
aws
SUMMIT
NEW YORK | JULY 26, 2023

AIM201

Build with generative AI on AWS, featuring Anthropic

Atul Deo (he/him)
Director and GM - Amazon Bedrock
AWS

Matt Bell (he/him)
Head of Product Research
Anthropic

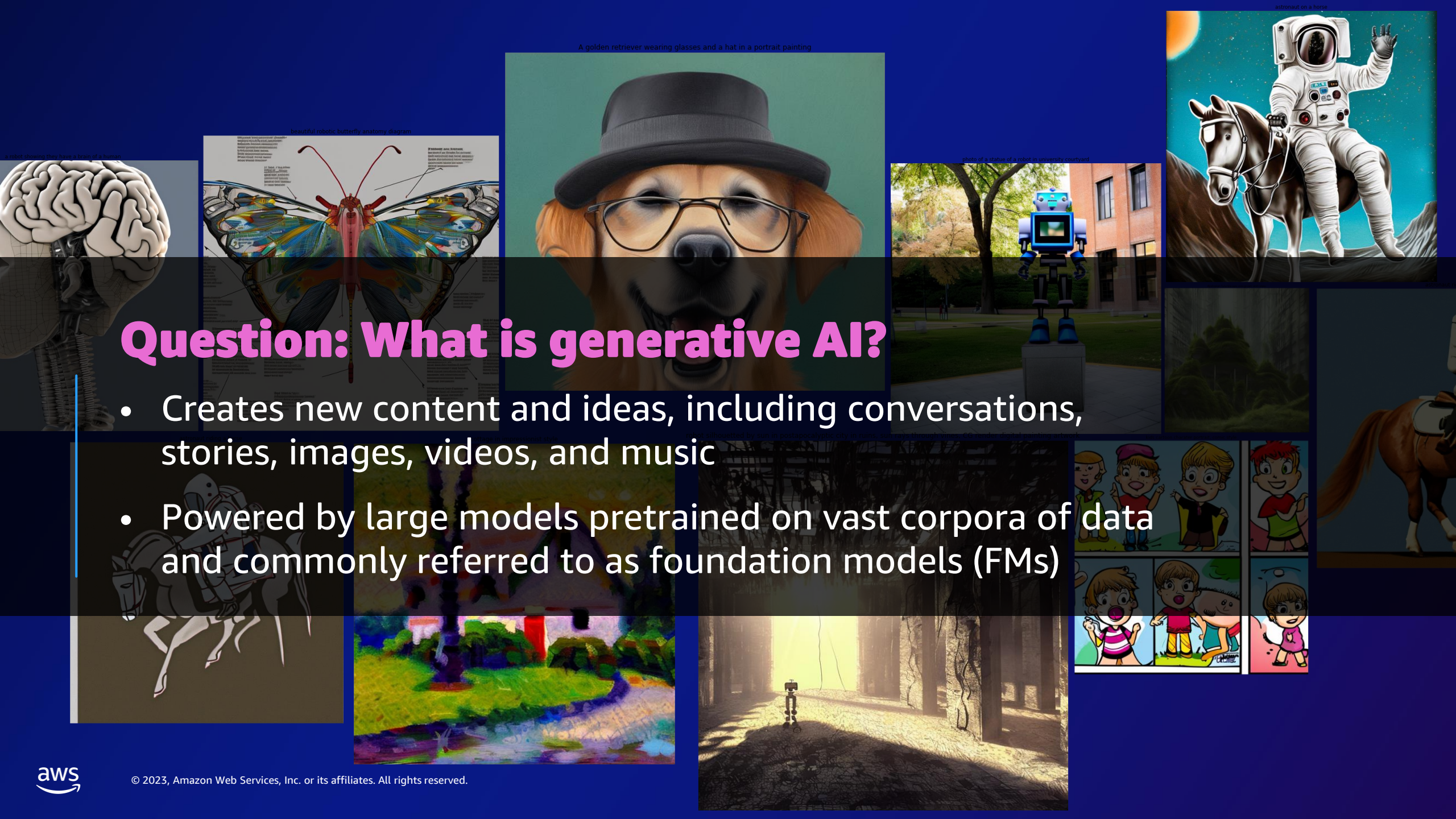
Agenda

Introduction to generative AI

Tools to build with generative AI on AWS

Anthropic Claude 2 deep dive

How to get started with generative AI



Question: What is generative AI?

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

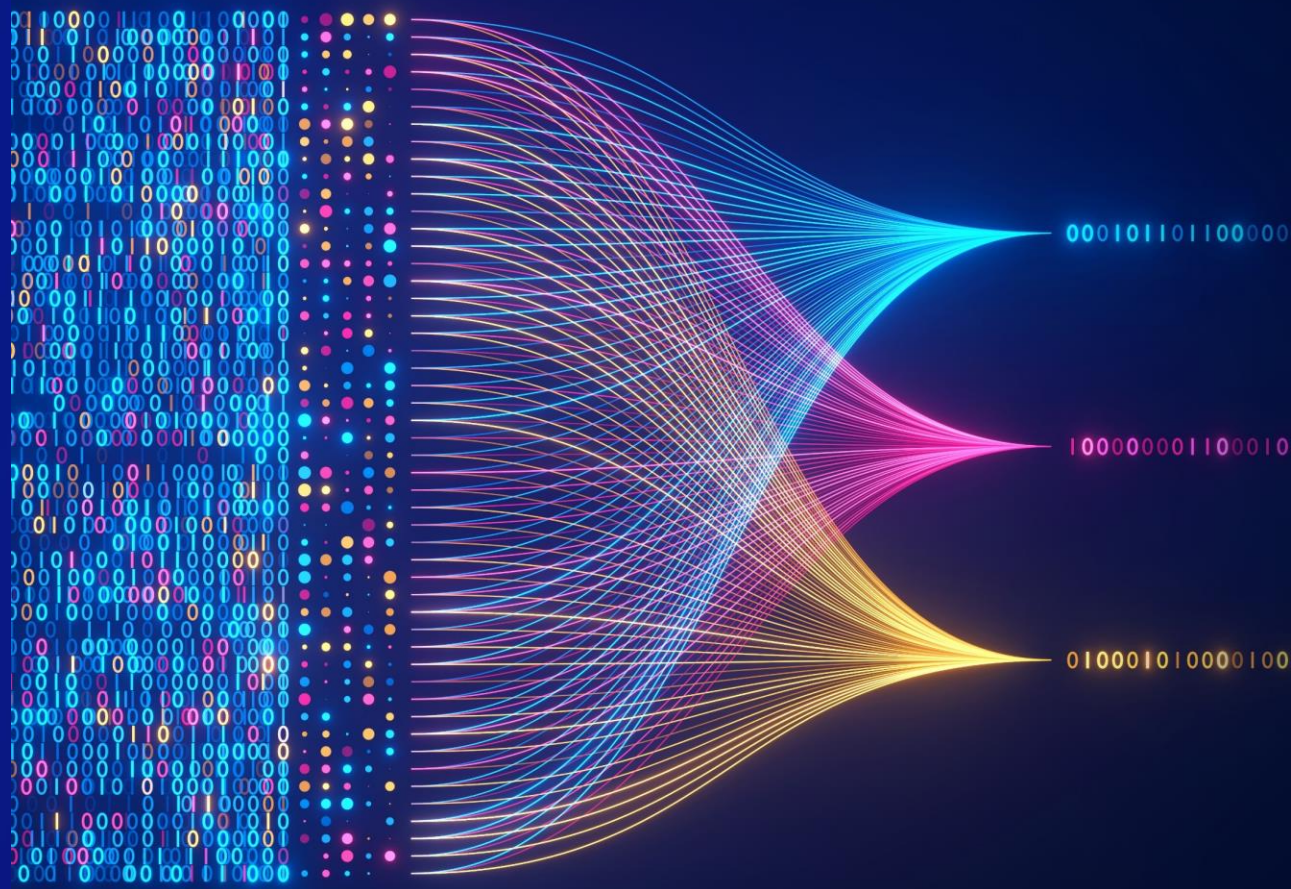
Generative AI is powered by foundation models

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

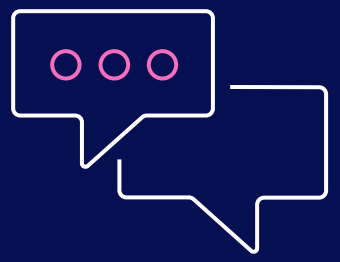
Customize FMs using your data for domain specific tasks



Types of foundation models

Input	FM	Output
"Summarize the articles on impact of walking on heart health"	**Text-to-text** Generate text from natural language prompts	"Ten thousand steps per day is optimum for maintaining a healthy heart"
"hand soap"	**Text-to-embeddings** Generate numerical representation of text	Numerical representation of "Hand soap refills Hand soap dispenser Hand soap antibacterial"
"a photo of an astronaut riding a horse on mars"	**Multimodal** Create and edit images using natural language prompts	

Generative AI creates significant business value



New experiences

Create new innovative and engaging ways of interacting with your customers and employees

Productivity

Radically improve productivity across all lines of business, for example Amazon CodeWhisperer helped complete tasks 57% faster

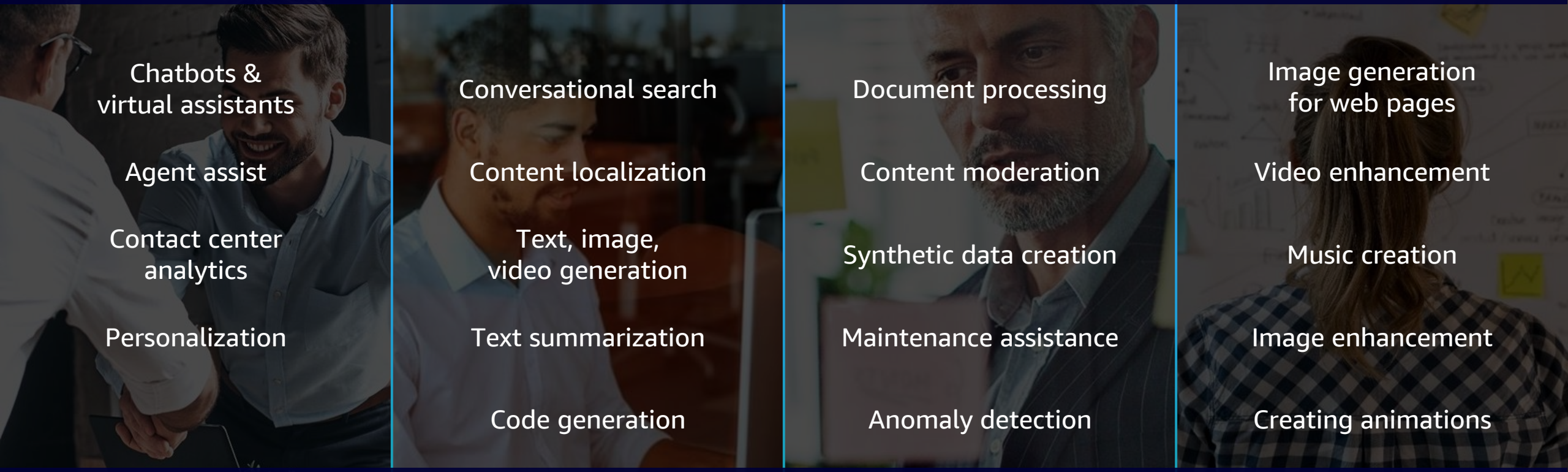
Insights

Extract insights and clear answers from all your corporate information, enabling faster and better decisions

Creativity

Create new content and ideas, including conversations, stories, images, videos, and music

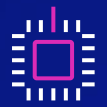
Generative AI can have a wide range of use cases



New experiences	Productivity	Insights	Creativity
Chatbots & virtual assistants	Conversational search	Document processing	Image generation for web pages
Agent assist	Content localization	Content moderation	Video enhancement
Contact center analytics	Text, image, video generation	Synthetic data creation	Music creation
Personalization	Text summarization	Maintenance assistance	Image enhancement
	Code generation	Anomaly detection	Creating animations

Tools to build with generative AI on AWS

The easiest way to build with FMs



The most price-performant infrastructure

Generative AI-powered applications

Selection of open-source FMs

Tools to build with generative AI on AWS

The easiest way to build with FMs

The most price-performant infrastructure

Generative AI-powered applications

Selection of open-source FMs

Amazon Bedrock

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

- Accelerate development of generative AI applications using FMs through an API
- No need to manage infrastructure
- Choice of industry-leading foundation models
- Privately customize FMs using your organization’s data
- Comprehensive AWS security capabilities

How Amazon Bedrock works

Amazon Bedrock

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

Choose an FM

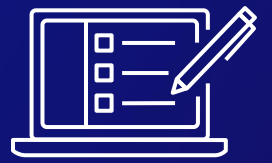
Use the playground to experiment with FMs and select the one that suits your needs

Use as is or customize

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

Send prompt

Use Amazon Bedrock API to send your prompts to the model



Integrate response

Integrate generated content into your application

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

new

Claude 2

LLM for thoughtful dialogue, content creation, complex reasoning, creativity, and coding, based on constitutional AI and harmlessness training

cohere

new

Command + Embed

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

stability.ai

new

Stable Diffusion XL 1.0

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

ANTHROP\C

Introducing Claude

Matt Bell
Head of Product Research
Anthropic

About Anthropic

AI progress = scaling + steerability + reliability + safety

We believe AI will permeate the economy within **2-4 years,** and we're focused on serving the world's leading enterprises

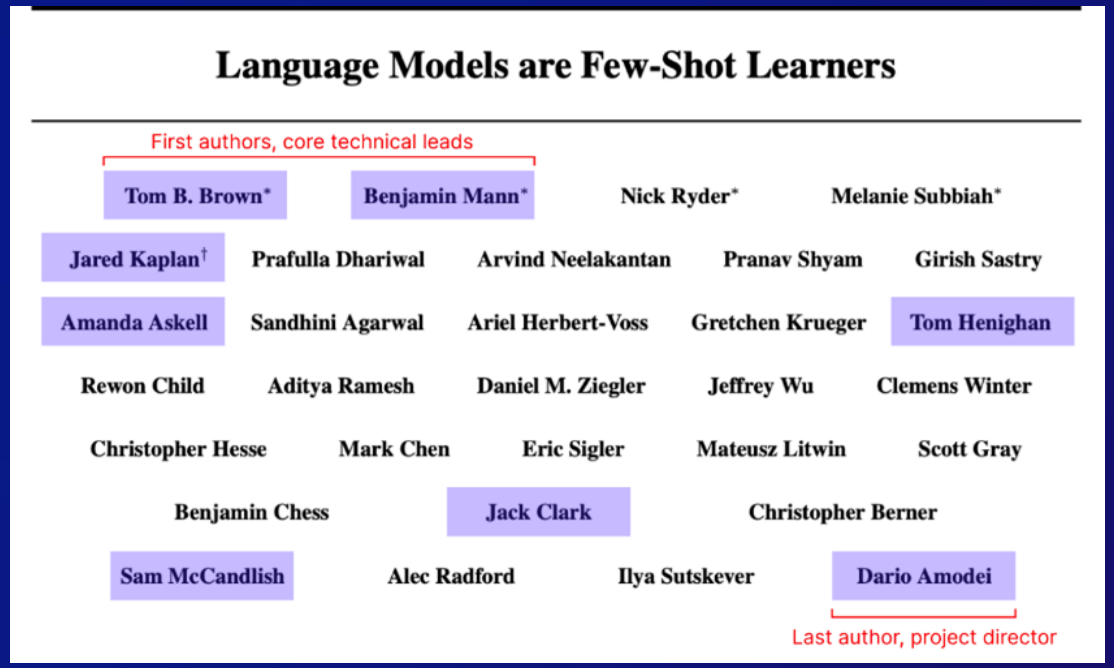
Who we are

Our technical leadership and co-founders are key authors in the groundbreaking GPT paper

Dario Amodei
CEO

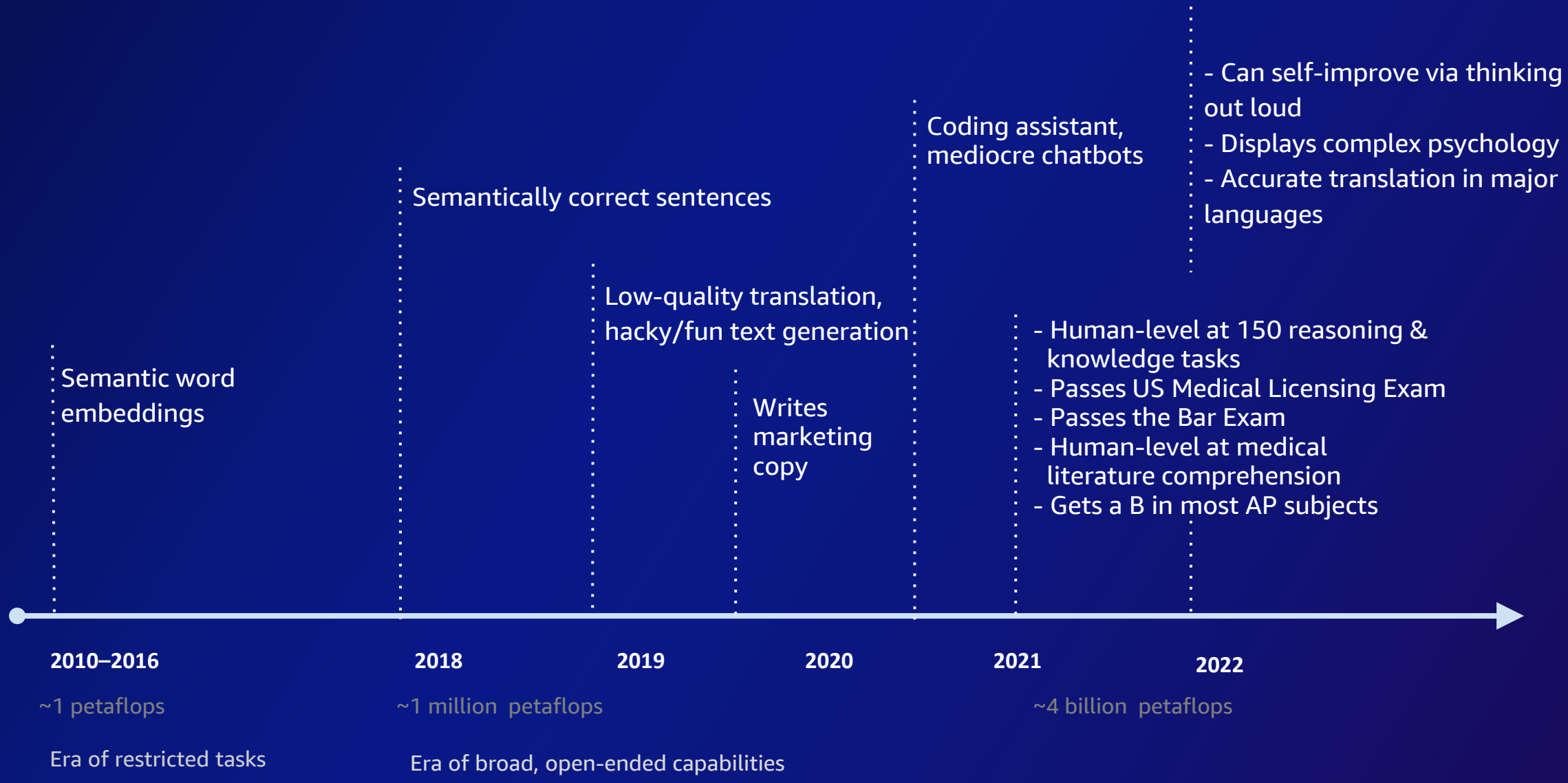
Daniela Amodei
President

Jared Kaplan
Alignment Science

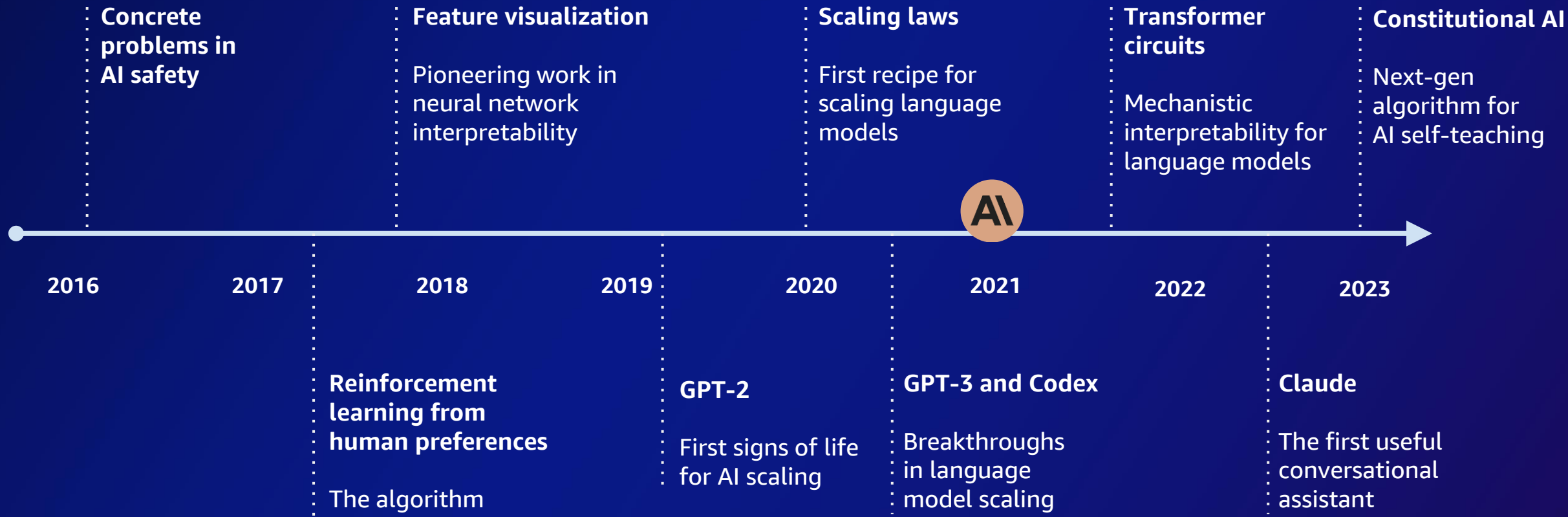


Authors in purple are founding members of Anthropic. Core technical leads (Tom Brown and Ben Mann) are Anthropic founders. Overall project was directed by Dario Amodei, Anthropic CEO.

An explosion of capabilities



Steerability and safety breakthroughs



Meet Claude 2

Today: Claude's state-of-the-art model

LLM leaderboard

Win rates vs. other models*

- GPT-4 (84%)
- Claude v1 (80%)
- Claude Instant v1 (75%)
- GPT-3.5 Turbo (72%)
- PaLM (53%)

Claude 2 data still to come!

*Source: Large Model Systems Organization is a research organization at UC Berkeley; https://chat.lmsys.org/?arena data as of 7/25/2023

Claude 2 compared to competitors

Claude 2 vs. GPT-4

- Safer
- 4 x as fast
- +50% cheaper
- 3-25x context window
- Much more customizable

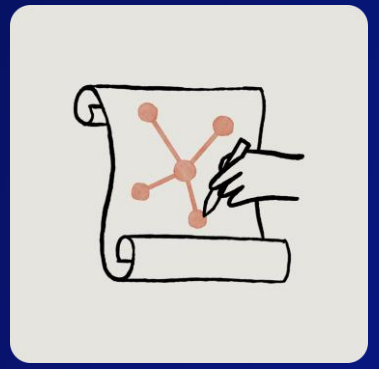
Powered by Claude:

Notion zoom slack Quora

What makes Claude different

Claude is safer

Constitutional AI



Constitutional principles

AI-generated datasets

Improved & aligned outputs

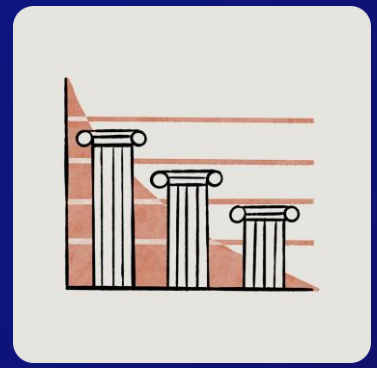
Helpful, honest, harmless (HHH)

Aims to help users

Only shares info it believes to be true

Won’t cooperate on harmful activities

Robust against jailbreaks



Significant, iterative red-teaming improvements for each model

Claude is flexible

Claude

Anthropic's most powerful model

Good for:

- Complex reasoning tasks
- Top quality creative work
- Sophisticated dialog
- Coding

Claude-instant

Very fast, very low cost model

Good for:

- Low latency tasks
- Lightweight dialog
- Content moderation
- Classification

Claude is cost-effective

ANTHROP\C	Input token count price/million token	Output token count price/million tokens
Claude-2 12k	$11.02	$32.68
Claude-2 100k	$22.04	$65.36

% price difference Claude vs. GPT-4	
Input	Output
-63%	-46%

Comparisons based on closest equivalent models. Claude-2 12k compared to GPT-4 8k; Claude-2 100k compared to GPT-4 32k

Claude has context

Claude-instant		GPT-3.5 Turbo
100k	vs.	16k
standard ~6x GPT-3.5 Turbo		standard

Claude 2		GPT-4
100k	vs.	8k/32k
standard ~4x GPT-4		standard /premium

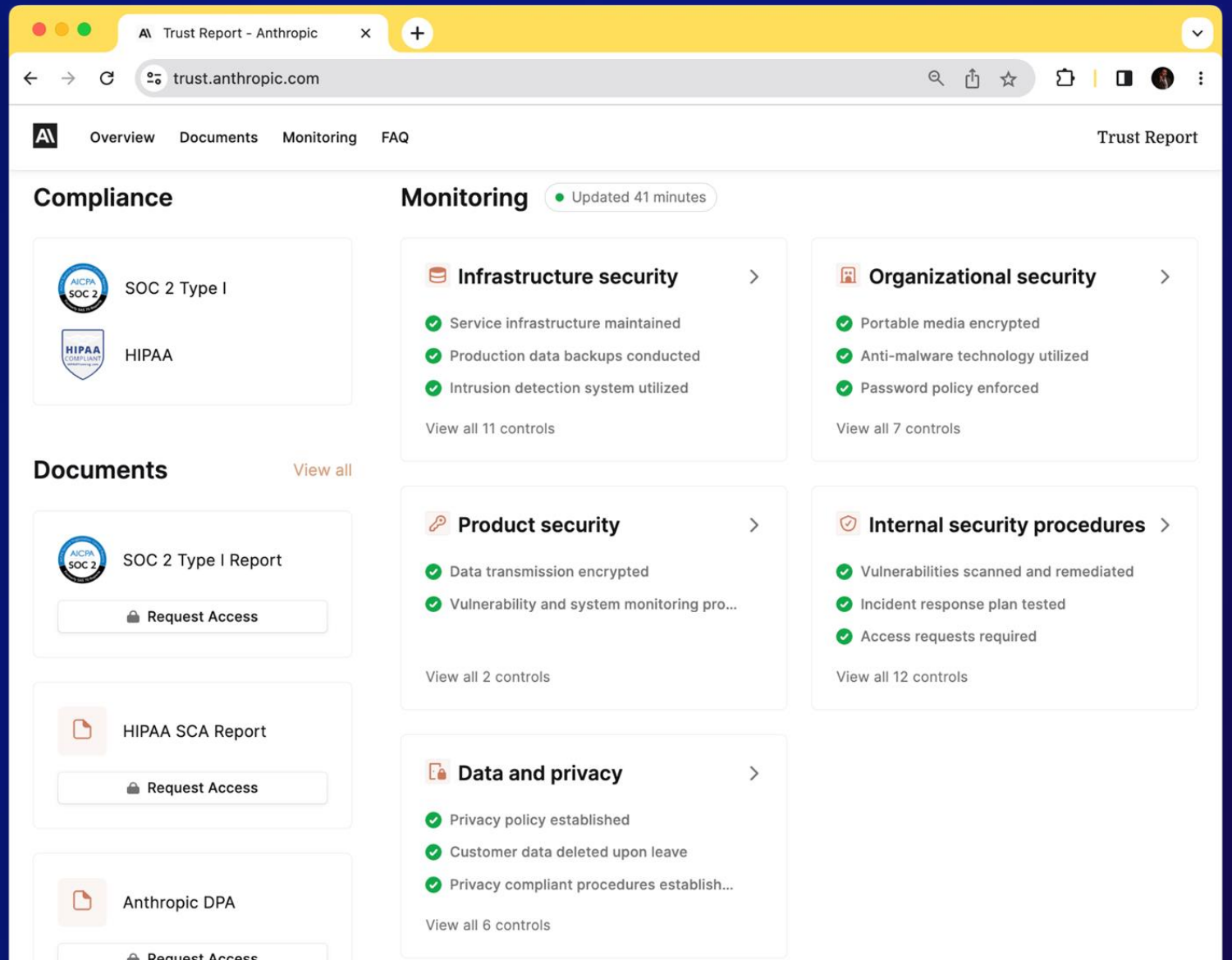
Context window length in tokens

Claude enables compliance

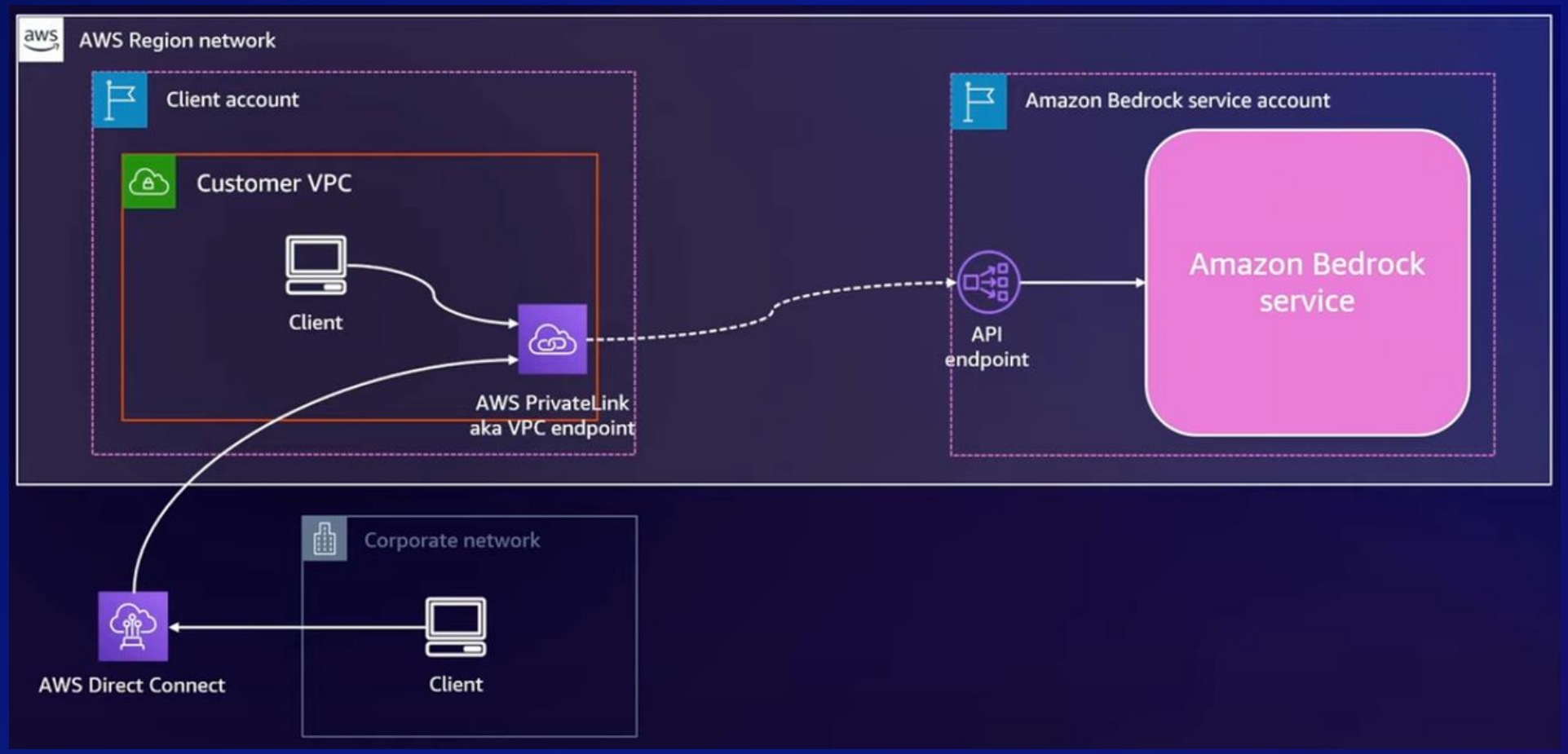
HIPAA COMPLIANT

AICPA SOC 2

Amazon Bedrock:
Multi or single tenant



Claude is secure – supported by AWS



Claude in action

Use cases

Core capabilities

- **Process mountains of text**
  - Documents, emails, FAQs, transcripts, records
- **Have natural conversations with deep knowledge**
  - Customer support, pre-sales, coaches, tutors, general advice “oracles”
- **Automate workflows**
  - Interface with search tools, APIs, and databases with autogenerated code

Combine them to solve complex tasks

- Customer service = Dialog agent + retrieval + document Q/A + API use + content moderation

Parts of a prompt

(0) “\n\nHuman:”

(1) Give task context

(2) Give detailed task description, including rules and exceptions.
Explain it as you would to a new employee with no context.

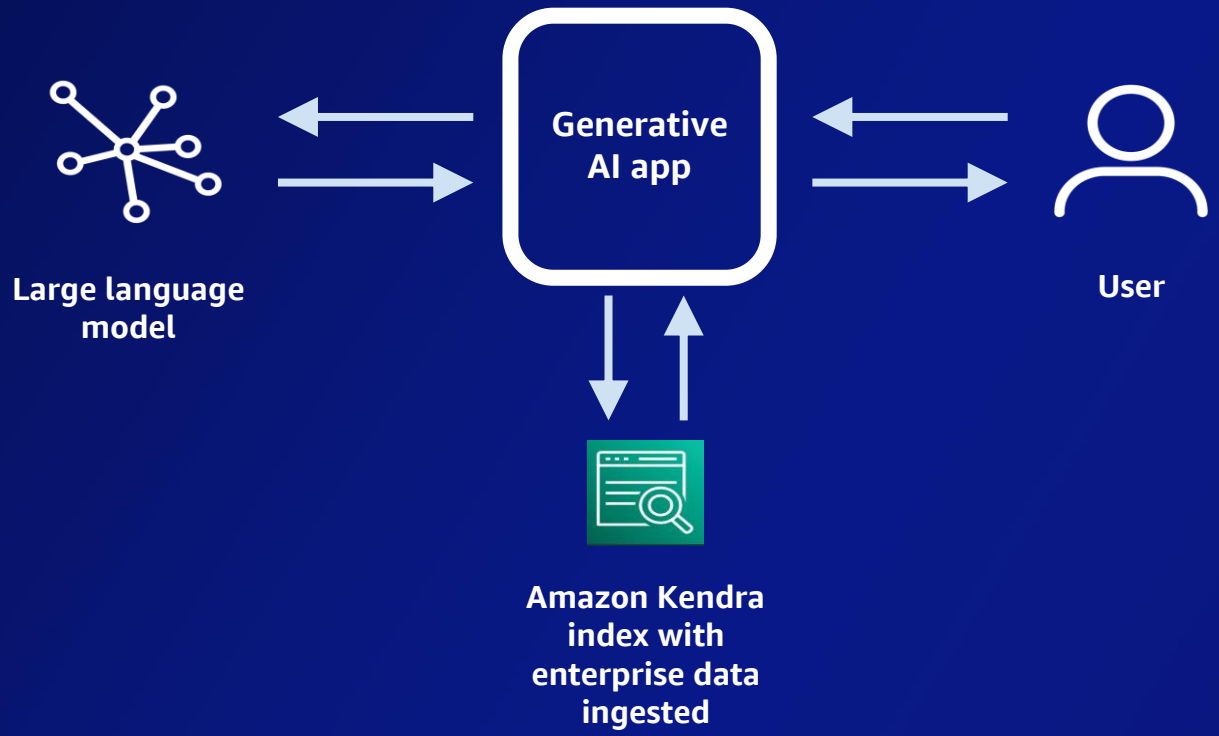
(3) Give example inputs & outputs
(optional, but improves formatting and accuracy)

(4) Provide the specific input to process

(5) Demonstrate the output formatting you want; deal with Claude’s natural chattiness by asking it to enclose the output in a specific format

(6) “\n\nAssistant:“ plus optional preamble to response

Retrieval



(1) **User**: What is denser, gold or tungsten?

(2) **Claude**: <thinking> First I need to find out how dense gold is. </thinking>

(3) **Claude**: <search_call> Gold density </search_call>

(4) **Amazon Kendra**: 19.3g/cm^3

(5) **Claude**: <thinking> Now I need to find out how dense tungsten is. </thinking>

(6) **Claude**: <search_call> Tungsten density </search_call>

(7) **Amazon Kendra**: 19.25g/cm^3

(8) **Claude**: <thinking> Now I have all the information I need to answer the question </thinking>

(9) **Claude**: <answer> Gold, with a density of 19.3g/cm^3, is very slightly denser than tungsten, which has a density of 19.25g/cm^3 </answer>

Try Claude 2 on Amazon Bedrock today

aws.amazon.com/bedrock

Amazon Bedrock deep dive

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

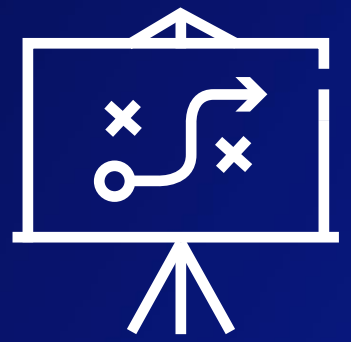
- Built with 20+ years of Amazon ML experience
- Automate language tasks such as summarization and text generation with Amazon Titan Text FM
- Enhance search accuracy and improve personalized recommendations with Amazon Titan Embeddings FM
- Support responsible use of AI by reducing inappropriate or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

Secure customization to drive differentiation

Driving innovation with Amazon Bedrock

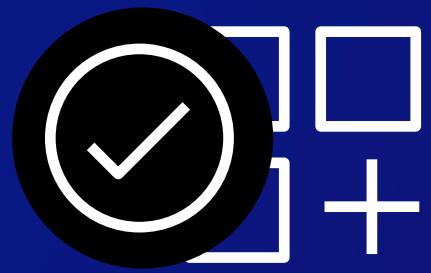
It takes a lot to get FMs to complete tasks

FOUNDATION MODELS ALONE CANNOT PERFORM TASKS

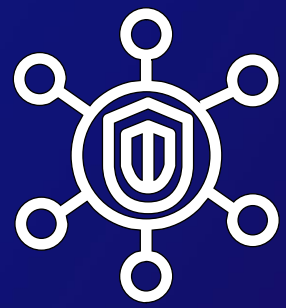


DEFINE INSTRUCTIONS AND ORCHESTRATION

CONFIGURE FM TO ACCESS DATA SOURCES



COMPLETE ACTIONS WITH API CALLS



MANAGE CLOUD HOSTING AND SECURITY

NEW

Agents for **Amazon Bedrock**

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

IN PREVIEW TODAY

Agents enable FMs to complete tasks

Breaks down and orchestrates tasks

Securely accesses and retrieves company data

Takes action by executing API calls on your behalf

Provides fully managed infrastructure

Set up agents for Amazon Bedrock in a few simple steps

1

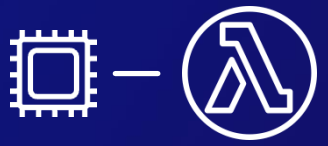
SELECT YOUR FOUNDATION MODEL

2

PROVIDE BASIC INSTRUCTIONS

3

SELECT RELEVANT DATA SOURCES



4

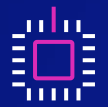
DEVELOPER SPECIFIES LAMBDA FUNCTIONS

Agents

for Amazon Bedrock

Tools to build with generative AI on AWS

The easiest way to build with FMs



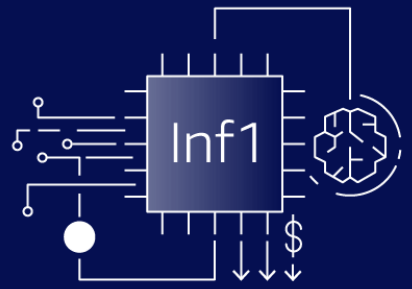
The most price-performant infrastructure

Generative AI-powered applications

Selection of open-source FMs

Purpose-built accelerators for generative AI

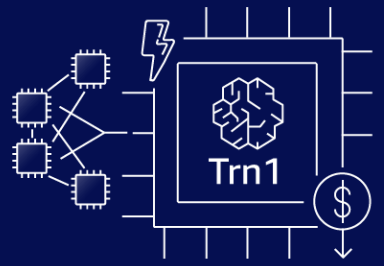
AWS Inferentia



Lowest cost per inference in the cloud for running deep learning (DL) models

Up to 70% lower cost per inference
than comparable Amazon EC2 instances

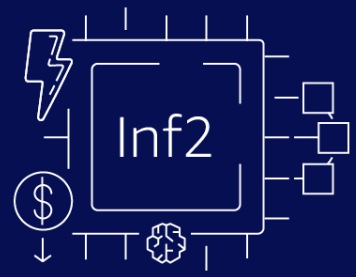
AWS Trainium



The most cost-efficient, high-performance training of LLMs and diffusion models

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

AWS Inferentia2



High performance at the lowest cost per inference for LLMs and diffusion models

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

Strong relationship with NVIDIA

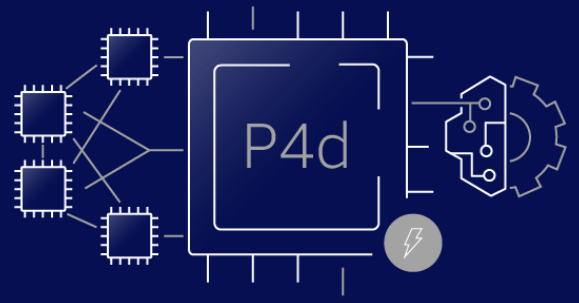
Amazon EC2 P5 instances

Coming soon!

Powered by NVIDIA H100 Tensor Core GPUs

Up to 6x faster and up to 40% cost-to-train savings
than previous generation GPU-based instances

Amazon EC2 P4d/P4de instances



Powered by NVIDIA A100 Tensor Core GPUs

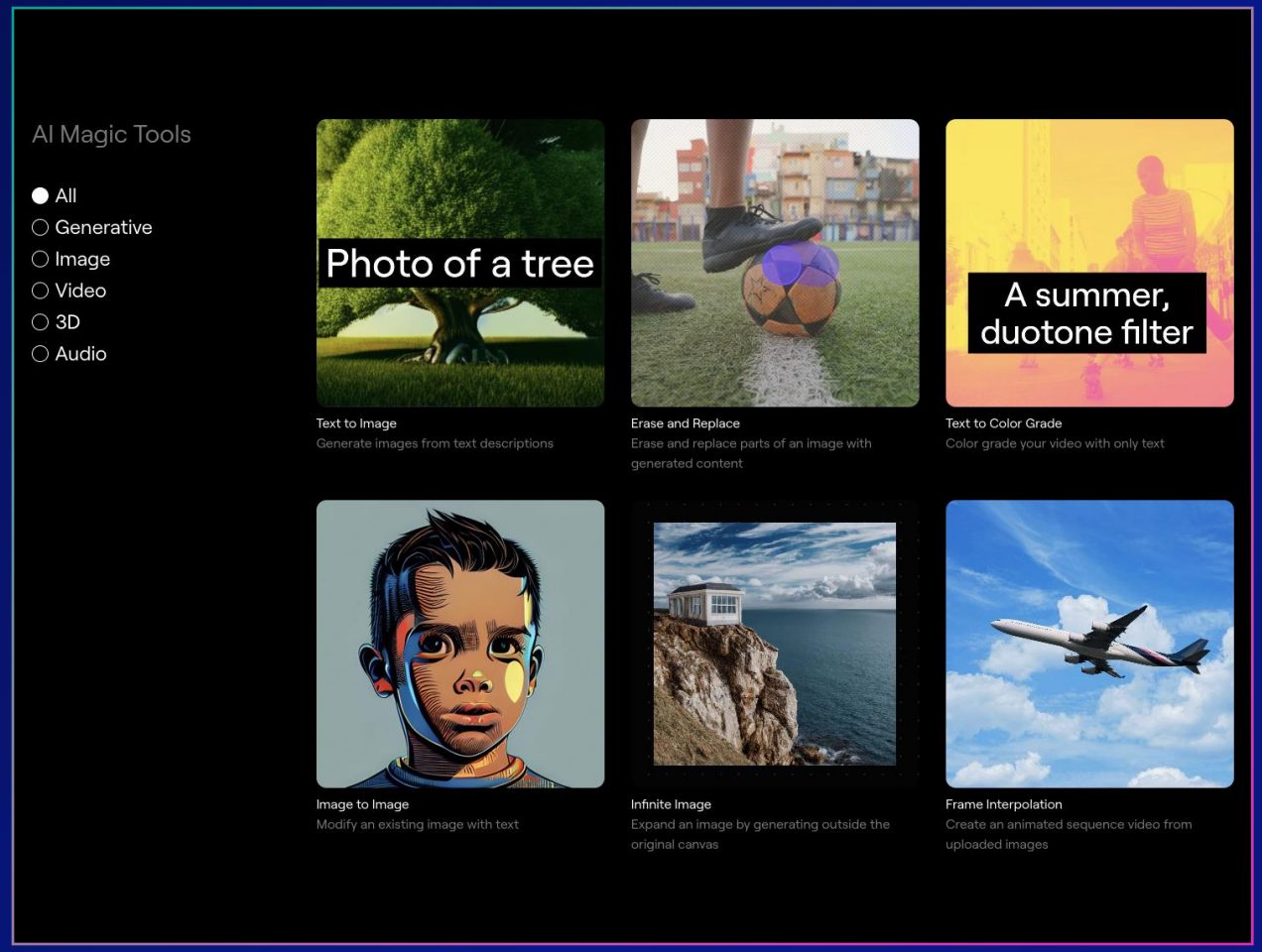
Up to 2.5x faster and up to 60% lower training costs
than previous generation P3/P3dn instances

Amazon EC2 G5 instances

Powered by NVIDIA A10G Tensor Core GPUs

Up to 3.3x higher performance
than previous generation G4dn instances

Runway delivers a better creator experience with AWS Inferentia2



"At Runway, our suite of AI Magic Tools enables our users to generate and edit content like never before. With Amazon EC2 Inf2 instances, we're able to run some of our models with up to 2x higher throughput. This enables us to introduce more features, deploy more complex models, and ultimately deliver a better experience for the millions of creators using Runway."

Cristobal Valenzuela, CEO

runway

Tools to build with generative AI on AWS

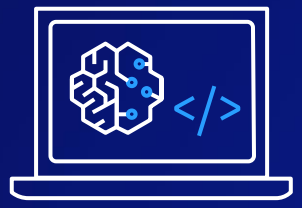
The easiest way to build with FMs

The most price-performant infrastructure

Generative AI-powered applications

Selection of open-source FMs

Amazon CodeWhisperer: Now generally available and free to use for individual developers



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

During preview Amazon ran a productivity challenge, and participants who used Amazon CodeWhisperer were 27% more likely to complete tasks successfully and did so an average of 57% faster than those who did not use Amazon CodeWhisperer

Accenture improves developer productivity with Amazon CodeWhisperer

"Accenture is using Amazon CodeWhisperer to accelerate coding as part of our software engineering best practices initiative in our Velocity platform. The Velocity team was looking for ways to improve developer productivity. After searching for multiple options, we came across **Amazon CodeWhisperer, which reduced our development efforts by 30%.** We are now focusing more on improving security, quality, and performance."

Balakrishnan Viswanathan

Tech Architecture Sr. Manager, Accenture

Tools to build with generative AI on AWS

The easiest way to build with FMs

The most price-performant infrastructure

Generative AI-powered applications

Selection of open-source FMs

aws

Build with publicly available FMs

AVAILABLE ON SAGEMAKER JUMPSTART

Meta AI

Models

Llama 2 7B, 13B, 70B

Tasks

Question answering
Chat
Summarization
Paraphrasing
Sentiment analysis
Text generation

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
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

databricks

Models

Dolly

Tasks

Question answering
Chat
Summarization
Paraphrasing
Sentiment analysis
Text generation

AI21 labs

Models

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

Tasks

Text generation
Long-form generation
Summarization
Paraphrasing
Chat
Information extraction

co:here

Models

Cohere
Command XL

Tasks

Text generation
Information extraction
Question answering
Summarization

LightOn

Models

Lyra-Fr 10B, Mini

Tasks

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

alexa

Models

AlexaTM 20B

Tasks

Machine translation
Question answering
Summarization
Annotation
Data generation

AI21 Labs built its FM faster with Amazon SageMaker

AI21 labs

CHALLENGE

AI21 Labs builds large language models. These are huge neural networks with tens to hundreds of billions of parameters and training them is a major challenge due to complexity and resource needs.

SOLUTION

AI21 Labs trained the Jurassic-Grande model with 17 billion parameters using Amazon SageMaker. Amazon SageMaker made the model training process easier and more efficient while working perfectly with DeepSpeed library.

OUTCOME

- ✓ Easier and more efficient model training
- ✓ Ability to scale the distributed training jobs easily to hundreds of NVIDIA A100 GPUs
- ✓ Lower inference costs

aws

Start your generative AI journey today

1

Start driving productivity with Amazon CodeWhisperer today

2

Explore FMs through Amazon Bedrock and other FMs on SageMaker JumpStart

3

Get started on a PoC for your top use cases

skillbuilder.aws

Your time is now

Build in-demand cloud skills your way

Thank you!

Please complete the session survey in the mobile app

Atul Deo
twitter: @atuldeo

Matt Bell
LinkedIn: /thismattbell