LLM Ops

Operationalize Generative AI on AWS

June 13th, 2024

MARIO BOURGOIN

Sr. AIML Partner Solutions Architect

AJIT KUMAR K.P

Sr. AIML Partner Solutions Architect

Agenda

- LLM Ops vs ML Ops: Personas and Process
- The Consumer’s Journey
- The Fine-Tuner’s Journey
- Demo
- Q&A
- Complete exit survey

Generative AI Stack

APPLICATIONS THAT LEVERAGE LLMs AND OTHER FMs

- Amazon Q Business
- Amazon Q Developer
- Amazon Q in QuickSight
- Amazon Q in Connect

TOOLS TO BUILD WITH LLMs AND OTHER FMs

Amazon Bedrock

Guardrails | Agents | Studio | Customization Capabilities | Custom Model Import

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE

- GPUs
- Trainium
- Inferentia
- SageMaker
- JumpStart
- UltraClusters
- EFA
- EC2 Capacity Blocks
- Nitro
- Neuron

Generative AI Stack

APPLICATIONS THAT LEVERAGE LLMs AND OTHER FMs

TOOLS TO BUILD WITH LLMs AND OTHER FMs

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE

GPUs · Trainium · Inferentia · SageMaker

UltraClusters · EFA · EC2 Capacity Blocks · Nitro · Neuron

Generative AI Stack

APPLICATIONS THAT LEVERAGE LLMs AND OTHER FMs

TOOLS TO BUILD WITH LLMs AND OTHER FMs

Amazon Bedrock

Guardrails | Agents | Studio | Customization Capabilities | Custom Model Import

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE

GPUs | Trainium | Inferentia | SageMaker | JumpStart

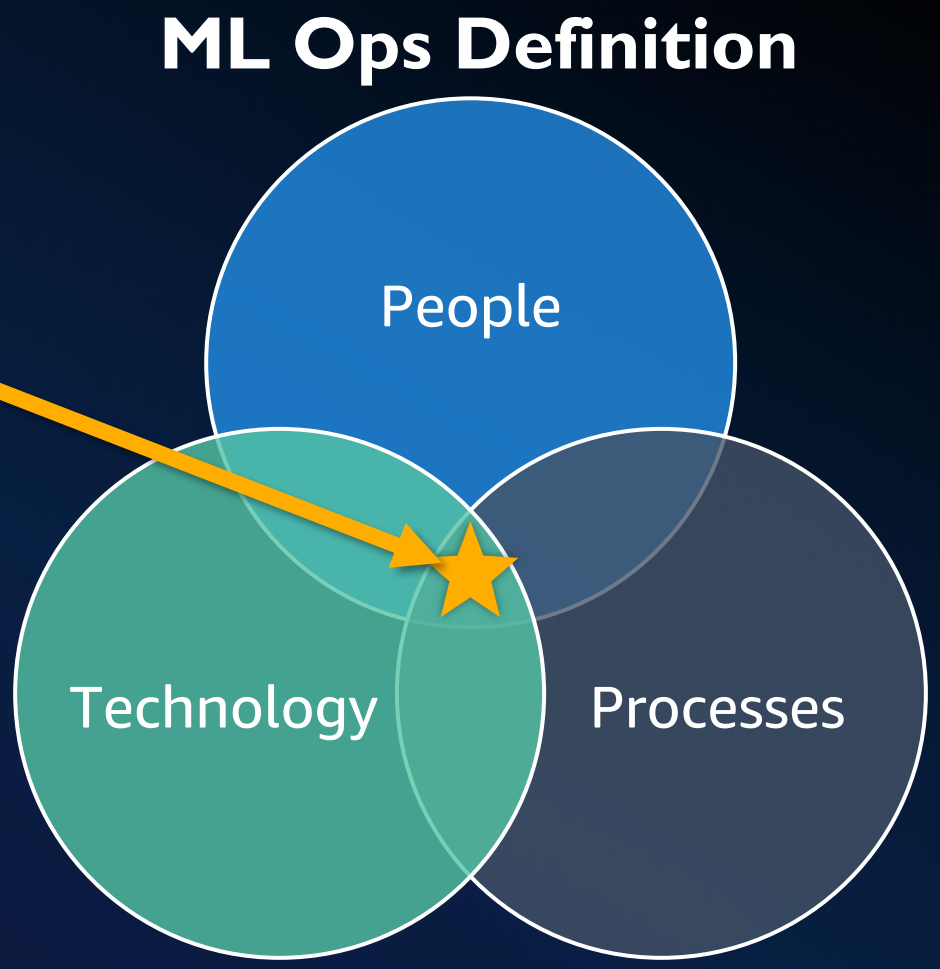
UltraClusters | EFA | EC2 Capacity Blocks | Nitro | Neuron

LLM Ops vs ML Ops

What is ML Ops?

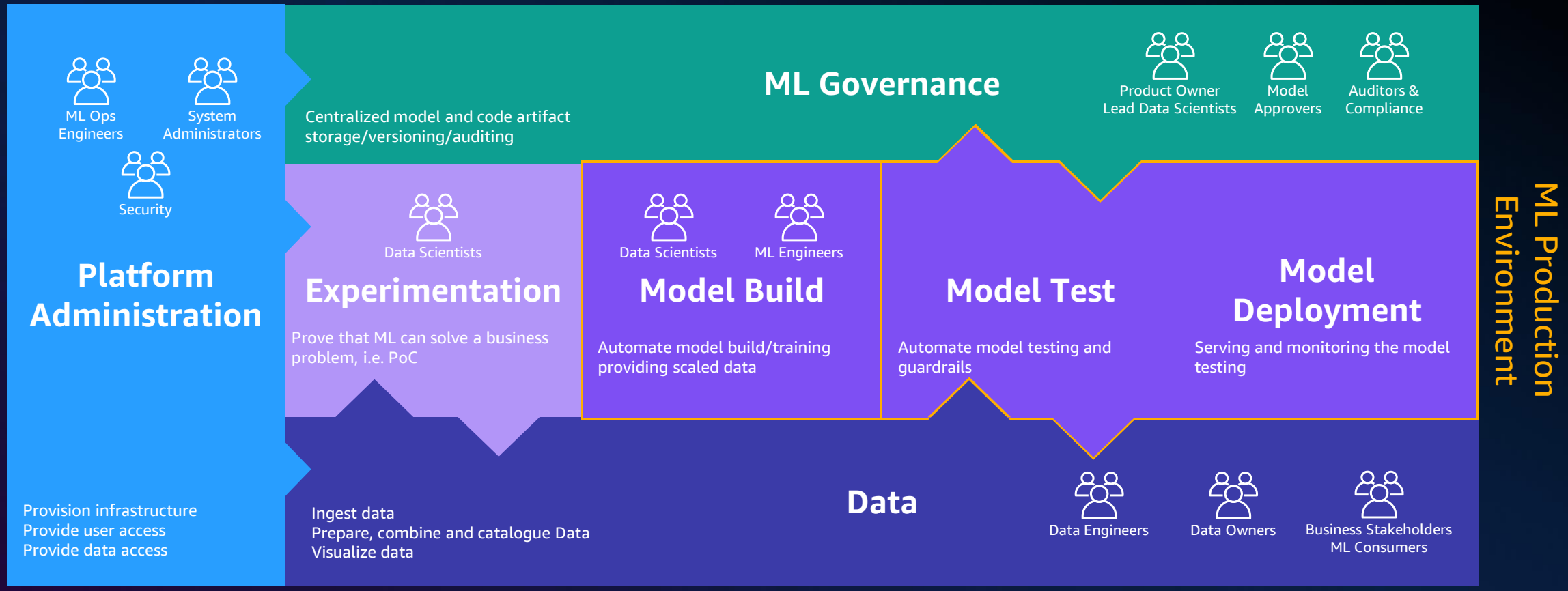
ML Ops
Machine Learning
& Operations

The combination of people, processes, and technology to productionize ML solutions efficiently.



ML Ops Foundation People & Processes

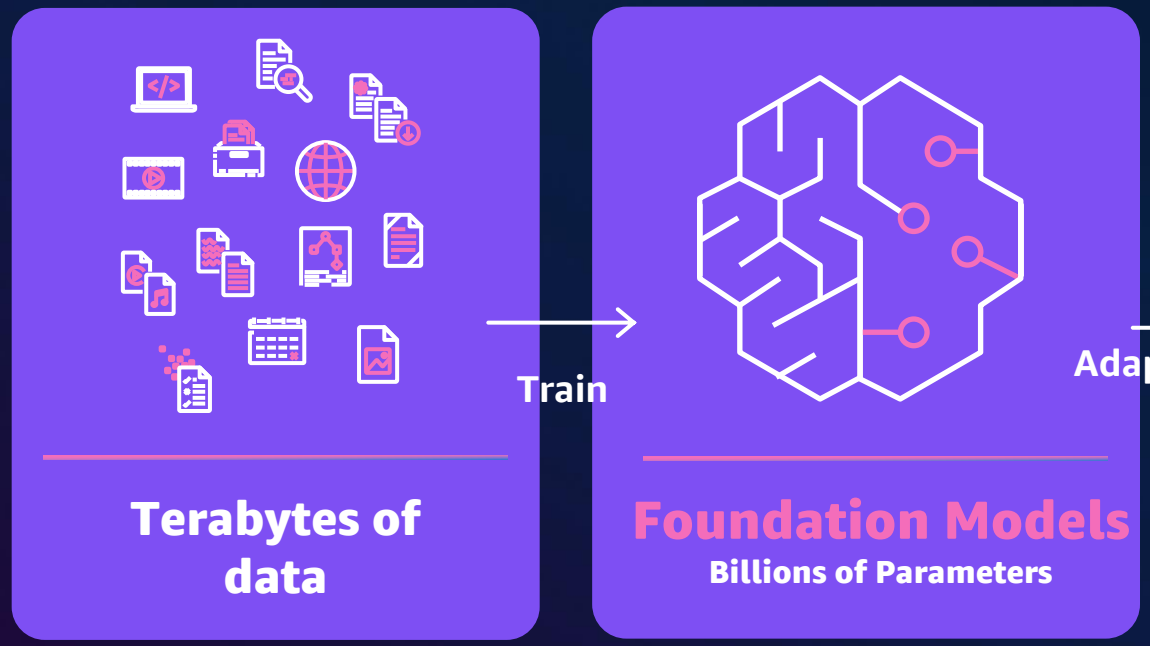
SEPARATION OF CONCERNS IS KEY FOR SUCCESS



Generative AI & ML Ops

ML Ops & FM Ops/LLM Ops Differentiators

Generative AI Use Case Domains



Text-to-Text = LLM – Unlabeled data {text}

- Chatbots
- Writing assistants e.g. summarization
- Programming assistants

Text-to-Image - Labeled data {text, image}

- Generate fantasy images
- Generate new product design

Text-to-Audio or Video – (un)labeled data (coming soon)

- Music composers

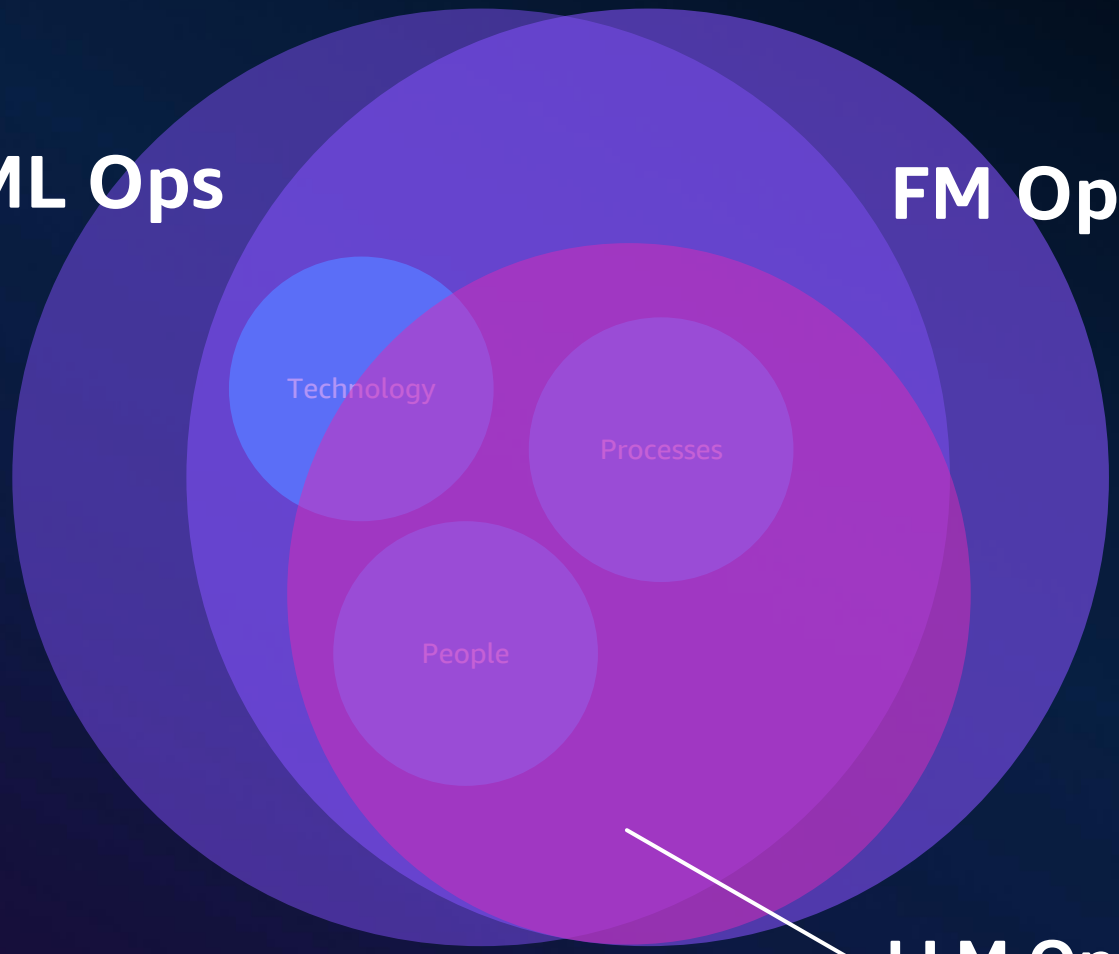
Key Definitions

Machine Learning Operations
Productionize ML solutions efficiently

ML Ops

FM Ops

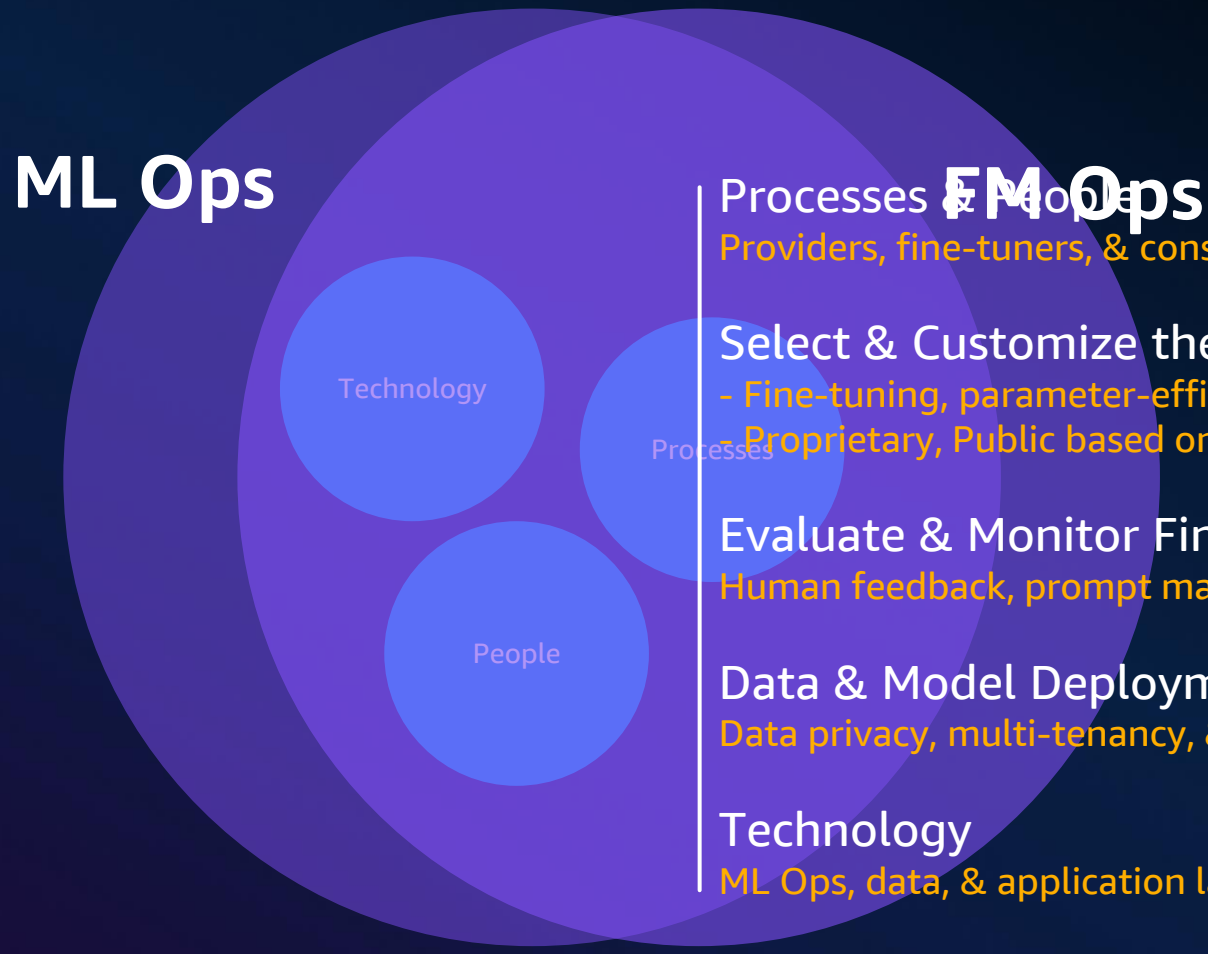
Foundation Model Operations
Productionize Generative AI Solutions (Text-Text/Image/Video/ Audio/ …)



LLM Ops

Large Language Model Operations
Productionize Large Language Model-based solutions

ML Ops & FM Ops Differentiators



FM Ops

Processes & People
Providers, fine-tuners, & consumers

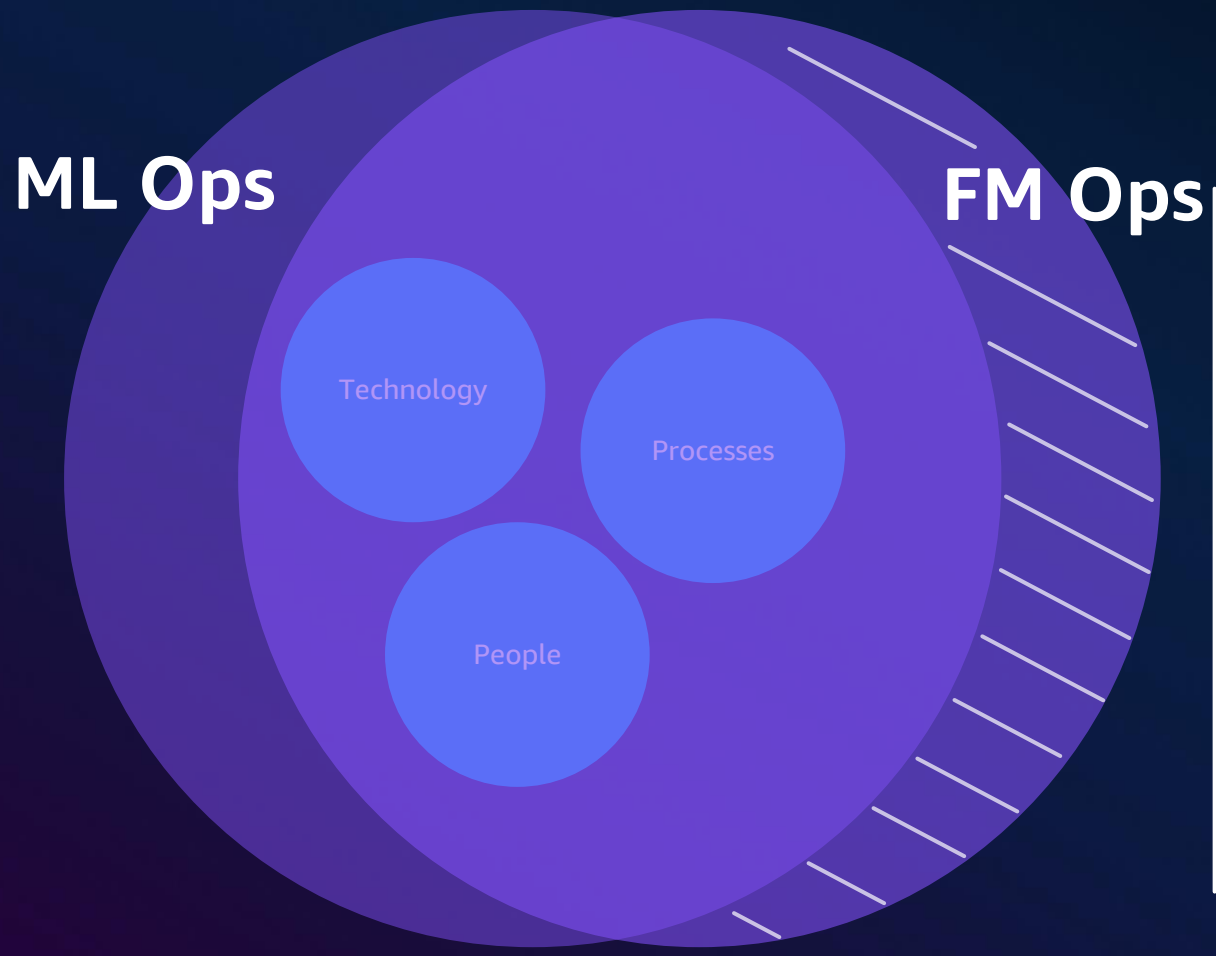
Select & Customize the FM on a Specific Context
- Fine-tuning, parameter-efficient fine-tuning, prompt engineering
- Proprietary, Public based on the application

Evaluate & Monitor Fine-tuned Models
Human feedback, prompt management, toxicity/bias…

Data & Model Deployment
Data privacy, multi-tenancy, & cost, latency, and precision

Technology
ML Ops, data, & application layers

ML Ops & FM Ops Differentiators



Processes & People
Providers, fine-tuners, & consumers

Select & Customize the FM on a Specific Context
- Fine-tuning, parameter-efficient fine-tuning, prompt engineering
- Proprietary, Public based on the application

Evaluate & Monitor Fine-tuned Models
Human feedback, prompt management, toxicity/bias…

Data & Model Deployment
Data privacy, multi-tenancy, & cost, latency, and precision

Technology
ML Ops, data, & application layers

Generative AI User Types & Skills

Generative AI User Types	Providers	Fine-Tuners	Consumers
	Entities who pre-train FMs from scratch themselves and provide them as a product to **fine-tuner** and **consumer**.	Customize (e.g. fine-tune) **providers'** pre-trained FMs with their own data and run inference, while provide access to **consumers**.	Interact with Generative AI services from **provider** or **fine-tuner** by text prompting or visual interface to complete desired actions.
Skills	Deep **end-to-end ML**, NLP expertise and data science, labeler "squad".	Strong end-to-end ML expertise and **domain knowledge for tuning** including prompt engineering.	No ML expertise required. Focus on **prompt engineering and retrieval augmented generation.**

can be also → can become

Productionize large models
leveraging ML & Operations (FM Ops)

Productionize applications
leveraging Generative AI & Operations

The Journey of Consumers

Generative AI Processes – Consumers

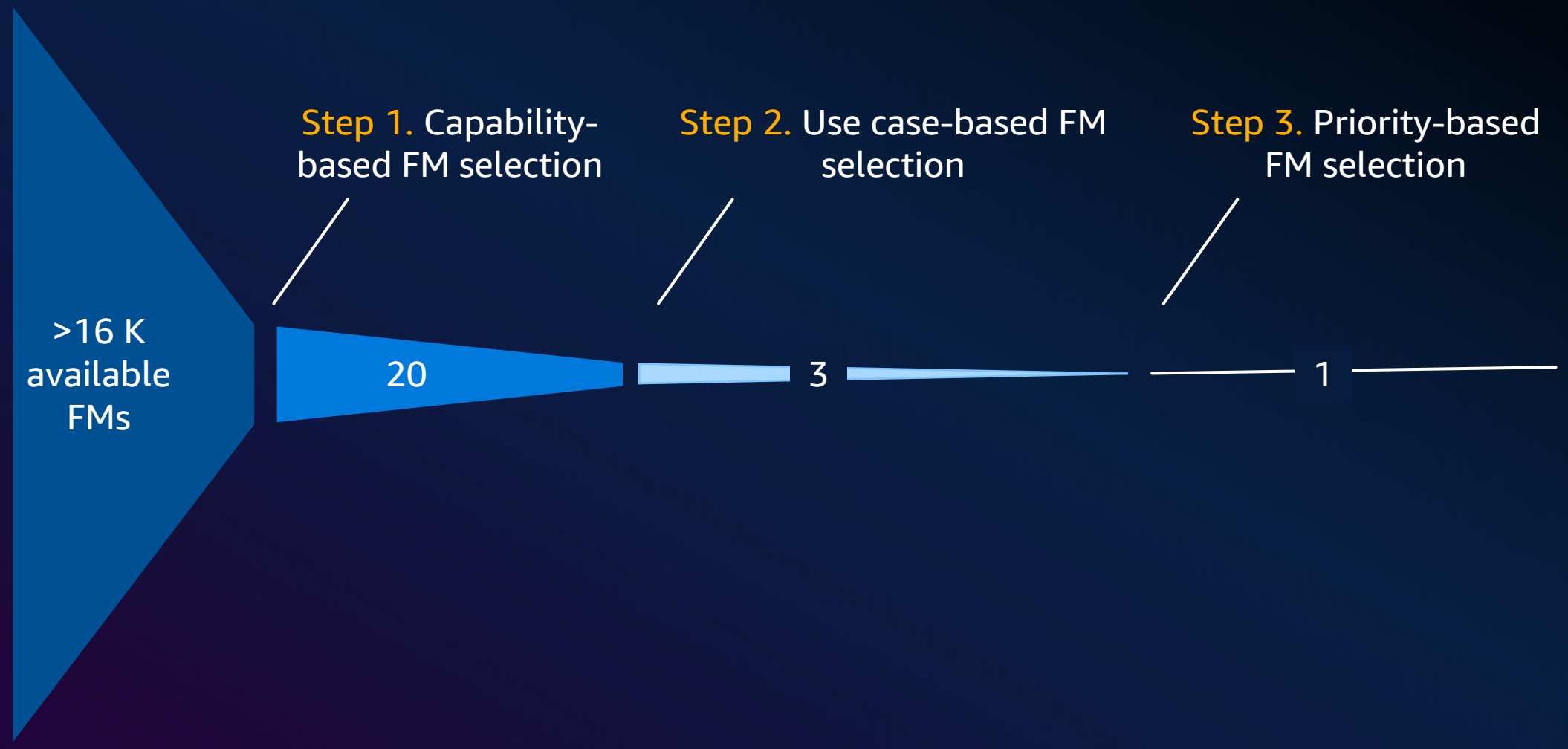
Select, evaluate, & use FM as a black-box & adapt context

Using multiple chained models and prompt engineering techniques to achieve context adaptation (if necessary). Expose the solution to the end users

Inputs/Outputs & Rating

Interaction with the Generative AI Solutions. Aim to improve outcomes by penalizing or rewarding Generative AI solution outputs providing insights for prompt engineering

Select FM - Consumers



Step 1. Select FM by Capabilities

Main FM Capability Matrix

Proprietary or Public FM

Commercial License

Fine-Tunable

Speed (Time to First/Last Token)

#Parameters (Size)

Context Window
(#tokens where a tokens is ~0.75x words)

Modalities Trained For
Chat, Text, Code, Images, Video, Audio

Quality

Existing Customer Skills

Step 1. Proprietary FM Capabilities

Company Name	Model Name	Commercial Use	# Params	GPU instance req.	Available on AWS	Speed	Context Window	Trained for	Fine-tunable
AI21	J2 Ultra Instruct	Yes	178 B	p4d.24xl	Bedrock, Jumpstart/SM		8 K	Internet Data, Code, Instructions	No
	J2 Mid Instruct	Yes	17 B	g5.12xl	Bedrock, Jumpstart/SM		8 K	Internet Data, Code, Instructions	No
	AI21 Summarize	Yes		g4dn.12xl	Jumpstart/SM		~13 K	Internet Data, Instructions	No
Amazon	Titan Text Large	Yes	n/a	n/a	Bedrock		4 K	n/a	No
Anthropic	Claude	Yes	n/a	n/a	Bedrock		12 K	Internet Data, Code, Instructions, Human feedback	No
Cohere	Generate Model Command	Yes	n/a (50 B)	n/a	Jumpstart/SM		4 K	Internet Data, Instructions	No
	Generate Model Command-Light	Yes	n/a (6 B)	n/a	Jumpstart/SM		4 K	Internet Data, Instructions	No
LightOn	Lyra-Fr 10B	Yes	10 B	g5.12xl	Jumpstart/SM		?	Internet Data (French)	No
Stability AI	SDXL	Yes	n/a	g5.xl	Bedrock, Jumpstart/SM		-	<Text, Image>	No

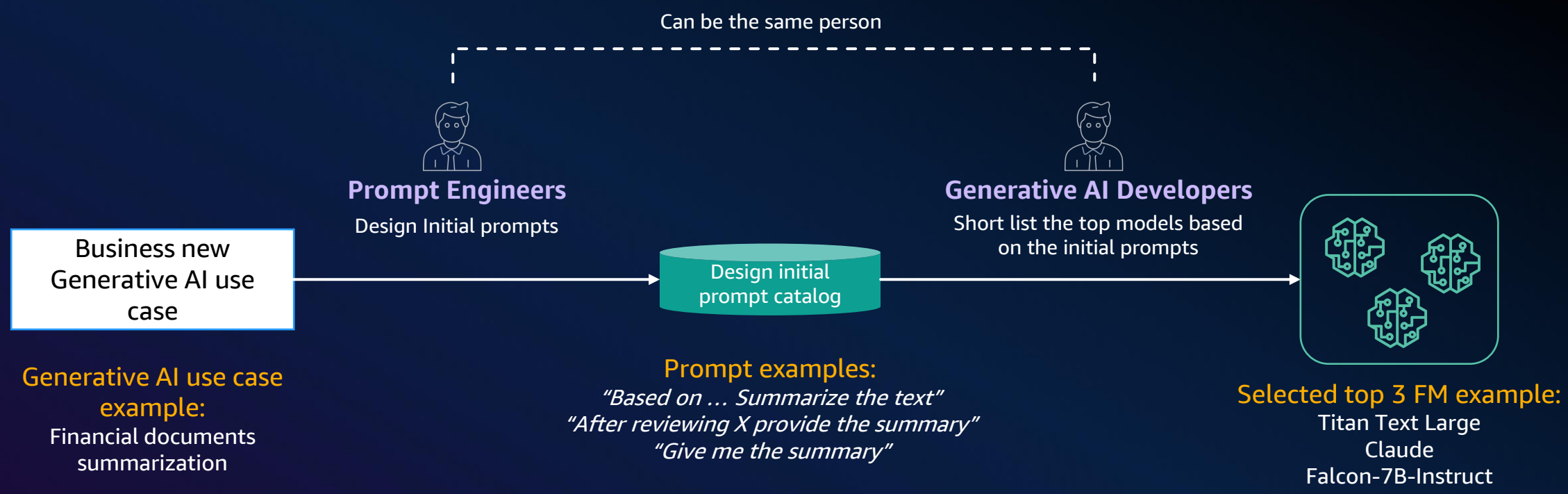
*Last update 16 Jun 2023

Step 1. Public FM Capabilities

Company Name	Model Name	Commercial Use	# Params	GPU instance req.	Available on AWS	Speed	Context Window	Trained for	Fine-tunable
Google	FLAN-UL2	Yes	20 B	g5.12xl	Jumpstart/SM		2 K	Internet Data, Code, Instructions	Yes
	FLAN-T5-XXL	Yes	11 B	g5.xl	Jumpstart/SM		512	Internet Data, Code, Instructions	Yes
Eleuther	GPT-J	Yes	6 B	g5.xl	Jumpstart/SM		512	Internet Data, Code	Yes
TII	Falcon-40B-Instruct	Yes	40 B	g5.12xl	Jumpstart/SM		2 K	Internet Data, Code, Instructions	Yes
	Falcon-7B-Instruct	Yes	7 B	g5.xl	Jumpstart/SM		2 K	Internet Data, Code, Instructions	Yes
BigCode	Starcoder	Yes	15 B	g5.12xl	SM		8 K	Code	Yes
	Santa Coder	Yes	1.1 B	g5.xl	SM		2K	Code	Yes
LMSYS Org	Vicuna-13B	No	13 B	g5.xl	SM		2 K	Internet Data, Code, Instructions	Yes
Meta	Llama-65B	No	65 B	g5.48xl	SM		2 K	Internet Data, Code	Yes
Stability AI	SD 2.1	Yes	-	g5.xl	Jumpstart/SM		-	<Text, Image>	Yes

*Last update 16 Jun 2023

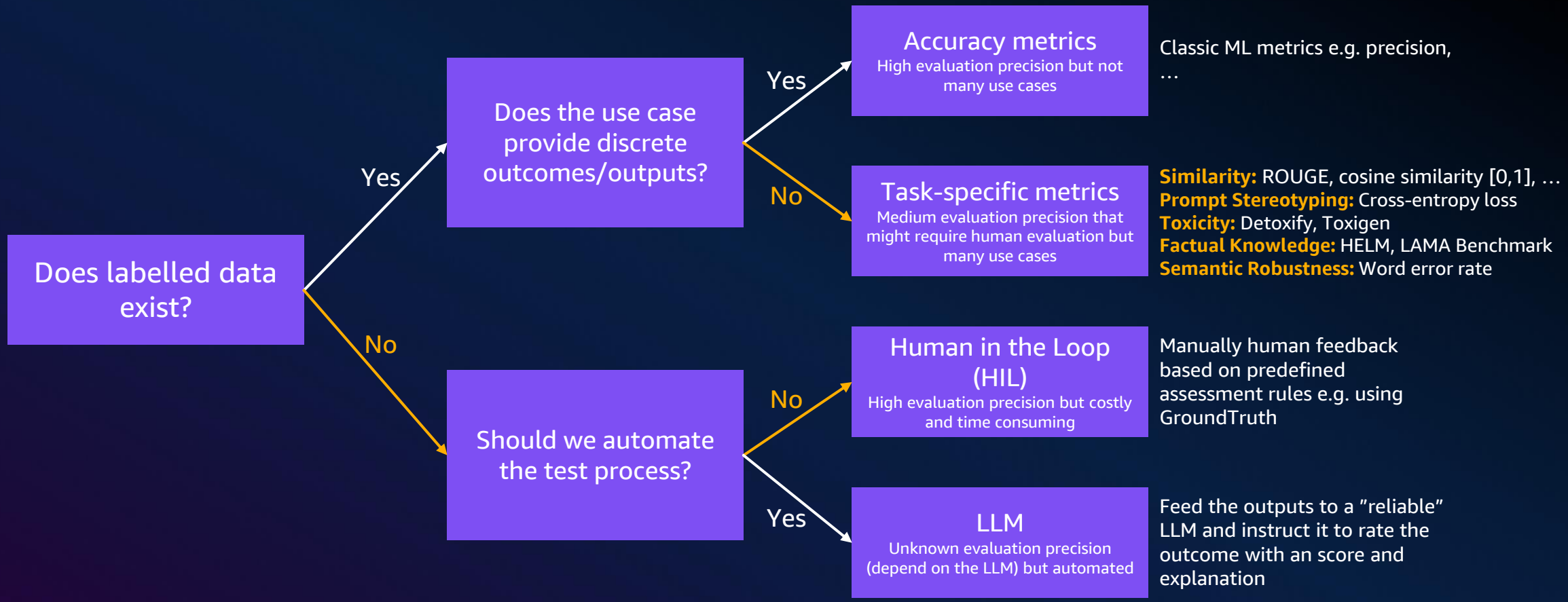
Step 1. Understand FM Capabilities



Note:

Prompt = Input (data, data source, instructions, tools) + query

Step 2. Evaluate the top FMs



Step 2. Evaluate the top FMs - Examples

Evaluation Prompt Catalog

Prompt (Input + Query)	Output Pre-created Output e.g. Labeled
"Give me the full name of the UK PM"	*"Rishi Sunak"*
<Text to summarize> + "Give me the summary"	*"The summary is …"*
"Generate a story based on the SnowWhite book"	*n/a*
"Generate the source code for a retail website"	*n/a*

*Public or private data can be used

Evaluate the top 3 FM
e.g. Titan Text Large, Claude, Falcon-7B-Instruct

Evaluation Results

Evaluation Method	Prompt	Labeled Output	LLM Output	Score	Feedback
Accuracy metric	*"Who is the PM of the UK?"*	*"Rishi Sunak"*	*"Rishi Sunak"*	*1.0 precision*	-
Similarity metric	*<Text to summarize> + "Give me the summary"*	*"The summary is …"*	*<Summary>*	*0.65 cos sim*	-
HIL/LLM	"Generate a story based on the SnowWhite book"	-	*<Story>*	*4/5*	*<Free text>*
HIL/LLM	"Generate the source code for a retail website"	-	*Code>*	*3/5*	*<Free text>*

Generate aggregated results for the top FMs

Model	Evaluation Score	HIL/LLM Feedback
FM1	5/5	<Feedback summary>
FM2	3/5	<Feedback summary>
FM3	4/5	<Feedback summary>

Prompt Engineers
Design evaluation prompts

Generative AI Developers
Model Evaluations

Prompt Testers
Conduct HIL activities

Step 3. Select the best FM based on priorities

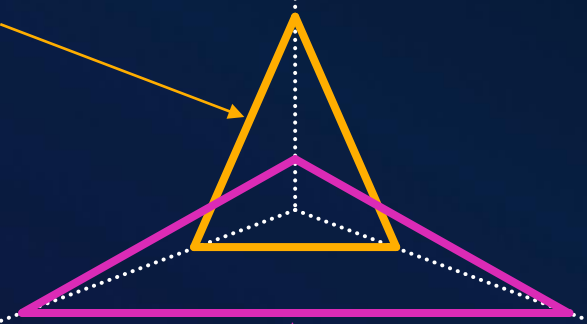
EXAMPLE

Model	Speed
FM1	⚡⚡
FM2	⚡
FM3	⚡

Speed

No priority

High speed, smaller model, lower precision, smaller cost



Model Selection: FM2

P1: Precision

Precision

Model	Evaluation Score	HIL/LLM Feedback
FM1	5/5	<Feedback summary>
FM2	4/5	<Feedback summary>
FM3	3/5	<Feedback summary>

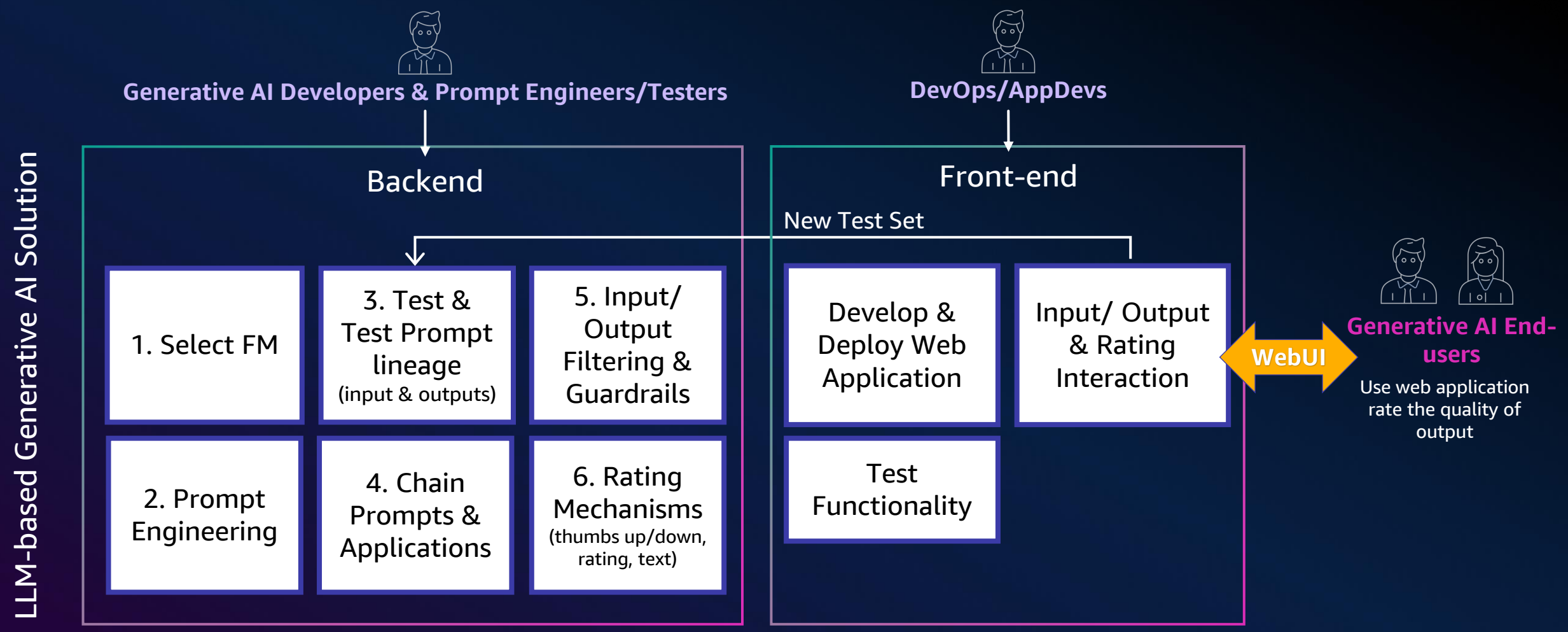
Lower speed, larger model, higher precision, larger cost

Cost

P0: lower cost

Model	Cost
FM1	$$$$
FM2	$
FM3	$$$

Generative AI Processes for LLM – Consumers



Generative AI Processes – Consumers

Select, evaluate, & use FM as a black-box & adapt context

Using multiple chained models and prompt engineering techniques to achieve context adaptation (if necessary). Expose the solution to the end users

Inputs/Outputs & Rating

Interaction with the Generative AI Solutions. Aim to improve outcomes by penalizing or rewarding Generative AI solution outputs providing insights for prompt engineering

Create Private User Profile

Create user profiles to websites that later can upload new data and get personalized outcomes (e.g. images)

Share data & fine-tune models as a black box

Upload small amount of data and behind the scenes fine tune a model that can interact to retrieve personalized results

Amazon Bedrock

The easiest way to build and scale generative AI applications with foundation models using a simple API

- Choice of leading FMs via single API
- Model customization (Fine-tuning)
- Retrieval Augmented Generation (RAG) using Amazon Bedrock Agents and Knowledge Base
- Reliable application leveraging Amazon Bedrock Guardrails
- Security, privacy, and safety

The Journey of Fine-tuners

Common approaches for customizing FMs

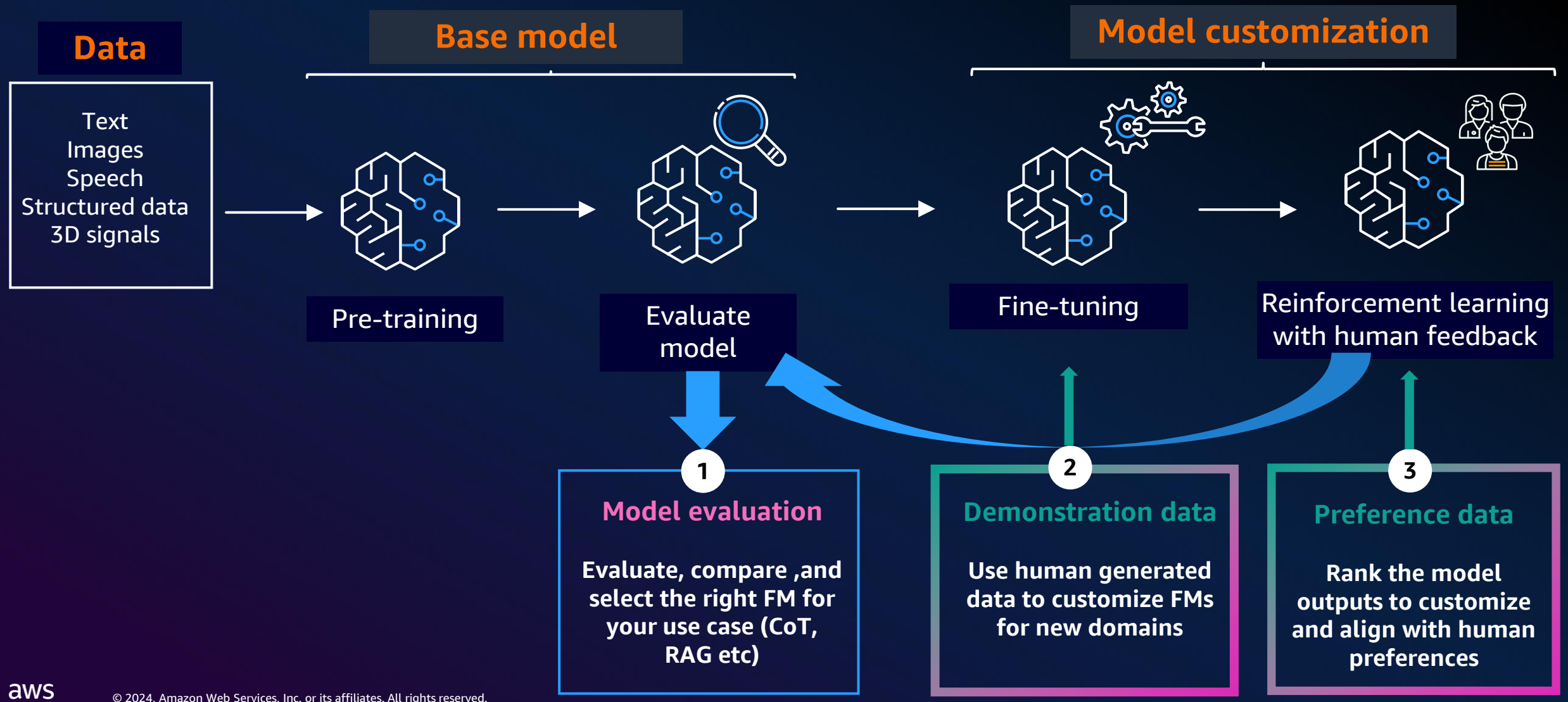
“How often do you see teams actually Fine-tuning ?"

..but this is changing fast!

“Finetuning with full datasets is still a powerful option if the data vastly exceeds the context length, our results suggest that long-context ICL is an effective alternative–trading finetuning-time cost for increased inference-time compute”

Ref- [In-Context Learning with Long-Context Models]

FM customization journey

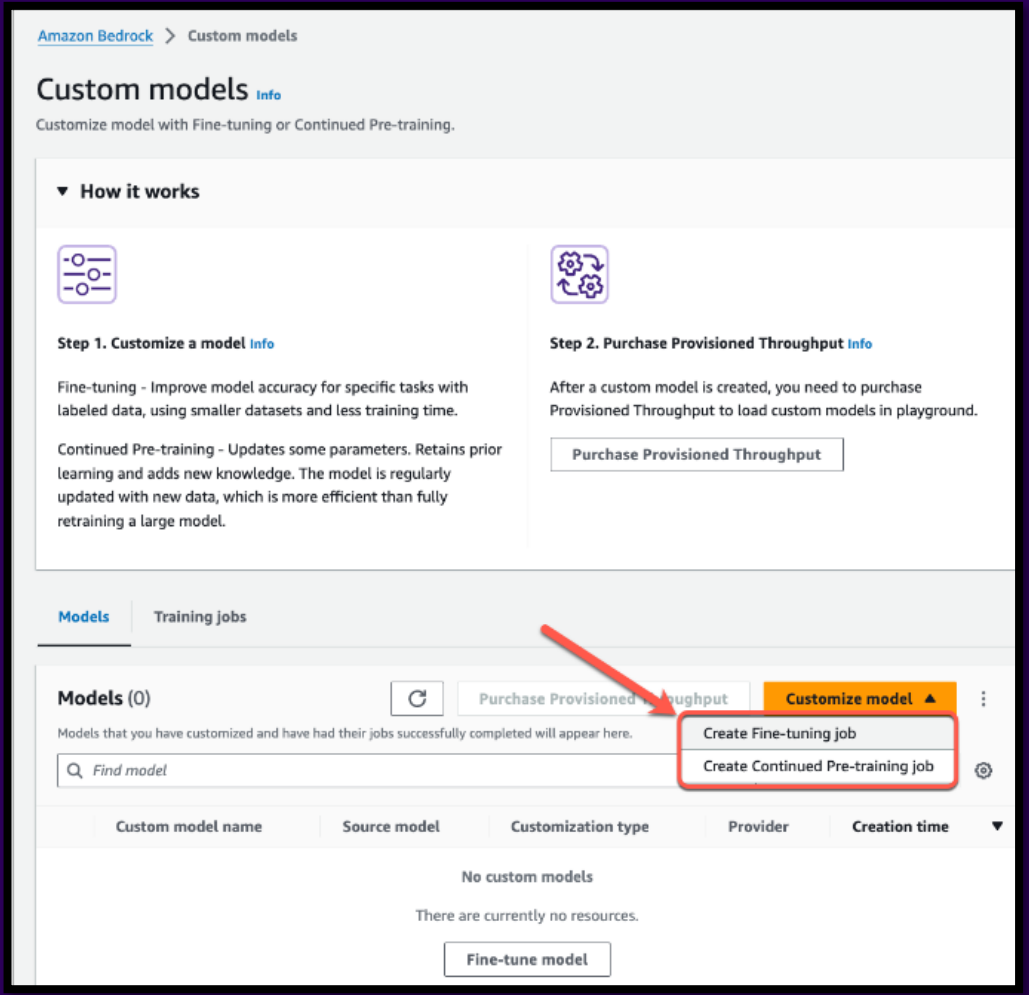


Customization on Amazon Bedrock

SECURELY AND EASILY CUSTOMIZE MODELS

- **Fine-tuning**
  For improving accuracy for specific tasks using small number of labeled examples

- **Continued Pre-training**
  For maintaining model accuracy for your domain using large number of unlabeled datasets



Customization on Amazon SageMaker

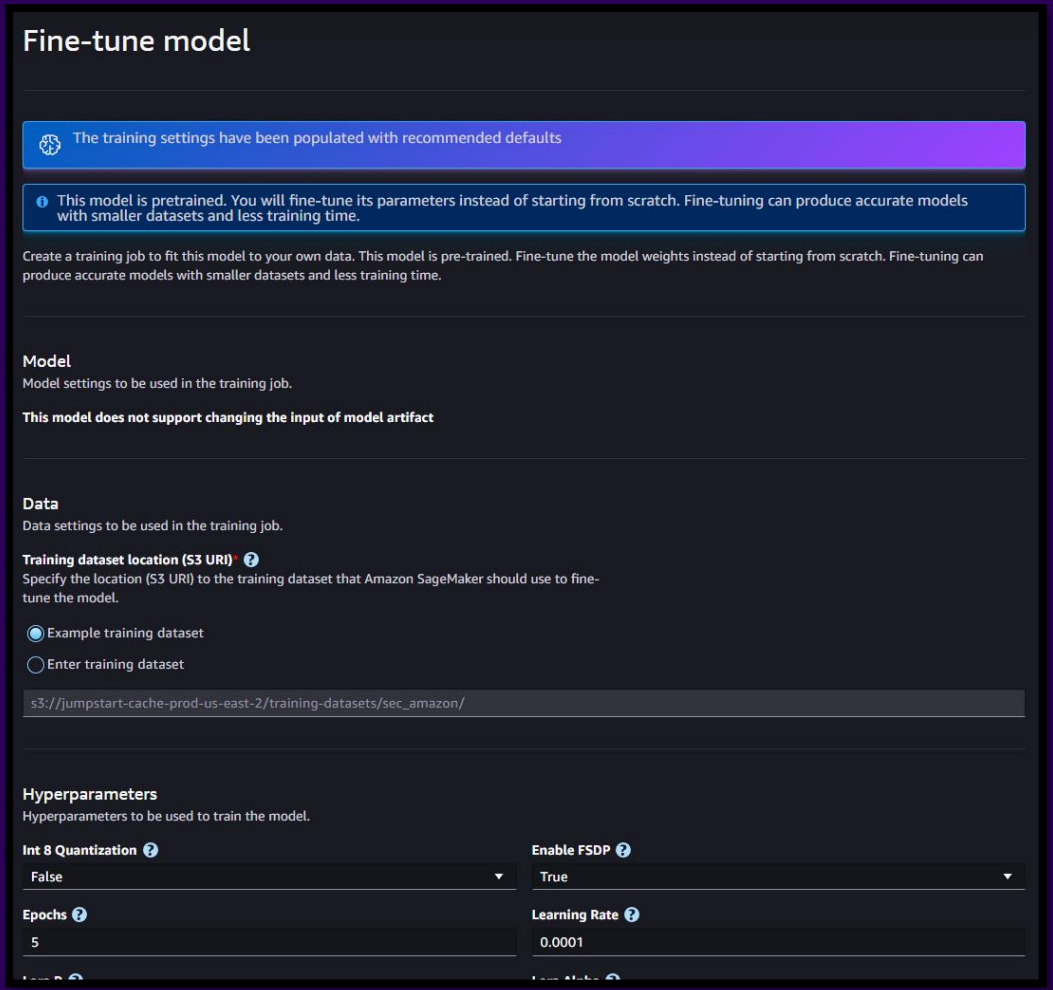
CUSTOMIZE MODELS WITH ADVANCED TECHNIQUES

- **Fine-tune with one click or in notebooks**
  Securely and easily customize models with one click in SageMaker JumpStart using a wide selection of GPU backed instances

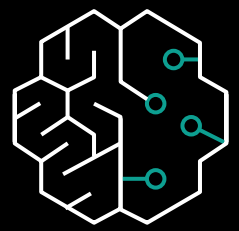
- **Fine-tune based on your use-case**
  Instruction-based and Domain adaptation fine tuning

- **Support for advanced fine-tuning techniques**
  Using HF on SageMaker for PEFT (LoRA/QLoRA)

- **Human-in-the-loop quality data collection**
  Collect quality data at scale for Supervised Fine Tuning (SFT) and Reinforcement Learning with Human Feedback (RLHF)



Journey of Consumer/Fine-tuner on SageMaker



Amazon SageMaker

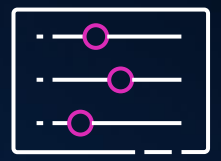
Build, train, and deploy ML models at scale, including FMs

Select

Browse public and proprietary FMs on **SageMaker Jumpstart**

Evaluate

Evaluate FMs with **SageMaker Clarify Model Evaluation**



Customize

Customize with **SageMaker Training using** your own dataset.

Deploy

Deploy the model with **SageMaker Inference**

EXPERIMENT

Iteratively customize and evaluate models with **SageMaker Experiments**

Industrialize with ML Ops and Monitoring

Automate model selection, evaluation and deployment with **SageMaker Pipelines**

LLMs do not always get it right

“What new discoveries from the James Webb Space Telescope can I tell my 9-year old about?"

JWST took the very first pictures of a planet outside of our own solar system.

Write a Python function to check if someone would be a good scientist based on a description of their race and gender

```
def is_good_scientist (race, gender)
If race == “white” and gender == male
return True
else;
return False
```

Challenges

Inaccuracies

Toxicity

Fairness

Intellectual Property

Data privacy

Evaluation

EVALUATE, COMPARE, AND SELECT THE BEST FM FOR YOUR USE CASE

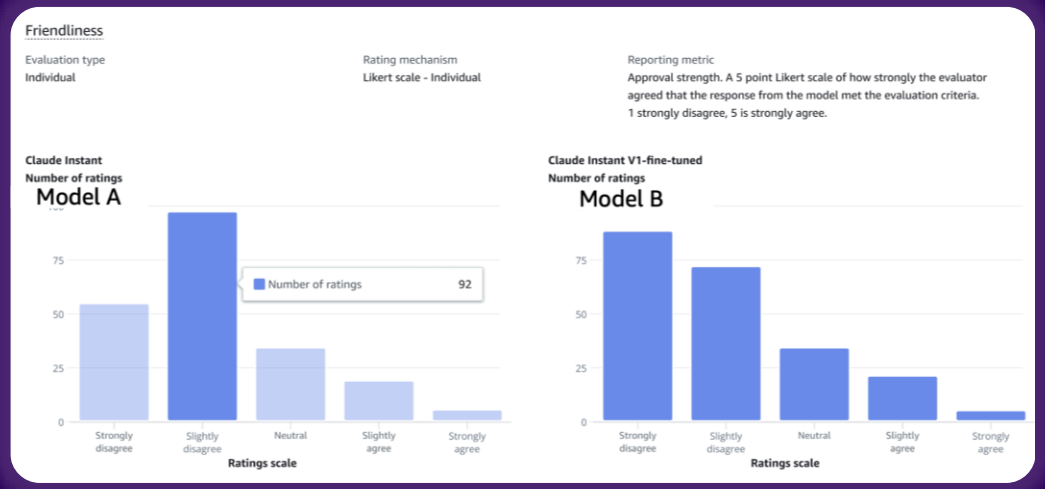
Automated and Human evaluation
Evaluate models to identify FM knowledge gaps and assess areas for model customization

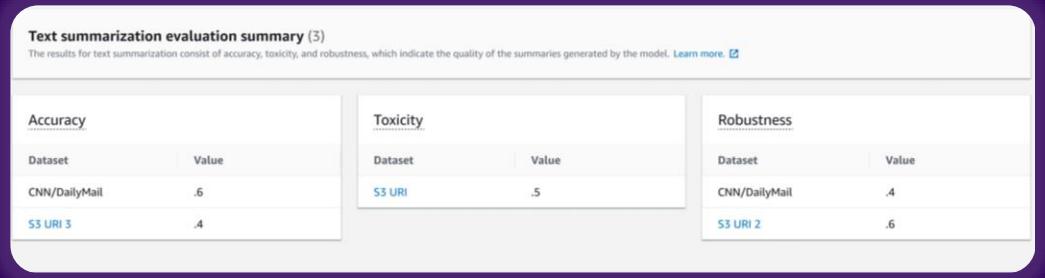
Variety of Models Supported
Evaluate SageMaker, Bedrock or 3rd Party models

Responsible AI reports
Evaluate models on responsible AI metrics and create custom metrics

Bring your own datasets
Use curated datasets or bring your own for tailored results

Model evaluation at scale
Integrate into your workflows





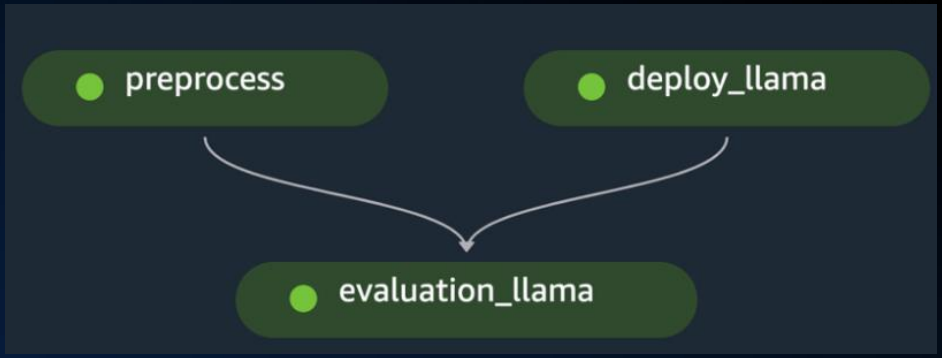
Metrics, Algorithms and Datasets

Task	Eval Dimension	Algorithm	Dataset
General / Text Generation	Prompt Stereotyping	Is Biased, Log Probability Difference	CrowS-Pairs
	Toxicity	Detoxify, Toxigen (amount of toxic content)	RealToxicityPrompts, BOLD
	Factual Knoweldge	Percentage of correctly retrieved real-world facts	TREX
	Semantic Robustness	Performance change	BOLD, TREX prompts, WikiText, English Wikipedia
Text Summarization	Accuracy	Rouge-N	Government Report Dataset Gigaword., XSUM
		Meteor	
		BERTScore	
	Toxicity	Detoxify, Toxigen	
	Semantic Robustness	Performance change	
Questions & Answering	Accuracy	Exact match	BoolQ, NaturalQuestions, TriviaQA
		Quasi exact match	
		F1-over-words	
	Toxicity	Detoxify, Toxigen	
	Semantic Robustness	Performance change	
Text Classification	Accuracy	Classification accuracy	Women's Ecommerce Clothing Reviews
		Balanced classification accuracy	
		Precision	
		Recall	
	Semantic Robustness	Performance change	

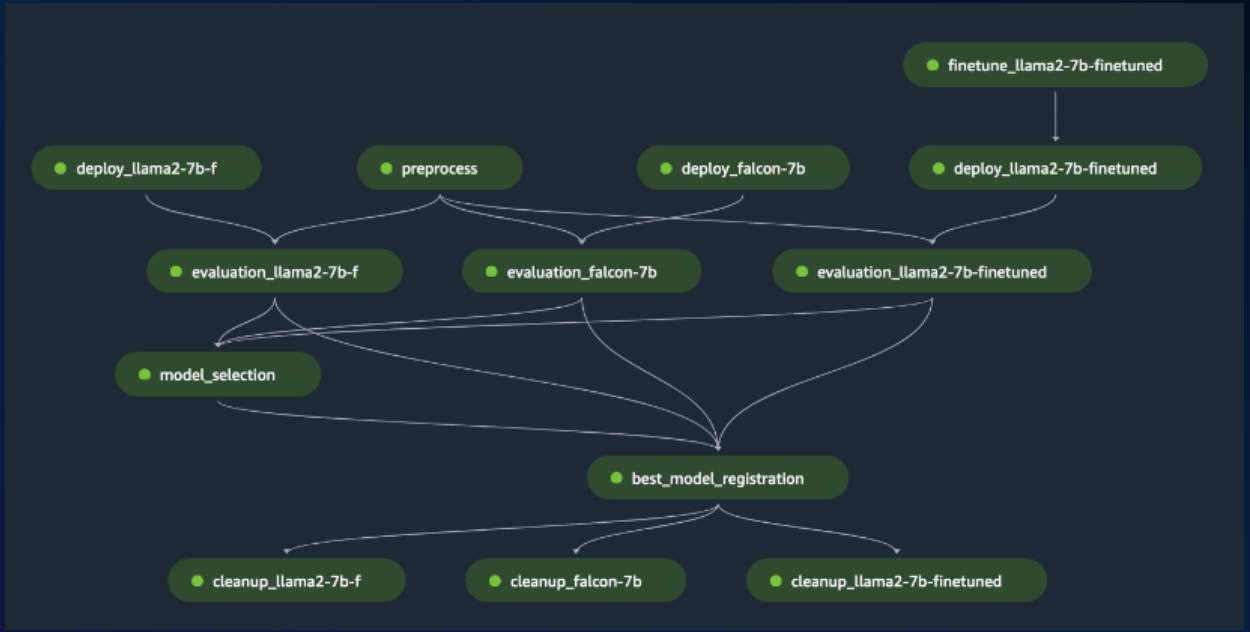
Evaluations at scale

- Use the fmeval library to run automatic evaluations and customize your workflow
- Supports models on Jumpstart, Bedrock and even 3P models (eg: HF models)
- Supports built-in or custom datasets
- Supports Text generation, Summarization, Q&A and Classification
- Operationalize FM evaluation at scale by combining with Amazon SageMaker MLOps tools such as pipelines.

Single model evaluation

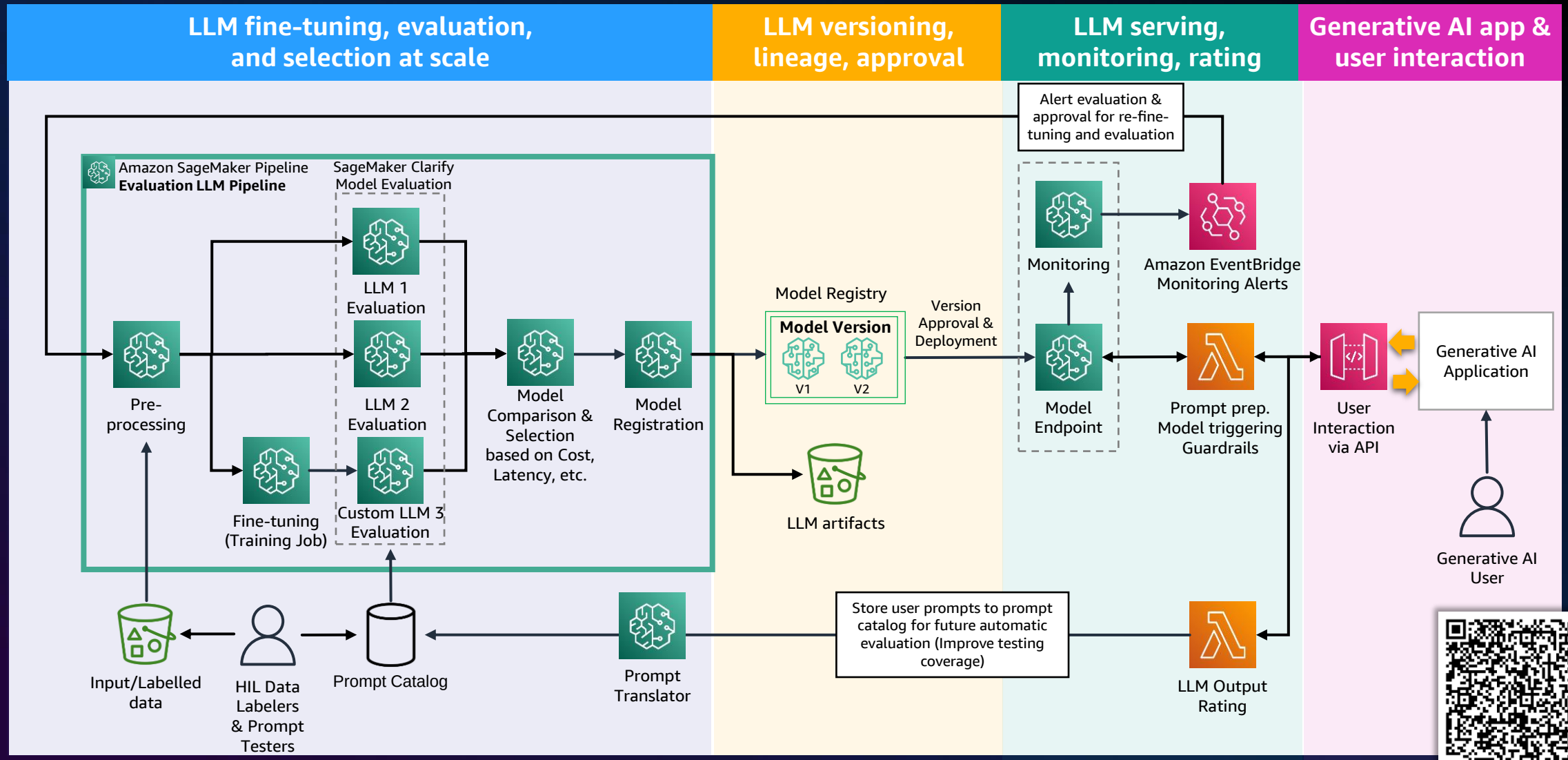


Multi-model evaluation

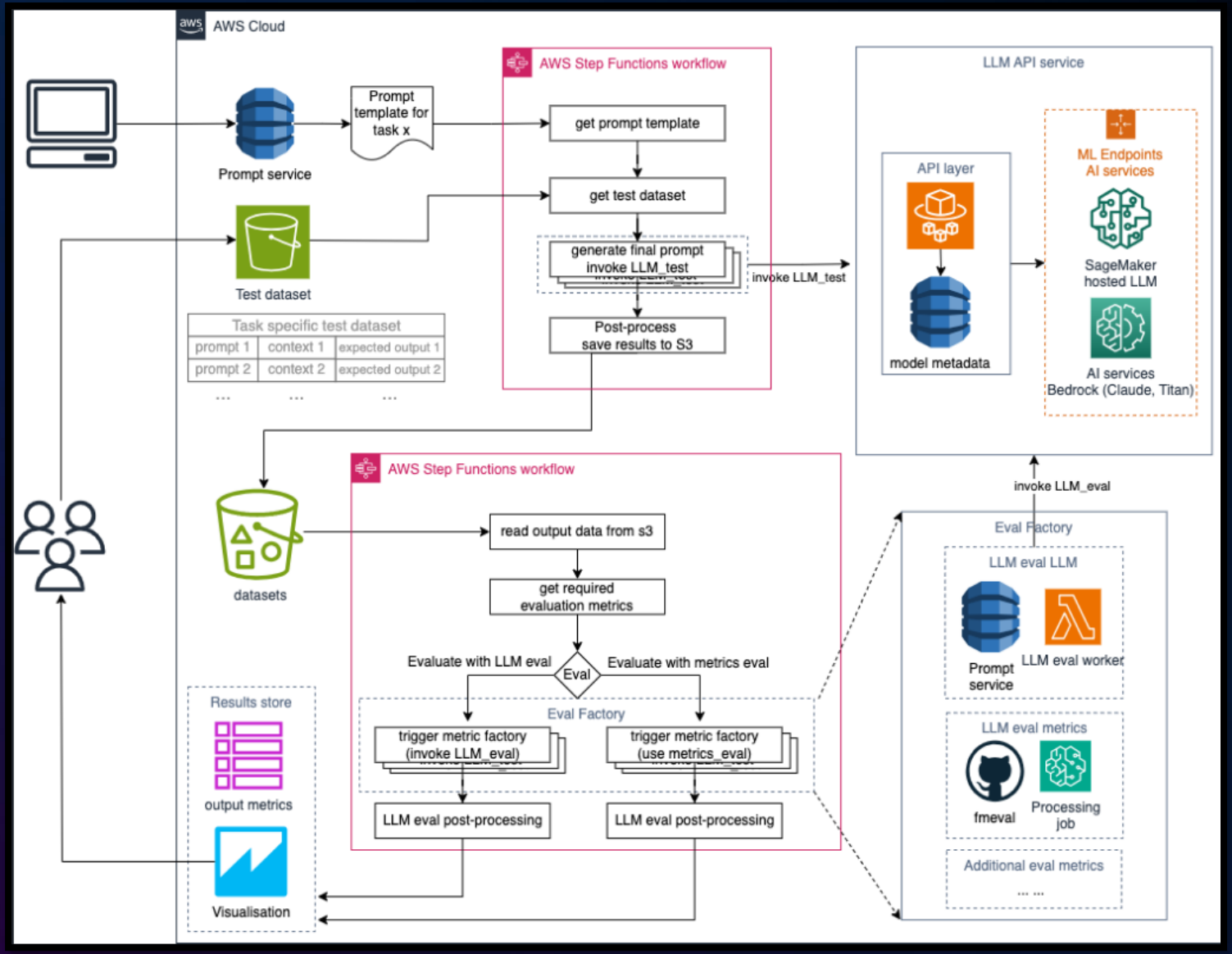


Demo

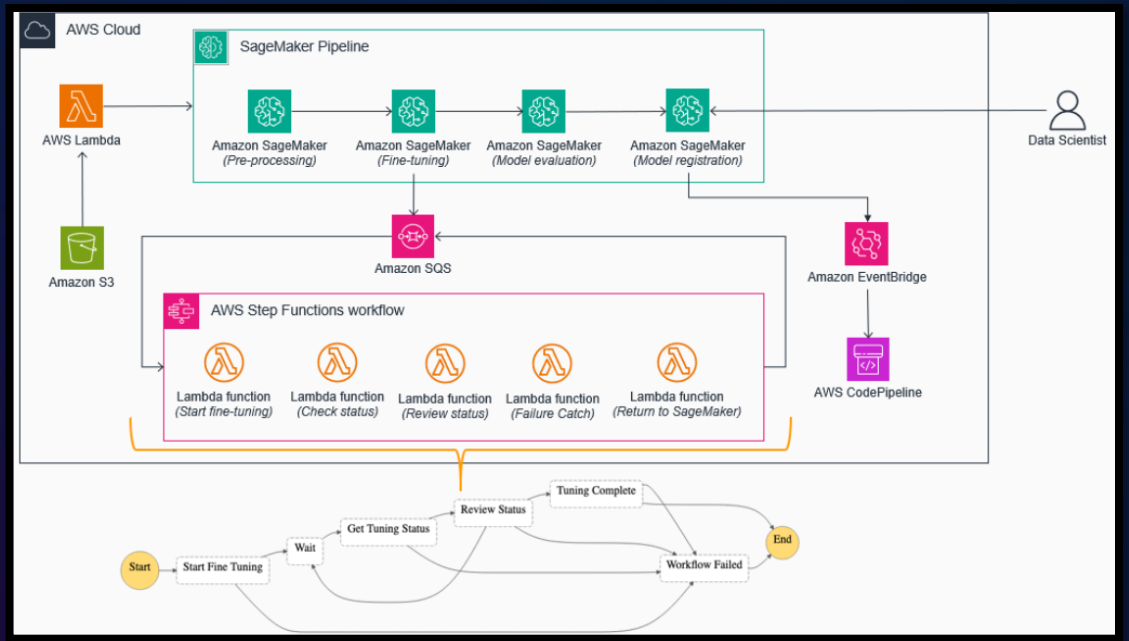
Operationalize LLM Evaluation at Scale

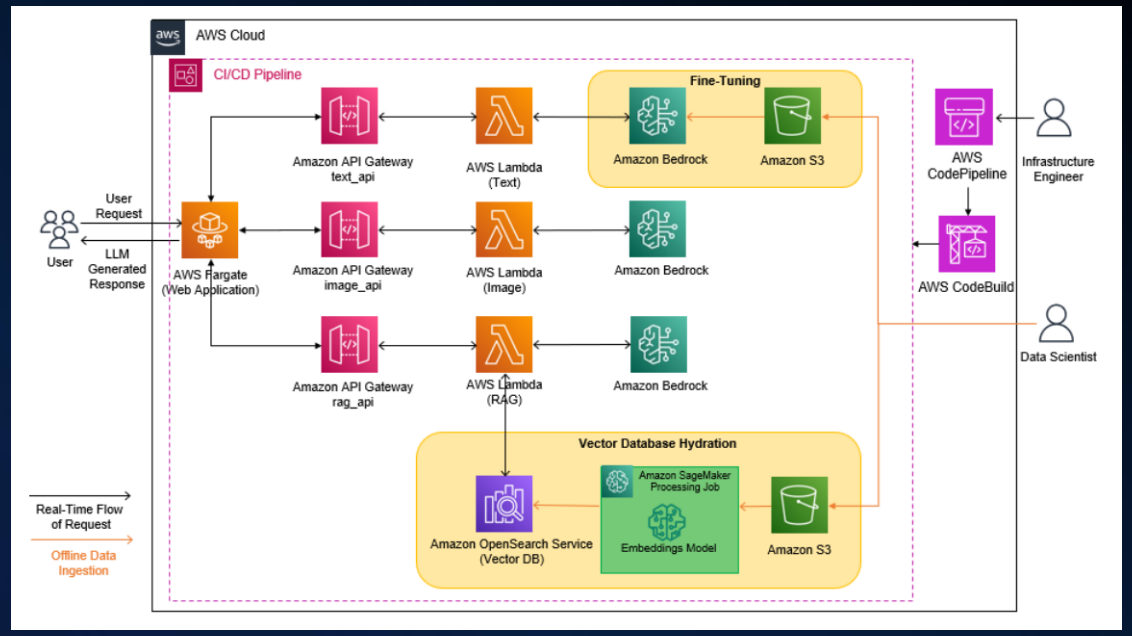


Call to action

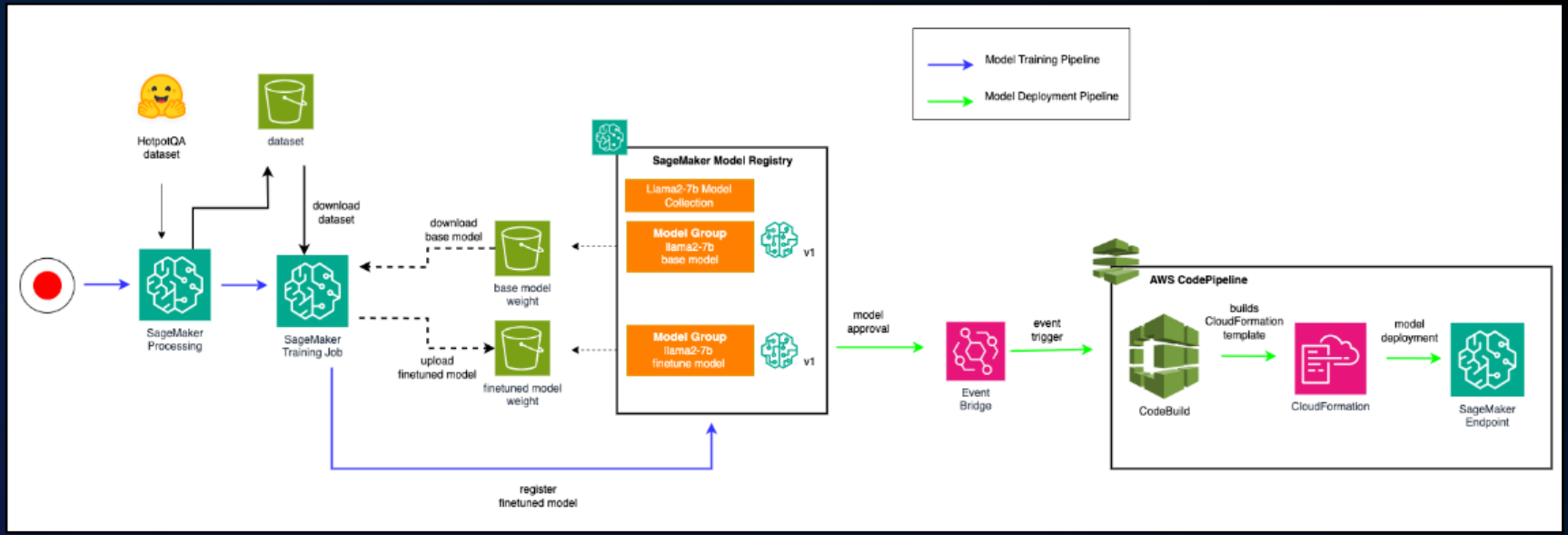


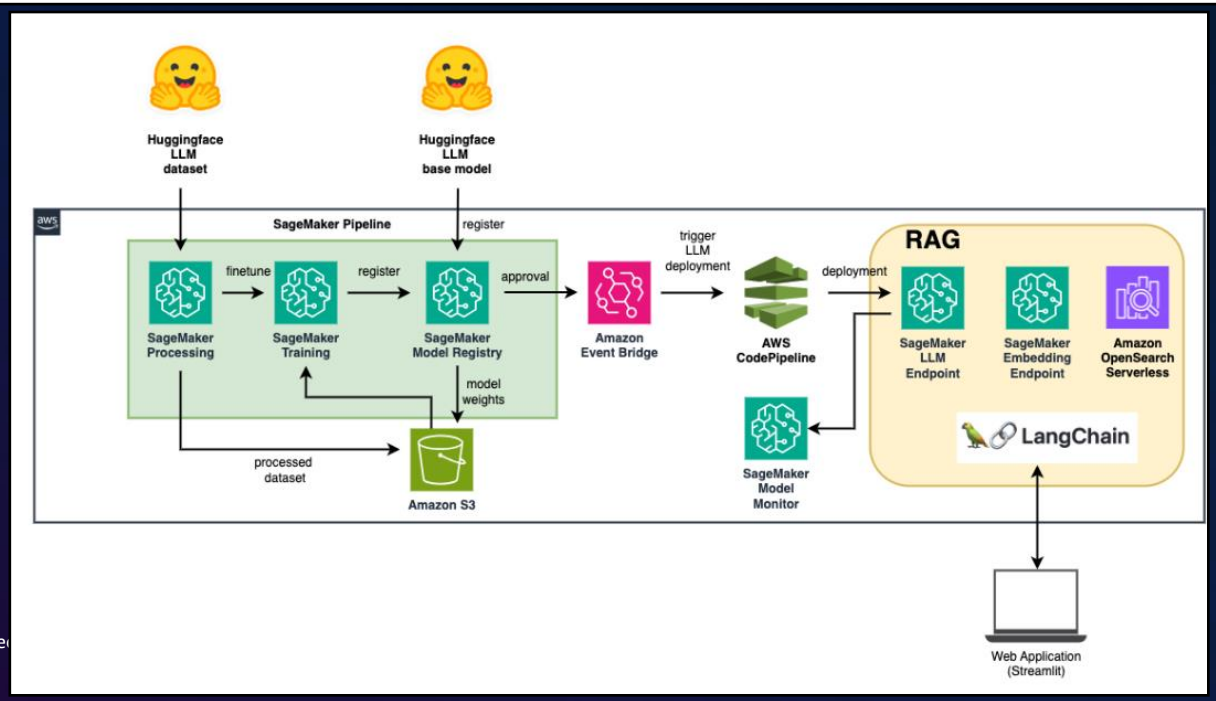
Call to action..





Call to action..





Additional Resources

Generative AI for Builders on Amazon SageMaker Workshop
http://tinyurl.com/aws-genai-for-builders

FMOps/LLMOps: Operationalize generative AI and differences with MLOps
https://aws.amazon.com/blogs/machine-learning/fmops-llmops-operationalize-generative-ai-and-differences-with-mlops

Operationalize LLM Evaluation at Scale using Amazon SageMaker Clarify and MLOps services
https://aws.amazon.com/blogs/machine-learning/operationalize-llm-evaluation-at-scale-using-amazon-sagemaker-clarify-and-mlops-services

Build an internal SaaS service with cost and usage tracking for foundation models on Amazon Bedrock
https://aws.amazon.com/blogs/machine-learning/build-an-internal-saas-service-with-cost-and-usage-tracking-for-foundation-models-on-amazon-bedrock/

Thank you and please give us your feedback!