FMOps/LLMLOps: Operationalise Generative AI using MLOps principles

Dr Sokratis Kartakis (he/him)  
Senior MLOps SA Architect, EMEA  
Amazon Web Services

Heiko Hotz (he/him)  
Senior LLM SA Architect, EMEA  
Amazon Web Services
Agenda

MLOps Foundation Overview
• MLOps KPIs, Maturity, People, Processes, Technology

Generative AI (GenAI) & MLOps
• Main Definitions

MLOps & FMOps/LLMOps Differentiators
• Processes & People
• Providers, fine-tuners, & consumers
• Select & Adapt the FM on a Specific Context
• Evaluate & Monitor Fine-tuned Models
• Data & Deployment
• Technology
What is MLOps?

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

MLOps

Machine Learning & Operations

MLOps Definition

People

Processes

Technology
# MLOps Foundation Expected Outcomes

**Standardize operations and infrastructure for your data science**

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

**Reduce platform, people and operation costs**

Customer references building MLOps foundation and business benefits:

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.
MLOps Maturity Model

- Initial: Establish the Experimentation Environment
- Repeatable: Standardize Code Repositories & ML Solution Deployment
- Reliable: Introduce Testing, Monitoring, and Multi-account Deployment
- Scalable: Templatize and Productionize Multiple ML Solutions by Multiple Teams

Models in Production

MLOps Maturity

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.
MLOps Key Personas and Roles

**Advance Analytics Team**
- **Data Engineer**
  - Prepare & Ingest data building ETL pipelines
- **Data Owners**
  - Manage data sharing and provide access
- **Data Scientist**
  - Create the best ML models to solve business problems
- **ML Engineer**
  - Collaborate with DS to productionize ML

**Data Science Team**
- **Experimentation & MLOps**
- **Data Lake**

**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

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.
MLOps Foundation People & Processes

Separation of Concerns Is Key for Success

Platform Administration

- MLOps Engineers
- System Administrators
- Security

Experimentation

Prove that ML can solve a business problem, i.e. PoC

Data

- Ingest data
- Prepare, combine and catalogue Data
- Visualize data

Data Governance

Centralized model and code artifact storage/versioning/auditing

Model Build

Automate model build/training providing scaled data

Model Test

Automate model testing and guardrails

Model Deployment

Serving and monitoring the model testing

Data Governance

Product Owner
- Lead Data Scientists
- Model Approvers
- Auditors & Compliance

Model and Code

- ML Governance
- Centralized model and code artifact storage/versioning/auditing

Provision infrastructure
- Provide user access
- Provide data access

Ingest data
- Prepare, combine and catalogue Data
- Visualize data

Prove that ML can solve a business problem, i.e. PoC

Automate model build/training providing scaled data

Automate model testing and guardrails

Serving and monitoring the model testing

Data
- Data Engineers
- Data Owners
- Business Stakeholders
- ML Consumers
MLOPs Scalable Phase

MULTIPLE TEAMS AND ML USE CASES ADOPT MLOPS
Generative AI (GenAI) & MLOps
MLOps & FMOps/LLMLOps Differentiators
GenAI Use Case Domains

Terabytes of data

Foundation Models
Billions of Parameters

Train

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

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

FMOps

Large Language Model Operations
Productionize Large Language Model-based solutions

LLMOps

Technology

Processes

People
MLOps & FMOps Differentiators

MLOps

Technology

People

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

FMOps
MLOps & FMOps Differentiators

MLOps
- Technology
- Processes
- People

FMOps
- 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

**Generative AI User Types**

- **Providers**: Entities who build foundation models from scratch themselves and provide them as a product to **tuner** and **consumer**.
- **Fine-Tuners**: Fine-tune foundational models from **providers** to fit custom requirements. Orchestrate the deployment of the model as a service for use by **consumers**.
- **Consumers**: Interact with Generative AI services from **provider** or **tuner** by text prompting or visual interface to complete desired actions.

**Skills**

- **MLOps is required**
  - Providers: Deep end-to-end ML, NLP expertise and data science, labeler “squad”
  - Fine-Tuners: Strong end-to-end ML expertise and knowledge of model deployment and inference. Strong domain knowledge for tuning including prompt engineering.
  - Consumers: No ML expertise required. Mostly application developers or end-users with understanding of the service capabilities. Only prompt engineering is required for better results.

**Productionize applications where DevOps/AppDev is more relevant than MLOps**
The Journey of Consumers
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
Select FM - Consumers

Step 1. Understand top proprietary and open source FM capabilities

Step 2. Test & evaluate the top selected FMs (e.g. top 3)

Step 3. Select the best FM based on your priorities

Quick short listing:
Use a small set of test prompts based on the task

Use case-based benchmarking:
Evaluate the models based on predefined prompts and outputs (prompt catalog)

Priority-based decision:
Select based on business priorities cost, latency, precision

13 K available FMs

20

3

1
## Step 1. Understand top FM capabilities

### Main FM Capability Matrix

<table>
<thead>
<tr>
<th>Proprietary or open-source FM</th>
<th>Commercial License</th>
<th>Fine-tunable</th>
<th>Speed</th>
<th>#Parameter</th>
<th>Context Window</th>
<th>Trained on</th>
<th>Quality</th>
<th>Existing Customer Skills</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(#tokens where a tokens is ~0.75x words)</td>
<td>(instructions, code, internal data)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Step 1. Proprietary FM Capabilities

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

*Last update 16 Jun 2023*
# Step 1. Open-source FM Capabilities

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

*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)

<table>
<thead>
<tr>
<th>Draft AI Act Requirements</th>
<th>GPT-4</th>
<th>Cohere</th>
<th>Stable Diffusion v2</th>
<th>Claude</th>
<th>PaLM 2</th>
<th>BLOOM</th>
<th>LLaMA</th>
<th>Jurassic-2</th>
<th>Luminous</th>
<th>GPT-NeoX</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data sources</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>22</td>
</tr>
<tr>
<td>Data governance</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>19</td>
</tr>
<tr>
<td>Copyrighted data</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>7</td>
</tr>
<tr>
<td>Compute</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>17</td>
</tr>
<tr>
<td>Energy</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>16</td>
</tr>
<tr>
<td>Capabilities &amp; limitations</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>27</td>
</tr>
<tr>
<td>Risks &amp; mitigations</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>16</td>
</tr>
<tr>
<td>Evaluations</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>15</td>
</tr>
<tr>
<td>Testing</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>10</td>
</tr>
<tr>
<td>Machine-generated content</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>21</td>
</tr>
<tr>
<td>Member states</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>9</td>
</tr>
<tr>
<td>Downstream documentation</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>24</td>
</tr>
</tbody>
</table>

[https://crfm.stanford.edu/2023/06/15/eu-ai-act.html](https://crfm.stanford.edu/2023/06/15/eu-ai-act.html)
Step 1. Understand FM Capabilities

Business new GenAI use case

GenAI use case example: Financial documents summarization

Prompt Engineers
Design Initial prompts

Design initial prompts

Prompt examples:
“Based on ... Summarize the text”
“After reviewing X provide the summary”
“Give me the summary”

Can be the same person

GenAI Developers
Short list the top models based on the initial prompts

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

Does labelled data exist?

- Yes
  - Does the use case provide discrete outcomes/outputs?
    - Yes
      - Accuracy metrics
        - High evaluation precision but not many use cases
        - Classic ML metrics e.g. precision, …
    - No
      - Similarity metrics
        - Medium evaluation precision that might require human evaluation but many use cases
        - ROUGE, cosine similarity [0,1], …
  - No
    - Human in the Loop (HIL)
      - High evaluation precision but costly and time consuming
      - Manually human feedback based on predefined assessment rules e.g. using GroundTruth

- No
  - Should we automate the test process?
    - Yes
      - LLM
        - Unknown evaluation precision (depend on the LLM) but automated
        - Feed the outputs to a "reliable" LLM and instruct it to rate the outcome with an score and explanation
    - No
      - Manually human feedback based on predefined assessment rules e.g. using GroundTruth
Step 2. Evaluate the top FMs - Examples

<table>
<thead>
<tr>
<th>Prompt (Input + Query)</th>
<th>Output Pre-created Output e.g. Labeled</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Give me the full name of the UK PM”</td>
<td>“Rishi Sunak”</td>
</tr>
<tr>
<td>&lt;Text to summarize&gt; + “Give me the summary”</td>
<td>“The summary is …”</td>
</tr>
<tr>
<td>“Generate a story based on the SnowWhite book”</td>
<td>n/a</td>
</tr>
<tr>
<td>“Generate the source code for a retail website”</td>
<td>n/a</td>
</tr>
</tbody>
</table>

*Public or private data can be used

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

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

Generate aggregated results for the top FMs
Step 3. Select the best FM based on priorities

### Table: Model Evaluation Score

<table>
<thead>
<tr>
<th>Model</th>
<th>Evaluation Score</th>
<th>HIL/LLM Feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>FM1</td>
<td>5/5</td>
<td>&lt;Feedback summary&gt;</td>
</tr>
<tr>
<td>FM2</td>
<td>4/5</td>
<td>&lt;Feedback summary&gt;</td>
</tr>
<tr>
<td>FM3</td>
<td>3/5</td>
<td>&lt;Feedback summary&gt;</td>
</tr>
</tbody>
</table>

### Table: Model Cost

<table>
<thead>
<tr>
<th>Model</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>FM1</td>
<td>$$$$</td>
</tr>
<tr>
<td>FM2</td>
<td>$</td>
</tr>
<tr>
<td>FM3</td>
<td>$$</td>
</tr>
</tbody>
</table>

### Diagram:
- **Precision**: FM2 (High speed, smaller model, lower precision, smaller cost)
- **Cost**: P0: lower cost
- **Model Selection**: FM2

EXAMPLE
GenAI Processes for LLM - Consumers

1. Select FM
2. Prompt Engineering
3. Test & Test Prompt lineage (input & outputs)
4. Chain Prompts & Applications
5. Input/Output Filtering
6. Rating Mechanisms (thumbs up/down, rating, text)

LLM-based GenAI Solution

GenAI Developers & Prompt Engineers/Testers

Backend

New Test Set

Front-end

Develop & Deploy Web Application
Input/Output & Rating Interaction
Test Functionality

DevOps/AppDevs

WebUI

GenAI End-users
Use web application rate the quality of output

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.
GenAI Technology for LLM - Consumers
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.
GenAI Processes for LLM - Consumers

GenAI Developers & Prompt Engineers/Testers

1. Select FM or Use Fine-tuned Mode
2. Prompt Engineering
3. Test & Test Prompt lineage (input & outputs)
4. Chain Prompts & Applications
5. Input/Output Filtering
6. Rating Mechanisms (thumbs up/down, rating, text)

Fine-tune FM using APIs

DevOps/AppDevs

New Test Set

Backend

Develop & Deploy Web Application
Input/Output & Rating Interaction
Test Functionality
Create User Account & Share Data
Fine-tune Personalized Models

Front-end

WebUI

GenAI End-users Use web application rate the quality of output

LLM-based GenAI Solution

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.
GenAI Technology for LLM - Consumers
GenAI Technology for LLM - Consumers
The Journey of Providers
GenAI Providers Productionize FM using MLOps
The Journey of Fine-tuners
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

Select the ‘right’ model

- Pre-trained FM
- Thousands of example input/outputs
- Tens of example input/outputs

Fine-Tuning (Training Job)

- Requires high computation power (high cost but lower than training) to calculate all the weights of Large FM (deep learning model)
- Higher accuracy

Parameter-Efficient Fine-Tuning (PEFT - Training Job)

- Requires low computation power (reduced cost e.g. 1/10 of fine-tuning) as it adds small new layers in the Large FM (deep learning model)
- Lower accuracy

Prompt (Inputs)

Completion (Outputs)
MULTIPLE TEAMS AND ML USE CASES ADOPT MLOPS
Multiple teams and ML use cases adopt MLOPs

**Select FM Models**

Only Open Source FM can be stored in Model Registry

Multi-tenancy deployment

Human in the loop: Manual testing of the fine-tuned models
Data & Open Source Fine-tuned FM Deployment

AWS Account – Fine-tuner

Model Registry

Main Model Group

V1 V2 V3

Preloaded Open-Source FM

Notebook or Code Repo

Customer Data

Amazon SageMaker Endpoint

Internet

Internet connection is required to get locally the FMs

Open Source FM

FM1

Source Code
Data & Proprietary Fine-tuned FM Deployment

AWS Account – Fine-tuner
- Notebook or Code Repo
- Customer Data
- Amazon SageMaker Endpoint
- Model

AWS Account – Proprietary FM Provider 1
- Model Registry
- Main Model Group
  - V1
  - V2
  - V3
  - Producer Production Ready FM

AWS Account – Proprietary FM Provider 2

AWS Account – Proprietary FM Provider N

Use model package

Internet
MLOPs & Generative AI Technology – Fine-tuner

Three main layers are interconnected:

- MLOps
- Data
- Generative AI Application
THREE MAIN LAYERS ARE INTERCONNECTED

Amazon Bedrock
People & Processes
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 Team**
- **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
MLOps & FMOps Key Personas and Roles

**Advance Analytics Team**
- **Data Engineer**
  - Prepare & Ingest data building ETL pipelines
- **Data Owners**
  - Manage data sharing and provide access

**Data Science Team**
- **Data Scientist**
  - Create the best ML models to solve business problems
- **ML Engineer**
  - Collaborate with DS to productionize ML

**Platform Team**
- **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**
- **Business Stakeholder**
  - **Product Owners**
    - Define business problem, business KPIs, and make business decisions
- **Data & ML Consumers**
  - Consumers of ML results from other BUs, driving business decision making

**Risk & Compliance**
- **Approve & Review Models**
  - Review models, data sources, code artifacts

**Business Viz Dashboards, ML Adoption, & ROI**
- **Business Stakeholder**
  - Define business problem, business KPIs, and make business decisions

---

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.
### MLOps & FMOps Key Personas and Roles

<table>
<thead>
<tr>
<th>Role</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Engineer</td>
<td>Prepare &amp; Ingest data building ETL pipelines</td>
</tr>
<tr>
<td>Data Scientist</td>
<td>Create the best ML models to solve business problems</td>
</tr>
<tr>
<td>ML Engineer</td>
<td>Collaborate with DS to productionize ML</td>
</tr>
<tr>
<td>Security &amp; Architects</td>
<td>Assess data, user, and service access creating policies and infrastructure</td>
</tr>
<tr>
<td>MLOps Engineer/Admin</td>
<td>Standardize CI/CD, user/service role, model consumption, testing and deployment methodology</td>
</tr>
<tr>
<td>Platform Team</td>
<td>Secure Cloud/Data/ML Platform</td>
</tr>
<tr>
<td>Data Owner</td>
<td>Manage data sharing and provide access</td>
</tr>
<tr>
<td>Labeler Team</td>
<td>Data Preparation at Scale</td>
</tr>
<tr>
<td>Data Lake</td>
<td>Data Preparation at Scale</td>
</tr>
<tr>
<td>Labelers/Editors</td>
<td>Label or edit billions of data for FM models and hundreds of data for fine tuning interacting with data lake using a dedicated website</td>
</tr>
<tr>
<td>Fine Tuners</td>
<td>Select the corresponding FM, evaluate the model &amp; design the deployment method/infrastructure</td>
</tr>
<tr>
<td>Application Developer Team</td>
<td>Integrate GenAI models in applications</td>
</tr>
<tr>
<td>Business</td>
<td>Viz Dashboards, ML Adoption, &amp; ROI</td>
</tr>
<tr>
<td>Business Stakeholder</td>
<td>Define business problem, business KPIs, and make business decisions</td>
</tr>
<tr>
<td>Product Owners</td>
<td></td>
</tr>
<tr>
<td>Data Consumers</td>
<td>Consumers of ML results from other BUs, driving business decision making</td>
</tr>
<tr>
<td>Data &amp; ML Consumers</td>
<td></td>
</tr>
<tr>
<td>End-Users</td>
<td>Consume Generative AI applications</td>
</tr>
<tr>
<td>Generative AI End-users</td>
<td>Consume Generative AI solutions as black box, share data and rate the quality of output</td>
</tr>
</tbody>
</table>
Generative AI Personas

- **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

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

- **Data Science Team Extension**
  - Context Adaptation

- **Application Developer Team**
  - Integrate GenAI models in applications

- **Generative AI Developers, AppDev, & Prompt Engineers/Tests**
  - 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 solutions as black box, share data and rate the quality of output

- **Generative AI End-users**
  - Consume Generative AI applications
Generative AI Personas

**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.

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

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

**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.

**Application Developer Team**
- Integrate GenAI models in applications.

**End-Users**
- Consume Generative AI applications as black box, share data and rate the quality of output.

**Data Science Team Extension**
- Context Adaptation.

**AppDev**
- Develop the front-end of the GenAI application.

**Labeler Team**
- Data Preparation at Scale.

**Data Preparation at Scale**
- Labeler Team
- Data Science Team Extension
- Application Developer Team
- End-Users

**Context Adaptation**
- AppDev

**Integrate GenAI models in applications**
- Prompt Engineers
- Prompt Testers

**Consume Generative AI applications**
- Fine Tuners
- Generative AI Developers

**Consume Generative AI solutions as black box, share data and rate the quality of output**
- Data Labelers/Editors
- Data Science Team Extension

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

**Design the input/output prompts to adapt the solution to the context and test the initial version**
- Prompt Engineers

**Test at scale the Generative AI solution (back-end/front-end) and feed their results to the prompt test repository**
- Prompt Testers
**MLOps Foundation People & Processes**

Separation of Concerns Is Key for Success

---

**Platform Administration**
- MLOps Engineers
- System Administrators
- Security

**Experimentation**
- Prove that ML can solve a business problem, i.e. PoC
- Data Scientists

**Model Build**
- Automate model build/training providing scaled data
- Data Scientists
- ML Engineers

**Model Test**
- Automate model testing and guardrails
- Data Scientists
- ML Engineers

**Model Deployment**
- Serving and monitoring the model testing
- Data Scientists
- ML Engineers

**ML Governance**
- Product Owner
- Lead Data Scientists
- Model Approvers
- Auditors & Compliance

**Data**
- Ingest data
- Prepare, combine and catalogue Data
- Visualize data
- Data Engineers
- Data Owners
- Business Stakeholders
- ML Consumers

---

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.
MLOps & GenAI Foundation People & Processes

Separation of Concerns Is Key for Success

Platform Administration
- MLOps Engineers
- System Administrators
- Security

Experimentation
- Data Scientists
- Fine-Tuners

Prove that ML can solve a business problem, i.e. PoC
- Ingest data
- Prepare, combine and catalogue Data
- Visualize data

Model Build
- Data Scientists
- Fine-Tuners
- ML Engineers

Automate model build/training providing scaled data
- Prove that ML can solve a business problem, i.e. PoC
- Automate model build/training providing scaled data

Model Test
- Data Scientists
- Fine-Tuners
- ML Engineers

Automate model testing and guardrails
- Automate model testing and guardrails

Model Deployment
- Product Owner
- Lead Data Scientists
- Model Approvers
- Auditors & Compliance

Centralized model and code artifact storage/versioning/auditing
- Centralized model and code artifact storage/versioning/auditing

Data
- Data Engineers
- Data Owners
- Business Stakeholders
- ML Consumers

Application
- Generative AI Developers
- Prompt Engineers/Testers
- AppDev

Web UI
- Data Labelers/Editors
- Generative AI End-users
Bonus: GenAI/LLM Vulnerabilities

https://owasp.org/www-project-top-10-for-large-language-model-applications/descriptions/
Prompt Injections: Bypassing filters or manipulating the LLM using carefully crafted prompts that make the model ignore previous instructions or perform unintended actions.

Data Leakage: Accidentally revealing sensitive information, proprietary algorithms, or other confidential details through the LLM’s responses.

Inadequate Sandboxing: Failing to properly isolate LLMs when they have access to external resources or sensitive systems, allowing for potential exploitation and unauthorized access.

Unauthorized Code Execution: Exploiting LLMs to execute malicious code, commands, or actions on the underlying system through natural language prompts.

SSRF Vulnerabilities: Exploiting LLMs to perform unintended requests or access restricted resources, such as internal services, APIs, or data stores.
GenAI/LLM Vulnerabilities 2/2

**Overreliance on LLM-generated Content:** Excessive dependence on LLM-generated content without human oversight can result in harmful consequences.

**Inadequate AI Alignment:** Failing to ensure that the LLM’s objectives and behavior align with the intended use case, leading to undesired consequences or vulnerabilities.

**Insufficient Access Controls:** Not properly implementing access controls or authentication, allowing unauthorized users to interact with the LLM and potentially exploit vulnerabilities.

**Improper Error Handling:** Exposing error messages or debugging information that could reveal sensitive information, system details, or potential attack vectors.

**Training Data Poisoning:** Maliciously manipulating training data or fine-tuning procedures to introduce vulnerabilities or backdoors into the LLM
Conclusion
MLOps & FMOps Differentiators

MLOps

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
A/B testing & human feedback

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

Technology
MLOps, data, & application layers

FMOps

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
A/B testing & human feedback

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

Technology
MLOps, data, & application layers
Thank you!

Dr Sokratis Kartakis (he/him)
Senior MLOps SA Architect, EMEA
Amazon Web Services

Heiko Hotz (he/him)
Senior LLM SA Architect, EMEA
Amazon Web Services

@sokratis.kartakis

Please complete the session survey