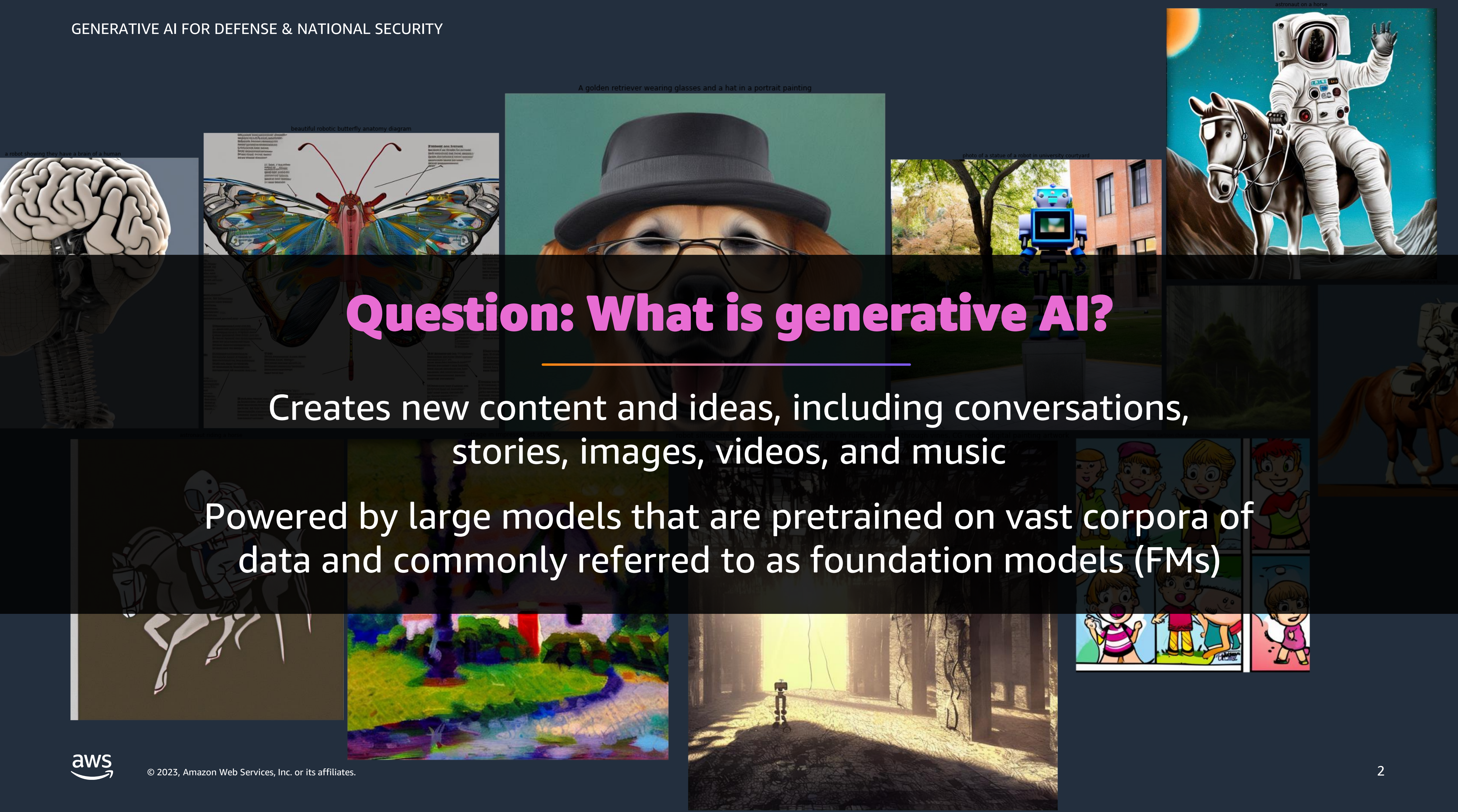
Generative AI

Generative AI Innovation Center

Frank Tanner

Manager, Generative AI Innovation Center
aifrank@amazon.com


a robot showing they have a brain of a human
beautiful robotic butterfly anatomy diagram
A golden retriever wearing glasses and a hat in a portrait painting
photo of a statue of a robot in university courtyard
astronaut on a horse

Question: What is generative AI?

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

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

Unlocking the potential of generative AI

- The easiest way to build with FMs
- The most price-performant infrastructure
- Generative AI-powered applications
- Flexibility to build with your own FMs

Amazon Bedrock

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

Benefits

- Accelerate development of generative AI applications using FMs through an API
- No need to manage infrastructure
- Choice of FMs from AI21 Labs, Anthropic, Cohere, Stability AI, and Amazon
- Privately customize FMs using your organization's data
- Comprehensive AWS security capabilities
- NEW: Enable generative AI apps to complete tasks in just a few clicks using agents for Amazon Bedrock

Amazon Bedrock supports leading FMs

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 Dutch, French, German, Italian, Portuguese, and Spanish

ANTHROP\C

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

More models on Amazon SageMaker JumpStart*

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

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

OpenLLaMA

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

* New models added on a weekly basis

Takeaway for BENGAL Performers

Wide variety of models & providers

ANTHROP\C

Hugging Face

AI21labs

cohere

Meta AI

databricks

LightOn

stability.ai

amazon

Managed infrastructure

Full control of your model training with managed and most price-performant infrastructure

Governance

Purpose-built governance tools to help you responsibly implement ML

Efficient distributed training

Complete distributed training up to 40% faster

Debugging and experimentation tools

Capture metrics and profile training jobs in real time to quickly correct performance issues

Track ML model iterations

Price-performant inference

Deploy models in production for any use case with best price-performance

Thank you!

Backup Content

Comprehensive data protection and privacy

Your data used with Amazon Bedrock is not used for service improvement and not shared with third-party model providers

Private connectivity between Amazon Bedrock service and your virtual private cloud (VPC)

Your data is encrypted in transit and at rest

Privately customize FMs, retaining control over how your data is used and encrypted

Support for governance and auditability

Comprehensive monitoring and logging capabilities

Track usage metrics and build customized dashboards using Amazon CloudWatch

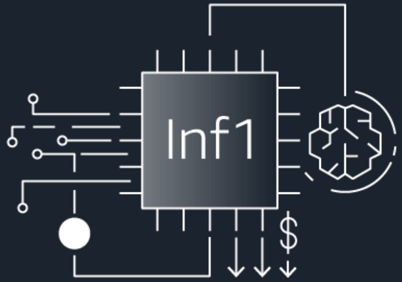
Monitor API activity and troubleshoot issues as you integrate other systems into your applications using AWS CloudTrail

Unlocking the potential of generative AI

- The easiest way to build with FMs
- The most price-performant infrastructure
- Generative AI-powered applications
- Flexibility to build with your own FMs

Purpose-built accelerators for generative AI

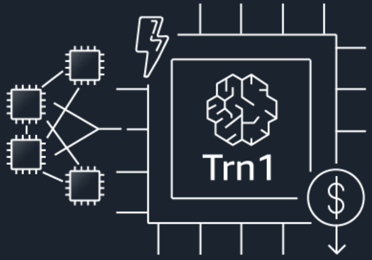
AWS Inferentia



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

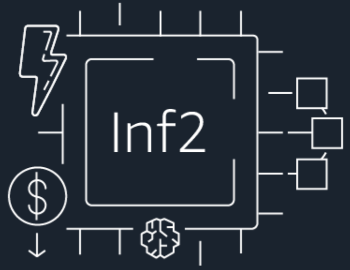
AWS Trainium



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

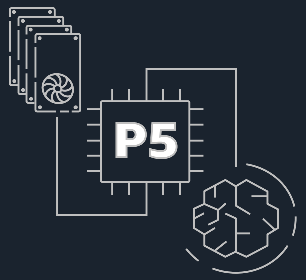


High performance at the lowest cost per inference for LLMs and diffusion models

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

Unmatched experience delivering GPU-based compute resources

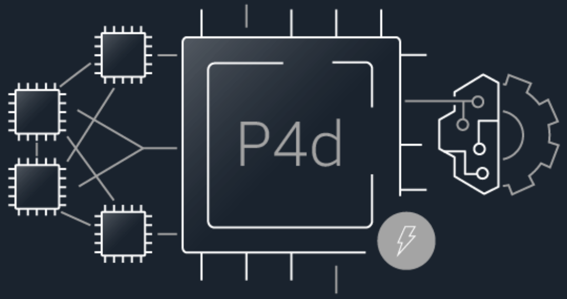
Amazon EC2 P5 instances



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 and 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 and P3dn instances

Amazon EC2 G5 instances

Powered by
NVIDIA A10G Tensor
Core GPUs

Up to 3.3x higher performance than previous-generation G4dn instances