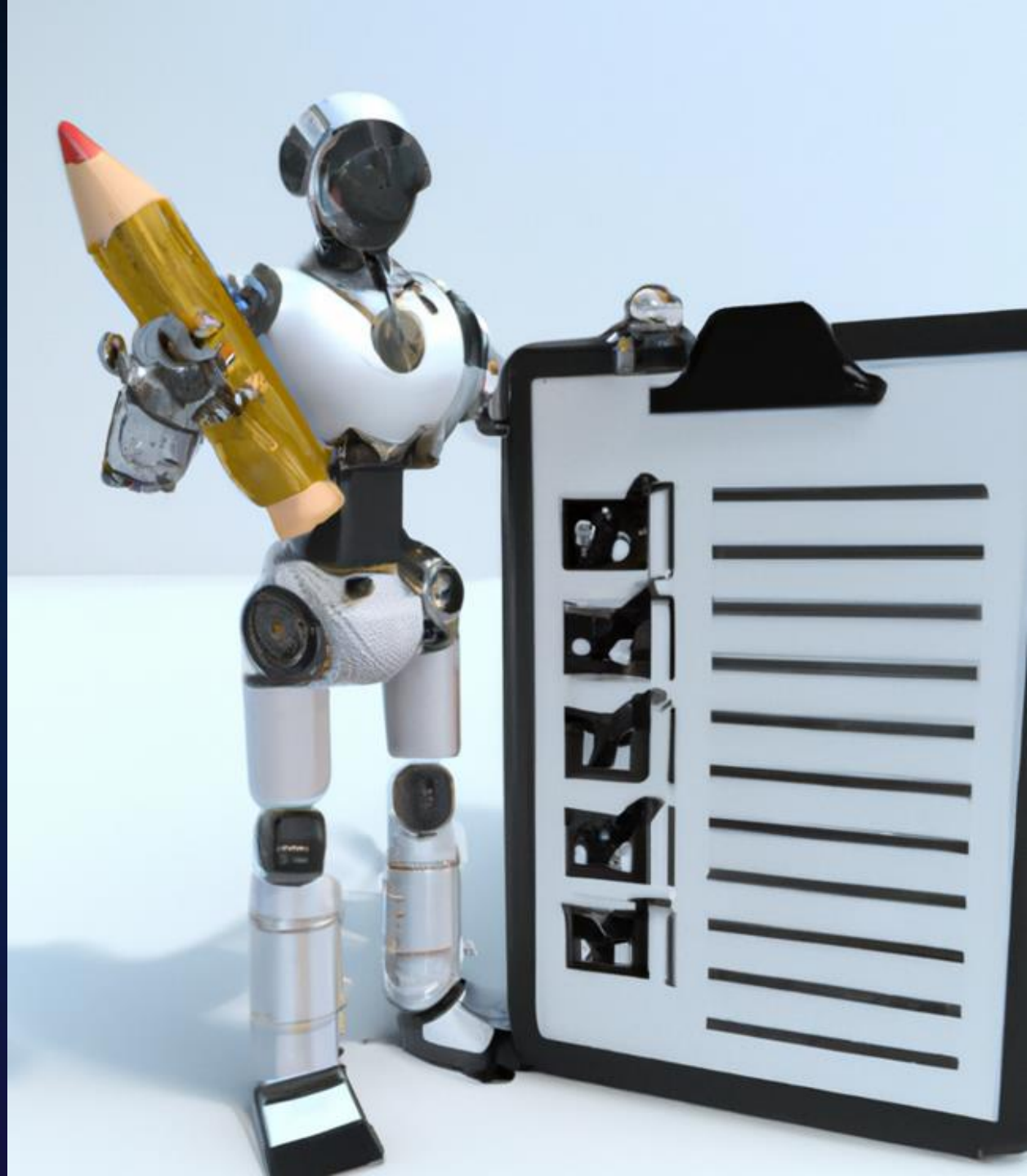
- Model is fine-tuned on your data to improve its responses
- Fine-tuned model can be offered as an API or can be hosted by you
- You can fine-tune an open sourced model or a closed-source model
- Examples: Amazon Bedrock customized models, Amazon SageMaker JumpStart



# 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

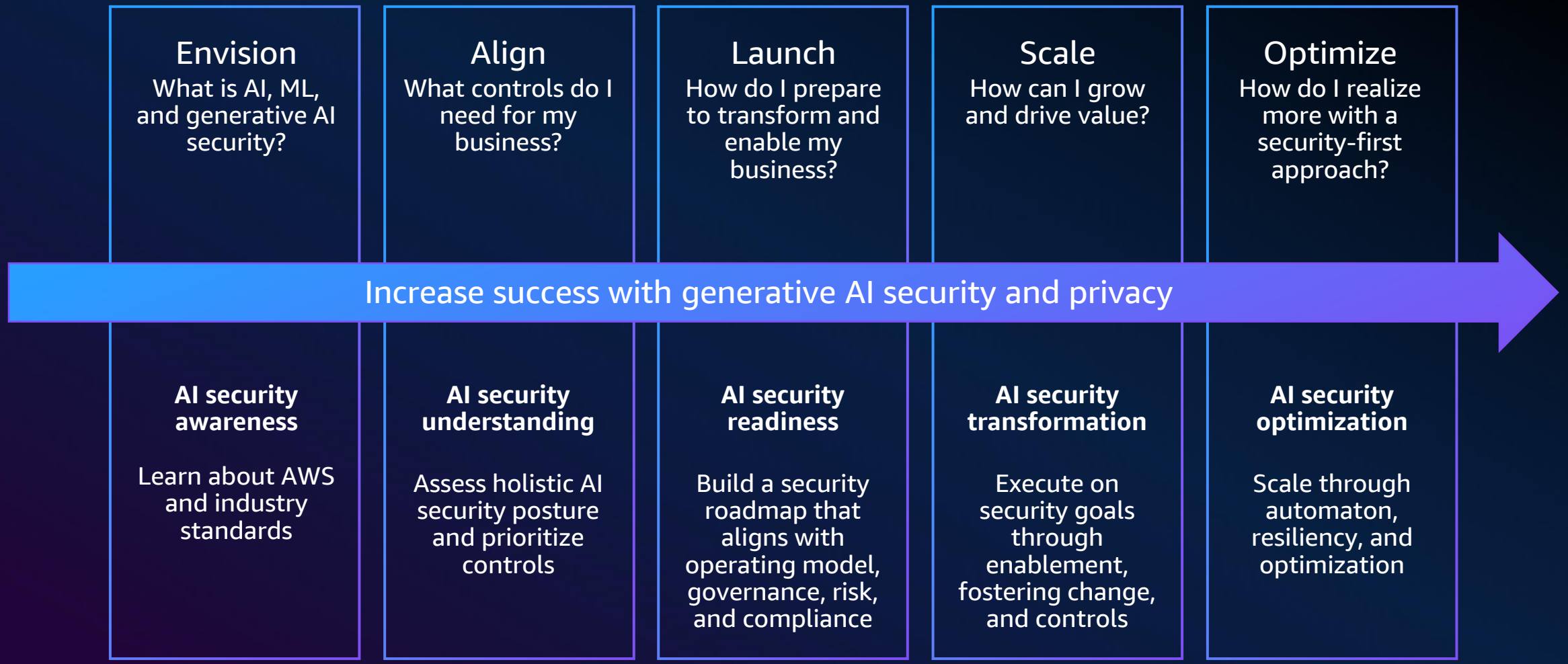
# ATLAS™

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

# Whiteboard

# AI security and privacy journey – Where are you today?



# Additional resources

Introduction to the Generative AI 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)

skillbuilder.aws

# Build beyond

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