



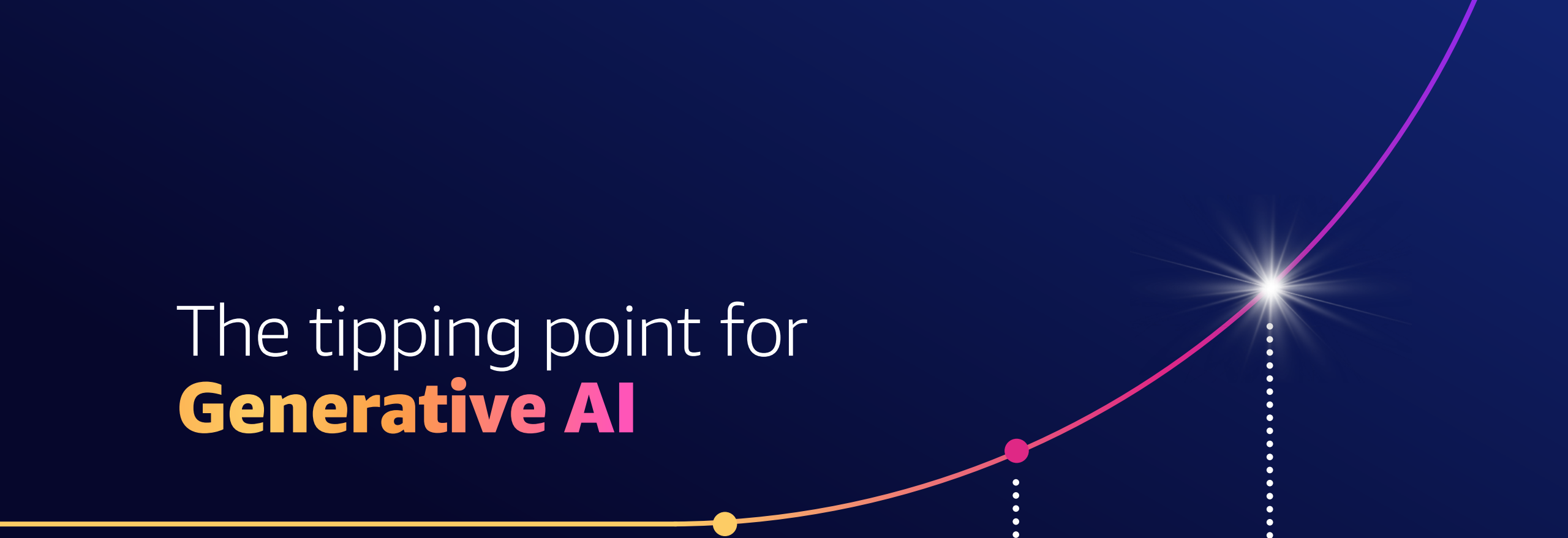
Generative AI on AWS

Marios Parthenios
Sr. Solutions Architect
Amazon Web Services

Question: What is generative artificial intelligence (AI)?

- Creates new content and ideas, including conversations, stories, images, videos, and music
- Powered by large models that are pretrained on vast corpuses of data and commonly referred to as foundation models (FMs)

The tipping point for **Generative AI**



MASSIVE PROLIFERATION
OF DATA

AVAILABILITY OF
SCALABLE COMPUTE
CAPACITY

MACHINE LEARNING
INNOVATION

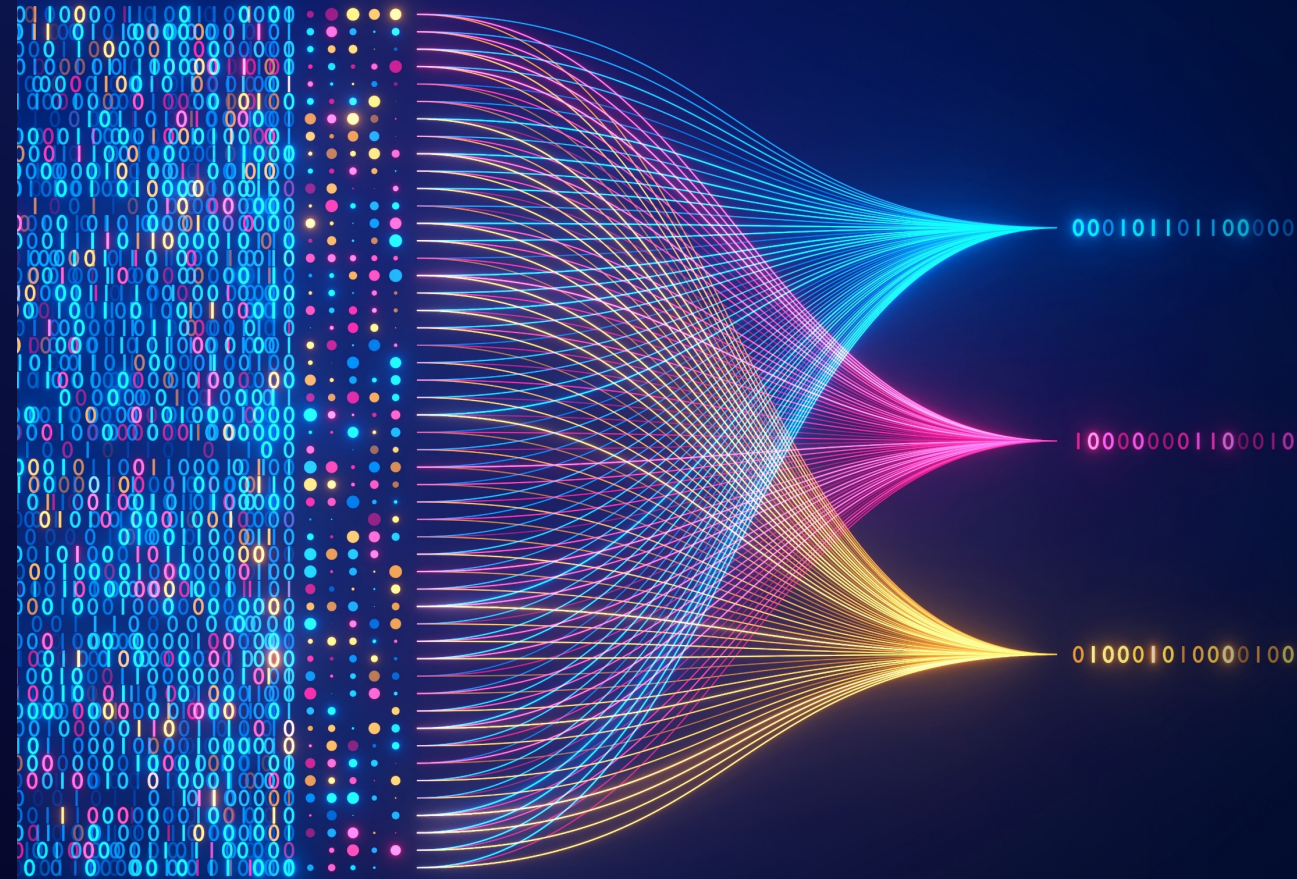
Generative AI is powered by foundation models

Pretrained on vast amounts of unstructured data

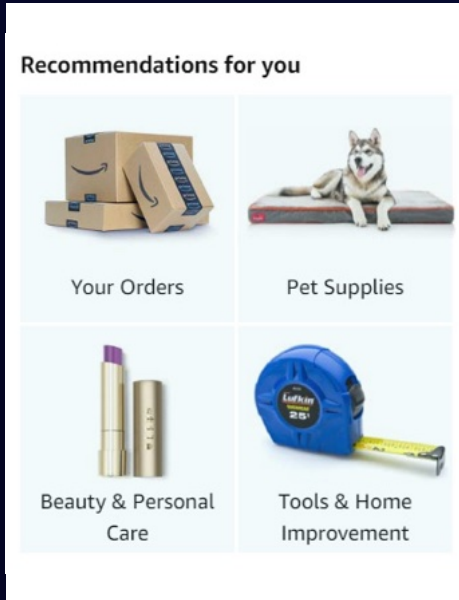
Contain large number of parameters that make them capable of learning complex concepts

Can be applied in a wide range of contexts

Customize FMs using your data for domain specific tasks



ML innovation is in Amazon's DNA



**4,000 products
per minute** sold
on Amazon.com



1.6M packages
every day

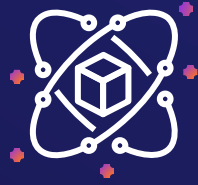


Billions of Alexa
interactions each week



Just Walk Out
technology in airports,
stadiums and more

Everything you
need to
accelerate
your generative
AI journey



Easiest way to build

with leading foundation models



Differentiate with your data

in a secure and private environment



Increase productivity

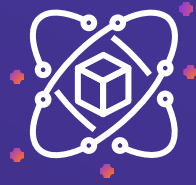
with generative AI applications and services



Most performant, low cost

infrastructure to scale generative AI

Everything you
need to
accelerate
your generative
AI journey



Easiest way to build

with leading foundation models



Differentiate with your data

in a secure and private environment



Increase productivity

with generative AI applications and services



Most performant, low cost

infrastructure to scale generative AI

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

Amazon Bedrock

Choice of foundation models

AI21labs

ANTHROPIC

co:here

∞ Meta AI

stability.ai

amazon

JURASSIC-2

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

CLAUDE 2

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

COMMAND

Text generation model for business applications like summarization, copywriting, dialog, extraction, and question answering

LLAMA 2

Pre-trained and fine-tuned LLMs for natural language tasks like question answering and reading comprehension

SDXL 1.0

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

AMAZON TITAN

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



Foundation models alone
cannot execute tasks





Agents for **Amazon Bedrock**

Enable generative AI applications to
complete tasks in just a few clicks

IN PREVIEW



Breaks down and
orchestrates tasks



Securely accesses and
retrieves company data



Takes action by executing
API calls on your behalf



Provides fully managed
infrastructure support

Driving innovation with Amazon Bedrock

Chegg

lonely planet

cimpress

PHILIPS

IBM | The Weather Company

nexxiot

Sun Life

Neiman Marcus

RYANAIR

hellmann
WORLDWIDE LOGISTICS

WPS Office
Make It Simple

twilio

BRIDGEWATER

Showpad

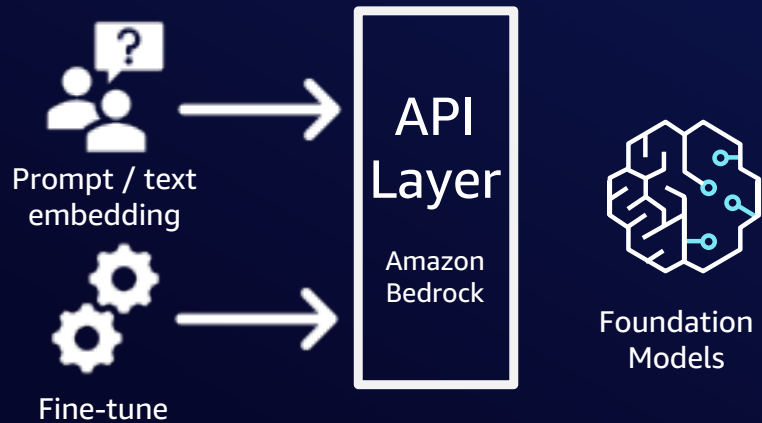
coda

Booking.com

Amazon SageMaker & SageMaker JumpStart

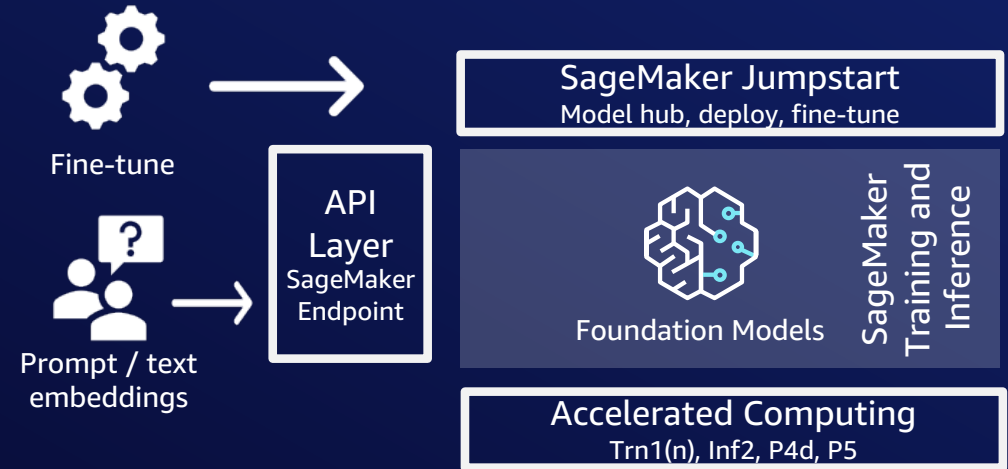


How do I access foundation models?



Amazon Bedrock

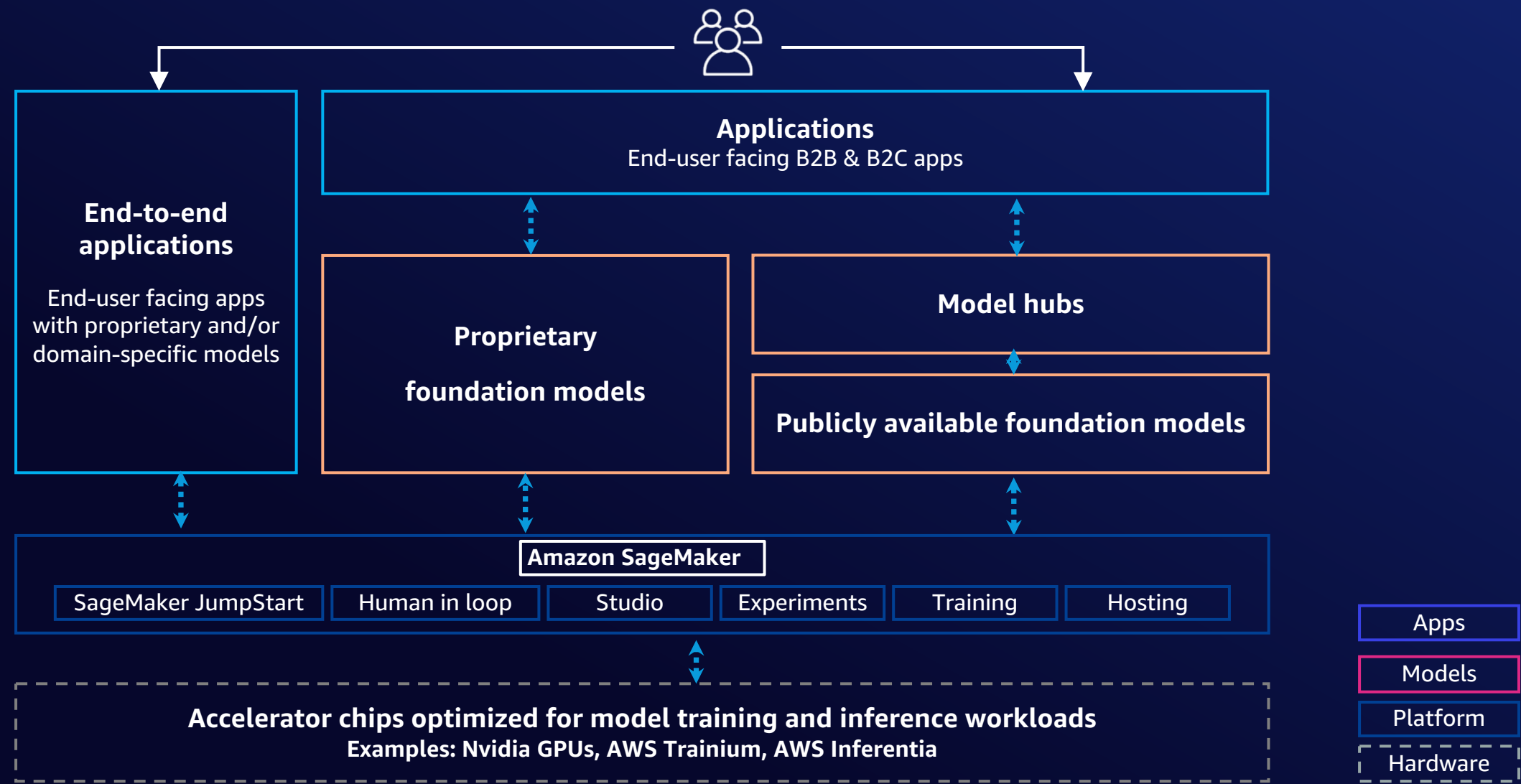
- The easiest way to build and scale generative AI applications with FMs
- Access directly or fine-tune foundation model using API
- Serverless



Amazon SageMaker JumpStart

- ML hub with FMs, built-in algorithms, and prebuilt ML solutions that you can deploy with just a few clicks
- Deploy FM as SageMaker endpoint (hosting)
- Fine-tuning leverages SageMaker training jobs
- Choose SageMaker managed accelerated computing instance

Generative AI: Lay of the land using SageMaker



Build with publicly available foundation models

AVAILABLE ON SAGEMAKER JUMPSTART

AI21labs

Models
Jurassic-2 Ultra, Mid
Contextual answers
Summarize
Paraphrase
Grammatical error
correction

Tasks
Text generation
Long-form
generation
Summarization
Paraphrasing
Chat
Information
extraction

Meta AI

Models
Llama 2 7B, 13B, 70B

Tasks
Question answering
Chat
Summarization
Paraphrasing
Sentiment analysis
Text generation

cohere

Models
Cohere
Command XL

Tasks
Text generation
Information
extraction
Question answering
Summarization

Hugging Face

Models
Falcon-7B, 40B
Open LLaMA
RedPajama
MPT-7B
BloomZ 176B
Flan T-5 models (8 variants)
DistilGPT2
GPT NeoXT
Bloom models
(3 variants)

Tasks
Machine translation
Question answering
Summarization

stability.ai

Models
Stable Diffusion XL 1.0
2.1 base
Upscaling
Inpainting

Tasks
Generate photo-realistic
images from text input
Improve quality of
generated images

Features
Fine-tuning on Stable
Diffusion 2.1 base
model

LightOn

Models
Lyra-Fr
10B, Mini

Tasks
Text generation
Keyword extraction
Information extraction
Question answering
Summarization
Sentiment analysis
Classification

databricks

Models
Dolly

Tasks
Question answering
Chat
Summarization
Paraphrasing
Sentiment analysis
Text generation

alexa

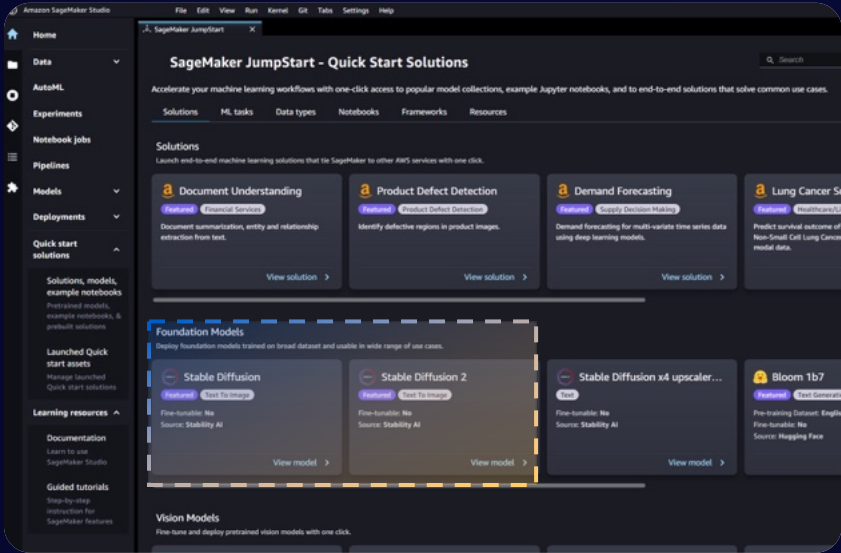
Models
AlexaTM 20B

Tasks
Machine translation
Question answering
Summarization
Annotation
Data generation

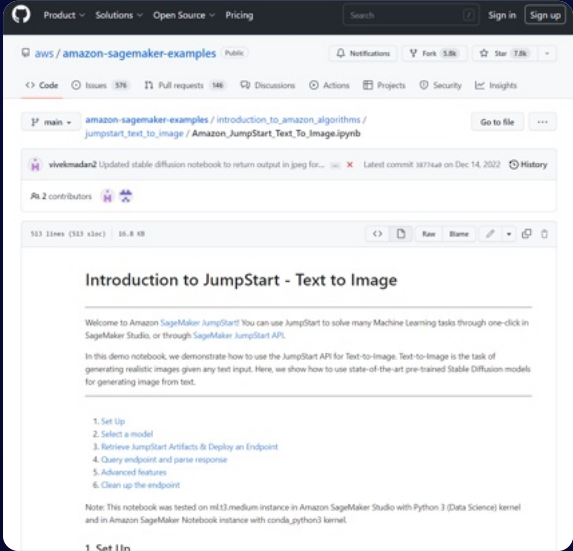


3 ways to use foundation models with SageMaker JumpStart

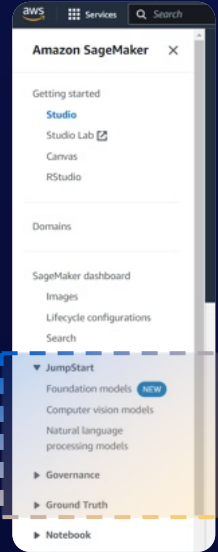
SageMaker Studio One-step deploy



SageMaker Notebooks



AWS Management Console Preview



Try-out experience

cohere

Cohere Generate Model - Medium

By Cohere

Try a product demo of the capabilities of this model from Cohere. Do not upload any confidential or sensitive information. Use of this feature is for demonstration purposes only. This demo may not accurately represent the actual response times of the product.

Prompt

Context:
The United Nations is an intergovernmental organization founded in 1945 with the mission of maintaining international peace and security, promoting human rights, and fostering social and economic development. It is composed of 193 member states and has its headquarters in New York City.
Question:
What is the mission of the United Nations?
Answer:

Generate text

Output

The mission of the United Nations is to maintain international peace and security, promote human rights, and foster social and economic development.

Copy output

General info

Temperature

0.9

Number of tokens

100

Top k

0

Top p

0.7

Presence Penalty

0

Frequency Penalty

0

- Try out the models and model prompts without running code or incurring costs
- Available for proprietary models in Top 10 in HELM benchmarks and public models for comparison purposes

Simple deploy experience

MODEL

Stable Diffusion 2.1 base

text · text to image · foundation models · featured

Open notebook Browse JumpStart

Deploy Train Notebook Model details

Deploy Model

Deploy a pretrained model to an endpoint for inference. Deploying on SageMaker hosts the model on the specified compute instance and creates an internal API endpoint. JumpStart will provide you an example notebook to access the model after it is deployed. [Learn more.](#)

- > Deployment Configuration
- > Security Settings

Deploy

- Training instance type
- Security settings



Easy fine-tune experience

Stable Diffusion 2.1 base

text · text to image · foundation models · featured

[Open notebook](#)[Browse JumpStart](#)

[Deploy](#) [Train](#) [Notebook](#) [Model details](#)

Create a training job to fit this model to your own data.

This model is pretrained, you will fine-tune its parameters instead of starting from scratch. Fine-tuning can produce accurate models with smaller datasets and less training time. [Learn more](#).

▼ Data Source

Select the default dataset, or use your own data to fine-tune this model.

Training data set ⓘ

s3://jumpstart-cache-prod-us-east-1/training-datasets/cats_sd_finetuning/

[Browse](#)

> Deployment Configuration

> Hyper-parameters

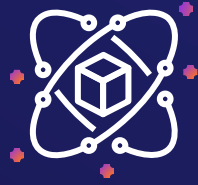
> Security Settings

[Train](#)

- Labeled data set path
- Training instance type
- Hyper-parameters & security settings



Everything you
need to
accelerate
your generative
AI journey



Easiest way to build

with leading foundation models



Differentiate with your data

in a secure and private environment



Increase productivity

with generative AI applications and services



Most performant, low cost

infrastructure to scale generative AI

Your data is
your differentiator



Privately customize foundation models using your organization's data



Fine-tune

PURPOSE

Maximizing accuracy for specific tasks

DATA NEED

Small number of labeled examples

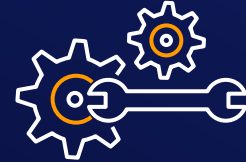
Keeping your data private and secure



None of the customer's data is used to train the underlying model

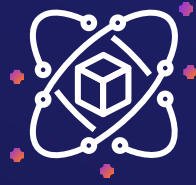


All data is encrypted at rest and PrivateLink support allows access to Bedrock APIs via customer's VPC endpoints



Customized foundation models and the customer-specific data that trains them remain private

Everything you
need to
accelerate
your generative
AI journey



Easiest way to build

with leading foundation models



Differentiate with your data

in a secure and private environment



Increase productivity

with generative AI applications and services



Most performant, low cost

infrastructure to scale generative AI

GENERALLY AVAILABLE

Amazon CodeWhisperer

Build apps faster and more
securely with an AI coding
companion



Generate code suggestions
in real-time



Scan code for hard-to-find
vulnerabilities



Flag code that resembles open-
source training data or filter by
default

FREE FOR INDIVIDUAL TIER

CodeWhisperer: ML-powered coding companion

Provides code recommendations based on contextual information like prior code and comments

GENERATES

- Entirely new code based on context
- Code from plain English comments
- Complete functions

Available in all major integrated development environments (IDEs) as an extension

```
# Write a function to upload a file to S3.
def upload_file_to_s3(file_name, bucket_name, object_name):
    """
    Uploads a file to an S3 bucket

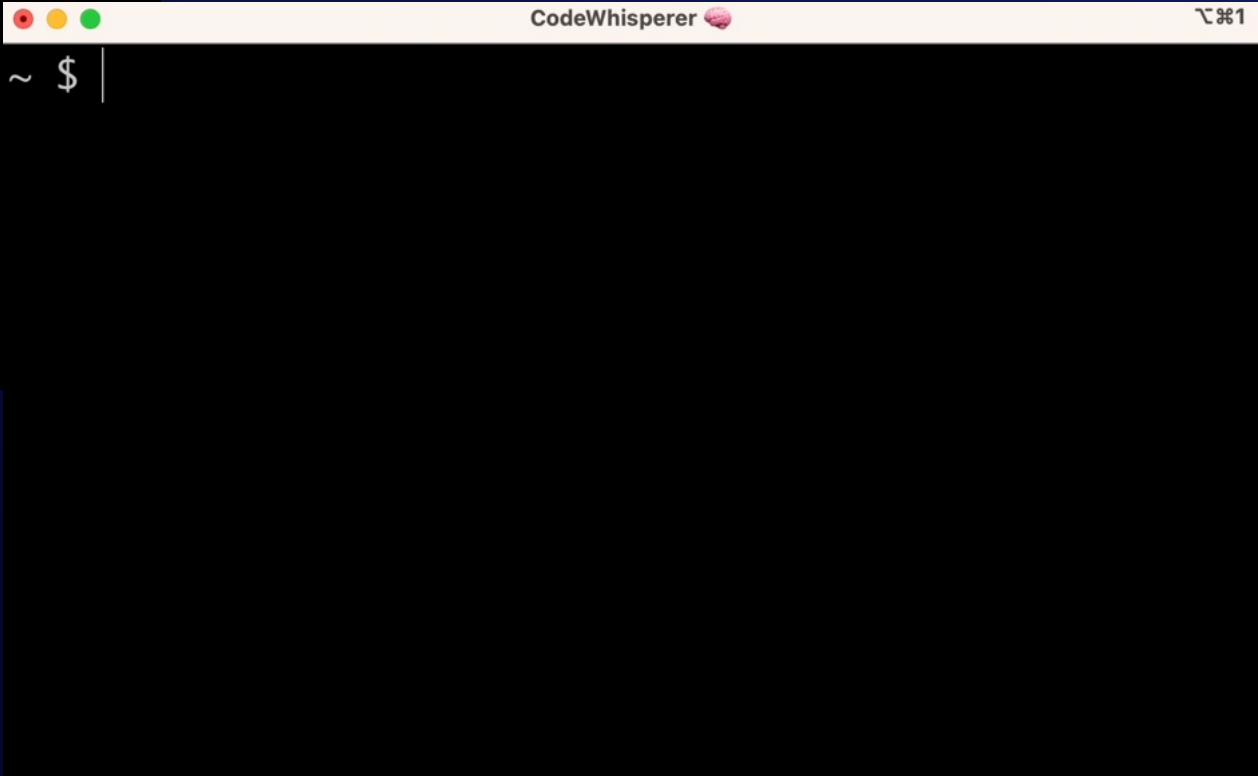
    :param file_name: File to upload
    :param bucket_name: Bucket to upload to
    :param object_name: S3 object name. If none then file_name is used
    :return: True if file was uploaded, else False
    """

    # Upload the file
    s3_client = boto3.client('s3',
                             aws_access_key_id=AWS_ACCESS_KEY_ID,
                             aws_secret_access_key=AWS_SECRET_ACCESS_KEY,
                             region_name=AWS_REGION_NAME)

    try:
        s3_client.upload_file(file_name, bucket_name, object_name)
        print(f'File {file_name} uploaded to S3 bucket {bucket_name} as {object_name}')
        return True
    except FileNotFoundError:
        print(f'File {file_name} not found')
```



CodeWhisperer for command line



CODEWHISPERER

Using
CodeWhisperer
to increase
productivity

accenture

KOCH

Infosys

HCLTech

SmugMug

publicis
sapient

amazon ads



COMING SOON

Amazon CodeWhisperer customization capability

Generate code recommendations based on
your internal codebases



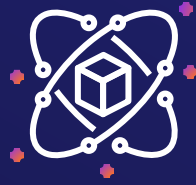
Generates organization-specific code
recommendations based on their
internal codebase



Will be available to customers as
part of a new CodeWhisperer
Enterprise Tier



Everything you
need to
accelerate
your generative
AI journey



Easiest way to build

with leading foundation models



Differentiate with your data

in a secure and private environment



Increase productivity

with generative AI applications and services



Most performant, low cost

infrastructure to scale generative AI

Deep investments **in global infrastructure**



Broad choice of ML
accelerators

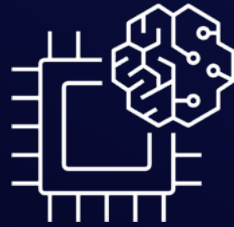


High performance,
low-cost ML infrastructure



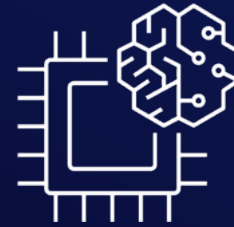
10+ years of silicon
innovation

Purpose-built accelerators for generative AI



AWS Trainium

Up to 50% savings on training costs
over comparable Amazon EC2 instances



AWS Inferentia2

Up to 40% better price performance
than comparable Amazon EC2 instances

Demo



AWS Skill Builder equips your teams with practical skills that can be applied immediately



80+ digital courses and learning resources on AI and ML, including generative AI, designed by AWS experts.



Practice with game-based learning and interactive challenges in a secure sandbox environment.



Validate expertise with an AWS Certification:
AWS Certified Machine Learning – Specialty



Generative AI training for executives

Learn how generative AI
can address your
business challenges,
drive growth, and
revolutionize industries.



**AWS is here to
help you get the
generative AI
skills you need
to transform
your business.**

aws.com/training

AWS Learning Needs Analysis:

Build a data-driven plan to accelerate learning



Learn more about AWS Skill Builder:





Thank you!

maparthe@amazon.com

<https://www.linkedin.com/in/marios-parthenios/>