Priya Arora

GLOBAL HEAD OF THE GENERATIVE AI CENTER OF EXCELLENCE FOR AWS PARTNERS

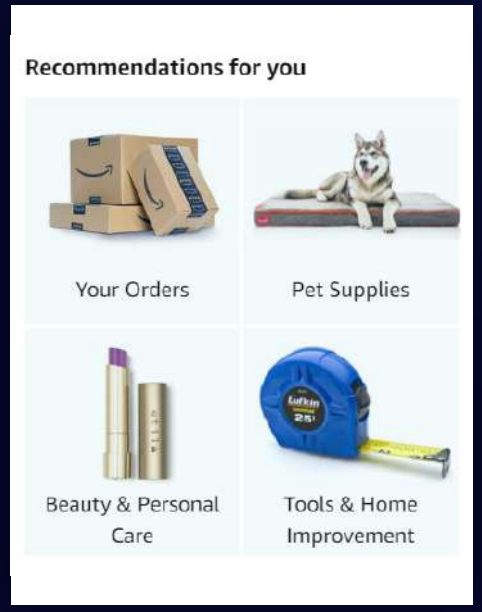
Andre Boaventura

GLOBAL GENERATIVE AI SPECIALIST

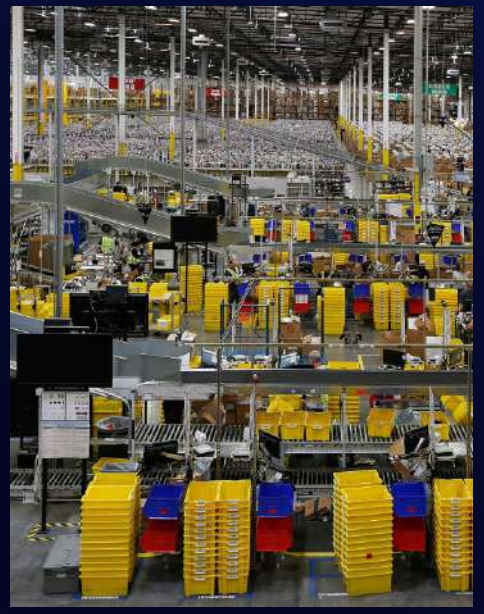
AMAZON WEB SERVICES

Innovation can **transform industries**

ML innovation is in **Amazon's DNA**



4,000 products per minute sold on Amazon.com



1.6M packages every day

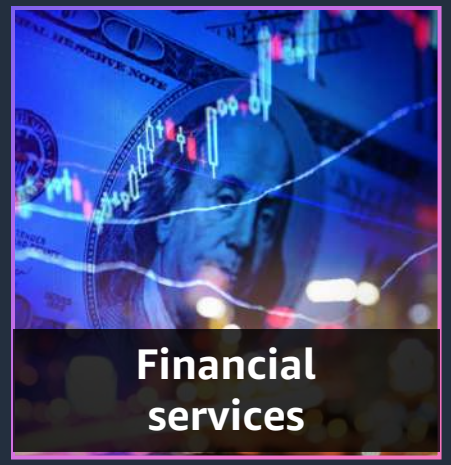
Billions of Alexa interactions each week

First Prime Air delivery on **December 7, 2016**

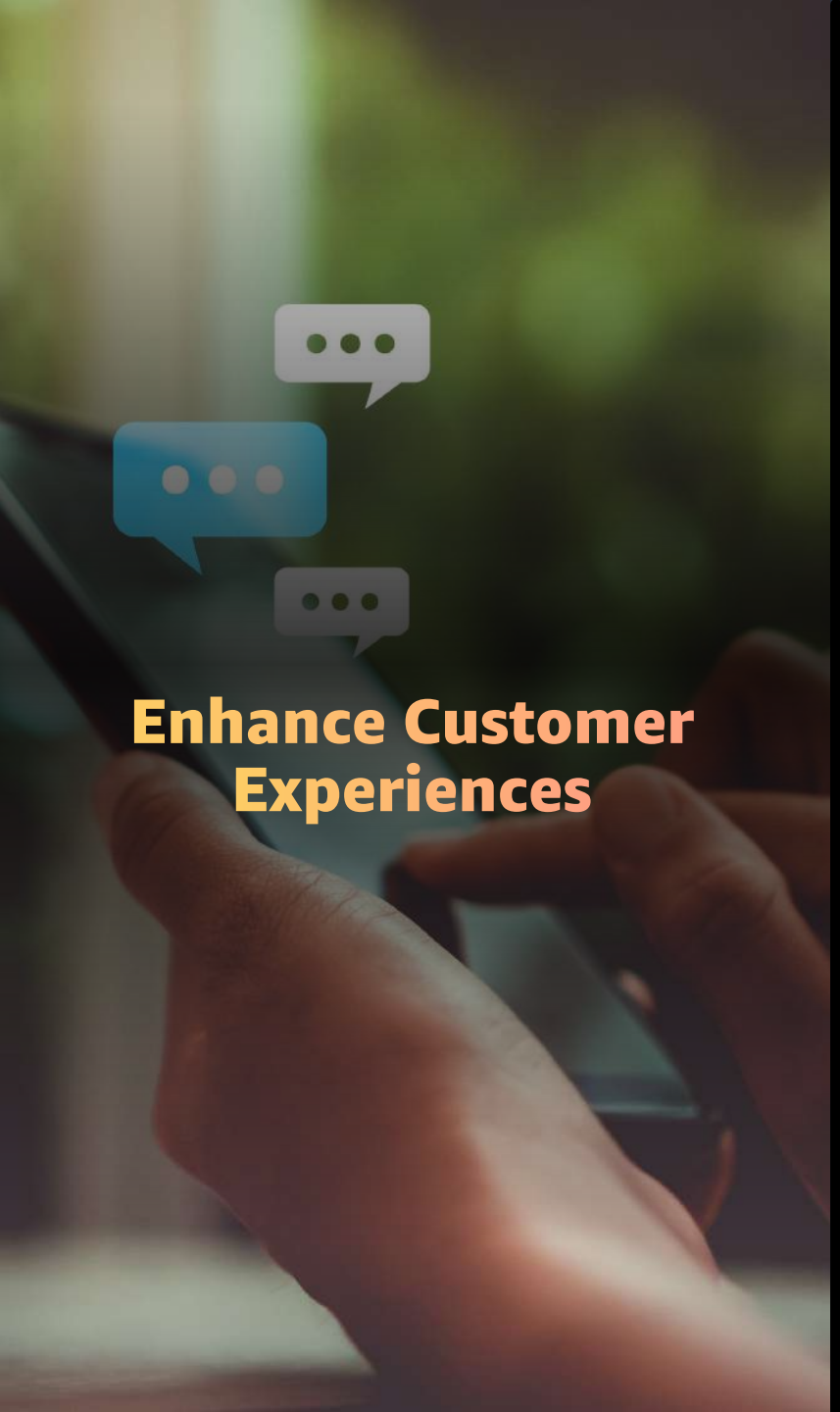
More than 100,000 customers use AWS for ML

The tipping point for **Generative AI**

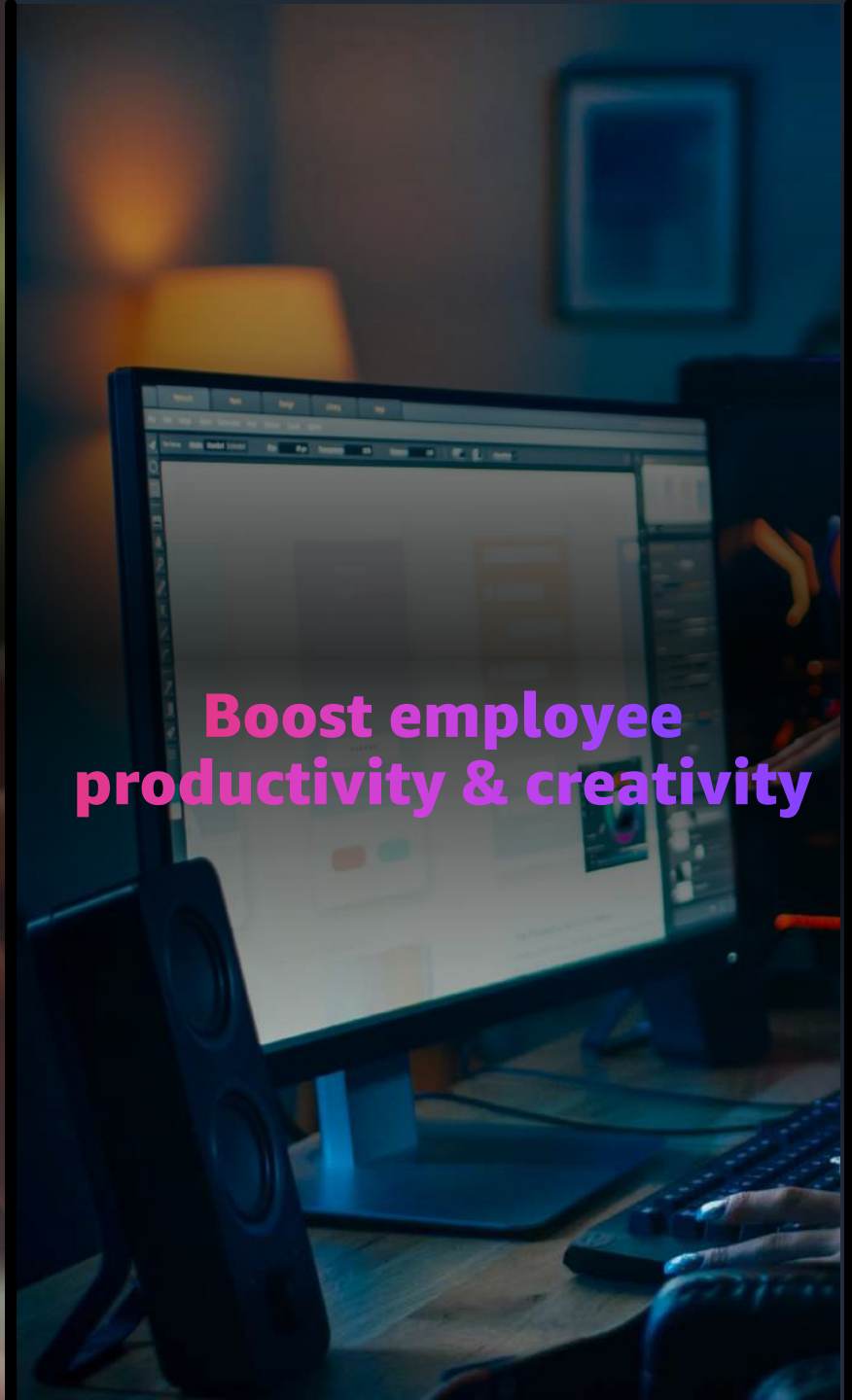
Generative AI is transforming all industries



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Enhance Customer Experiences



Boost employee productivity & creativity

Optimize business processes

Unlocking the potential of generative AI

Easiest way to build
with Foundation Models

The most performant
infrastructure for generative AI

GenAI-powered
applications on AWS

Flexibility to build
your own Foundation Models

NOW GENERALLY AVAILABLE

Amazon **Bedrock**

The easiest way to build and scale generative AI applications with foundation models (FMs)

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

Choose FMs from Amazon, AI21 Labs, Anthropic, Cohere, Meta, and Stability AI to find the right FM for your use case

Privately customize FMs using your organization's data

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

Bedrock Overview

Foundation Model updates

On-demand and Provisioned throughput

Pricing model

Feature Highlights

Demos

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Building generative AI applications is challenging

Accessing multiple FMs and newer versions

Customizing FMs is not easy

Data privacy and security

Getting FMs to execute tasks

Connecting to data sources

Difficult to manage infrastructure

Amazon Bedrock

THE EASIEST WAY TO BUILD AND SCALE GENERATIVE AI APPLICATIONS WITH FMS

Access a range of leading FMs via a single API

Privately customize FMs with your own data

Enable data security and compliance

Build agents that execute complex business tasks by dynamically invoking APIs

Extend the power of FMs with your data using retrieval augmented generation (RAG)

Get the best price performance without managing infrastructure

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

How Amazon Bedrock works

Amazon Bedrock

Build generative AI applications using FMs through a serverless API service

Choose a FM

Use the playground to experiment with FMs and select the one that suits your needs

Use as is or customize

Fine-tune FMs as needed; Bedrock will automatically deploy the FM for inference

Send prompt

Use Bedrock API to send your prompts to the model

Receive response

Receive model response in your application

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Benefits

- Choice of leading foundation models
- Easy model customization
- Fully managed agents to execute tasks
- Native support for RAG
- Security and compliance

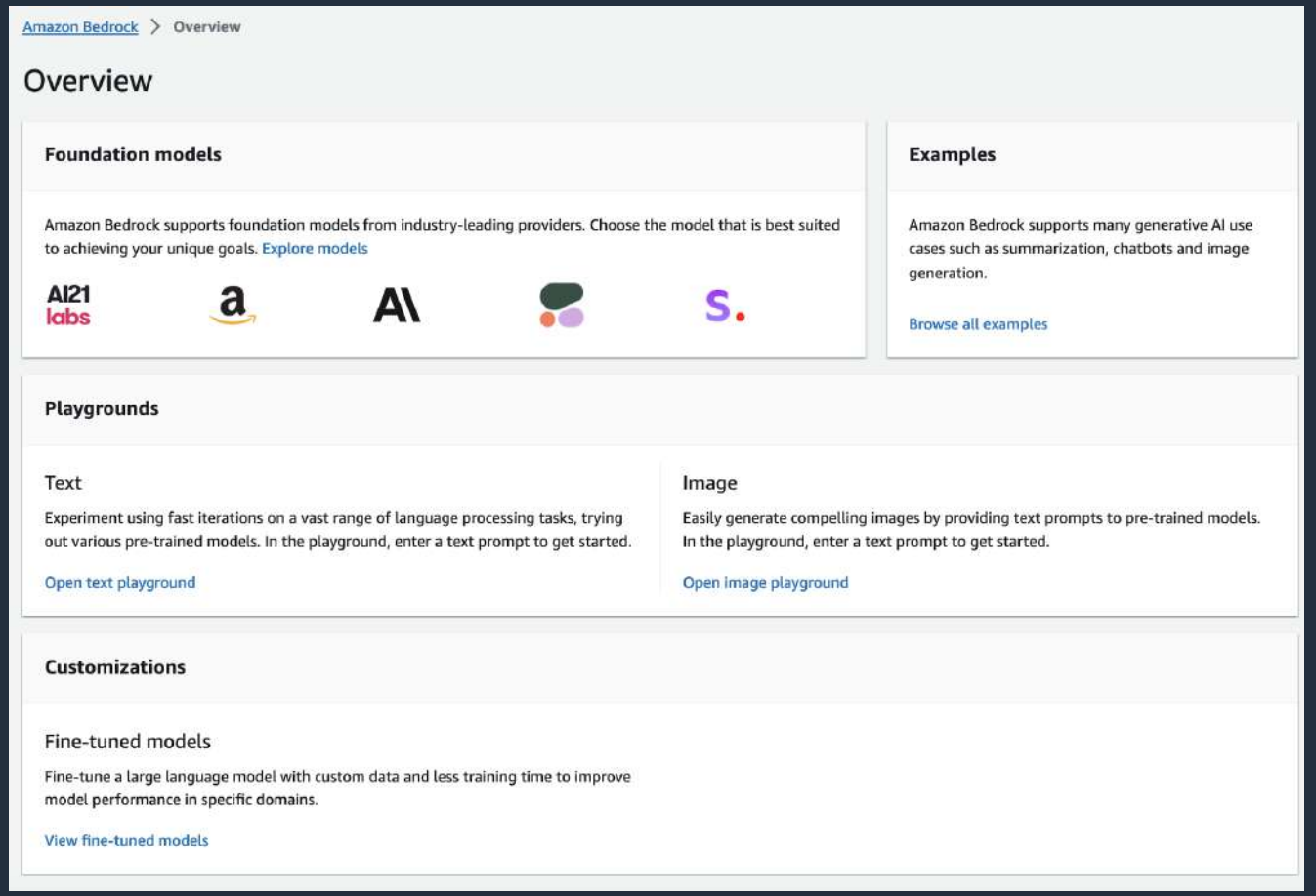
© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Work with a range of leading FMs

Use a single API access to FMs from Amazon and leading AI companies

Privately customize FMs for a particular task without annotating large volumes of data

Scale with managed infrastructure



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Foundation model updates

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Amazon Bedrock supports leading foundation models

amazon

Amazon Titan

Text summarization, generation, classification, open-ended Q&A, information extraction, embeddings and search

AI21labs

Jurassic-2

Multilingual LLMs for text generation in Spanish, French, German, Portuguese, Italian, and Dutch

ANTHROPIC

Claude 2

LLM for conversations, question answering, and workflow automation based on research into training honest and responsible AI systems

cohere

Command and embed

Text generation model for business applications and embeddings model for search, clustering, or classification in 100+ languages

Meta

COMING SOON

Llama 2

Fine-tuned models ideal for dialogue use-cases and language tasks

stability.ai

Stable Diffusion

Generation of unique, realistic, high-quality images, art, logos, and designs

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Tools for storing and retrieving **vector embeddings**

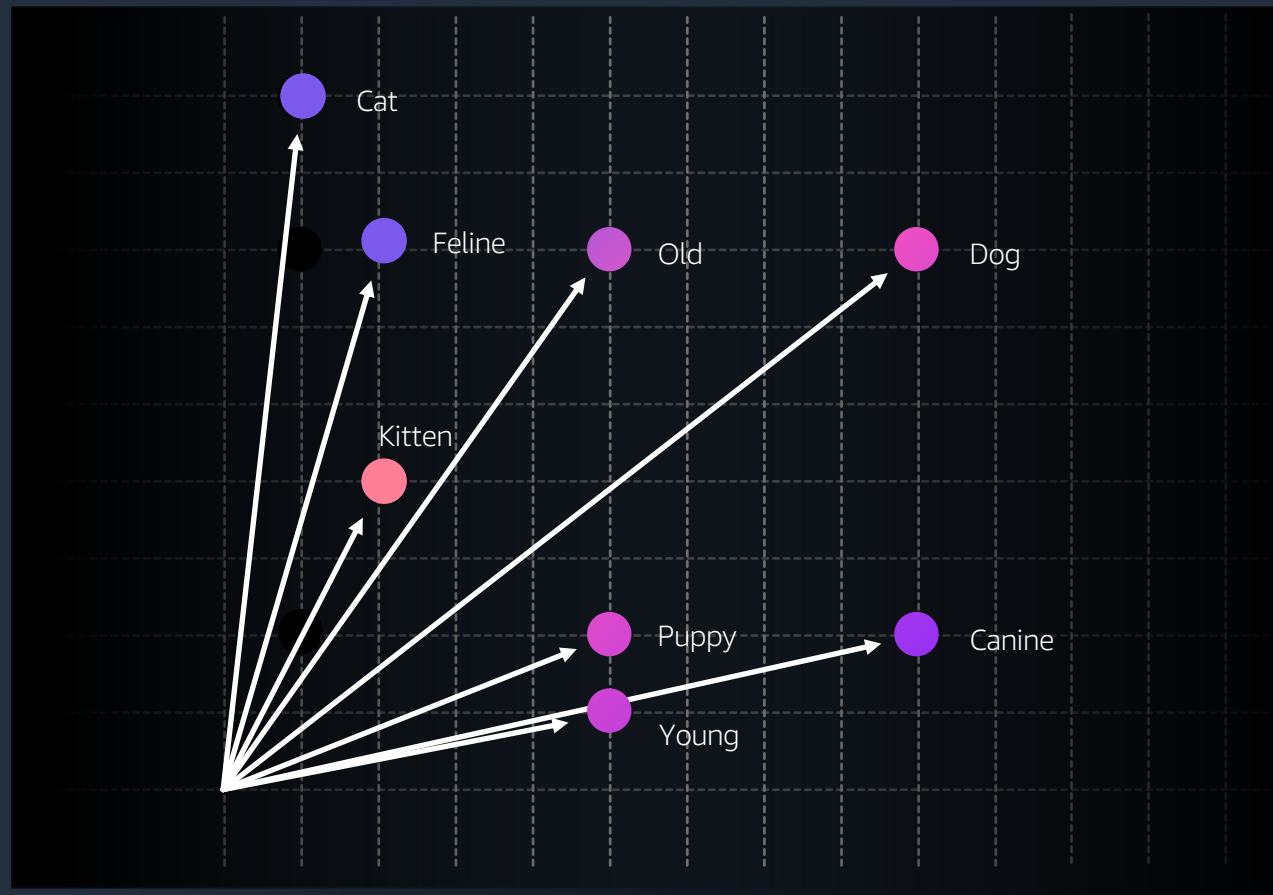
© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

HUMAN TEXT

You may return or exchange your order within 30 days of when the order was shipped. Full refunds will be processed back to the original payment method. After 30 days, we offer a store credit only. Customers will be responsible for the cost to ship returns or exchanges after 30 days.

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Words as **vectors**



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

WITHOUT SEMANTIC SEARCH

Result 1

10 results

1

Not applicable

text: Genuine FO Shoes Bag Zipped Sports Bag Shoe Caase -Black Color. G JOY Golf Shoes Bag & Sports Bag Descriptions: P : FOOTJOY Golf Shoes Bag Material: PU Color: Black Size: 24 (W) x 35 (L) x 11 (H) (cm) Side mesh type.

2

Not applicable

text: NIKE Genu g Zipped Shoe Bag Golf Shoe Sports Gym Pouc NIKE Shoes Bag Information: Model: Nike golf shoe bag (hoes Pouch) Color: Black Size: 36 x 23 x 15 (cm) Material: Polyester

3

Not applicable

text: 1 Pairs S Stretcher for Golf Shoes. This pair of shoes stretc ep your golf shoes in style! It is made of high quality solid and durable in use. used inside the shoe to preserve its shape, stop it from

4

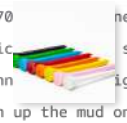
Not applicable

text: Nike Cont Golf Shoes (3 Y, White/Metallic Silver/Bright C >#1.5 Swoosh design trad embroidery at l hest and Nike Athletic Department logo clip label at front lower right hem. Dark grey heather jersey at

5

Not applicable

text: 100Pcs 70 ner Mixed Color Plastic Golf TEES. Made of plastic strong;Packed in reseal able bag for ease of use;Inn ign, four corners to support;Flat end even can clean up the mud on shoes or golf;Material: ABS



Golf Shoes Bag

Golf Shoes Bag

Golf Shoes Bag

Golf Shoe Case

Stretcher for Golf Shoes

Golf Shoes

Plastic Golf Tees.

WITH SEMANTIC SEARCH

Result 2

10 results

1

Not applicable

tJoy adies Summer Series Golf Shoes 98847 White/Tan oseou . The FootJoy Womens Superlite Spikeless golf desi ned to provide versatility and style on and off the course for all lady golfers. This ne spikeless sole design

Golf Shoes

2

Not applicable

tJoy TJ54252-13 Wide Contour Casual Mens Golf Sh n - 1 Wide. Pro Am Golf has specialized in the sale of roduc s for over 35 years Our products are designed to that it helps you to improve the game Utmost

golf

Golf Shoes

3

Not applicable

FootJoy uperlites Athletic Spikeless Golf Shoes 2016 White Grey/Black Medium 8.5. Pro Am Golf has specilized le of golfing products for over 35 yea ducts ned t make sure that it helps you to improve the game

Golf Shoes

4

Not applicable

FootJoy uperlites Athletic Spikeless Golf Shoes 2016 Dark lue/Lime Medium 10.5. Pro AM Gol cialized le of golfing products for over 35 years Our products ned t make sure that it helps you to improve the game

Golf Shoes

5

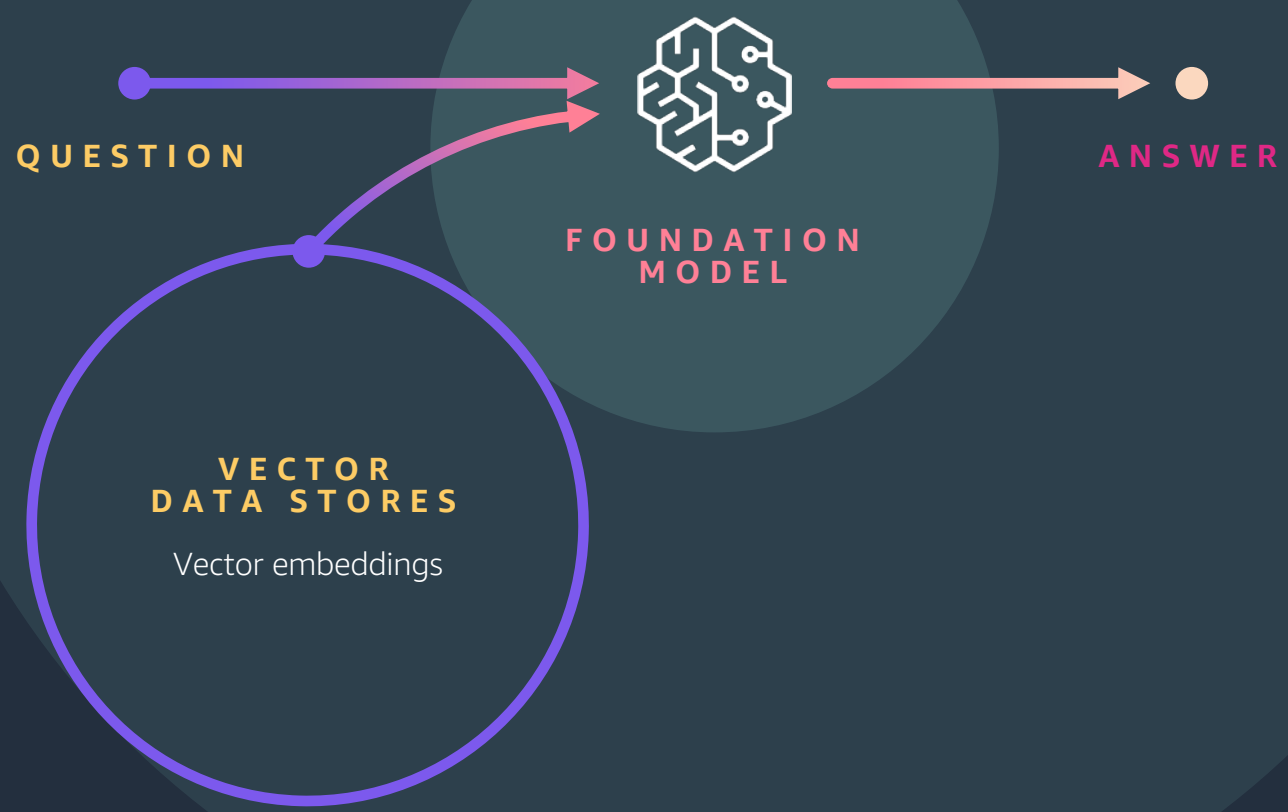
Not applicable

text: FootJoy S perlites Athletic Spikeless Golf Shoes 2016 White/Grey/Black Medium 11.5. Pro AM Gol specialized le of olfing products for over 35 years Our products ed make sure that it helps you to improve the game

Golf Shoes

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Vector embeddings support RAG



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Titan Text Embeddings

Titan Text Embeddings NEW

V2.0

Translates text inputs (words, phrases) into numerical representations (embeddings). Comparing embeddings produces more relevant and contextual responses than word matching.

Max Tokens: **8,000**
Output Vectors**: 1,536**
Language: **Multilingual** (25 languages)

Model ID: *amazon.titan-embed-g1-text-02*

Highlights

- Titan Text Embeddings offers fast, cost effective, high-performance, accurate embeddings in 25 languages.
- Optimized for text retrieval tasks, semantic similarity and clustering.
- Applications of this model includes semantic search and personalization.

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Understanding Amazon Bedrock inference rates, limits, and throughput

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Summary of Amazon Bedrock Inference options

On demand

Pay-as-you-go; no usage commitments

- Pricing based on input and output token count for LLMs
- Great for prototyping, POC's, small workloads with more relaxed requirements for throughput and latency
- Requests per minute (RPM) and Tokens per minute (TPM) enforced

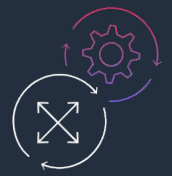
Provisioned throughput

Provision sufficient throughput to meet your application's performance requirements

- Guaranteed throughput at a fixed cost
- Higher throughput available
- Flexible commitment term of 1 month or 6 months
- Pay hourly rate, discounted for extended commit
- Great for production workloads, or inference on custom models

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Understanding on-demand RPM and TPM limits



On demand

- Hard limits on runtime quotas enforced
- Supports variety of workload types
  - High number of requests with low number of tokens
  - High number of tokens with low number of requests
  - And anywhere in between

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Provisioned Throughput (PT) model units

Provisioned throughput

- Model Units provide throughput for a specific model, measured in terms of:

  Max Input TPM | Max Output TPM

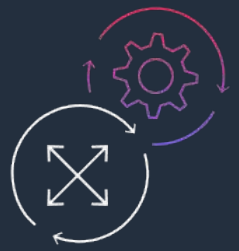
- Workloads use both input and output throughput
- Workloads use model units to achieve either one of those maximums at a time, or some portion of each
- Customers buy the number of model units that match their requirements
- Customers can flexibly use on-demand and multiple model units in the same account

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Model pricing

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Pricing dimensions



On demand

Pay-as-you-go; no usage commitments

Provisioned throughput

Provision sufficient throughput to meet your application's performance requirements

Model customization

Enable personalization and optimize solutions tailored to your business needs.

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

On demand pricing for text inference

Modality: Text and Text Embeddings

Model Provider	Model Name	Context Window	Prompt [per 1M tokens]	Completions [per 1M tokens]
AI21	J2 Mid	8,191	$12.50	$12.50
AI21	J2 Ultra	8,191	$18.80	$18.80
Amazon	Titan Text – Lite	4,000	$0.30	$0.40
Amazon	Titan Text – Express	8,000	$1.30	$1.70
Anthropic	Claude Instant	100,000	$1.63	$5.51
Anthropic	Claude 2	100,000+	$11.02	$32.68
Cohere	Command	4,096	$15.00	$20.00

Model Provider	Model Name	Context Window	Output vector	Input [per 1M tokens]
Amazon	Titan Embeddings – Text	8,000	1,536	$0.10

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

On demand pricing for image inference

Modality: Image generation

Model Provider	Model Name	Image Resolution	Standard Quality [per 1,000 images]	Premium Quality [per 1,000 images]
Stability AI	SDXL 0.8	512 x 512 or smaller	$18.00	$36.00
Stability AI	SDXL 0.8	Larger than 512 x 512	$36.00	$72.00

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Pricing for custom model training & inference

Modality: Text

Model Provider	Model Name	Price to train [1M tokens]	Price to infer from a custom model for 1 mu/hr.* [No commitment]	1 mo. commitment $/hr./mu	6 mos. commitment $/hr./mu
Amazon	Titan Text – Lite	$0.40	$7.10	$6.40	$5.10
Amazon	Titan Text – Express	$0.80	$20.50	$18.40	$14.80

Price to store each custom model per month: **$1.95**

* No-commit option of 1 model unit is only available for inferencing from custom models

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data Privacy & Security

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Data privacy

You are always in control of your data

- Customer data is not used to improve Amazon Titan models for other customers, and is not shared with other foundation model providers
- Customer data (fine-tuned models) remain in the region where they are created

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data security

You are always in control of your data

- Support for **AWS PrivateLink** so customers can establish private connectivity between virtual private clouds (VPCs) and the Bedrock service using VPC Endpoints
- Integration with AWS Identity and Access Management Service (IAM) to manage inference access, deny access for specific models, and enable Console access
- You can use CloudTrail to monitor API activity and troubleshoot issues as you build solutions
- Fine-tuned (customized) models are encrypted and stored using service managed keys; only you have access to your customized models through an endpoint
- Support for **Customer Managed Keys (CMK)** so customers can create and control keys to encrypt fine-tuned models
- Support for VPC configuration of fine-tuning jobs

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Additional feature highlights

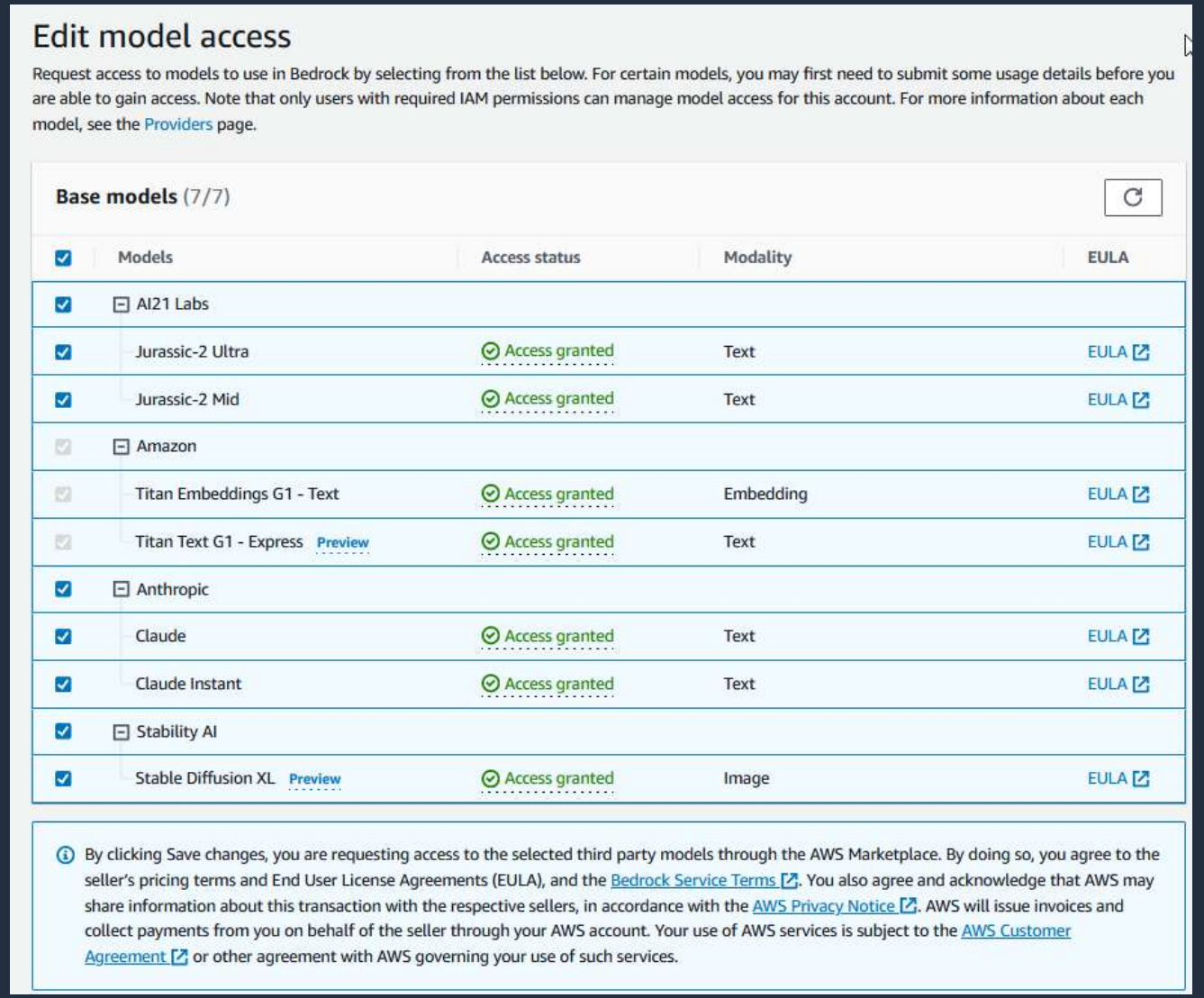
© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Model access

Fine grained model access

Customers can now easily subscribe to specific foundation models that best fit their use cases

For Claude models, customers will continue to ‘Request’ access by completing *Anthropic’s models usage details* form.



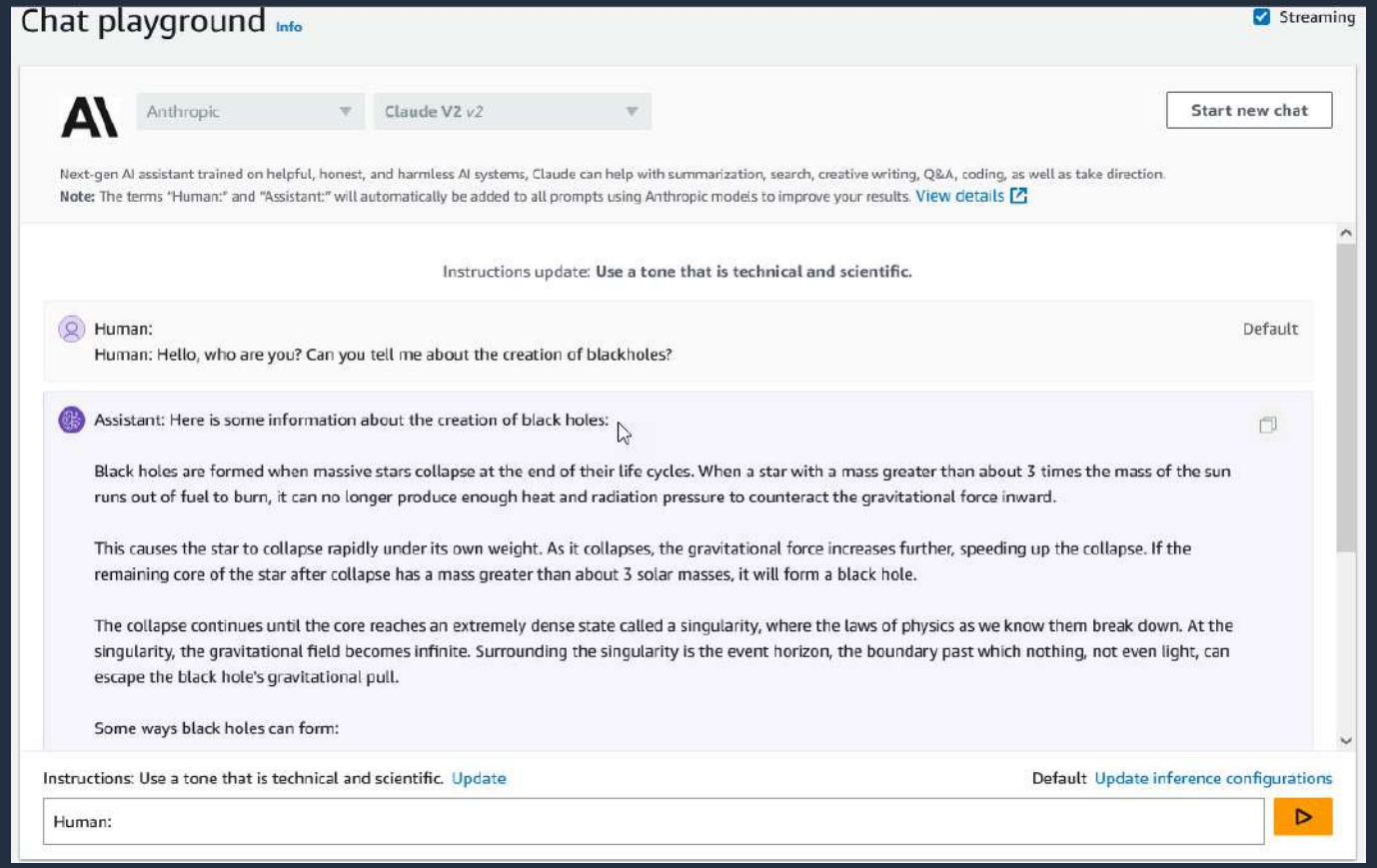
© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon Bedrock Playground

Quickly evaluate Foundation models using the Chat playground

Provide instructions to assume a role throughout the conversation in its entirety

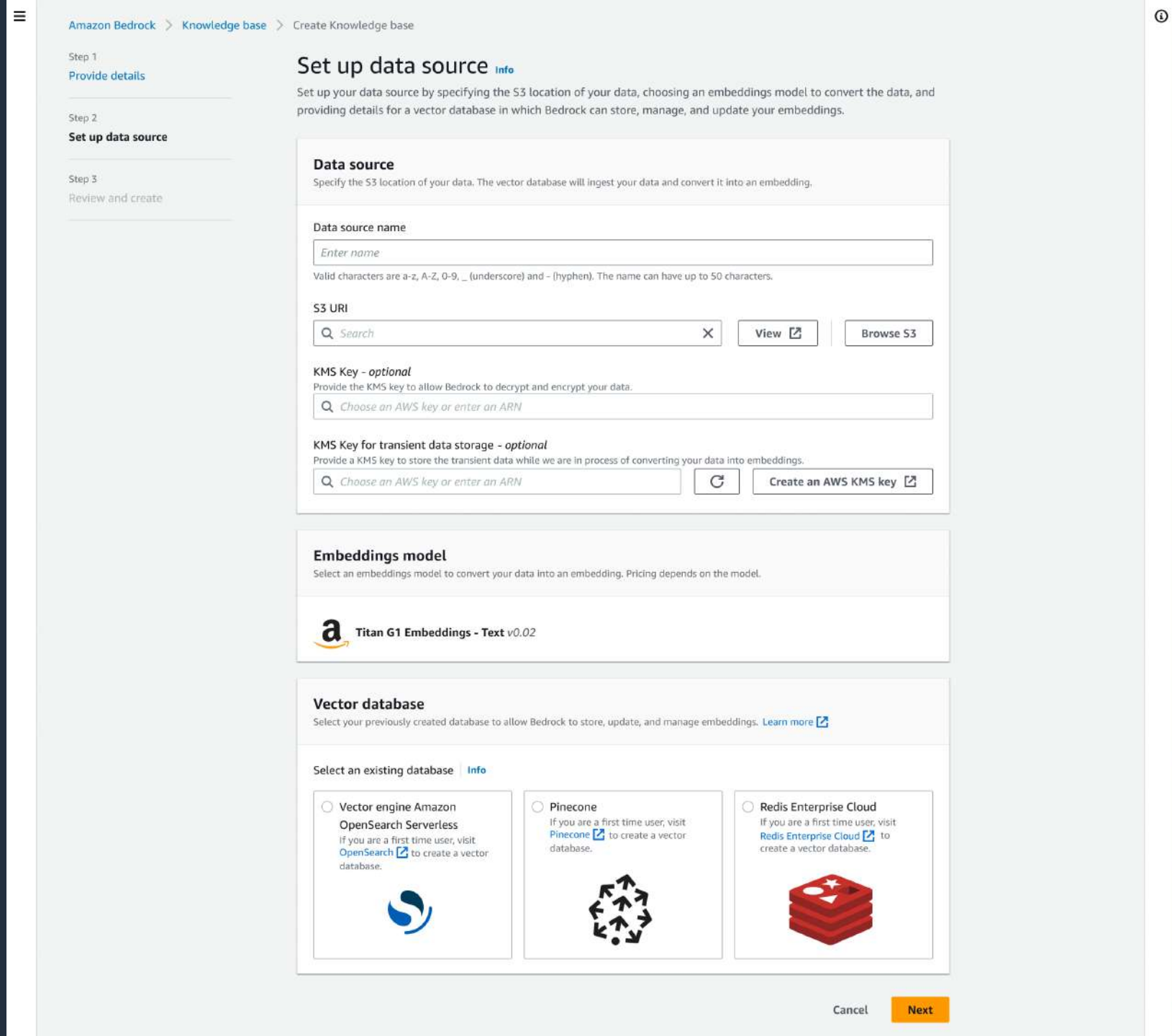
Prompts and responses are not retained once the browser session is complete, or by clicking on ‘Start new chat’



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Knowledge base for retrieval augmented generation (RAG) PREVIEW

- Connect FMs to data sources including vector engine for Amazon OpenSearch Serverless, Pinecone, and Redis Enterprise Cloud
- Enable automatic data source detection
- Provide source attribution



Agents enable generative AI applications to complete tasks in just a few clicks

PREVIEW

1

SELECT YOUR FOUNDATION MODEL

2

PROVIDE BASIC INSTRUCTIONS

3

SELECT RELEVANT DATA SOURCES

4

DEVELOPER SPECIFIES LAMBDA FUNCTIONS

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

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Custom models

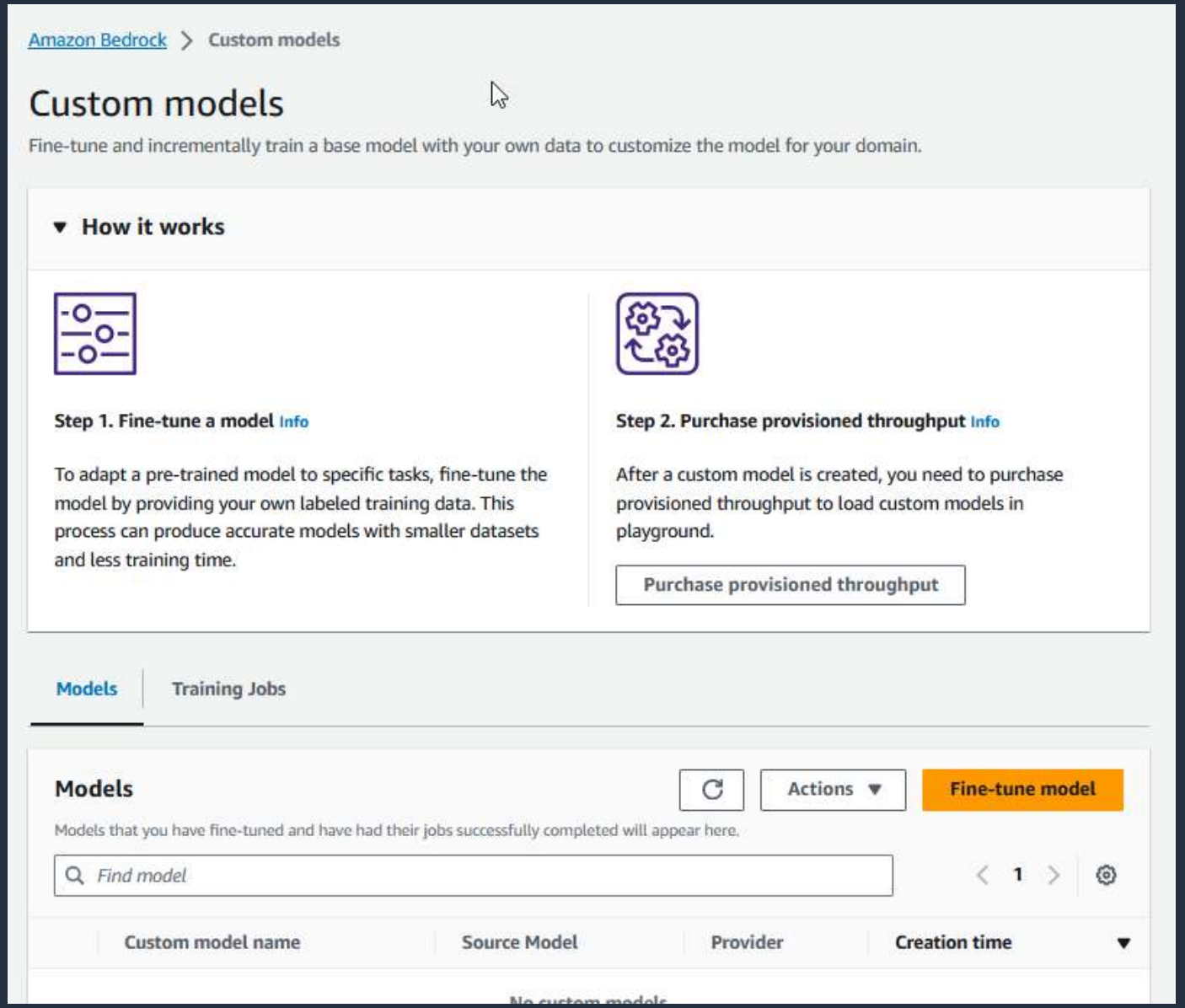
Create fine-tuned models using APIs

Fine-tuned models can be invoked the same way as base models, via the playground, or through the API on Provisioned throughput

Currently, fine-tuning is supported for Titan Text models. Support for other models will be enabled over time.

Customers with access to Amazon Titan Text models will now have access to fine tuning NEW

Tagging lets you easily manage, identify, organize, search for, and filter fine tuning jobs and models



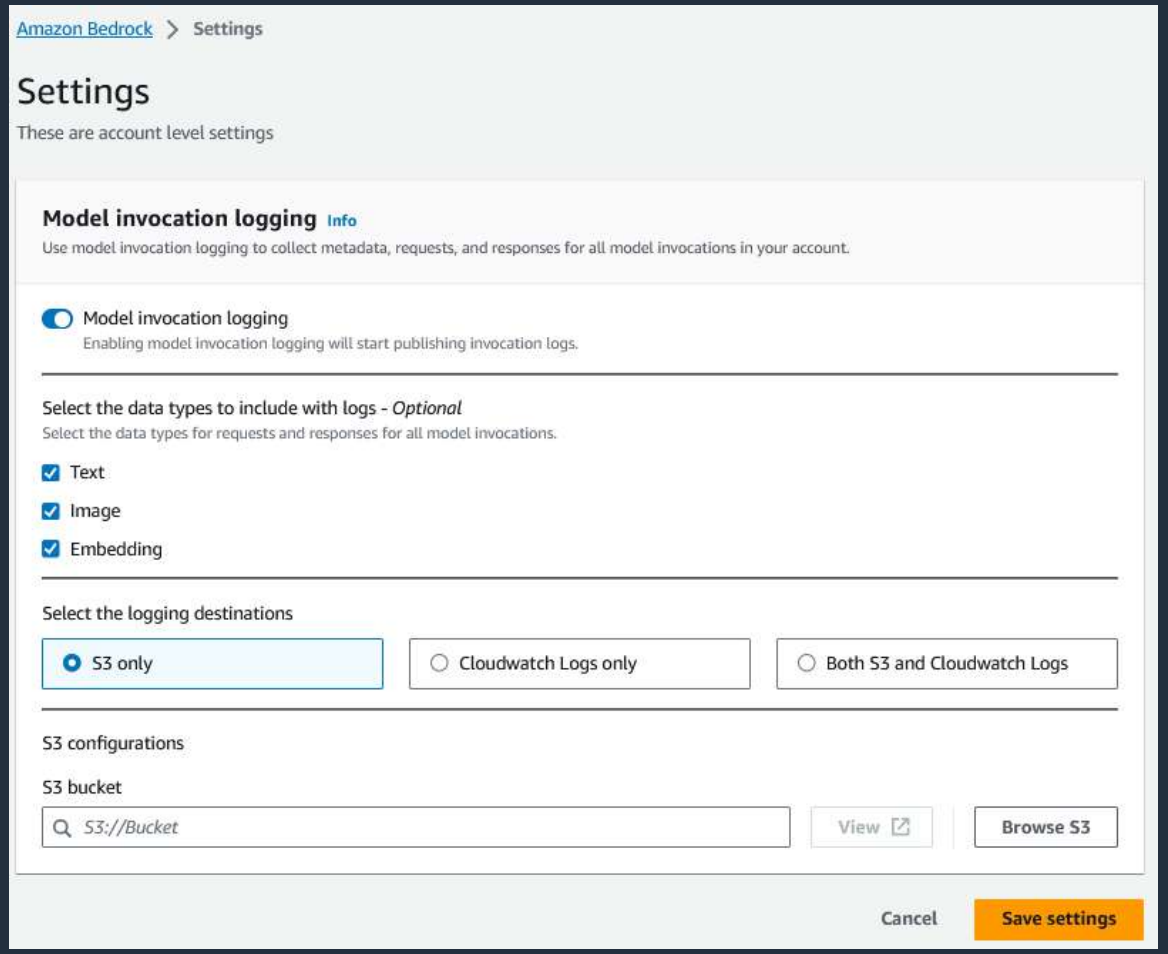
© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Model invocation logging

PREVIEW

Use model invocation logging to collect metadata, requests, and responses for all model invocations in your account.

Supports both S3 and CloudWatch logs



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

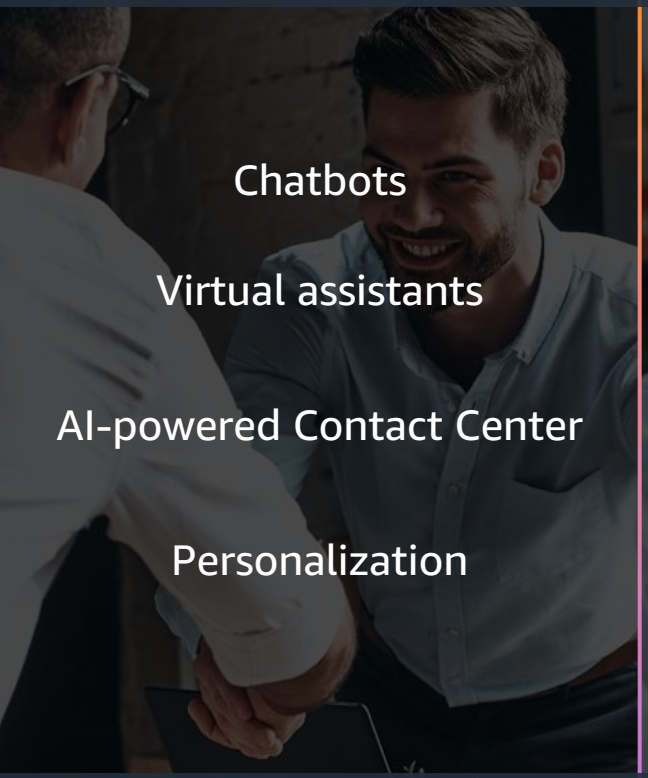
Bedrock Demo

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Applications built on top of Bedrock

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Generative AI can be used for a wide range of use cases



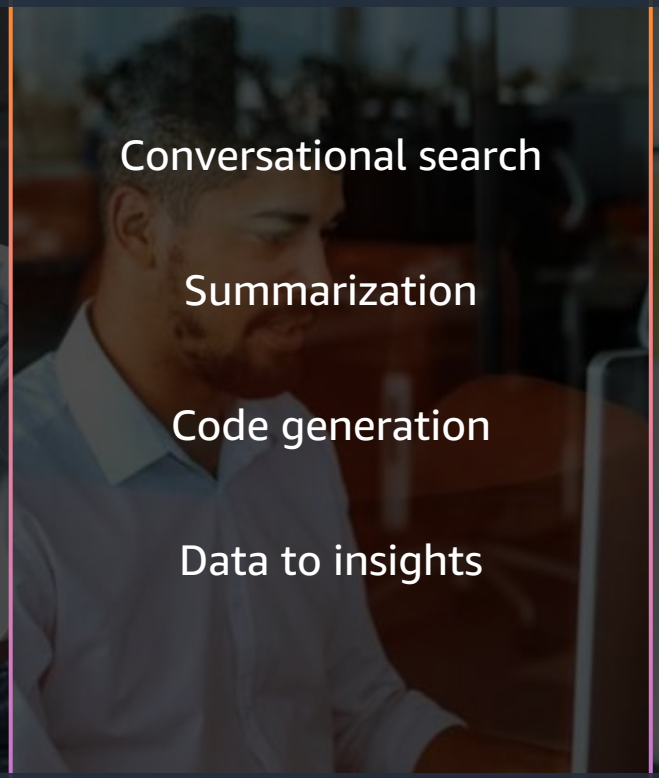
Chatbots

Virtual assistants

AI-powered Contact Center

Personalization

Enhance customer experience



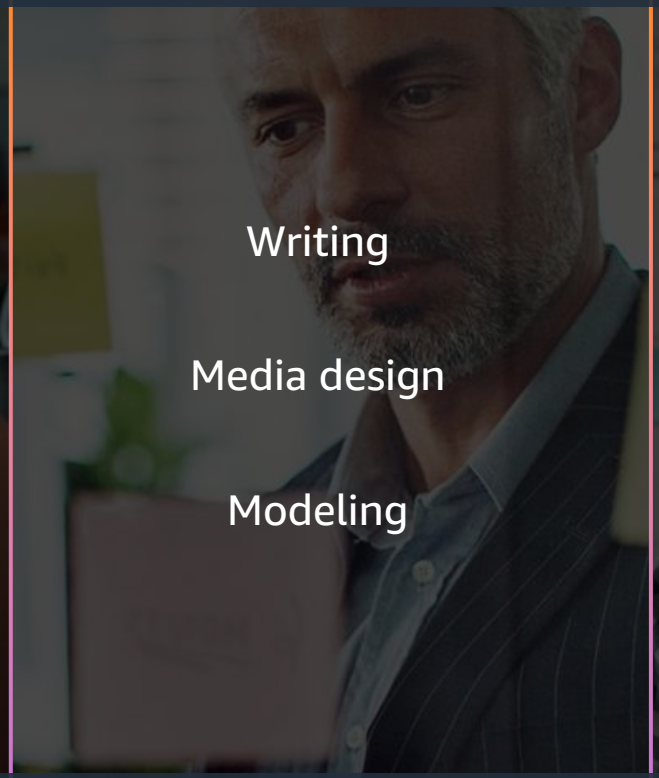
Conversational search

Summarization

Code generation

Data to insights

Boost employee productivity

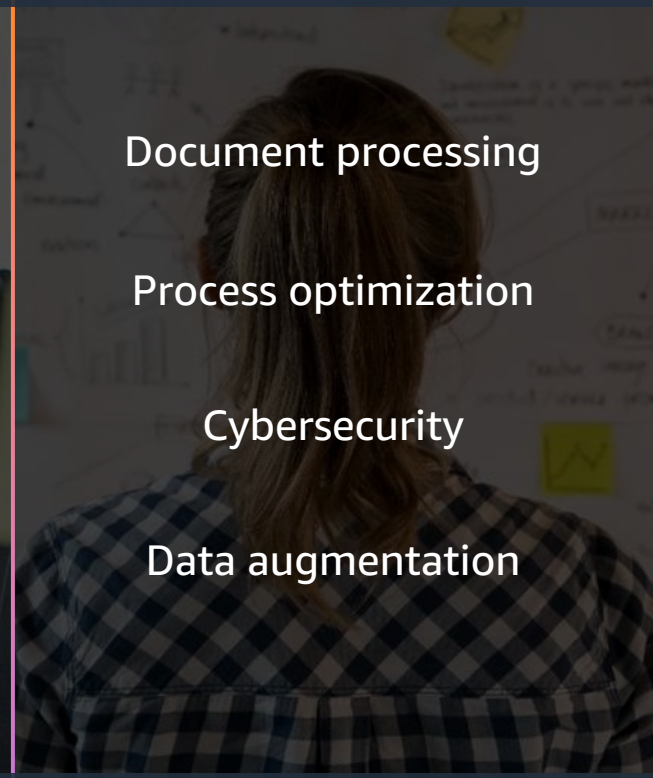


Writing

Media design

Modeling

Creativity and content creation



Document processing

Process optimization

Cybersecurity

Data augmentation

Improve business operations

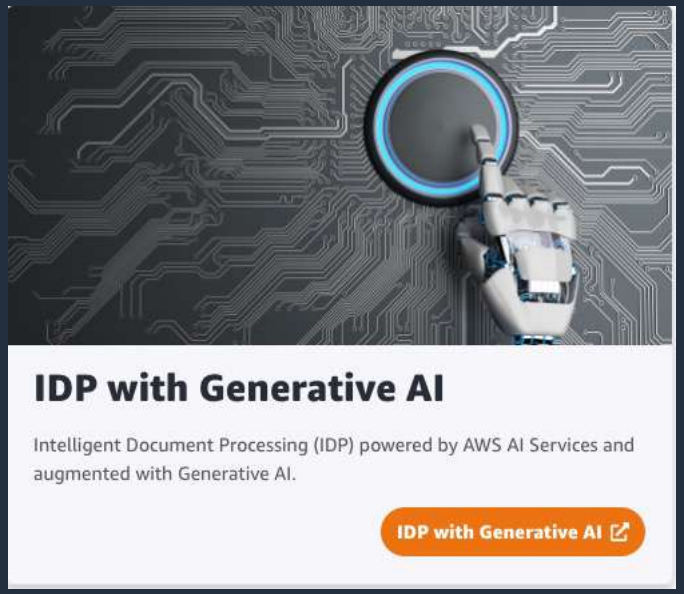
© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Demo: Intelligent Document Processing (IDP)

Improve employee productivity and make faster decisions with intelligent document processing

Organizations in various sectors, including healthcare, finance, legal, retail, and manufacturing, often handle a substantial volume of documents critical to their daily operations.

IDP with AWS AI services empowers organizations to automate information extraction from diverse document types and formats swiftly and accurately, eliminating the need for machine learning expertise, thereby enabling timely, cost-effective, and high-quality decision-making.

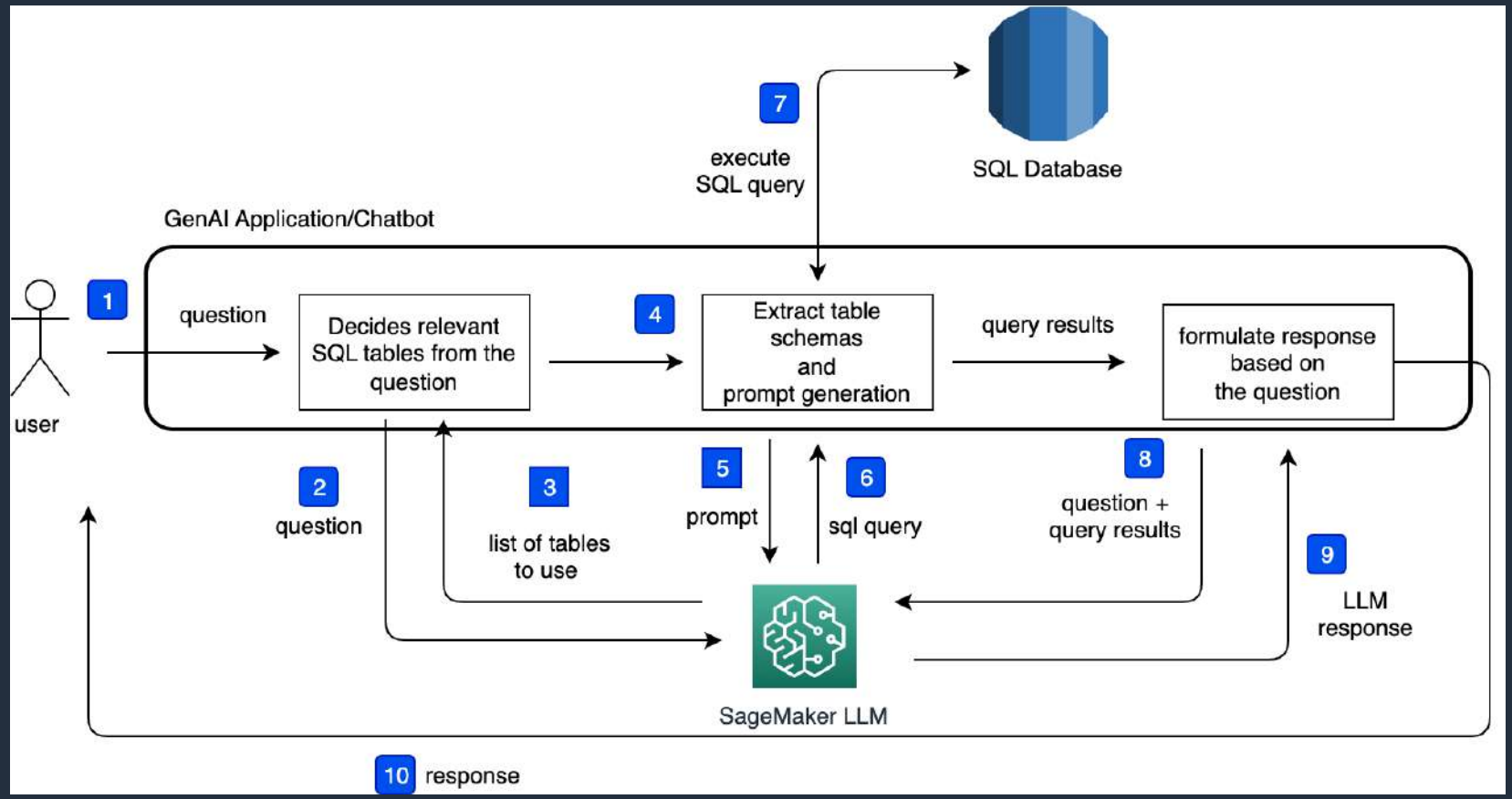


© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Demo: Gather business insights with Generative AI

Empowers business users to gain insights into historical, current and future states 24/7 without writing any code or SQL query.

- Use natural language to interact with enterprise data
- Customize the interface – text or speech
- Reduce dependency on database and programming experts
- Powered by an agent and a pipeline using LangChain
- Chatbot has memory buffer
- Cross-industry application



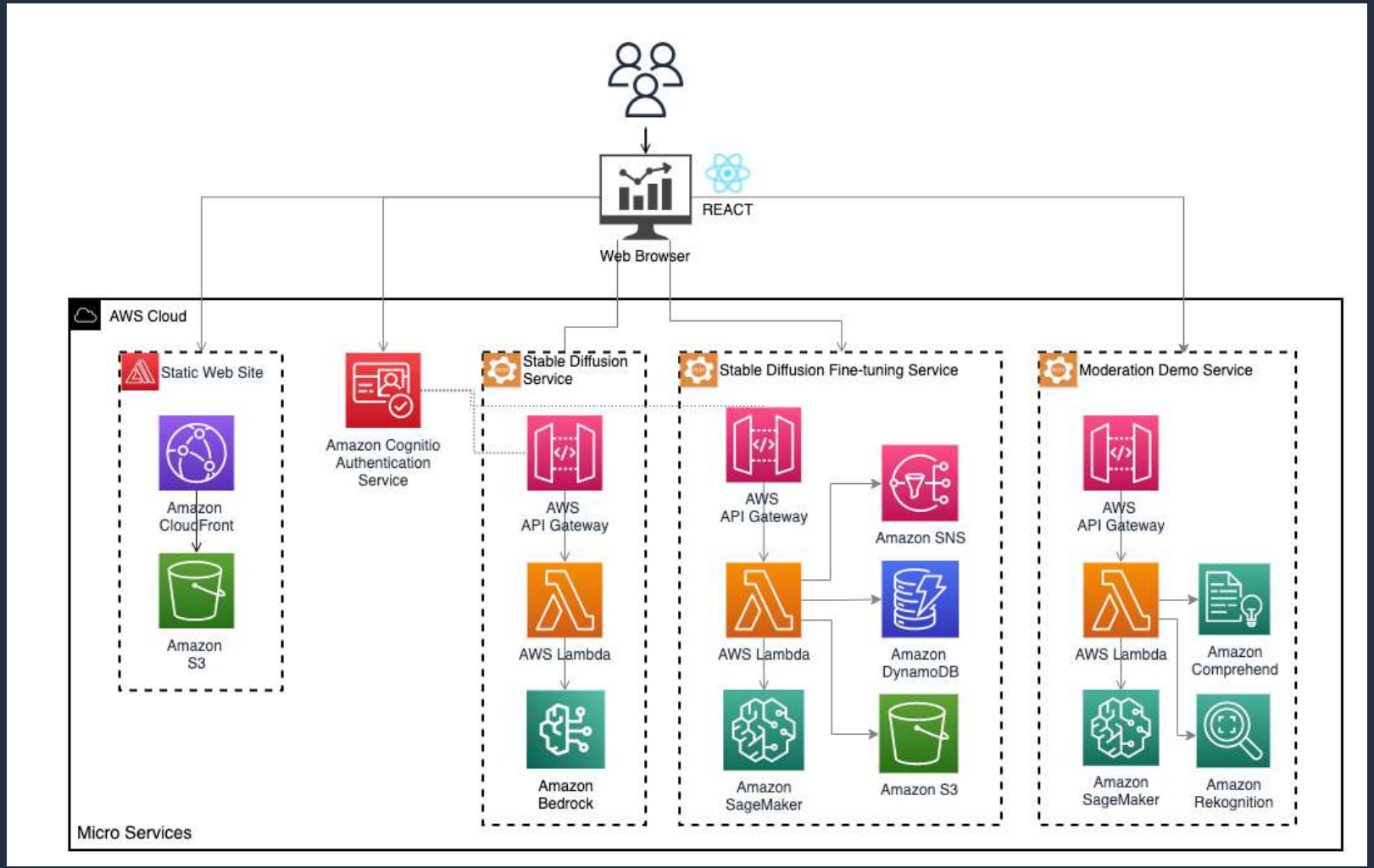
© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Demo: Avatar generation and diffusion moderation

Discover Stable Diffusion features & its implementation on AWS with Amazon Bedrock & Amazon SageMaker

Features
- Text-to-image Generation
- Image-to-image Generation
- Masking / inpainting
- Fine-tuning
- Stable diffusion moderation

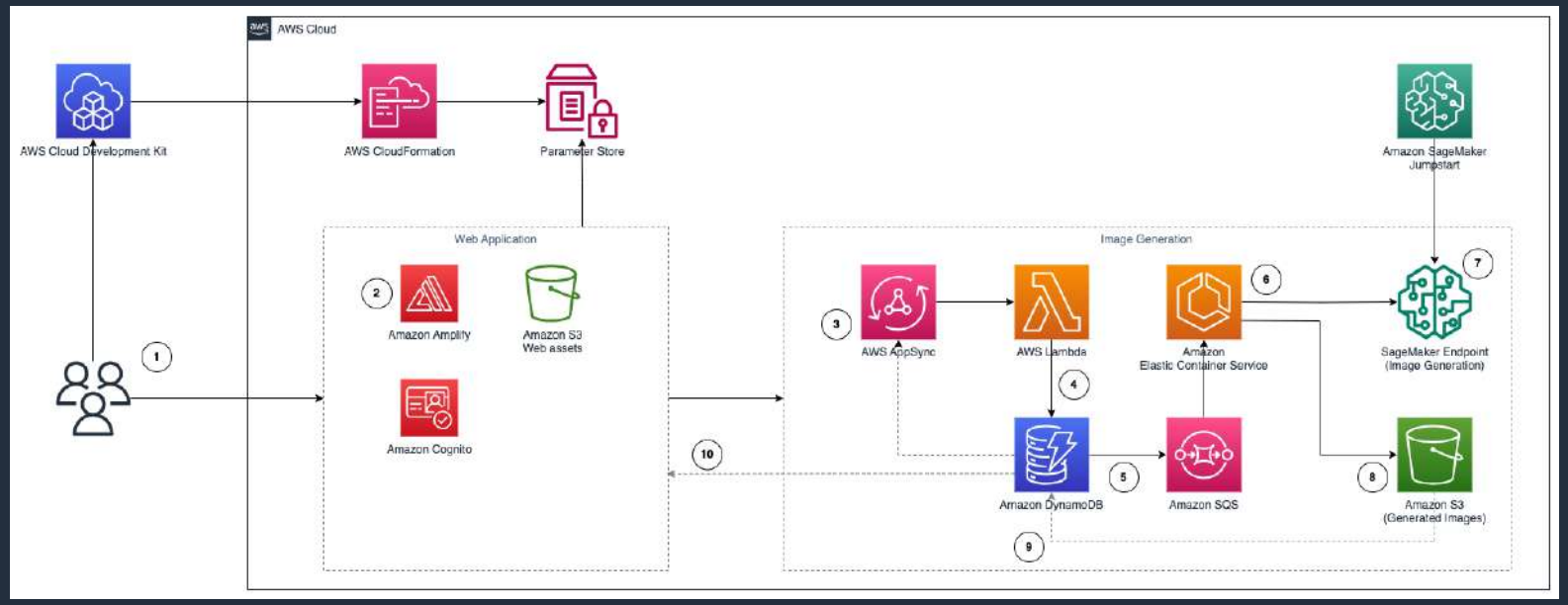
Use Cases
- Write marketing copy
- Ad campaign generation
- Content personalization
- Social media post generation



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Demo: Generate Images for Advertising

Create high quality images for Ad campaigns using first party data and Generative AI

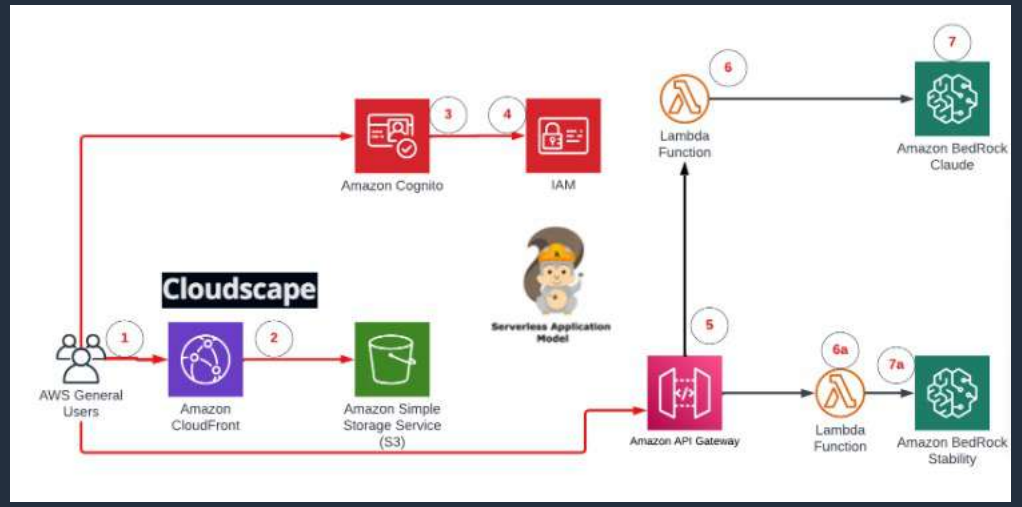


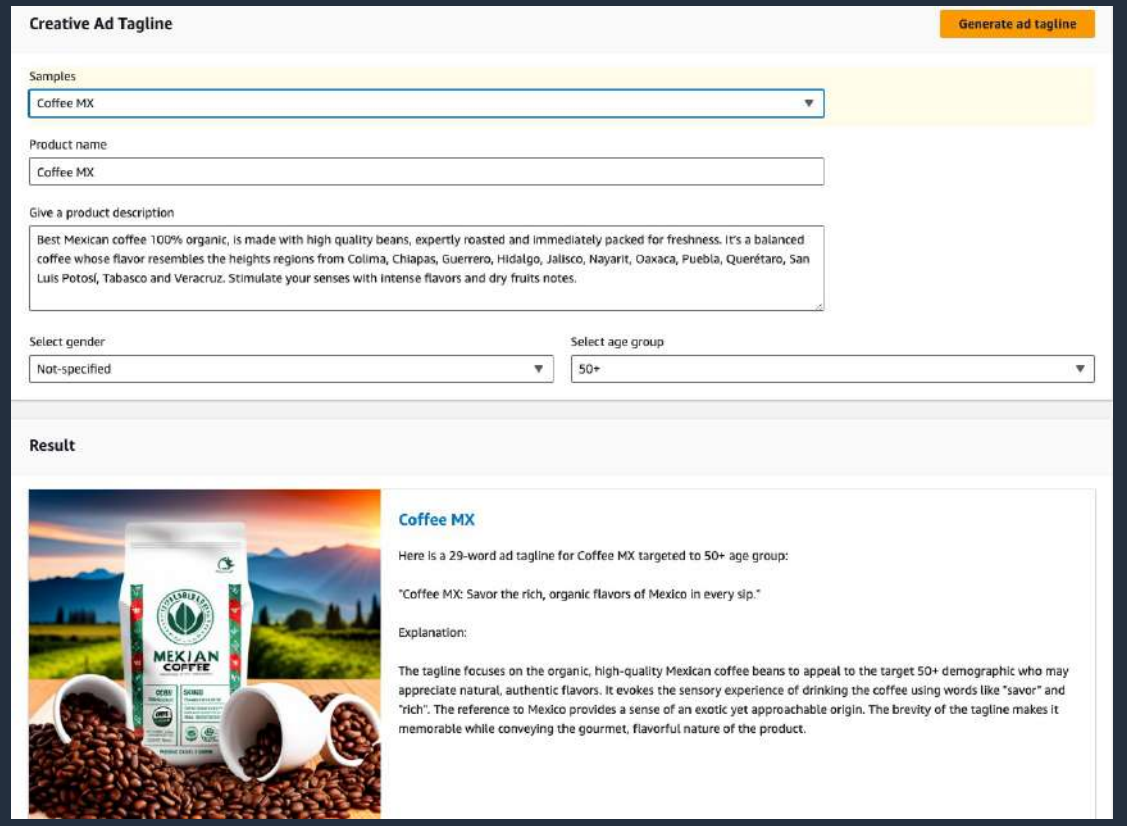
Metallic green colored car, sunset time, color photo, outdoors in pleasant landscape on a mountain road.

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Demo: Transform Advertising Text with Generative AI

Let your Imagination be your Best Advertising Tool





Best Mexican coffee 100% organic, is made with high quality beans, expertly roasted and immediately packed for freshness. It's a balanced coffee whose flavor resembles the heights regions from Colima, Chiapas, Guerrero

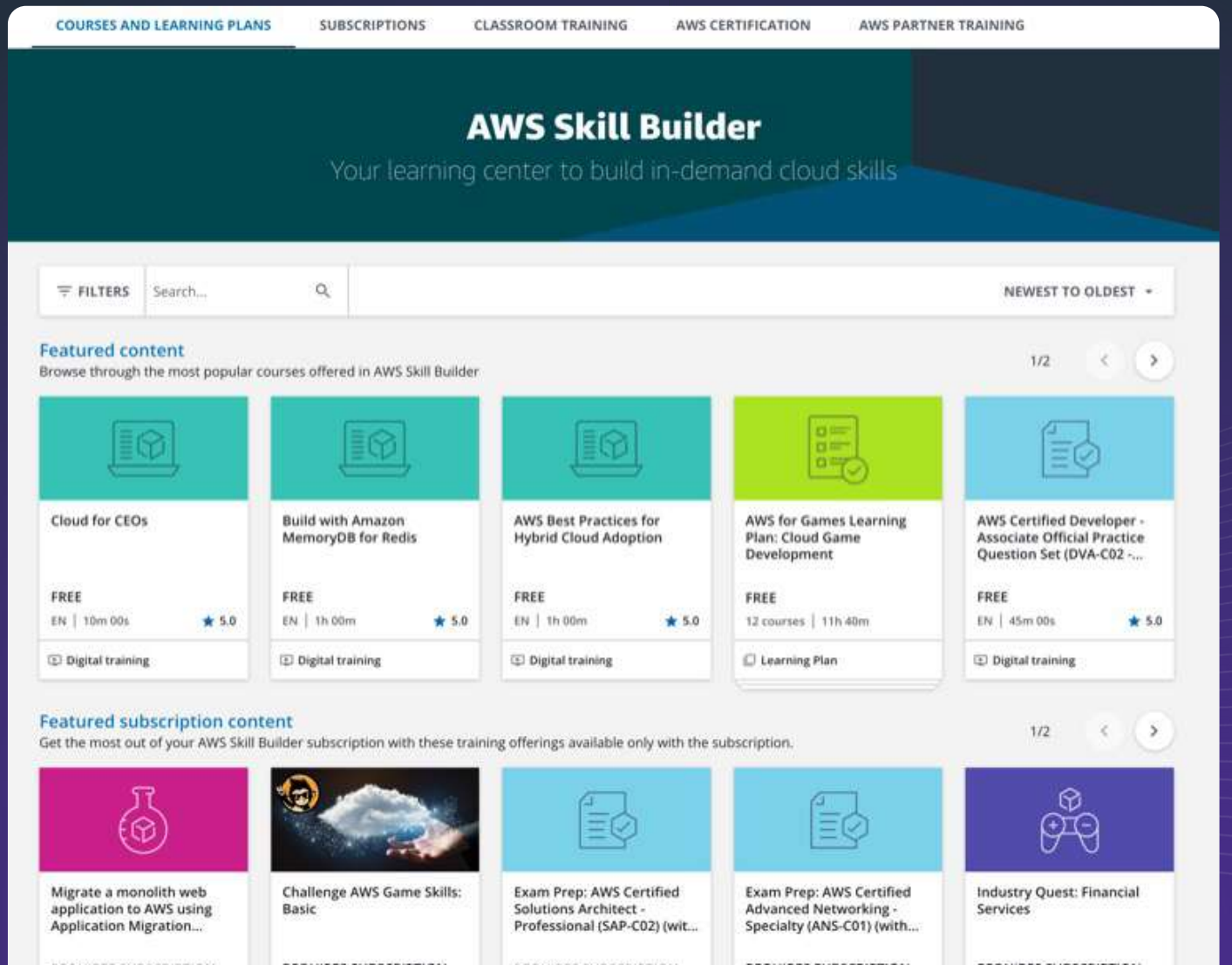
© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Other resources

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

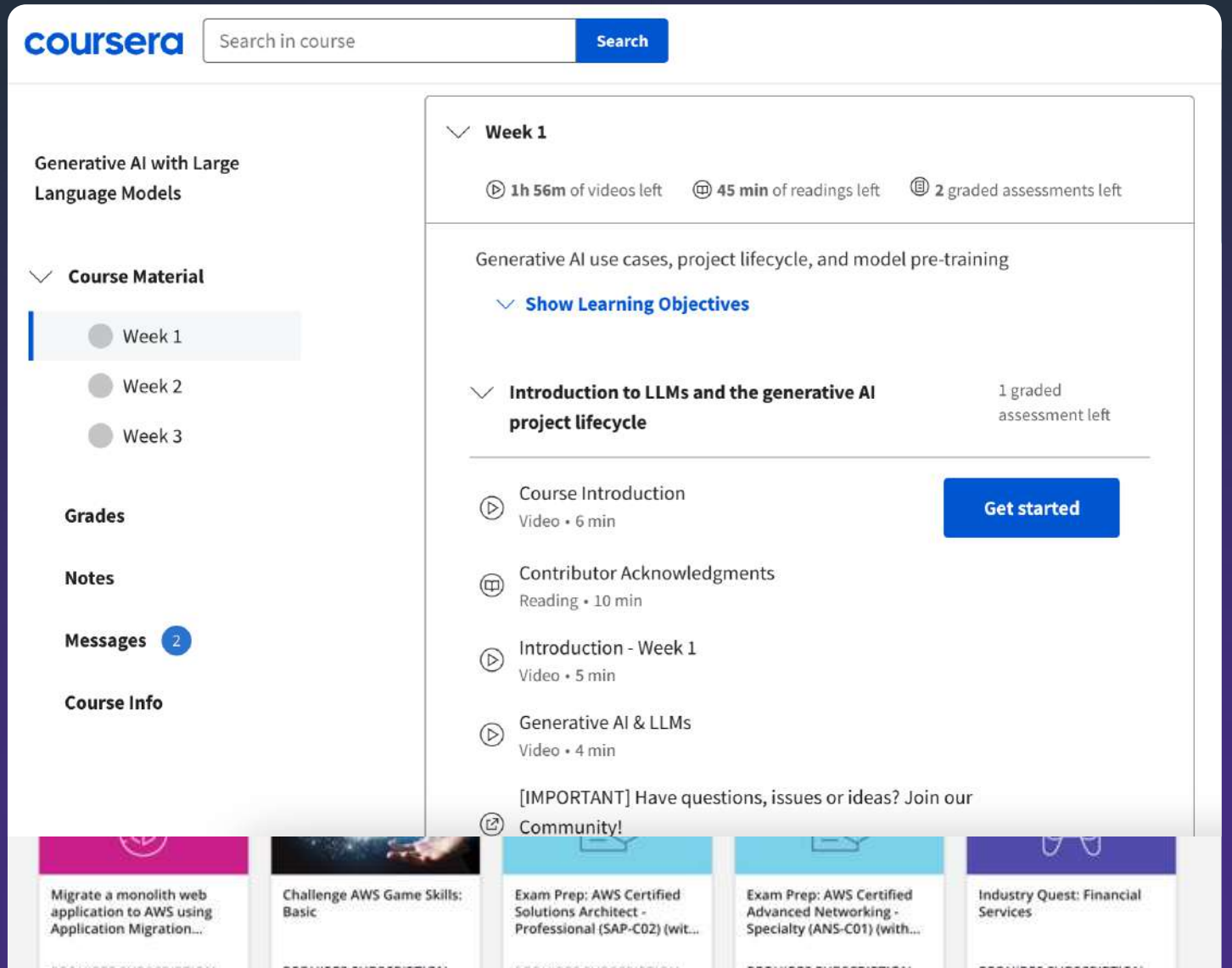
SKILLBUILDER.AWS

600+ digital courses



Hands-on generative AI training opportunities

Learn the fundamentals of generative AI for real-world applications



Generative AI for Business Professional (3+ hours)

This specialized plan offers a strategic overview of generative models, enabling you to grasp the pivotal role they play in driving innovation across industries. From understanding the technology's impact on market dynamics to identifying opportunities for differentiation, this learning experience empowers participants to make informed decisions that harness the power of generative AI for business's success.

Link to enroll

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Generative AI for Technical Professionals (9+ hours)

Dive into the realm of AWS Generative AI with our specialized learning plan crafted for technical professionals, including professional services, solutions architects, data analysts and engineers, and Technical Pre-Sales Architects. This comprehensive plan offers a deep exploration of generative models, equipping participants with the skills to architect innovative solutions that push the boundaries of what's possible. From honing your understanding of foundations models to orchestrating seamless integrations of generative AI into complex architectures, participants will gain practical insights and hands-on experience.

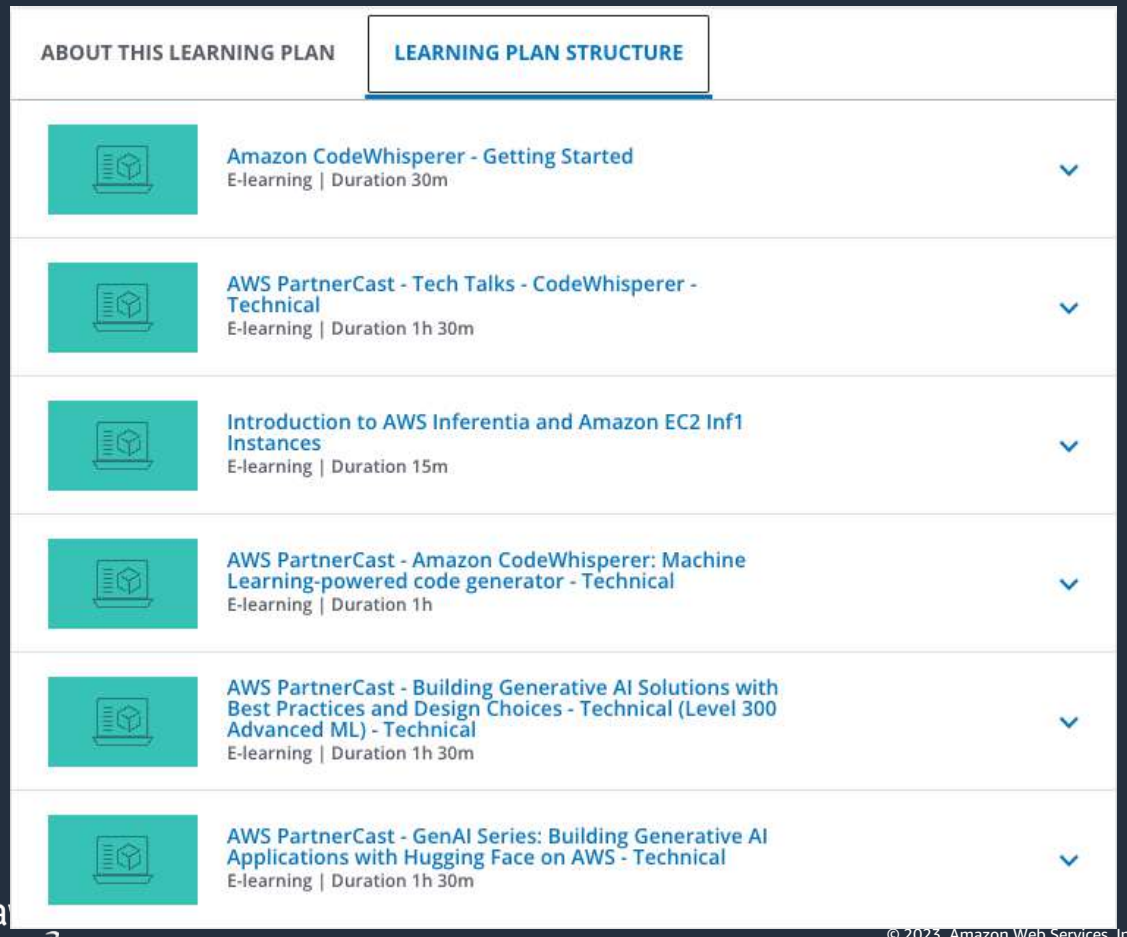
- AWS Partner: Generative AI on AWS Essentials (Business)
  E-learning | Duration 1h 5m
- AWS PartnerCast - Building Generative AI on AWS: Key Service Features and Demos (Technical Focused)
  E-learning | Duration 1h 27m
- AWS PartnerCast - Breaking the Limits with AWS Generative AI - Technical
  E-learning | Duration 1h 2m
- AWS PartnerCast - GenAI Series: Building Generative AI Applications with AI21 on AWS - Technical
  E-learning | Duration 1h
- Introduction to AWS Inferentia and Amazon EC2 Inf1 Instances
  E-learning | Duration 15m
- AWS PartnerCast - Building Generative AI Solutions with Best Practices and Design Choices - Technical (Level 300 Advanced ML) - Technical
  E-learning | Duration 1h 30m
- AWS PartnerCast - GenAI Series: Building Generative AI Applications with Hugging Face on AWS - Technical
  E-learning | Duration 1h 30m
- Introduction to Amazon SageMaker
  E-learning | Duration 13m
- AWS PartnerCast - GenAI Series: Distributed Training on AWS SageMaker - Technical
  E-learning | Duration 1h 30m

Link to enroll

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Generative AI for Developers (6+ hours)

Explore the cutting-edge realm of AWS Generative AI through our comprehensive learning plan tailored for developers. From understanding the fundamentals of generative models to mastering the intricacies of AWS tools such as Codewhisperer, this learning plan offers hands-on experiences, real-world use cases, and expert guidance.

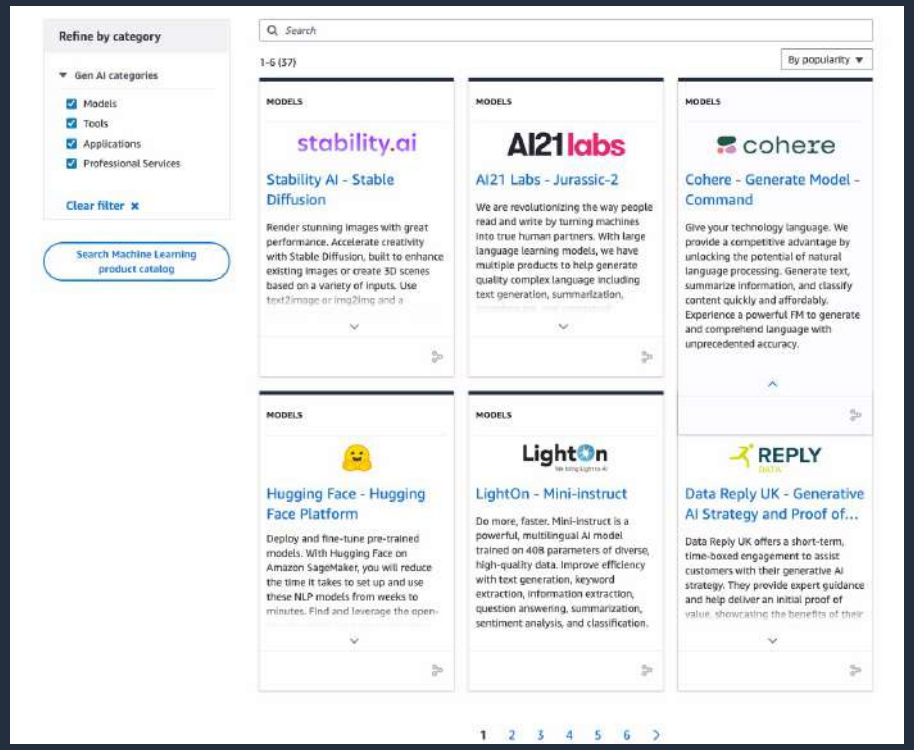


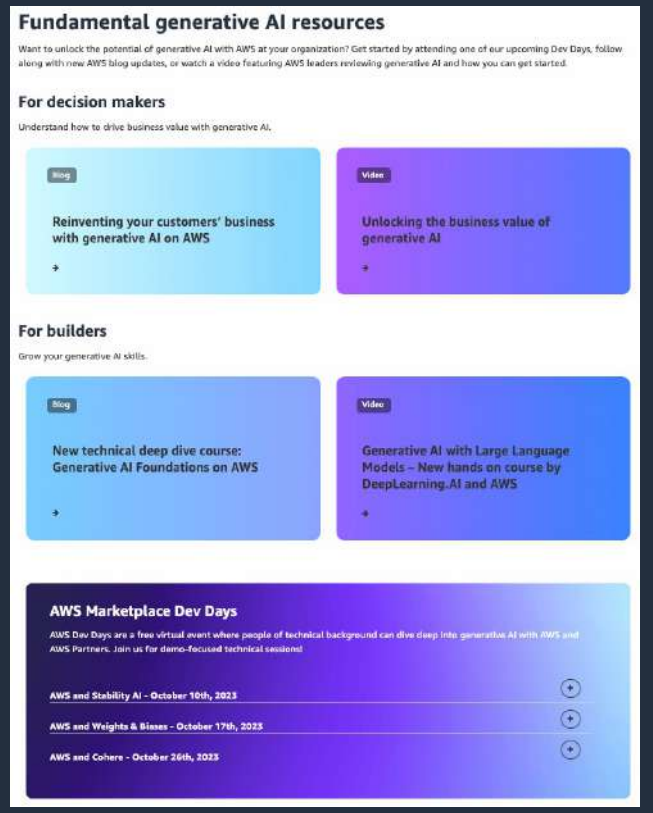
Link to enroll

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Marketplace

AWS Marketplace helps organizations discover, procure, experiment with, provision, and govern third-party software, data, and services. From FMs to tooling, applications to consultants, find a generative AI solution provider to kick-start your transformation journey.





https://aws.amazon.com/marketplace/solutions/machine-learning/generative-ai

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon Bedrock Workshop

Workshop studio

Code in Github

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Q&A

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Thank you!

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Amazon Web Services, AWS, the Powered by AWS logo, and all AWS service names used in this slide deck are trademarks of Amazon.com, Inc. or its affiliates.