LOS ANGELES | MAY 22, 2024

aws SUMMIT

SEC301

Building secure generative AI applications on AWS

Raghavarao Sodabathina
He/Him
Principal Solutions Architect
AWS

Brian Soper
He/Him
Senior Solutions Architect
AWS

Agenda

01 What is generative AI?

02 Generative AI design patterns

03 Building generative applications on AWS

04 Generative AI application security controls

05 Chalk time – Generative AI reference architectures

06 Best practices and key takeaways

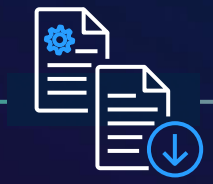
07 How we can help you

What is generative AI?

What is generative AI?

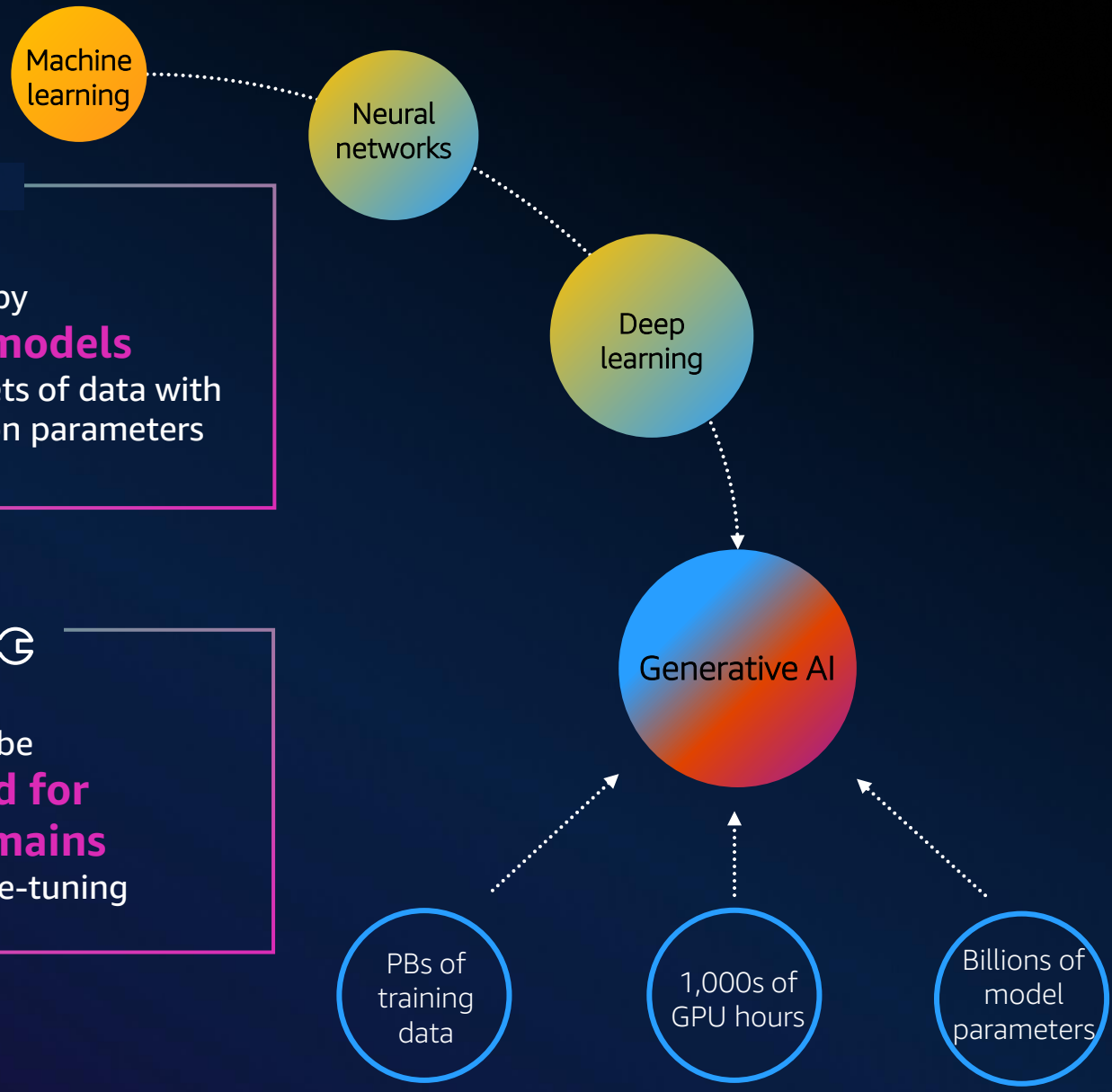
AI that can
generate content
close enough to human-created content for real-world tasks

Powered by
foundation models
pre-trained on large sets of data with several hundred billion parameters



Applicable to
many use cases
like text summarization, question answering, digital art creation, code generation, and so on

Tasks can be
customized for specific domains
with minimal fine-tuning



Source: https://bit.ly/456imEb

Generative AI has the potential to transform all industries

FSI

- Personalized financial advice through conversational assistance
- Life insurance underwriting and pricing through unstructured data synthesis

Retail

- Conversational chatbot for hyper personalization and shopping guidance
- Improved marketing with enhanced customer segmentation and personalized content creation

Manufacturing

- Faster and cheaper part design through generative design
- Automation of manual tasks through text summary and synthesis

Travel & hospitality

- Connected, personalized guest journeys
- Inventory optimization and demand forecasting

Healthcare/life sciences

- Accelerate drug discovery and research
- Synthetic data generation for research

Consumer goods

- Expedite formulation design by quickly testing combinations of components
- Generate personalized marketing content based on (un)structured data from consumer profiles and community insights

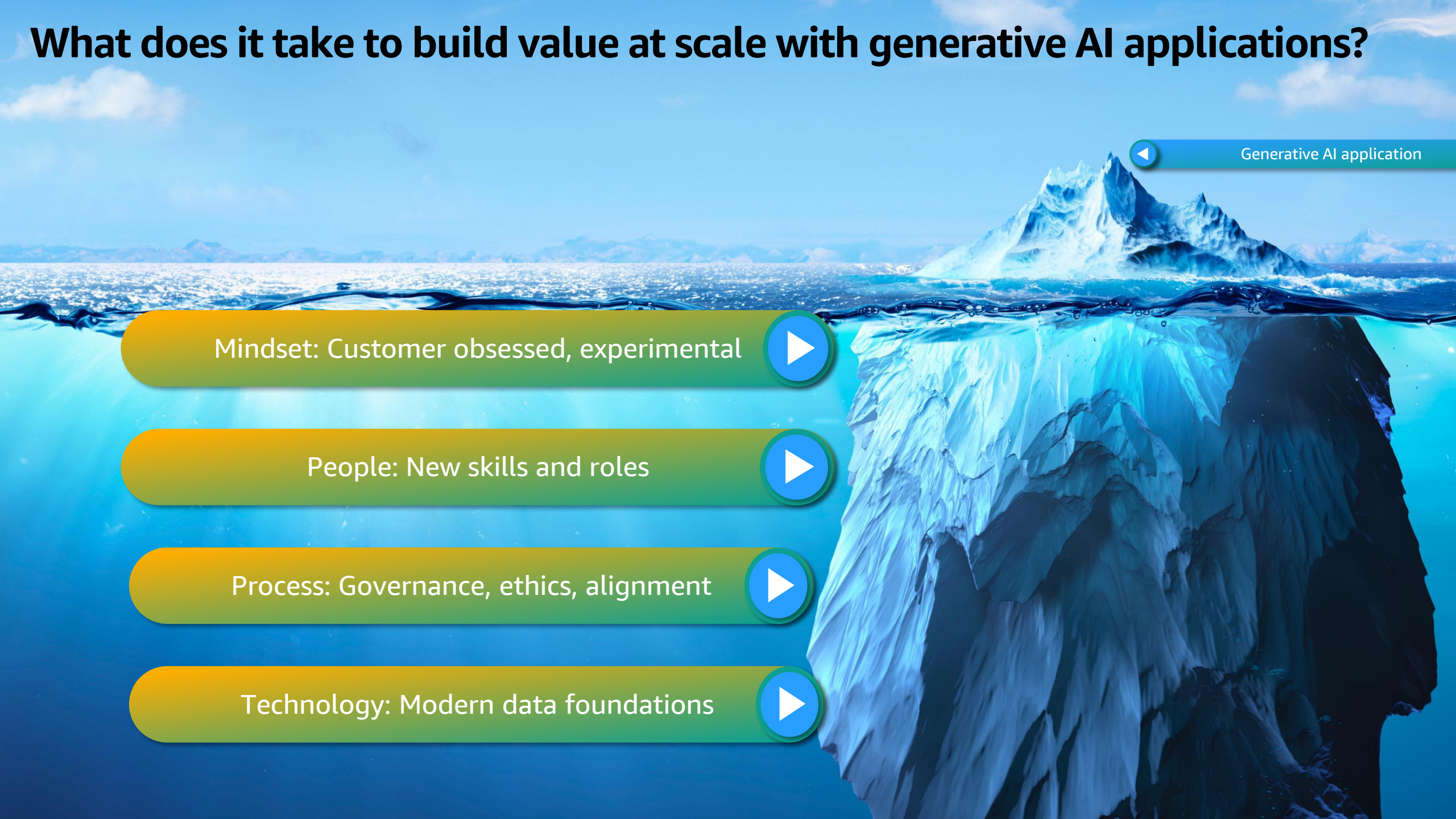
Media & entertainment

- Hyper personalization and dynamic content placement
- Content creation

Gaming

- Quickly adapting existing games with new themes
- Personalized sports bet recommendations

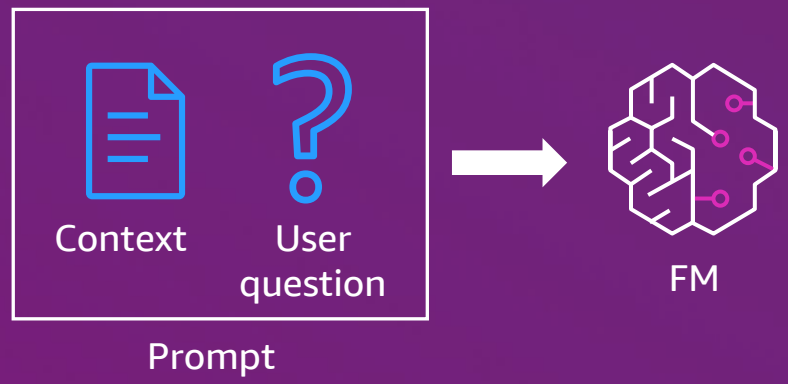
What does it take to build value at scale with generative AI applications?



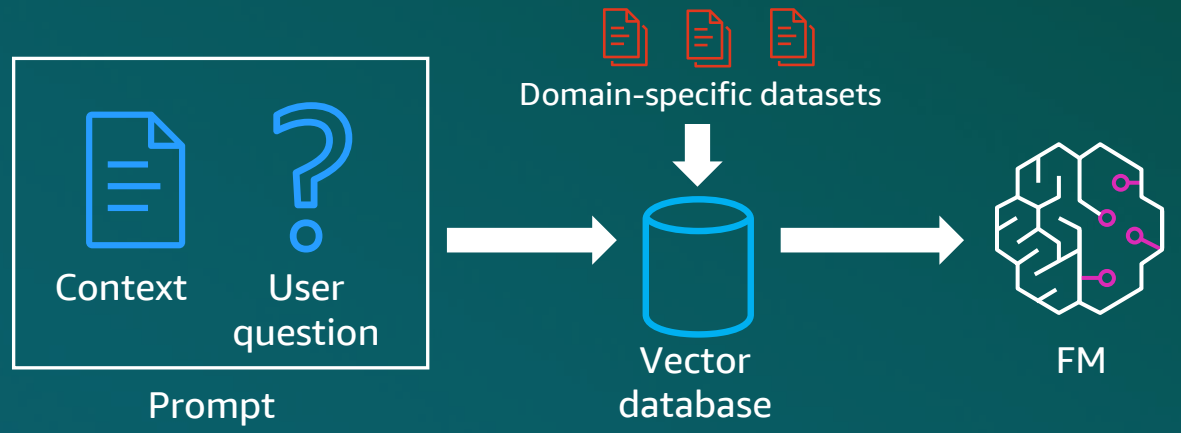
Generative AI design patterns

Emerging generative AI application design patterns

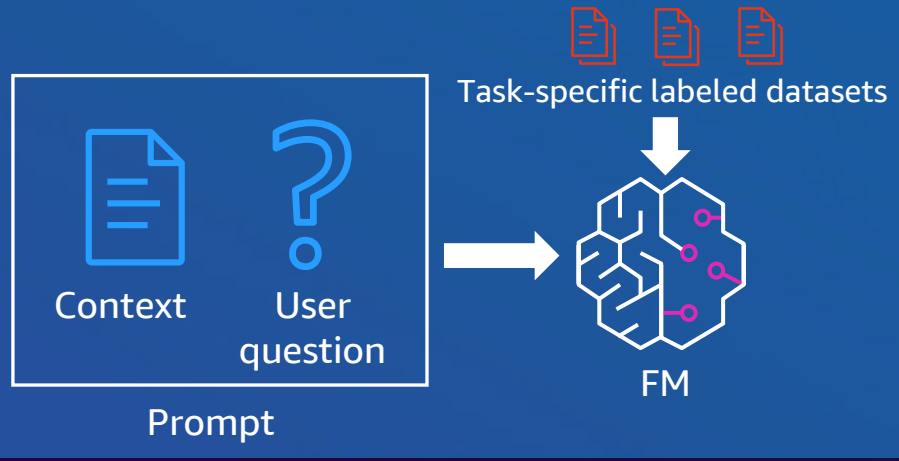
Prompt engineering



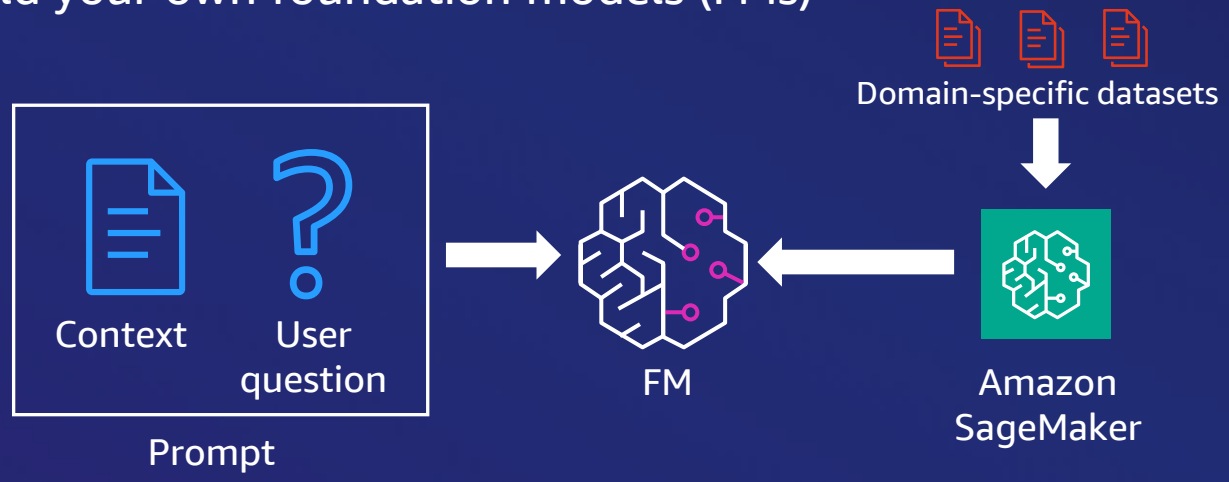
Retrieval Augmented Generation (RAG)



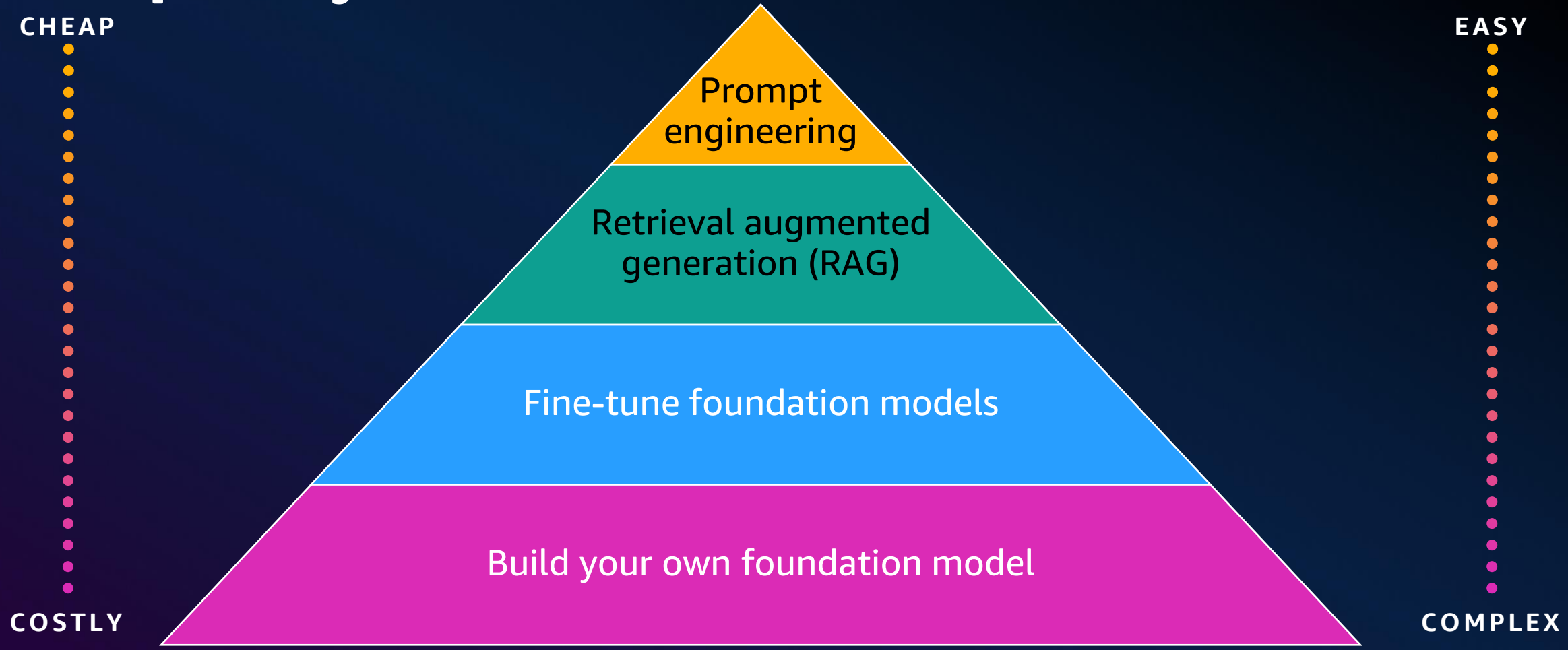
Fine-tune generative AI foundation models



Build your own foundation models (FMs)



Emerging GenAI design patterns, anticipated cost & complexity



Choosing the right generative AI design pattern

IDEAL USAGE PATTERNS FOR VARIOUS GENERATIVE AI DESIGN PATTERNS

	Prompt eng.	RAG	Fine-tune FM	Build your own FM
Cost	Low	Low-Medium	Medium	High
Training duration	Not required	Not required	Minutes to hours	Days to weeks to months
Organization maturity	Basic development capability	Strong development capability	Machine learning, data, and development capability	Strong LLM and data capability with access to large amount of training data
Skills	API integration	Data engineering Embedding tuning Vector DB performance tuning	Experience in training, tuning, and hosting LLM models	LLM model training and operations
AWS services	Amazon Bedrock, Amazon SageMaker JumpStart	Amazon Bedrock, Amazon SageMaker JumpStart	Amazon Bedrock, Amazon SageMaker JumpStart	Amazon SageMaker

Building generative applications on AWS

Amazon Bedrock

Accelerate development of generative AI applications using FMs through an API without managing infrastructure

Enable generative AI apps to complete tasks with agents

Find the right FM for your use case

amazon

ANTHROP\C

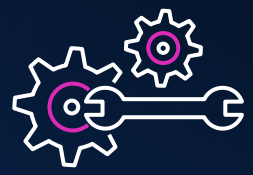
AI21labs

cohere

stability.ai

Meta

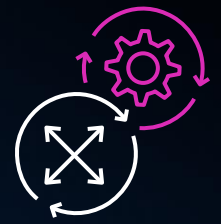
Mistral AI



Privately customize FMs using your organization's data

```
bedrock.invoke_model(
    modelId = model_id,
    contentType = "...",
    accept = "...",
    body = body)
```

Enhance your data protection using comprehensive AWS security capabilities



Responsible AI provided by supported model providers, Amazon Titan supports AI best practices

https://go.aws/48r8iIE

Amazon Bedrock

Broad choice of models

AI21 labs	amazon	ANTHROPIC	cohere	Meta	MISTRAL AI_	stability.ai
Contextual answers, summarization, paraphrasing	Text summarization, generation, Q&A, search, image generation	Summarization, complex reasoning, writing, coding	Text generation, search, classification	Q&A and reading comprehension	Text summarization, Q&A, text classification, text completion, code generation	High-quality images and art
Jurassic-2 Ultra	Amazon Titan Text Lite	Claude 3 Opus	Command	Llama 3 8B	Mistral Large	Stable Diffusion XL1.0
Jurassic-2 Mid	Amazon Titan Text Express	Claude 3 Sonnet	Command Light	Llama 3 70B	Mistral 7B	Stable Diffusion XL 0.8
	Amazon Titan Text Embeddings	Claude 3 Haiku	Embed English	Llama 2 13B	Mixtral 8x7B	
	Amazon Titan Text Embeddings V2	Claude 2.1	Embed Multilingual	Llama 2 70B		
	Amazon Titan Multimodal Embeddings	Claude 2	Command R+			
	Amazon Titan Image Generator	Claude Instant	Command R			

7 providers and 29 models!

aws

Agents for Amazon Bedrock enable generative AI applications to complete tasks in a few quick steps

1

SELECT YOUR FOUNDATION MODEL

2

PROVIDE BASIC INSTRUCTIONS

3

SELECT RELEVANT DATA SOURCES

4

SPECIFY AVAILABLE ACTIONS

- Breaks down and orchestrates tasks
- Securely accesses and retrieves company data
- Takes action by invoking API calls on your behalf
- Provides fully managed infrastructure

New additions to Amazon Bedrock

Providing extensive capabilities for building generative AI apps

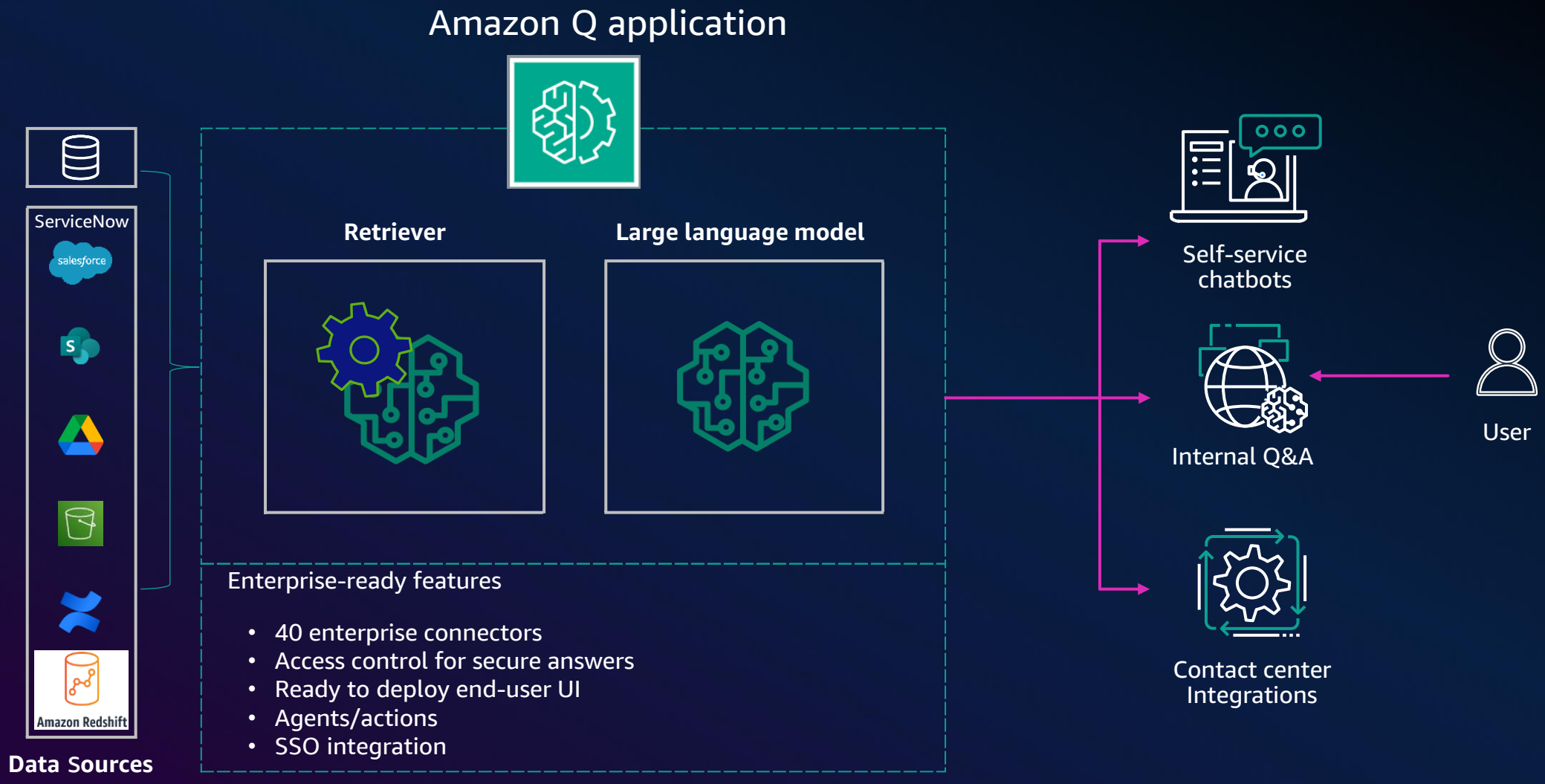
- Model evaluation
- Custom model import
- Agents – Claude 3 Sonnet and Haiku support
- Agents – Quick create
- Zero-setup RAG
- Multi-data source support for knowledge bases

Building highly accurate generative AI applications

ENABLING ENTERPRISES TO BUILD HUMAN-LIKE CONVERSATIONAL EXPERIENCES

Challenges adopting generative AI	Amazon Q
Hallucinations & traceability	Restricted to trusted enterprise data and references to original content sources
Multiple data silos	Connector service to aggregate your data, allowing users to generate answers from multiple documents
Enterprise security	Results incorporate user access permissions, including ACL support
Data relevance	Responses generated from most recent data in your knowledge base
Time to value	Fully managed service, low-code interface for quick deployment

Amazon Q high-level architecture



Why use foundation models on SageMaker JumpStart

1

Choose foundation models offered by model providers

Meta AI

AI21 labs

Lighton
We bring Light to AI

stability.ai

co:here

alexa

2

Deploy model

Deploy the model for inference using SageMaker hosting options includes single node

3

Fine tune model and automate ML workflow

Only selected models can be fine-tuned

Automate ML workflow

Data stays in your account including model, instances, logs, model inputs, model outputs

Fully integrated with Amazon SageMaker features

Build your own FM at scale using Amazon SageMaker

Managed infrastructure
Full control of your model training with managed and price-performant infrastructure

Efficient distributed training
Complete distributed training up to 40% faster

Debugging and experimentation tools
Capture metrics and profile training jobs in real time to quickly correct performance issues. Track ML model iterations easily.

Price-performant inference
Deploy models in production for any use case while optimizing cost

Repeatable and reproducible MLOps
Automate and standardize processes across the ML lifecycle

Governance
Purpose-built governance tools to help you implement ML responsibly

Human-in-the-loop support
Create high-quality datasets and align model outputs with human preferences

Key factors in decision-making

Cost

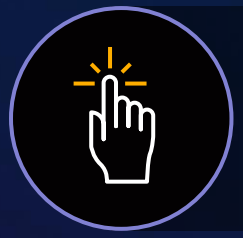
Optimize for cost with a variety of models for your needs with AWS pay-as-you go pricing

Accuracy

Use highly accurate models per HELM benchmarks, and Model Evaluation on Amazon Bedrock (preview)

Speed (latency)

Optimize for performance with different model and types



Ease of use

Choose the right generative AI design pattern based on your use case

Data security

Establish private connectivity between your virtual private clouds (VPCs) and Amazon Bedrock and Amazon SageMaker

Hot themes for generative AI applications

Develop and use AI that is fair, transparent, accountable, and aligned with human values while considering the potential impact on individuals, society, and the environment

Mitigate and address the unintended discrimination or unfairness that can arise from biased data, algorithms, or system design, in order to ensure equitable and unbiased outcomes

Safeguard sensitive information, protect against unauthorized access or misuse, and ensure the responsible handling and protection of intellectual property rights in generated content

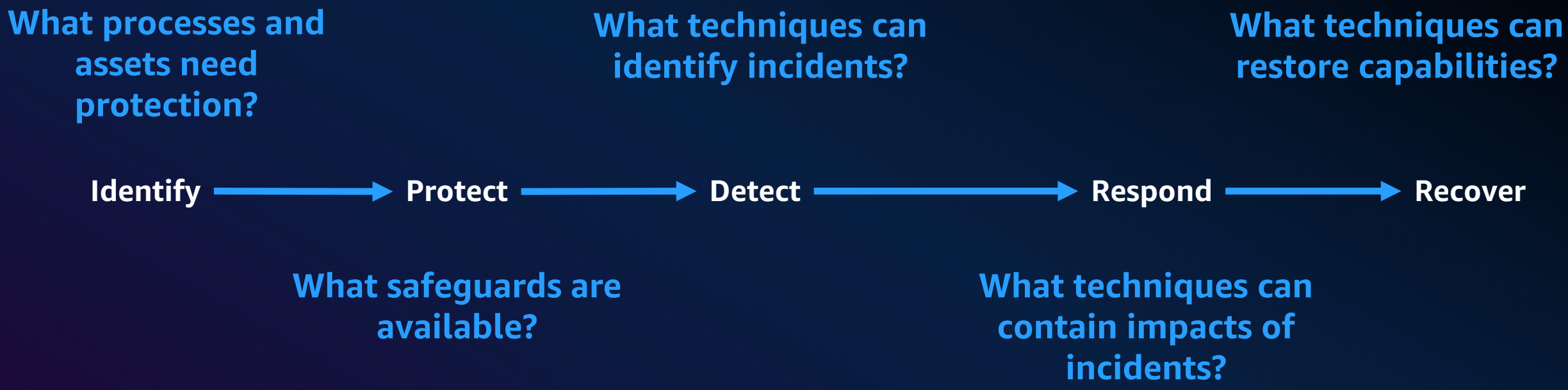
Prevent the generation of misleading or unrealistic outputs that could potentially deceive users or lead to the dissemination of false information

Ensure that the goals, values, and intentions of AI systems align with those of human users, mitigating the risks of unintended or harmful behavior

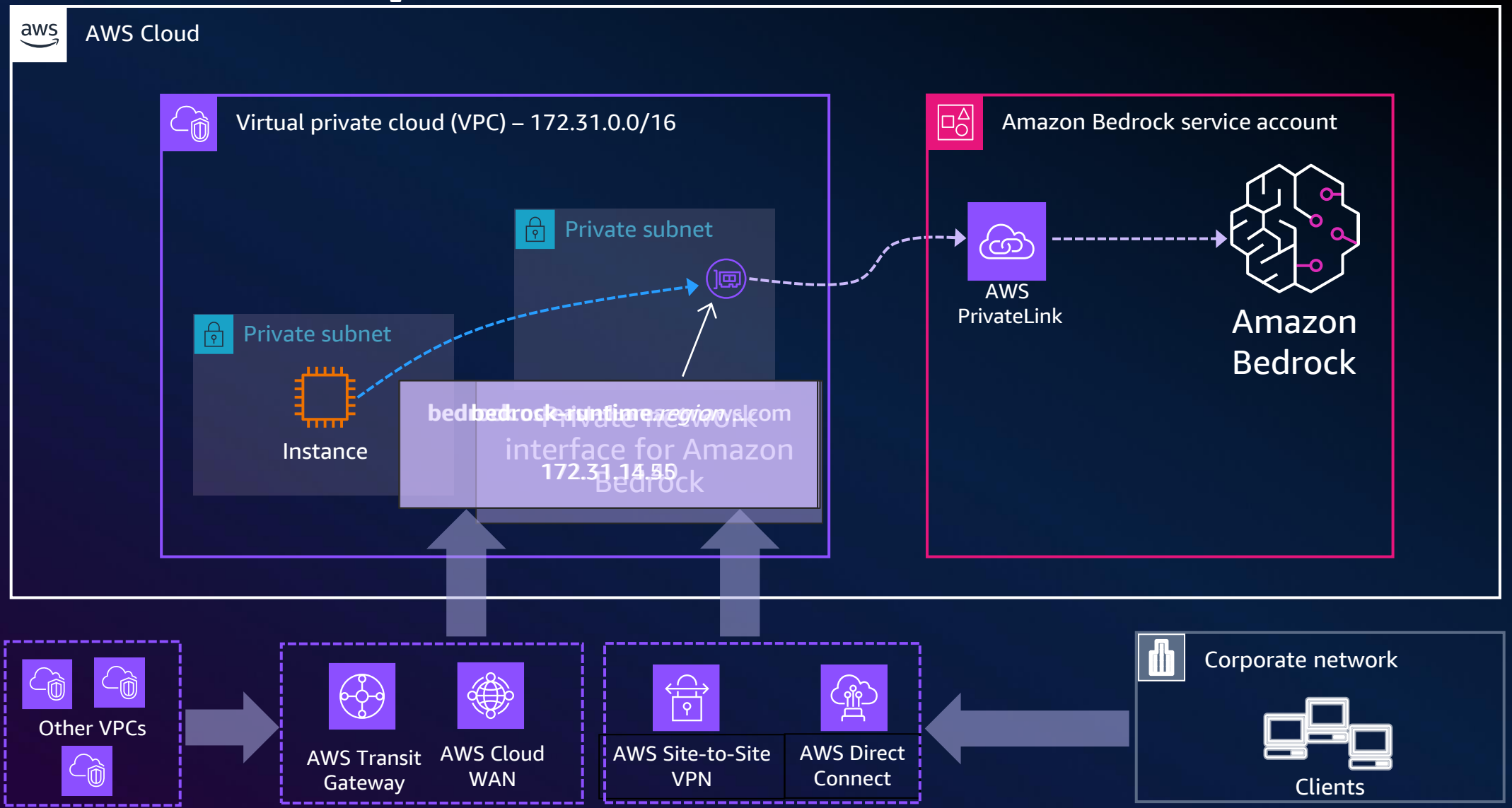
Keep up with the rapid pace of advancements and research in the field, ensuring that AI models and systems remain up to date and effective in a constantly evolving landscape

Generative AI application security controls

NIST Cybersecurity Framework



Infrastructure protection with Amazon Bedrock



Authentication and authorization with Amazon Bedrock

IAM

- Identity-based policies
- Actions
- Resources
- SCP for model invocation access controls
- Tags (ABAC)

IAM policy with actions & resources

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "BedrockConsole",
            "Effect": "Allow",
            "Action": [

"bedrock:ListFoundationModels",
                "bedrock:InvokeModel",

"bedrock:AcceptUserAcknowledgement,

"bedrock:GetUserFeedback",

"bedrock:SendUserFeedback",
                "bedrock:GetPrompt",
                "bedrock:UpdatePrompt",
                "bedrock:ListPrompts",
                "bedrock:DeletePrompt",

"bedrock:GetModelPermission"
],
            "Resource": "*"
        }
]
```

IAM fine-grained controls

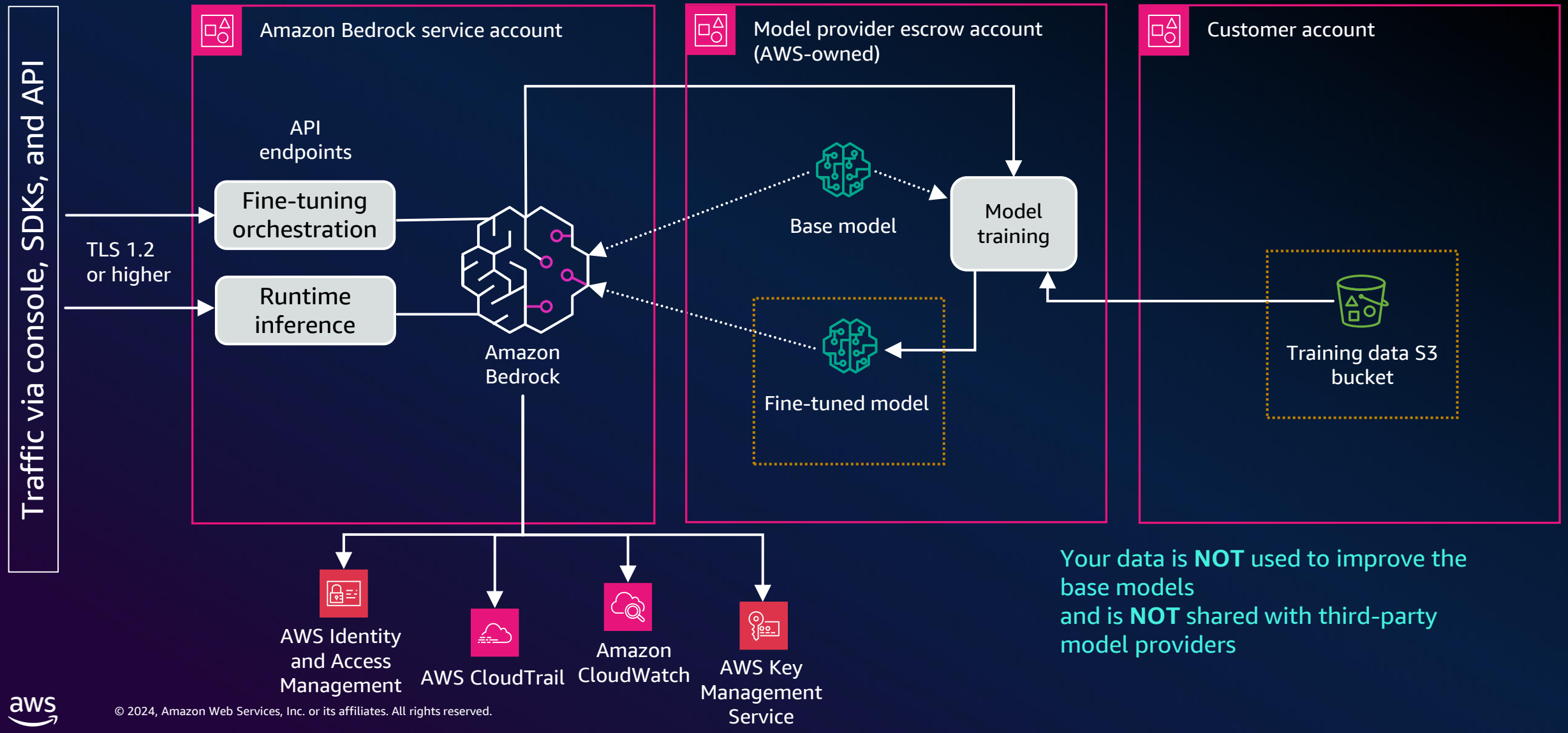
```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "BedrockConsole",
            "Effect": "Allow",
            "Action": [
        "bedrock:ListFoundationModels",
        "bedrock:InvokeModel",
        "bedrock:GetModelPermission"

],
            "Resource":
"arn:aws:bedrock:*::foundation-model/model-a"
        }
]
```

Service control policy

```
{
    "Version": "2012-10-17",
    "Statement":
    {
        "Sid": "DenyInferenceForModelX",
        "Effect": "Deny",
        "Action": "bedrock:InvokeModel",
        "Resource":
 "arn:aws:bedrock:*:*:foundation-model/model-a"
    }
}
```

Data protection with Amazon Bedrock

YOUR DATA IS ALWAYS WITHIN YOUR CONTROL



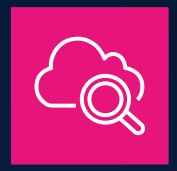
Your data is **NOT** used to improve the base models
and is **NOT** shared with third-party model providers

Amazon Bedrock – Key security controls

Amazon Bedrock

AWS CloudTrail

Amazon Bedrock will write API actions to AWS CloudTrail



Amazon CloudWatch

CloudWatch metrics supported

"AWS/Amazon Bedrock" namespace, and each metric is per model ("ModelId" dimension)

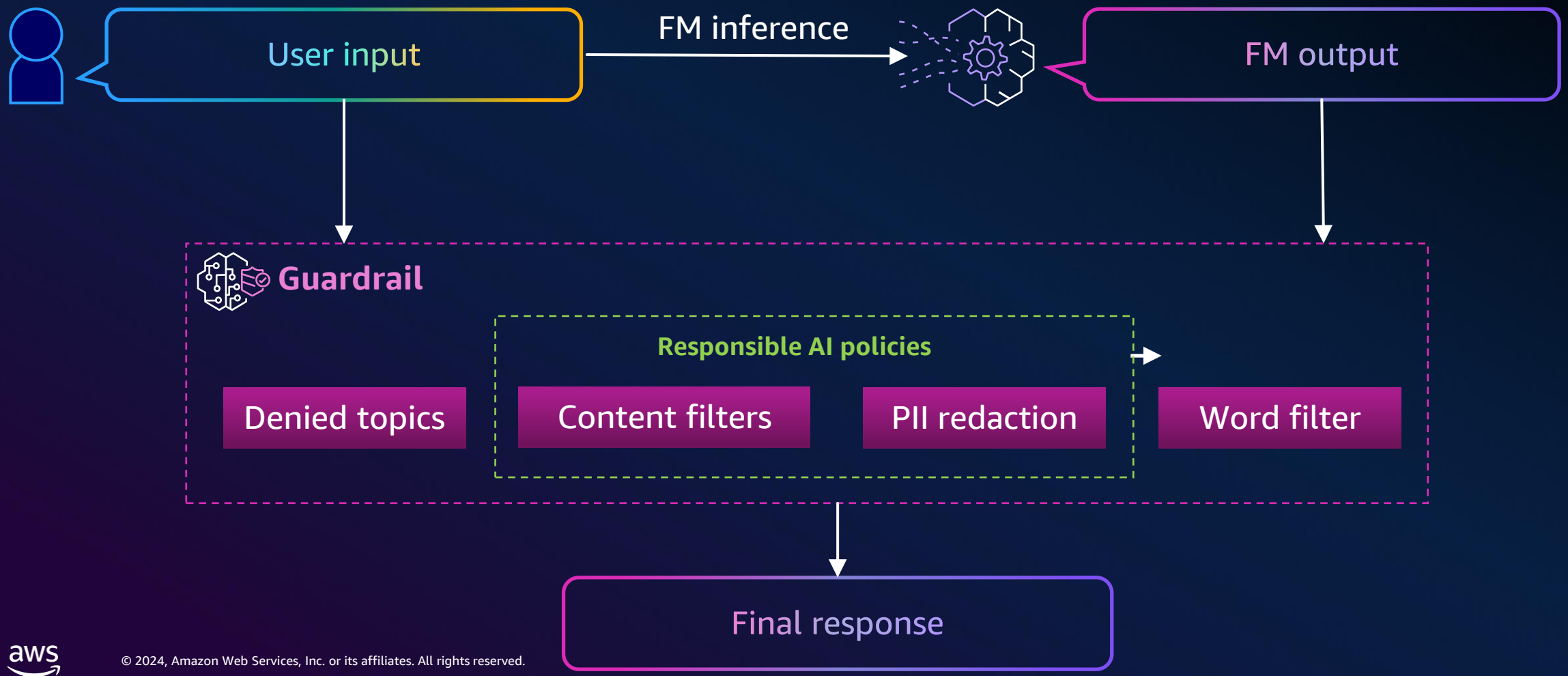
AWS Audit Manager

Audit Manager controls

"generative AI best practices framework v1"

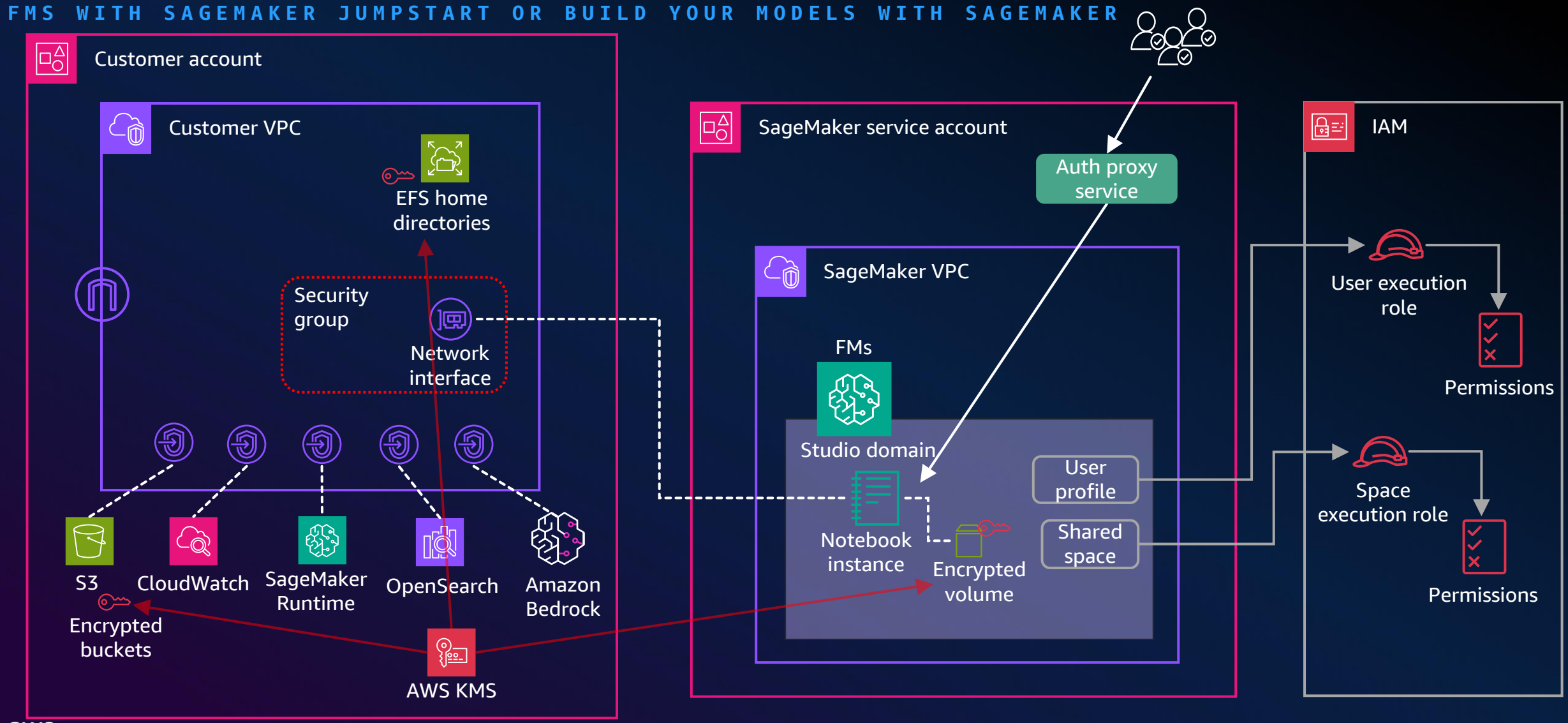
Guardrails for Amazon Bedrock

IMPLEMENT SAFEGUARDS TAILORED TO YOUR APPLICATION REQUIREMENTS AND RESPONSIBLE AI POLICIES



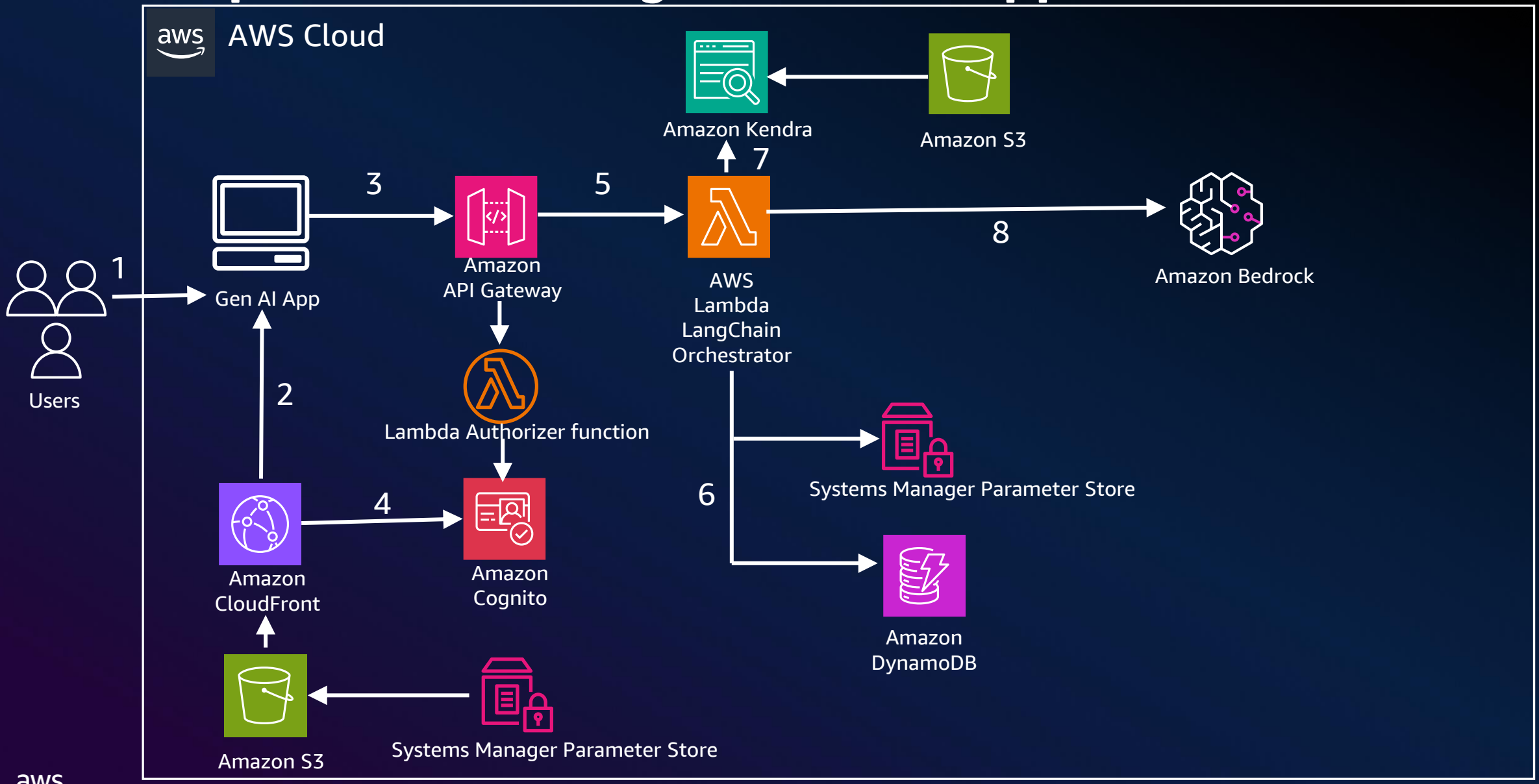
Amazon SageMaker security controls

FMS WITH SAGEMAKER JUMPSTART OR BUILD YOUR MODELS WITH SAGEMAKER



Chalk time – An example RAG generative AI applications architecture

An example chatbot RAG generative AI application architecture



https://go.aws/3sU4YWs

Key takeaways and best practices

Key takeaways and best practices

1. Your data is not used for Amazon Bedrock service improvement and not shared with third-party model providers
2. You can integrate with AWS Identity and Access Management Service (IAM) to manage inference access, allow/deny access for specific models, and enable AWS Management Console access
3. Fine-tuned (customized) models are encrypted and stored using the customer's AWS KMS key. Only you have access to your customized models.
4. Use AWS CloudTrail to monitor API activity and troubleshoot issues as you integrate with generative AI applications: https://docs.aws.amazon.com/awscloudtrail/latest/userguide/cloudtrail-user-guide.html
5. You must first establish a business value and then identify your users in order to apply generative AI to produce the desired business outcome
6. Your next step in developing generative AI applications should be data discovery: https://docs.aws.amazon.com/wellarchitected/latest/analytics-lens/data-discovery.html
   In order for fundamental models to generate the required results, domain-specific data is essential.
7. Use AWS security services to form your defense-in-depth security strategy
8. Use Amazon Bedrock Model Evaluation to compare, and select the best foundation model for your use case
9. Implement safeguards customized to your application requirements and responsible AI policies with Guardrails for Amazon Bedrock

Learn more

Generative AI on AWS
http://go.aws/48nPoSV

Amazon Bedrock resources
https://go.aws/46qvLaB

Amazon Bedrock Workshop
https://bit.ly/3FqL1sY

The role of vector datastores in generative AI applications
http://go.aws/3t77M2o

Amazon Bedrock new capabilities
https://bit.ly/3K2Vpd3

AWS Generative AI application builder
https://go.aws/3sU4YWs

An introduction to the Generative AI Security Scoping Matrix
https://go.aws/49UjIVW

OWASP Top 10 for LLMs
https://bit.ly/4doZCFo

Generative AI security Controls
https://bit.ly/3JH0M11

4 biggest questions about generative AI security
https://bit.ly/4dnUdyg

Build a Secure Enterprise Machine Learning Platform on AWS
https://go.aws/43zz9PA

aws

How we can help you ?

How AWS enables your data and generative AI journey?

Envision

Art of The Possible/Think Big

Learning from Amazon

Digital Innovation

EBCs

Enable

Training/enablement
ML University

Immersion Days
Data Driven Everything (D2E)
Use Case Discovery

ML Competency Center

Execute

ML Solutions Lab prototyping

AWS ProServe/partners

skillbuilder.aws

Build beyond

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

Thank you!

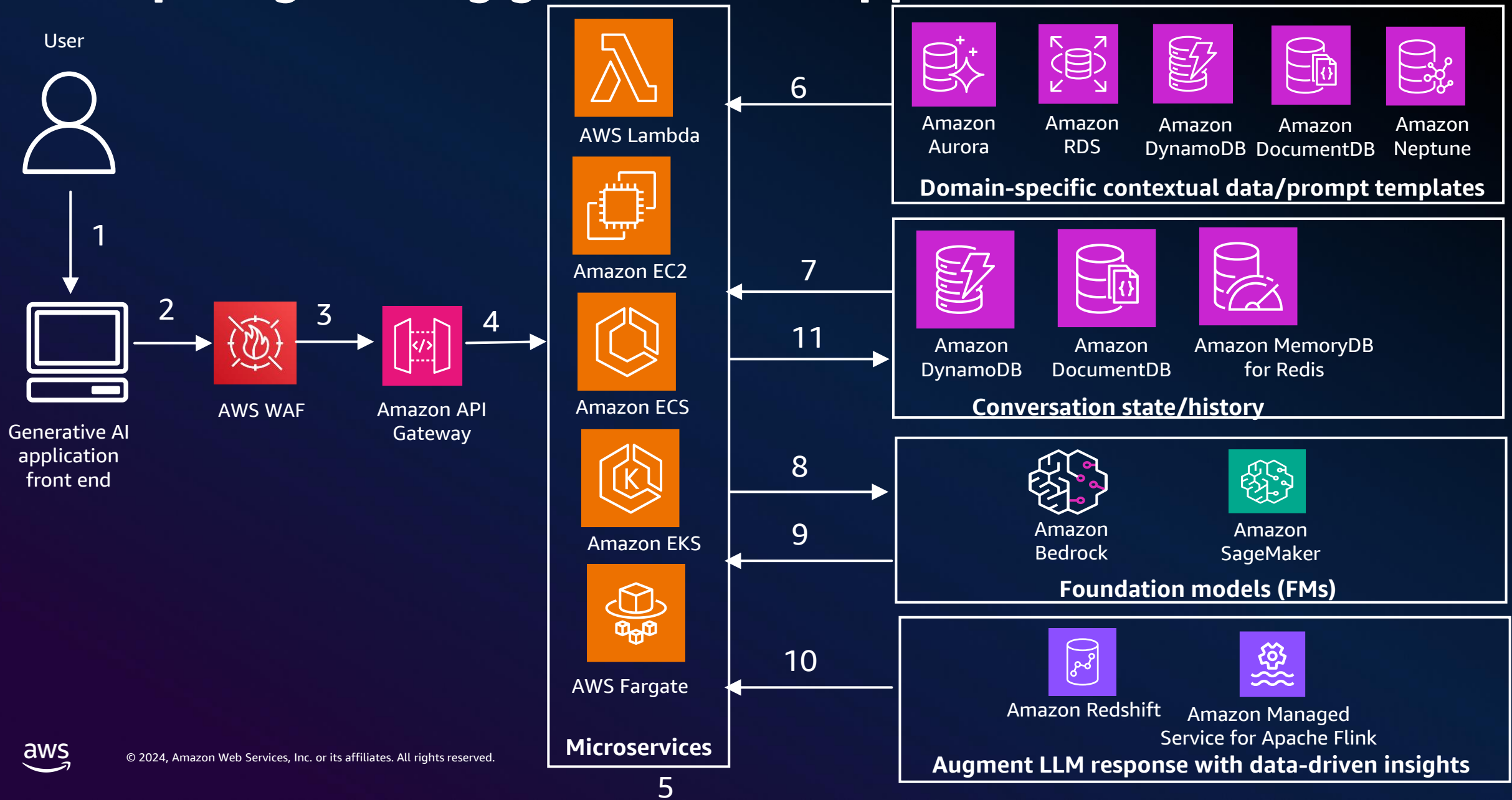
Please complete the session survey in the mobile app

Raghavarao Sodabathina
sodabar@amazon.com
https://www.linkedin.com/in/raghavaraos/

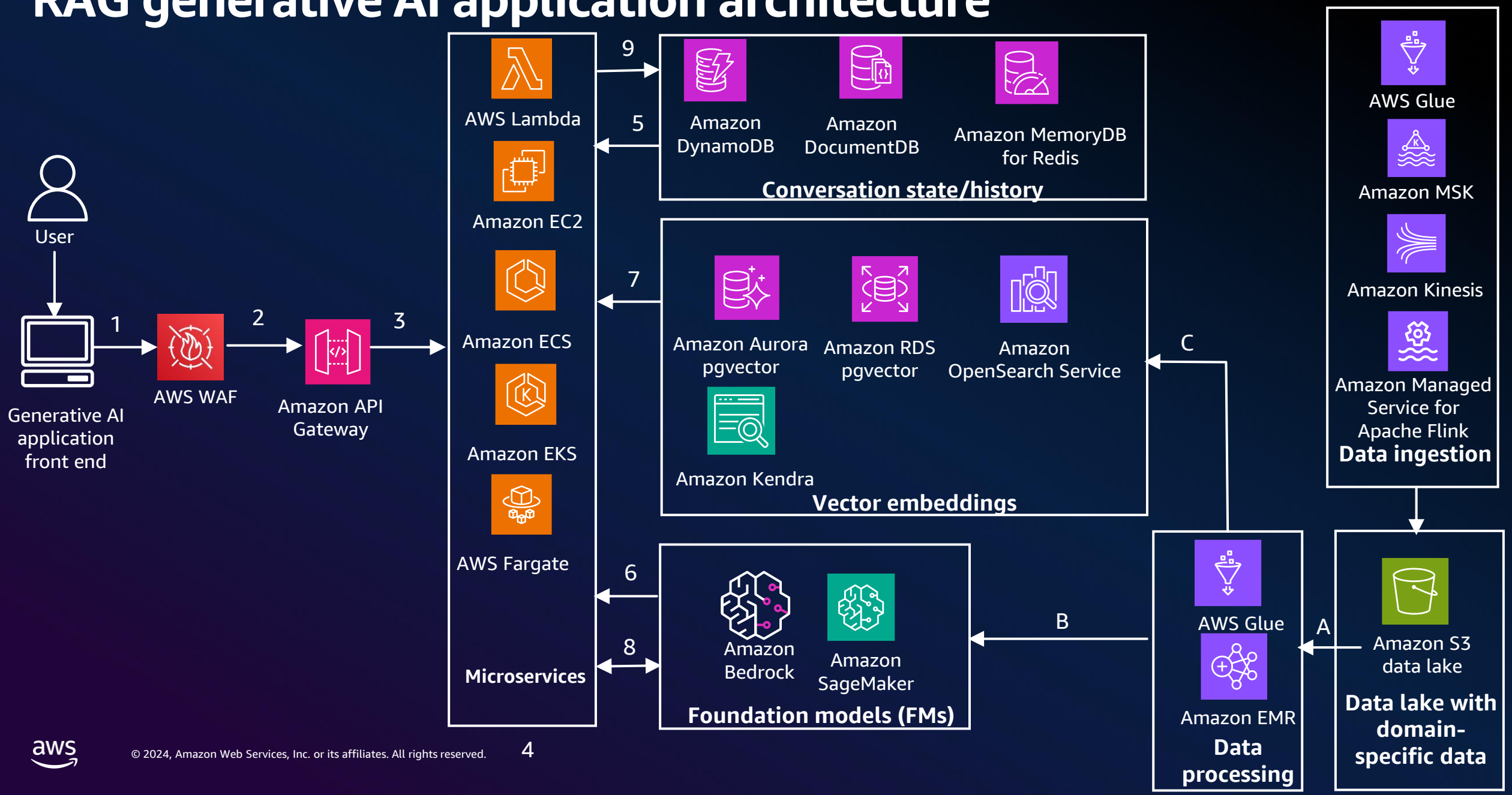
Brian Soper
sopeb@amazon.com
https://www.linkedin.com/in/briansoper/

Appendix

Prompt engineering generative AI application architecture



RAG generative AI application architecture



Fine-tuning generative AI application architecture