Democratizing Gen AI Large Language Model (LLM) Deployment: Amazon SageMaker JumpStart Unleashed

23th Oct, 2023

Said Nechab

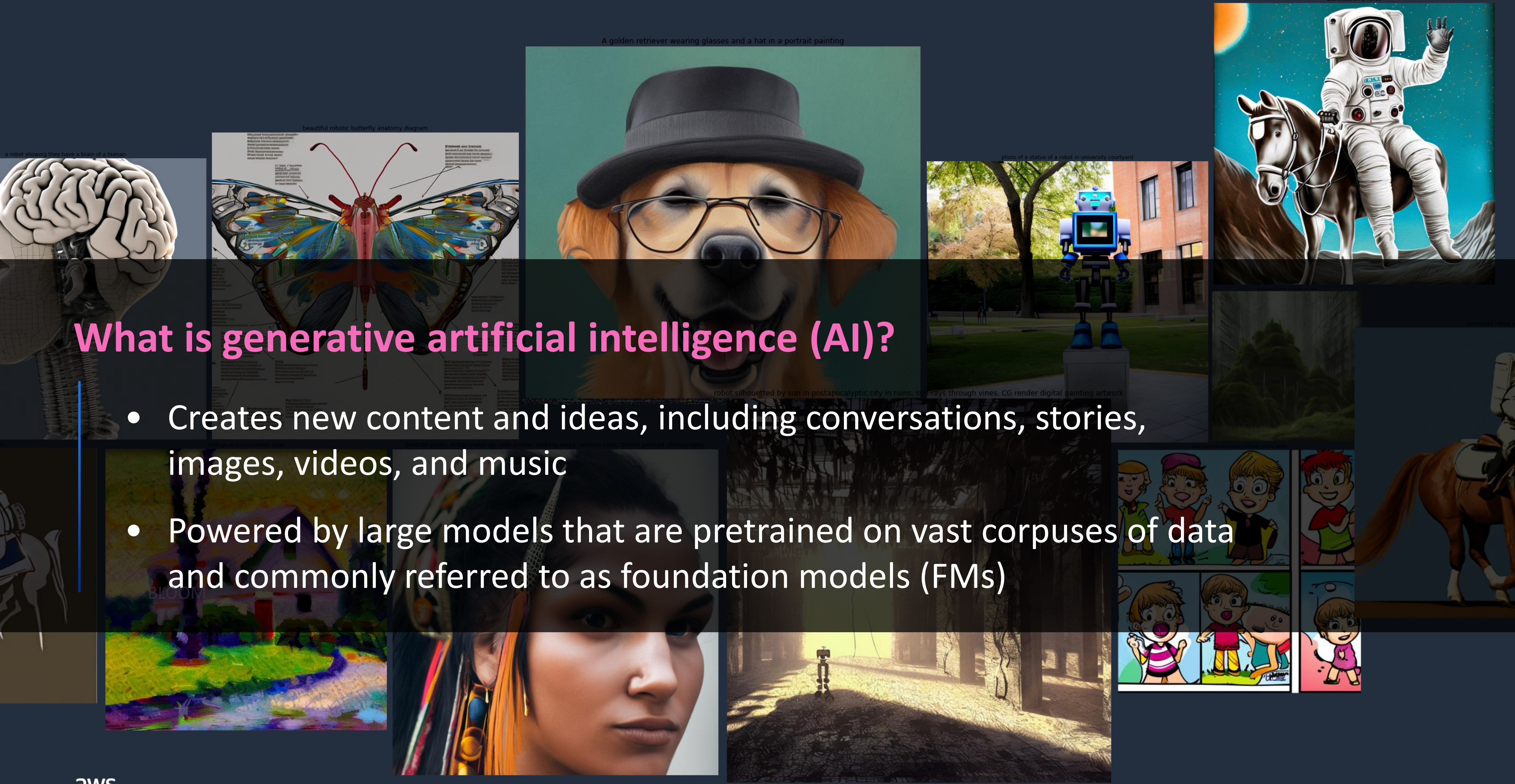
AWS Partner AI/ML Solution Architect

Ajit Kumar

AWS Partner AI/ML Solution Architect

Agenda

- Overview of Generative AI
- Generative AI offerings on AWS
- Overview of SageMaker JumpStart
- Customization of LLMs using SageMaker Jumpstart
- How to get started and CTA
- Exit Survey + Q&A



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)

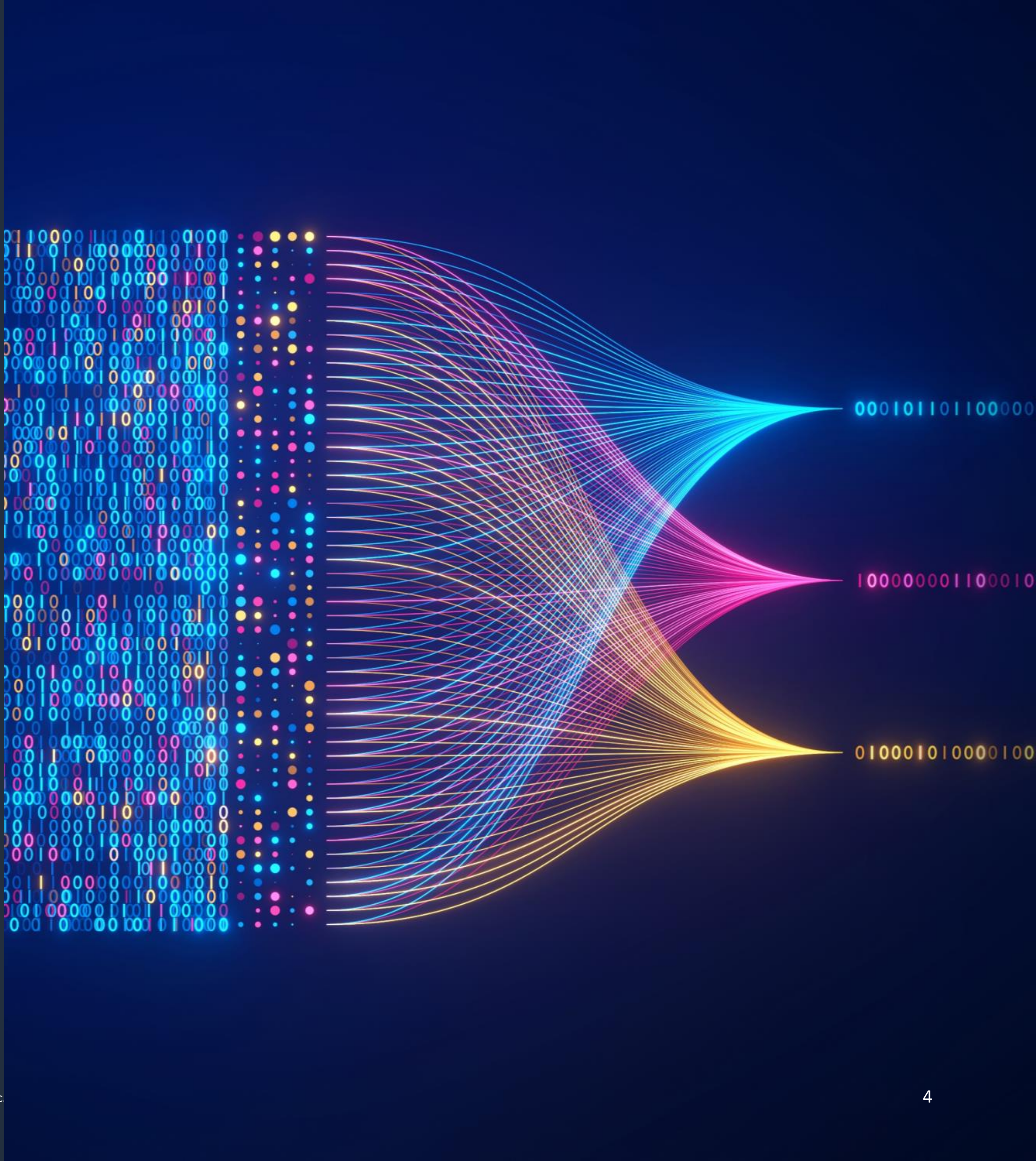
Generative AI is powered by foundation models

Pretrained on vast amounts of unstructured data

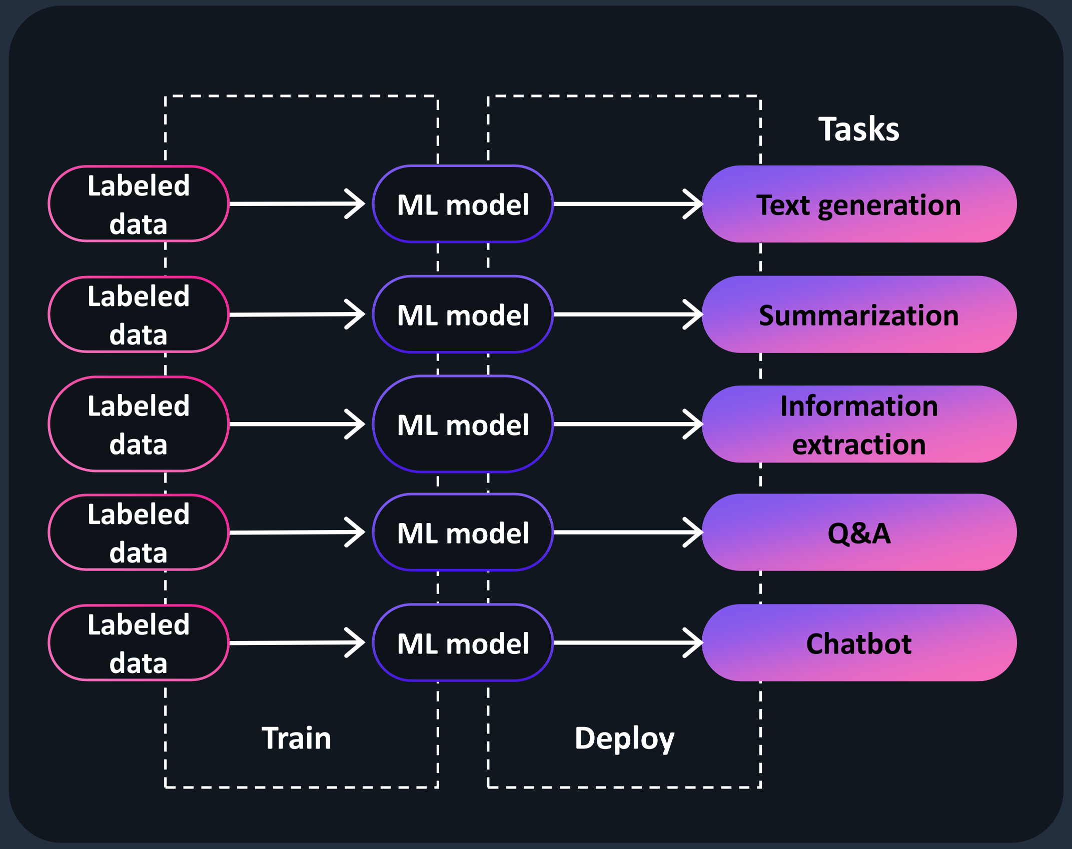
Contain large number of parameters that make them capable of learning complex concepts

Can be applied in a wide range of contexts

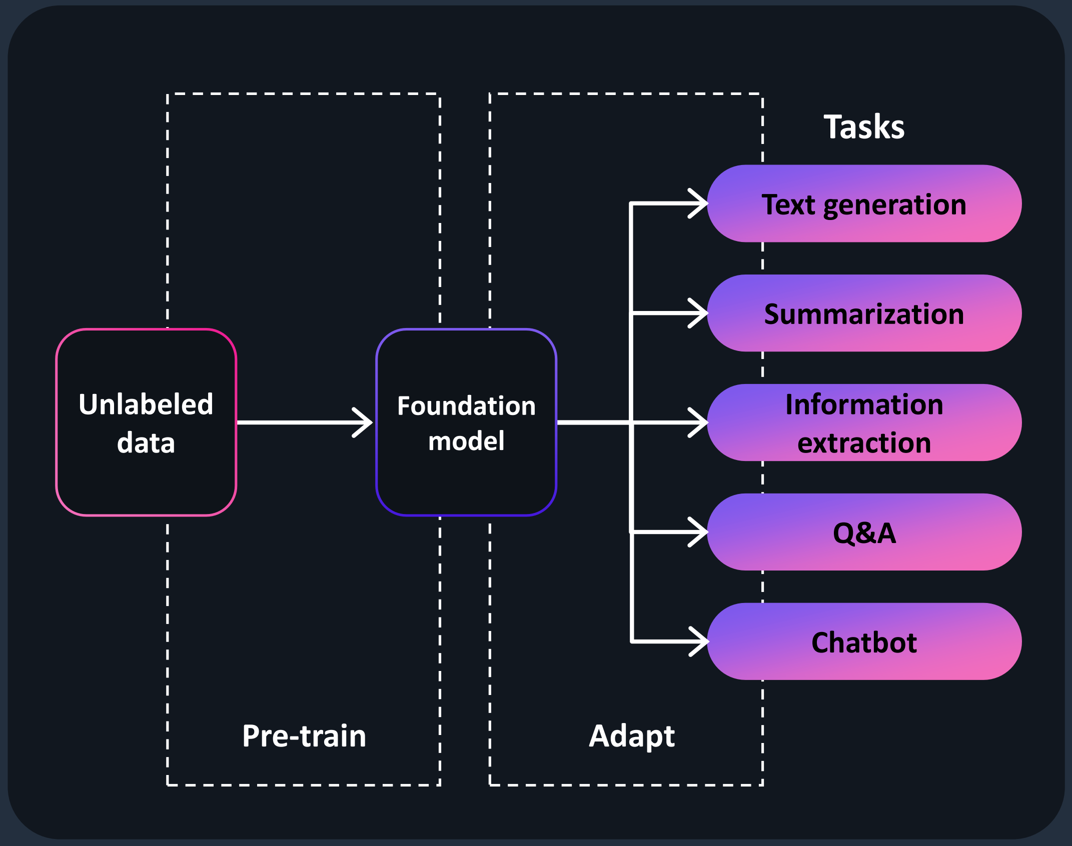
Customize FMs using your data for domain specific tasks



How foundation models differ from other ML models

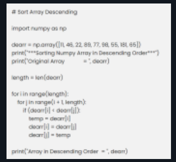


Traditional ML models



Foundation models

Types of foundation models

Input	FM	Output
[Text] “Summarize this article …….”	Text-to-text Generate text from simple natural-language prompts for various applications	[Text] “Ten thousand steps per day is optimum for maintaining a healthy heart”
[Text] “hand soap”	Text-to-embeddings Generate numerical representation of text that reflect the semantic meanings	[Vectors] [0.21, 0.18, 0.92, 0.47, 0.85,…] Hand soap [0.19, 0.15, 0.93, 0.45, 0.82,…] Liquid soap [0.15, 0.19, 0.99, 0.49, 0.80,…] Shower gel
[Text] “a photo of an astronaut riding a horse on mars”	Text-to-Image Generate and edit images from natural-language prompts	[Image] 
[Image]  + (optional) “a rose”	Image-to-Image Generate a new image using another image as guidance and (optionally) a prompt	[Image] 
[Text] “A young couple walking in rain.” “Children singing nature songs” “Write Python code to sort array …”	Multimodal Video Audio Code generation model	[Video]  [Audio]  [Code] 

Democratizing AI/ML through collaborations

AWS has collaborated with Meta PyTorch to help enterprise customers to move DL models from research into production seamlessly.

AWS has collaborated with Hugging Face to easily fine-tune and deploy next generation ML models on EC2 and SageMaker.

AWS has collaborated with Anthropic selects to become its primary cloud provider

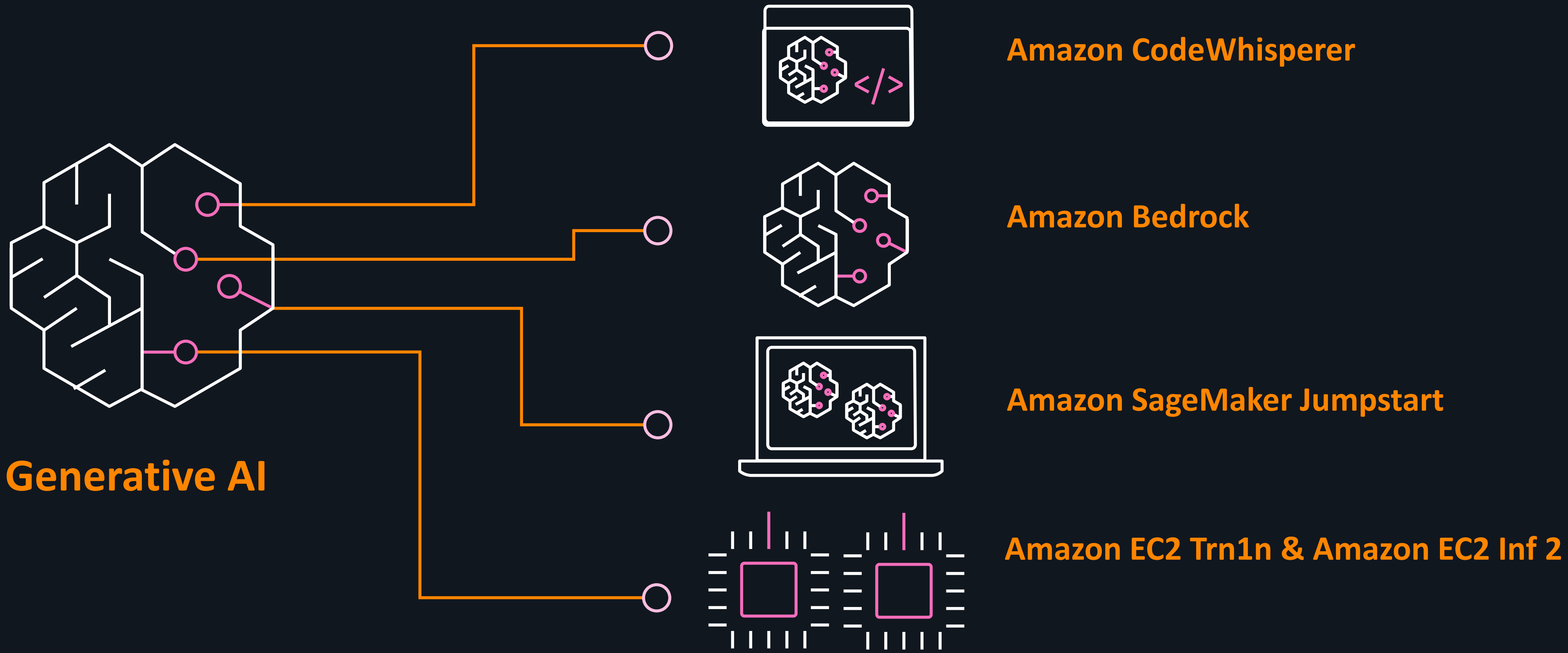
Meta and AWS collaborate to build, train, and deploy ML models with PyTorch: PR Link

Hugging Face and AWS collaborate to make open source models and AI more accessible: PR Link

Anthropic and AWS announce strategic collaboration to advance generative AI: PR Link

Platforms for sharing ML models & datasets

AWS offers a broad choice of Generative AI capabilities



Amazon SageMaker JumpStart

SageMaker JumpStart Overview

Machine learning hub

Browse through 400+ built-in algorithms with pretrained models, pretrained foundation models, solutions, and example notebooks

Pre-built training and inference scripts

Compatible with SageMaker and configurable with custom dataset

UI as well as API-based

Use the user interface for single click model deployment or API for the Python SDK-based workflow

Notebooks with examples

Jump into notebooks to use selected model with examples to guide you through the entire ML workflow

Share and collaborate within your organization

Share models and notebooks with others within your organization, and allow them to train with their own data or deploy as-is for inferencing

Foundation Models available on 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

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

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

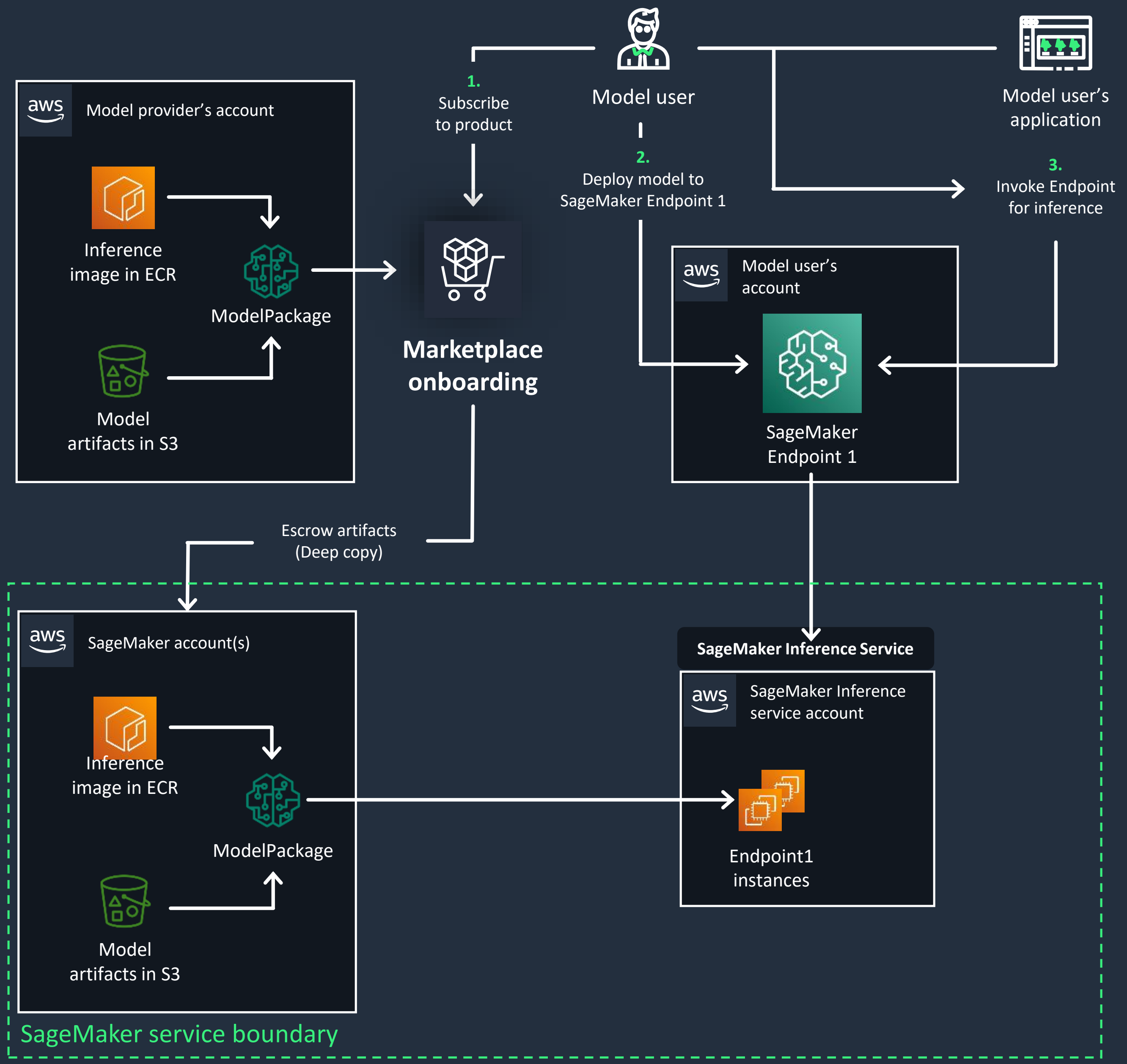
Recent updates/releases for public models

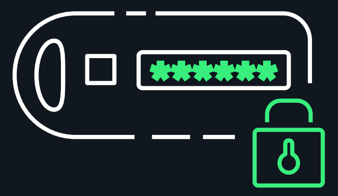
Latest inference only models

- Stable Diffusion XL 1.0 and 0.8
- Llama 2 and Llama 2 chat 70B, 13B and 7B
- Falcon and Falcon Instruct 40B
- Dolly V2
- RedPajama

Fine-tunable models

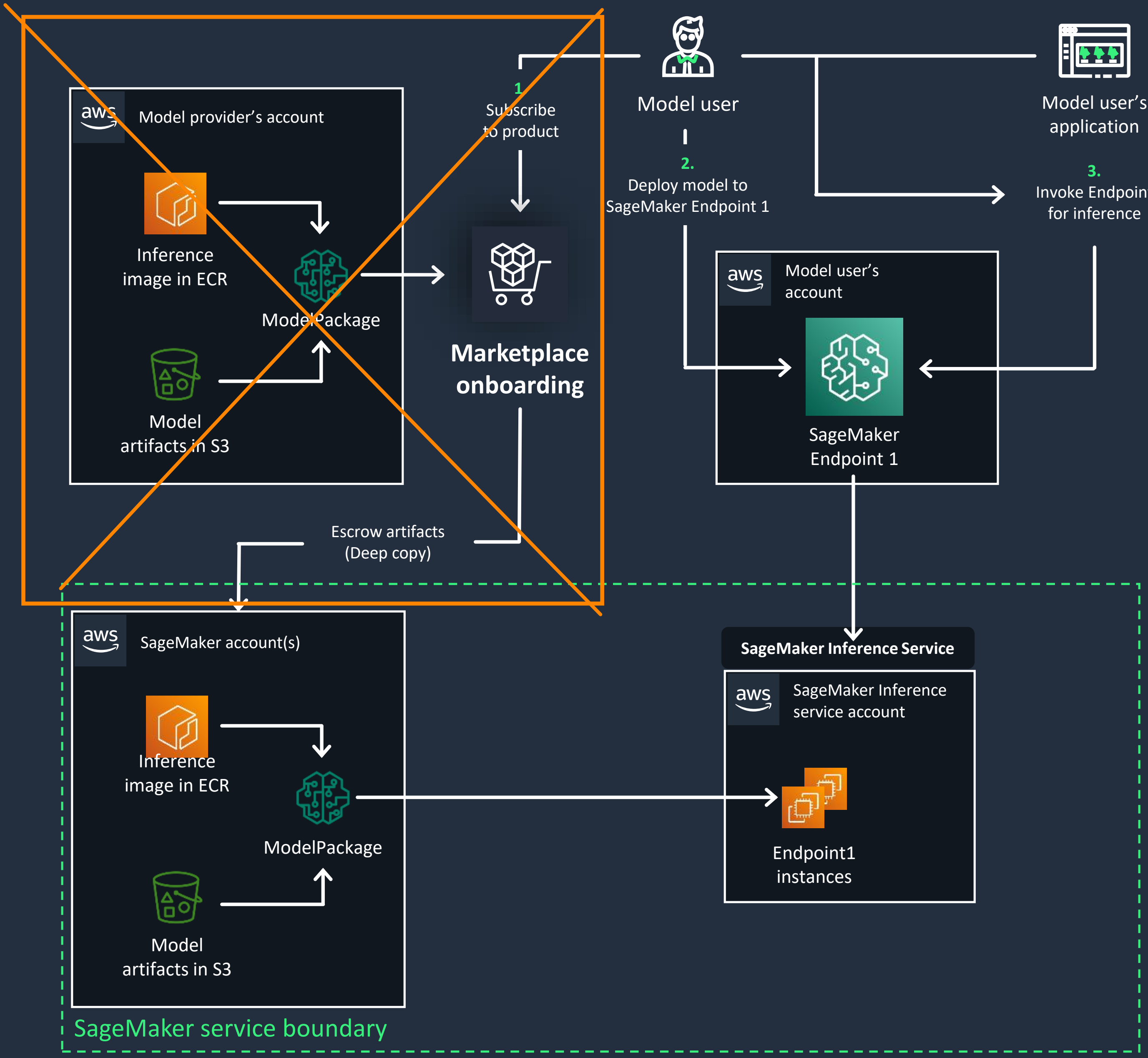
- Falcon 7B
- Red Pajama
- LightGPT
- FLAN T5 XL, XXL
- GPT-J 6B, GPT-NeoX
- Stable Diffusion 2.1





Deploying Proprietary Models through SageMaker JumpStart

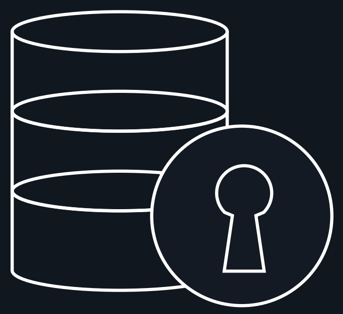
- Proprietary model package and endpoint is hosted in SageMaker owned account
- Containers have no outbound network access; user data and model provider IP is protected the same time
- No data is used to update/train the base model that JumpStart provides to customers



Deploying Publicly Available Models through SageMaker JumpStart

- Public model package and endpoint is hosted in SageMaker owned account
- SageMaker distributes Open-source model artifacts in world readable S3 buckets
- Containers can be enabled to run without network access

Data Privacy and security is our #1 priority



Customer is always in control of their data

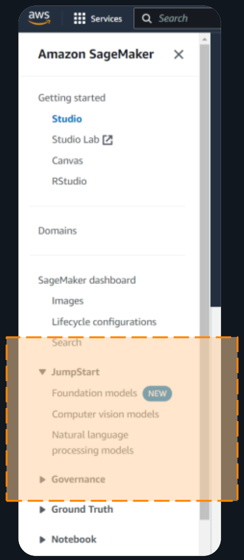
Customer data is not used for service improvement - training or re-training of 3rd Party models

Customer data (prompts or responses) not shared with Amazon or 3rd Party model providers

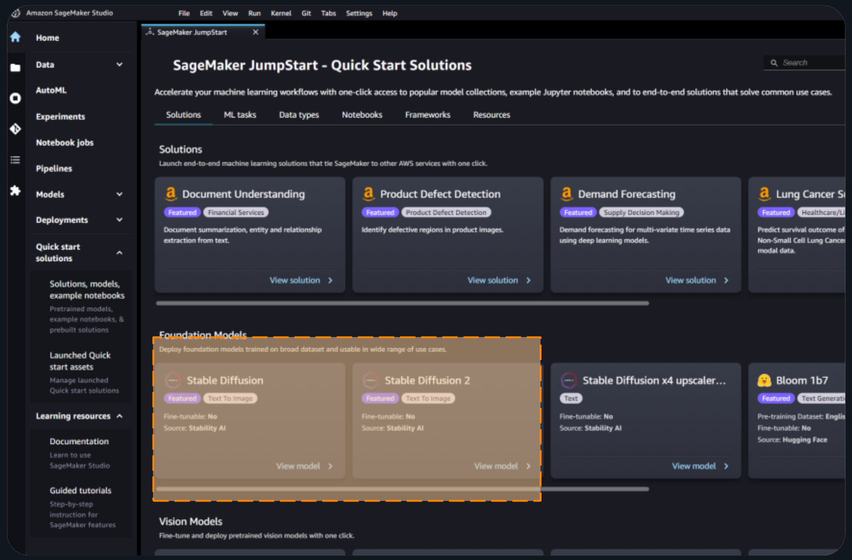
Customer data (prompts, responses, fine-tuned models) are kept in the region where they were created

3 ways to use FMs with SageMaker JumpStart

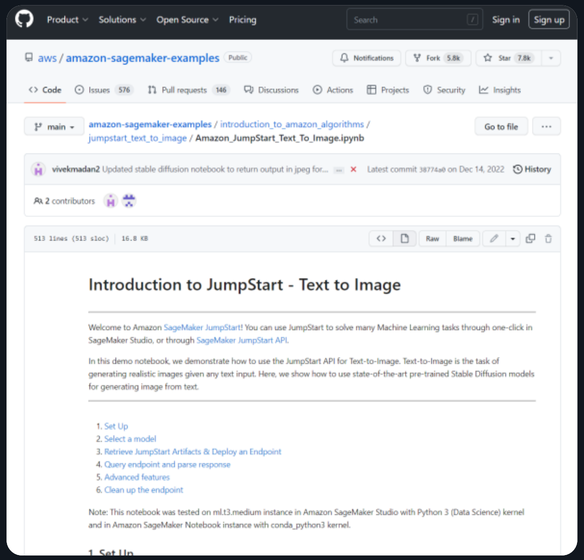
AWS console
Preview



SageMaker Studio
One-click deploy



SageMakerNotebooks
SDK



Pre-trained Foundation Models

Pre-Training

Large corpus + custom task or corpus

↓

Unsupervised /Self-Supervised Learning

↓

Pre-Trained Model

Models
1- DIY or 2: Proprietary and public models

Transformers based architecture:

- Based on a [neural network architecture](#) in processing sequential natural language data.
  - Encoder-only/Autoencoder Models eg: BERT, ROBERTA
  - Decoder-only/Autoregressive Models eg: GPT, BLOOM
  - Sequence-to-Sequence Models eg: T5, BART

- Far better than traditional like RNN based models

Advantages:

- Gives the LLM a strong foundation
- Teaches the LLM general language understanding

Demo 1–

Overview of SageMaker JumpStart and run inference on a pre-trained text to image stable diffusion model

What are challenges with foundation models?

- Larger budget
- Infrastructure requirements
- Front-end development
- Lack of comprehension
- Unreliable answers
- Bias

LLM Customizations methods

Emerging LLM customisation patterns

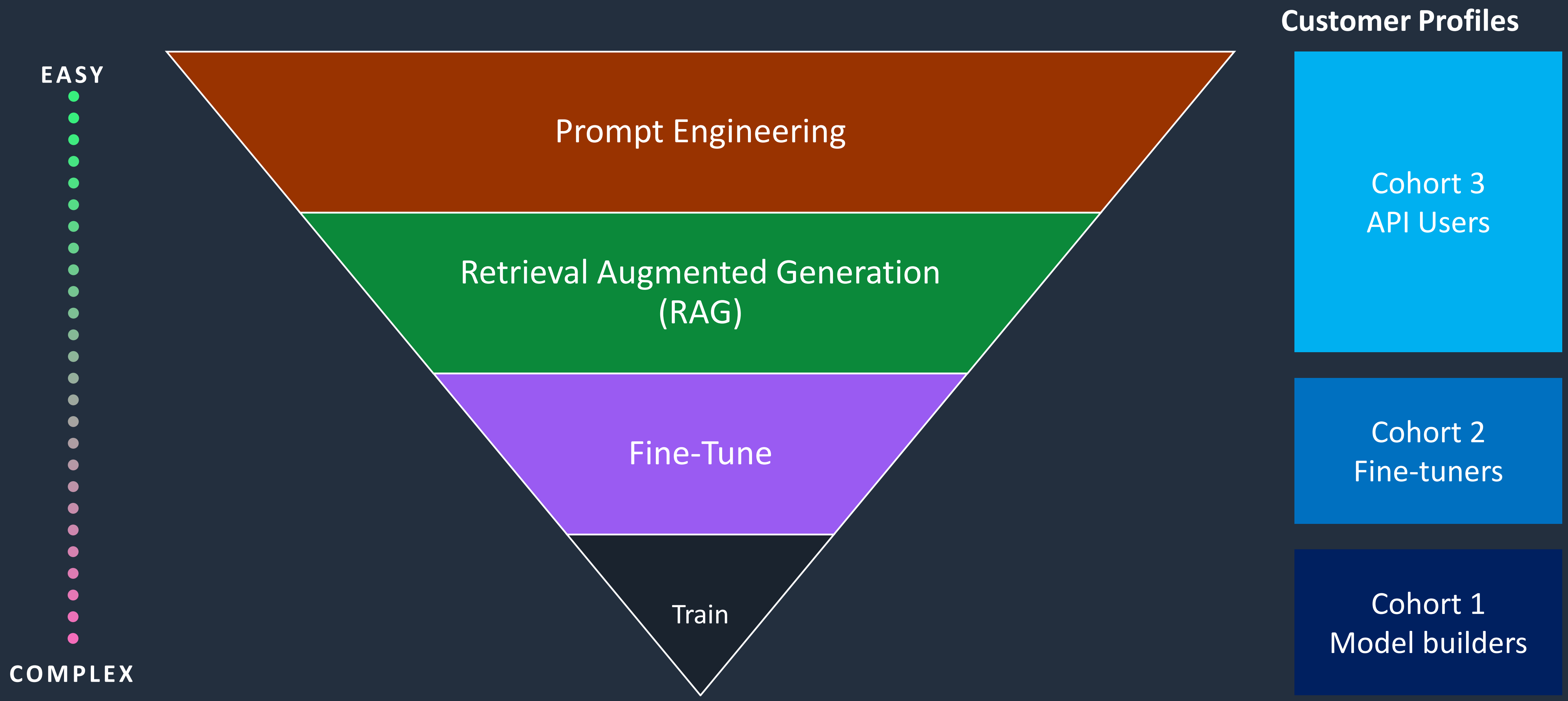
Prompt engineering (In-context learning)

Retrieval Augmented Generation (RAG)

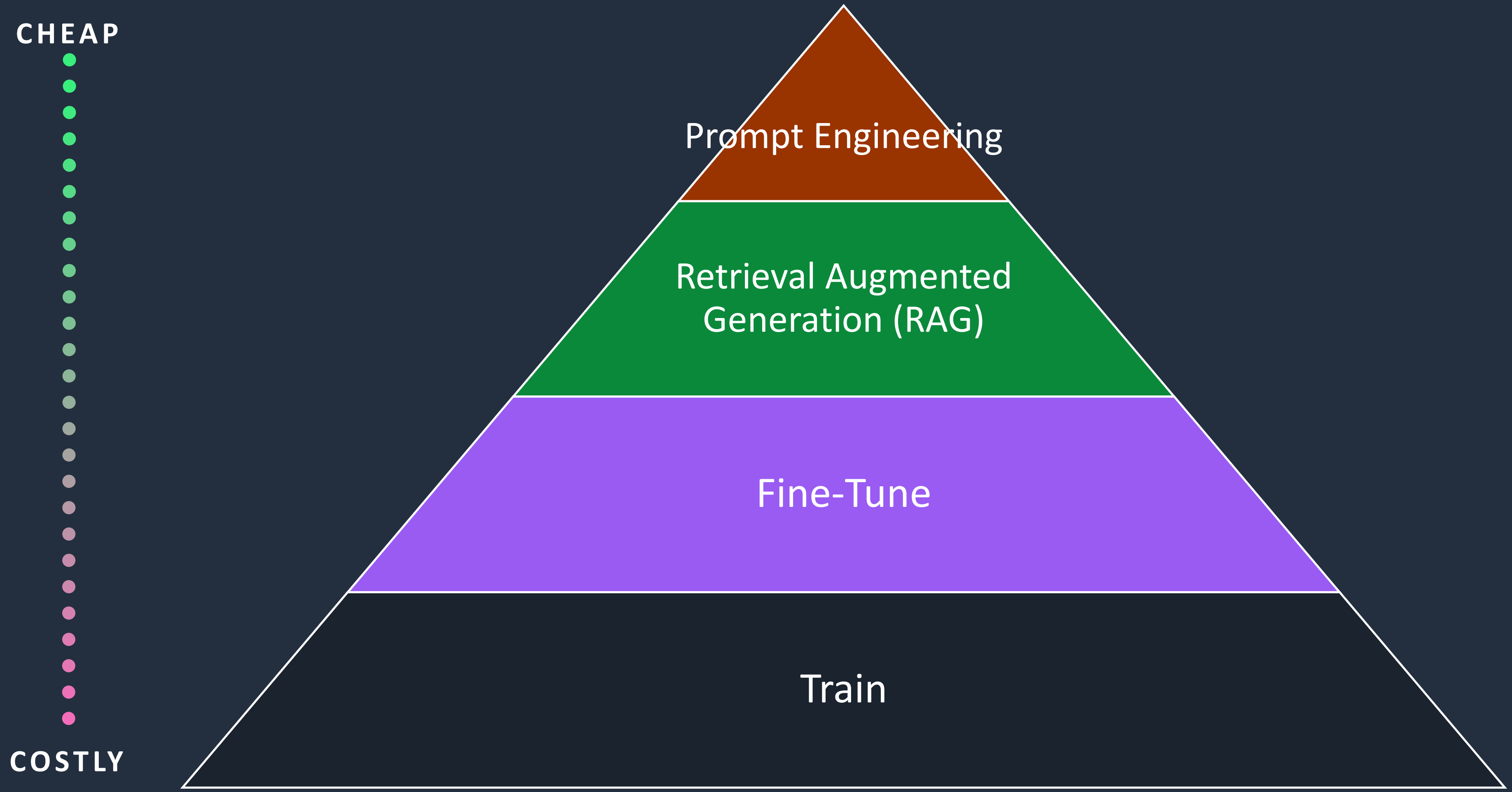
Fine-tuning

Training your own LLM

LLM customisation skills required



LLM customisation cost



Customer Profiles

Cohort 3
API Users

Cohort 2
Fine-tuners

Cohort 1
Model builders

LLM In-Context Learning (Prompt Engineering)

1

In-Context Learning

In-Context Learning (ICL)

- Learn from examples (demonstrations) given during inference. Called few-shot learning.
- ICL is similar to the decision process of human beings by learning from analogy.
- No parameter updates.
- Performance relies on the demonstration format and the order of examples.

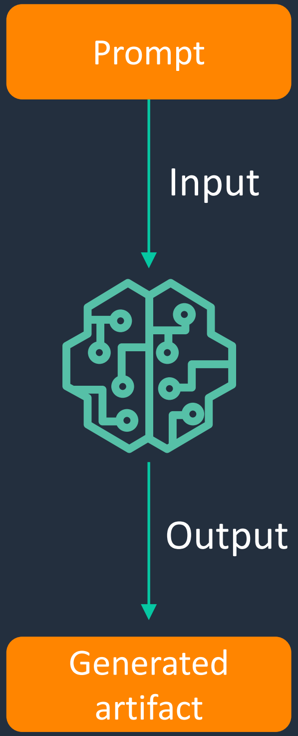
Prompt engineering types

1

In-Context Learning

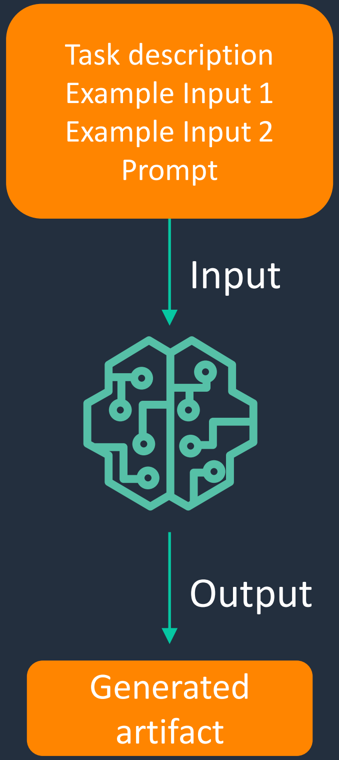
Zero shot prompts

- Direct request with sufficient context



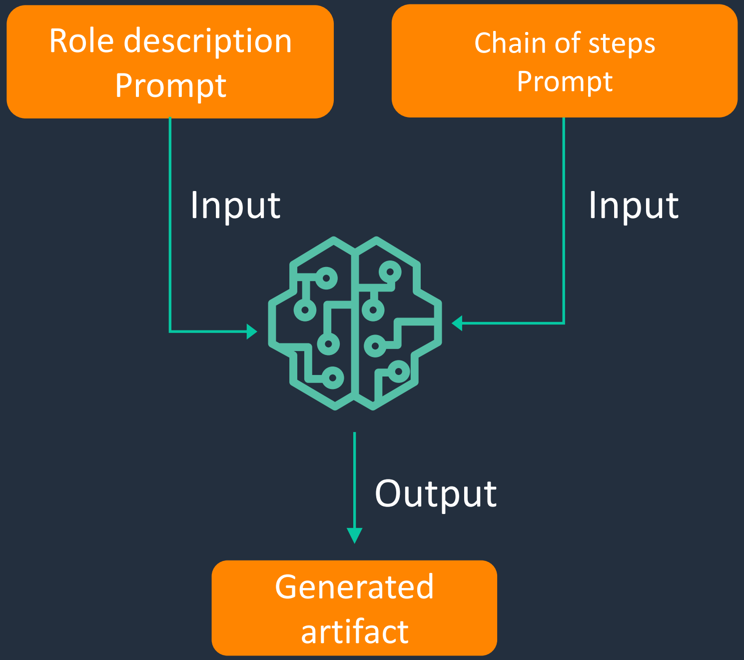
One shot or few shot prompts

Provide one or more examples with a request



Role or Chain of Thought prompts

- Provide the model with a role or persona for the task
- Provide a chain of steps for the model to follow



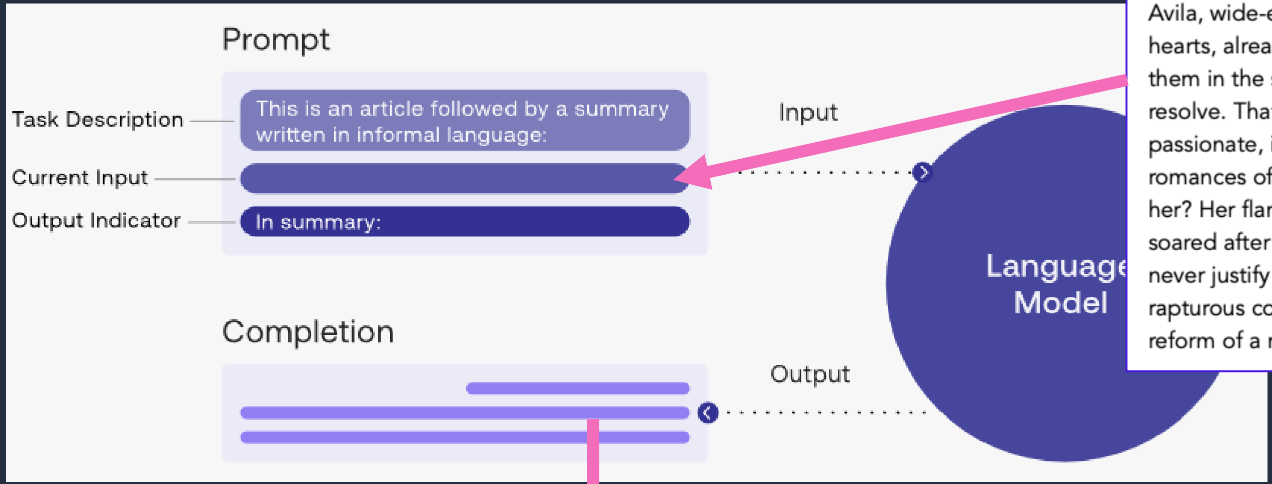
Zero shot prompt: prompting by instruction

1

In-Context Learning

- Zero-shot prompting allows language models to perform tasks for which they have not been explicitly trained on
- Using LLM out of the box

Who that cares much to know the history of man, and how the mysterious mixture behaves under the varying experiments of Time, has not dwelt, at least briefly, on the life of Saint Theresa, has not smiled with some gentleness at the thought of the little girl walking forth one morning hand-in-hand with her still smaller brother, to go and seek martyrdom in the country of the Moors? Out they toddled from rugged Avila, wide-eyed and helpless-looking as two fawns, but with human hearts, already beating to a national idea; until domestic reality met them in the shape of uncles, and turned them back from their great resolve. That child-pilgrimage was a fit beginning. Theresa's passionate, ideal nature demanded an epic life: what were many-volumed romances of chivalry and the social conquests of a brilliant girl to her? Her flame quickly burned up that light fuel; and, fed from within, soared after some illimitable satisfaction, some object which would never justify weariness, which would reconcile self-despair with the rapturous consciousness of life beyond self. She found her epos in the reform of a religious order.



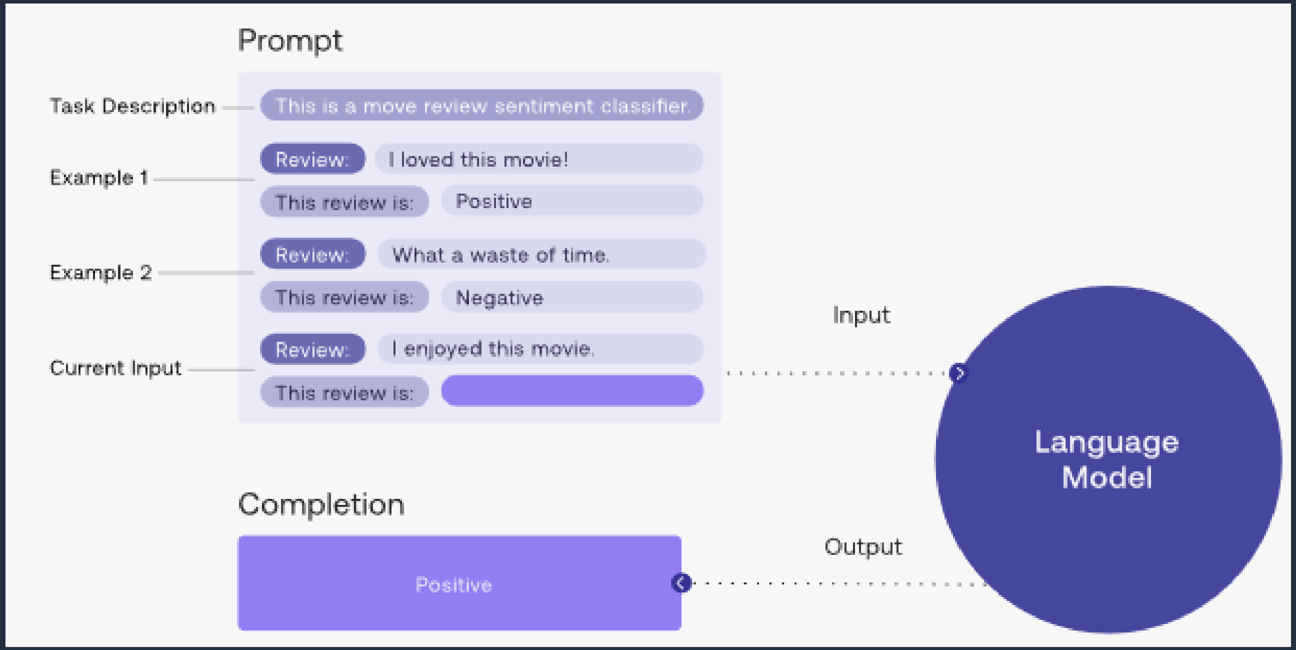
Theresa was a young girl of Spanish origin who became a member of a religious order in Italy.

Prompt Engineering: N shots example

1

In-Context Learning

- Same as standard prompt but A few-shot prompt normally includes n examples of (problem, solution) pairs known as "shots".
- Help to guide model performance.

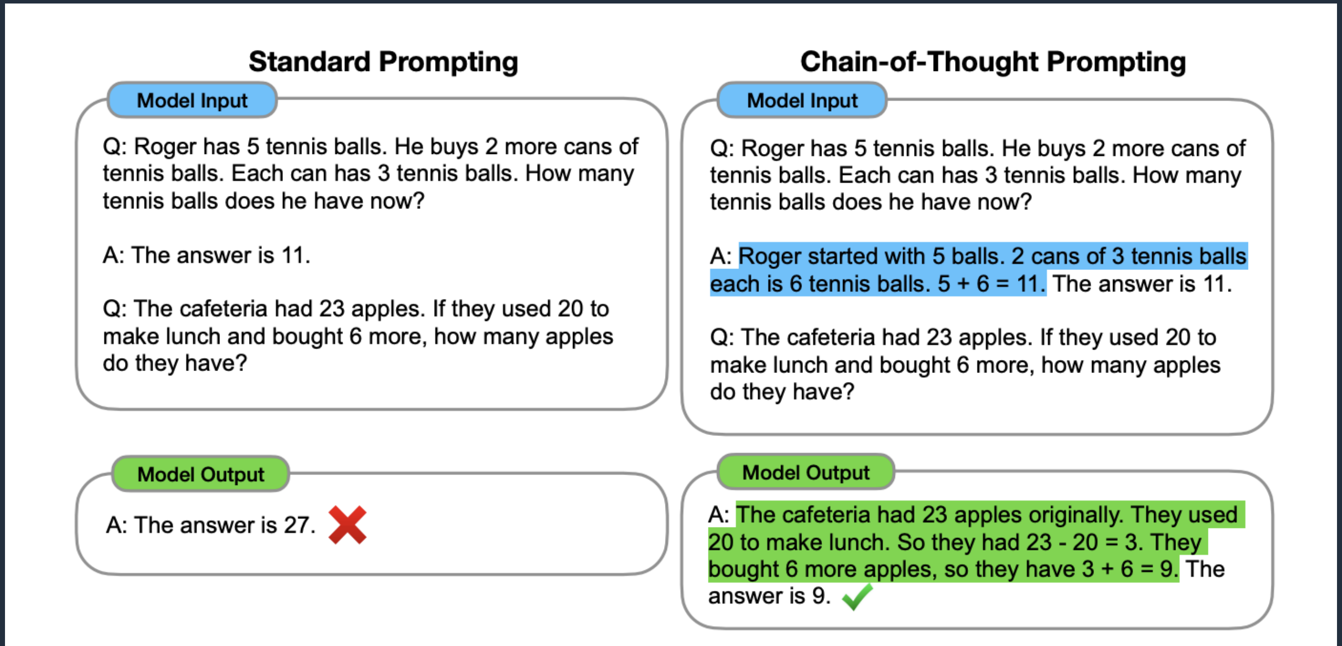


Prompt Engineering: Chain of Thought Prompting Examp

1

In-Context Learning

- Improves *reasoning* abilities in foundation models
- Addresses *multi-step problem-solving* challenges in arithmetic and *commonsense reasoning* task
- Generates intermediate reasoning steps, mimicking *human train of thought*, before providing the final answer.



Source: https://arxiv.org/pdf/2201.11903.pdf

1

In-Context Learning

Important parameters

Parameter settings to customize results:

- Temperature:
  - controls randomness.
    - Lower values pick probable tokens
    - higher values add randomness and diversity.

    => Use lower for factual responses, higher for creative

- Top-p: also adjusts determinism with "nucleus sampling".
  - Lower values give exact answers
  - higher values give diverse responses

Note:

- Only adjust one parameter at a time.
- Outcomes vary between language model types

Demo 2 – Prompt engineering

1

In-Context Learning

LLM and prompting limiation

- What about complex and knowledge-intensive tasks,
- accessing external knowledge sources to complete tasks.

=> Retrieval Augmented Generation method

LLM In-Context Learning (RAG)

2

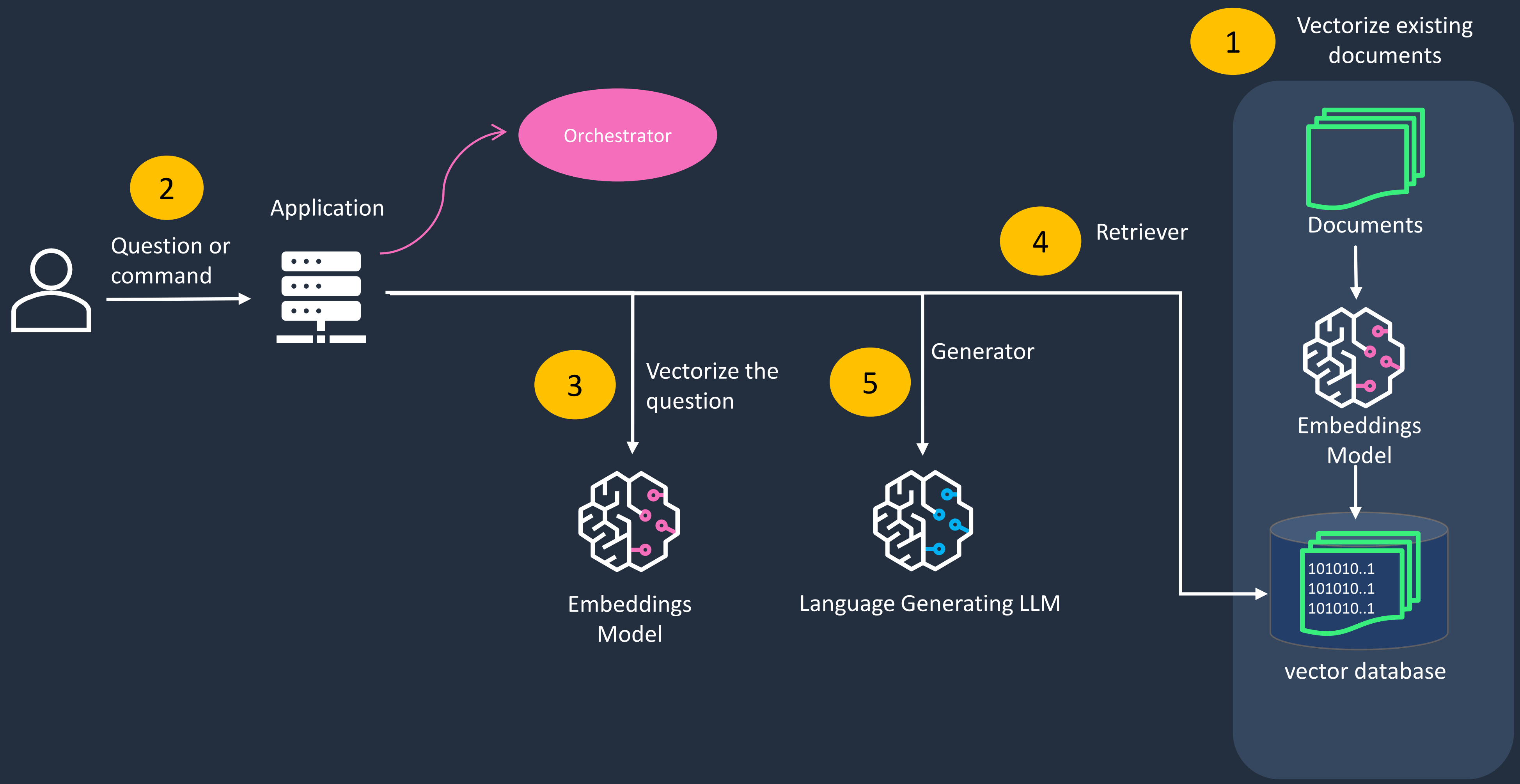
RAG

Retrieval Augmented Generation (RAG)

A sweet spot for many organisations

- Concise and relevant context
- Evolve knowledge base on the fly
- No need of complex LLM training and retraining
- No need to host a dedicated LLM

Architecture components in a RAG solution



Which Vector Database to use in AWS?

Looking for low-code no-code solution and rapid deployment	Already using Amazon OpenSearch Service and comfortable with NoSQL	Already using Amazon RDS/Aurora PostgreSQL and prefer using SQL
Use out-of-the-box 40+ connectors to ingest data from many data sources	Need low latency with HNSW algorithm, and lower memory with product quantization	Keep application and AIML data and vectors in the same DB for better governance, faster clones
Do not want to deal with management of data chunking, embeddings, and indexing algorithm choices	Need higher search accuracy using larger vector dimensions; flexibility to pick any vector embedding model	Need transactional and immediate consistency

Role Based Access Controls, Encryption, Authentication, Authorization, Auditing, and Fully Managed Services with Serverless Option

Amazon Kendra	Amazon OpenSearch Service or Amazon OpenSearch Serverless	Amazon RDS for PostgreSQL or Amazon Aurora PostgreSQL

Retrieval Augmented Generation (RAG)

Key Limitations:

- Increased Complexity => adding retriever component to generation model
- Limited Creativity => constrained by the retrieved information

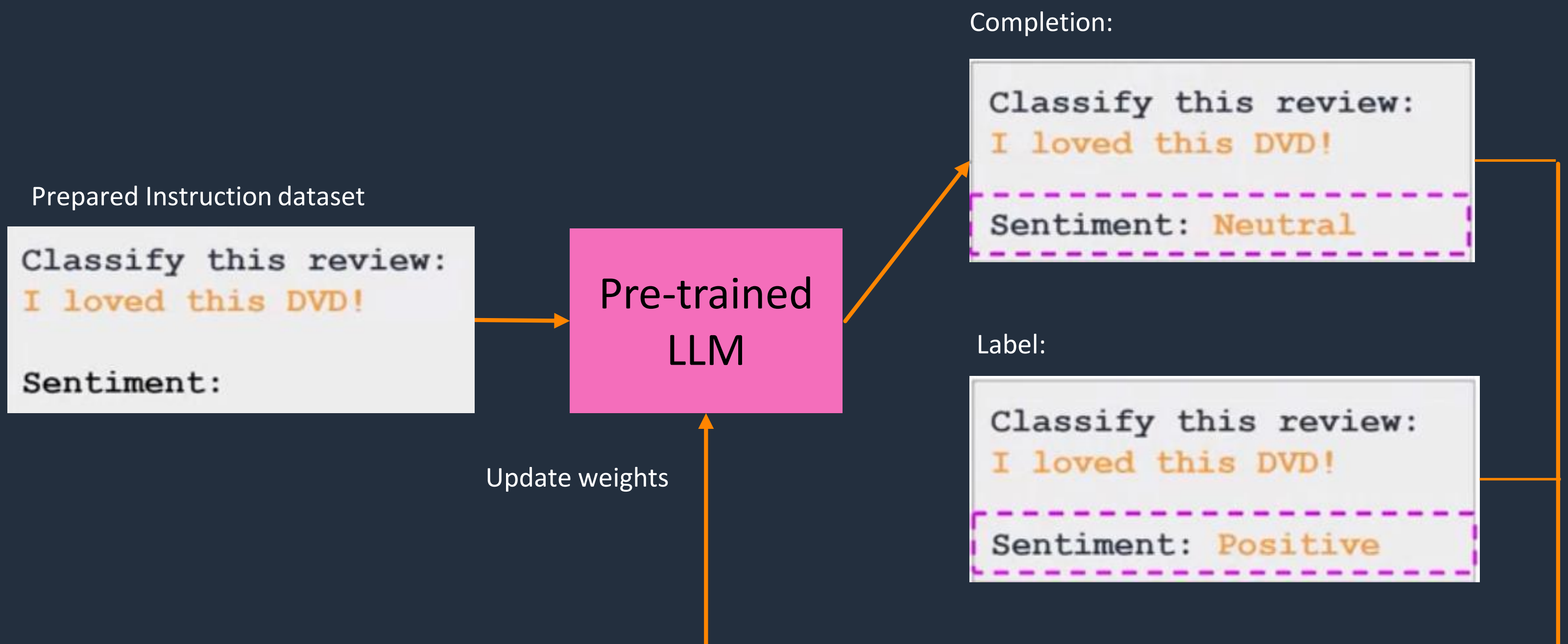
Fine-tuning might help overcome the above limitations

aws partnercast

LLM - Fine Tuning

aws

Instruction(Instruct) Fine-Tuning



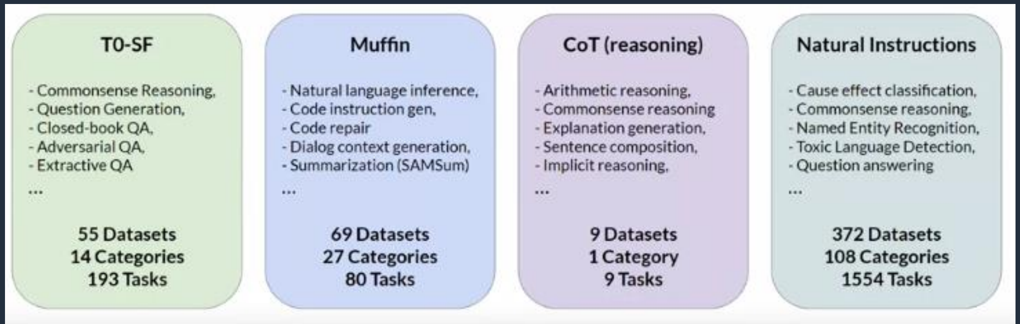
3

Fine-Tuning

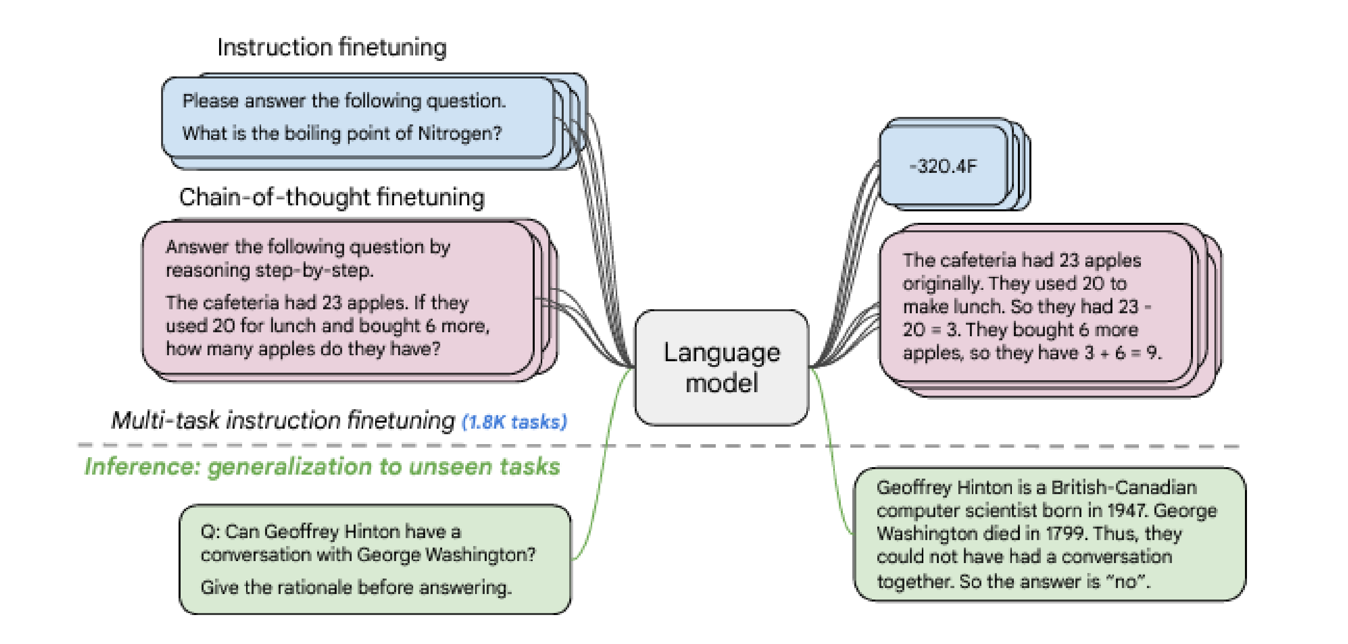
- Supervised learning on a Pre-Trained FM
- Turn pre-trained model into a specialized model
- Better performance with fewer labeled examples
- Smaller (fine-tuned) model can often outperform larger (more expensive) models
- Catastrophic forgetting- Single task fine tuning

FLAN (Fine-tuned Language Net)

- Multi Task FT- Overcome Catastrophic forgetting
- Multiple-tasks such as summarization, review rating, code translation, and entity recognition
- Fine tuned on 473 datasets, 146 categories, 1.8K tasks:



- Requires lots of data
- Varients such as FLAN- T5, FLAN- PALM, FLAN- UL2
  - Flan-PaLM 540B on 1.8K tasks outperforms PALM 540B by 9.4%.
  - Flan-PaLM 540B achieves 75.2% on five-shot MMLU
- Further fine-tune for specific use-case



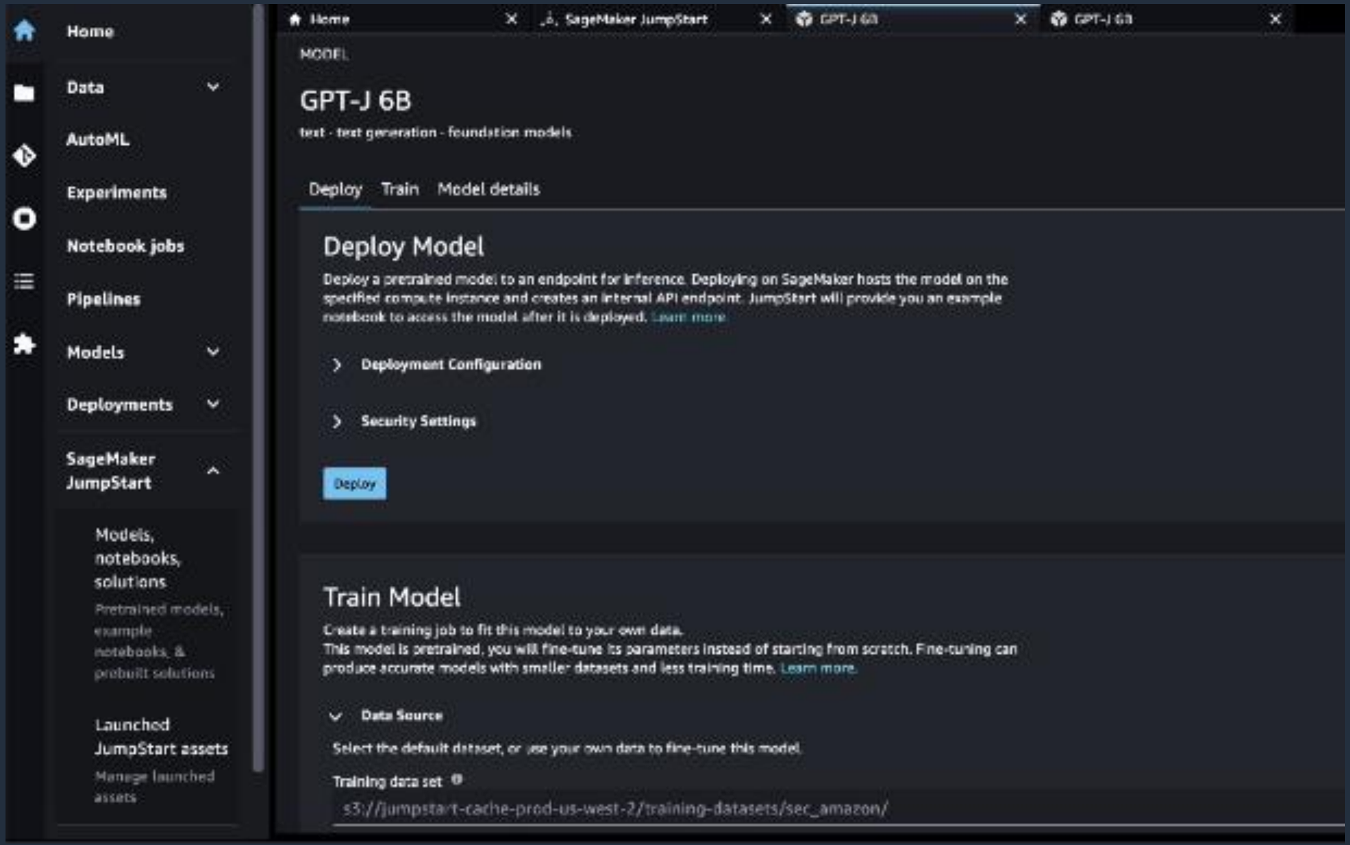
	MMLU	BBH-nlp	BBH-alg	TyDiQA	MGSM
Prior best	69.3[a]	73.5[b]	73.9[b]	81.9[c]	55.0[d]
PaLM 540B					
- direct prompting	69.3	62.7	38.3	52.9	18.3
- CoT prompting	64.5	71.2	57.6	-	45.9
- CoT + self-consistency	69.5	78.2	62.2	-	57.9
Flan-PaLM 540B					
- direct prompting	72.2	70.0	48.2	67.8	21.2
- CoT prompting	70.2	72.4	61.3	-	57.0
- CoT + self-consistency	75.2	78.4	66.5	-	72.0

Source- https://arxiv.org/abs/2210.11416

Domain Adaptation Fine-Tuning on Amazon SageMaker

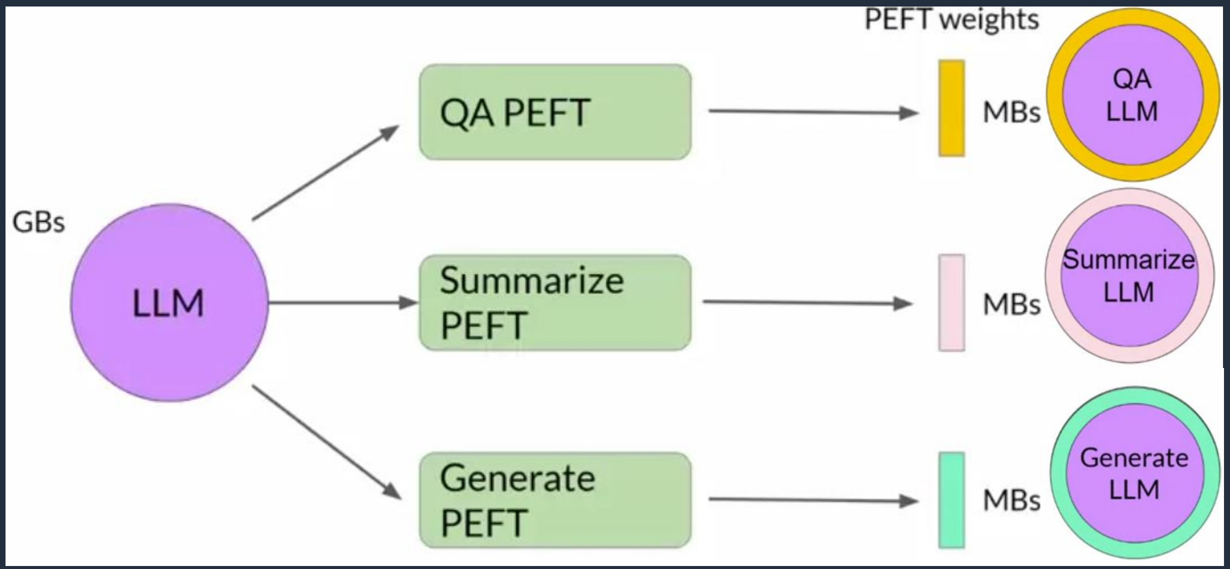
- Fine tuning GPT-J 6B
- On SEC filing dataset
- Use either JumpStart SDK or Amazon SageMaker Studio UI

Input	Output from GPT-J 6B Before Fine-Tuning	Output from GPT-J 6B After Fine-Tuning
What drives sales growth at Amazon?:	Amazon is the world's largest online retailer. It is also the world's largest online marketplace. It is also the world'	Sales growth at Amazon is driven primarily by increased customer usage, including increased selection, lower prices, and increased convenience, and increased sales by other sellers on our websites.



Domain-adaptation Fine-tuning of Foundation Models in Amazon SageMaker JumpStart on Financial data

Parameter-Efficient Fine-Tuning (PEFT)



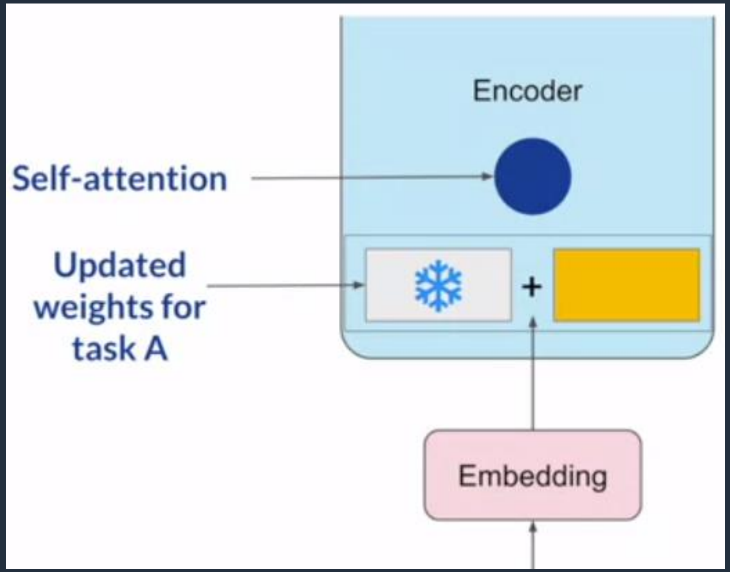
- A novel approach for fine-tuning
- Open-source library from HuggingFace
- Fine-tune a small number of (extra) model parameters
- State-of-the-Art PEFT achieve full fine-tuning performance
- Supported methods
  - LoRA & QLoRA - are most widely used and effective.
  - Prefix Tunning
  - AdaLora

  Etc..

[Github - HuggingFace - PEFT]

Low-Rank Adaptation (LoRA)

Training method that accelerates the training of large models while consuming less memory



- Freeze most of the original LLM weights eg: d x k = 512x64
- Inject 2 decomposition matrices (rank= 8) eg: r x k = 8 x 64, d x r = 512 x 8
- Train only the smaller matrices
- Add to the original weights

- GPT-3 175B- Reduces # of trainable parameters by 10000x and the GPU memory by 3x.
- Llama-2 7B- Fine-tune less than 1% of the parameters.

LORA: LOW-RANK ADAPTATION OF LARGE LANGUAGE MODELS

HuggingFace - Low-Rank Adaptation of Large Language Models (LoRA)

Quantized Low-Ranking Adaptation (QLoRA)

QLoRA extends LoRA to enhance efficiency by quantizing weight values

- Key optimizations:
  - 4-bit NormalFloat (FP32 → NF4)
  - Double Quantization (Constant unit variance)
  - Paged Optimizers (Prevent memory spikes during gradient checkpointing)

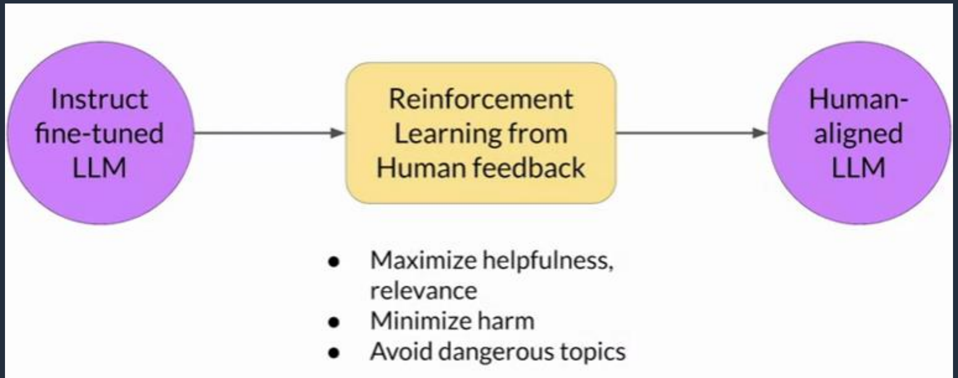
- Enables finetuning a 65B parameter model on a single 48GB GPU
- Matches the performance of fine-tuning and achieves state-of-the-art results on language tasks

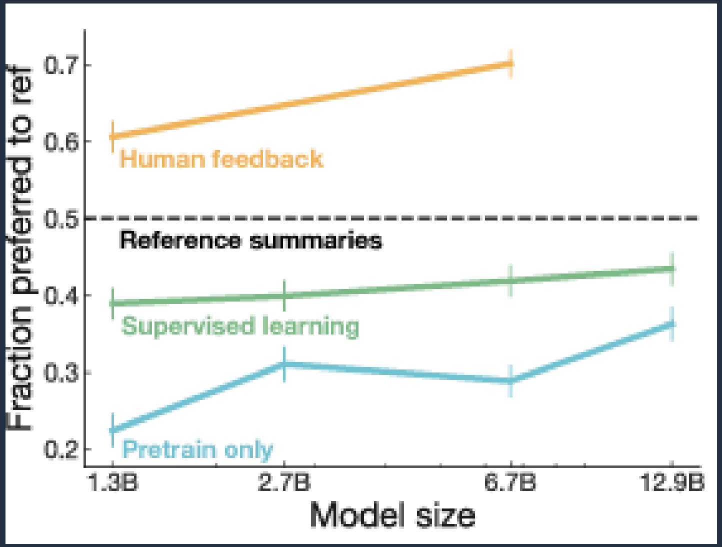
QLORA: Efficient Finetuning of Quantized LLMs
Fine Tuning LLM: Parameter Efficient Fine Tuning (PEFT) — LoRA & QLoRA

Responsible AI using RLFH

Reinforcement Learning with Human Feedback

- Align models with human values- helpfulness, honesty, and harmlessness (HHH)
- RLHF is popular technique for finetuning LLMs with human feedback
- RLHF shows better responses than a pretrained LLM, instruct fine-tuned LLM, and reference human baseline.
- RLHF is a complex and often unstable procedure.

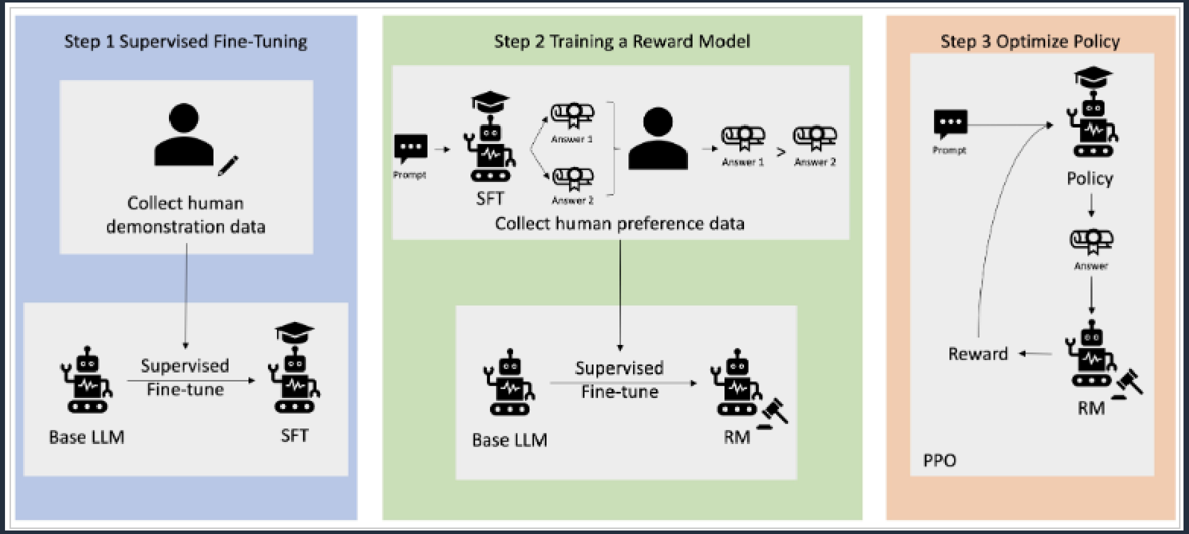




https://arxiv.org/abs/2009.01325

RLHF on Amazon SageMaker

- Human data annotators are tasked with authoring responses to various prompts.
- The collected responses (referred to as demonstration data) are used for supervised fine-tuning (SFT).
- Annotators rank model outputs based on HHH
- Human preference data is used to train a reward model (RM)
- RM is used by Proximal Policy Optimization (PPO) to train the supervised fine-tuned model



[Improving your LLMs with RLHF on Amazon SageMaker]

The Human Evaluation approach is defined, launched, and managed by the Amazon SageMaker Ground Truth Plus labeling service.

Demo 4 – Fine-tuning

53

How to get started / Call To Action

- AWS Digital Course - **AWS Partner: Generative AI Essentials (Business) - Gen AI skill badge launch**
- Partner Learning Plans on Generative AI
  - **Generative AI for Business Professionals**
  - **Generative AI for Technical Professionals**
  - **Generative AI for Developers**
  - **Mastering Amazon SageMaker**
- **Workshop Studio**
  - **Discover and participate in AWS workshops and GameDays**
- Labs
  - **Amazon CodeWhisperer workshop**
  - **Amazon BedRock workshop**
  - **Amazon SageMaker Jumpstart**

aws partnercast

Help Us Improve Our Sessions & Presentation Delivery!

Your feedback helps us become better presenters, plan upcoming PartnerCast sessions, and modify content.

Scan code to leave feedback. You will also be re-directed to the survey after the session.

Never miss a session!

Scan to save the Tech Talks landing page URL!

Next session:

Oct 31 | Amazon Transcribe

aws partnercast

You may also be interested in…

[Making better business decision with no-code Machine Learning using Amazon SageMaker Canvas]

Tue, Nov 7

10:30 AM IST | 1:00 PM SGT | 4:00 PM AEDT

Q&A

Thank you!

Please join us again for another PartnerCast session

https://aws.amazon.com/partners/training/partnercast/