

AWS RE:INVENT

re:Cap



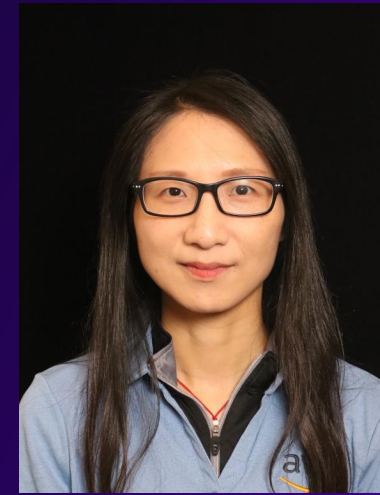
Artificial Intelligence and Machine Learning (AI/ML)



Koushik Konjeti
Sr. Solutions Architect
Federal Financials



Raju Rangan
Sr. Solutions Architect
Federal Financials



Sherry Ding
Sr. AI/ML Specialist SA
Federal

Agenda

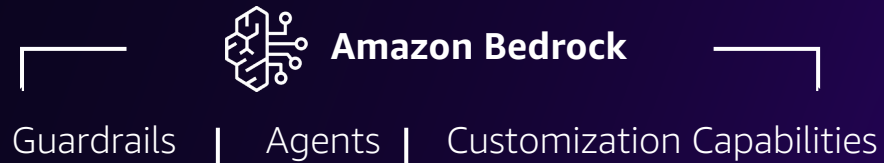
- Part 1: Generative AI
 - Amazon Bedrock
 - PartyRock (Bedrock playground)
 - Amazon Q
- Part 2: Amazon SageMaker
 - New SageMaker Studio
 - SageMaker HyperPod
 - SageMaker Canvas updates
- Part 3: Other AI Services
 - Amazon Personalize
 - Amazon Transcribe
 - Amazon Lex
 - Amazon CodeWhisperer

Generative AI Stack

APPLICATIONS THAT LEVERAGE LLMs AND OTHER FMs



TOOLS TO BUILD WITH LLMs AND OTHER FMs



INFRASTRUCTURE FOR FM TRAINING AND INFERENCE



Amazon Bedrock



THE EASIEST WAY TO BUILD AND SCALE GEN AI APPLICATIONS



- 21+ best in class FMs from 6 providers
- Privately customize models with your data
- Native support for Retrieval Augmented Generation (RAG)
- Agents that execute multistep tasks
- Model Evaluation
- Guardrails (preview)
- Security, privacy, and safety



Amazon Bedrock

BROAD CHOICE OF MODELS



AI21labs

amazon

ANTHROPIC

cohere

Meta

stability.ai

Jurassic

Contextual answers,
summarization,
paraphrasing

Amazon Titan

Text summarization,
generation, Q&A,
search

Claude

Summarization, complex
reasoning, writing, coding

Command + Embed

Text generation,
search, classification

Llama 2

Q&A and reading
comprehension

Stable Diffusion

High-quality images
and art

Amazon Bedrock

BROAD CHOICE OF MODELS

AI21labs

amazon

ANTHROPIC

cohere

Meta

stability.ai

Jurassic

Amazon Titan

Claude
Claude 2.1

Command + Embed

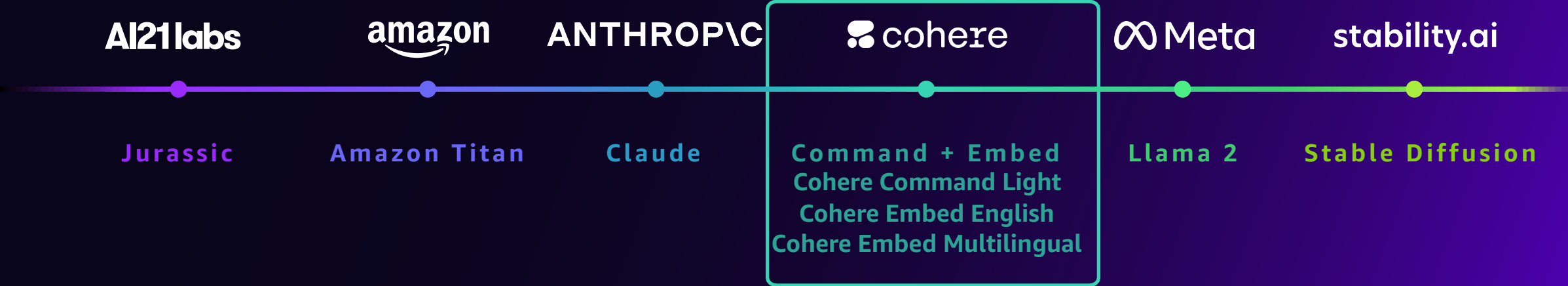
Llama 2

Stable Diffusion



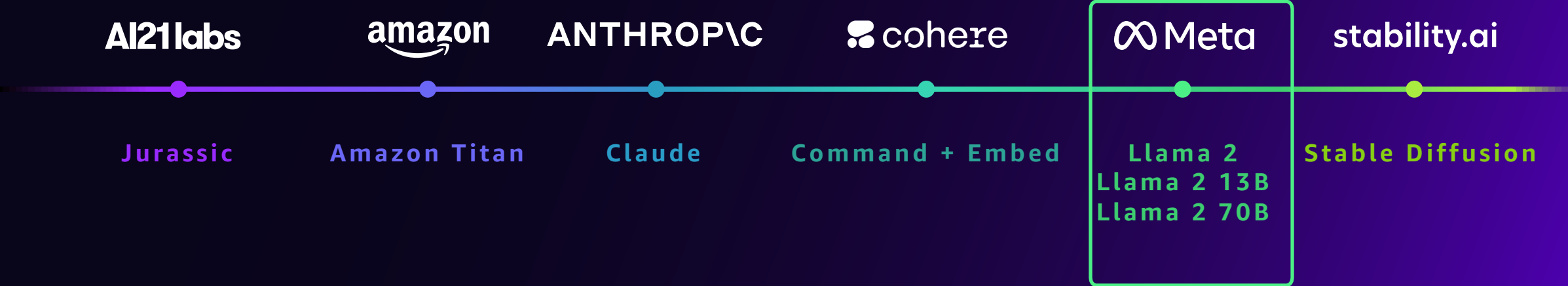
Amazon Bedrock

BROAD CHOICE OF MODELS



Amazon Bedrock

BROAD CHOICE OF MODELS



Amazon Bedrock

BROAD CHOICE OF MODELS

AI21labs

amazon

ANTHROPIC

cohere

Meta

stability.ai

Jurassic

Amazon Titan

Claude

Command + Embed

Llama 2

Stable Diffusion XL 1.0



Amazon Bedrock

EMBEDDINGS FOR TEXT AND IMAGES



NEW

Amazon Titan Multimodal Embeddings

Search, recommendation, and personalization



Accepts text, image, or a combination of text-image to generate embeddings



Adapts to unique and proprietary data



Built-in mitigation to help reduce biased search results

Amazon Bedrock

COST-EFFECTIVE TEXT MODELS



NEW

 Amazon
Titan Text Lite

Summarization, copywriting,
and ideal for fine-tuning

NEW

 Amazon
Titan Text Express

Open-ended text generation,
conversational chat, and RAG support

Amazon Bedrock

GENERATE IMAGE FROM TEXT PROMPT

Preview



NEW

Amazon Titan Image Generator

Generate realistic, studio-quality images



Generate studio-quality images using natural language prompts



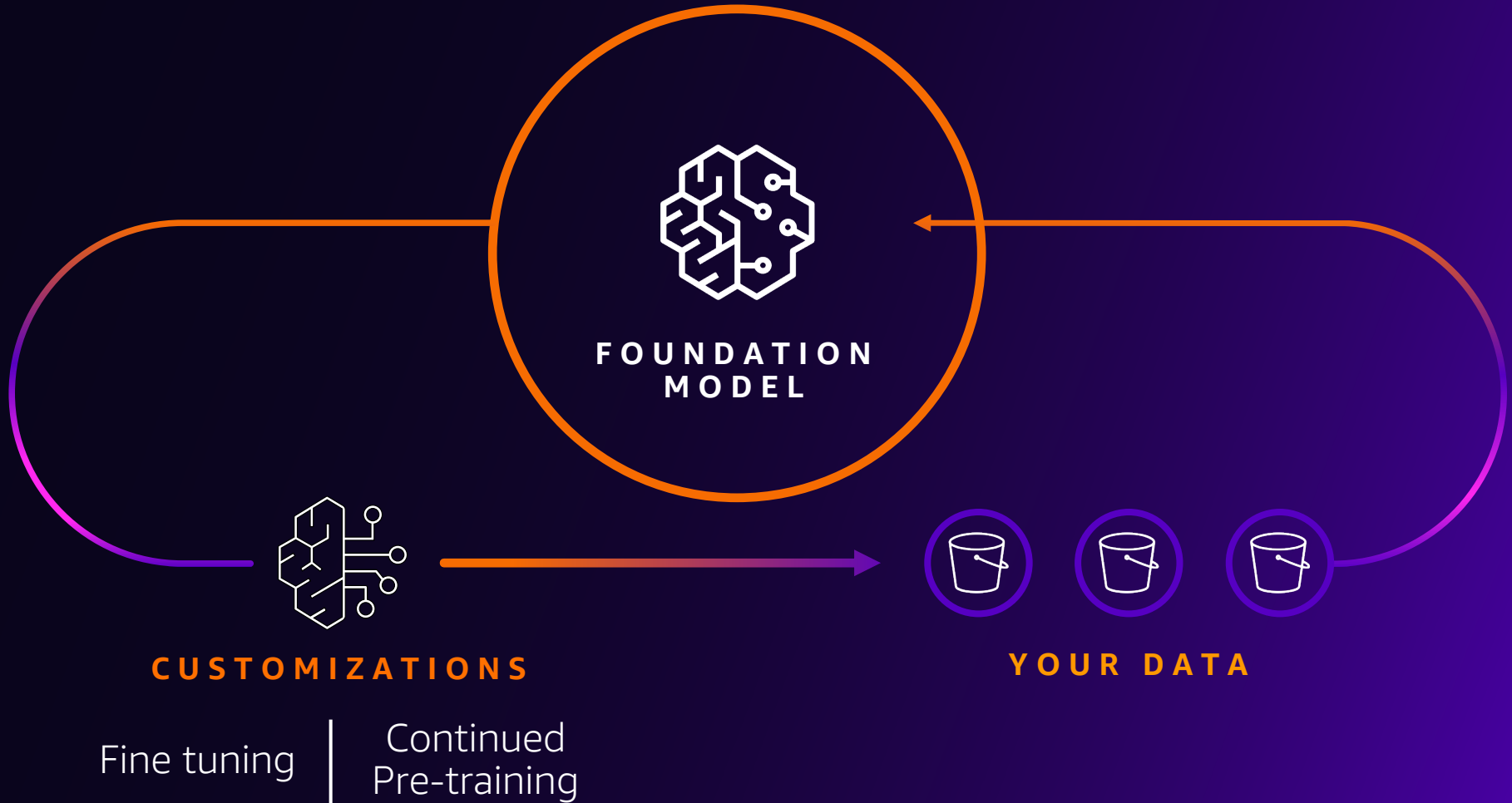
Customize images with proprietary data to match your brand style



Higher scores for text-to-image alignment compared to several leading models

Generative AI Customizations

EASILY CUSTOMIZE IN BEDROCK



Privately customize models with your data

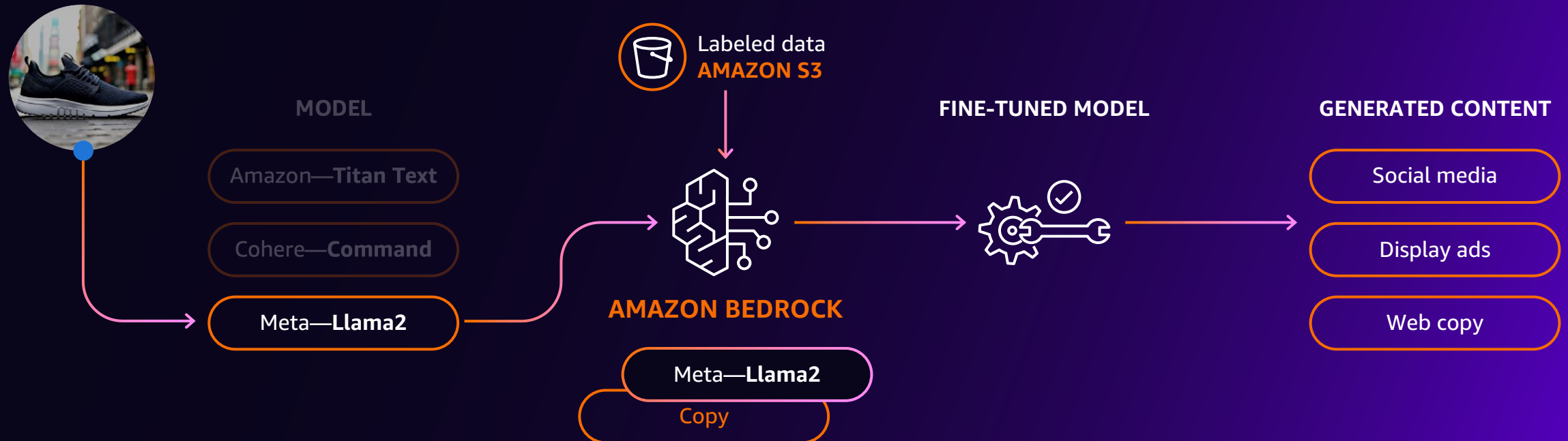
FINE-TUNING AND CONTINUED PRE-TRAINING

Deliver tailored, differentiated tail user experiences with customized FMs

Fine-tune Llama 2, Command, and Titan FMs for specific tasks with labeled data

Use continued pre-training to adapt Titan Text FMs to your domain with unlabeled data

None of your inputs to or outputs from Amazon Bedrock will be used to train the original base models



Model Evaluation on Amazon Bedrock

Preview



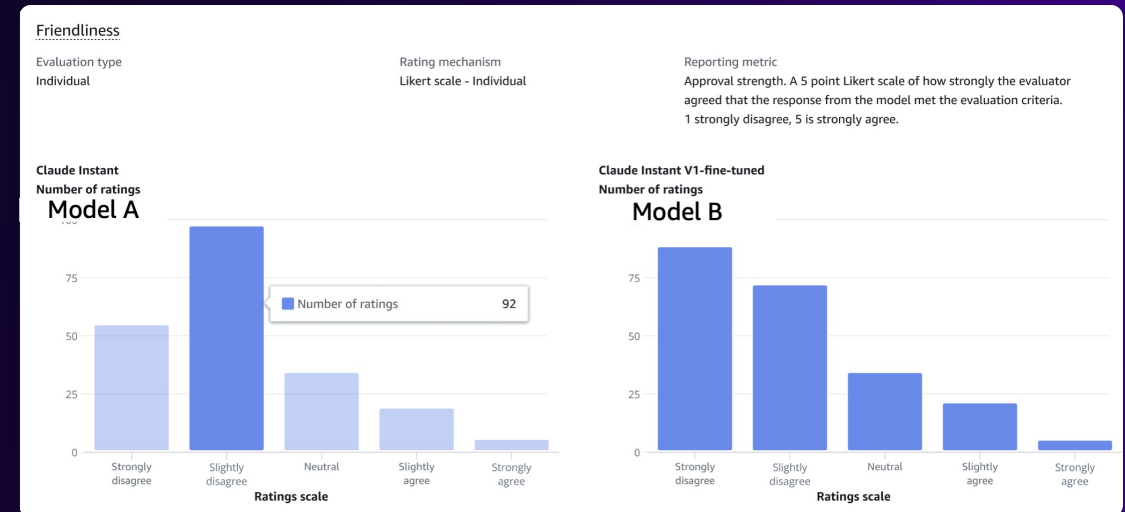
EVALUATE FMS TO SELECT THE BEST ONE FOR YOUR USE CASE

Choose automatic or human evaluation method

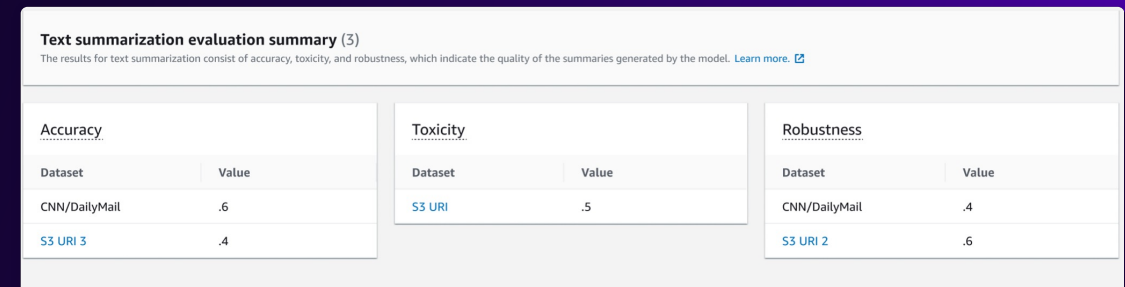
Curated datasets or bring your own

Pre-defined and custom metrics

Human evaluation report



Automatic evaluation report



Knowledge bases for Amazon Bedrock

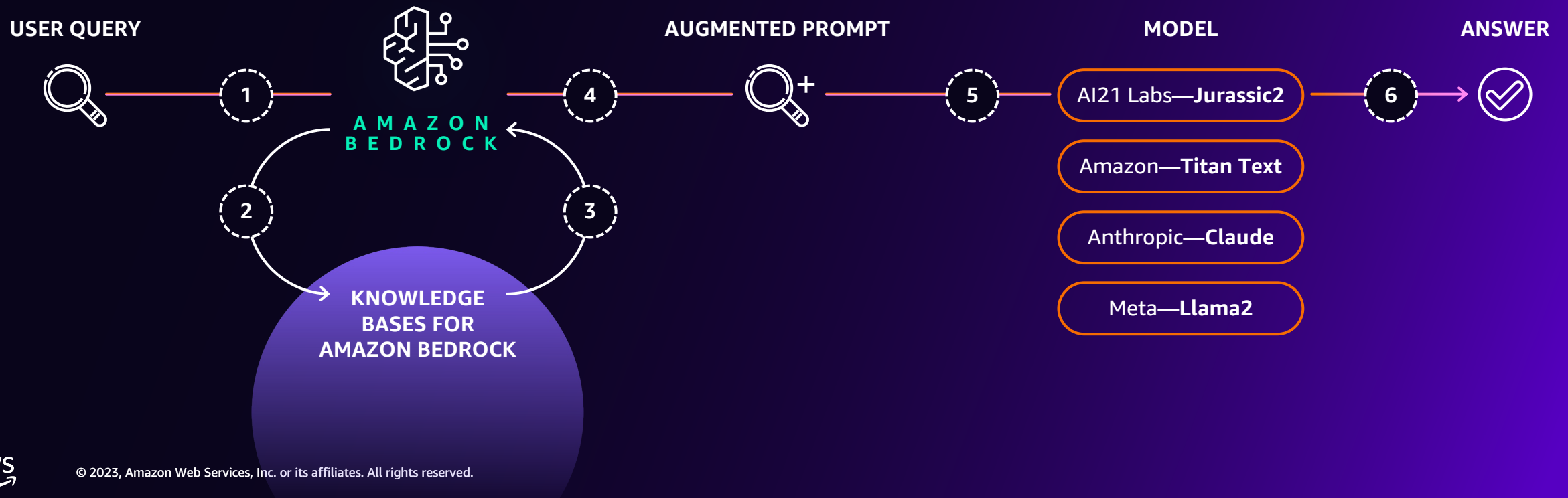
NATIVE SUPPORT FOR RETRIEVAL AUGMENTED GENERATION (RAG)

Securely connect FMs to data sources for RAG to deliver more relevant responses

Fully managed RAG workflow including ingestion, retrieval, and augmentation

Built-in session context management for multi-turn conversations

Automatic citations with retrievals to improve transparency



Amazon Bedrock



OPTIONS FOR NATIVE AND MANAGED RAG SOLUTION

Vector databases for **Amazon Bedrock**



Vector Engine For
Amazon OpenSearch
Serverless



Redis Enterprise
Cloud



Pinecone

COMING SOON



Amazon
Aurora

COMING SOON



MongoDB

Agents for Amazon Bedrock

ENABLE GENERATIVE AI APPLICATIONS TO EXECUTE MULTISTEP TASKS USING COMPANY SYSTEMS AND DATA SOURCES



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 for RAG |

| Takes action by invoking API calls on your behalf |

| Chain-of-thought trace and ability to modify agent prompts |

Responsible AI

INTRODUCING NEW RESPONSIBLE AI INNOVATION

Identifying the best FM for a specific use case: Model Evaluation in Amazon Bedrock

Implementing safeguards: Guardrails for Amazon Bedrock (preview)

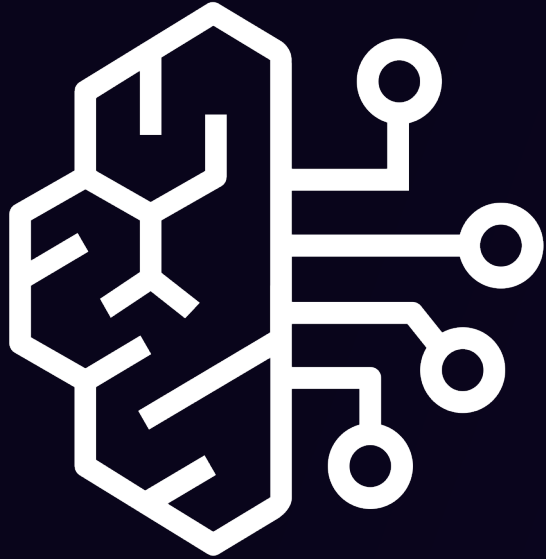
Combating disinformation: Watermarking in Amazon Titan

Building trust: Standing behind our models and applications with indemnification

Enhancing transparency: AWS AI Service Card for Amazon Titan Text



Amazon Bedrock - Other New Features



GDPR, SOC, ISO, CSA compliance and HIPAA eligibility

Bedrock Integration with AWS Step Function

Batch Inference added to Bedrock API

Vector Engine for Amazon OpenSearch Serverless (GA)

Vector Search for Amazon DocumentDB

PartyRock (Free trial)

PartyRock

AN AMAZON BEDROCK PLAYGROUND



PartyRock My Apps

PartyRock

Everyone can build AI apps.

Haiku Creator

Weird Tour Guide

Podcast Generator

Build your own app

An Amazon Bedrock Playground

App builder

→ Generate a 5-7-5 syllable haiku poem and a visual image based on user input.

User Input

winter

Generated Haiku

Here is a 5-7-5 syllable haiku poem about winter:

Snowflakes falling down
Blanketing the ground in white
Winter's cold embrace

Visual

- A shareable generative AI app building playground
- Learn prompt engineering
- Build mini-apps
- Share them with your friends
- Free to use for a limited time
- No credit card or AWS account required



Amazon Q

Preview

YOUR GENERATIVE AI ASSISTANT DESIGNED FOR WORK

- Engages in conversations to solve problems, generate content, and take action
- Understands your company information, code, and system
- Personalizes interactions based on your role and permissions
- Built to be secure and private

The image displays the Amazon Q user interface. At the top left, a window titled 'ASK Q' shows a line chart for 'MoM free trials' with a 'BUILD' button and an 'ADD TO ANALYSIS' button. The chart is interpreted as 'Total free trials by month'. In the center, a woman with her hair in a bun is smiling. To her right is the Amazon Q logo. Below her, a 'New Conversation' window shows a 'Draft a professional email' example from November 15, 2023. The main Amazon Q interface features a greeting: 'Hello, I am your AI assistant. Enter a prompt or choose a suggestion to start a conversation. I'll respond using data from within your organization.' Below the greeting are six buttons: 'Draft a professional email', 'Summarize a report', 'Help plan a workshop', 'Analyze market trends', 'Write a blog post', and 'Brainstorm taglines'. At the bottom of the main interface is a text input field labeled 'Enter a prompt' and an 'Enable Streaming' button. On the right side, a chat window titled 'Amazon Q Preview' shows a greeting: 'Hello! I'm Amazon Q, your AWS generative AI assistant. Ask me anything about AWS services and features or select a sample question below to start a conversation.' Below the greeting are five suggested questions: 'Why can't I SSH to my EC2 instance?', 'What is the CLI command to list all the t3 micro instances in my account?', 'How do I create an Application Load Balancer?', 'When should I use Amazon DynamoDB vs AWS RDS?', and 'I'm learning more every day. Help me improve by providing feedback'. At the bottom of the chat window is an 'Ask a question' input field with a 'Max 200 words' limit and a note that 'Use of Amazon Q is subject to the AWS Responsible AI Policy'.



Amazon Q

Preview



EMPOWER YOUR WORKFORCE WITH GENERATIVE AI



Your generative AI-powered assistant designed for work that can be tailored to your business, code, data, and operations

**Your
business**



**Building
on AWS**



**Amazon
QuickSight**



**Amazon
Connect**



**AWS
Supply Chain**



Amazon Q

Preview

TOP USE CASES



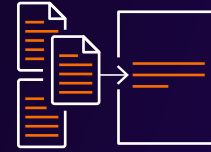
Streamline search experience

Connect internal wikis, SharePoint sites, Confluence, Salesforce, and 40+ other applications for a unified, conversational search experience.



Extract key insights

Perform comparative analysis on your documents and unlock powerful insights so that your employees can make faster, better data-driven decisions.



Generate summaries

Quickly understand the essence of documents by creating summaries of uploaded files or existing enterprise documents.



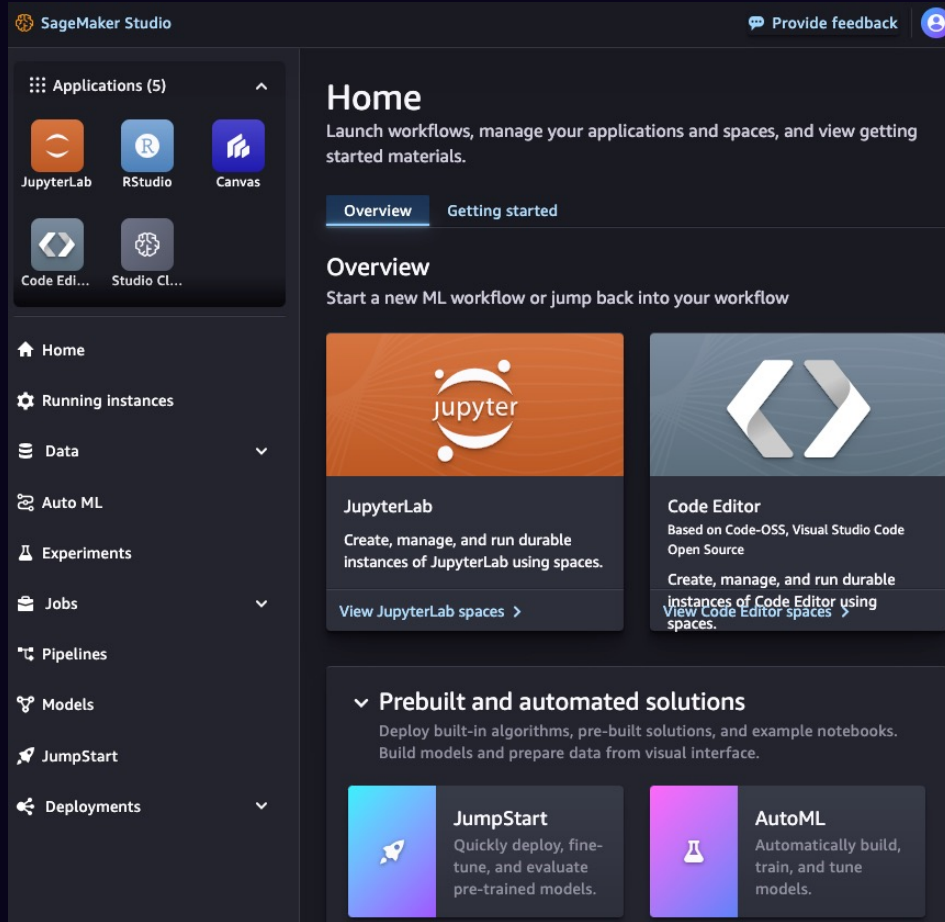
Accelerate content creation

Empower employees to create content faster and smarter across departments - from marketing to sales to engineering.



Amazon SageMaker Studio

FULLY MANAGED INFRASTRUCTURE, TOOLS, AND WORKFLOWS FOR ML-LIFECYCLE



New setup and onboarding experience

Run Jupyterlab 4 application within Studio

VS Code support

Bring your own EFS volume

Includes latest SageMaker Distribution image



Amazon SageMaker HyperPod

DEVELOP AND TRAIN FMS CONTINUOUSLY FOR WEEKS AND MONTHS



Resilient environment

Self-healing clusters reduce training time by up to 20%



Streamline distributed training

SageMaker distributed training libraries improve performance by up to 20%



Optimized resources utilization

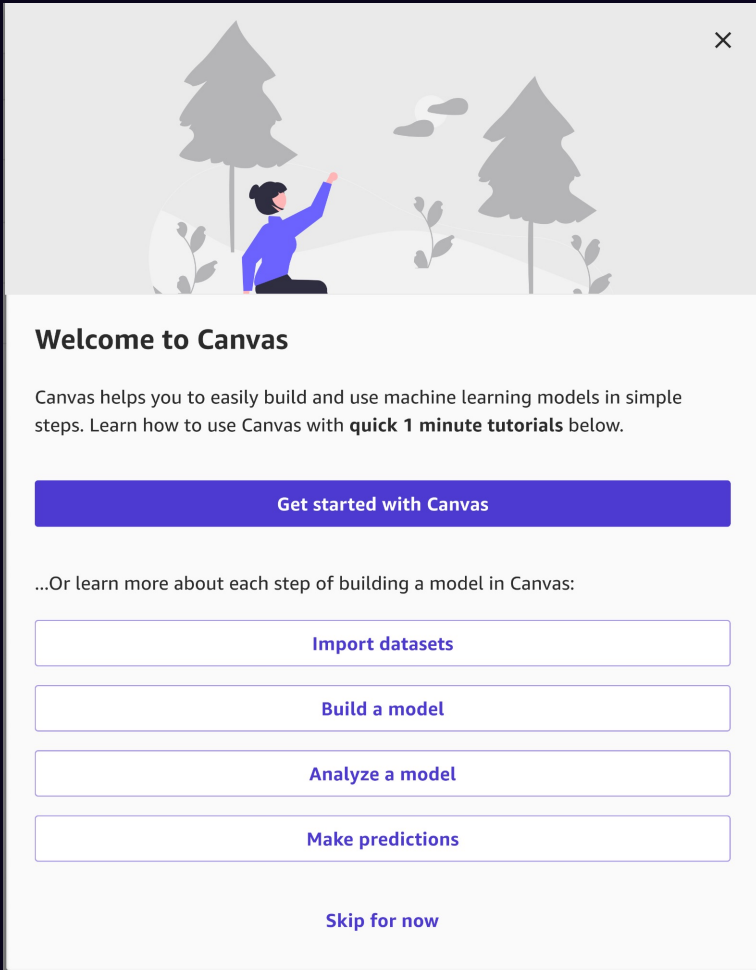
Control over computing environment and workload scheduling





Amazon SageMaker Canvas

BUILD ML MODELS USING A VISUAL INTERFACE, NO CODE REQUIRED

A screenshot of the Amazon SageMaker Canvas welcome interface. At the top, there is a header image showing a person in a blue shirt interacting with a stylized landscape of trees and a bird. Below the image, the text reads "Welcome to Canvas" followed by a paragraph: "Canvas helps you to easily build and use machine learning models in simple steps. Learn how to use Canvas with quick 1 minute tutorials below." A large blue button labeled "Get started with Canvas" is prominently displayed. Below this, the text says "...Or learn more about each step of building a model in Canvas:" followed by four white buttons with blue text: "Import datasets", "Build a model", "Analyze a model", and "Make predictions". At the bottom, there is a "Skip for now" link.

Finetune Foundation Models

Comprehensive Data Preparation (Data Wrangler)

Use Natural Language to Explore and Prepare data



SageMaker Other New Features



A SINGLE WEB-BASED INTERFACE FOR END-TO-END ML DEVELOPMENT

Smart shifting of data for model training (Preview)

Improved SDK tooling and UX for model deployment

New inference capabilities reduce costs and latency

Simplified developer experience for AI/ML workflows

Use SageMaker Clarify to evaluate FMs (Preview)

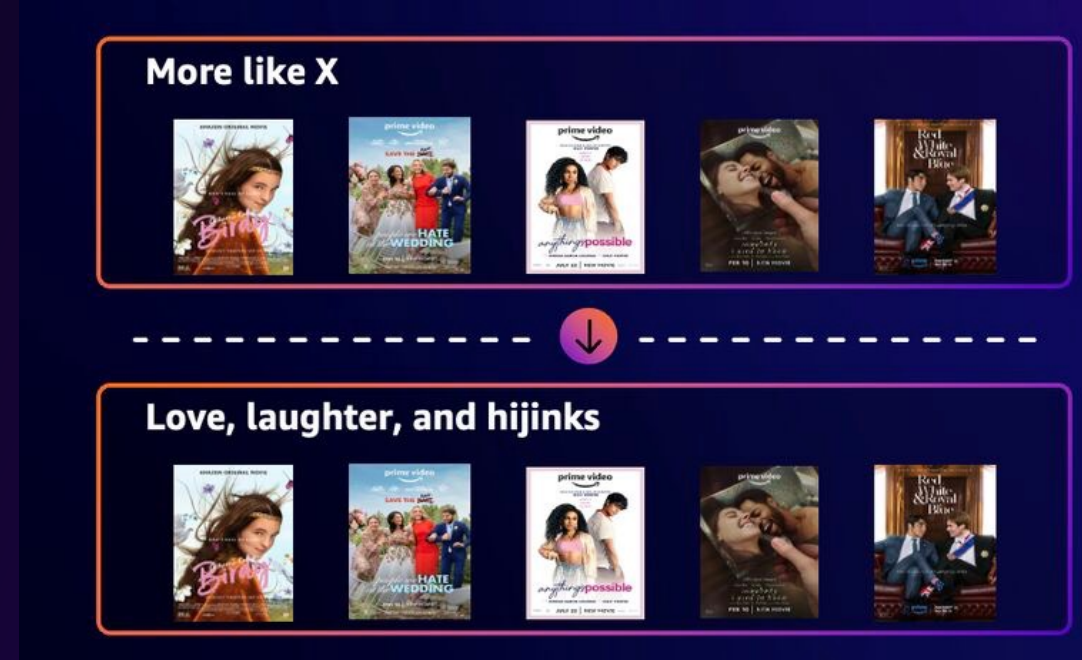
New large model inference DLC with NVIDIA's Tensor-RT-LLM support

Amazon Personalize



DRIVE HYPER-PERSONALIZED EXPERIENCES WITH GENERATIVE AI

- Content Generator to make recommendations more compelling
 - Carousel titles for movie collections
 - Subject lines for marketing emails
- LangChain integration
- Feed prediction response metadata to improve generative AI workflow
- The new Next Best Action recipe helps determine the best user actions to increase brand loyalty and conversion



Amazon Transcribe



AUTOMATICALLY CONVERT SPEECH TO TEXT

- A new speech foundation model-powered ASR system that expands support to over 100 languages
- Automatic language identification for multi-lingual real-time audio streams
- Toxicity Detection for spoken conversations
- Amazon Transcribe Call Analytics now offers generative call summarization (**preview**)





Amazon Lex

OPTIMIZE BOT CREATION AND PERFORMANCE WITH GENERATIVE AI

Build-time

- Descriptive Bot building
- Sample Utterance generation

Descriptive Bot Builder - GenAI [Info](#)

You must have Amazon Bedrock set up in order to use this feature. Please ensure you have requested access to Anthropic's V2 model.

Create a description of the bot you like to build. For example "We want a bot to help customers order food (using item id, quantity, size), check order status, and cancel an order. Use Order ID for indexing order."

Describe your use case

Create a description of the bot you like to build. For example "We want a bot to help customers order food (using item id, quantity, size), check order status, and cancel an order. Use Order ID for indexing order."

Minimum characters: 100, Maximum characters: 2000

Select model

AI Anthropic Claude V2

Run-time

- Assisted Slot resolution
- Conversational FAQ (**preview**)

Runtime generative AI features

These features improve responses to the users of the bots

Assisted slot resolution [Info](#)

Increase the accuracy in how your slots capture information by using LLMs. You can enable assisted slot resolution for AMAZON.Number, AMAZON.Alphanumeric, AMAZON.Date, AMAZON.Confirmation, AMAZON.City, AMAZON.Country and AMAZON.PhoneNumber slots.

Select model

AI Anthropic Claude V2





Amazon CodeWhisperer

BUILD APPLICATIONS FASTER AND MORE SECURELY WITH YOUR AI CODING COMPANION

- Gen AI-powered code remediation for hard-to-find security vulnerabilities
- Supports Customization
- Code suggestions for Infrastructure-as-Code tools:
 - AWS CloudFormation (YAML, JSON)
 - AWS CDK (Typescript, Python)
 - HashiCorp Terraform (HCL)
- Now available in Visual Studio 2022 (**preview**) and supports C#
- Supports hundreds of popular CLIs
 - Git, npm, AWS CLI and Docker



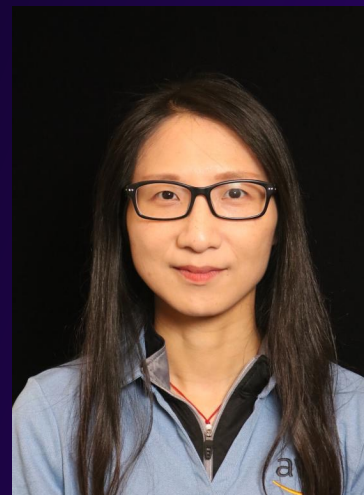
Thank you!



Koushik Konjeti
Sr. Solutions Architect
Federal Financials



Raju Rangan
Sr. Solutions Architect
Federal Financials



Sherry Ding
Sr. AI/ML Specialist SA
Federal



Please complete the session survey

