AIM201

Demystifying Generative AI on AWS

Americo Carvalho (He/Him)
Head AIML Specialty
AWS, WW Public Sector

Sam Palani (He/Him)
Head AIML Architecture
AWS, WW Public Sector

Overview of Generative AI

Applications in Public Sector

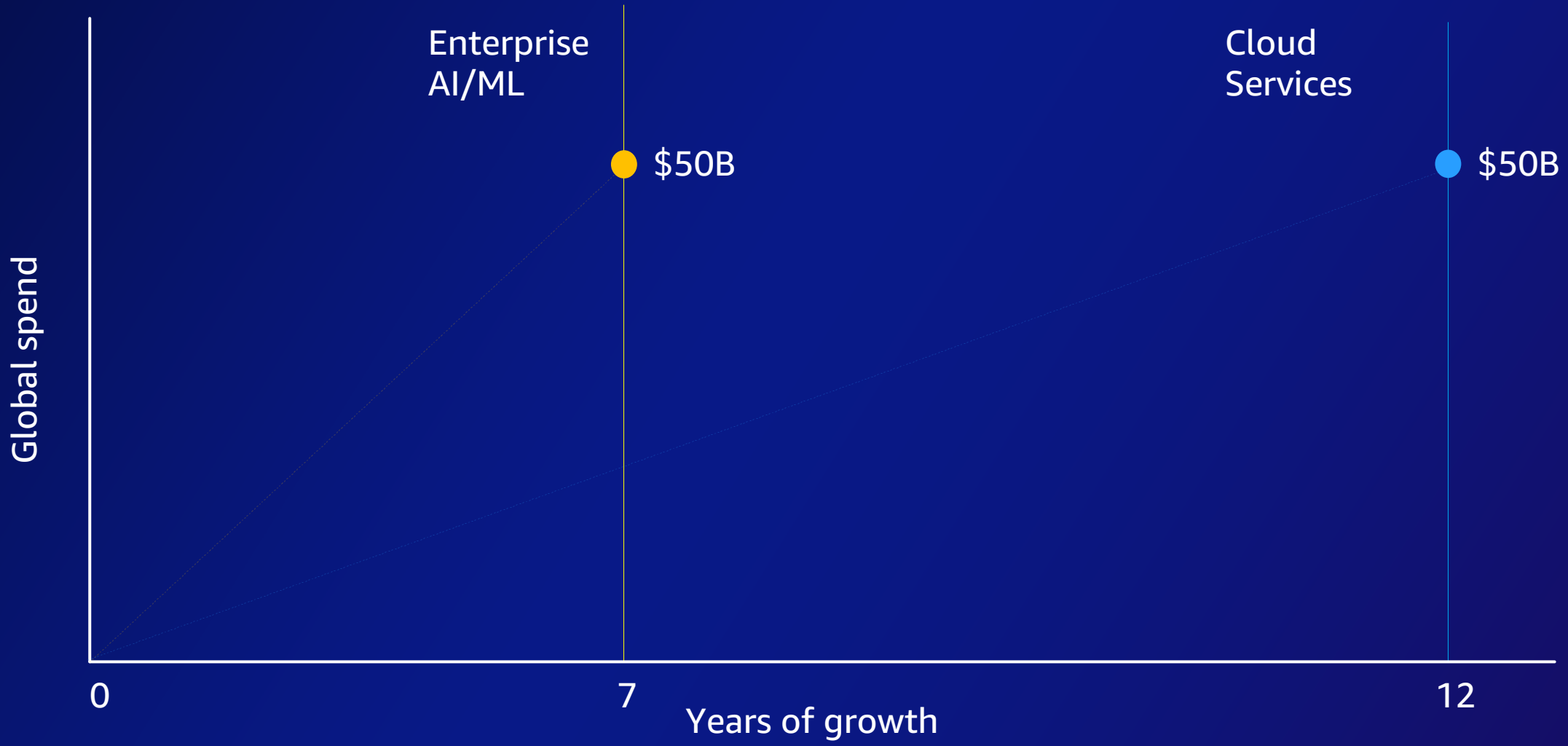
Technology Overview

Generative AI on AWS

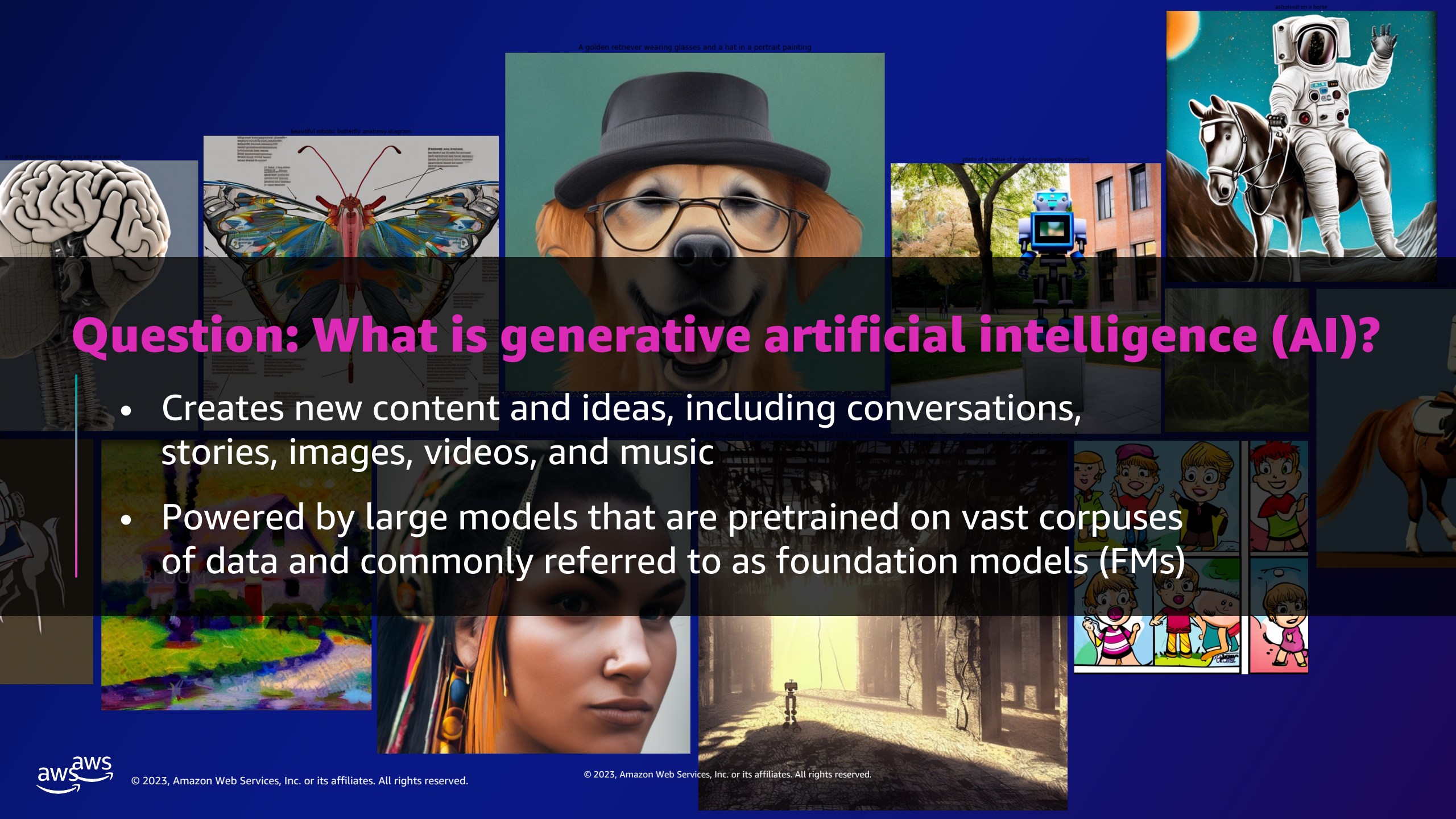
Resources & Getting Started Today

Q/A

Machine Learning is key to innovation



Source: IDC Worldwide Semiannual Artificial Intelligence Spending guide, Publication August 2021; IDC Semiannual Public Cloud Services Tracker, 1H2021, November 11, 2021
Note: Enterprise AI/ML and Cloud Services (Infrastructure and platform services) categories are not mutually exclusive



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)

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

Common use cases

Text generation

Q&A

Text summarization

Text extraction

Paraphrase rephrase

Search

Code generation

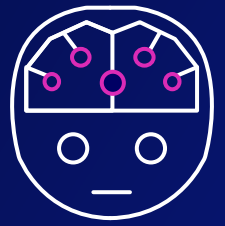
Image generation

Image classification

Audio generation

Video generation

Where does generative AI fit?

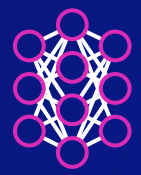


Artificial intelligence (AI)

Any technique that allows computers to mimic human intelligence using logic, if-then statements, and machine learning

Machine learning (ML)

A subset of AI that uses machines to search for patterns in data to build logic models automatically



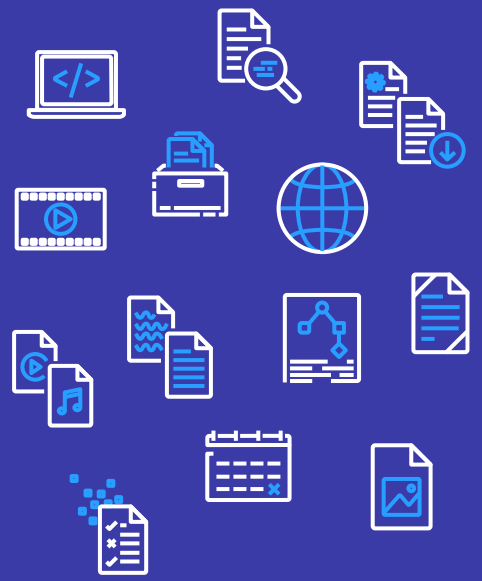
Deep learning (DL)

A subset of ML composed of deeply multi-layered neural networks that perform tasks like speech and image recognition

Generative AI

Powered by large models that are pretrained on vast corpora of data and commonly referred to as foundation models (FMs)

How foundation models work



Unlabeled data

Pretrain

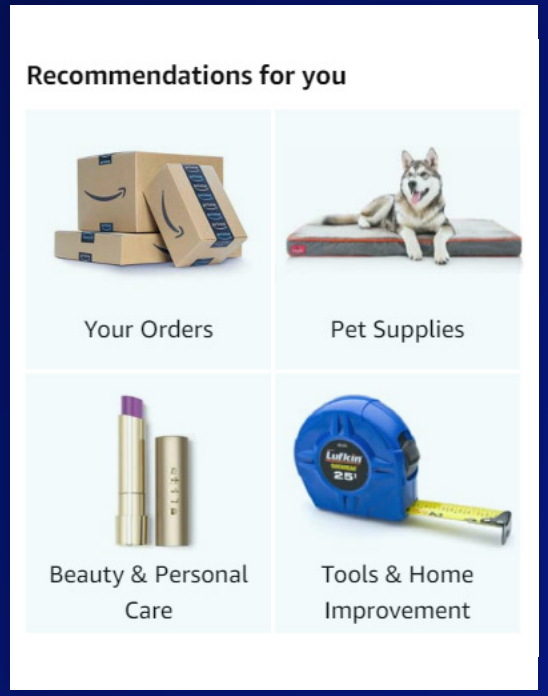
FM

Adapt

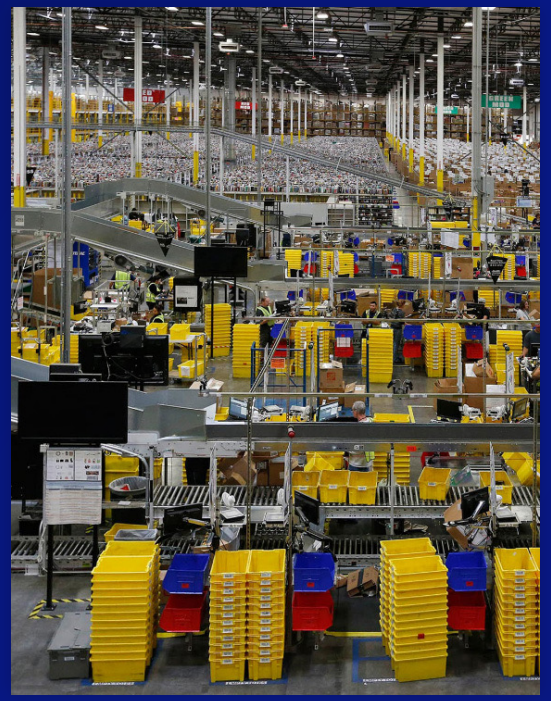
- Text generation
- Summarization
- Information extraction
- Q&A
- Chatbot

Broad range of general tasks

Amazon machine learning innovation at scale



4,000 products per minute sold on Amazon.com



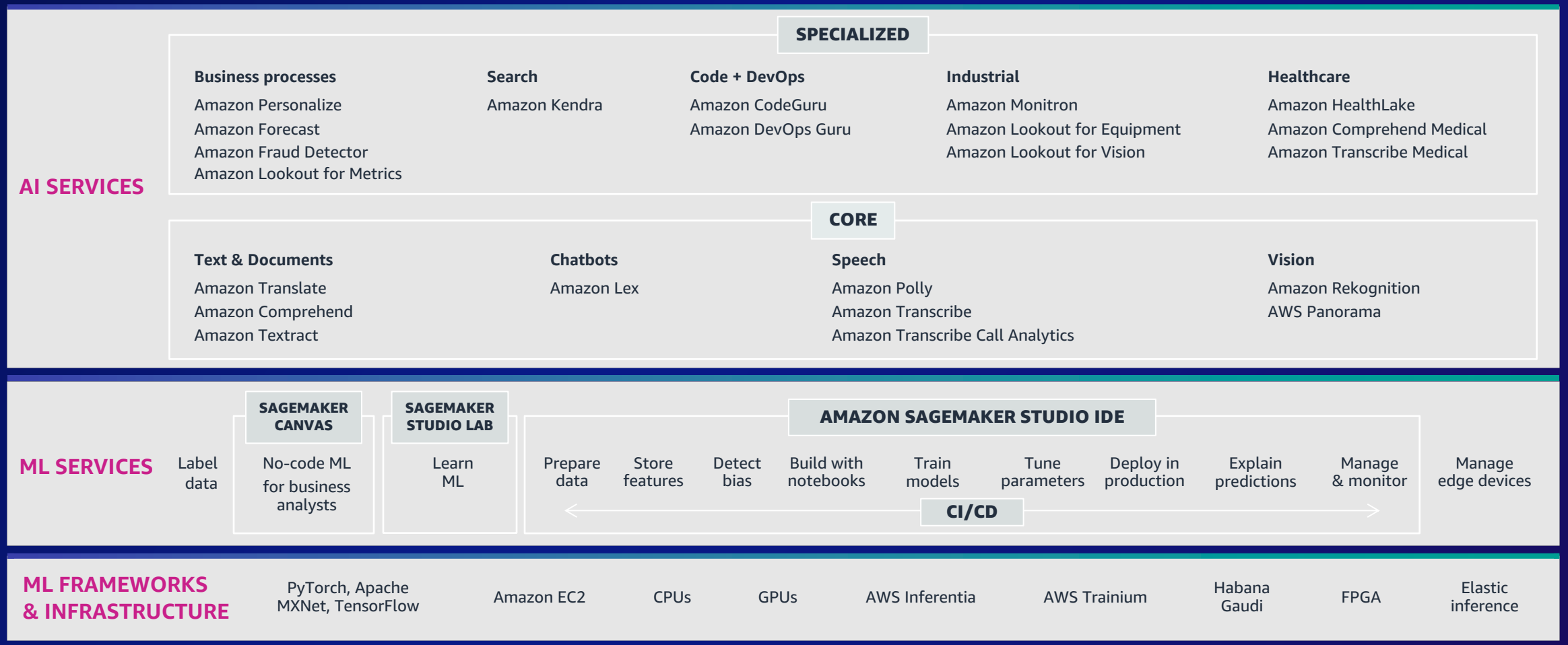
1.6 million packages every day

Billions of Alexa interactions each week

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

The AWS ML stack

Broadest and most complete set of machine learning capabilities



Amazon Generative AI Portfolio

Choice of many Foundation Models

AI SERVICES

Bedrock

In Preview

Amazon

Titan Text

Titan Embeddings

3rd Party

Jurassic-2

Claude

Stable Diffusion

Code Whisperer

Amazon

AI coding companion

ML SERVICES

Sagemaker

Jumpstart

Publicly available

stability.ai

Text2Image
Upscaling

alexa

AlexaTM 20B

Flan T-5 models
DistilGPT2, GPT2
Bloom models

Proprietary models

co:here

Cohere generate-med

Lighton

Lyra-Fr10B

AI21labs

Jurassic-1 Grande 17B

In Preview only

ML FRAMEWORKS & INFRASTRUCTURE

Self Managed ML

Meta PyTorch

3-way collaboration to move models to production on EC2 and Sagemaker

Hugging Face

CodeWhisperer:
ML-powered coding companion

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

```
# 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')
```

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

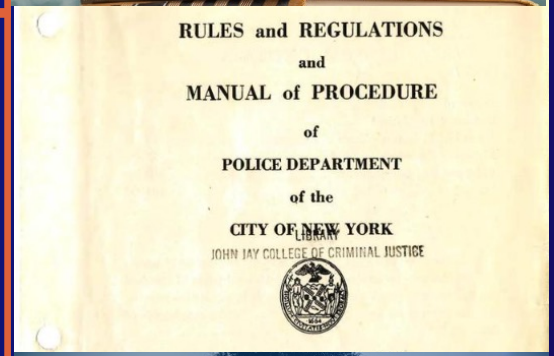
Generative AI use case in public sector

Education

- Flash card generation, quizzes, personalized learning
- Accessibility for impaired people
- Educational assistant (Conversational AI)

Gov.

- Information retrieval and synthesis (ex: legal docs)
- **Policy analysis and recommendation**
- Law enforcement - Sketches generation (individuals. Vehicles)

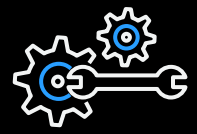


Healthcare

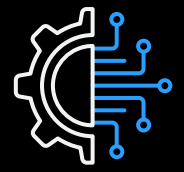
- Clinical coding assistance
- Ease diagnostic, care comprehension for patients
- Preventive care content generation

With the growth of AI comes the recognition that we must all use it **responsibly**

Our commitment to develop AI and machine learning in a responsible way is integral to our approach



Transforming responsible AI from theory to practice



Integrate responsible AI into the end-to-end ML lifecycle

Nurture and educate a more diverse generation of leaders in ML

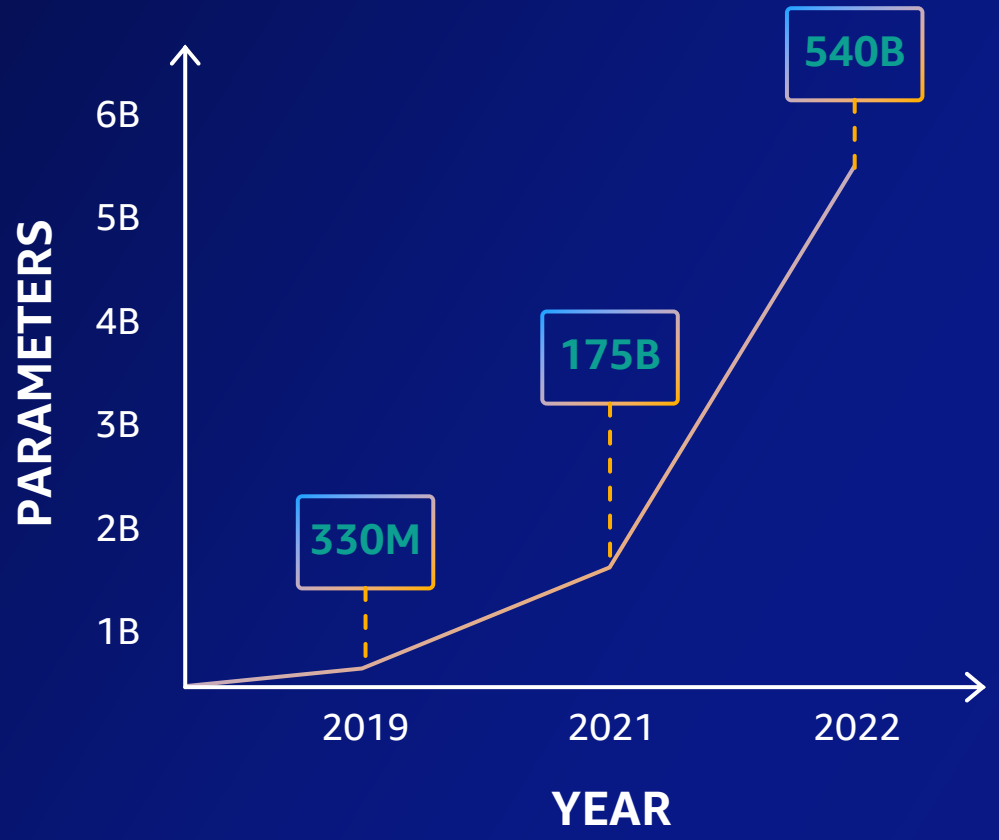
Advance the science behind responsible AI

Technology & AWS Capabilities

Generative AI

"Generative AI is a type of AI that can create new content and ideas, including conversations, stories, images, videos, and music. Like all AI, generative AI is powered by ML models—very large models that are pre-trained on vast amounts of data and commonly referred to as Foundation Models (FMs)."

Rise of foundation models – what has changed?

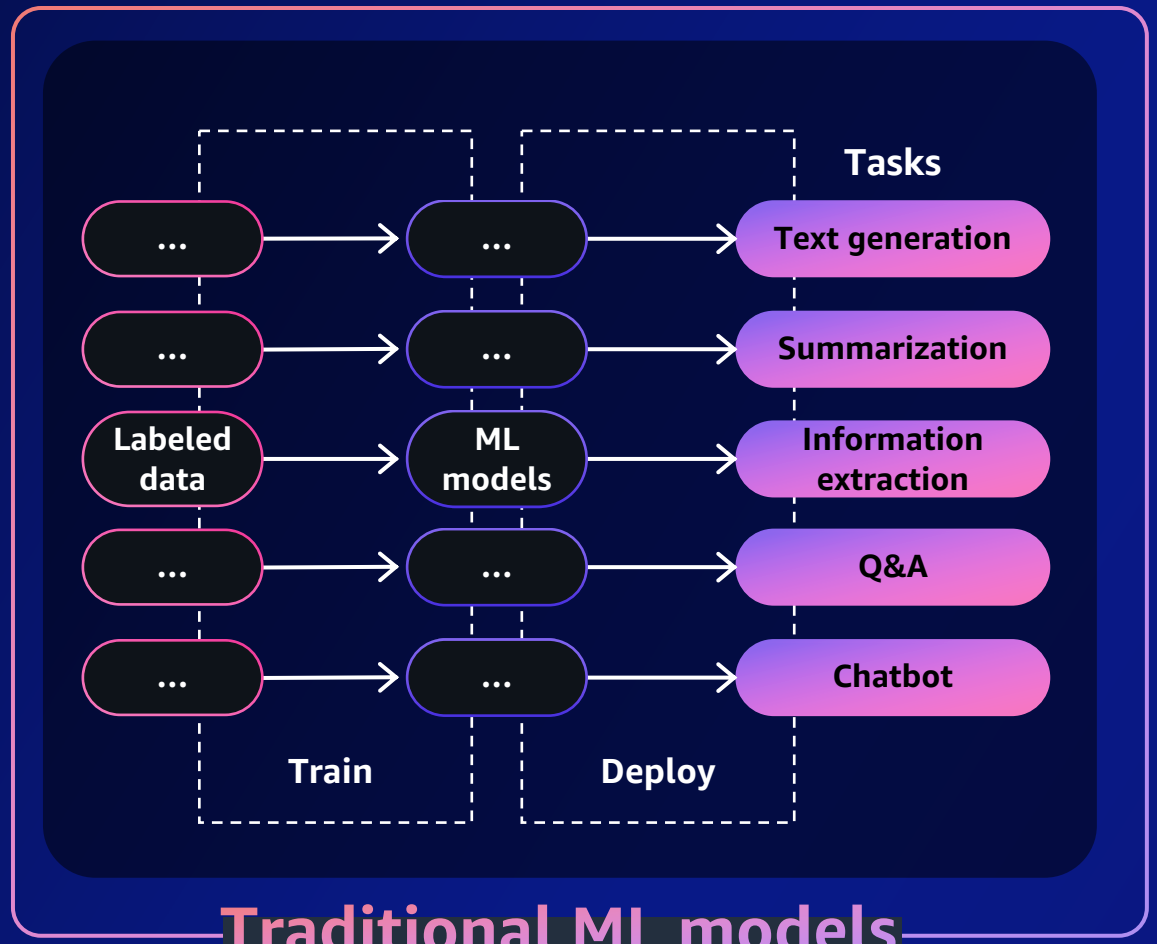


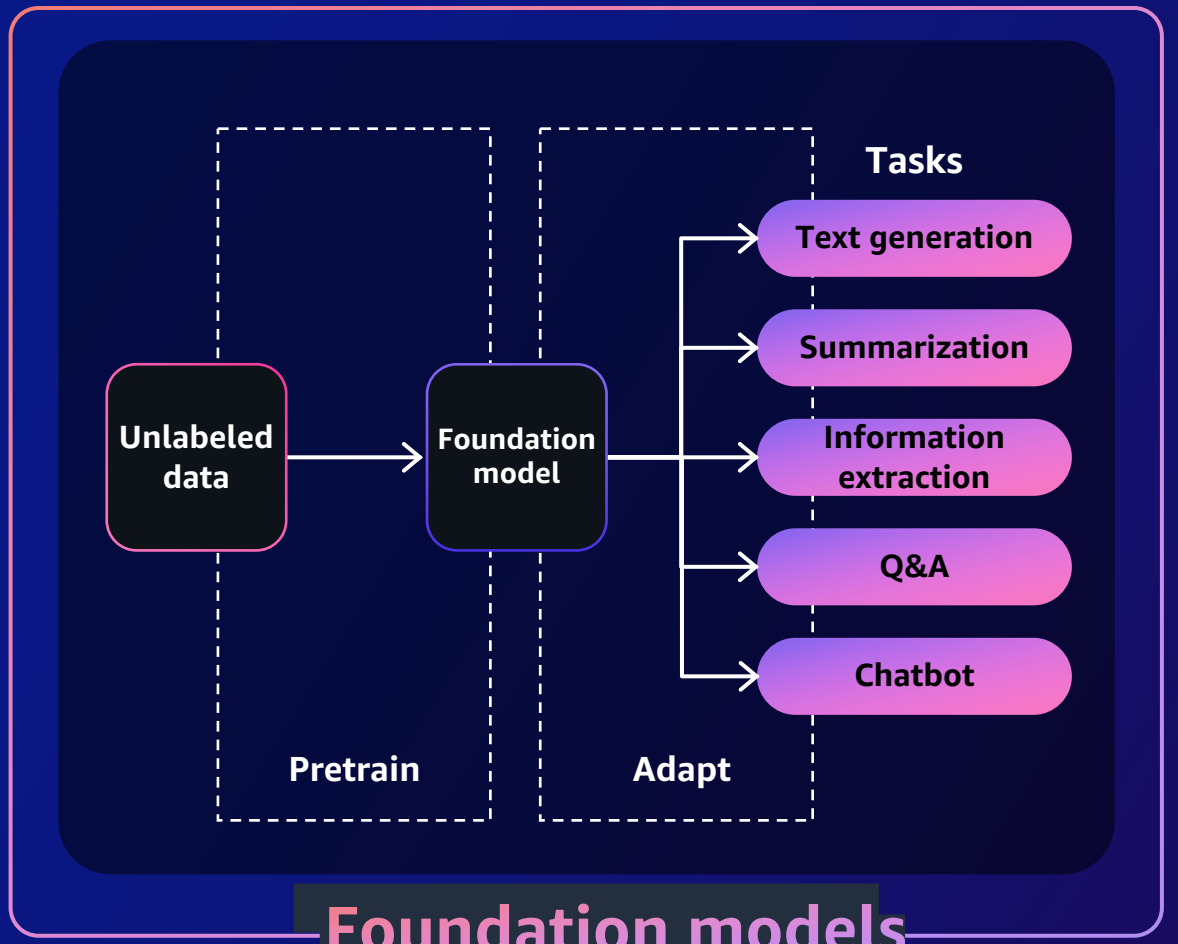
2019–2022

1,600x

increase in size of model as measured by number of parameters

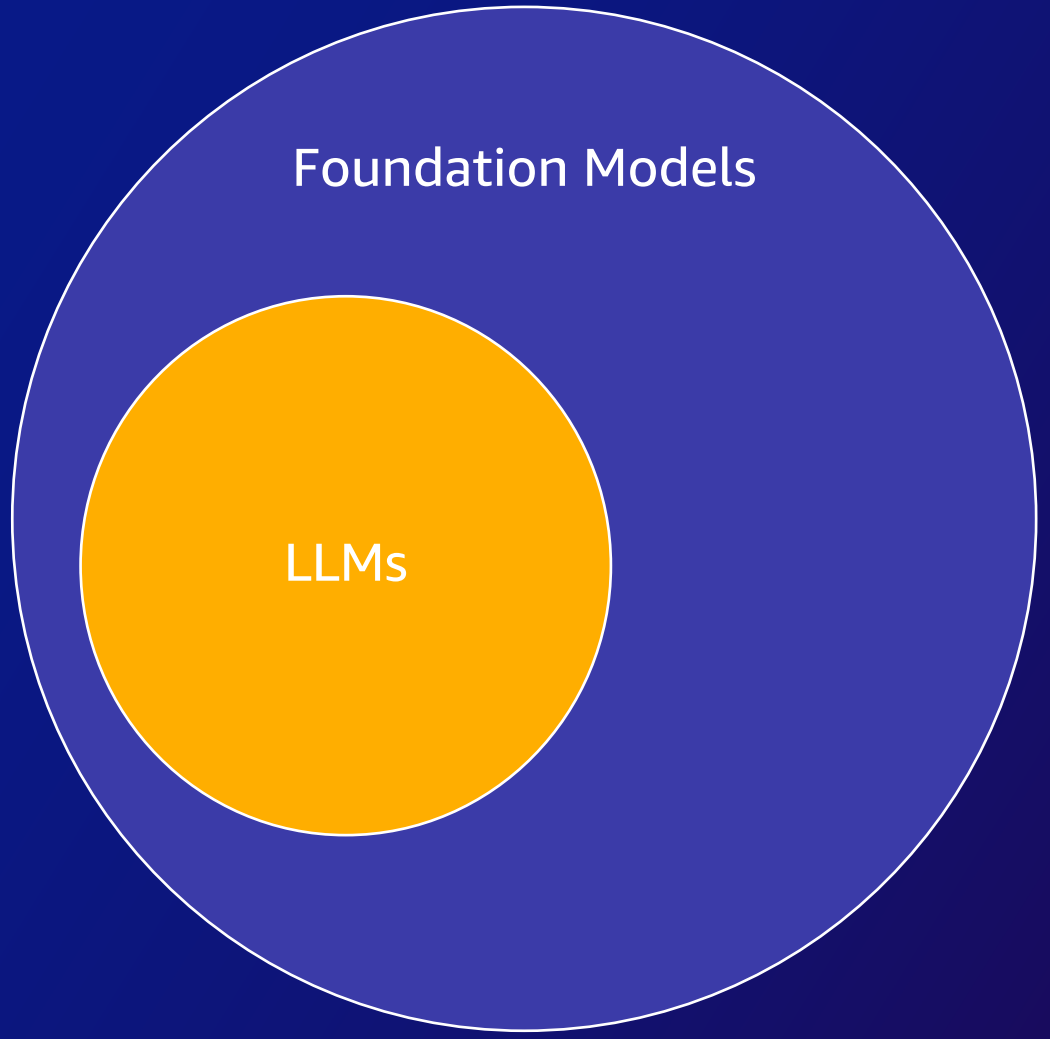
Traditional Model vs Foundation Model





What are Large Language Models (LLMs)

- Trained on text
- Excel at natural language prompts
- Have accelerated use cases such as question answering, code generation/explanation, summarization etc.



What do LLMs do?

- Like all ML models, LLMs make predictions
- The next ‘token’ in a sequence
- They produce a reasonable continuation of text based on the training data.

Building honest and responsible AI system requires

Bias Mitigation	4.5%
Transparency	3.6%
Ethics	3.2%
Accountability	3.1%
Privacy	2.7%

LLMs – Art vs Science

- Picking the most probable token returns flat/repetitive text

Building honest and responsible AI systems requires bias mitigation. Honest and responsible AI systems require bias mitigation to maintain public trust in technology. Bias mitigation is the cornerstone of developing honest and responsible AI systems. Building honest and responsible AI systems requires bias mitigation to maintain public trust in technology.

- 'Temperature' – parameter controls entropy

^^ Temperature = 0

Temperature = 0.8 →

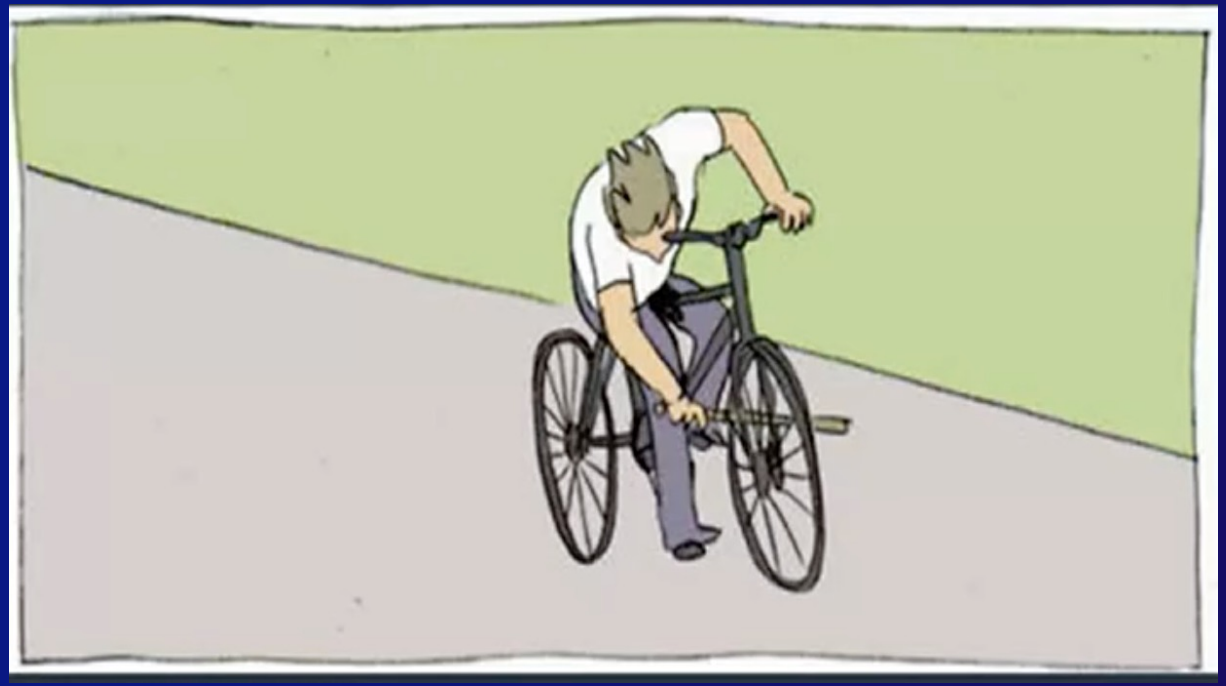
Constructing fair and accountable AI systems hinges on bias mitigation, a complex dance that intertwines with the very algorithms we program, asking us to remain vigilant and open to continual adjustment. Every line of code, every data point we feed into our AI, can either uphold or challenge systemic bias - that's the pivotal role of bias mitigation in constructing AI systems that are both honest and responsible.

Superpowers

- Aggregate and summarize complex content.
- Generate new content / infer on evidence.
- Improved with limited amount of domain data.
- Prompt engineering to tailor output.
- Human feedback to tune further.

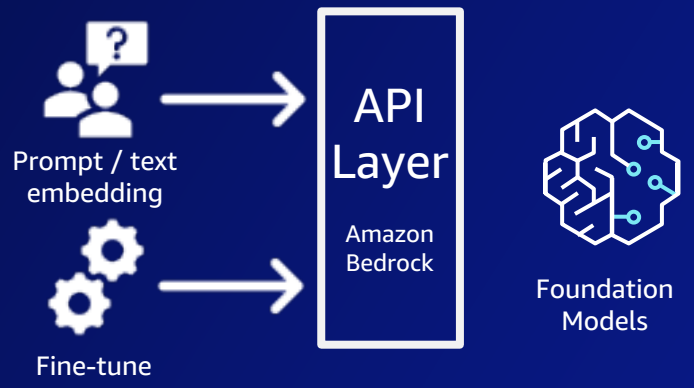
Limitations

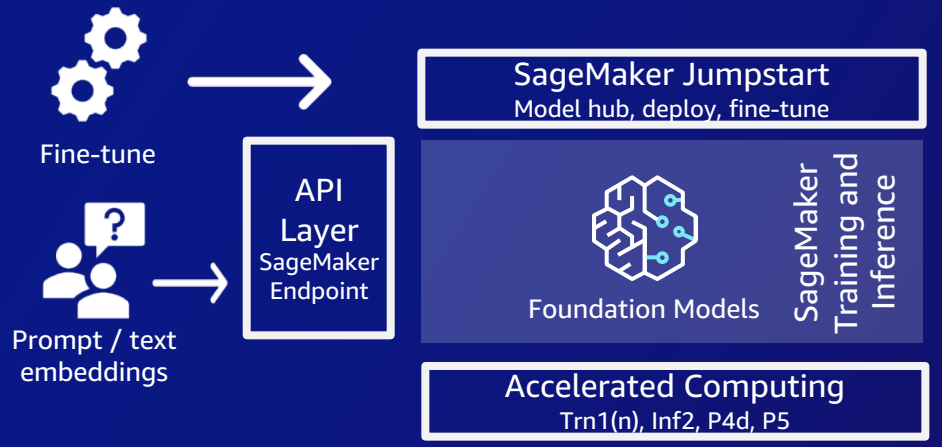
- Results can be unpredictable – model hallucination.
- Massive compute for building and using LLMs
- Output only as good as the prompt.
  - Retrieval Augmented Generation
  - Chat & Sessions



Generative AI on AWS

How do I access foundation models?





Amazon Bedrock

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

Amazon SageMaker JumpStart

- Machine learning (ML) hub with foundation models, 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

NEW

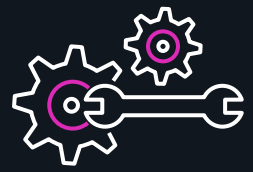
Amazon Bedrock

The easiest way to build and scale generative AI applications with FMs

Amazon Bedrock key benefits

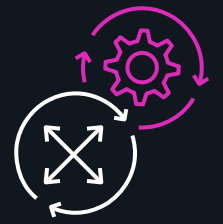
Accelerate development of generative AI applications using FMs through an API, without managing infrastructure

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



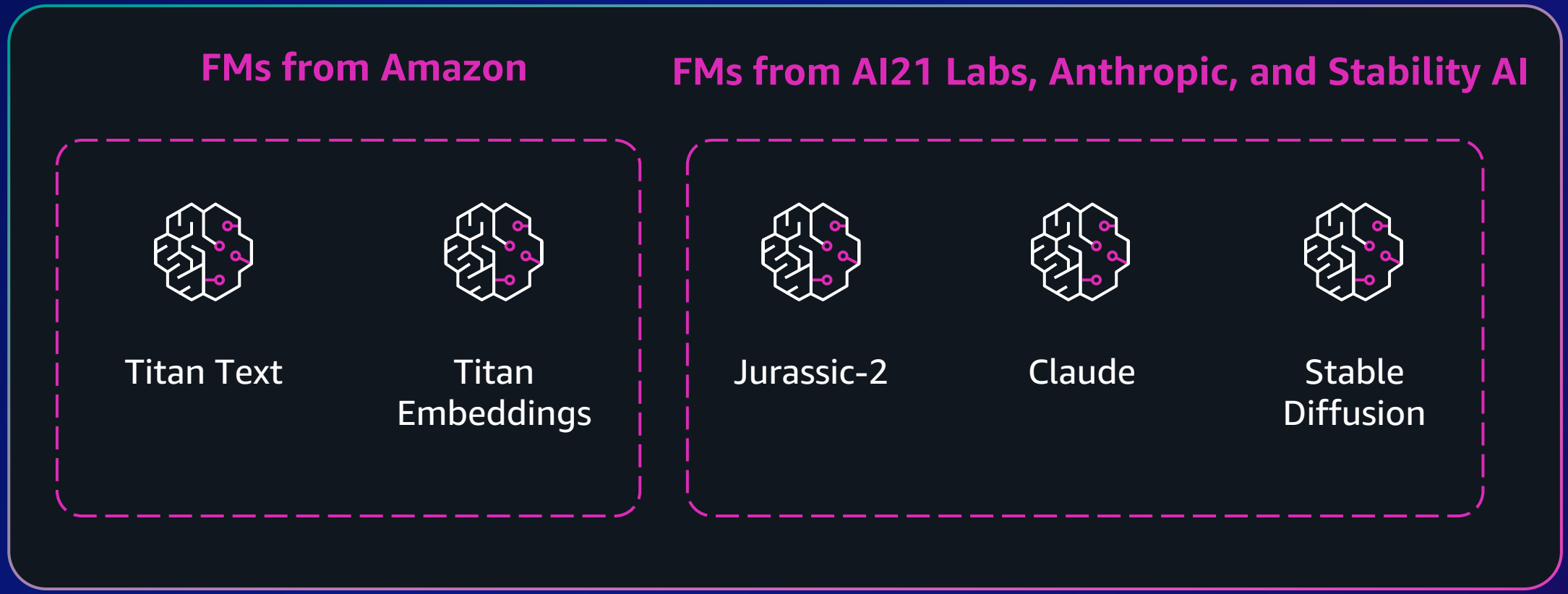
Privately customize FMs using your organization's data

Enhance your data protection using comprehensive AWS security capabilities



Use AWS tools and capabilities that you are familiar with to deploy scalable, reliable, and secure generative AI applications

Bedrock supports a wide range of foundation models



Amazon Titan

INNOVATE RESPONSIBLY WITH HIGH-PERFORMING FMs FROM AMAZON

Titan Text
focused on NLP tasks

Titan Embeddings
for enterprise tasks such as search and personalization

Benefits

- Built with 20+ years of Amazon ML experience
- Automate language tasks such as summarization and text generation with Amazon Titan Text FM
- Enhance search accuracy and improve personalized recommendations with Amazon Titan Embeddings FM
- Support responsible use of AI by reducing inappropriate or harmful content

Foundation models from top AI startups

Jurassic-2

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

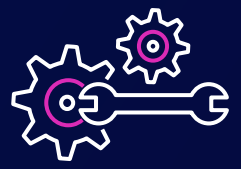
Claude

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

Stable Diffusion

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

Privately customize foundation models using your organization's data



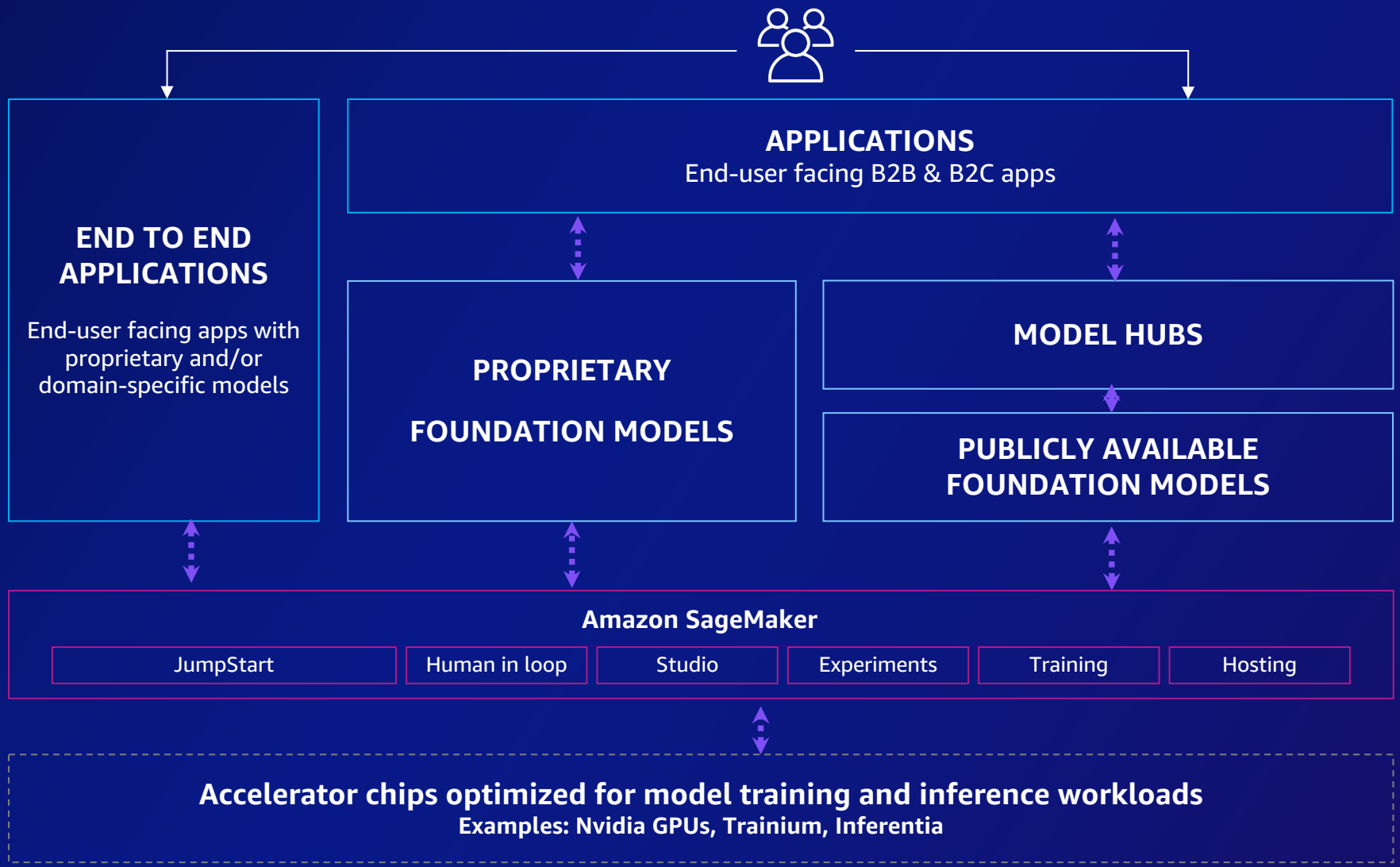
Fine-tune

PURPOSE — Maximizing accuracy for specific tasks

DATA NEED — Small number of labeled examples

Amazon SageMaker & SageMaker Jumpstart

Generative AI: Lay of the land using Amazon SageMaker



SageMaker JumpStart models and features

Publicly available

stability.ai

Models

Text2Image

Upscaling

Tasks

Generate photo-realistic images from text input

Improve quality of generated images

Features

Fine-tuning on SD 2.1 model

alexa

Models

AlexaTM 20B

Tasks

Machine translation

Question answering

Summarization

Annotation

Data generation

Models

Flan T-5 models (8 variants)

DistilGPT2, GPT2

Bloom models (3 variants)

Tasks

Machine translation

Question answering

Summarization

Annotation

Data generation

Proprietary models

co:here

Models

Cohere generate-med

Tasks

Text generation

Information extraction

Question answering

Summarization

Lighton

Models

Lyra-Fr 10B

Tasks

Text Generation

Keyword extraction

Information extraction

Question answering

Summarization

Sentiment analysis

Classification

AI21labs

Models

Jurassic-1 Grande 17B

Tasks

Text generation

Long-form generation

Summarization

Paraphrasing

Chat

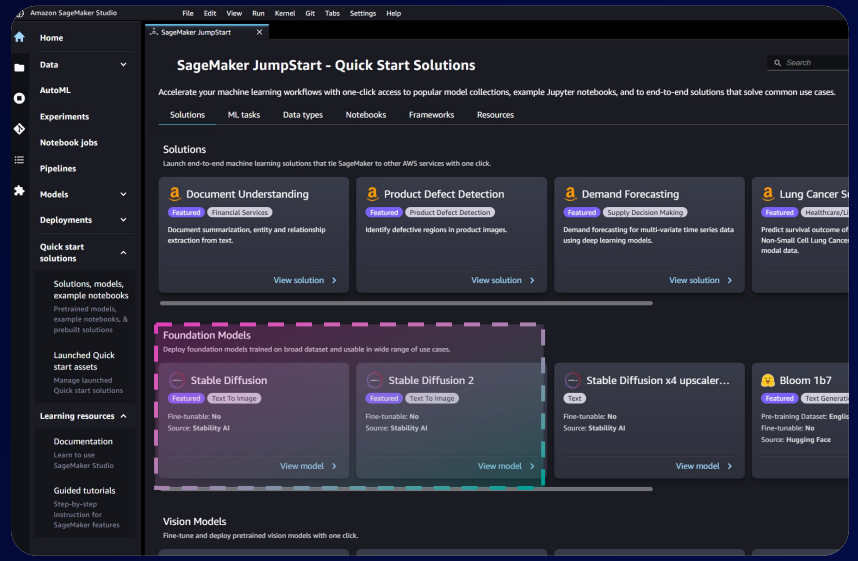
Information extraction

Question answering

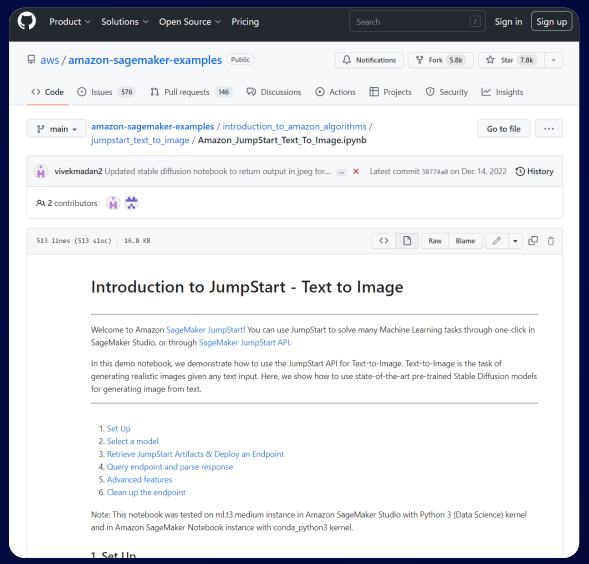
Classification

3 ways to use foundation models with SageMaker JumpStart

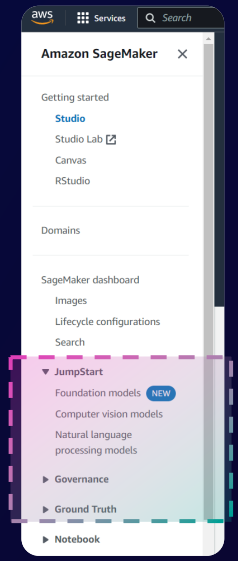
SageMaker Studio
One-step deploy



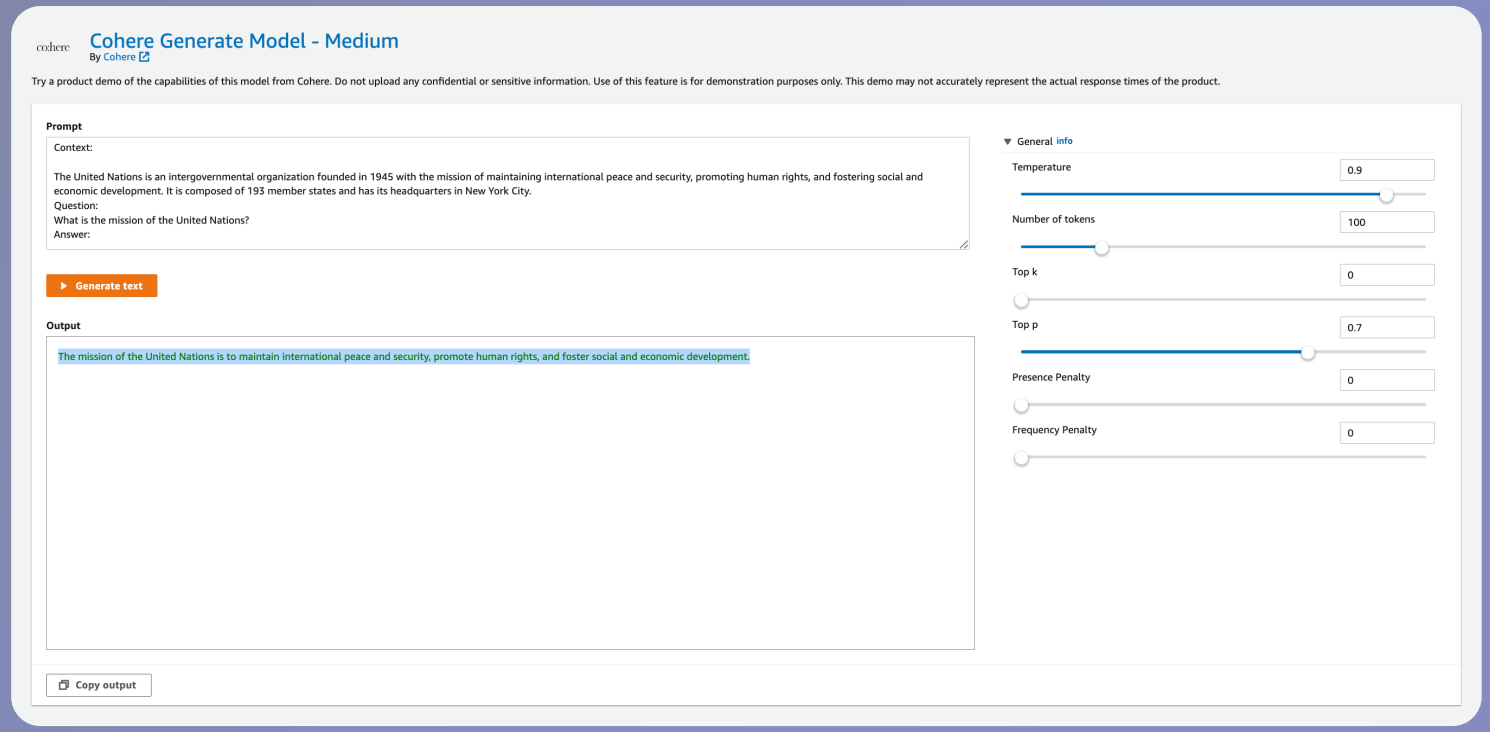
SageMaker Notebooks



AWS Management Console Preview

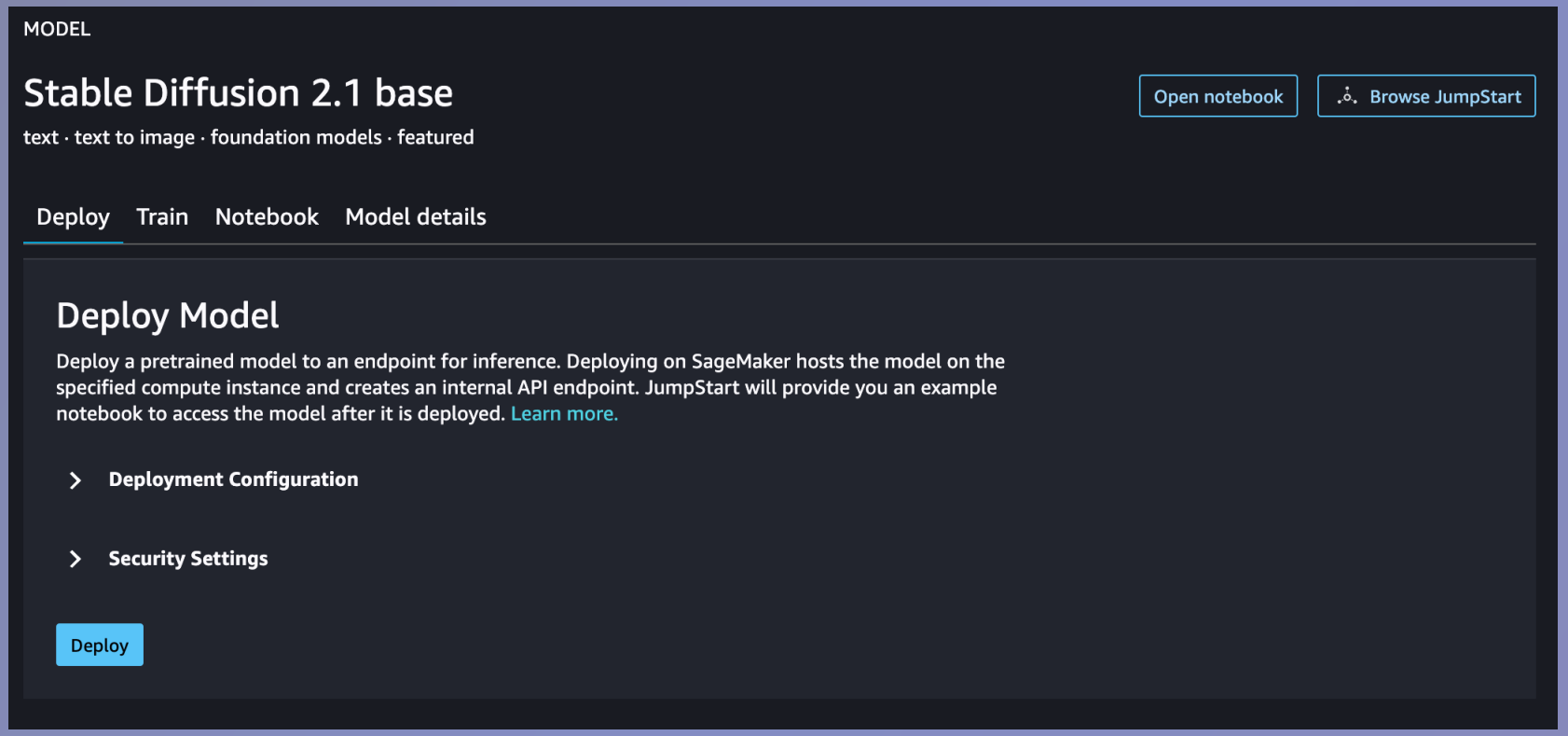


Try-out experience



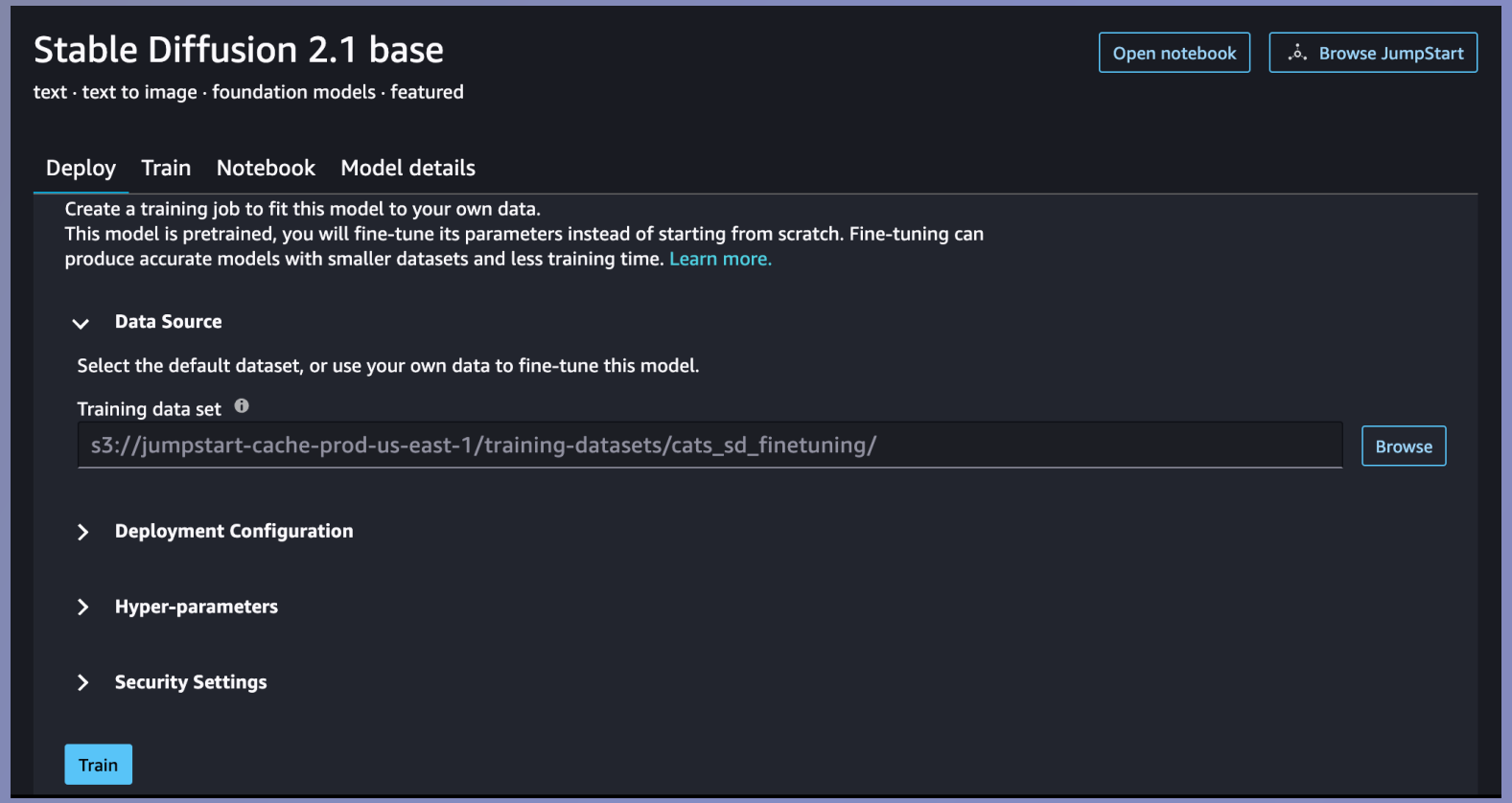
- Try out the models and model prompts without running code or incurring costs
- Available for proprietary models in Top 10 in HELM benchmarks and public models for comparison purposes

Easy deploy experience



- Training instance type
- Security Settings

Easy fine-tune experience



Labeled data set path

Training instance type

Hyper-parameters & Security settings

Resources & Getting Started Today

Start your generative AI journey today

Check out the generative AI webpage

Read announcement blog post

Watch Werner Vogels, AWS VP and CTO, explain generative AI

Examples: Retrieval-Augmented Generation

QUESTION AND ANSWER USING DOMAIN SPECIFIC DATASET

Amazon SageMaker Jumpstart + VectorDB as Amazon SageMaker KNN and Opensource (langchain)

Amazon SageMaker Jumpstart + VectorDB as Amazon OpenSearch

Example: Domain adaptation and fine-tuning

QUESTION AND ANSWER USING DOMAIN SPECIFIC DATASET

Domain Adaption Fine Tuning using Amazon SageMaker JumpStart on Financial Data

Thank you!

Please complete the session survey in the mobile app

Americo Carvalho

Email – americoc@amazon.com

Sam Palani

Email – sampal@amazon.com

LinkedIn & Twitter - @samx18