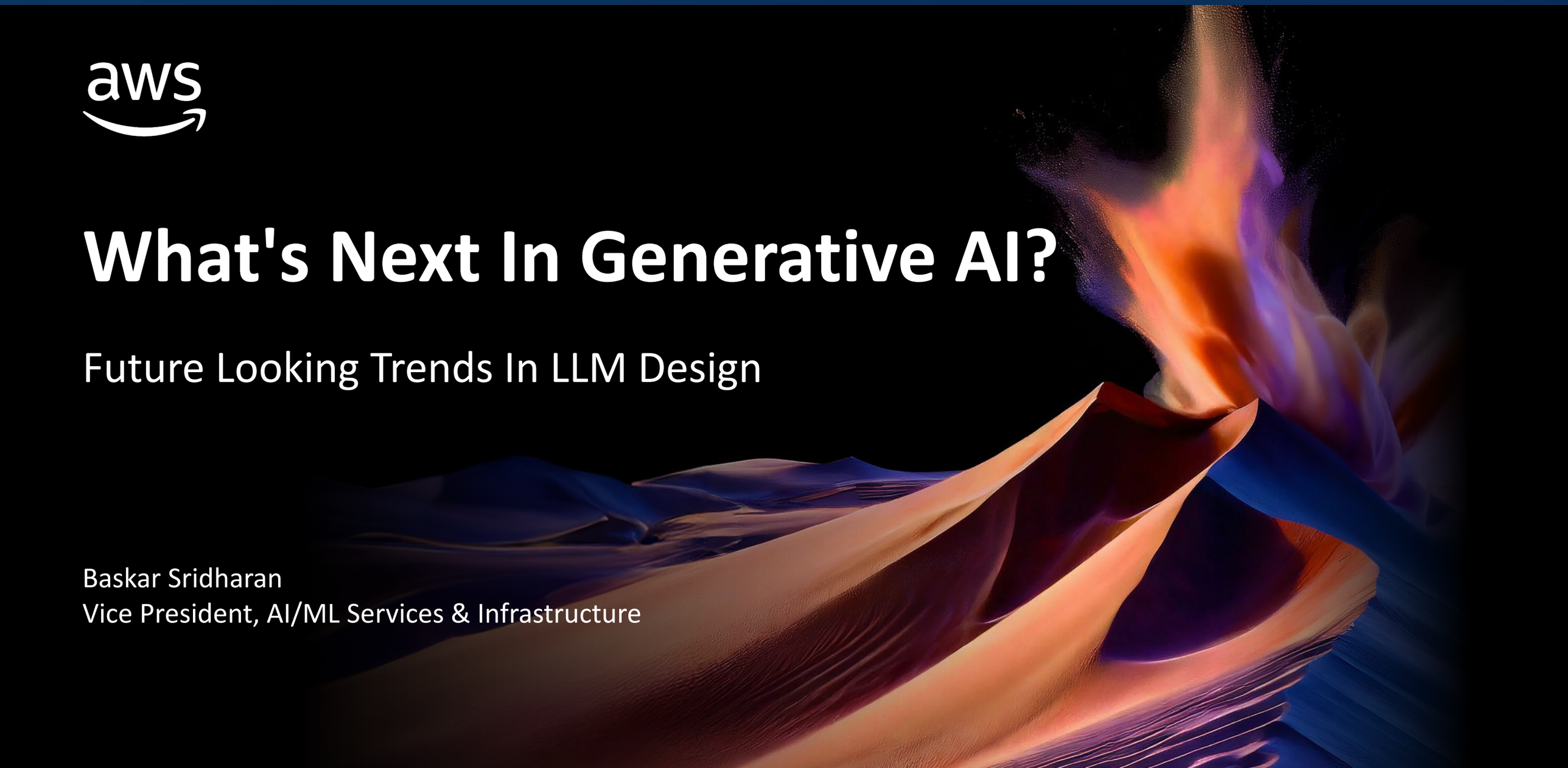
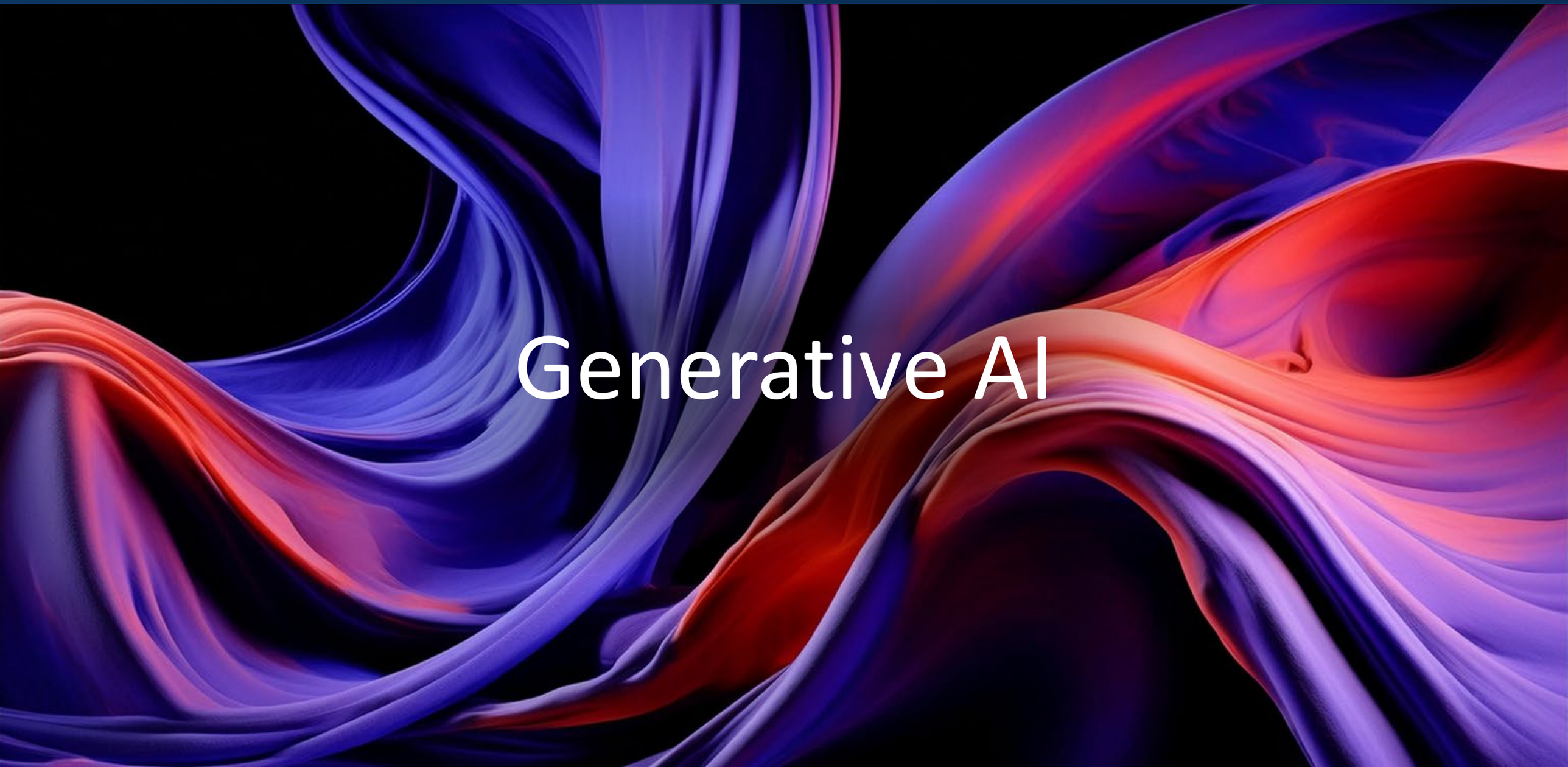
aws

What's Next In Generative AI?

Future Looking Trends In LLM Design



Baskar Sridharan
Vice President, AI/ML Services & Infrastructure


Generative AI

2023

The year of POCs

What is generative AI?

Is this secure?

Do I need to become a prompt engineer?

How do I choose a model?

Where do I get started?

What does this mean for my business?

What is a foundation model?

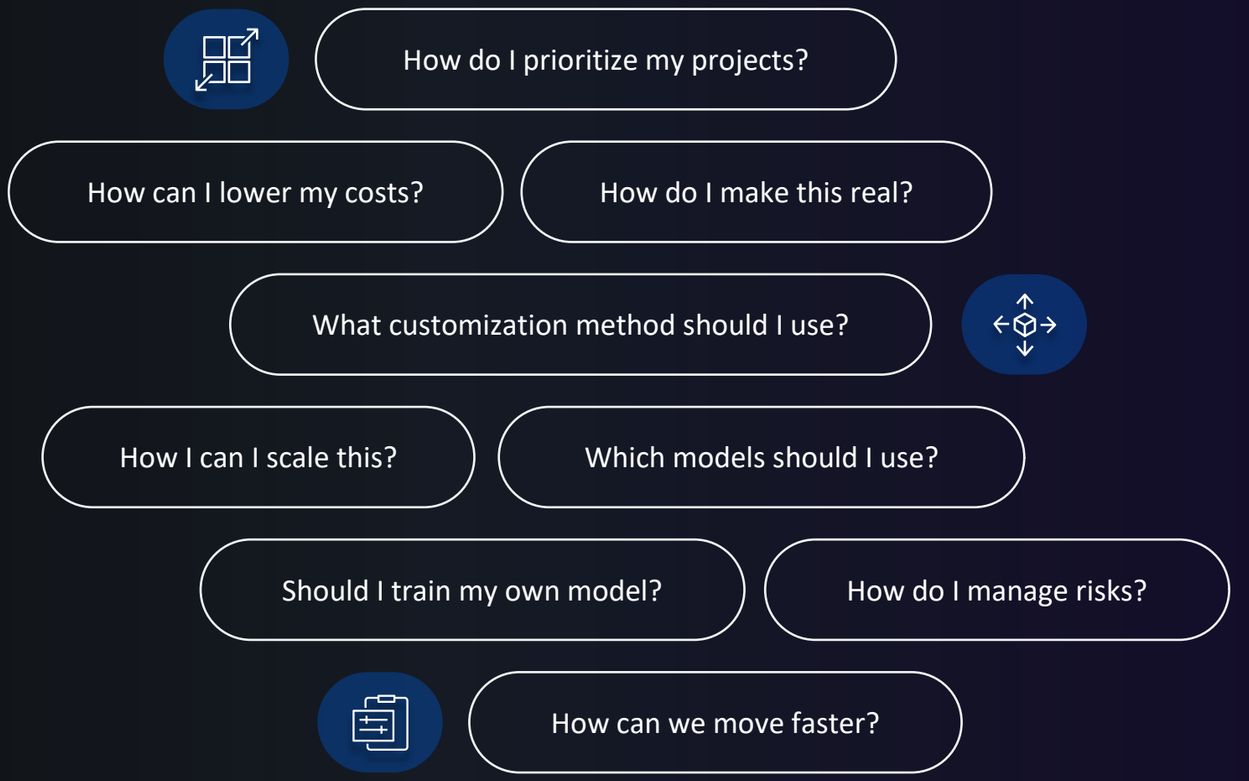
Which models should we try out?

What is FM?

What is a large language model?

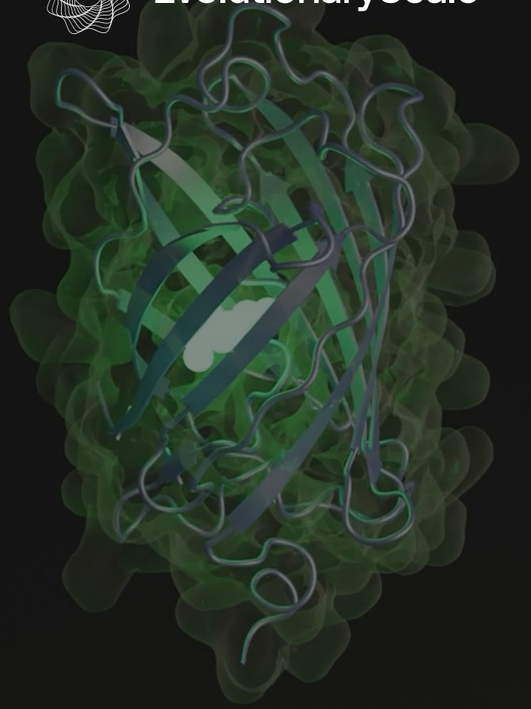
2024

The year of production

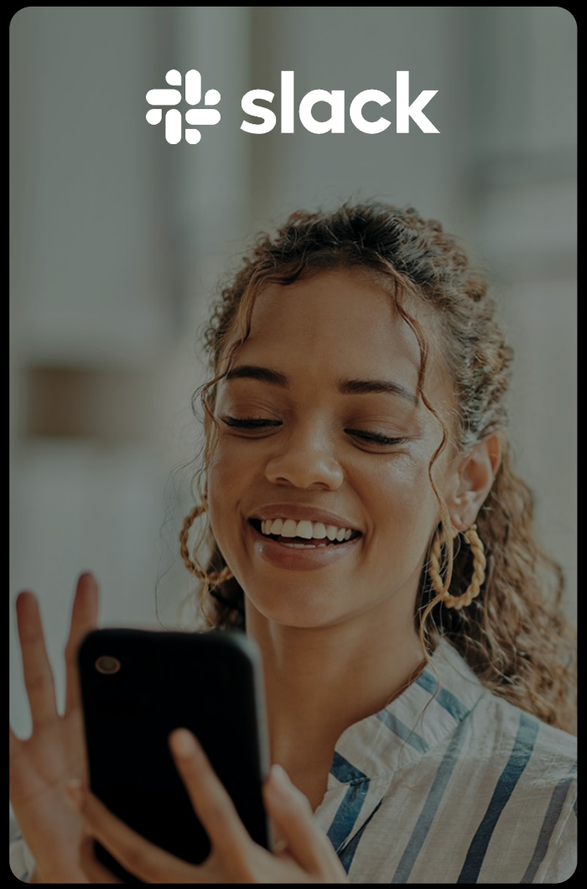



NFL
NEXT GEN STATS
powered by aws
5 years
Historical NFL data for training
1M+
Data points for every NFL game
20+
Features for each defender
10
New NFL stats for the 2025 season

EvolutionaryScale



slack



Thomson Reuters

96%

of all AI/ML unicorns
run on AWS

90%

of 2024 Forbes AI 50
run on AWS

Generative AI stack

APPLICATIONS THAT LEVERAGE LLMs & FMs

Amazon Q

AWS App Studio

TOOLS TO BUILD WITH LLMs & OTHER FMs

Amazon Bedrock

Guardrails | Agents | Studio | Customization | Custom Model Import | Amazon Models

INFRASTRUCTURE FOR FM TRAINING & INFERENCE

GPUs

Trainium

Inferentia

SageMaker

EFA

UltraClusters

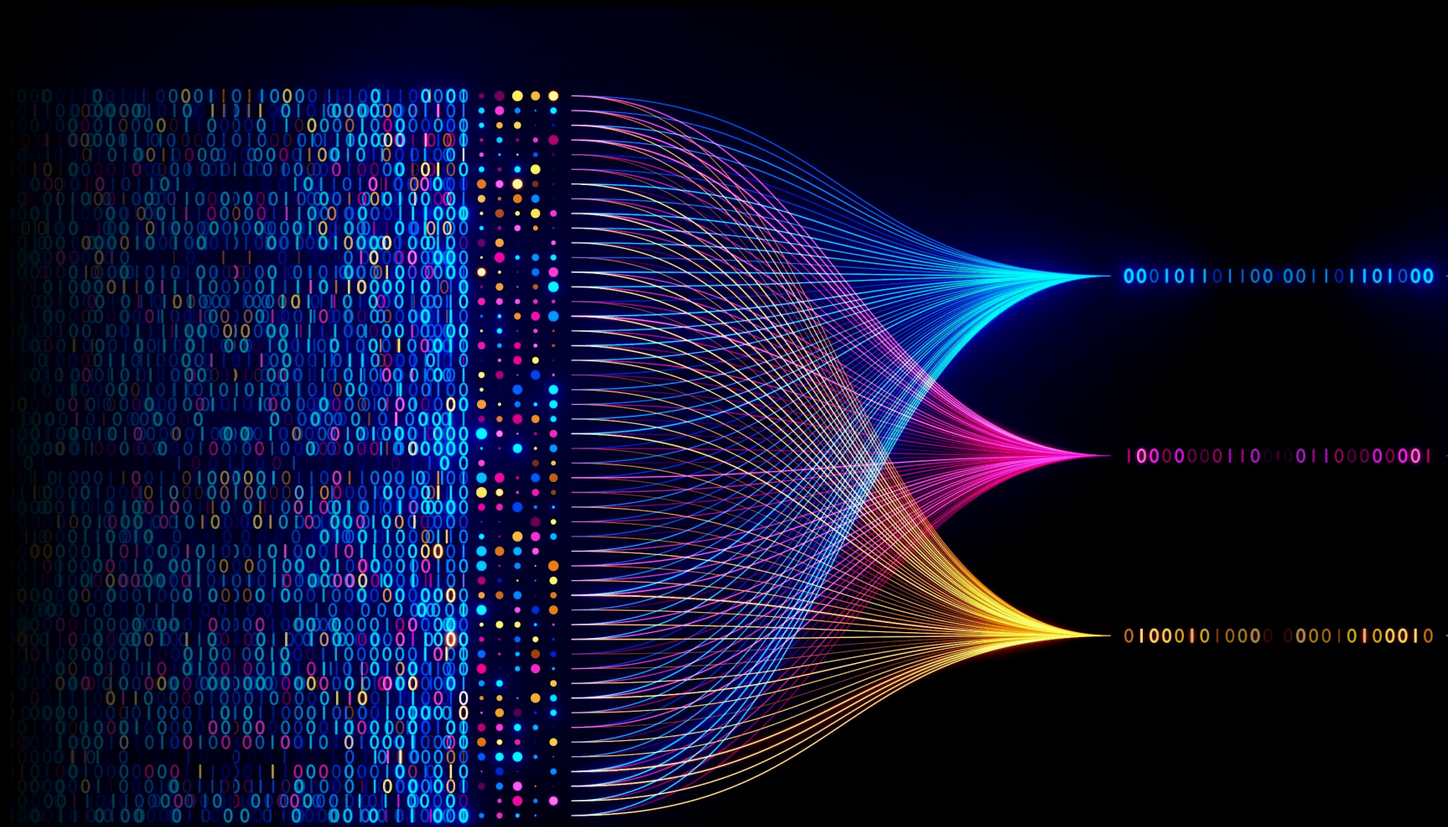
EC2 Capacity Blocks

Nitro

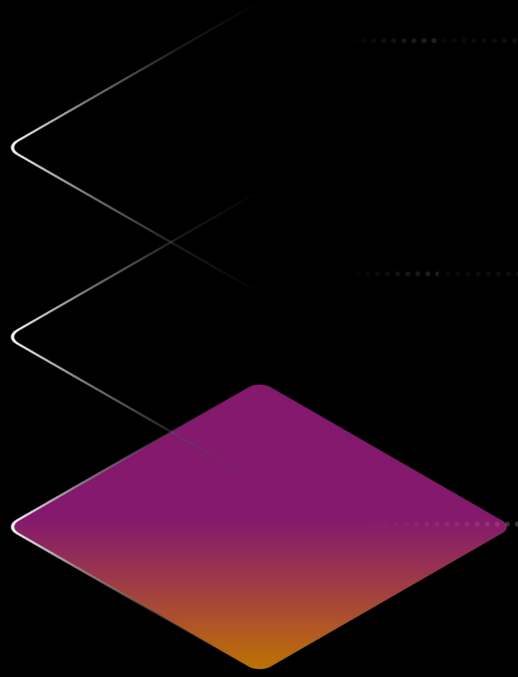
Neuron

Generative AI
powered by FMs

1. Pretrained on vast amounts of unstructured data
2. Contain large number of parameters that make them capable of learning complex concepts
3. Can be applied in a wide range of contexts
4. Customize FMs using your data for domain specific tasks



Generative AI stack



APPLICATIONS THAT LEVERAGE LLMs & FMs

Amazon Q

AWS App Studio

TOOLS TO BUILD WITH LLMs & OTHER FMs

Amazon Bedrock

Guardrails | Agents | Studio | Customization | Custom Model Import | Amazon Models

INFRASTRUCTURE FOR FM TRAINING & INFERENCE

GPUs

Trainium

Inferentia

SageMaker

EFA

UltraClusters

EC2 Capacity Blocks

Nitro

Neuron

Large scale FM training challenges

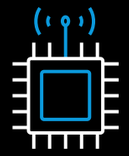
Clusters provision & management

Infrastructure stability

Strategies to optimize training performance

Generative AI inference challenges

Performance lag with user experience impact



Expensive GPUs and accelerators

Complex optimizations require months of time

NEW

Amazon SageMaker

Elastic Kubernetes Service (EKS) support for HyperPod

Remove the heavy lifting to scale across thousands of AI accelerators

A fully resilient infrastructure purpose-built for foundation model development

Optimize utilization of cluster's compute, memory, and network resources between training and inference workloads

NEW

Amazon SageMaker

Inference Optimization Toolkit

Speculative decoding, compilation, and quantization

Fully managed across Studio, Jumpstart, SDKs, and AWS command-line interface

Generative AI stack

APPLICATIONS THAT LEVERAGE LLMs & FMs

Amazon Q

AWS App Studio

TOOLS TO BUILD WITH LLMs & OTHER FMs

Amazon Bedrock

Guardrails | Agents | Studio | Customization | Custom Model Import | Amazon Models

INFRASTRUCTURE FOR FM TRAINING & INFERENCE

GPUs

Trainium

Inferentia

SageMaker

EFA

UltraClusters

EC2 Capacity Blocks

Nitro

Neuron

Amazon Bedrock

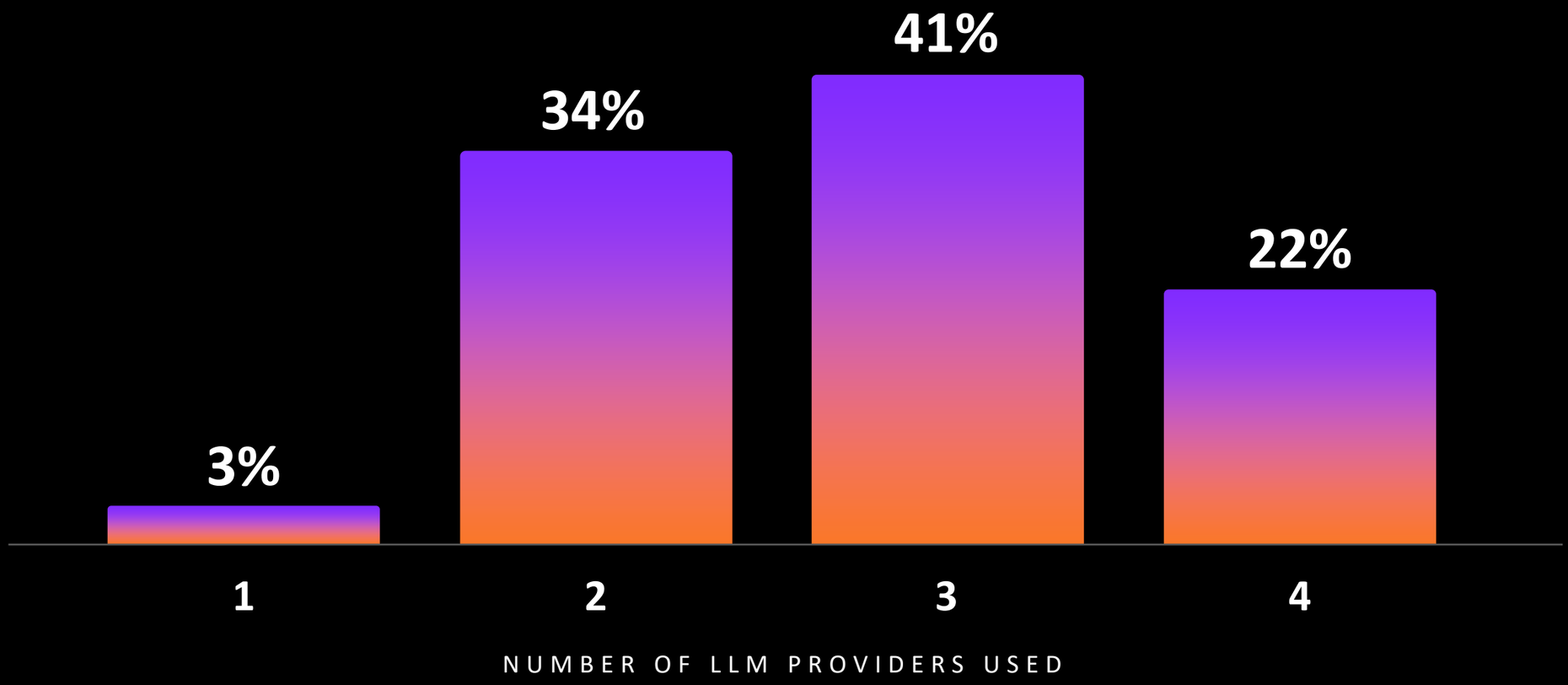
Broadest selection of models

CUSTOM MODEL IMPORT
Leverage your customized models on Amazon Bedrock

AI21labs	amazon	ANTHROPIC	cohere	Meta	MISTRAL AI_	stability.ai
JURASSIC-2 JAMBA-INSTRUCT	TITAN: TEXT,LITE,EXPRESS TEXT PREMIERE IMAGE GENERATOR EMBEDDINGS V2 MULTIMODAL EMBEDDINGS	CLAUDE 3.5 SONNET CLAUDE 3 HAIKU, SONNET, OPUS	COMMAND R+ COMMAND R COMMAND EMBED	LLAMA 3.1	MISTRAL 7B MIXTRAL 8x7B MISTRAL LARGE MISTRAL SMALL	STABLE DIFFUSION XL 1.0 & 3 LARGE STABLE IMAGE ULTRA STABLE IMAGE CORE

No one model to rule them all

Enterprises are deploying models from multiple model providers



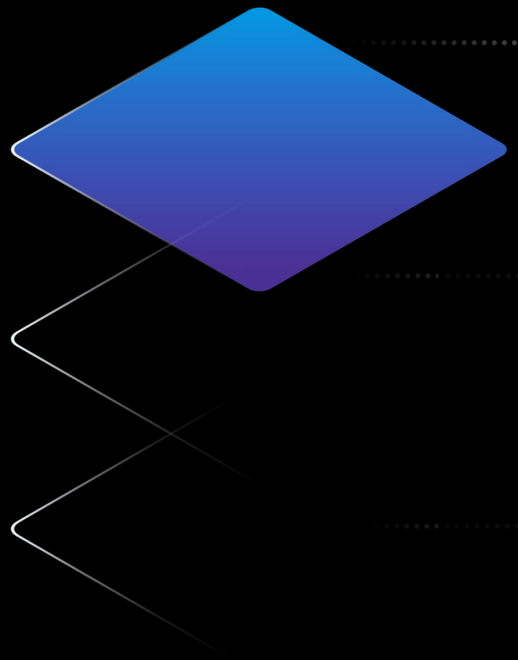

Responsible AI

Amazon Bedrock Guardrails

Implement safeguards customized to your application requirements and responsible AI policies

- Word filters
- Topic filters
- Harmful content filters
- PII filters
- Security
- Prompt injection
- Detect and block hallucinations

Generative AI stack



APPLICATIONS THAT LEVERAGE LLMs & FMs

Amazon Q

AWS App Studio

TOOLS TO BUILD WITH LLMs & OTHER FMs

Amazon Bedrock

Guardrails | Agents | Studio | Customization | Custom Model Import | Amazon Models

INFRASTRUCTURE FOR FM TRAINING & INFERENCE

GPUs Trainium Inferentia SageMaker EFA

UltraClusters EC2 Capacity Blocks Nitro Neuron

NEW

Amazon Q

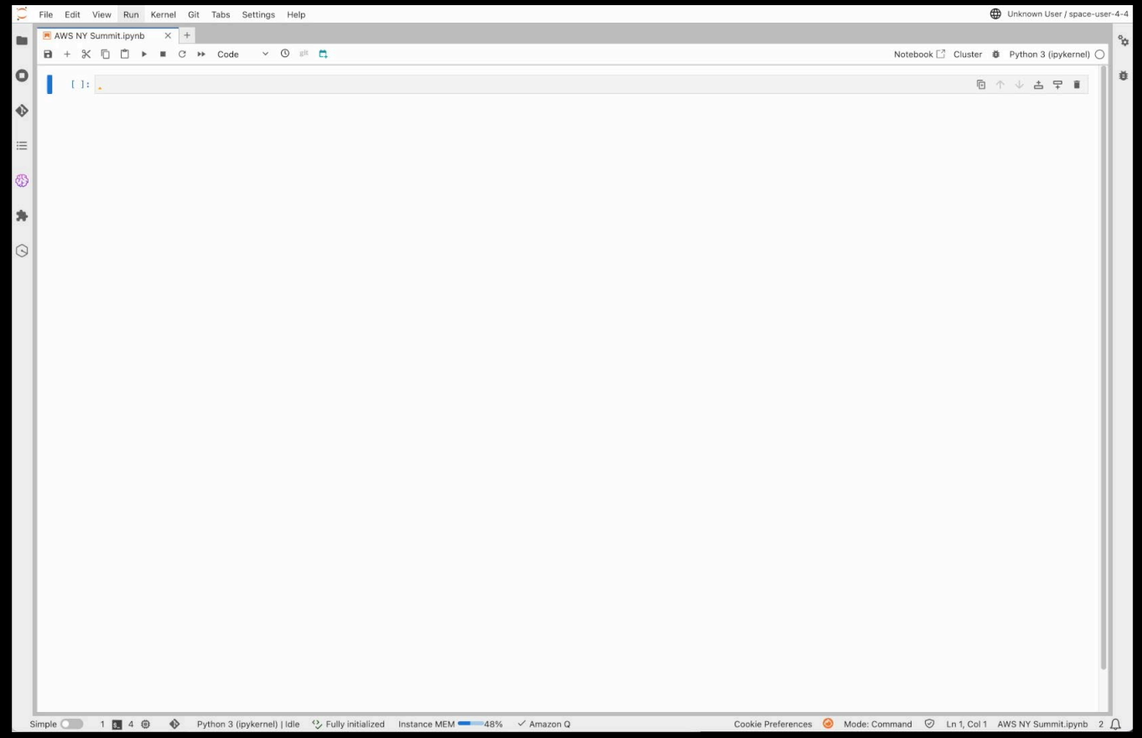
in SageMaker Studio

Build ML models using natural language

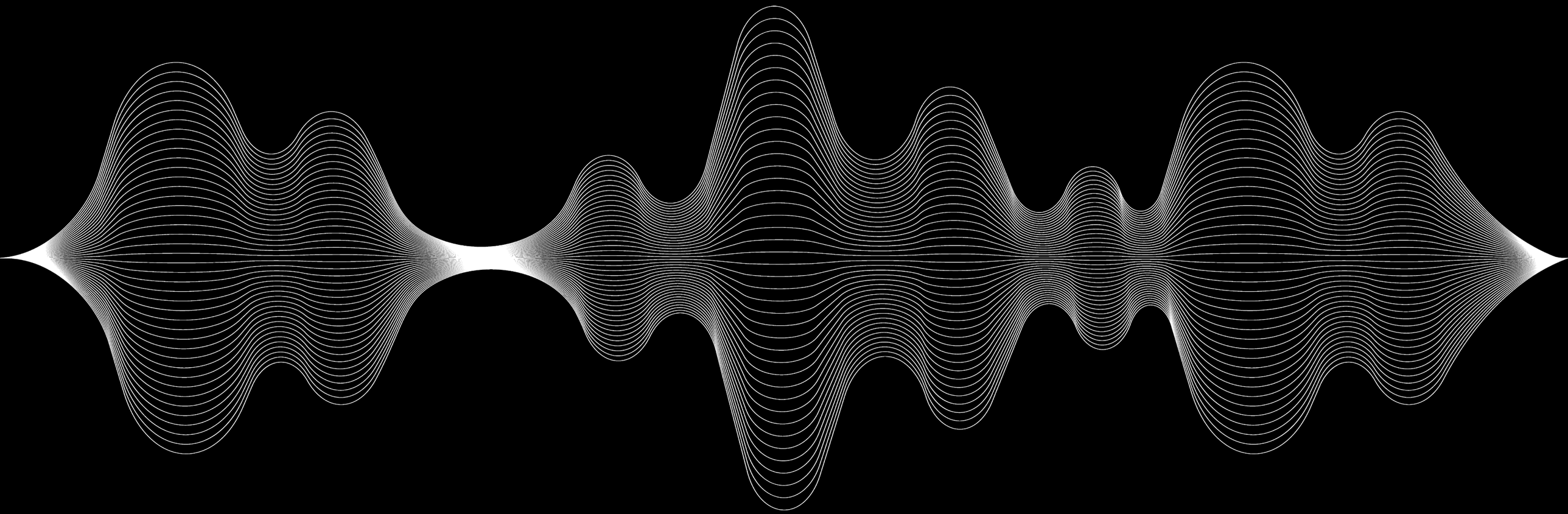
Tailored, step-by-step recommendations inside your SageMaker Studio notebooks

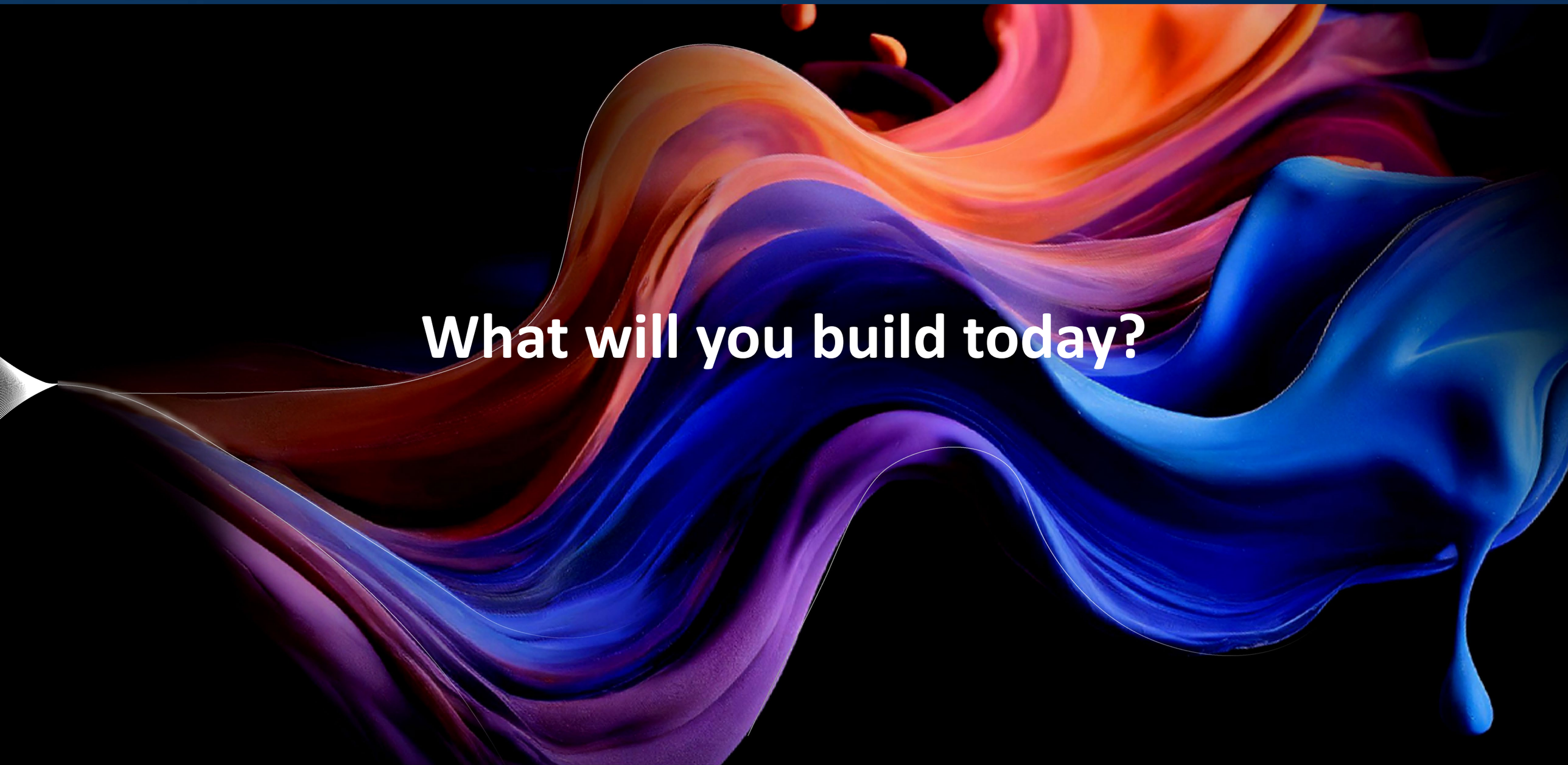
Amazon Q
in SageMaker Studio

1. Product guidance and support
2. Code generation
3. Troubleshooting



There has never been a better time to be a builder




What will you build today?