FMOps/LLMOps: Operationalise Generative AI using MLOps principles

Dr Sokratis Kartakis (he/him)
Snr MLOps Solutions Architect, EMEA
Amazon Web Services

Heiko Hotz (he/him)
Snr LLM Solutions Architect, EMEA
Amazon Web Services

Rahul Srivastava (he/him)
Snr Enterprise Solutions Architect, EMEA
Amazon Web Services

MLOps Foundation Expected Outcomes

STANDARDIZE OPERATIONS AND INFRASTRUCTURE FOR YOUR DATA SCIENCE

	Business Goal	Technical Metric	Before MLOps	MLOps Expected Outcomes	Business Value
1	Be more efficient in delivery	Time to value (from idea to production)	up to 12 months	< 3 months	Improve Speed-to-Value by 4x
2	Simplify route-to-live	Time to productionize existing ML use cases	3-6 months	< 2 weeks	Reduce FTE overhead in average 8x
3	Standardize infrastructure, data, & code	% Template driven development	n/a	> 85%	Focus on innovation increasing re-usability by 85%
4	Standardize onboarding of new teams and ML use cases	Time to instantiate a new MLOps infrastructure & ML projects	40 days	< 1 hours	Accelerate ML adoption across all business areas
5	Ensure high security standards	Execute the ML solutions without internet access in a private cloud	n/a	No internet	Your data is safe in your private cloud
	Reduce platform, people and operation costs				

Customer references building MLOps foundation and business benefits:

- NatWest: https://aws.amazon.com/solutions/case-studies/natwest-group-case-study
- BP: https://aws.amazon.com/solutions/case-studies/bp-machine-learning-case-study

MLOps Key Personas and Roles

Advance Analytics Team
Data Lake

Data Engineer
Prepare & Ingest data building ETL pipelines

Data Owners
Manage data sharing and provide access

Data Science Team
Experimentation & MLOps

Data Scientist
Create the best ML models to solve business problems

ML Engineer
Collaborate with DS to productionize ML

Platform Team
Secure Cloud/Data/ML Platform

MLOps Engineer/Admin
Standardize CI/CD, user/service role, model consumption, testing and deployment methodology

Security
Assess data, user, and service access creating policies and guardrails

Architects/ SysOps Engineer
Standardize account infrastructure, connectivity, user roles implementation

Business
Viz Dashboards, ML Adoption, & ROI

Business Stakeholder
Product Owners
Define business problem, business KPIs, and make business decisions

Business Stakeholder
Data & ML Consumers
Consumers of ML results from other BUs, driving business decision making

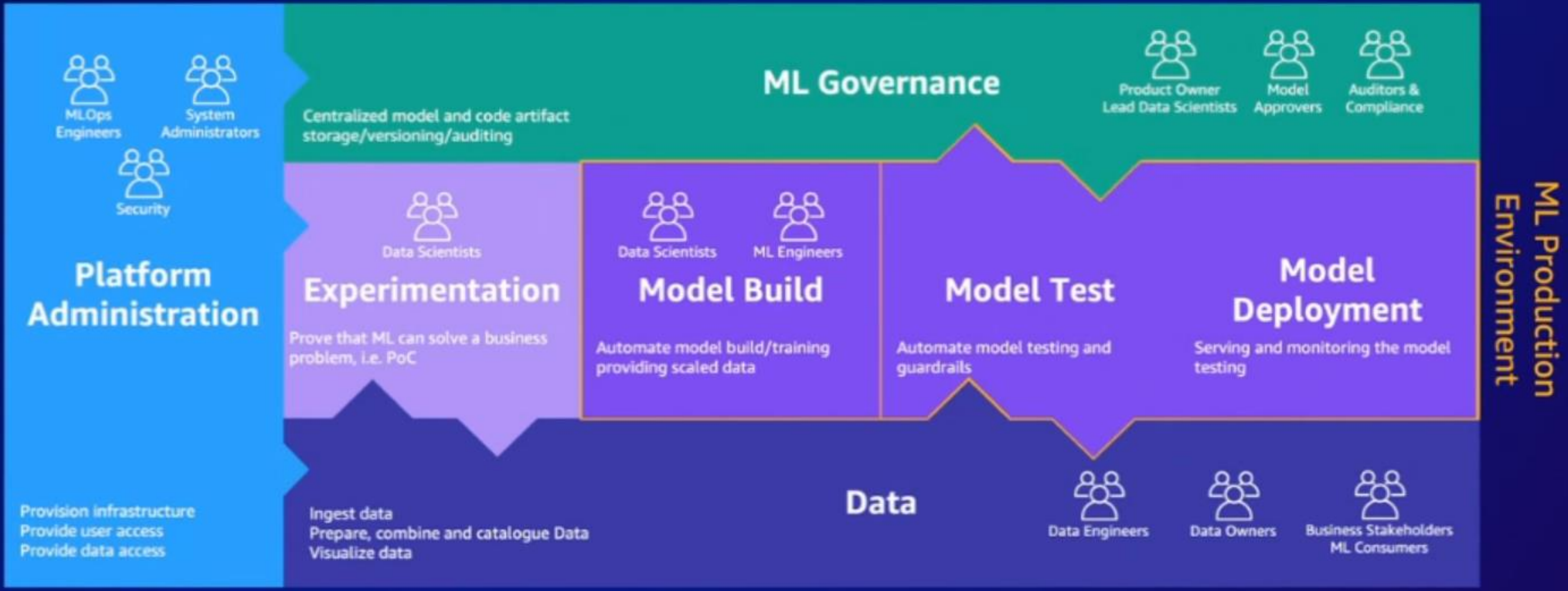
Risk & Compliance
Approve & Review Models

Auditors/Risk & Compliance
Review models, data sources, code artifacts

aws

MLOps Foundation People & Processes

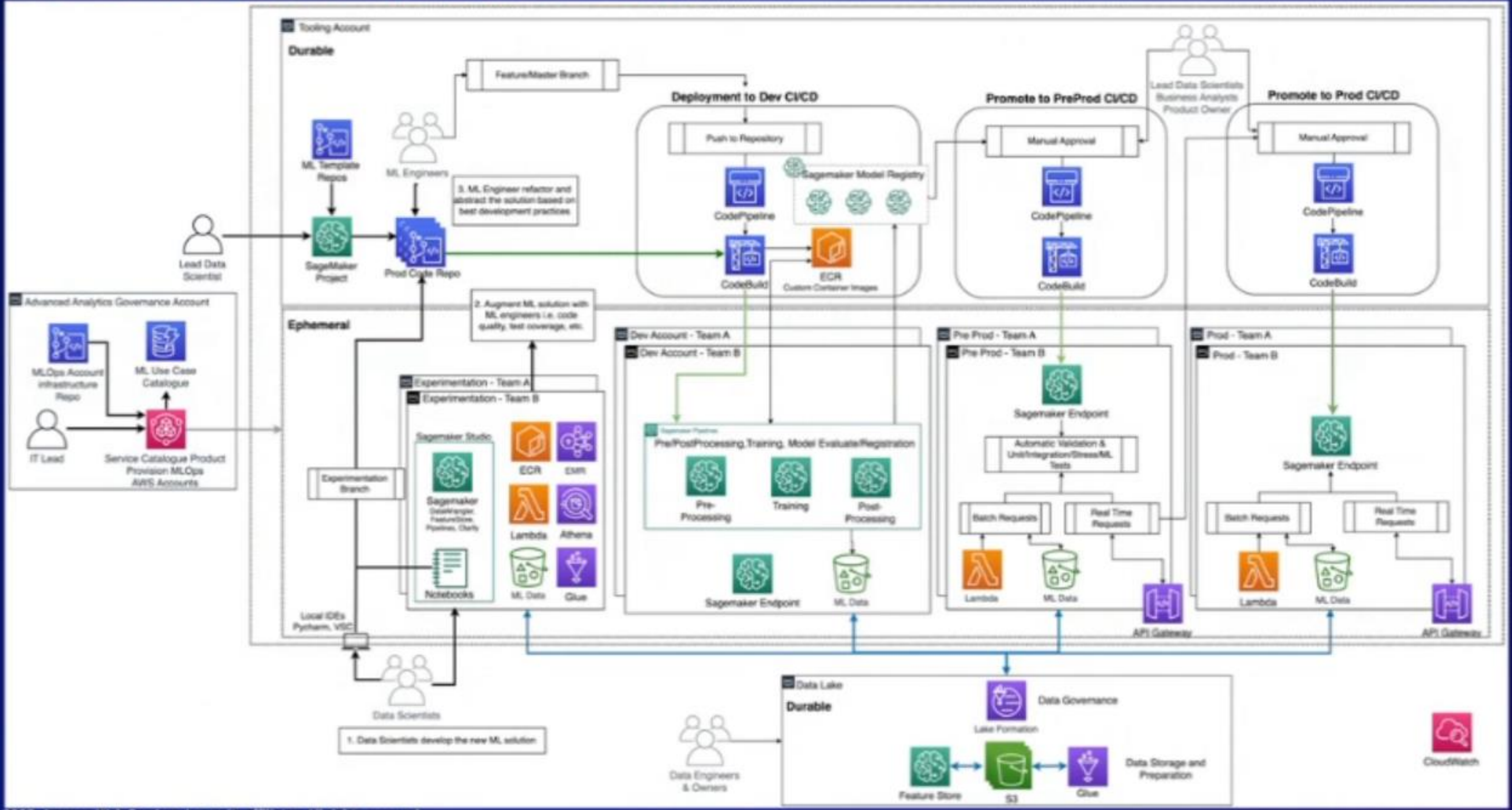
SEPARATION OF CONCERNS IS KEY FOR SUCCESS



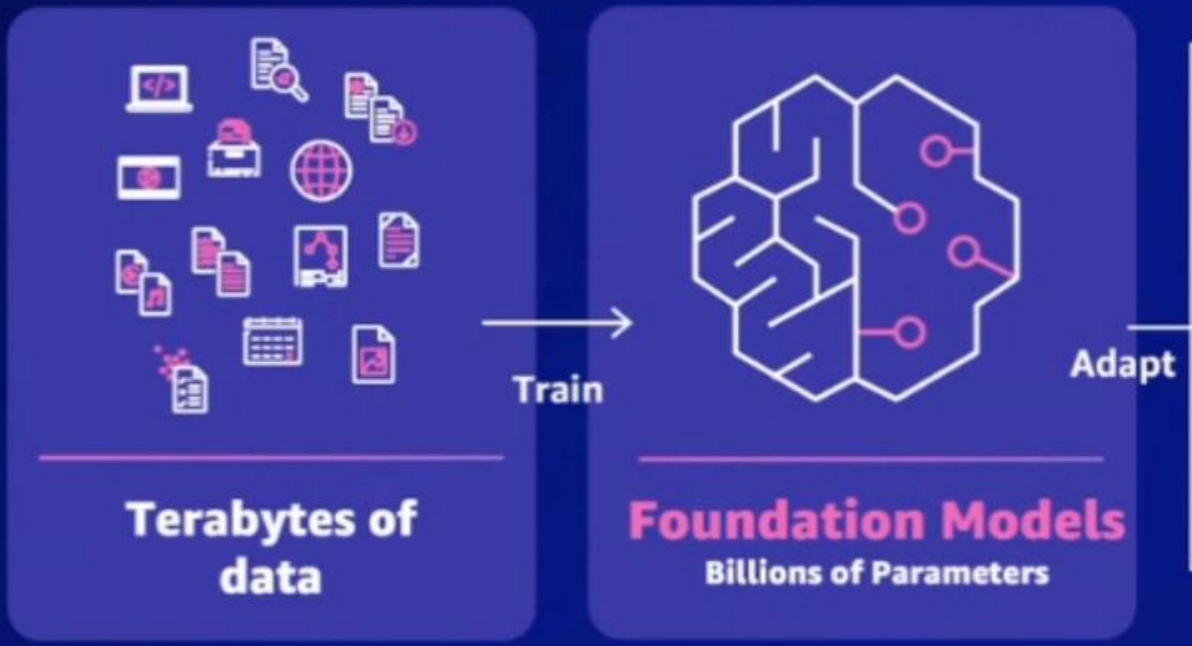
MLOPs Scalable Phase

MULTIPLE TEAMS AND ML USE CASES ADOPT MLOPS

MLOps foundation roadmap for enterprises
https://aws.amazon.com/blogs/machine-learning/mlops-foundation-roadmap-for-enterprises-with-amazon-sagemaker



GenAI 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

MLOps

FMOps

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

Key Definitions

Machine Learning Operations
Productionize ML solutions efficiently

MLOps

FMOps

Foundation Model Operations
Productionize GenAI Solutions (Text-Text/ Image/ Video/ Audio/ ...)

LLMOps

Large Language Model Operations
Productionize Large Language Model-based solutions

MLOps & FMOps Differentiators

Processes & People
Providers, fine-tuners, & consumers

Select & Adapt the FM on a Specific Context
- Fine-tuning, parameter-efficient fine-tuning, prompt engineering
- Proprietary, open source 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
MLOps, data, & application layers

GenAI User Types & Skills

Providers — can be also → Fine-Tuners ← can become — Consumers

	Providers	Fine-Tuners	Consumers
Generative AI User Types	Entities who build foundation models from scratch themselves and provide them as a product to **tuner** and **consumer**.	Fine-tune foundational models from **providers** to fit custom requirements. Orchestrate the deployment of the model as a service for use by **consumers**.	Interact with Generative AI services from **provider** or **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 knowledge of model deployment and inference. Strong domain knowledge for tuning including prompt engineering.	No ML expertise required. Mostly application developers or end-users with understanding of the service capabilities. Only prompt engineering is required for better results.

MLOps is required

Productionize applications where DevOps/AppDev is more relevant than MLOps

aws

GenAI 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 GenAI Solutions. Aim to improve outcomes by penalizing or rewarding GenAI solution outputs providing insights for prompt engineering

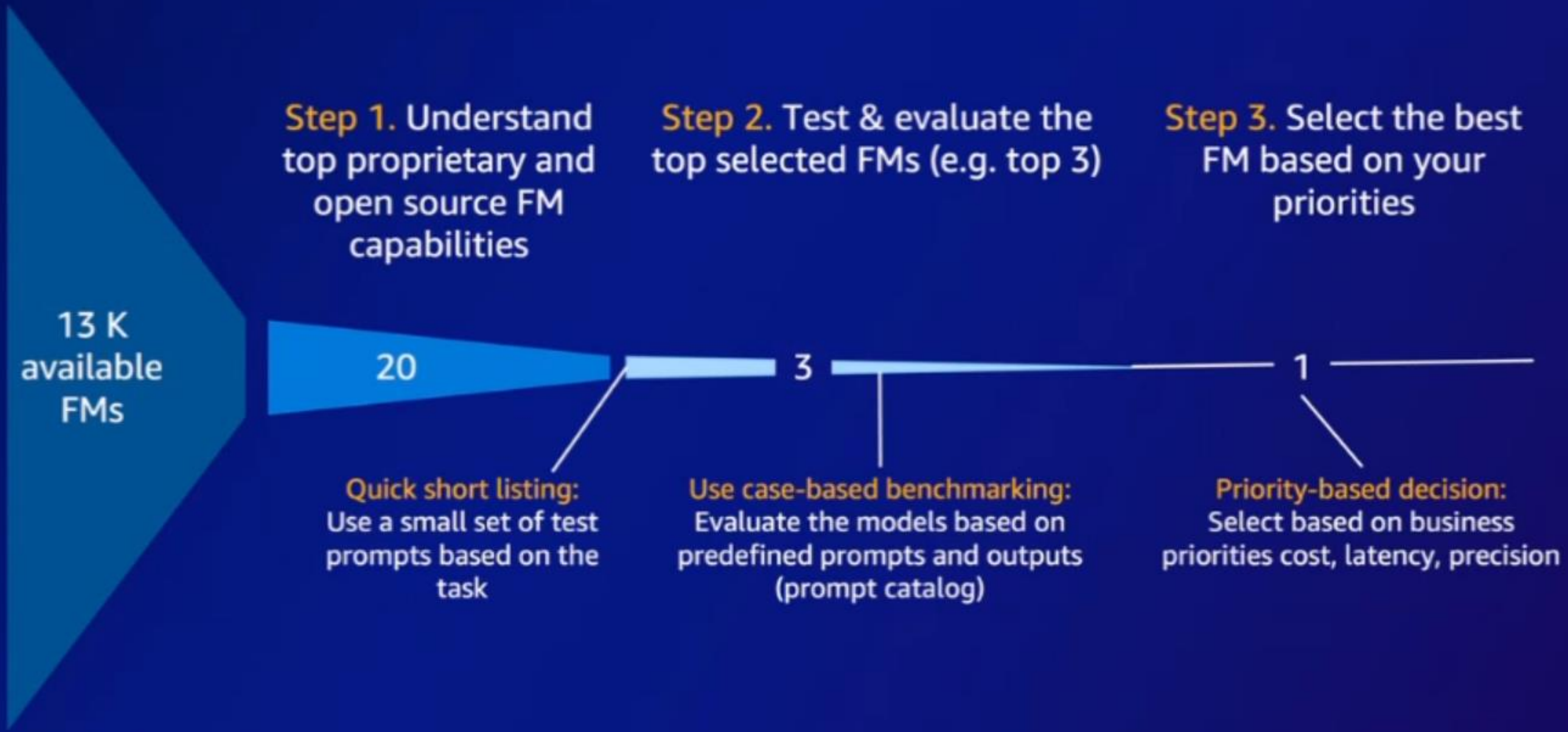
Create Private User Accounts

Create user accounts 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

Select FM - Consumers



Step 1. Understand top FM capabilities

Main FM Capability Matrix

Proprietary or open-source FM

Commercial License | Fine-tunable

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

Trained on (instructions, code, internet data) | Quality

Existing Customer Skills

Step 1. Proprietary FM Capabilities

Company Name	Model Name	Can be used Commercially	# Params	GPU instance req.	Available on AWS	Speed	Context Window	Trained on	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. Open-source FM Capabilities

Company Name	Model Name	Can be used Commercially	# Params	GPU instance req.	Available on AWS	Speed	Context Window	Trained on	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. EU AI Act Matters for FM Selection

Grading Foundation Model Providers' Compliance with the Draft EU AI Act

Source: Stanford Research on Foundation Models (CRFM), Institute for Human-Centered Artificial Intelligence (HAI)

Draft AI Act Requirements	GPT-4	Cohere Command	Stable Diffusion v2	Claude	PaLM 2	BLOOM	LLaMA	Jurassic-2	Luminous	GPT-NeoX	Totals
Data sources	●○○○	●●●○	●●●●	○○○○	●●○○	●●●●	●●●●	○○○○	○○○○	●●●●	22
Data governance	●●○○	●●●○	●●○○	○○○○	●●●○	●●●●	●●○○	○○○○	○○○○	●●●○	19
Copyrighted data	○○○○	○○○○	○○○○	○○○○	○○○○	●●●○	○○○○	○○○○	○○○○	●●●●	7
Compute	○○○○	○○○○	●●●●	○○○○	○○○○	●●●●	●●●●	○○○○	●○○○	●●●●	17
Energy	○○○○	●○○○	●●●○	○○○○	○○○○	●●●●	●●●●	○○○○	○○○○	●●●●	16
Capabilities & limitations	●●●●	●●●○	●●●●	●○○○	●●●●	●●●○	●●○○	●●○○	●○○○	●●●○	27
Risks & mitigations	●●●○	●●○○	●○○○	●○○○	●●●○	●●○○	●○○○	●●○○	○○○○	●○○○	16
Evaluations	●●●●	●●○○	○○○○	○○○○	●●○○	●●●○	●●○○	○○○○	●○○○	●○○○	15
Testing	●●●○	●●○○	○○○○	○○○○	●●○○	●●○○	○○○○	●○○○	○○○○	○○○○	10
Machine-generated content	●●●○	●●●○	○○○○	●●●○	●●●○	●●●○	○○○○	●●●○	●○○○	●●○○	21
Member states	●●○○	○○○○	○○○○	●●○○	●●●●	○○○○	○○○○	○○○○	●○○○	○○○○	9
Downstream documentation	●●●○	●●●●	●●●●	○○○○	●●●●	●●●●	●●○○	○○○○	○○○○	●●●○	24
Totals	25 / 48	23 / 48	22 / 48	7 / 48	27 / 48	36 / 48	21 / 48	8 / 48	5 / 48	29 / 48	

https://crfm.stanford.edu/2023/06/15/eu-ai-act.html

Step 1. Understand FM Capabilities

Can be the same person

Prompt Engineers

Design Initial prompts

GenAI Developers

Short list the top models based on the initial prompts

Business new GenAI use case

Design initial prompt catalog

GenAI use case example :

Financial documents summarization

Prompt examples:

"Based on … Summarize the text"

"After reviewing X provide the summary"

"Give me the summary"

Selected top 3 FM example:

Titan Text Large

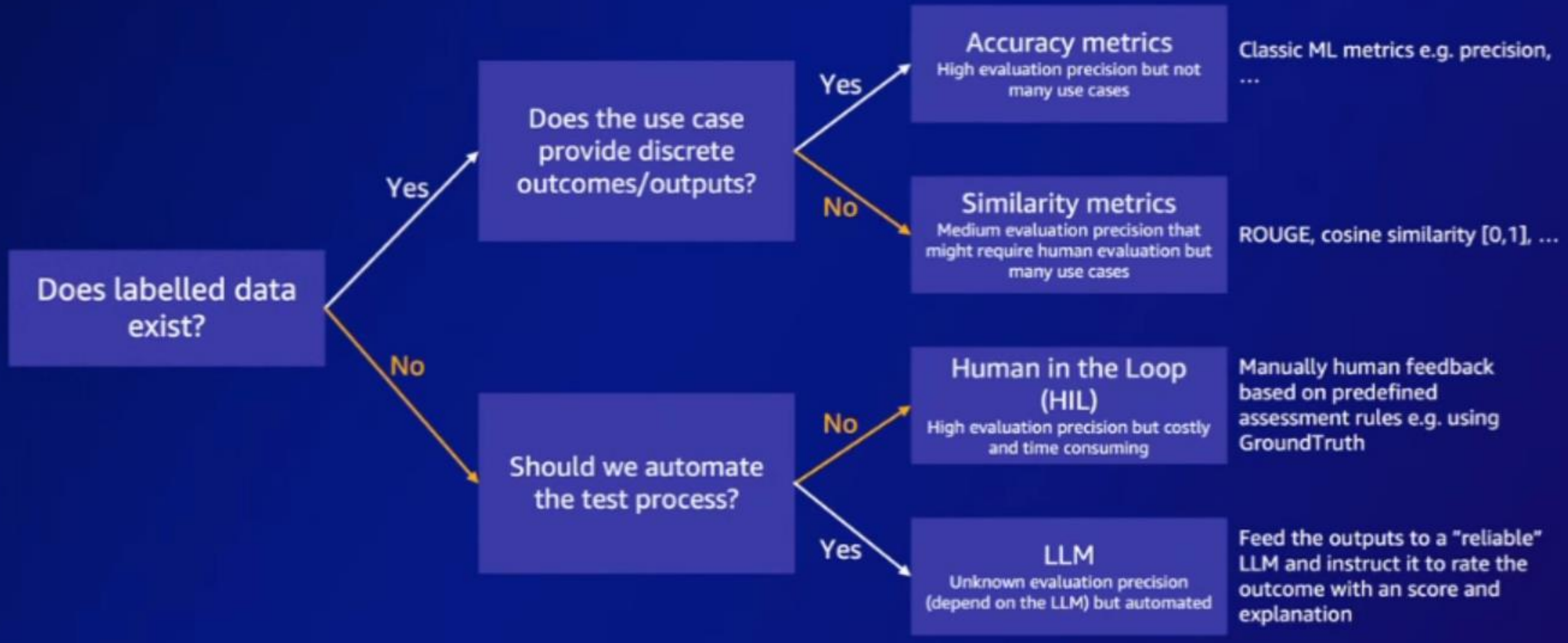
Claude

Falcon-7B-Instruct

Note:

Prompt = Input (data, data source, & instructions) + 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 ..."*	*<Summ ary>*	*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

GenAI 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

Lower speed, larger model, higher precision, larger cost

P1: Precision — Precision

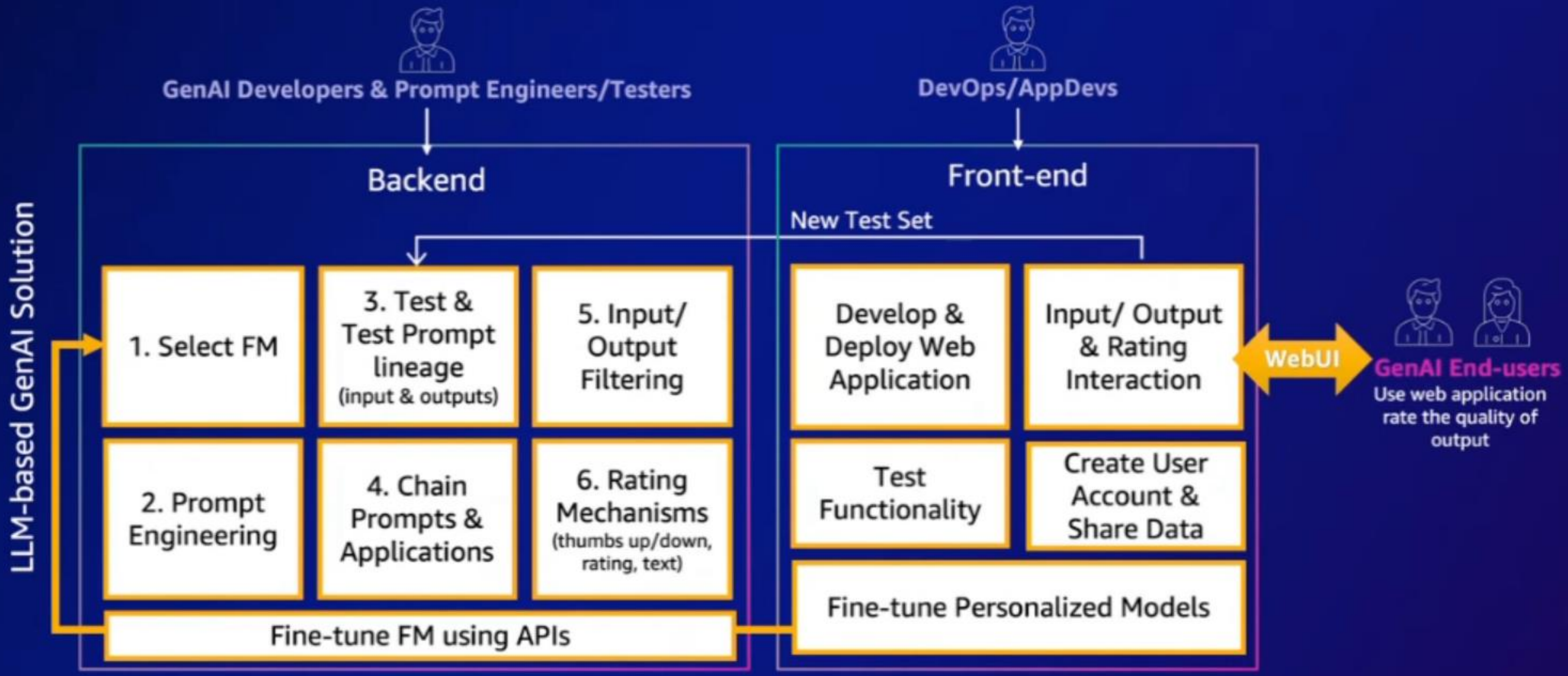
Model	Evaluation Score	HIL/LLM Feedback
FM1	5/5	<Feedback summary>
FM2	4/5	<Feedback summary>
FM3	3/5	<Feedback summary>

Cost — P0: lower cost

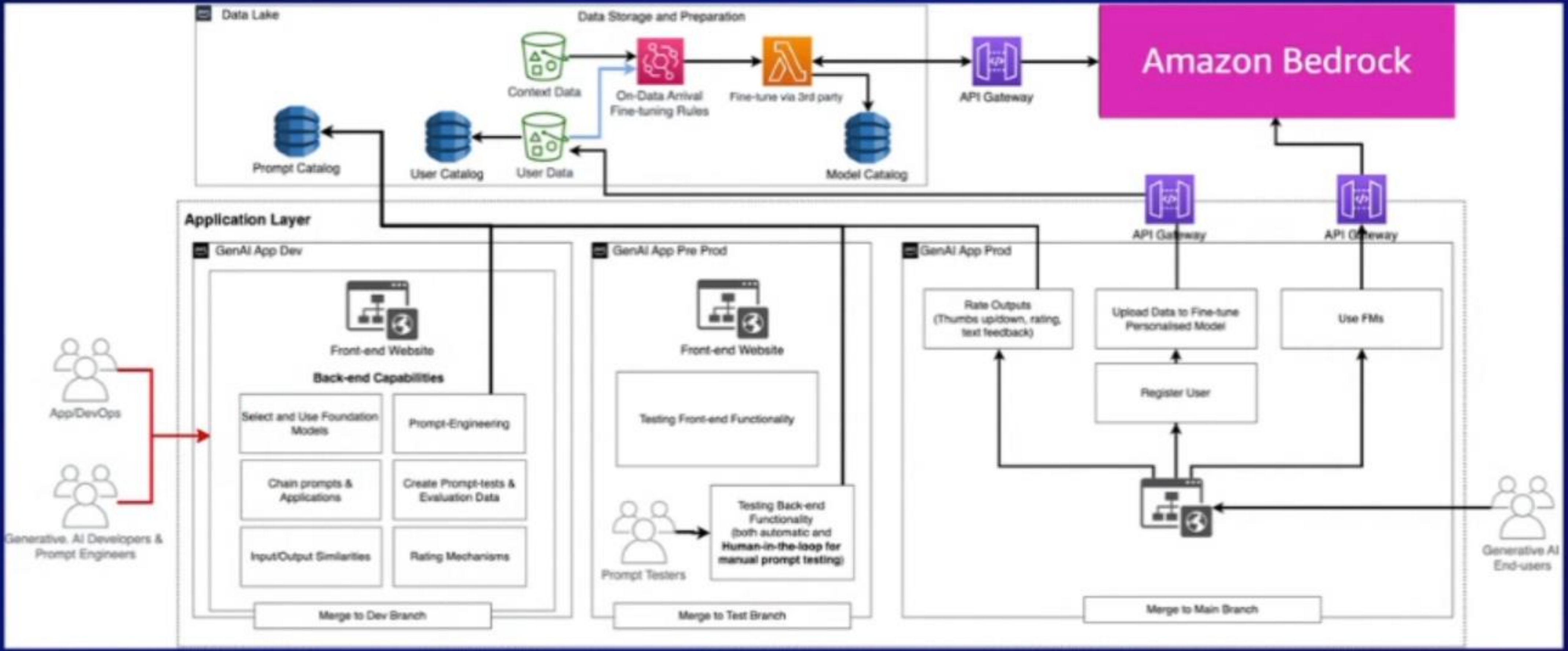
Model	Cost
FM1	$$$$
FM2	$
FM3	$$$

Model Selection: FM2

GenAI Processes for LLM - Consumers

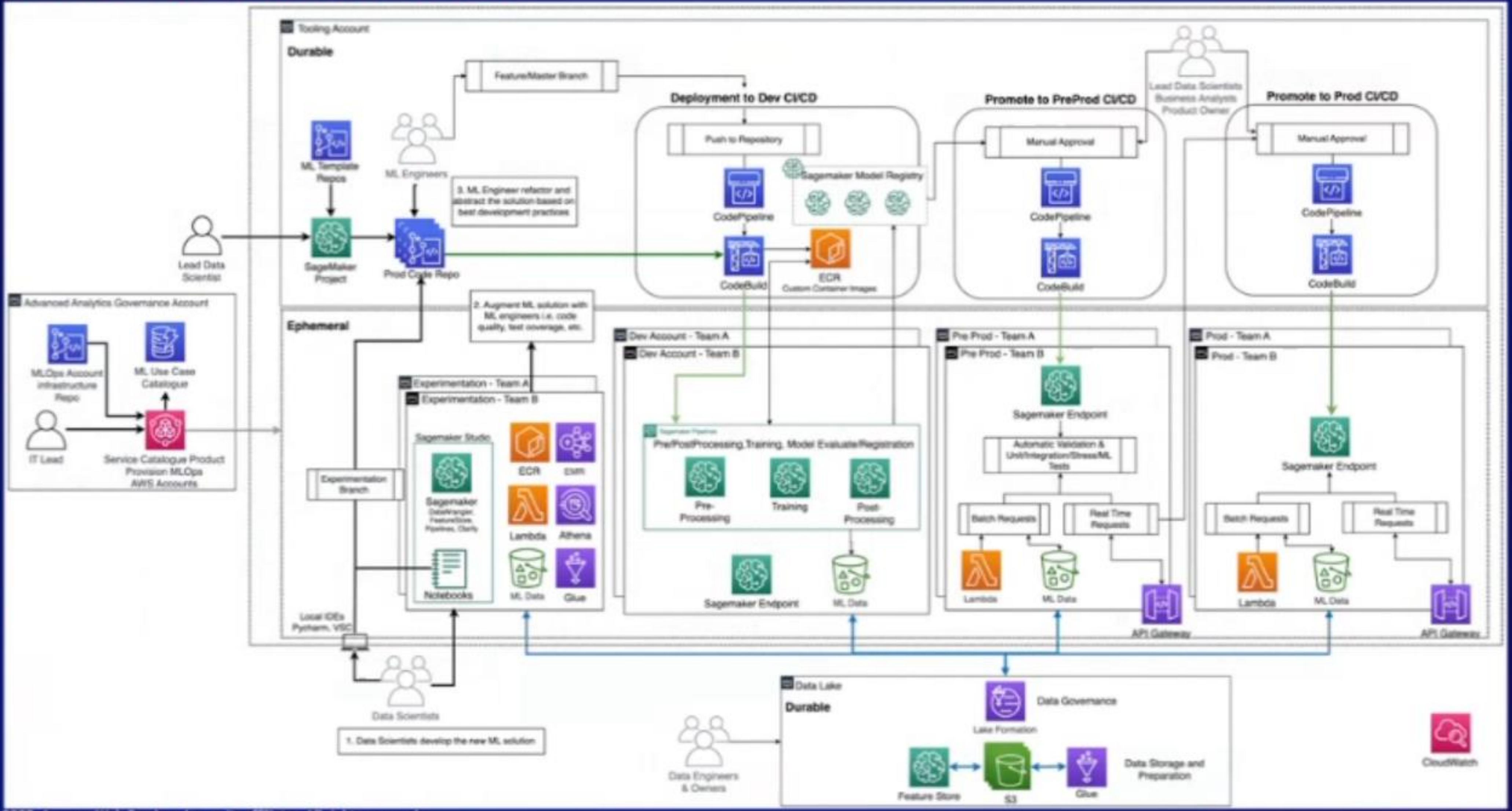


GenAI Technology for LLM – Consumers



GenAI Providers Productionize FM using MLOps

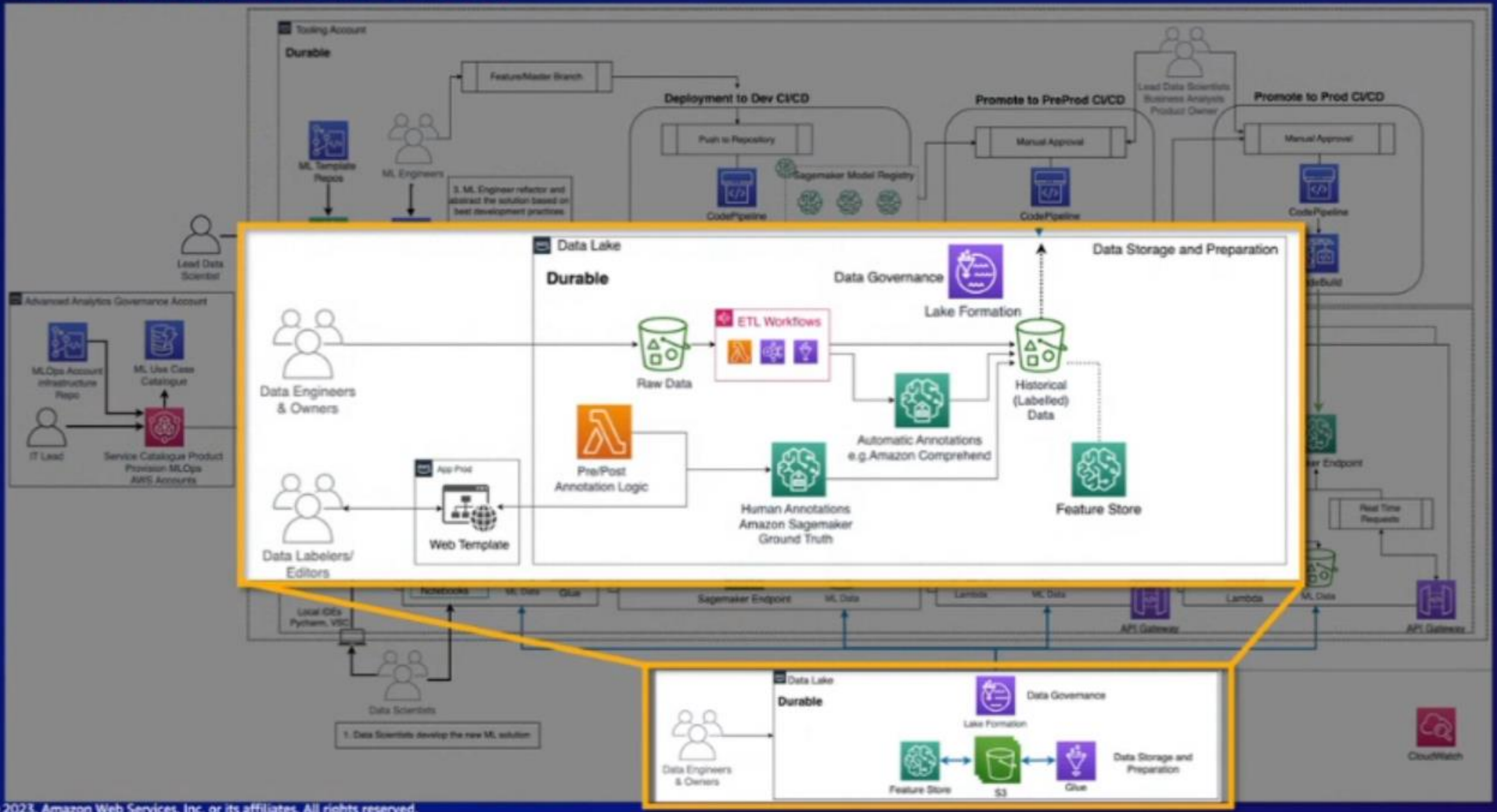
TRAIN MULTIPLE FOUNDATIONS MODELS



aws

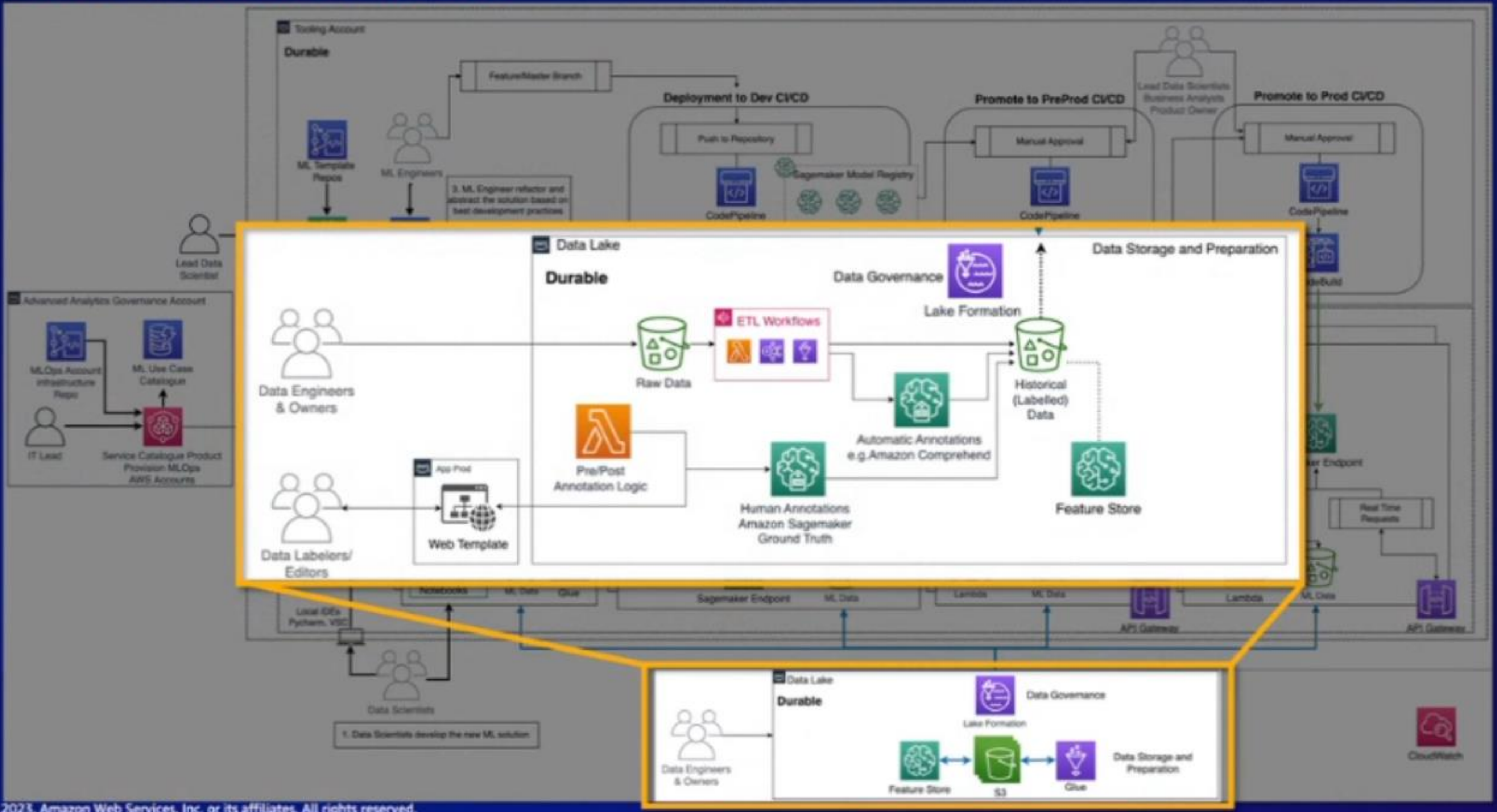
GenAI Providers Productionize FM using MLOps

TRAIN MULTIPLE FOUNDATIONS MODELS



GenAI Providers Productionize FM using MLOps

TRAIN MULTIPLE FOUNDATIONS MODELS



GenAI Processes - Fine-Tuners

Data Labeling
Human in the loop to label trillions, thousands, or hundreds of data

→

Fine-tune
Customization for specific domains

→

Deployment & Prompt Engineering
Trade-off between cost, precision, & latency

Chaim of models

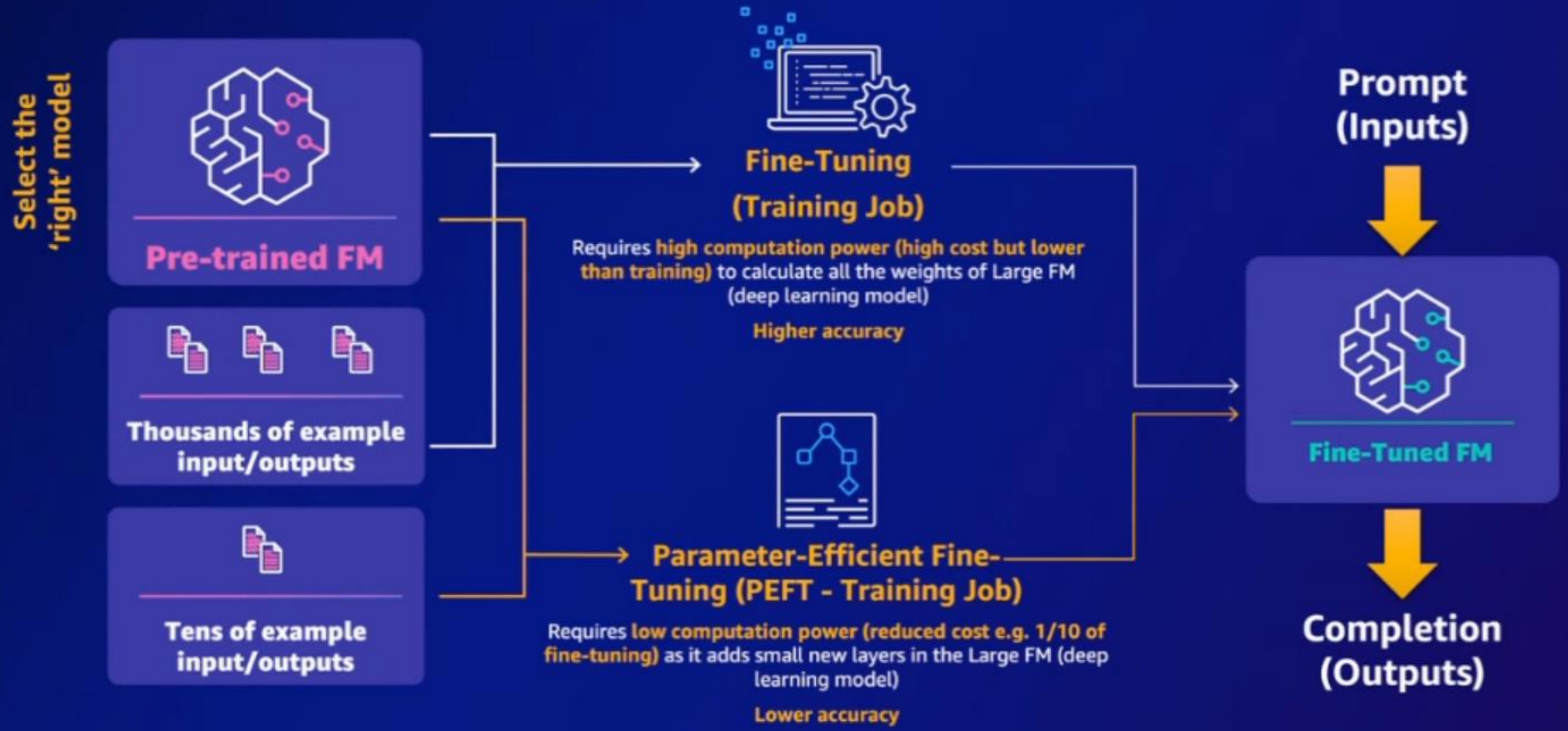
Prompt designing and engineering

Filtering input prompts and results using embeddings

→

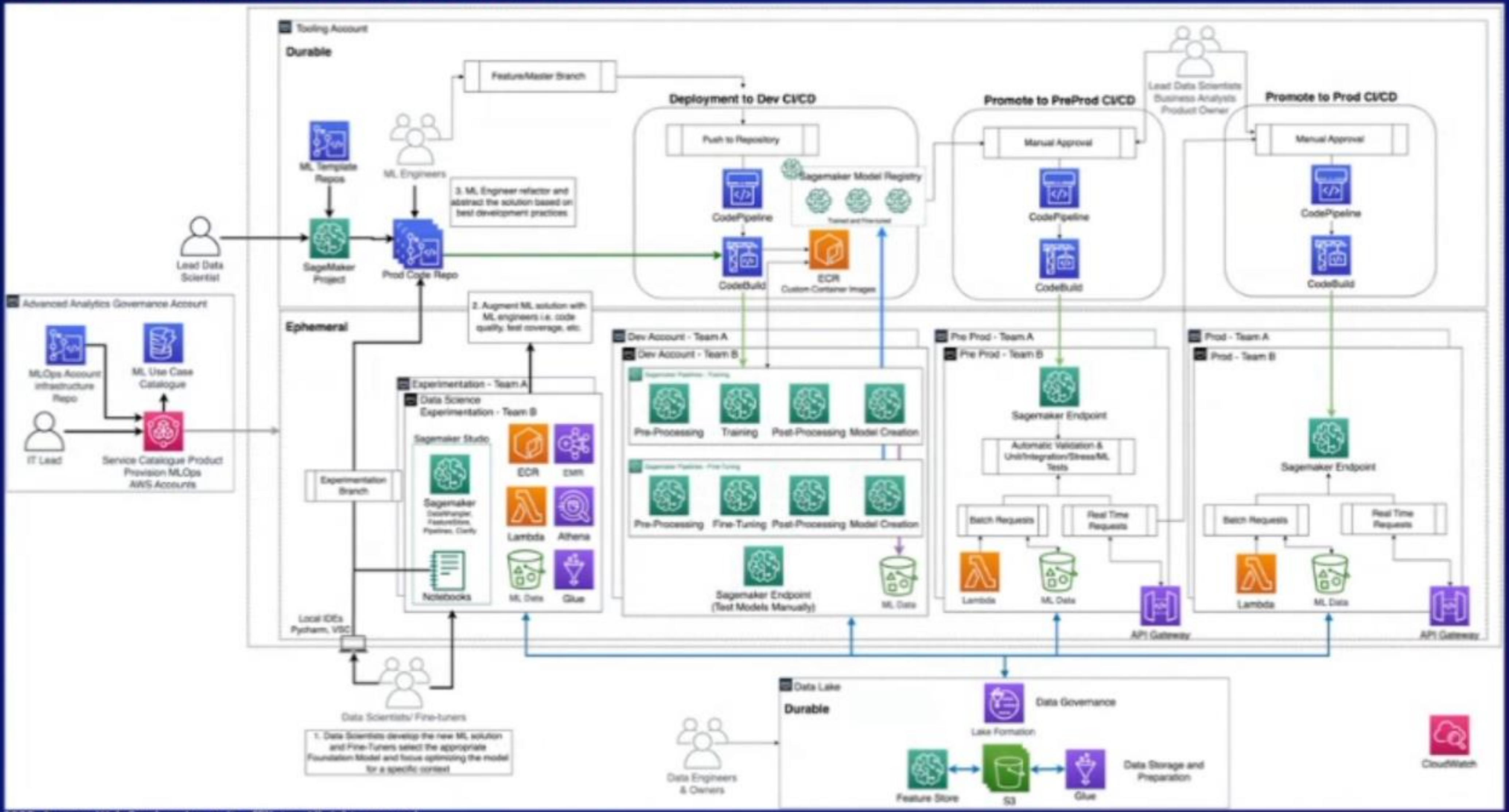
Monitor
Human in the loop feedback/rating, result similarities, toxicity rate, new methods under research...

Fine-Tuning, PEFT & Training



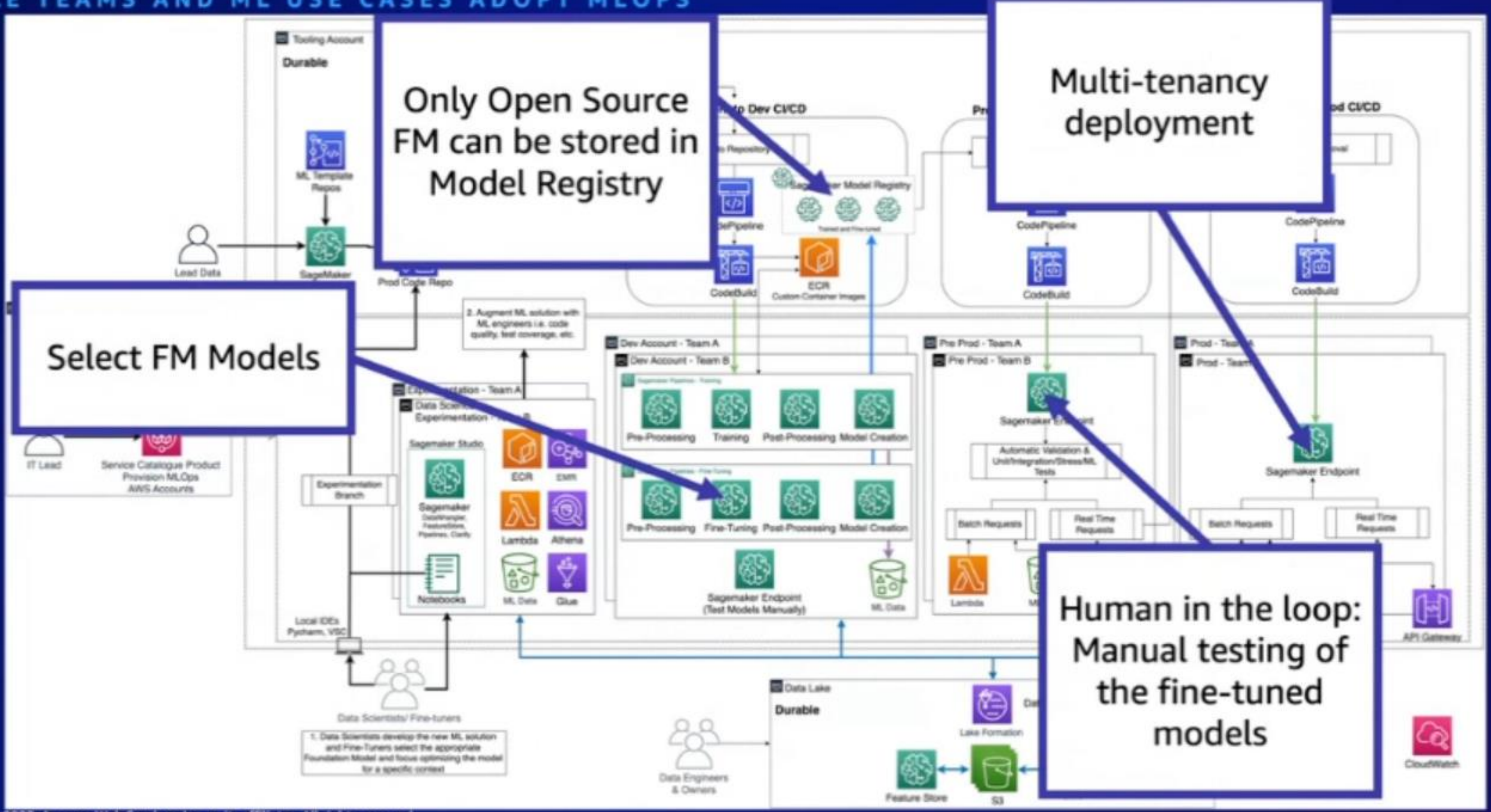
MLOPs & GenAI Technology - Fine-tuner

MULTIPLE TEAMS AND ML USE CASES ADOPT MLOPS



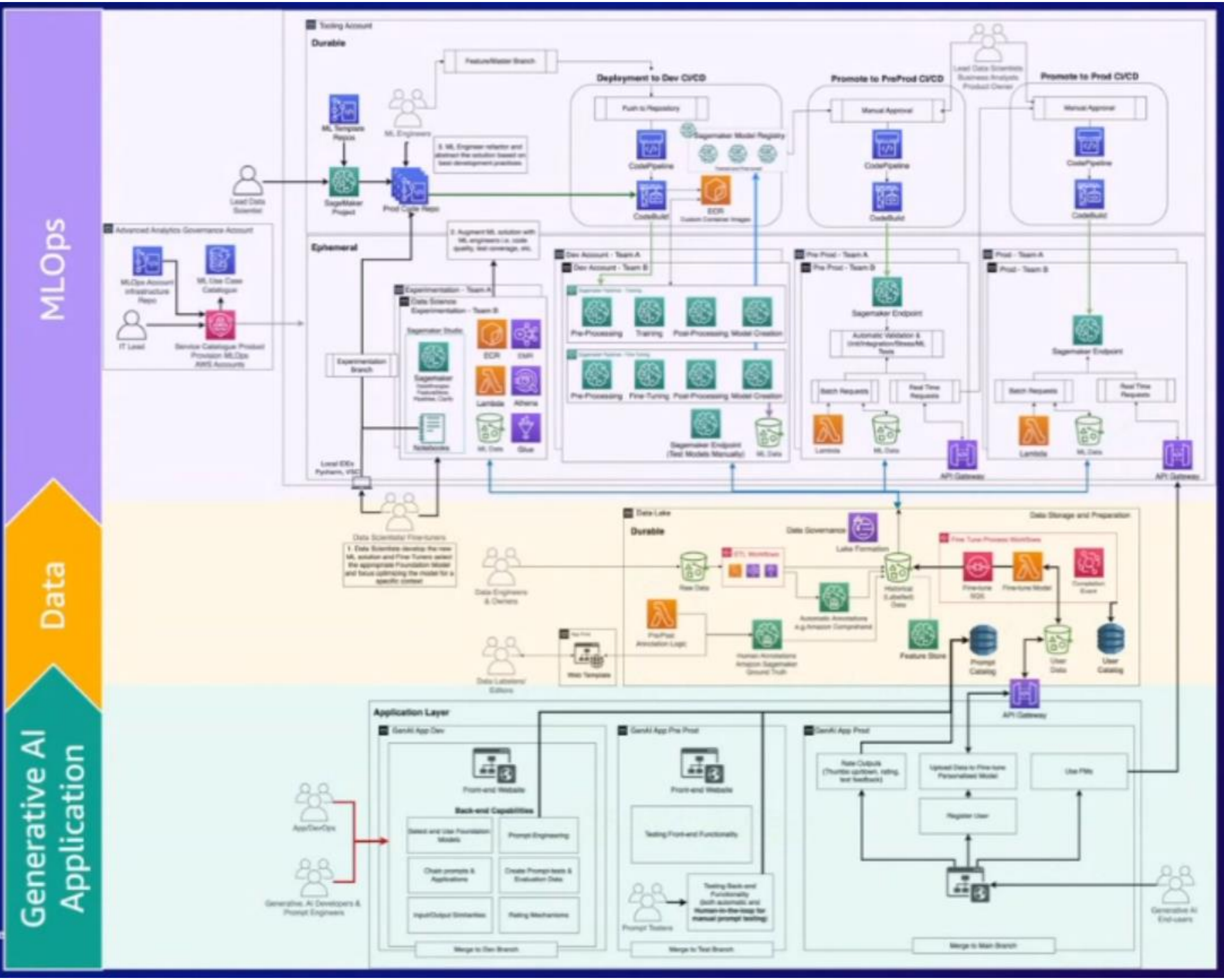
MLOPs & GenAI Technology - Fine-tuner

MULTIPLE TEAMS AND ML USE CASES ADOPT MLOPS



MLOPs & Generative AI Technology – Fine-tuner

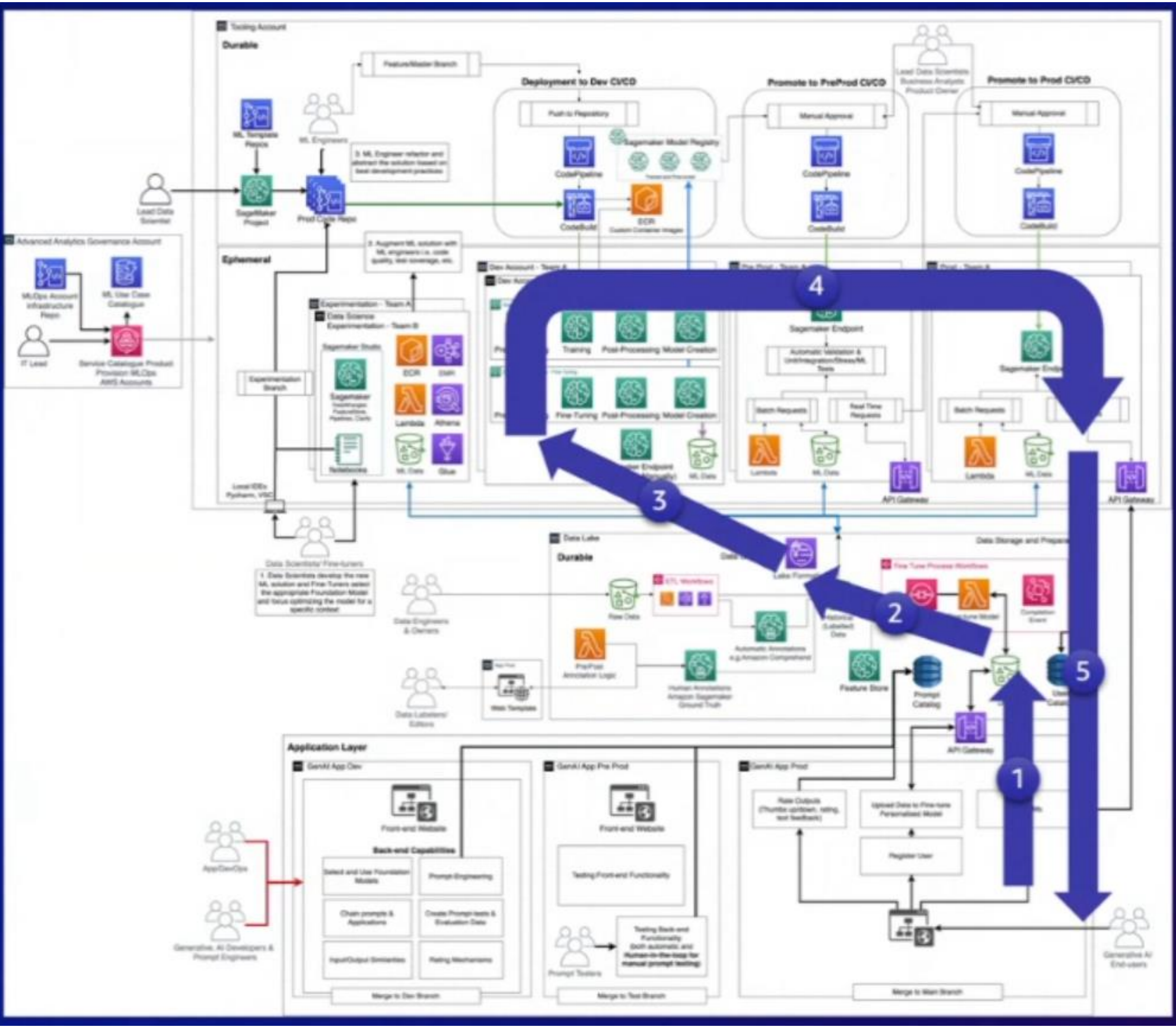
THREE MAIN LAYERS ARE INTERCONNECTED



aws

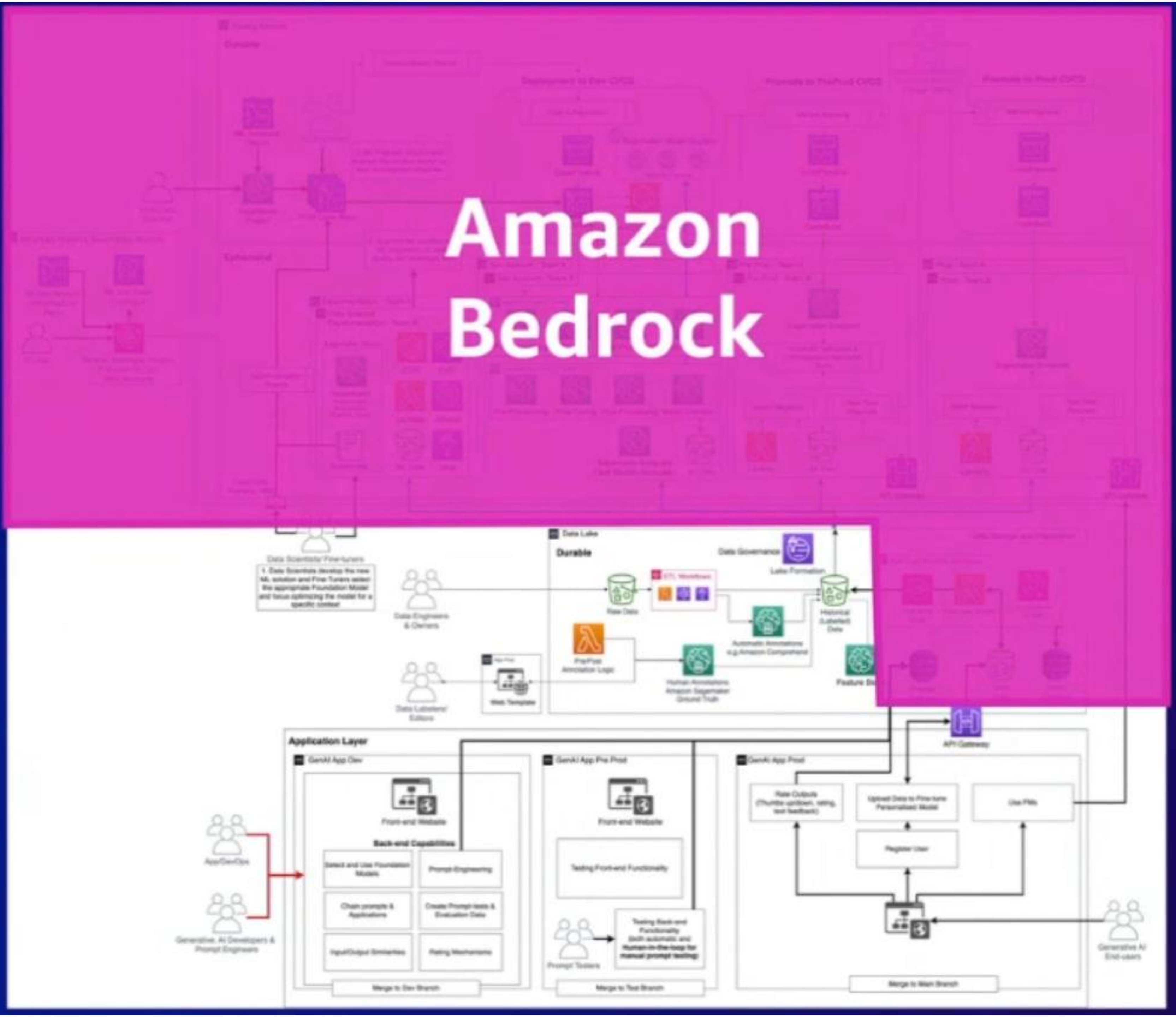
MLOPs & Generative AI Technology – Fine-tuner

THREE MAIN LAYERS ARE INTERCONNECTED



MLOPs & Generative AI Technology – Fine-tuner

THREE MAIN LAYERS ARE INTERCONNECTED



MLOps & FMOps Key Personas and Roles

Advance Analytics Team
Data Lake

Data Engineer
Prepare & Ingest data building ETL pipelines

Data Owners
Manage data sharing and provide access

Data Science Team
Experimentation & MLOps

Data Scientist
Create the best ML models to solve business problems

ML Engineer
Collaborate with DS to productionize ML

Platform Team
Secure Cloud/Data/ML Platform

MLOps Engineer/Admin
Standardize CI/CD, user/service role, model consumption, testing and deployment methodology

Security & Architects
Assess data, user, and service access creating policies and infrastructure

Business
Viz Dashboards, ML Adoption, & ROI

Business Stakeholder Product Owners
Define business problem, business KPIs, and make business decisions

Business Stakeholder Data & ML Consumers
Consumers of ML results from other BUs, driving business decision making

Labeler Team
Data Preparation at Scale

Data Labelers/Editors
Label or edit billions of Data for FM models and hundreds of data for fine tuning interacting with data lake using a dedicated website

Data Science Team Extension
Context Adaptation

Fine Tuners
Select the corresponding FM, evaluate the model & design the deployment method/infrastructure

Application Developer Team
Integrate GenAI models in applications

Generative AI Developers, AppDev, & Prompt Engineers/Testers
Design prompt inputs, create examples of prompt input/outputs, and test the engineered prompts, develop the GenAI application and front-end

End-Users
Consume Generative AI applications

Generative AI End-users
Consume Generative AI solutions as black box, share data and rate the quality of output

Generative AI Personas

Labeler Team
Data Preparation at Scale

Data Labelers/Editors

Label trillions of Data for FM models and hundreds of data for fine tuning interacting with data lake using a dedicated website

Data Science Team Extension
Context Adaptation

Fine Tuners

Select the corresponding FM, evaluate the model & design the deployment method/infrastructure

Application Developer Team
Integrate GenAI models in applications

Generative AI Developers

Select, test, evaluate the FM, filter inputs/outputs, and develop the GenAI application back-end (e.g. LangChain Experts)

AppDev

Develop the front-end of the GenAI application

Prompt Engineers

Design the input/output prompts to adapt the solution to the context and test the initial version

Prompt Testers

Test at scale the Generative AI solution (back-end/ front-end) and feed their results to the prompt test repository

End-Users
Consume Generative AI applications

Generative AI End-users

Consume Generative AI solutions as black box, share data and rate the quality of output

MLOps & GenAI Foundation People & Processes

SEPARATION OF CONCERNS IS KEY FOR SUCCESS