Build with generative AI on AWS

Vienna Data Science tools meetup

VIENNA | 21 SEPTEMBER 2023

Paolo Di Francesco, PhD
Sr Solutions Architect
AWS

What we'll talk about today

- An overview of AIML in AWS
- An overview of foundation model hosting options on AWS
- Adding your data to generative AI
- Tips and resources

The AWS AI/ML stack

AI SERVICES

GENERATIVE AI

preview

AMAZON BEDROCK
Foundation Models

AMAZON CODEWHISPERER
AI Coding Companion

SPECIALIZED

BUSINESS PROCESSES
- Amazon Personalize
- Amazon Forecast
- Amazon Fraud Detector
- Amazon Lookout for Metrics

SEARCH
- Amazon Kendra

CONVERSATION
- Amazon Lex
- Amazon Transcribe Call Analytics
- Contact Lens
- Voice ID

CODE + DEVOPS
- Amazon CodeGuru
- Amazon DevOps Guru

INDUSTRIAL
- Amazon Monitron
- Amazon Lookout for Equipment
- Amazon Lookout for Vision

HEALTH
- Amazon HealthLake
- Amazon Comprehend Medical
- Amazon Transcribe Medical
- Amazon Omics

CORE

TEXT
Amazon Translate | Amazon Comprehend

SPEECH
Amazon Polly | Amazon Transcribe

VISION
Amazon Textract | Amazon Rekognition | AWS Panorama

AMAZON SAGEMAKER

CANVAS
No-code ML for business analysts

STUDIO LAB
Learn ML

GROUND TRUTH
Label data

STUDIO IDE

Prepare data | Geospatial ML | Build with notebooks | Train models | Deploy in production

Store features | Tune parameters | Manage and monitor

CI/CD | GOVERNANCE | RESPONSIBLE ML

JUMPSTART
Catalog of models and solutions

ML FRAMEWORKS & INFRASTRUCTURE

PyTorch, Apache MXNet, TensorFlow | Amazon EC2 | CPUs | GPUs | AWS Inferentia | AWS Trainium | Habana Gaudi | FPGA

aws

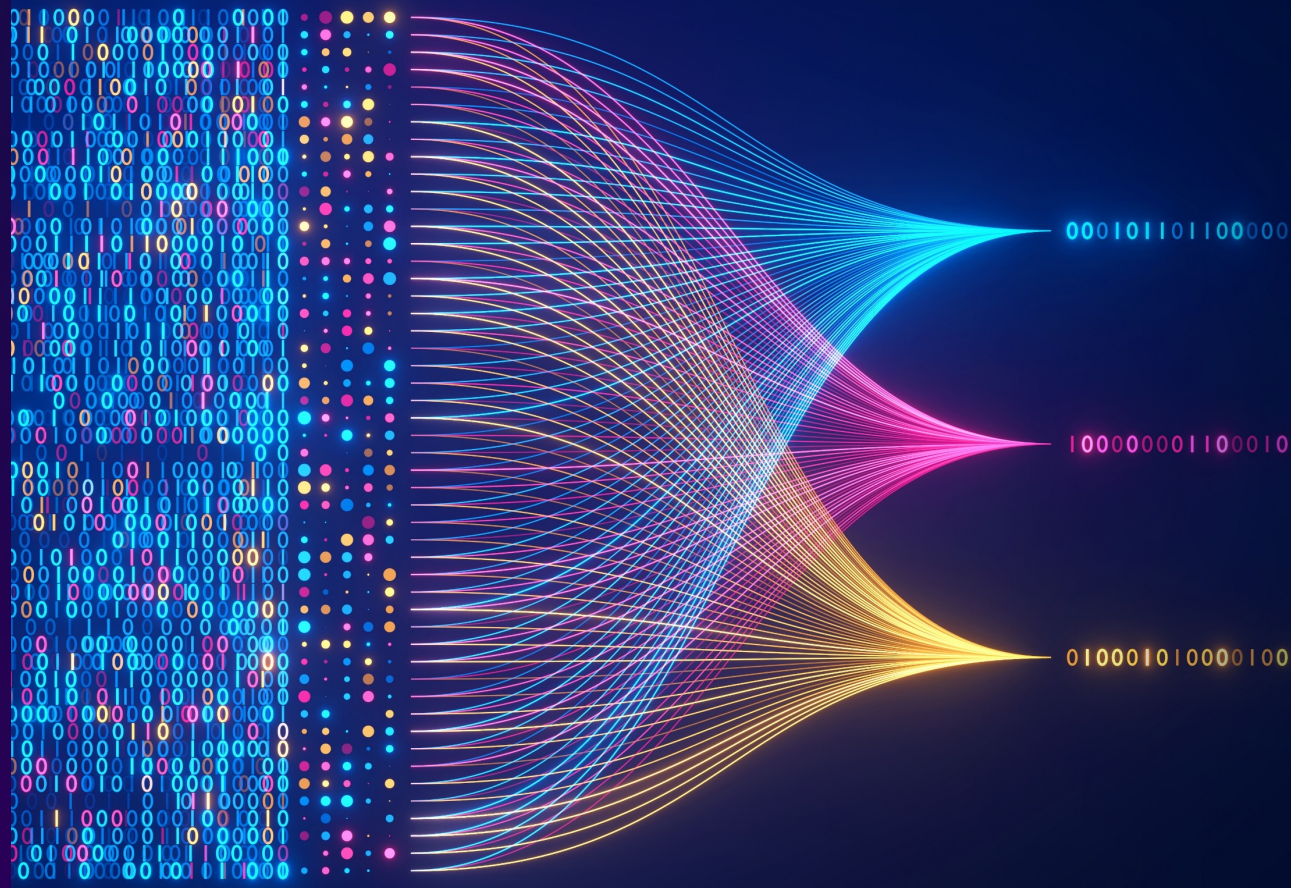
Generative AI is powered by FMs

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



Generative AI is everywhere

Chatbots

Virtual assistants

AI-powered Contact Center

Personalization

Conversational search

Summarization

Code generation

Data to insights

Writing

Media design

Modeling

Document processing

Process optimization

Cybersecurity

Data augmentation

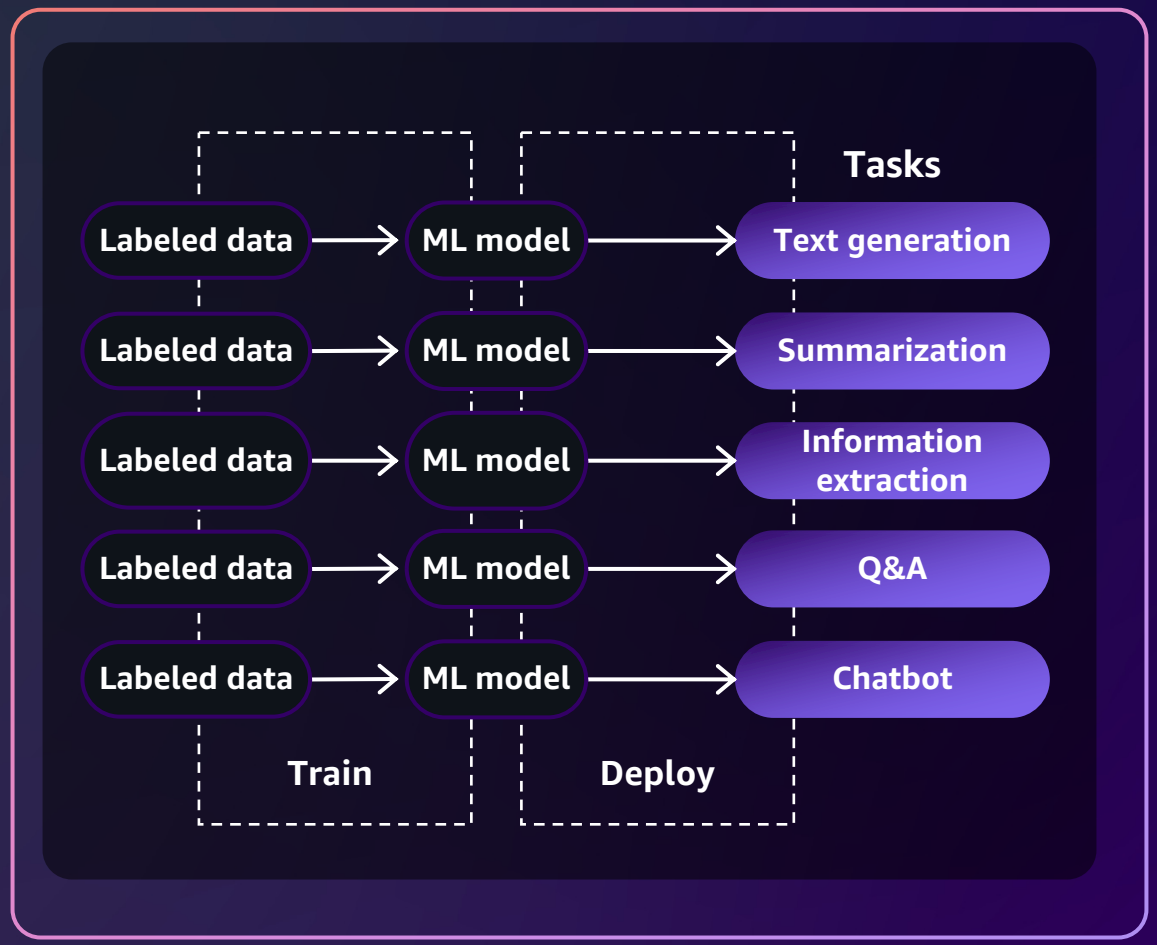
Enhance customer experience

Boost employee productivity

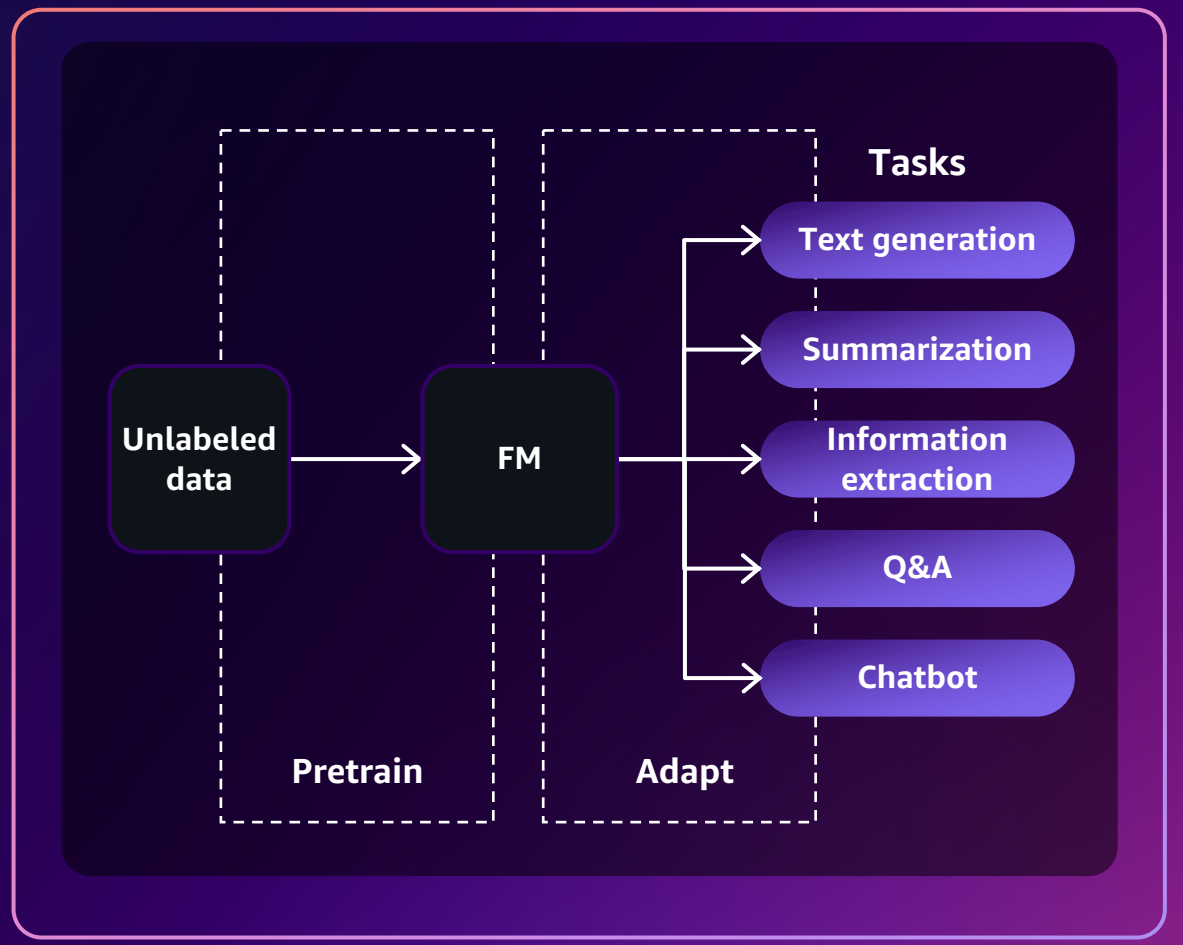
Creativity and content creation

Improve business operations

How FMs differ from other ML models

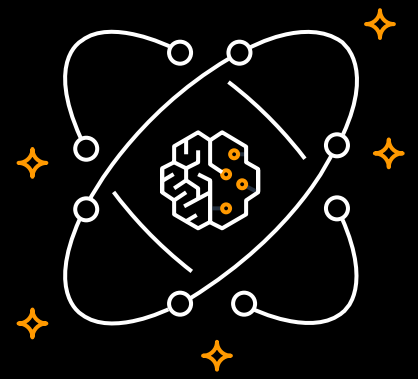


Traditional ML models



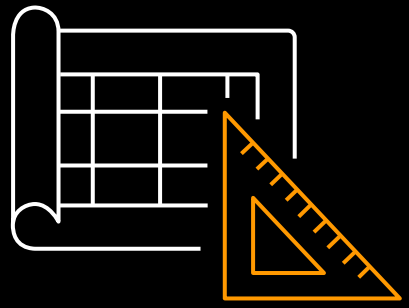
Foundation models

3 ways to leverage Foundation Models



Use AI service APIs or deploy publicly available models

Adapt pre-trained models



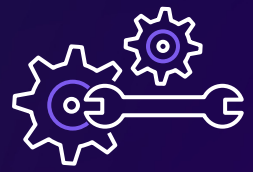
Train your own model

Hosting foundation models on AWS

Amazon Bedrock preview

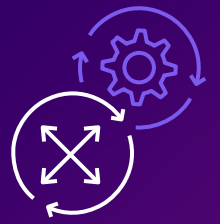
Easy to use, “Serverless” experience

Select foundation models from: AI21 Labs, Anthropic, Stability AI, Cohere and Amazon



Privately customize FMs using your data

Comprehensive AWS security capabilities



Build scalable, reliable, and secure generative AI applications

Amazon Bedrock preview

Choice of foundation models

AI21labs

JURASSIC-2

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

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

NEW

CLAUDE 2

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

amazon

AMAZON TITAN

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

aws

Amazon SageMaker JumpStart

Amazon SageMaker JumpStart

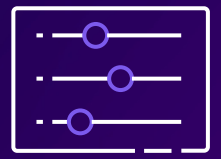
Foundation Model Hub

Browse

Browse public and proprietary FMs

Experiment

Experiment with FMs before choosing a model for deployment



Customize

Customize with your own dataset without training from scratch

Deploy

Deploy the model and run inference

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

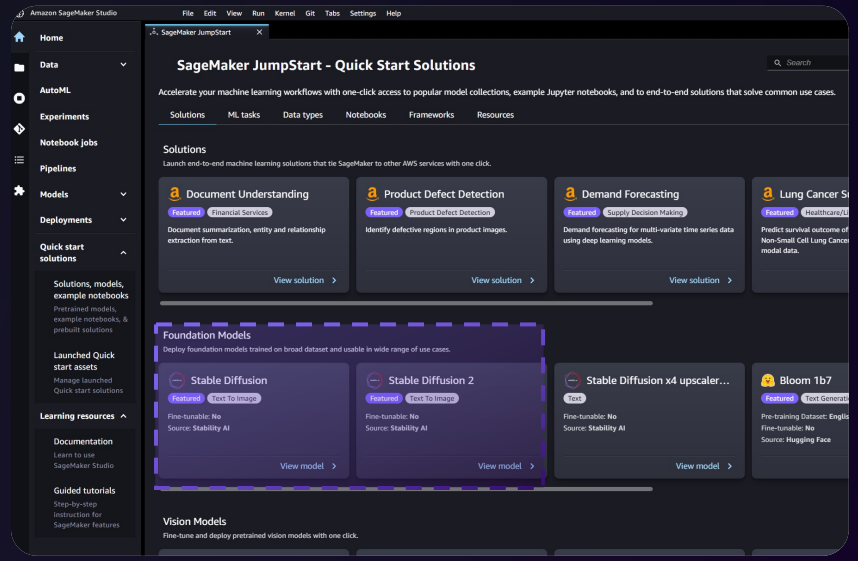
Annotation

Data generation

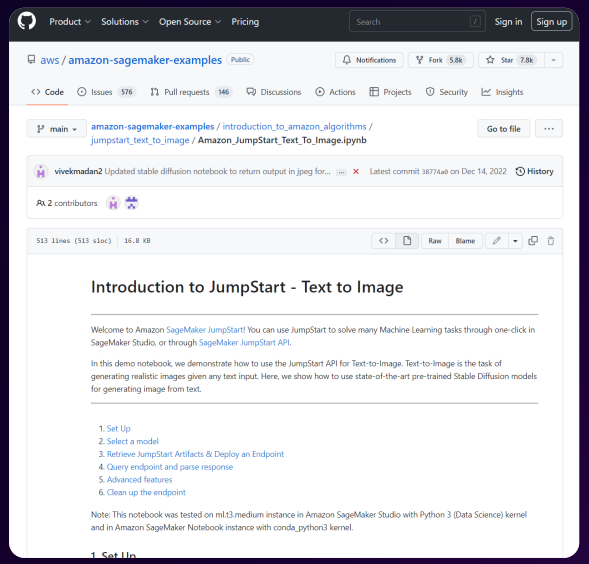
aws

3 ways to use foundation models with SageMaker JumpStart

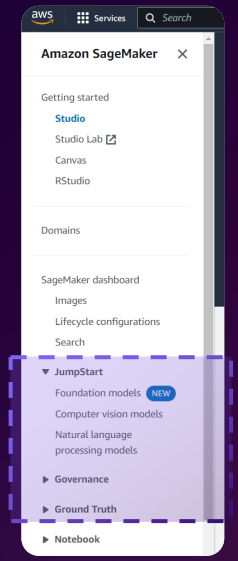
SageMaker Studio
One-step deploy



SageMaker Notebooks

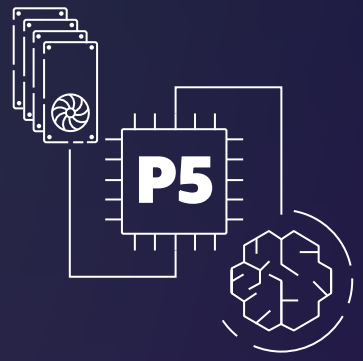


AWS Management Console Preview



Hardware choice

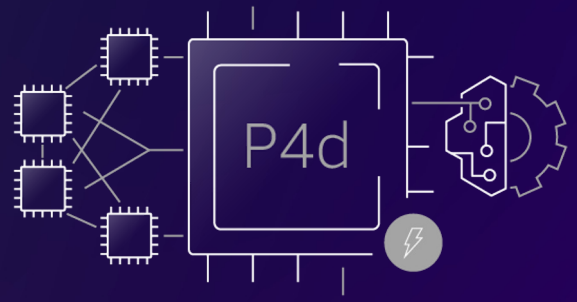
Amazon EC2 P5 instances



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/P4de instances



Powered by NVIDIA A100 Tensor Core GPUs

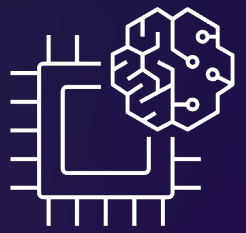
Up to 2.5x faster and up to 60% lower training costs
than previous generation P3/P3dn instances

Amazon EC2 G5 instances

Powered by NVIDIA A10G Tensor Core GPUs

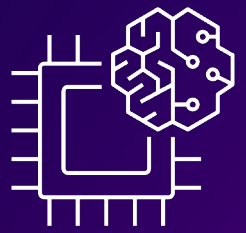
Up to 3.3x higher performance
than previous generation G4dn instances

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

bit.ly/3PIs5Ms

Which option should I choose?

Amazon Bedrock

"I want to build GenAI apps easily!"

- Easiest option: Simple API
- Serverless experience
- Growing model selection
- Simple fine-tuning

Amazon SageMaker JumpStart

"I want to start simple, with more customization options"

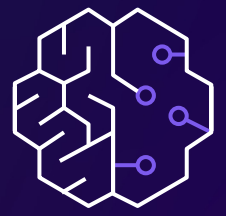
- Easy setup on Amazon SageMaker
- Self-managed endpoints
- More models available
- Fine-tuning supported

Self-hosted EC2 / Containers

"I want full control over everything"

- If needed
- Self-managed everything
- But: need to roll your own
- Beware of cost and complexity

Five Examples to make FMs work for you



Foundation Model

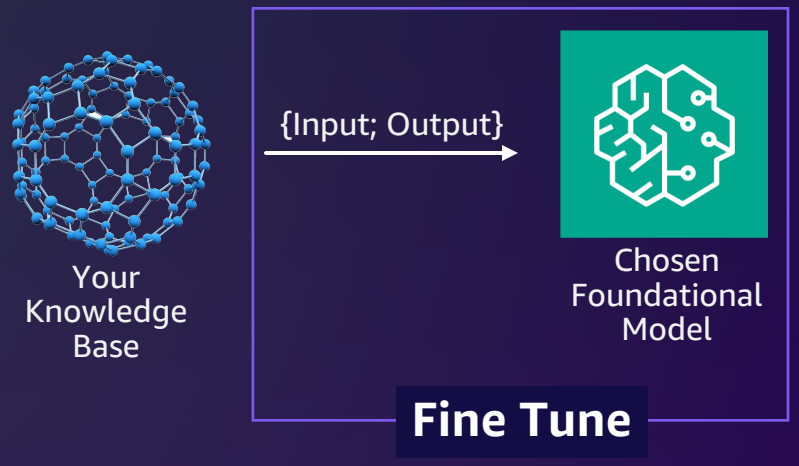
- **Prompt Engineering**
- **Fine-Tuning**
- **Retrieval Augmented Generation**
- **Modular Reasoning**
- **Content Moderation**

Your data is
your differentiator

Fine tuning or RAG?

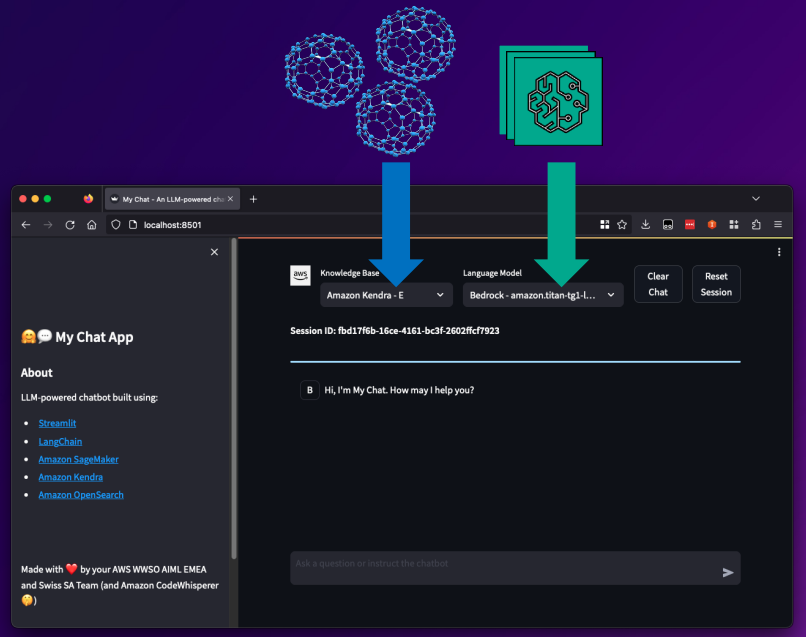
Prepare (Optional)

Fine tune with pairs of inputs and outputs (from 40 to a few hundreds)



Run

Giving the model more of your data for answering questions, aka **Retrieval Augmented Generation (RAG)**



- User asks a question
- Model rephrases question to query knowledge base
- Knowledge base provides extracts
- Model uses extracts to craft answer

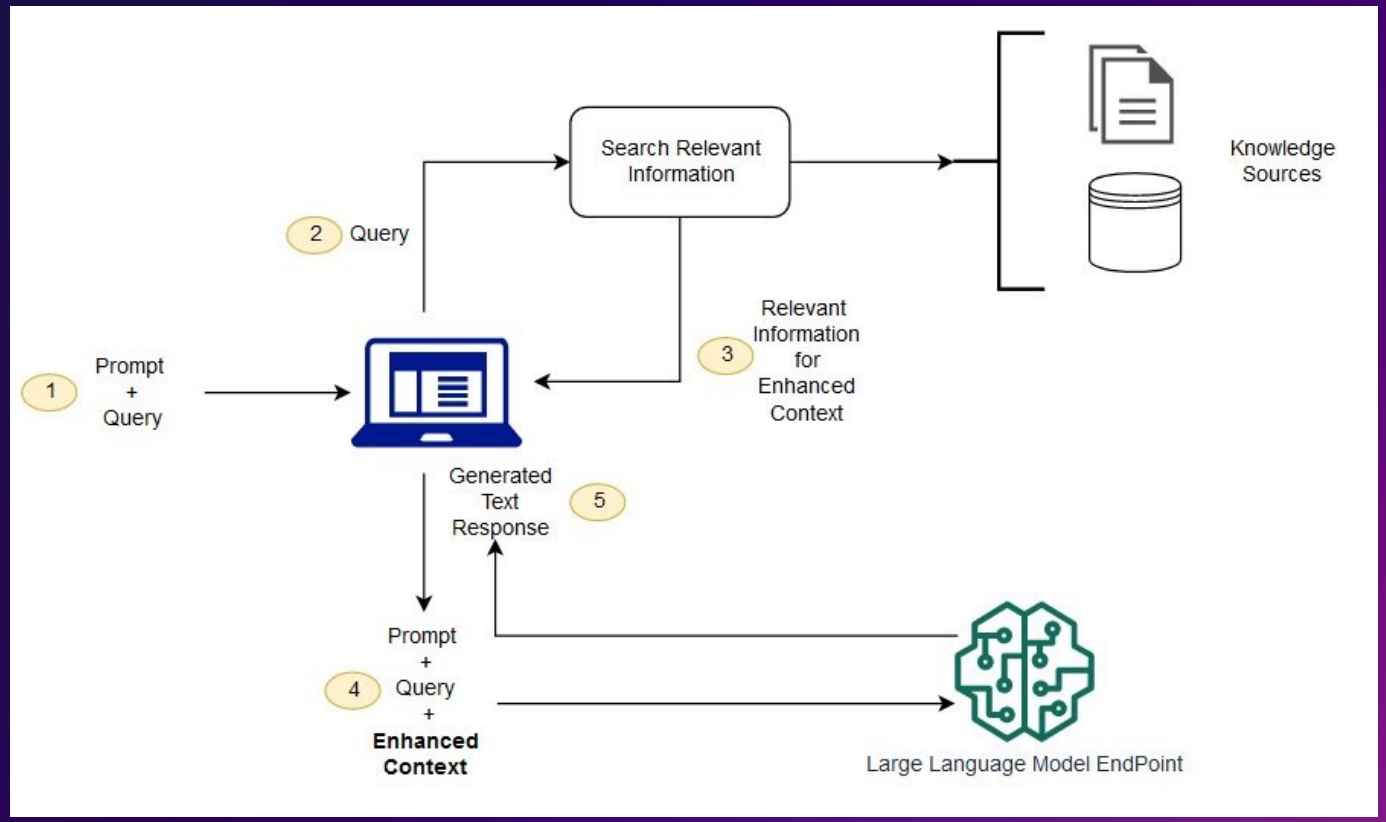
Try it with your data and the model that suits you best

RAG and large language models

- Split up documents into fragments
- Index them in a search engine or database
- Examples:
  - Facebook AI similarity search (FAISS)
  - Amazon OpenSearch Service
  - Amazon RDS PostgreSQL with pgvector extension
  - Amazon ElastiCache Redis
  - Amazon Kendra

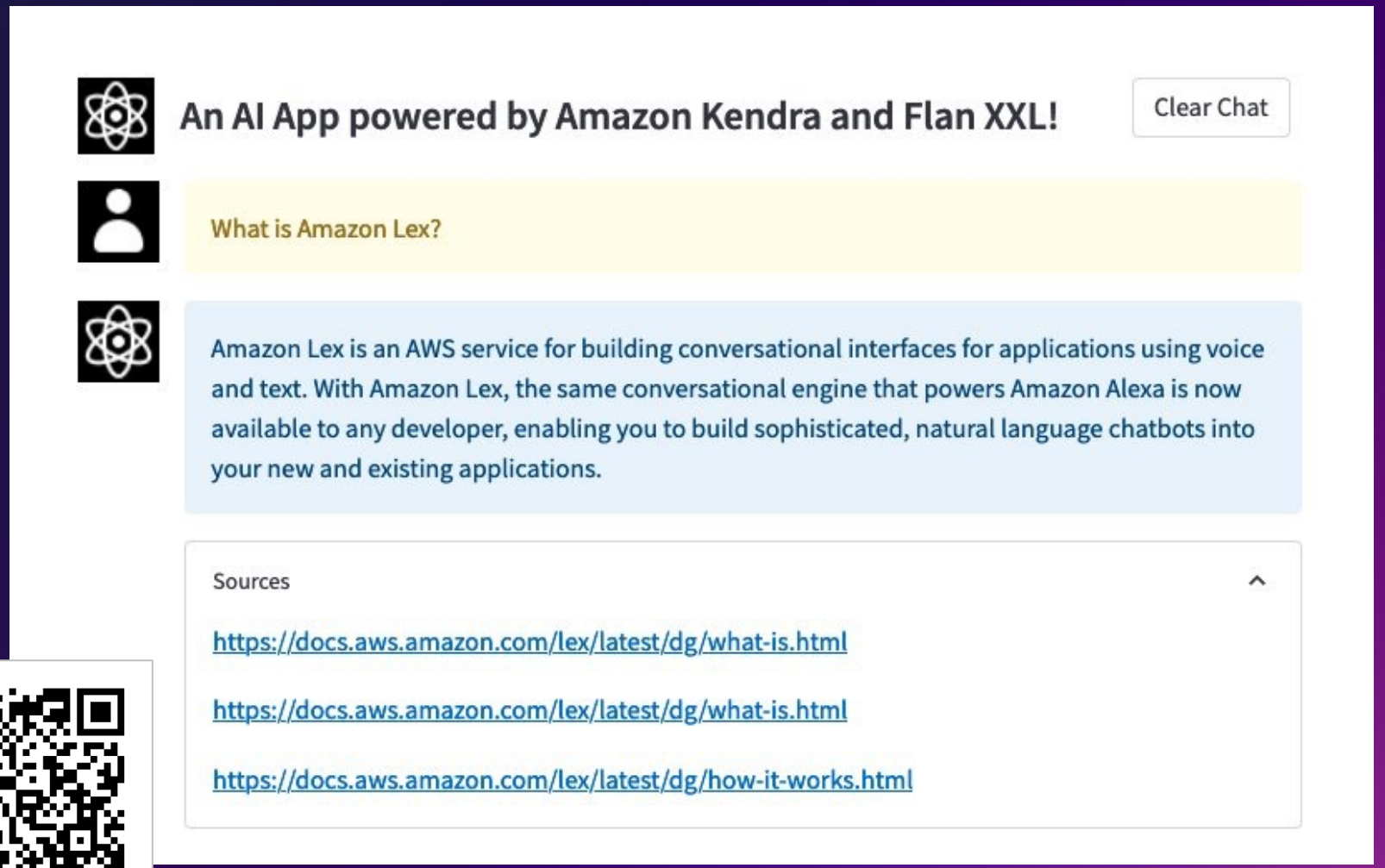
Sample notebooks

- Available in the Amazon SageMaker documentation
- Using JumpStart models and LangChain



bit.ly/3OKOvwc

RAG example with Amazon Kendra and Flan XXL

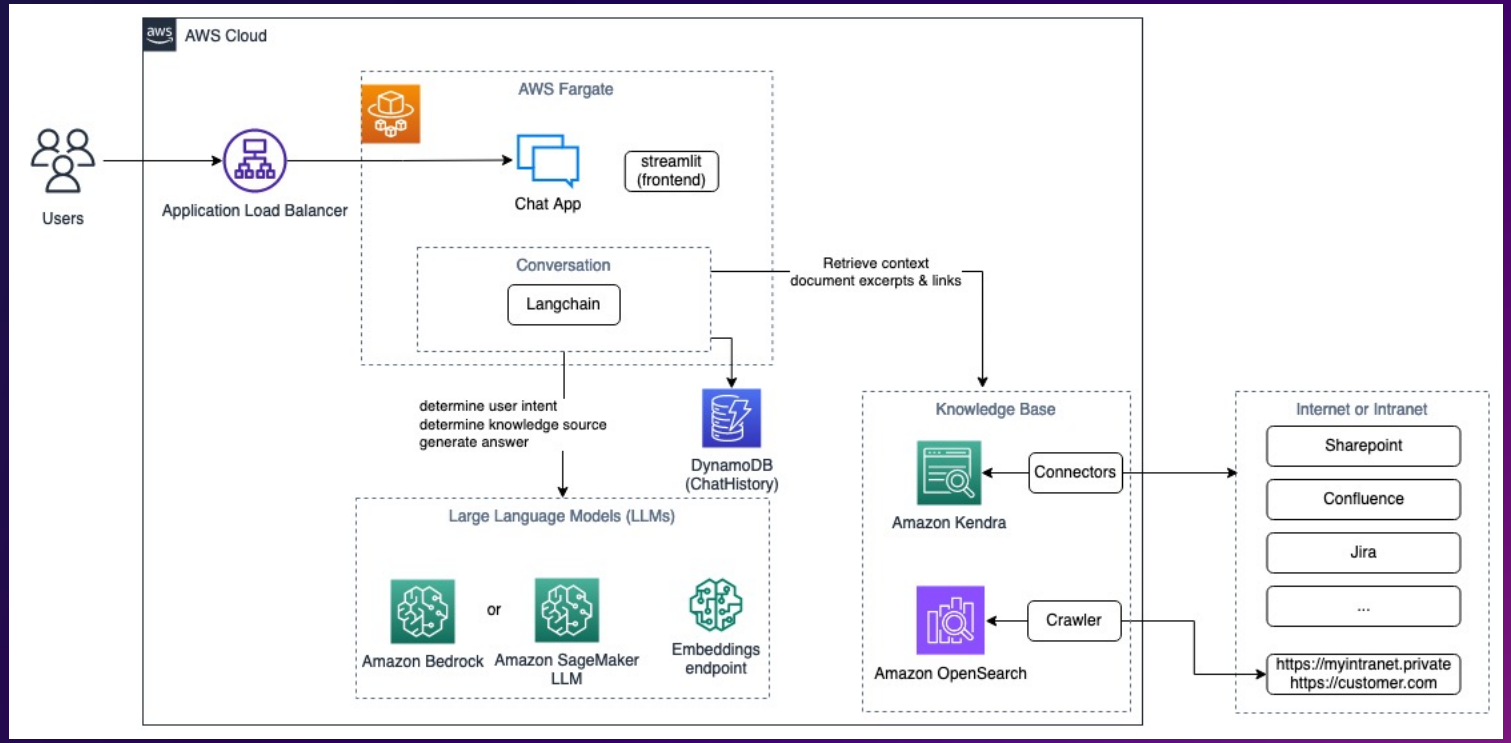


bit.ly/3OmpkP5

Harness the power of your data

A practical example*

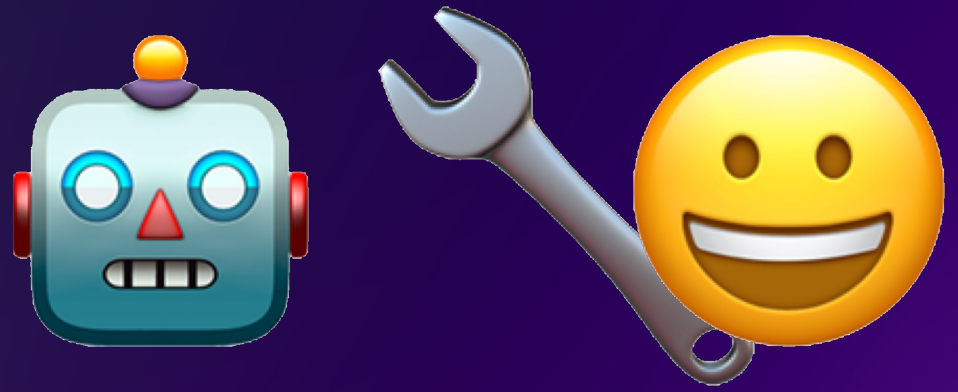
Main components:
- Streamlit app
  - Langchain connectors
- LLMs
  - SageMaker endpoints
  - Amazon Bedrock
- Knowledge Base
  - Amazon Kendra
  - OpenSearch + SageMaker Embedding endpoint



Solutions fully deployable via IaC

*repo not yet available in the public

Useful tools and resources

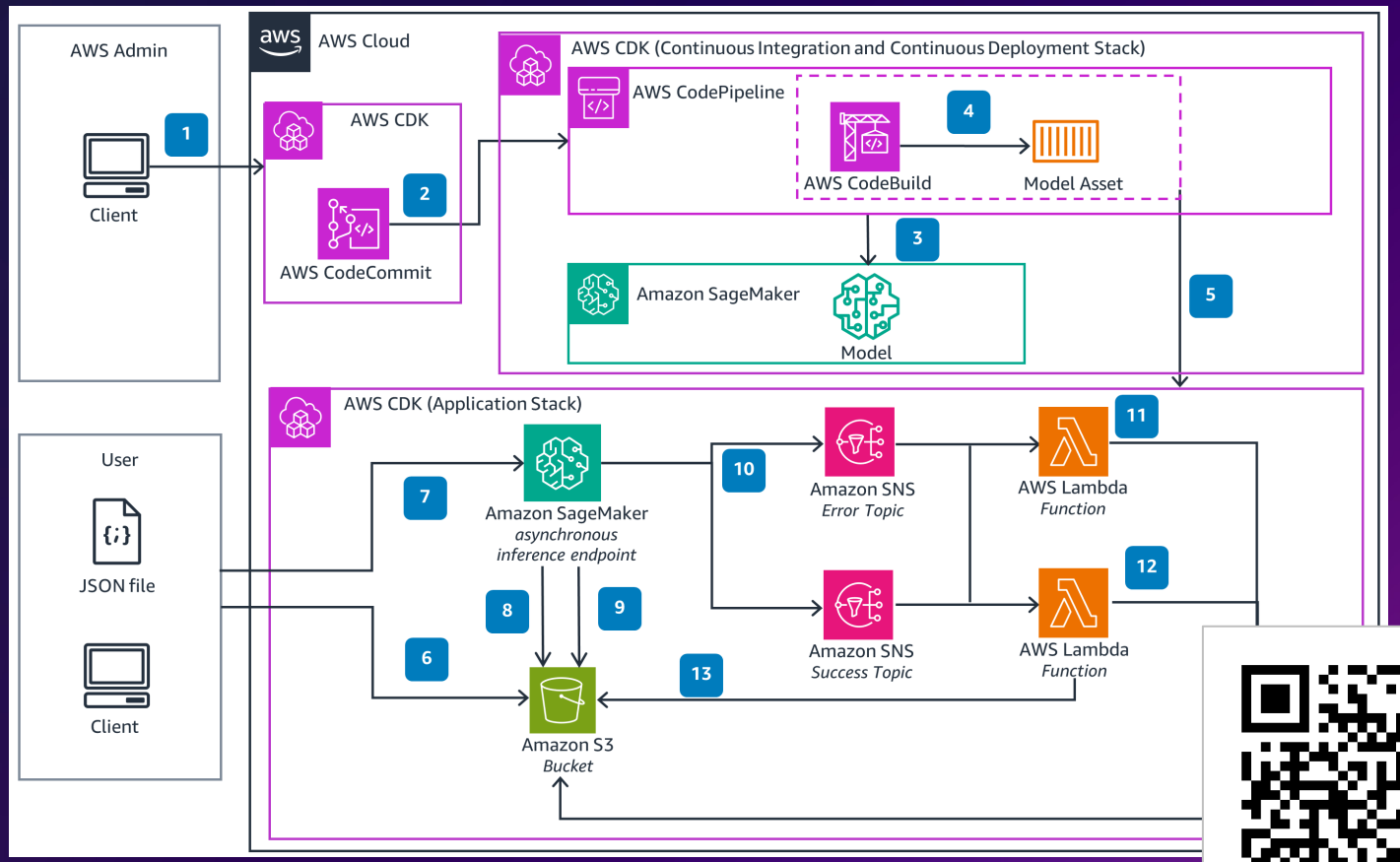


Useful tools and resources

- 🦜🔗 LangChain: flexible abstractions and extensive toolkit
-  Streamlit.io: fast and simple data app building
- AWS Machine Learning Blog
- Amazon.science

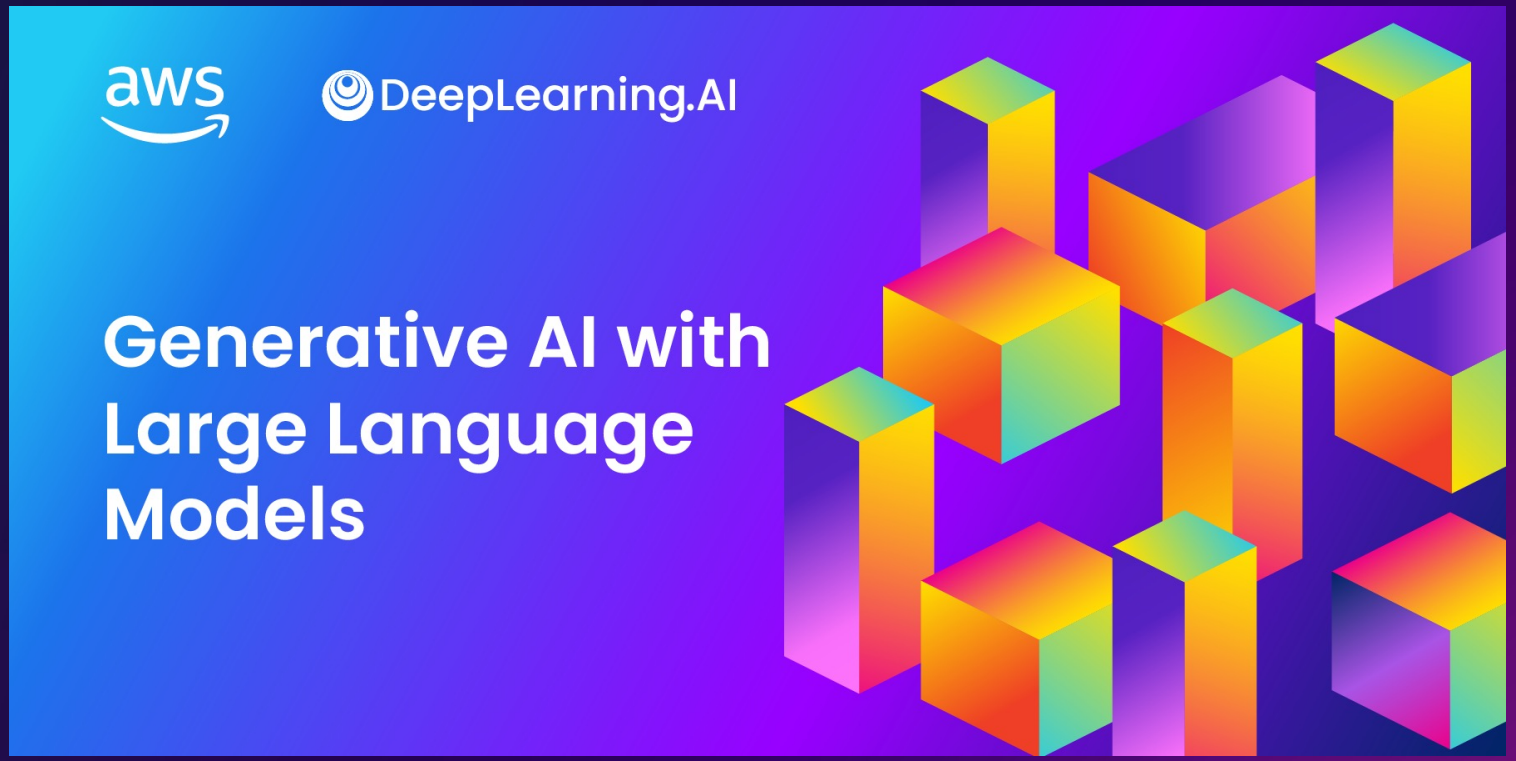
Deploy JumpStart models easily with AWS CDK

- Choose a JumpStart model
- Write one line of AWS CDK (Python)
- Deploy automatically
- CI/CD pipeline included
- go.aws/3KrQuTw



Learning tip!

- 3 weeks
- Hands-on
- On-demand
- Covering the full generative AI project lifecycle



go.aws/47aFAv4

Learning tip #2!

- 7 chapters, 8 hours
- Deep dive, hands-on
- Free
- Pre-train, fine-tune and deploy foundation models on AWS

bit.ly/3rTl1Db

Thank you!

Paolo Di Francesco

frpaolo@amazon.at
paolo-di-francesco