AWS State, Local, and Education Learning Days

Phoenix

3:15pm – 4:15pm

300 level

Deep Dive - Securing Generative AI

Securing Generative AI :
The Security Scoping Matrix

aws Learning Days
State, Local, and Education

Securing Generative AI : the security scoping matrix

Anthony Harvey (he/him)

Sr. Security Specialist SA
Amazon Web Services
aharveyr@amazon.com

aws Learning Days
State, Local, and Education

Agenda

Overview of generative AI

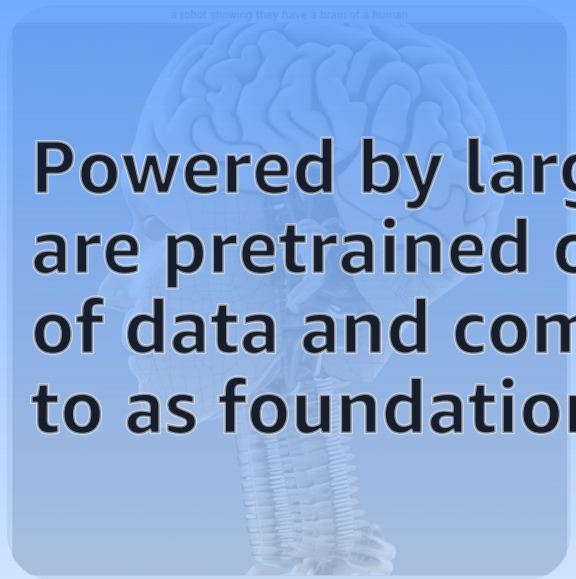
Mental model to scope generative AI use-cases

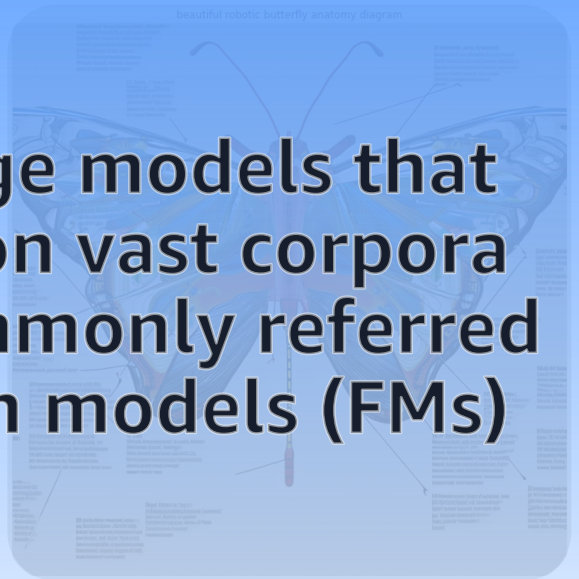
Securing generative AI use-cases

Responsible AI

Key take-aways

What is generative AI?



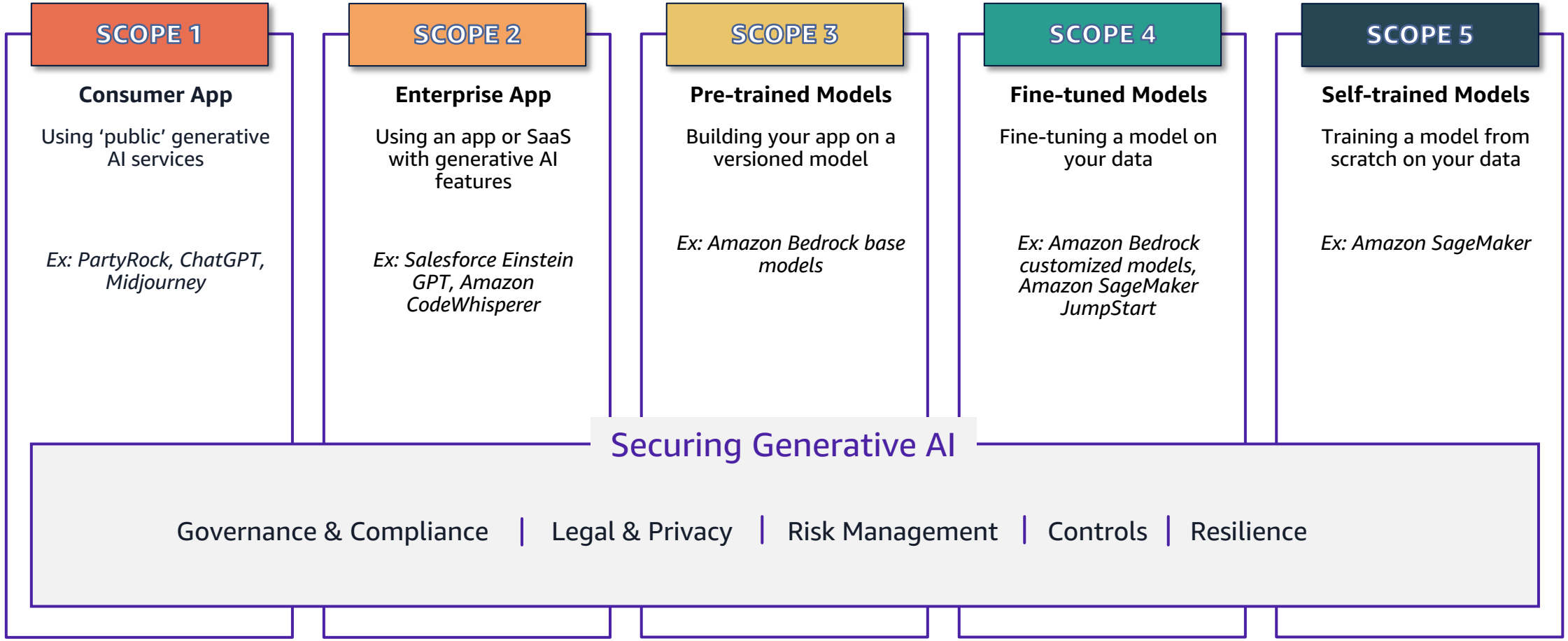


Powered by large models that are pretrained on vast corpora of data and commonly referred to as foundation models (FMs)

Creates new content and ideas, including conversations, stories, images, videos, and music

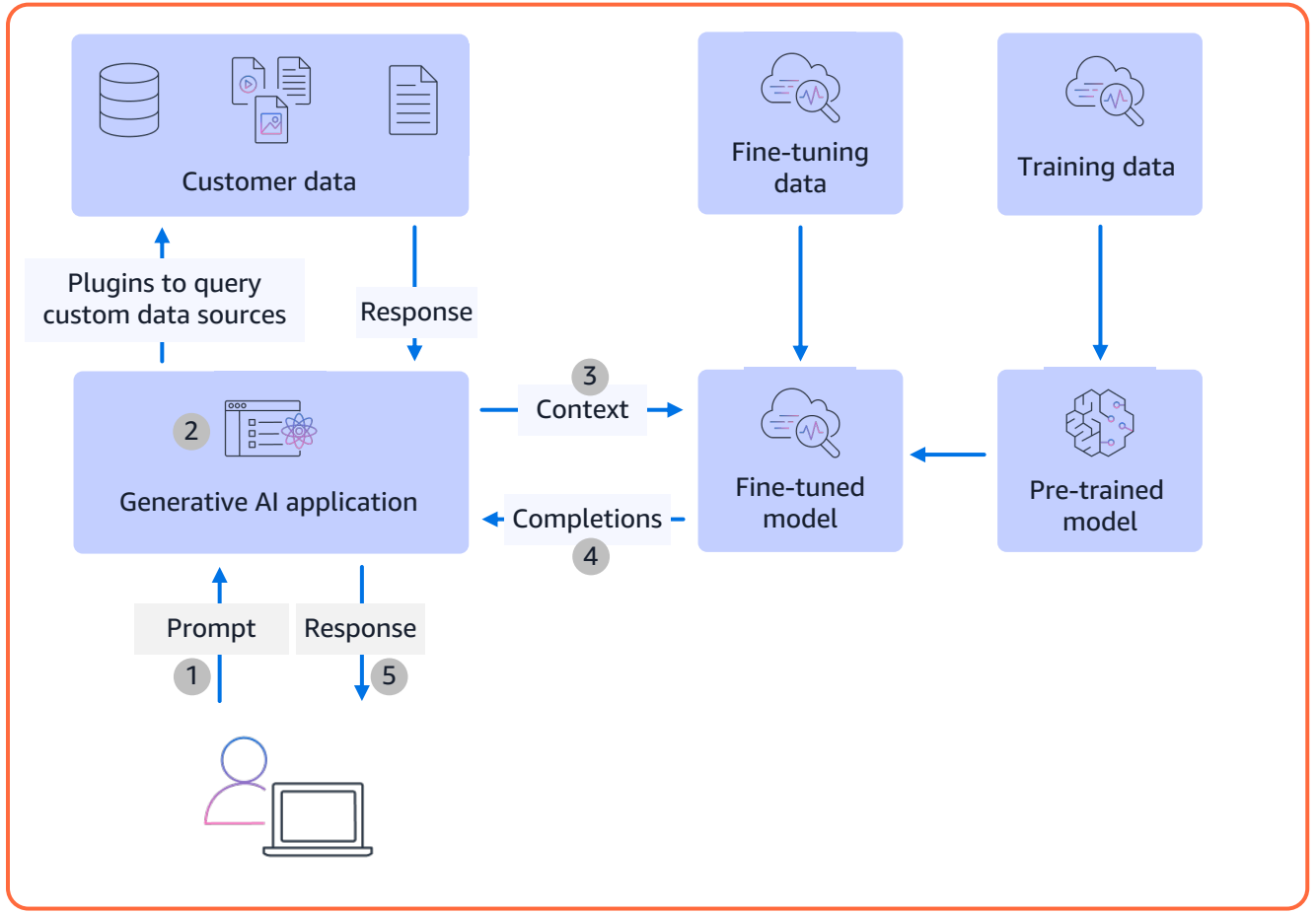
Generative AI Security Scoping Matrix

A MENTAL MODEL TO CLASSIFY USE-CASES



Data flows in a generative AI application

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



Common guidance across all Scopes

Securing the use of generative AI in your organization

Don’t

Outright ban generative AI technologies

Do

Empower users by creating policies for the use of generative AI technologies

Refer to and reinforce your existing data policies

Implement controls to remove harmful/inappropriate/incorrect content from inputs and outputs

Threat model your generative AI applications

Track model versions as part of your software bill of materials (SBOM)

Generative AI compliance concerns

AI compliance is an evolving space

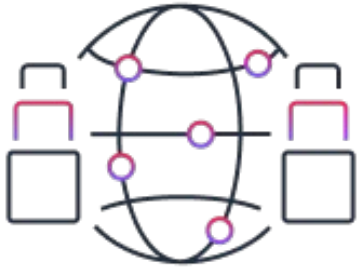
ISO 42001 – First global standard

Currently over 800 AI policy initiatives from 69 countries, territories and the EU (OECD.AI) including the EU Artificial Intelligence (AI) Act and others

Existing general privacy regulations (eg: GDPR, CCPA, and others)

Existing standards frameworks (eg: ISO27090, ISO38507, ISO23053:2022)

Resilience considerations



Prompt Engineering:

- Monitor input size against limits for your models
- Store a copy of your prompt and output data if needed

Additional considerations:

- Apply resilient app design patterns (backoffs and retries, graceful degradation)
- High Availability and Disaster Recovery strategy for vector databases

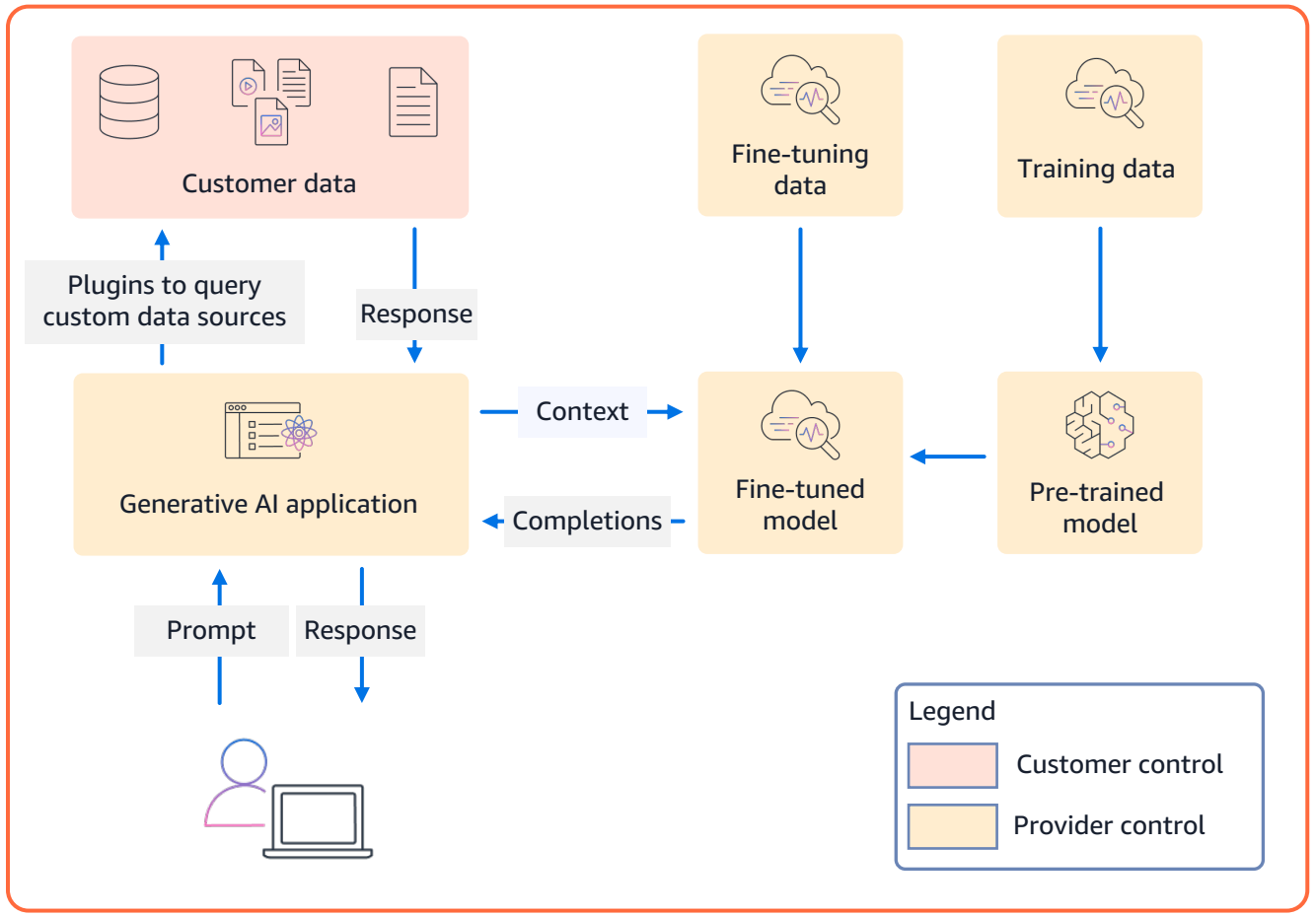
Scope 1: Consumer App

Scope 1: Consumer App

SCOPE 1

DATA FLOW AND DATA OWNERSHIP

- Consumer off-the-shelf apps, aimed at home and non-enterprise users
- Can be free or paid for, utilizing standard contract terms but not considered an enterprise agreement
- Typically provided as a web UI
- Examples include: PartyRock, ChatGPT, Midjourney, Google Bard



Governance & Compliance

SCOPE 1

- Create generative AI usage guidelines and educate workforce on the acceptable use of consumer services
- Establish process/guidelines for output validation
- Develop compliance monitoring & reporting processes

Governance & Compliance

SCOPE 1

Example – Team 8 Acceptable use policy

Acceptable Use Policy

The following is a list of guidelines for employees to follow when making use of GenAI generically, including ChatGPT. Employees should be trained on the appropriate use of the GenAI system and the relevant policies and regulations governing its use.

Violations of GenAI usage policies may result in disciplinary action, up to and including termination of employment.

- Employees must not disclose confidential or proprietary information to a GenAI technology, directly or through a third party application, unless through following the guidelines of the policy.
- Employees must use GenAI in a respectful and professional manner, refraining from using profanity, discriminatory language, or any other form of communication that could be perceived as offensive.
- Employees must comply with all relevant laws and regulations, including those related to data privacy and information security, according to our internal policy [policy name, link to the policy].
- Employees should report any concerns or incidents related to the use of GenAI to their supervisor or the appropriate department.

https://team8.vc/wp-content/uploads/2023/04/Team8-Generative-AI-and-ChatGPT-Enterprise-Risks.pdf

Legal & Privacy

SCOPE 1

- Treat prompts and outputs as public
- Don’t input any PII, confidential, proprietary, or company IP data (refer to your data classification and handling policy)
- Understand service provider’s terms of service and privacy policy, including who has access to the data
- Understand any legal implications of using outputs commercially
- Recognize terms of service and privacy policy on consumer apps can change without notice at any time

Risk Management

SCOPE 1

- Perform third-party risk assessment with existing risk management **framework**
- Establish security responsibility model with third party
- **Understand if and how the third party will use the your inputs/outputs and usage data**
- **Understand ownership of data, especially prompts and generated responses**

Controls

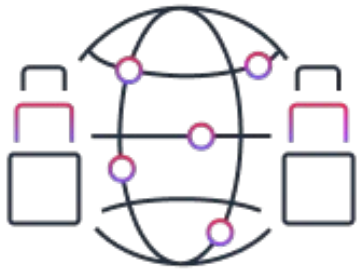
SCOPE 1

Most controls for third-party consumer-oriented services will be coarse-grained & perimeter based, such as:

- Cloud Access Security Brokers (CASB)
- Web proxies
- Data Loss Prevention (DLP) services

Resilience

SCOPE 1



- Incorporate third-party SLA's if available in availability goals, however consumer apps may not offer an SLA
- Increase client timeouts if necessary for extended latency for complex prompt completions

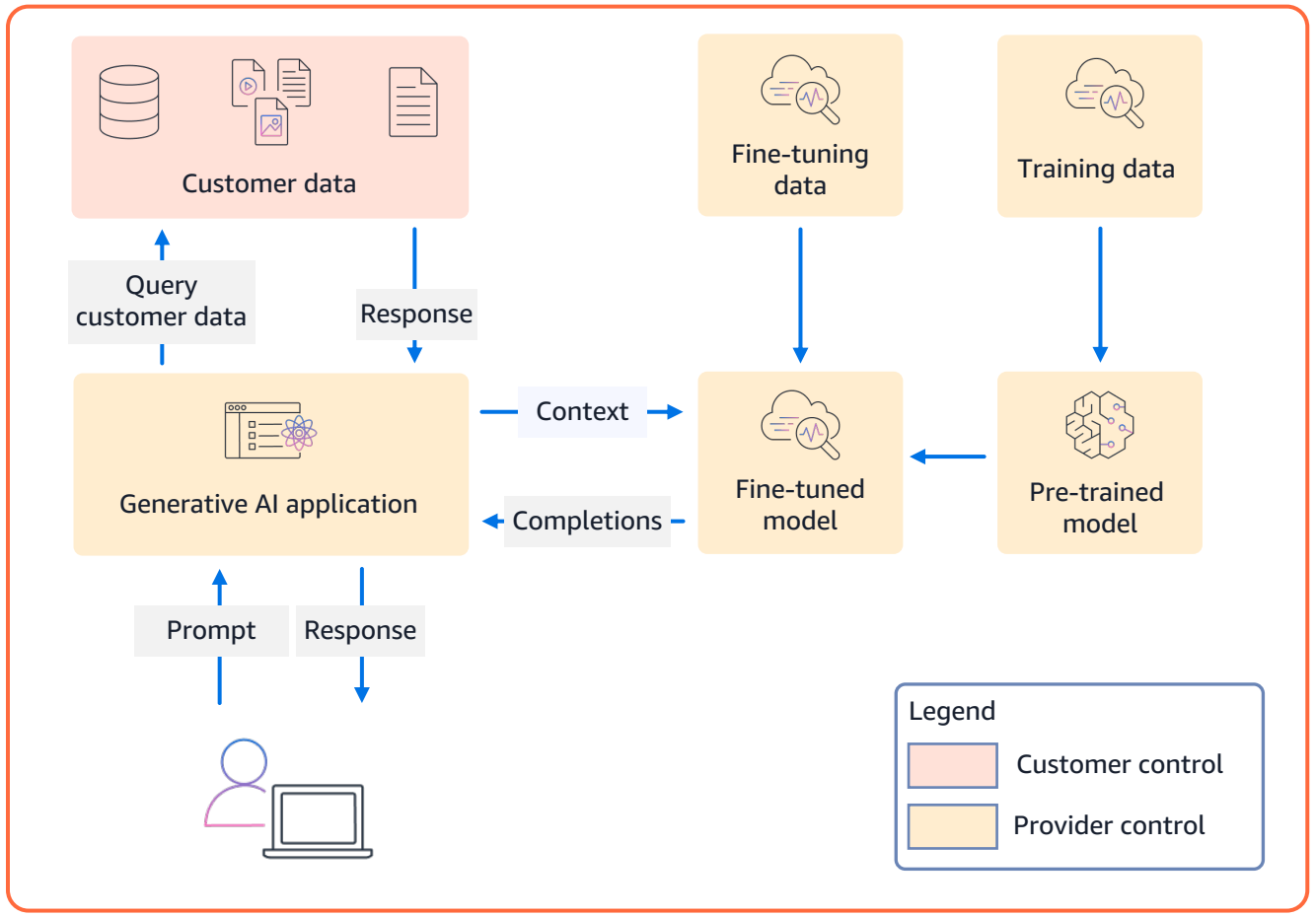
Scope 2: Enterprise App

Scope 2: Enterprise App

SCOPE 2

DATA FLOW AND DATA OWNERSHIP

- Generative AI features built into Enterprise Apps (desktop or SaaS)
- Enterprise level services, aimed at businesses and organizations for professional use
- Paid for under enterprise agreements or standard business contract terms
- Examples include: Amazon CodeWhisperer, Salesforce Einstein GPT



Governance & Compliance

SCOPE 2

- Create generative AI usage guidelines and educate workforce on the acceptable use of enterprise AI services
- Establish process/guidelines for output validation
- Develop compliance monitoring & reporting processes
- Understand the data flow of the service: does the service use downstream third party services?
- Align usage to regulatory requirements

Legal & Privacy

SCOPE 2

- Understand service provider's terms of service and privacy policy, including who has access to the data
- Understand any legal implications of using outputs commercially
- Determine acceptable data classification
- Data residency: where is the data stored and processed?
- Exercise any opt-out mechanisms to avoid enterprise data from being used for training or shared with other entities

Risk Management

SCOPE 2

- Perform third-party risk assessment with existing risk management **framework**
- Establish security responsibility model with third party
- **Understand if and how the third party will use the your inputs/outputs and usage data**
- **Understand ownership of data, especially prompts and generated responses**

Controls

SCOPE 2

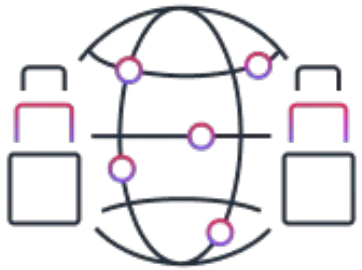
Make use of existing perimeter based controls such as:

- Cloud Access Security Brokers (CASB)
- Web proxies
- Data Loss Prevention (DLP) services

Enterprise apps may provide fine-grained access controls integrated into the app

Resilience

SCOPE 2



- Incorporate third-party SLA's if available in availability goals
- Increase client timeouts if necessary for extended latency for complex prompt completions

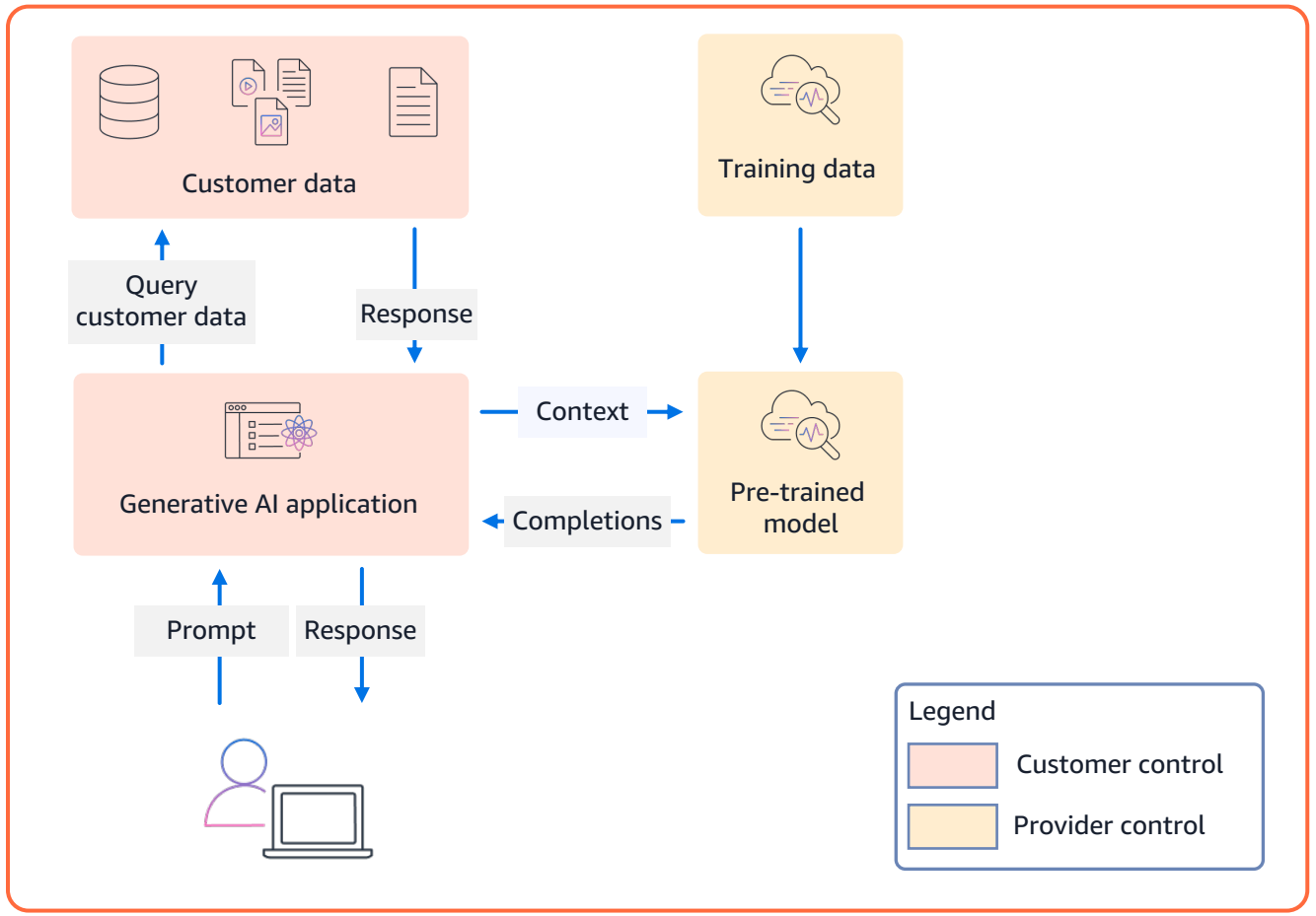
Scope 3: Pre-trained Models

Scope 3: Pre-trained Models

SCOPE 3

DATA FLOW AND DATA OWNERSHIP

- App uses pre-trained models provided by a model provider
- Models can be offered as an API service or can be hosted by you
- Models can be open-source or closed-source
- Examples include: Amazon Bedrock base models
  (*e.g., Amazon Titan, Cohere, Meta, Anthropic Claude, AI21 Labs Jurassic, Stability AI Stable Diffusion, etc.)*



Governance & Compliance

SCOPE 3

- Establish process/guidelines for output validation
- Develop compliance monitoring & reporting processes
- Align usage to regulatory requirements
- Understand the data used to train the model: ownership and quality

Legal & Privacy

SCOPE 3

- How are prompts and outputs protected when using a 3rd party model
- Is your data being used by the provider – understand why, and how your data is being protected
  - *“Amazon Bedrock doesn't use your prompts and continuations to train any AWS models or distribute them to third parties. Your training data isn't used to train the base Amazon Titan models or distributed to third parties” ***
- Is your data shared with other customers?
- What is the source of the data used to train the model – understand ownership and copyright challenges

*** https://docs.aws.amazon.com/bedrock/latest/userguide/data-protection.html*

Risk Management

SCOPE 3

- Include threat modeling in risk management
- Consider the following for your application's existing threat model:
  - Prompt injection
  - Insecure output handling
  - Sensitive information disclosure
  - Insecure plugin design
  - Excessive agency
  - Supply chain vulnerabilities
  - Model denial of service

Source: OWASP Top 10 for LLMs

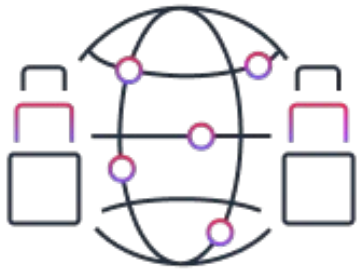
Controls

SCOPE 3

- Control who can use specific foundation models
- Control access to **inference endpoints**
- Fine-grained controls on access to data inside an LLM are not possible given current LLM technology
- Technologies such as WAF or DLP can be useful to filter malicious & sensitive inputs

Resilience

SCOPE 3



Self-hosted models:

- Be flexible on compute (such as EC2 Instance types)
- Reserve or pre-provision instances for static stability

API service:

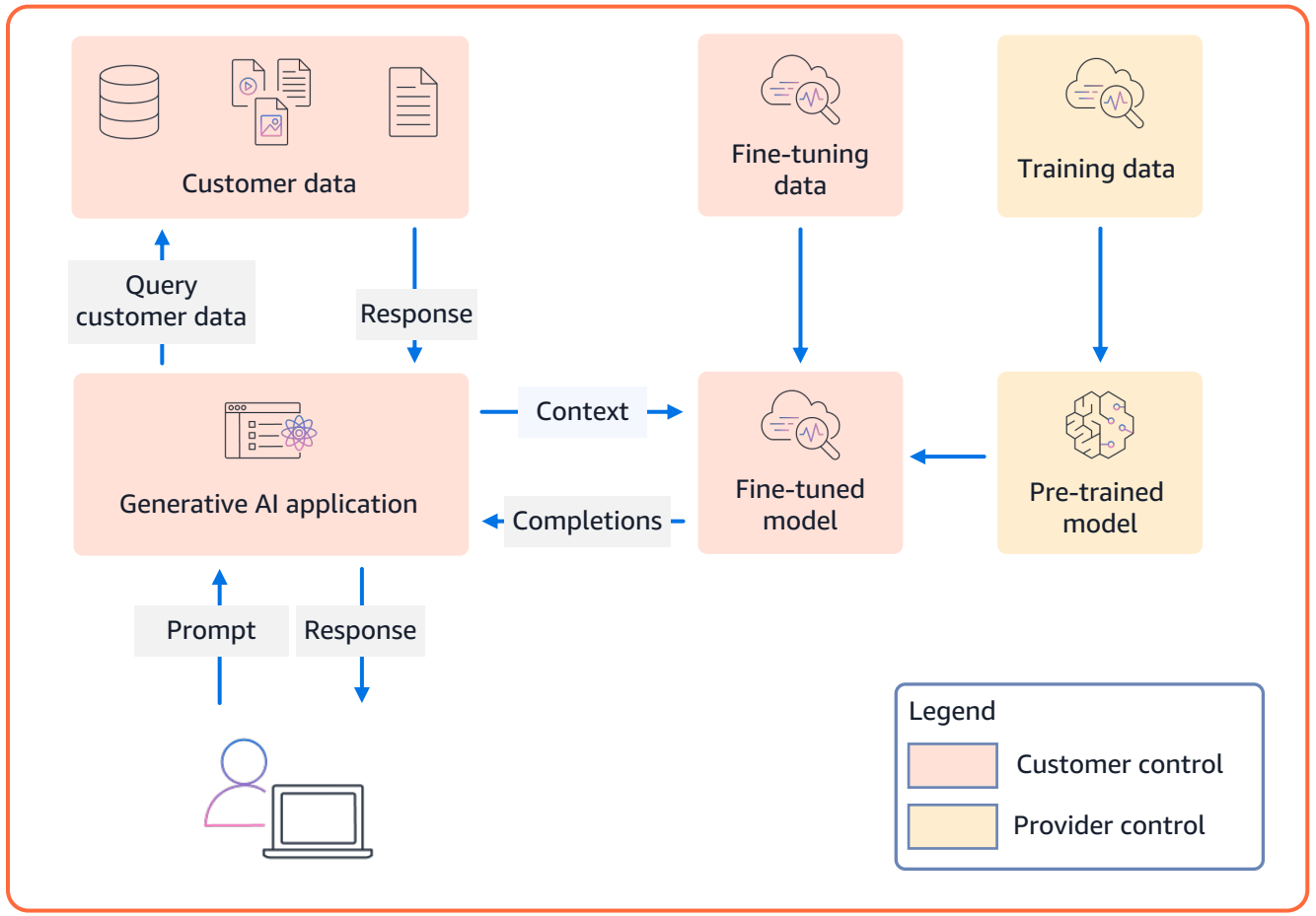
- Ensure service is available in your chosen regions (e.g. Amazon Bedrock)

Scope 4: Fine-tuned Models

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 include: Bedrock customized models, Amazon SageMaker JumpStart



Governance & Compliance

SCOPE 4

- Understand the data used to fine-tune the model: ownership and quality
- Establish process/guidelines for output validation
- Develop compliance monitoring & reporting processes
- Align usage to regulatory requirements
- Control access to fine-tuned model

Legal & Privacy

SCOPE 4

- What is the source of the data used to fine-tune the model – understand ownership and copyright challenges
- Fine-tuned model inherits the data classification of the data used for fine-tuning
- Avoid tuning a model on PII directly; it is not currently possible to "unlearn" data in a model without completely retraining
- Restrict access to the fine tuned model given its data classification

Risk Management

SCOPE 4

- Include threat modeling in risk management
- Consider the following for your application's existing threat model:
  - Prompt injection
  - Insecure output handling
  - Sensitive information disclosure
  - Insecure plugin design
  - Excessive agency
  - Supply chain vulnerabilities
  - Model denial of service
  - Training data poisoning
  - Model theft

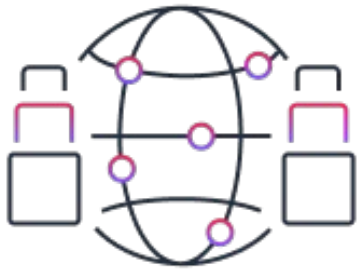
Controls

SCOPE 4

- Control who can use specific foundation models
- Control access to inference endpoints
- Fine-grained controls on access to data inside an LLM are not possible given current LLM technology
- Technologies such as WAF or DLP can be useful to filter malicious & sensitive inputs
- Protect the model artifacts and the inference endpoints
  - Identity and access management
  - Encryption
  - Monitoring

Resilience

SCOPE 4



Self-hosted models:

- Be flexible on compute (such as EC2 Instance types)
- Reserve or pre-provision instances for static stability
- Data management strategy (i.e. copying models across regions)

API service:

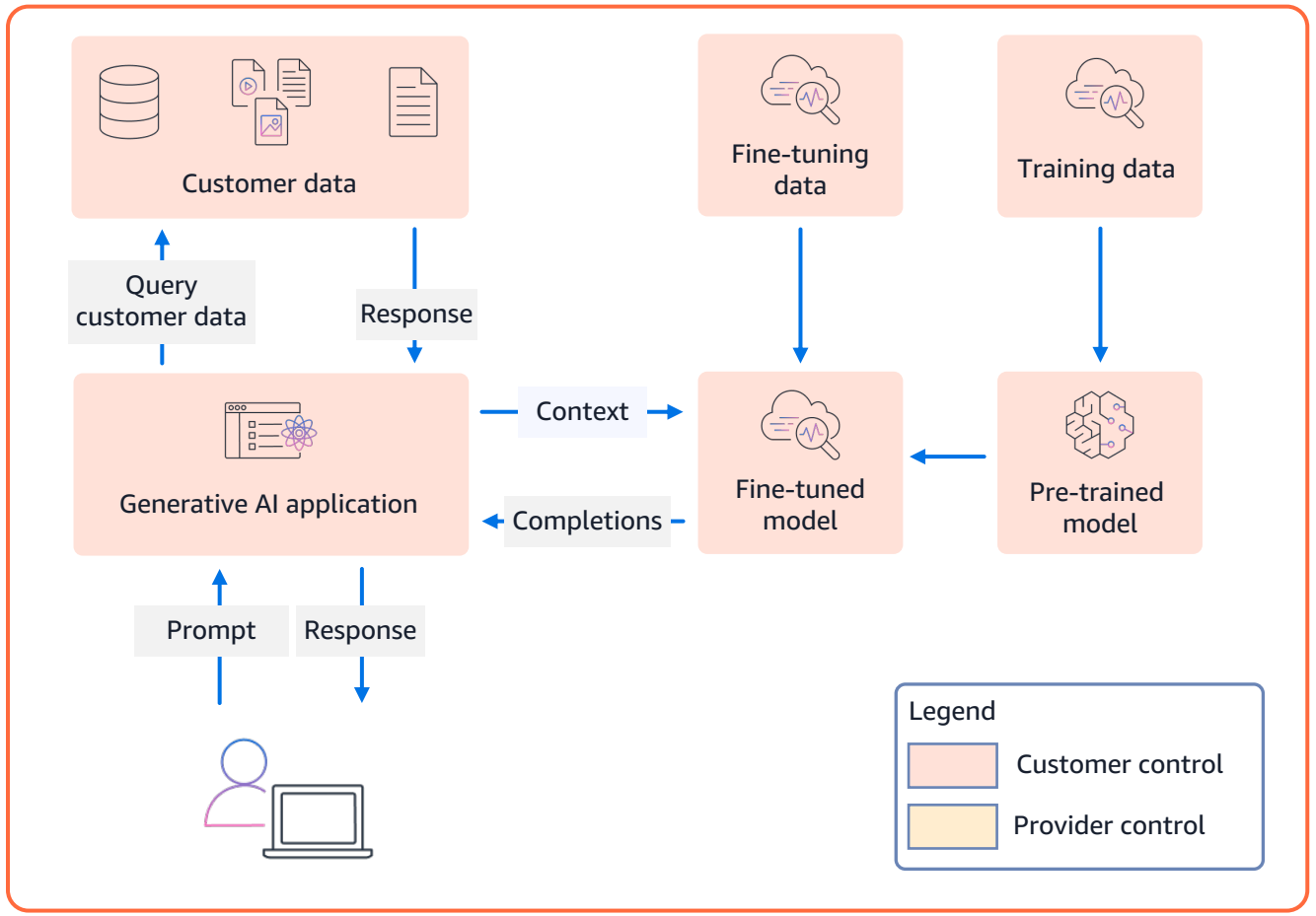
- Ensure availability of service in your chosen regions (e.g. Amazon Bedrock)
- May need to fine-tune in multiple regions

Scope 5: Self-trained Models

Scope 5: Self-trained Models

DATA FLOW AND DATA OWNERSHIP

- You train a model from scratch using your training data
- You control all aspects of the training process and optionally the fine-tuning process
- Examples include: Amazon SageMaker



Governance & Compliance

SCOPE 5

- Govern and protect the training data according to your existing data policies
- Trained model inherits the data classification of the training data

Legal & Privacy

SCOPE 5

- Avoid training a model on PII directly; it is not currently possible to "unlearn" data in a model without completely retraining
- You are the model provider and must take on the responsibility to clearly communicate how data will be used, stored, and maintained through a EULA
- Limit use of customer data (prompts and outputs) to the minimum needed, to limit exposure and risk

Risk Management

SCOPE 5

- Include threat modeling in risk management
- Consider the following for your application's existing threat model:
  - Prompt injection
  - Insecure output handling
  - Sensitive information disclosure
  - Insecure plugin design
  - Excessive agency
  - Training data poisoning
  - Supply chain vulnerabilities
  - Model denial of service
  - Model theft

Controls

SCOPE 5

Fine-grained controls on access to data inside an LLM are not possible given current LLM technology

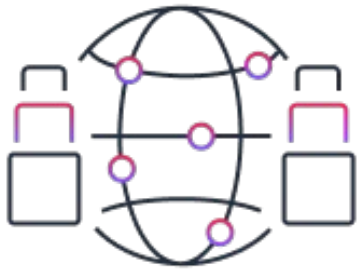
Technologies such as WAF or DLP can be useful to filter malicious & sensitive inputs

Protect the model artifacts and the inference endpoints:

- **Identity and access management**
- Encryption
- Monitoring

Resilience

SCOPE 5



- Be flexible on compute (such as EC2 Instance types)
- Reserve or pre-provision instances for static stability
- Save checkpoints frequently during training

Responsible AI

Responsible AI Dimensions

FAIRNESS

Considering impacts on different groups of stakeholders

EXPLAINABILITY

Understanding and evaluating system outputs

CONTROLLABILITY

Having mechanisms to monitor and steer AI system behavior

SAFETY

Preventing harmful system output and misuse

PRIVACY & SECURITY

Appropriately obtaining, using and protecting data and models

GOVERNANCE

Incorporating best practices into the AI supply chain, including providers and deployers

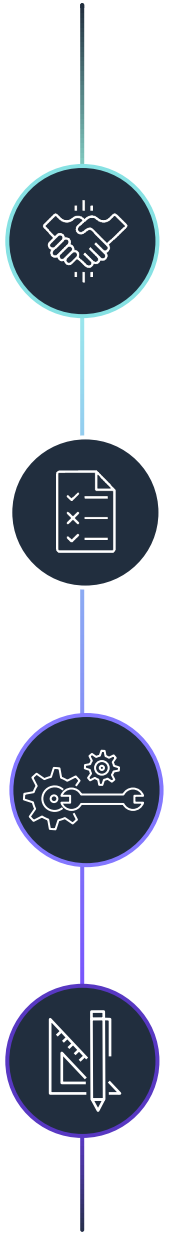
TRANSPARENCY

Enabling stakeholders to make informed choices about their engagement with an AI system

VERACITY & ROBUSTNESS

Achieving correct system outputs, even with unexpected or adversarial inputs

Responsible AI: Best practices



Put your people first

Assess risk on a (use) case-by-case basis

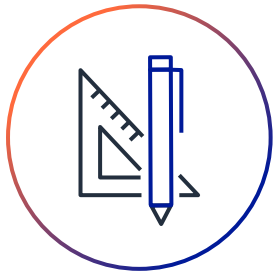
Iterate across the AI lifecycle

Test, test again, and then test again

Educating people on responsible AI is paramount

AWS Machine Learning University (MLU)—free to all

Responsible use of machine learning guide



AWS AI and ML Scholarship program

Training and certification

New AWS Machine Learning University (MLU) Course on Fairness and Bias Mitigation

NEW

Free, public, hands-on training on fairness and bias

Taught by the same Amazon scientists that train our own AWS employees on machine learning

Over 9 hours of content available to everyone

>> Get started today

What can you do now?

Continue to explore responsible AI

Check out learning resources from AWS, including training, guidance, and research.

Educate your organization

Pass along what you learn to members of your team.

Identify needs & consider risks

Think carefully about your organization's needs and where AI fits. Engage with ML experts at AWS to get started.

Choose diverse talent

Strengthen your AI team by reflecting diversity within it.

The Generative AI Innovation Center can help

Art of the possible	Explore and experiment safely, reliably, and cost-effectively
Responsible use	Guardrails to reduce risk—more control and focus across bias and fairness, privacy and security, explainability, and governance for enterprise content and workflows
Model selection	Select the best foundation models taking advantage of purpose-built infrastructure for training and inference
Time to value acceleration	Apply foundation models to real-world use cases with the Innovation Center as a trusted capable partner to demonstrate viability and accelerate path to production
MLOps	Balance efficiency, scalability, and risk reduction with a framework to manage the development and deployment of generative AI models

Thank you!

Anthony Harvey (he/him)
Sr. Security Specialist SA
Amazon Web Services
aharveyr@amazon.com

Please complete the survey for this session

Securing Generative AI : the security scoping matrix

aws Learning Days
State, Local, and Education