AWS re:Inforce

JUNE 10 - 12, 2024 | PHILADELPHIA, PA

APS201

Accelerate securely: The Generative AI Security Scoping Matrix

Matt Saner
He/Him
Senior Manager, Security Specialists
AWS

Mike Lapidakis
He/Him
Senior Manager, Specialists
AWS

Meet your speakers!

Matt Saner

Home: Indianapolis

Role: Security Specialist Leader

Hobbies: General aviation, home automation, and being the best dad I can be!

Mike Lapidakis

Home: Denver

Role: Specialist Leader

Hobbies: Chasing his kids, getting outside, learning by breaking stuff

Today, we'll 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. What are some **examples** of how security requirements change depending on scope?

Innovation can **transform industries**

Code generation

Document processing

Personalization

Summarization

Transforming data into insights

Disambiguation

Traditional AI/ML

Typically used to **predict** based on data

- Used to identify class membership, apply labels, forecast future performance, and other tasks

Examples

- Enterprise search for documentation
- Speech recognition for phone calls
- Computer vision for remote inspection
- Forecasting for future performance
- Anomaly detection in logs

Generative AI

Typically used to **create** based on data

- Used to generate new things (text, images, audio) based on large corpora (data sets) of existing examples

Examples

- Generate a draft report or summary
- Generate a chat for customer support
- Generate a photo or video for training
- Generate a voice for IVR directory
- Generate code for DevSecOps

Eating your security vegetables

Log monitoring
Broccoli

Data protection & classification
Peas

User education
Carrots

Patching
Spinach

Incident response planning
Bell peppers


Don’t boil the ocean

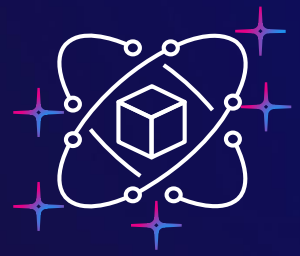
The “3 leg stool” of generative AI security

The “3 leg stool” of generative AI security

Security of generative AI

The “3 leg stool” of generative AI security

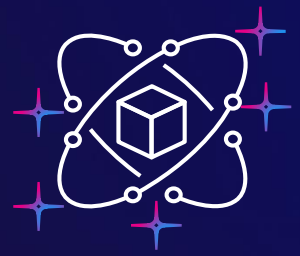
Security of generative AI



Generative AI for security

The “3 leg stool” of generative AI security

Security of generative AI



Generative AI for security

Security from generative AI

The “3 leg stool” of generative AI security

Security of generative AI

Generative AI for security

Security from generative AI

Generative AI Security Scoping Matrix

A MENTAL MODEL TO CLASSIFY USE CASES

Generative AI Security Scoping Matrix

A MENTAL MODEL TO CLASSIFY USE CASES

Buy

Generative AI Security Scoping Matrix

A MENTAL MODEL TO CLASSIFY USE CASES

SCOPE 1

Consumer app

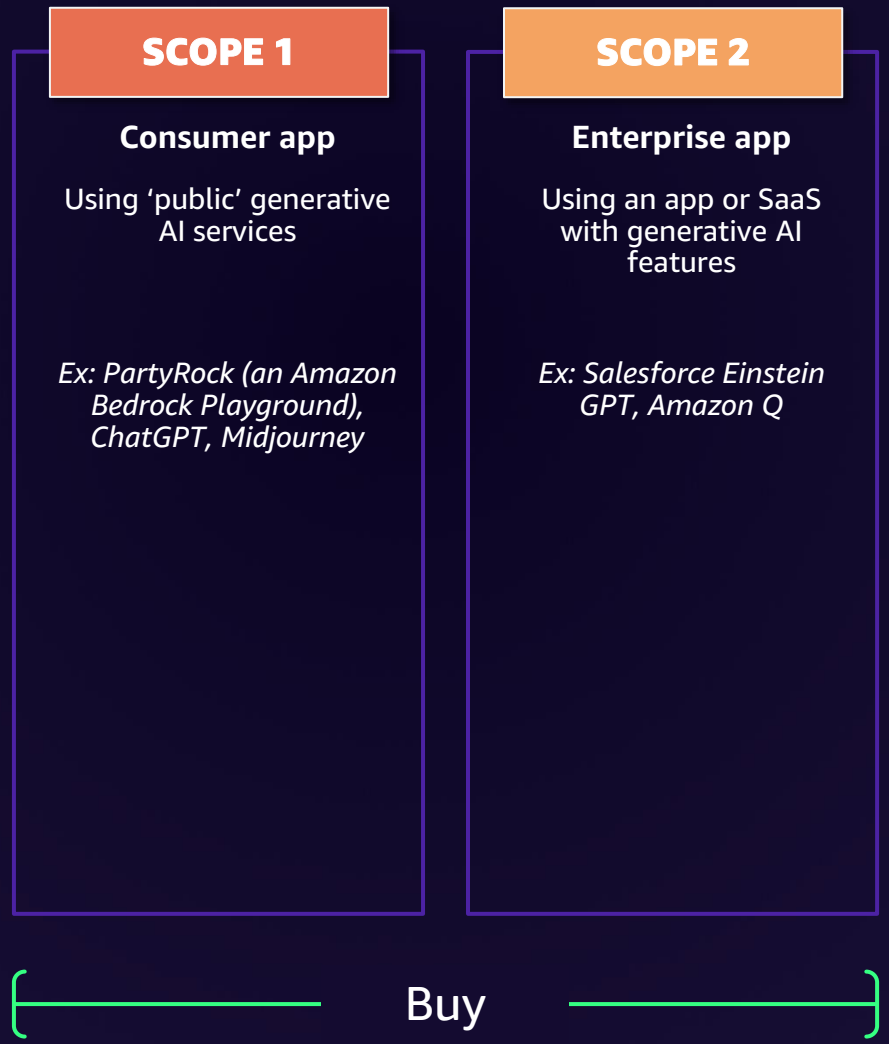
Using ‘public’ generative AI services

Ex: PartyRock (an Amazon Bedrock Playground), ChatGPT, Midjourney

Buy

Generative AI Security Scoping Matrix

A MENTAL MODEL TO CLASSIFY USE CASES



Generative AI Security Scoping Matrix

A MENTAL MODEL TO CLASSIFY USE CASES

Generative AI Security Scoping Matrix

A MENTAL MODEL TO CLASSIFY USE CASES

Generative AI Security Scoping Matrix

A MENTAL MODEL TO CLASSIFY USE CASES

SCOPE 1	SCOPE 2	SCOPE 3	SCOPE 4
Consumer app	**Enterprise app**	**Pre-trained models**	**Fine-tuned models**
Using 'public' generative AI services	Using an app or SaaS with generative AI features	Building your app on a versioned model	Fine-tuning a model on your data
Ex: PartyRock (an Amazon Bedrock Playground), ChatGPT, Midjourney	*Ex: Salesforce Einstein GPT, Amazon Q*	*Ex: Amazon Bedrock base models*	*Ex: Amazon Bedrock customized models, Amazon SageMaker JumpStart*

Buy

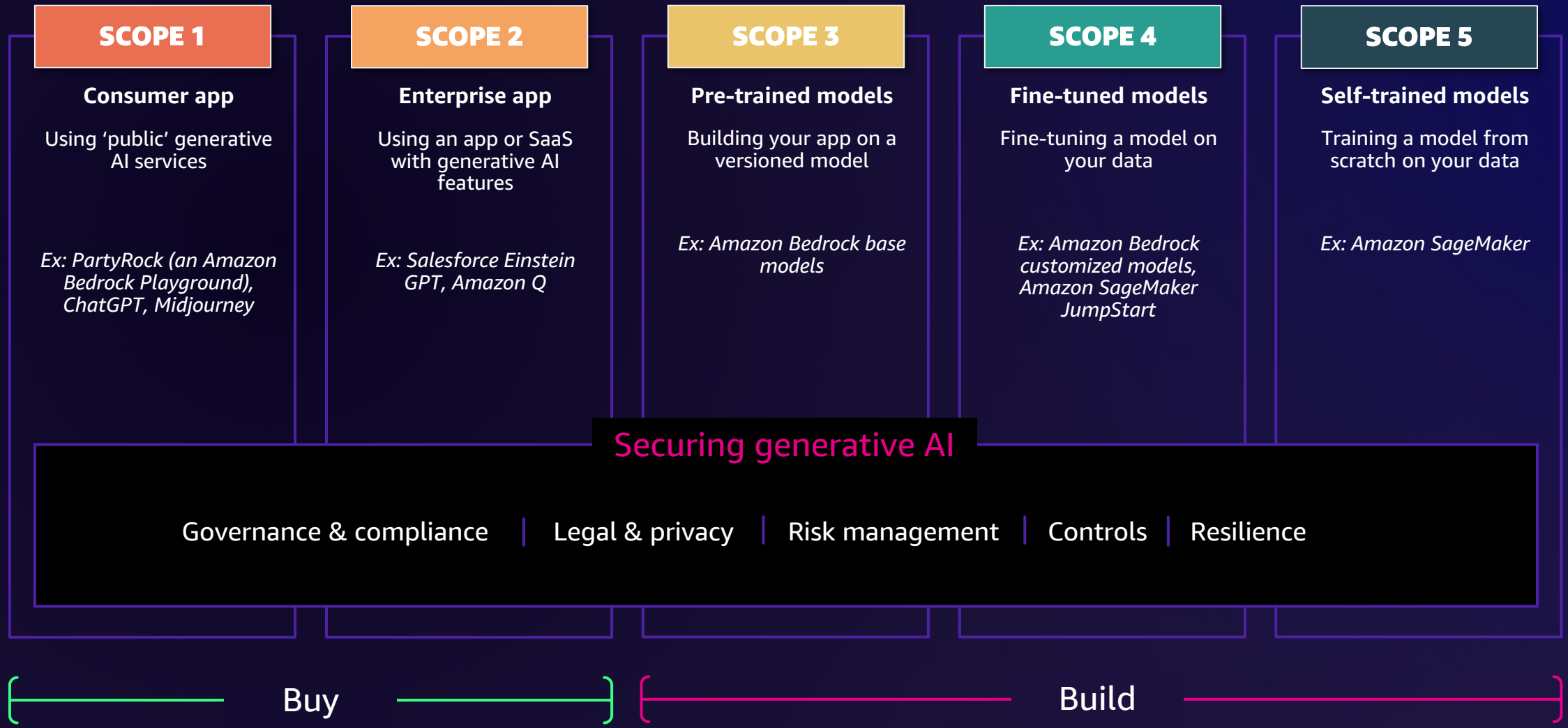
Build

Generative AI Security Scoping Matrix

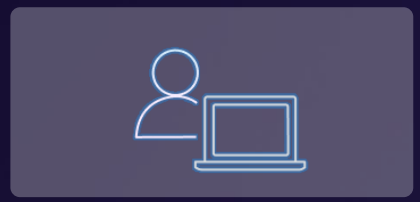
A MENTAL MODEL TO CLASSIFY USE CASES

Generative AI Security Scoping Matrix

A MENTAL MODEL TO CLASSIFY USE CASES

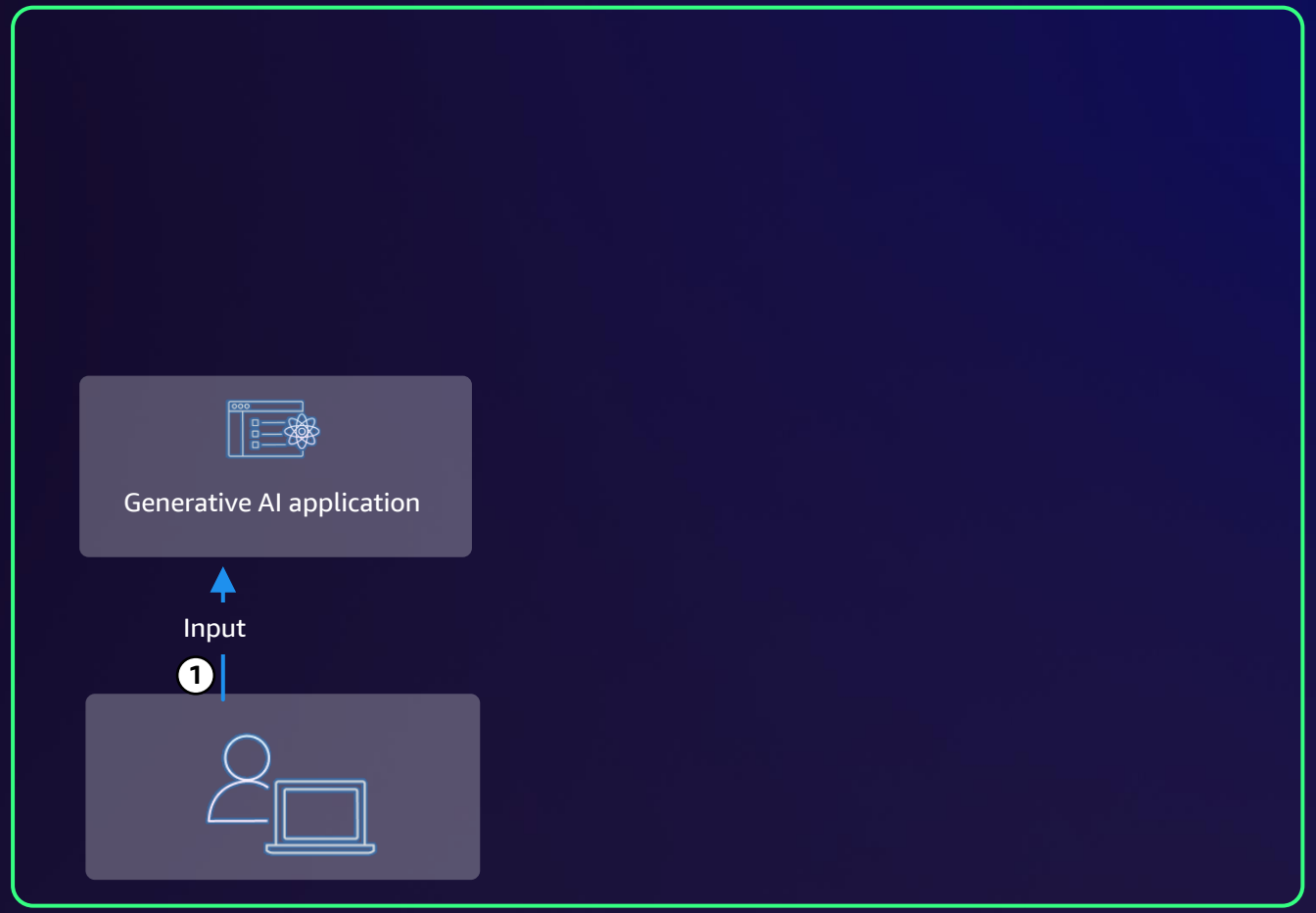


Data flows in a generative AI application



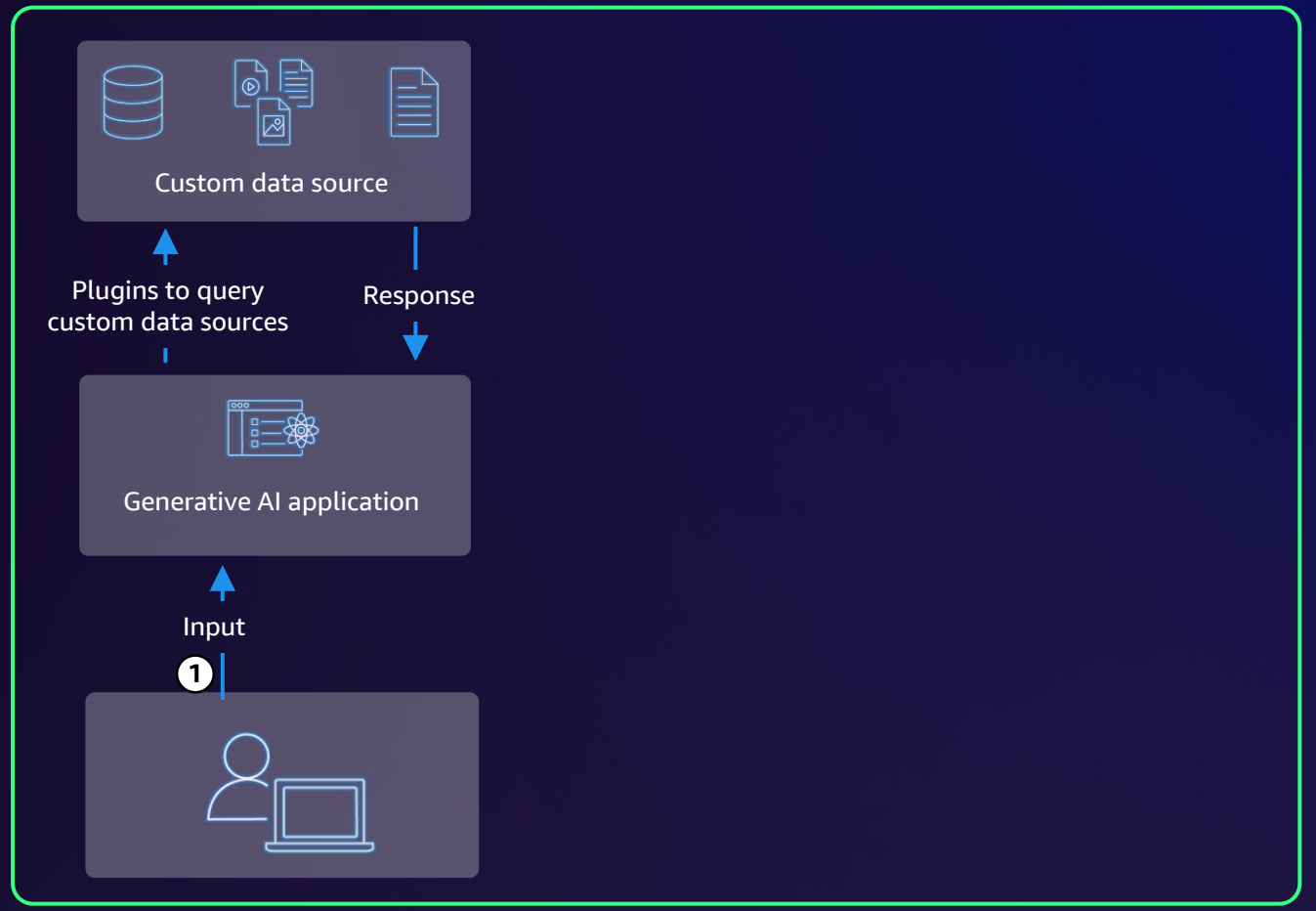
Data flows in a generative AI application

1. App receives input from user



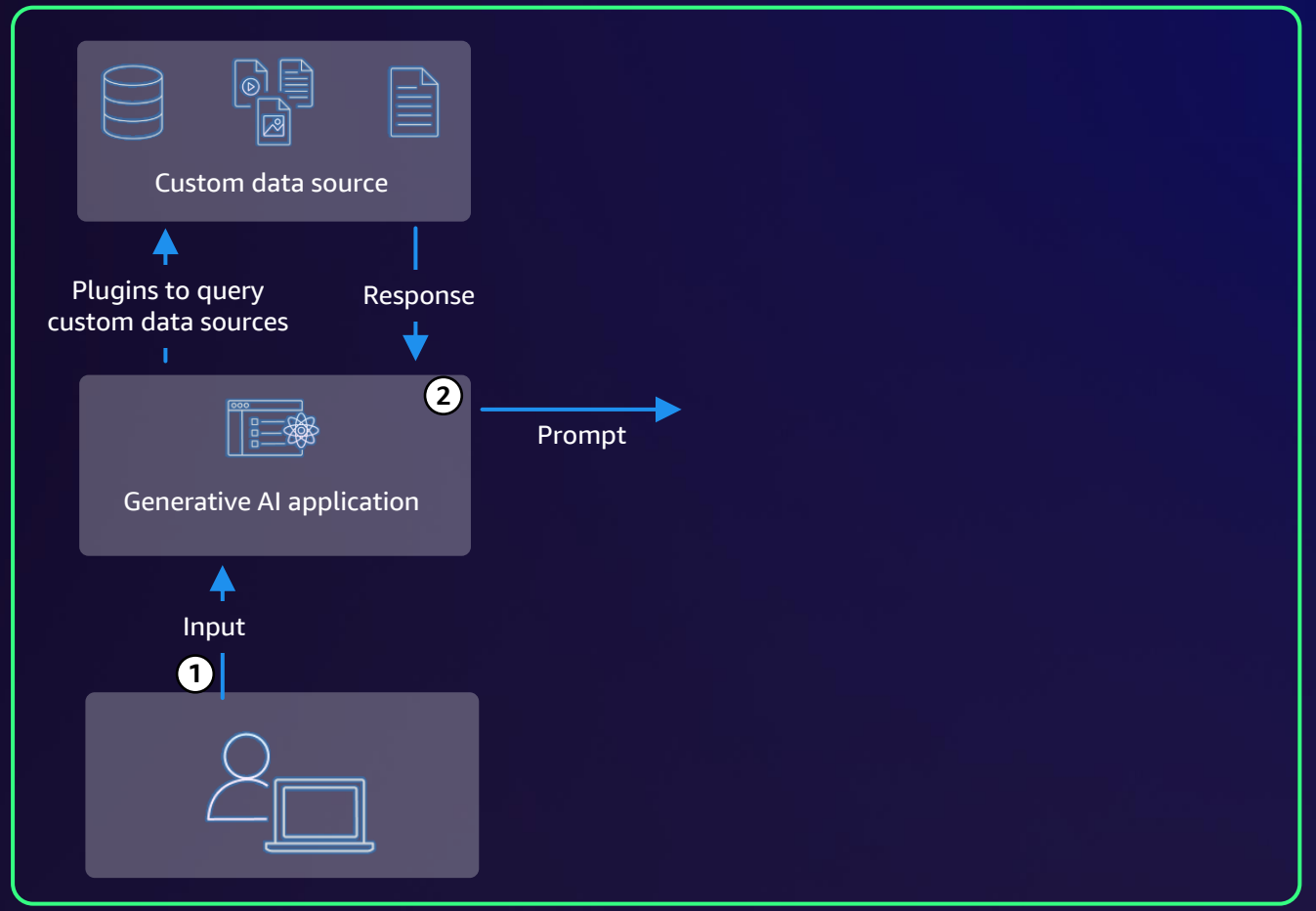
Data flows in a generative AI application

1. App receives input from user
    - [Optional] App queries data from custom data sources



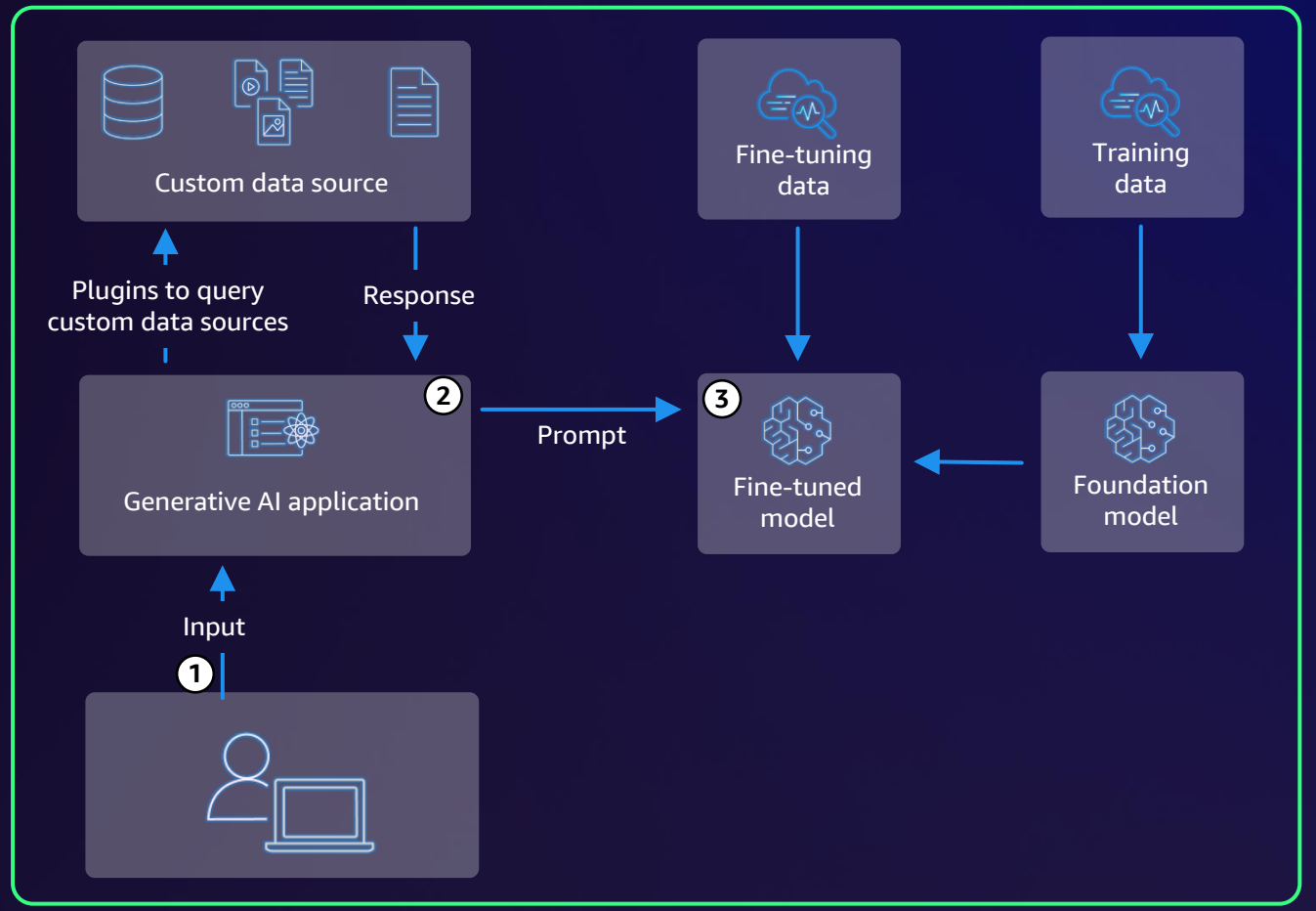
Data flows in a generative AI application

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



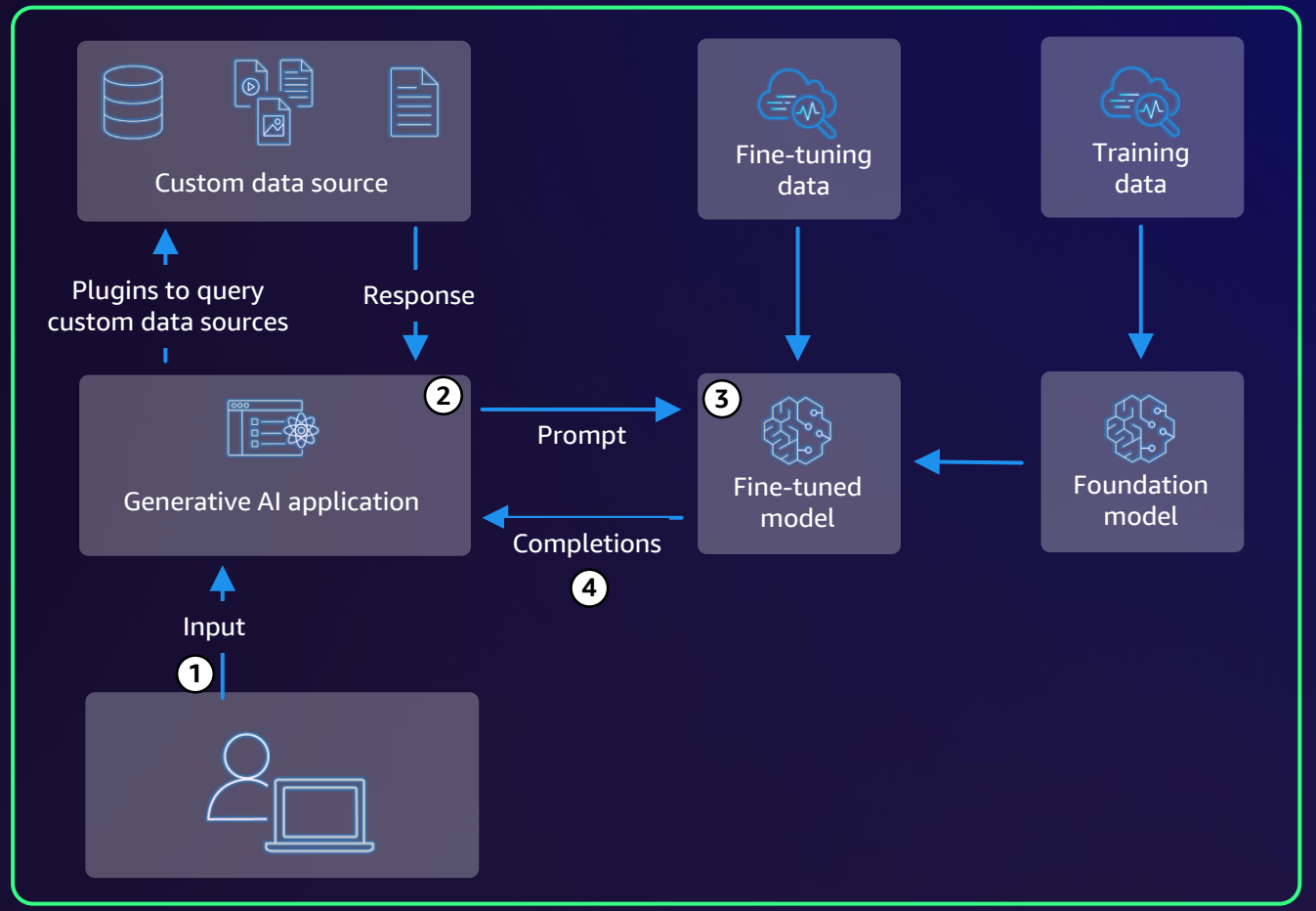
Data flows in a generative AI application

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 processed by a model (fine-tuned or pre-trained)



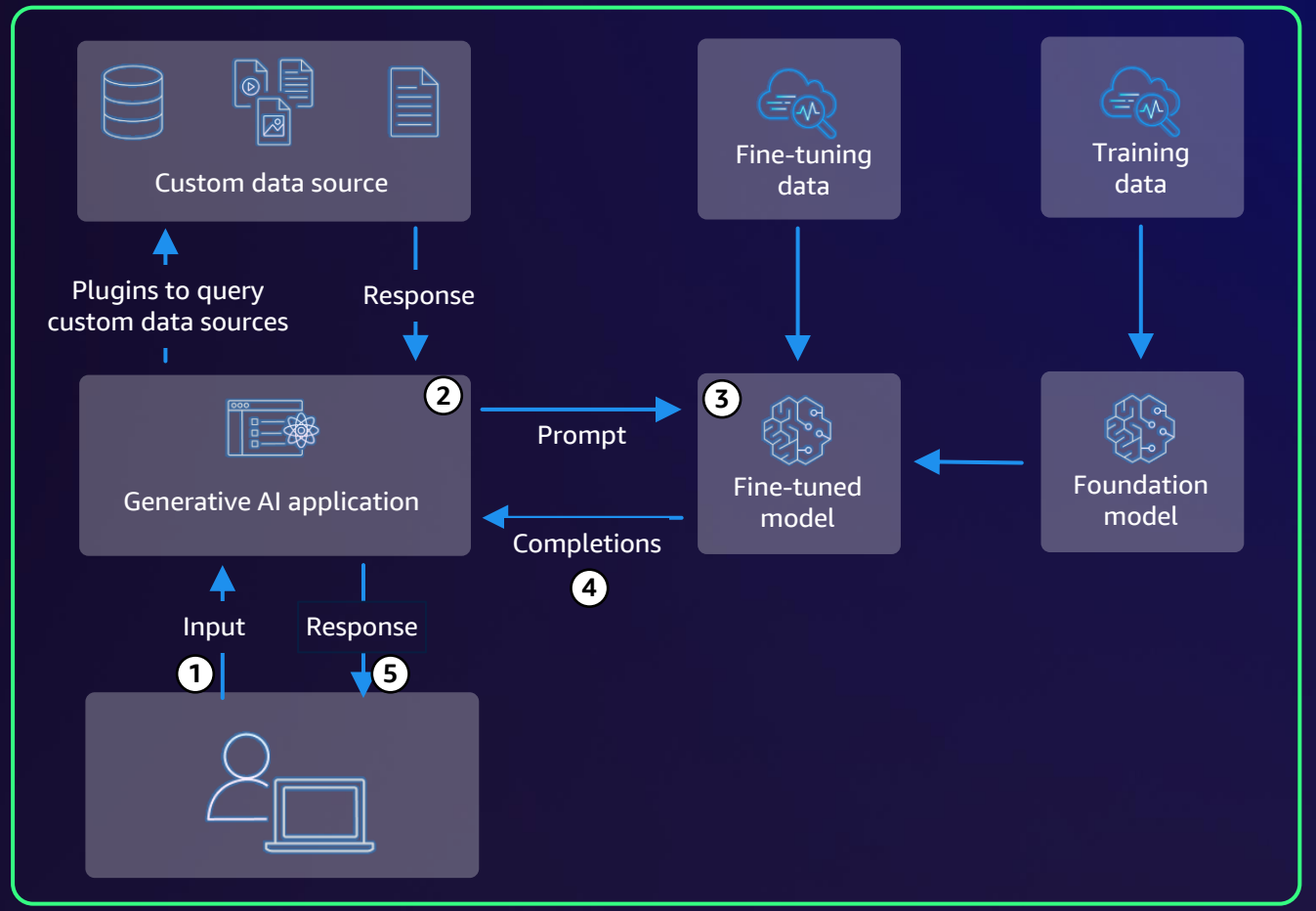
Data flows in a generative AI application

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 processed by a model (fine-tuned or pre-trained)
4. Completion is processed by app



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 processed 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

No global approach to govern the use of generative AI

Currently over 1000 AI policy initiatives from 70 countries, territories and the EU (OECD.AI) including the EU Artificial Intelligence (AI) Act (coming in 2024), Canadian Artificial Intelligence and Data Act (AIDA), and others are under review. 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

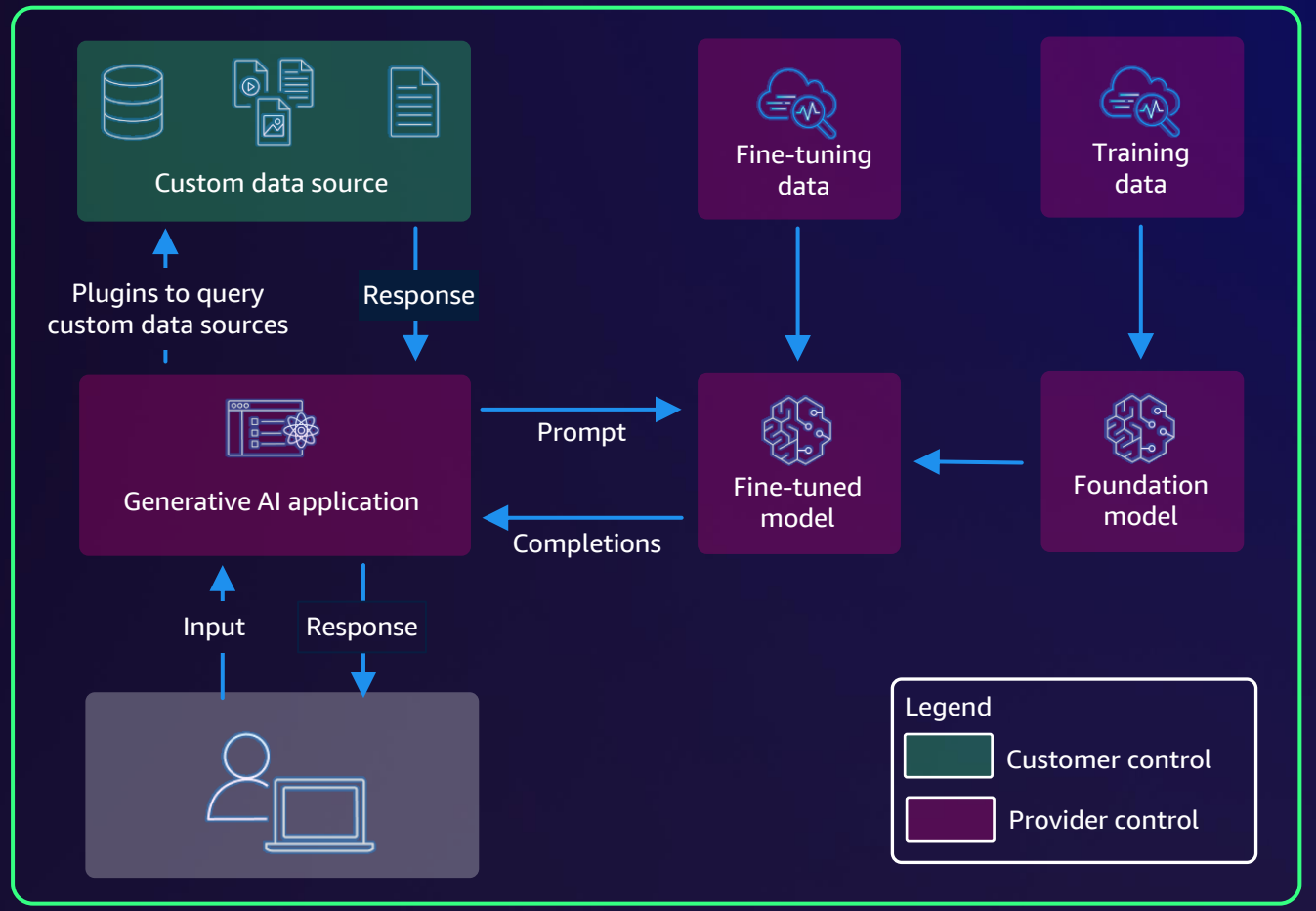
Scope 1: Consumer app

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 (an Amazon Bedrock Playground), OpenAI ChatGPT, Midjourney, Google Gemini



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

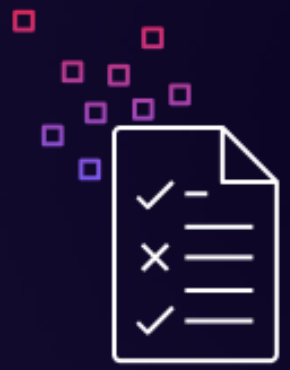
Legal & privacy

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

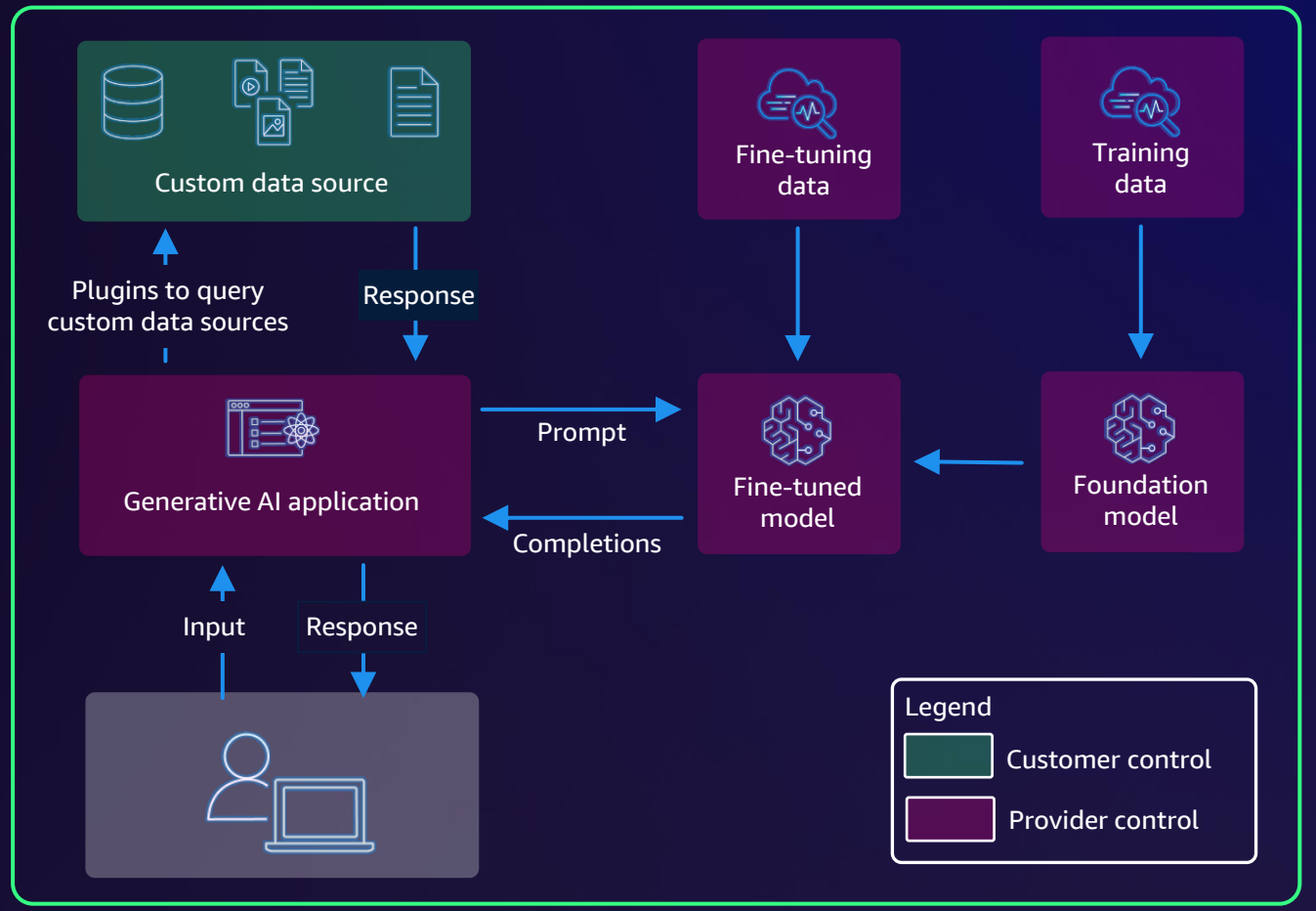
Scope 2: Enterprise app

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 Q, 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

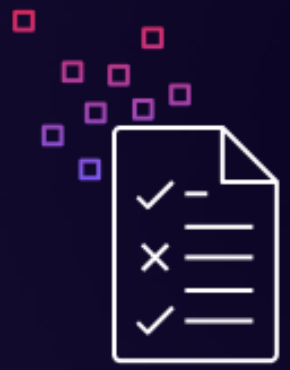
Legal & privacy

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
  - Provider more likely to provide SLAs
- Increase client timeouts if necessary for extended latency for complex prompt completions

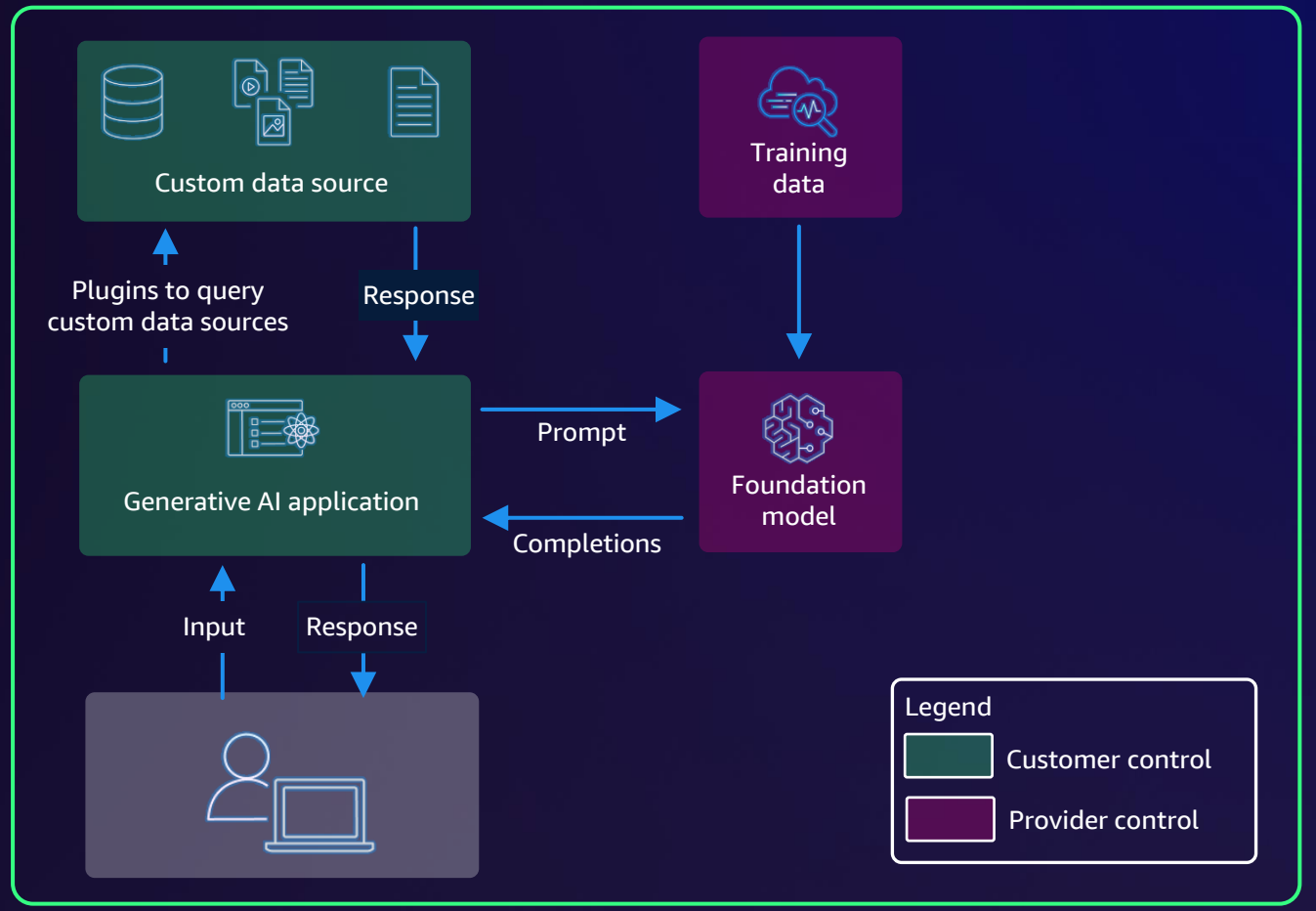
Scope 3: Pre-trained models

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

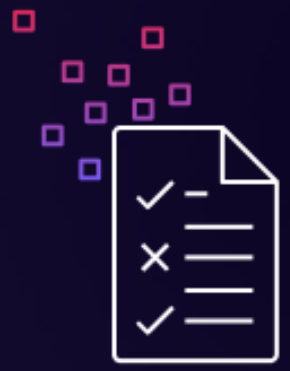
SCOPE 3

- 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" [1]
- Is your data shared with other customers?
- What is the source of the data used to train the model – understand ownership and copyright challenges

[1] 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

Guardrails for **Amazon Bedrock**

Safeguard your generative AI applications with your custom responsible AI policies

1. Apply guardrails to any foundation model (incl. fine-tuned models) and agents for Amazon Bedrock
2. Configure harmful content filtering based on your responsible AI policies
3. Define and disallow denied topics with short natural language descriptions
4. Redact or block sensitive information such as PII, and use of custom Regex

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)

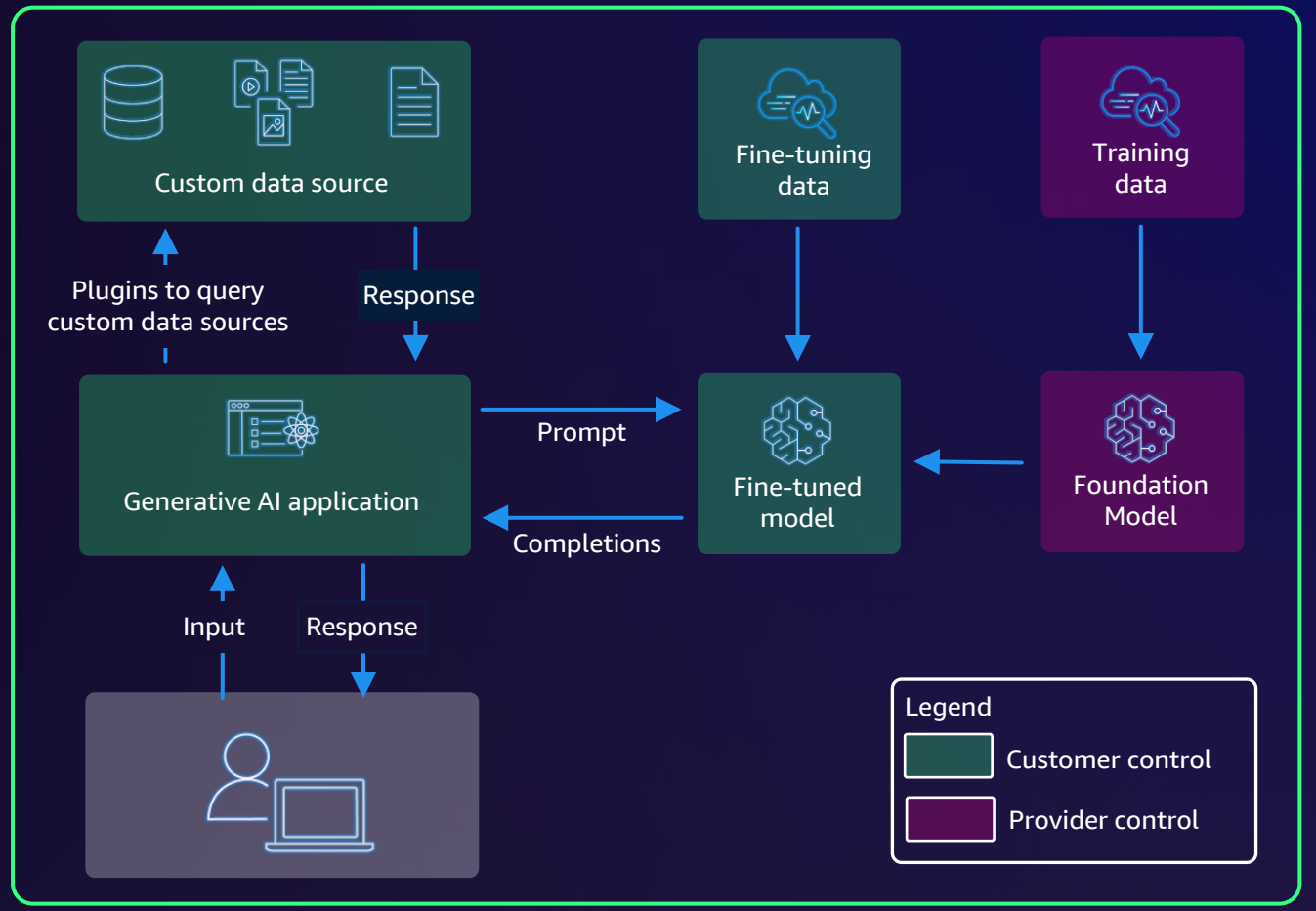
Scope 4: Fine-tuned models

Scope 4: Fine-tuned models

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 include: Amazon 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

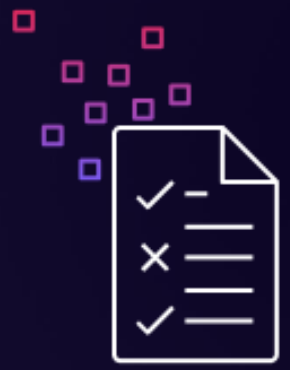
Legal & privacy

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

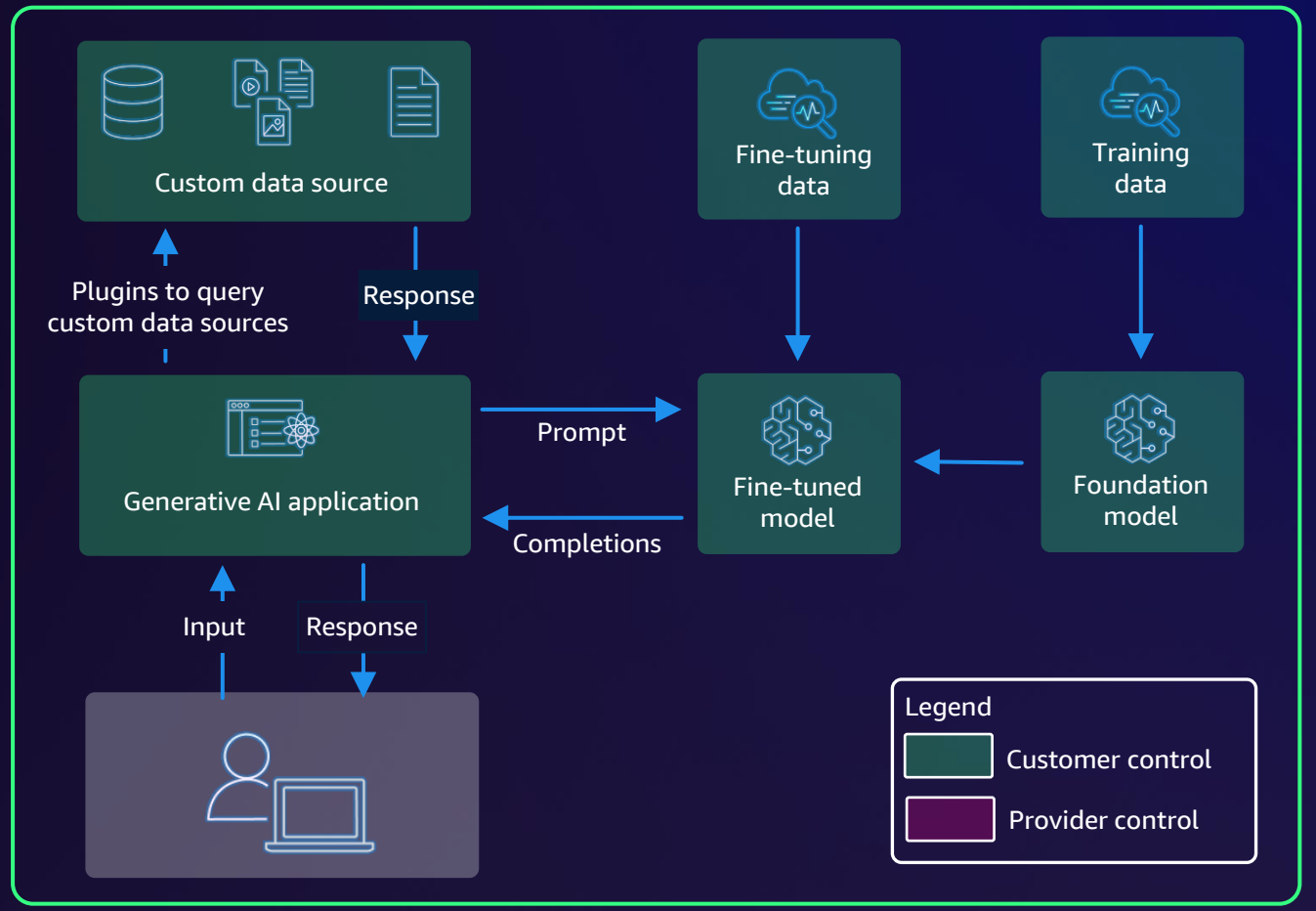
Scope 5: Self-trained models

Scope 5: Self-trained models

SCOPE 5

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

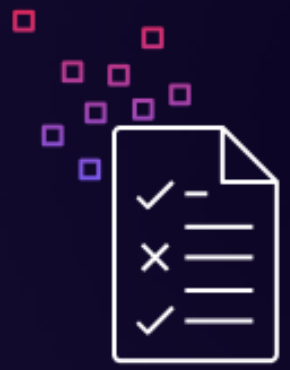
Legal & privacy

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

Key takeaways

1. Does generative AI **require me to change my approach** to security?

   No! But… understand what is unique to generative AI workloads

2. What are the **intersections** of generative AI and security?

   Securing generative AI + Using generative AI to secure + Securing from generative AI threats

3. What **mechanism** can I use to understand what **risks, security, and compliance requirements** impact my use of generative AI?

   Start with the Generative AI Security Scoping Matrix

4. What are some **examples** of how security requirements change depending on scope?

   Provided key examples across 5 scopes and 5 dimensions

Learn more – Generative AI security blog series

Dive deeper via these blogs…

Securing generative AI: An introduction to the Generative AI Security Scoping Matrix

Securing generative AI: data, compliance, and privacy considerations

Securing generative AI: Applying relevant security controls

Designing generative AI workloads for resilience