DATA+AI SUMMIT
BY databricks

FINE-TUNING LLMs with Precision and Speed

Manasa Parvathipuram & Sonali Guleria
June, 2024

LLM: A TRILLION SQUARE RUBIKS CUBE

43,000,000,000,000,000,000 possibilities!!

AGENDA

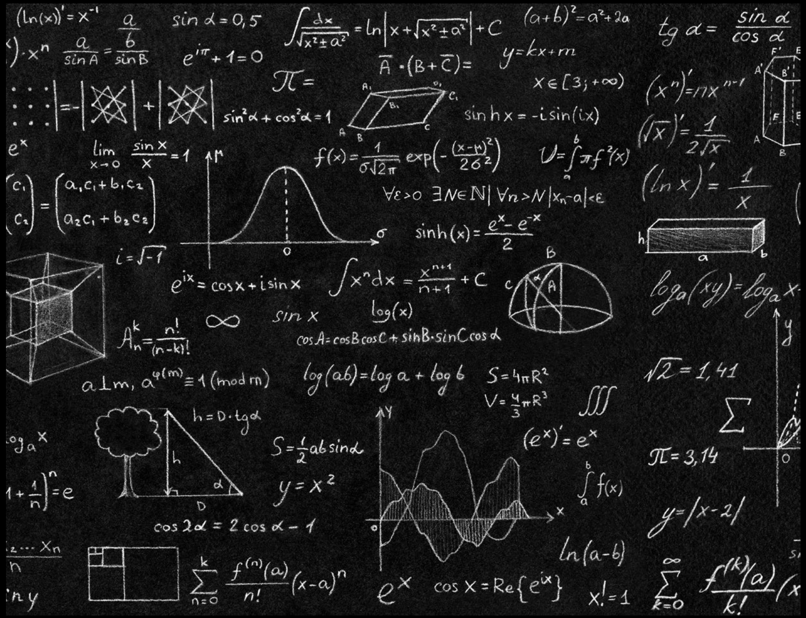
Section

- 01 **Business Scenario:** *Executive Styled Summaries*
- 02 **Version 1.0:** Architecture & Scaling Considerations
- 03 **Version 2.0:** Architecture, Data Curation & Fine-Tuning
- 04 **Demo:** Show & Tell
- 05 **Finetuning Deep Dive**
- 06 **Solution with Databricks**
- 07 **Best Practices**
- 08 **Q&A Session**

BUSINESS SCENARIO: EXECUTIVE STYLED FINANCIAL SUMMARIES

Transform complex content to actionable summaries

Gen AI-powered tool designed to empower executive-level decision-making by curating and summarizing content into concise, impactful summaries, articulated in a manner that aligns with executive communication styles.



VERSION 1.0: ARCHITECTURE DECISION

Prompt Engineering

Prompt Engineering is the art of **crafting effective prompts** to extract the desired output from AI language models

Fine Tuning

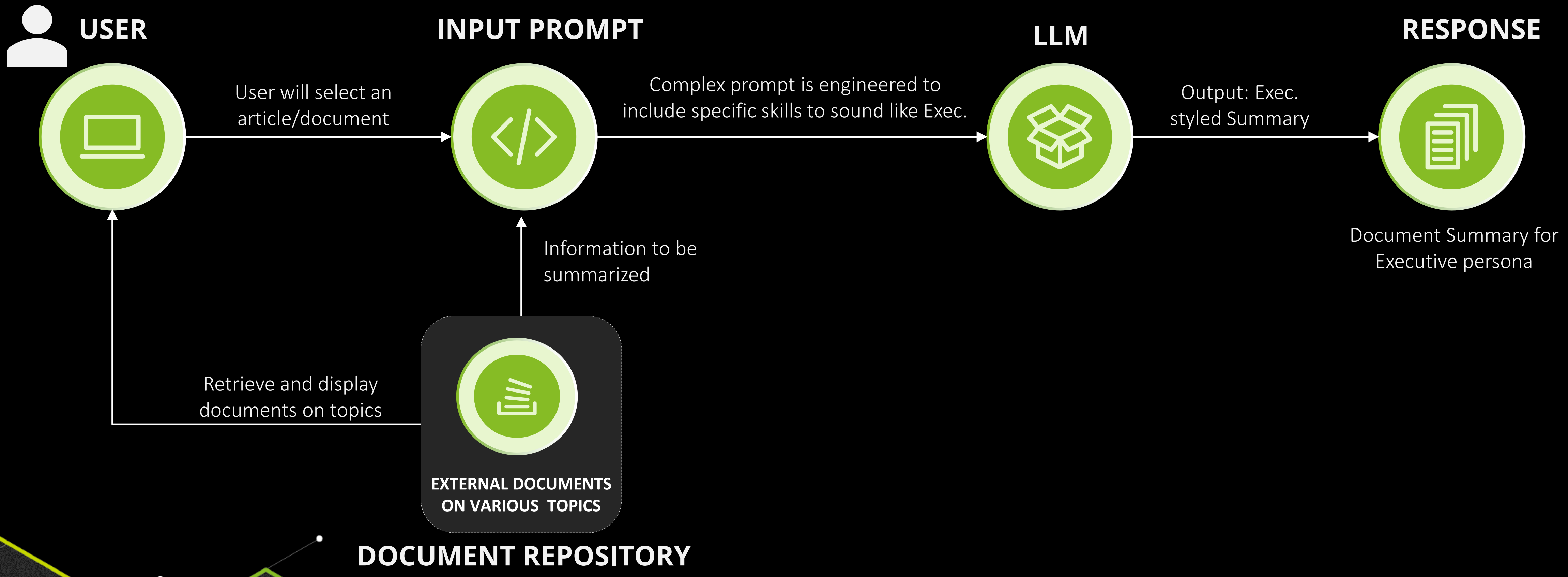
It involves **adjusting and adapting a pre-trained model** to perform specific tasks or to cater to a particular domain

Build your Own LLM

Building a Foundational model on broad data at scale such that it can be adapted and deliver task as per design

VERSION 1.0: PROMPT ENGINEERING

Initial approach leveraged PROMPT ENGINEERING



VERSION 1.0: SCALING CONSIDERATIONS

While pre-trained LLMs offer a powerful starting point, they may not be fully optimized

Response Times

- Complex prompts extend response times.
- multi step prompts increase processing times

Financial Efficiency

- Long prompts drive up token usage
- Potential high costs at scale

Customization

- Tailored solution for specific needs
- Ensure operational flexibility and growth potential

VERSION 1.0: SCALING CONSIDERATIONS

While pre-trained LLMs offer a powerful starting point, they may not be fully optimized

Output Reliability

- Ensure consistent quality in results
- Minimize bias and hallucinations

Control Over Training Data

- Maintain quality of training data
- Monitor data for consistent accuracy

Data Confidentiality

- Intellectual Property Risk
- Regulatory Compliance Risk
- Data Sovereignty Risk

VERSION 2.0: ARCHITECTURE DECISION

Prompt Engineering

Prompt Engineering is the art of **crafting effective prompts** to extract the desired output from AI language models

Fine Tuning

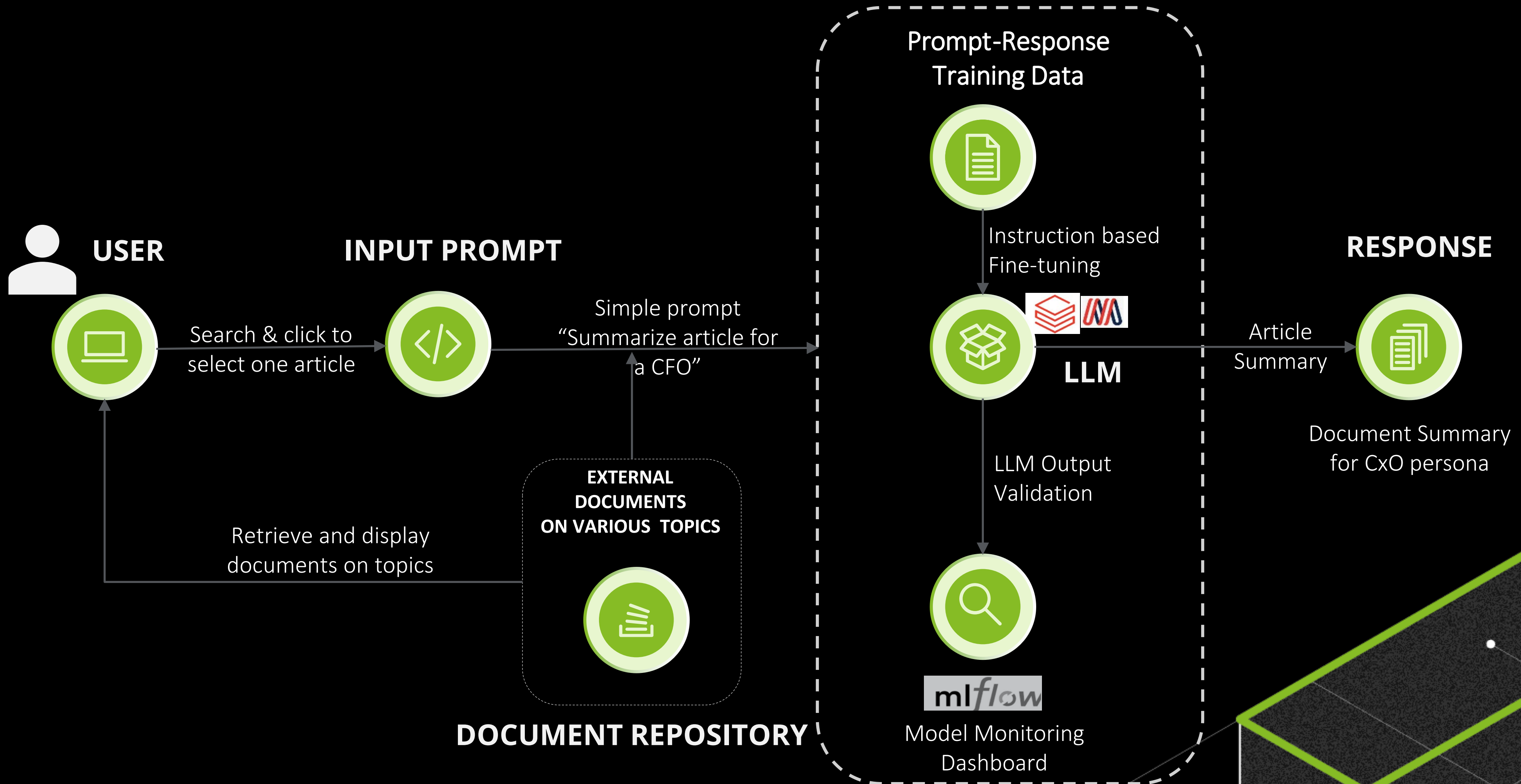
It involves **adjusting and adapting a pre-trained model** to perform specific tasks or to cater to a particular domain effectively

Build your Own LLM

Building a Foundational model on broad data at scale such that it can be adapted and deliver task as per design

VERSION 2.0: DATABRICKS FINE-TUNING

Adapting pre-trained model to our business scenario



DEMO : EXECUTIVE STYLED SUMMARIES

Version 2.0

RESULTS SUMMARY

Overview

We were able to fine-tune pre-trained models like Mistral 7B in less than 30 minutes. This efficient fine-tuning process, which used a combination of human-generated and model-generated samples, led to promising results, demonstrating the power of Databricks in accelerating fine tuning on LLMs.

Data Preparation & Curation:

- Model: Llama3
- Golden summaries = 3
- Training Data set : 5000 samples
- LLM-as-a-judge: DBRX

Model Training:

Input. Dataset [Generated]

- Training: 4500 samples
- Evaluation: 500 samples (10 golden samples)
- LLM-as-a-judge: GPT-4

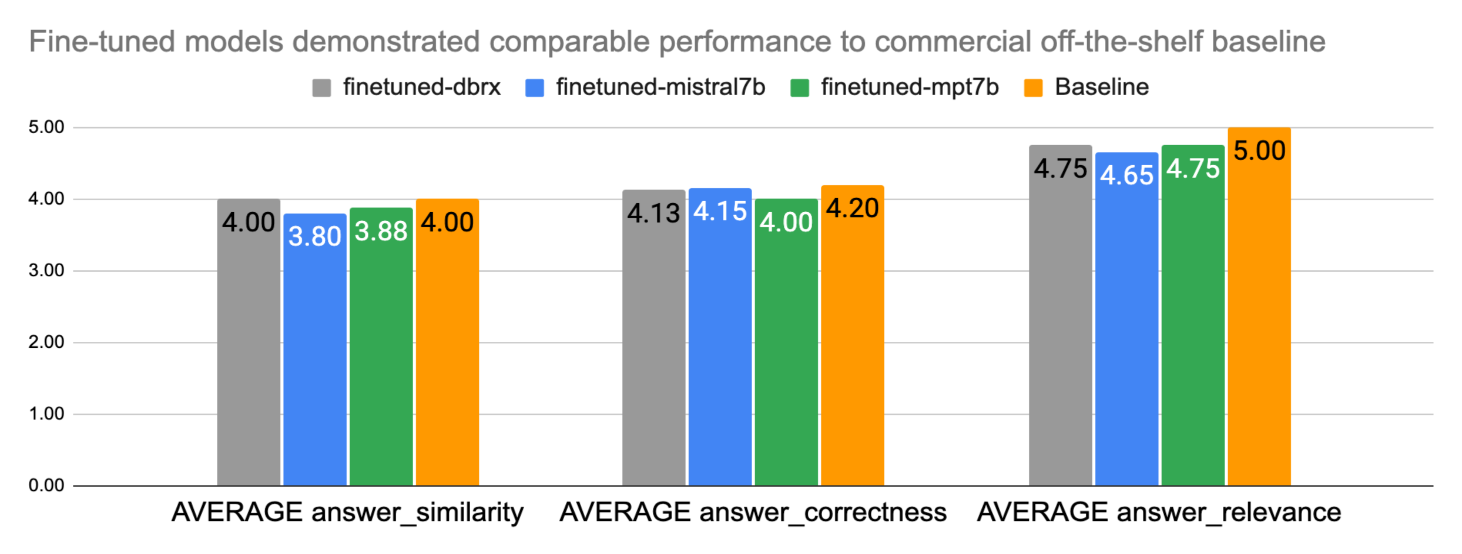
Models Fine-tuned: MPT-7B, Mistral 7B, DBRX

Number of runs: 12

Fine Tuning Results and Stats

- **MPT 7B** Initial test run showed promising results
- **Mistral 7B and DBRX** was used as the main model for Fine tuning

Performance Summary



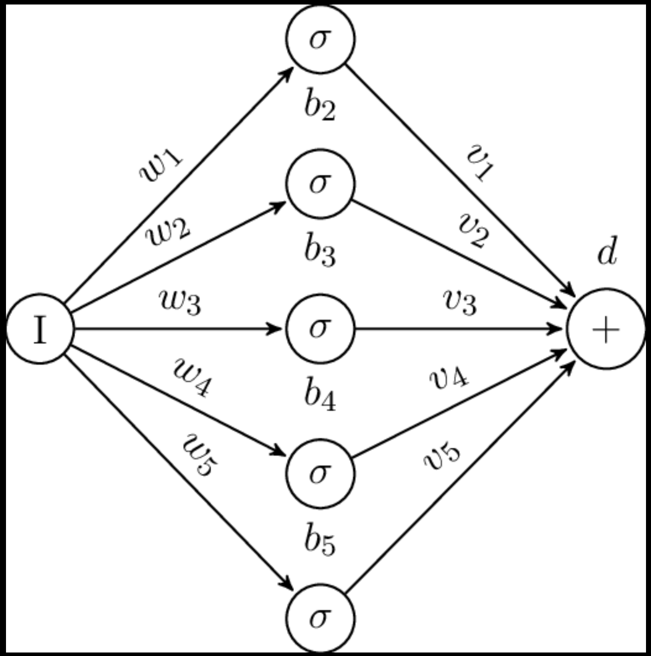
FINE TUNING: FUNDAMENTALS

A smaller fine-tuned model can outperform a larger model

What?

Fine-tuning is a **machine learning technique** that involves **further training a pre-trained model** on a smaller specific dataset to adapt it to a **new, specific task or domain**.

But, What?



The weights and bias gets adjusted

Train from the scratch?

- A lot of Data (Billions to **Trillions** worth of Data)
- And time (days to several **weeks**)
- Higher **specialization** is needed
- Need **Expensive** GPUs

FINE TUNING: Use-cases

Continued Pretraining

- Data format: Text
- Data size: *Billions* of tokens
- Resources: 100s of GPUs
- Time: Days of training

Gives Factual knowledge

Objective: Domain specific language or knowledge

Instruction Fine-tuning

- Data format: Prompt & Response
- Data size: *Millions* of tokens
- Resources: 10s of GPUs
- Time: Minutes of training

Change response style

Objective: Specialized model for specific tasks

FINE-TUNING CONSIDERATIONS

Qualifying Fine-tuning

When to Fine-tune?

- Want full ownership over a custom model for data privacy
- Tried **few-shot learning** and **prompt engineering**
- You are latency-sensitive or cost-sensitive and want to use a smaller, cheaper model with your task-specific data.
- Accessing **up-to-date data**
  - **Do you need to incorporate a lot of latest information?**

FINE TUNING: CHALLENGES

Data

- For supervised learning tasks, high quality labeled data is needed
- Data distribution design
- Manual tokenization and pre-processing to make it model compatible
- Optimizing for memory

Operational

- Open-source libraries requires **extensive knowledge of model architecture.**
- **Too many training knobs** - Batch size, epochs, optimizer, regularization
- Manual Management of intermediate model states, checkpointing, etc,.
- Integration with ML stack

Infrastructure

- **Complex** Networking and run configurations
- Bad Configuration can lead to inferior performance and hardware failures [OOMs]
- For multi-node trainings, additional settings such as sharding
- GPU availability and cost
- Scalability

DATA PREPARATION & CURATION

Fine-tuning often requires thousands of high-quality, human-labeled question and answer pairs

Synthesize machine-generated article & summarization pairs

Conduct Tone Analysis on business-generated examples to identify key tones and summary elements

DBRX & Llama3-70B

Develop Prompt for model-generated executive summarizations

Generate model summaries for article dataset

Llama3-70B

Evaluate model summaries with LLM Judge

DBRX

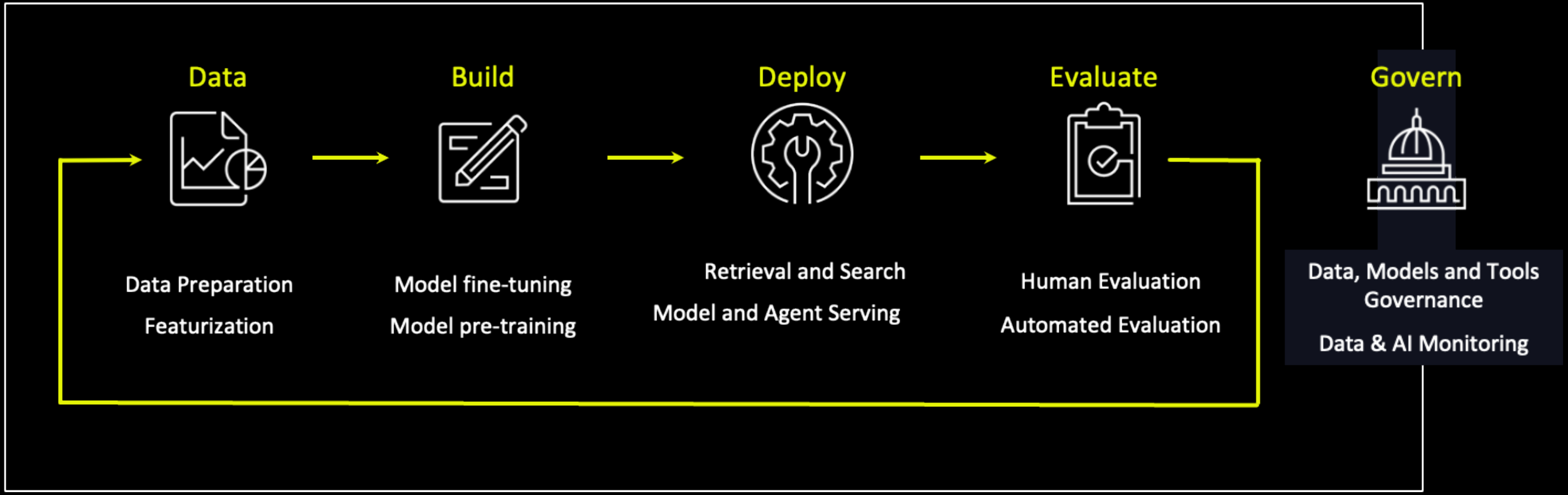
Collect feedback from business users

Iterate as necessary

DATABRICKS MOSAIC AI FINE-TUNING

Up to 7X faster and cheaper training of large AI Models

A **simple** application interface to fine-tune AI models with **minimal configuration** in a **secure**, **scalable** and **cost-effective** way.



EVALUATING FINE-TUNING MODELS

Looking for quality, coherence, and relevance

Our comprehensive evaluation Framework was three-fold comprising of simple metric evaluation, LLM as a judge and human feedback.

Intrinsic Evaluation

- **Token Accuracy:** overlap of n-grams between generated and actual summaries. Higher the better.
- **Perplexity:** Measures how well the language model predicts the next word in a sequence, with lower perplexity indicating better performance

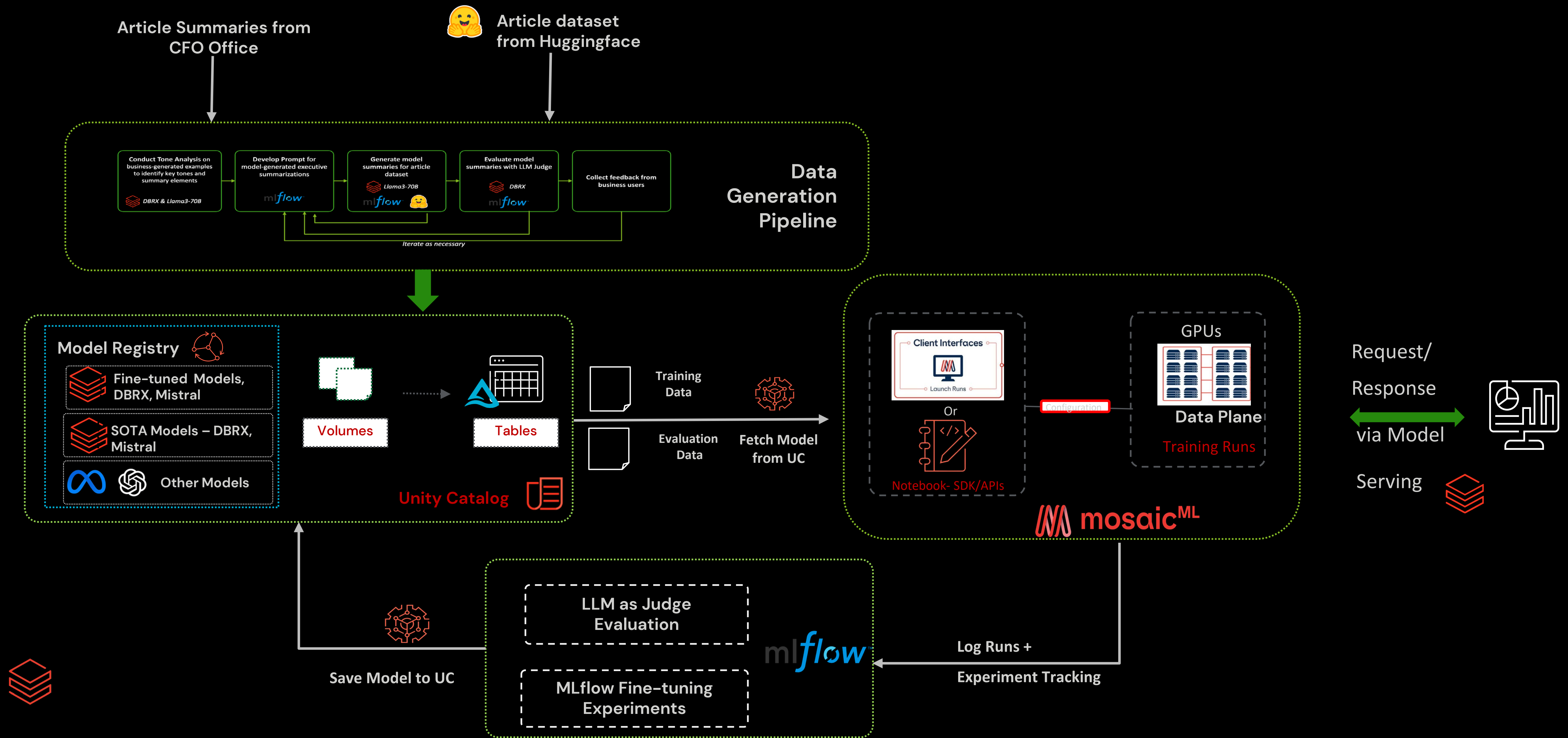
LLM-as-a-Judge (MLflow)

- **answer_similarity():** How semantically similar the model prediction is to the ground truth label (business summary).
- **answer_correctness():** How Factually correct the generated summaries are.
- **answer_relevance():** How relevant the model summaries are to the input article.

Human Feedback

- Conciseness
- Contextual Understanding
- Coherence
- Completeness
- Customization

Let's put it together



IMPACT OF FINE-TUNING A MODEL

For the article summarization problem, a model before and after fine-tuning

Input

A 2500-word article: “Organizations can leverage industry clouds to enable strategic transformation and stay on the cutting edge Competition is at the heart of transformation. It requires continuous innovation and new ways to grow profits. Digital transformation helps organizations innovate faster and better. And therein lies one of the greatest challenges: pursuing digital transformation means building a race car and entering a never-ending race....”

Prompt Engineered Output

Industry clouds are becoming a strategic tool for organizations looking to maintain a competitive edge in the digital landscape. These solutions standardize foundational ...g businesses to concentrate on differentiating their unique capabilities. ...lability, and adaptability, with potential access to a significant portion of ... market. Differentiation can be achieved through strate... industry clouds. If a suitable solution isn't available, org... partner with cloud providers to cocreate industry cloud services. Implement... new operating model, aligning engineering teams with business product own... clouds mature, they not only allow organizations to focus on strategic abilities but also significantly enhance their ca... akeaways include the importance of understanding the value of indu... a new operating model, and exploring cocreation if needed. Action... g areas for differentiation, aligning engineering and business tea... ering industry clouds as a part of the digital transformation strategy.

Hard to follow

Missing Stats

Inconsistent Responses

Finetuned Output

Overview: The article discusses the use of industry clouds to e... ormation and stay competitive in the digital race.

Key Takeaways:
1. Industry clouds are modular building blocks that speed up the development of industry-specific digital solutions, allow... odernize and innovate incrementally.
2. Organizations should define... d focus on the user journeys that matter the most, using industry cloud... pment for their top strategic use cases.
3. Implementing industry clou... new operating model, with internal resources focusing on the 20% where the organization plans to differentiate.

Action Items for CFO:
1. Evaluate the potential benefits of industry clouds for your organiza... top strategic use cases to focus on.
2. Consider the new operating model required for implementing indu... allocate resources accordingly.

Adapted Style

Digestible bullets

Clear action items

BENEFITS OF DATABRICKS FINE-TUNING

Fine-tune models faster, cheaper and at scale easily and securely

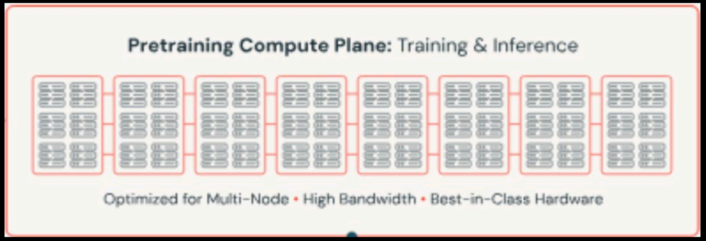
Ease of Use

- Simplified API
- Fully managed and Serverless
- Multi-model support
- Fully Integrated with Mlflow and Mosaic AI such as model serving

```
run = ft.create(
  model='databricks/dbrx-base',
  train_data_path='dbfs:/Volumes/main/schema_name/ift/train.jsonl',
  register_to='main.schema_name',
  training_duration='2ba',
  task_type='INSTRUCTION_FINETUNE'
)
```

Scalability

- Auto-scalability
- Auto-checkpointing
- Full abstraction and management of low-level infrastructure settings such as sharding etc



Governance

- Full control of the data and model
- End to end governance with Unity Catalog
- Lakehouse monitoring
- Databricks AI Security Framework (DASF)

BEST PRACTICES

Make the Best out of your runs!

Data Preparation & Curation

- Still Garbage in Garbage out
- High quality > high quantity; but more data is (almost) always better
- Follow ML Data preparation best practices
- Pay attention to the Data format
- Multiple Data source:
  - **Tune hyperparameters and up-sample**

Training

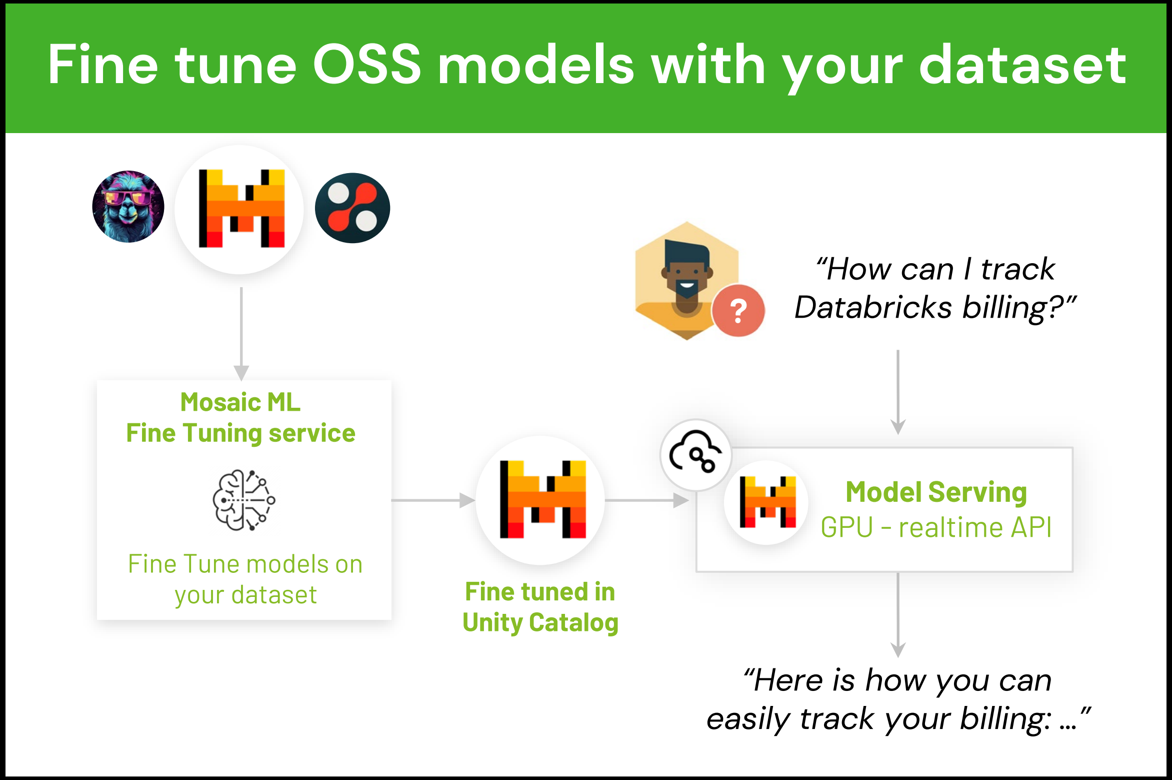
- Hyperparameter sweep with altering Training Duration and Learning Rates
- Experiment with wide variety of Learning Rates- 1e-4, 3e-5, 1e-5, 3e-6, 1e-6, 3e-7
- Intrinsic Metrics alone are not enough
- Create metrics that are specific to your task

Iterative Experimentation

- This will be your best friend and true guide
- The only way to find what works, is to work it
- Always monitor your experiments

CALL TO ACTION!

Try Mosaic AI & LLM Fine Tuning now!




```
dbdemos.install('llm-fine-tuning')
```

[Open the demo page]