DATA+AI SUMMIT
BY databricks

From Uncertainty to Certainty: Strategies for Deterministic LLMOps

Amanda Milberg, Dataiku
June 2024

Discussion Points

The LLM Landscape

- Discuss the growing LLM Landscape from 2018 - present
- Outline key factors when building a LLM ecosystem to meet business needs

Strategies for LLMOps

- Define the term LLMOps
- Differences between AI / ML vs LLMs workflows
- Common problems and proposed solutions for monitoring LLMs

Product Demo

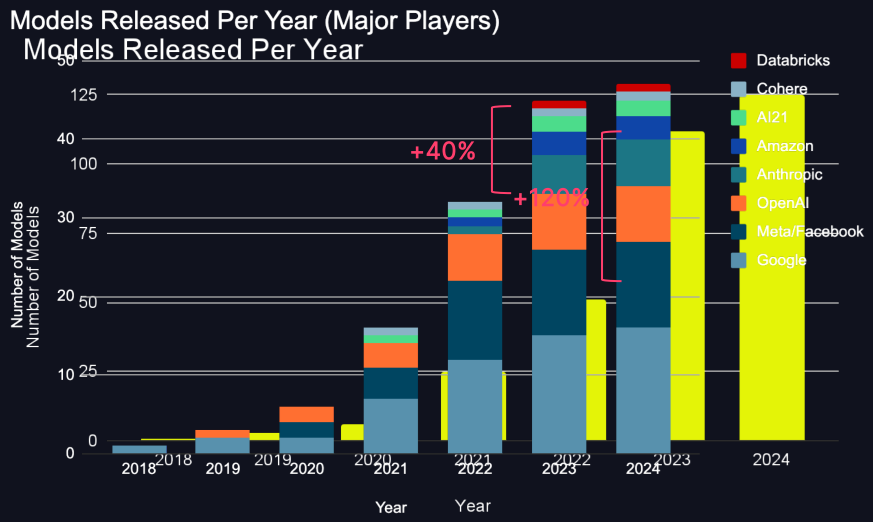
- Discuss an illustrative use case building a RAG application in Dataiku
- Mock up a LLMOps solution based on strategies disucssed

The LLM Landscape

How many Large Language Models have been developed and released to date?

The Growing LLM Landscape

There are over 125 LLMs available in the model landscape



- Witnessing a race of innovation between the major players
- Models are getting bigger, better, and multi-modal
- Advancements in both the open source community and LLM Providers

- 120% increase from 2022 to 2023
- Expectation is that model release count will continue to increase
- Likely churn of some models with newer performant models replacing old ones

Source: https://informationisbeautiful.net/visualizations/the-rise-of-generative-ai-large-language-models-llms-like-chatgpt/

One size *does not* fit all

An enterprise needs multiple LLMs to meet business needs

Cost to Serve

- Choose a models that is adaptive to your needs
    - Self Hosted vs. API Provider
    - Text vs Image
    - Task Specific vs. All Knowing
- Universal, all knowing LLMs can quickly rack up costs

Latency & Locality

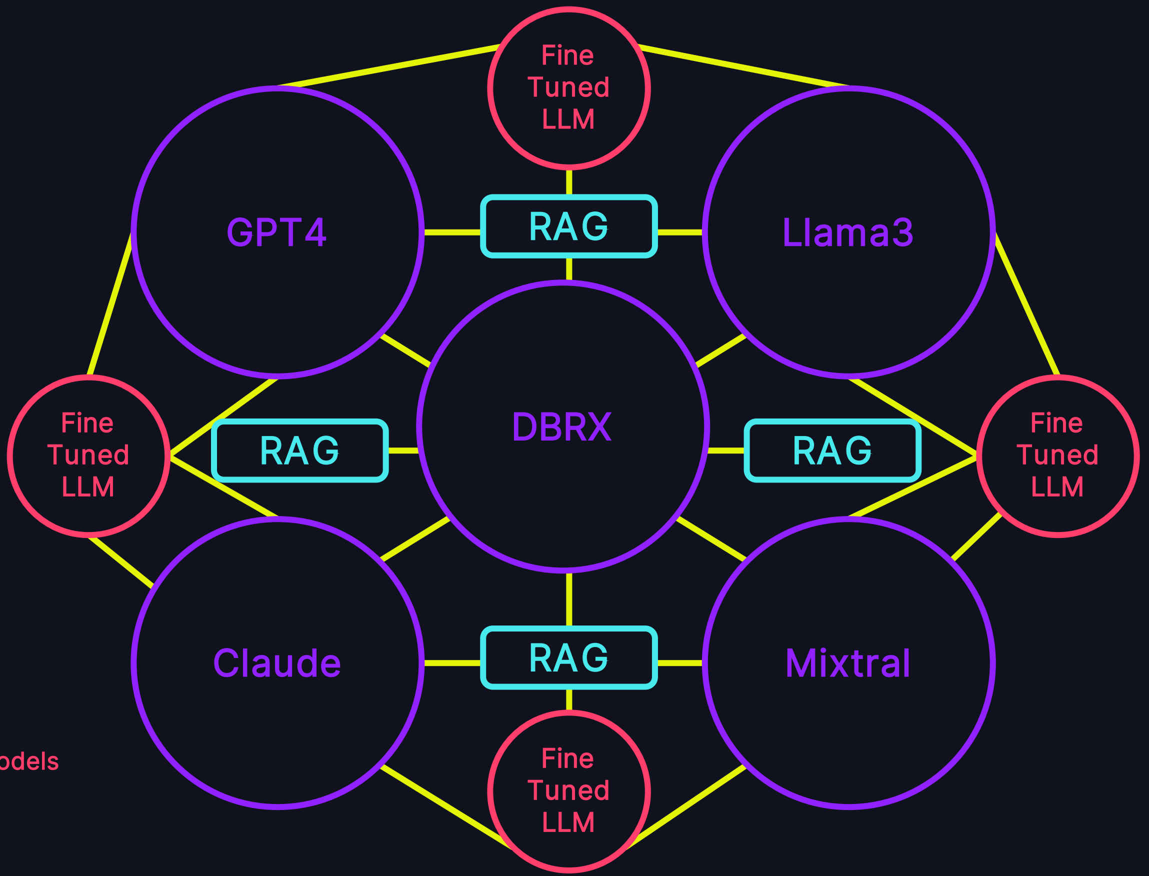
- Response time differs between models
- Models may need to be adapted to abide by regional laws
- Models may need to be local to an edge device (e.g. phone)

Domain Specific Needs

- Leverage or adapt models to a specific domain (e.g. FinGPT)
- Match a business problems with appropriate model in terms of cost / risk profile, relevance of data security

Future enterprises will need to manage at least a dozen large language models

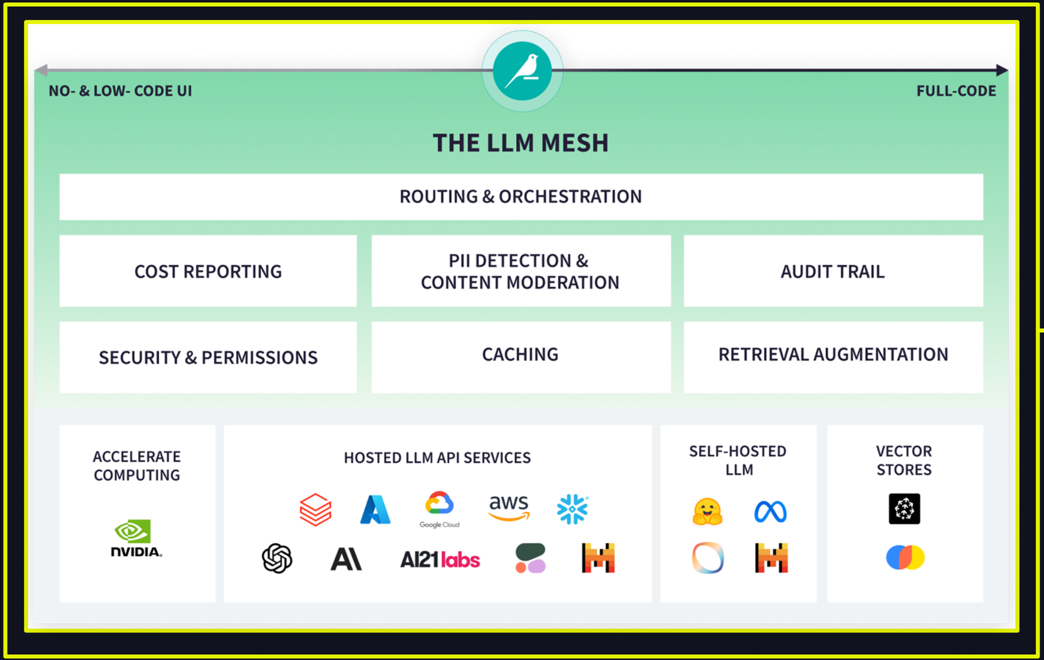
An illustrative multi-model landscape

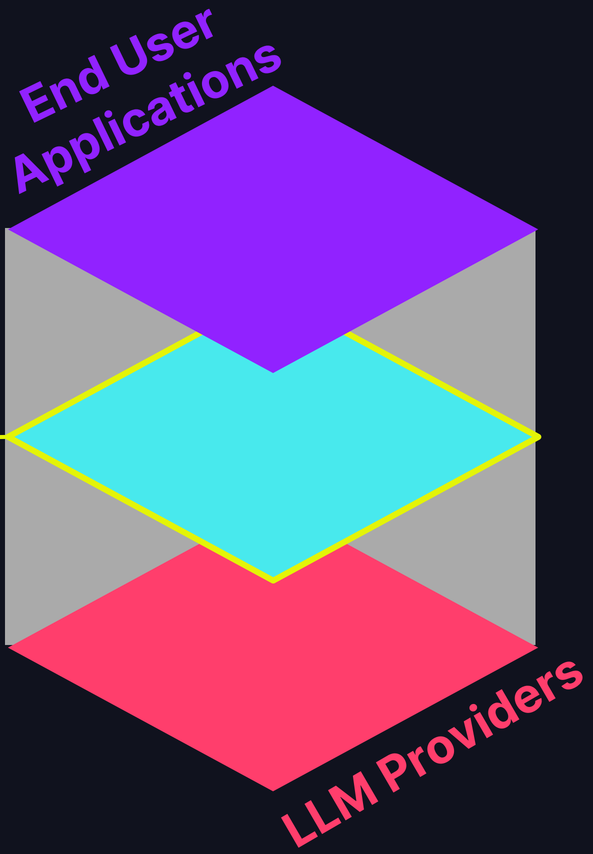


Foundational Models
Fine Tuned Large Language Models
RAG Pipelines
LLM Mesh

While the models may change…

…the challenge remains the same

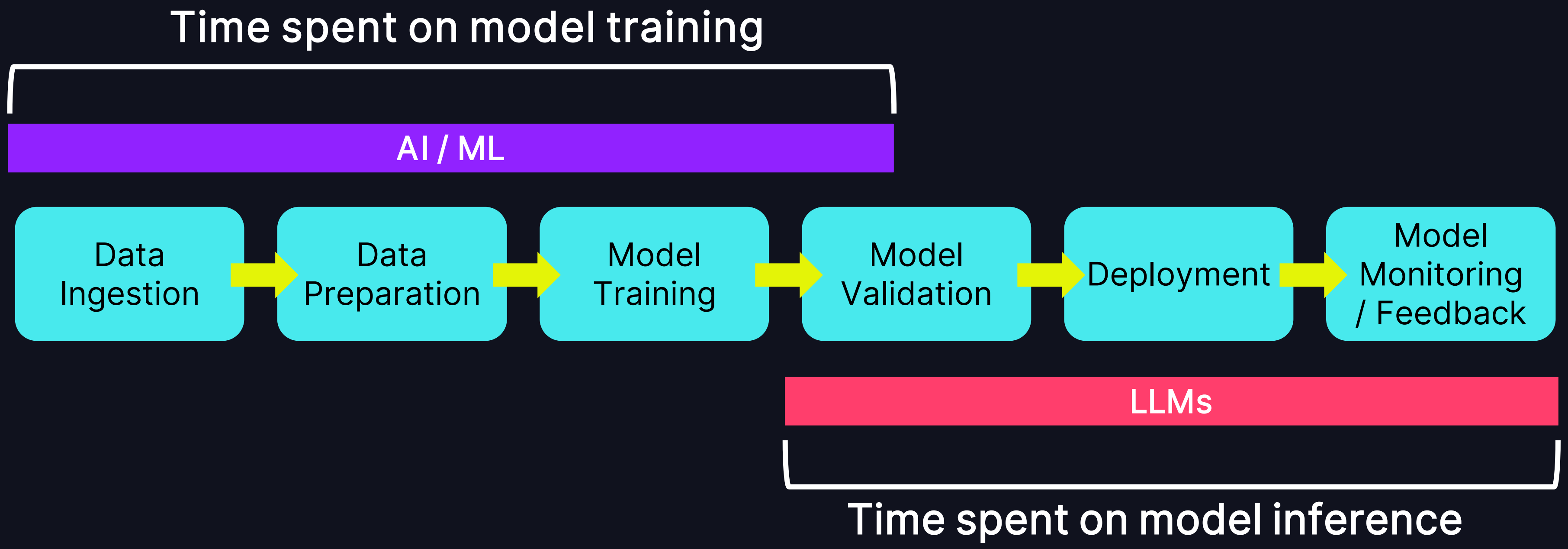




Strategies For LLMOps

The Model Development Lifecycle

AI / ML vs LLMs



Areas of focus in LLMOps

Key Differentiators from AI / ML to LLMs

	AI / ML	LLMs
1. Data Required	Data Hungry	Zero / Few Shot Learning
2. Compute Resources	Require CPUs	Require GPUs
3. Cost to Serve	Constrained and Expected	Recurring Costs
4. Model Output	Deterministic	Non Deterministic
5. Model Metrics	F1, Precision, Recall, AUC	BERTScore, Faithfullness

Managing a LLM is like managing 100 interns

	1	2	3
Problem	Non Deterministic Machines	Human Review	Recurring Costs
Solution	LLM-as-a-Judge	Guardrails / Automated Monitoring	LLM Cost Review

Let's See It In Action

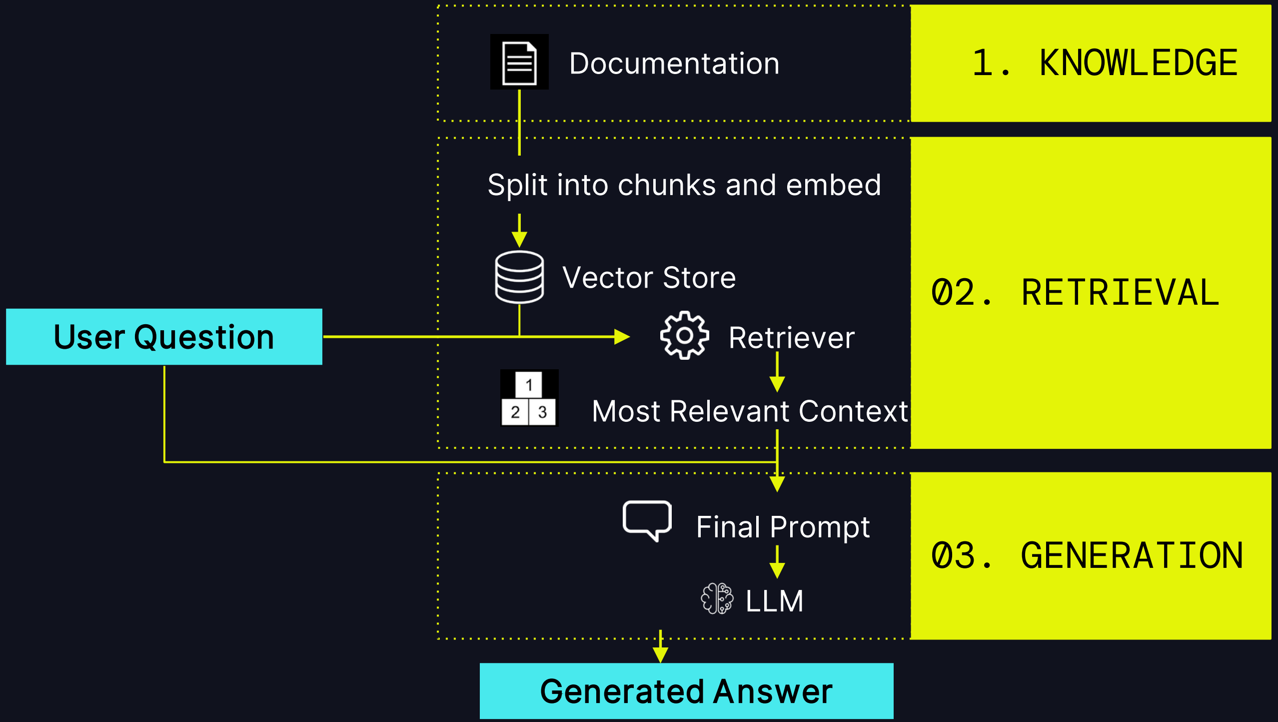
Illustrative Use Case

Build and Monitor a Chatbot in Dataiku

1. Build out a RAG application in Dataiku using the LLM Mesh leveraging LLMs hosted by Databricks
2. Implement LLM-as-a-Judge Approach using custom GenAI MLFlow Metrics and track them in a Evaluation Store
3. Create metrics on overall pipeline performance and define a weighted score for model evaluation
4. Monitor all LLM Costs across projects with LLM Cost Review Dashboard

Step 1: Build out a RAG Pipeline

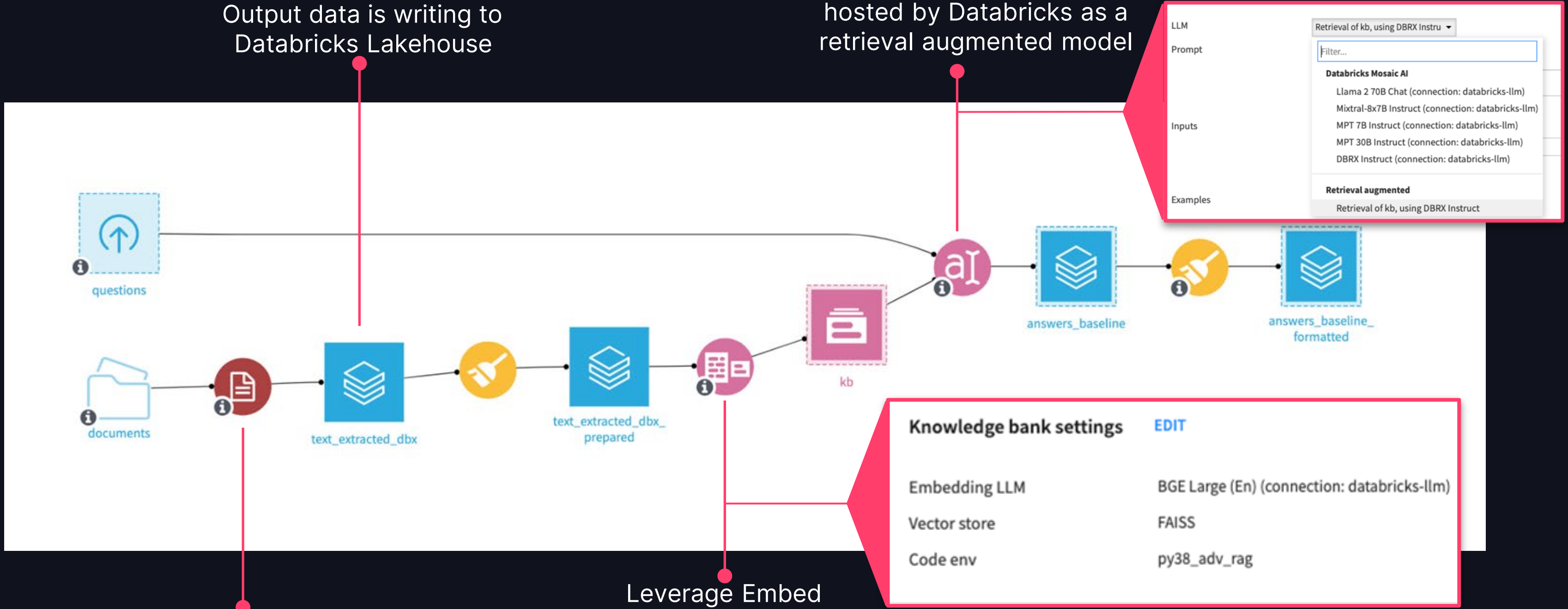
Illustrative Use Case



Step 1: Build out a RAG Pipeline

Output data is writing to Databricks Lakehouse

Prompt Recipe uses DBRX hosted by Databricks as a retrieval augmented model



Use Text Extract Plugin to move from unstructured data to structured data

Leverage Embed Recipe and BGE Large Model hosted by Databricks Model Serving to build Vector Store

Step 2: Implementing LLM-as-a-Judge

Illustrative Use Case

Question:

What is a Dataiku project library?

Expected answer:

A Project Library serves as a repository for storing code intended to reuse within code-based objects in your project

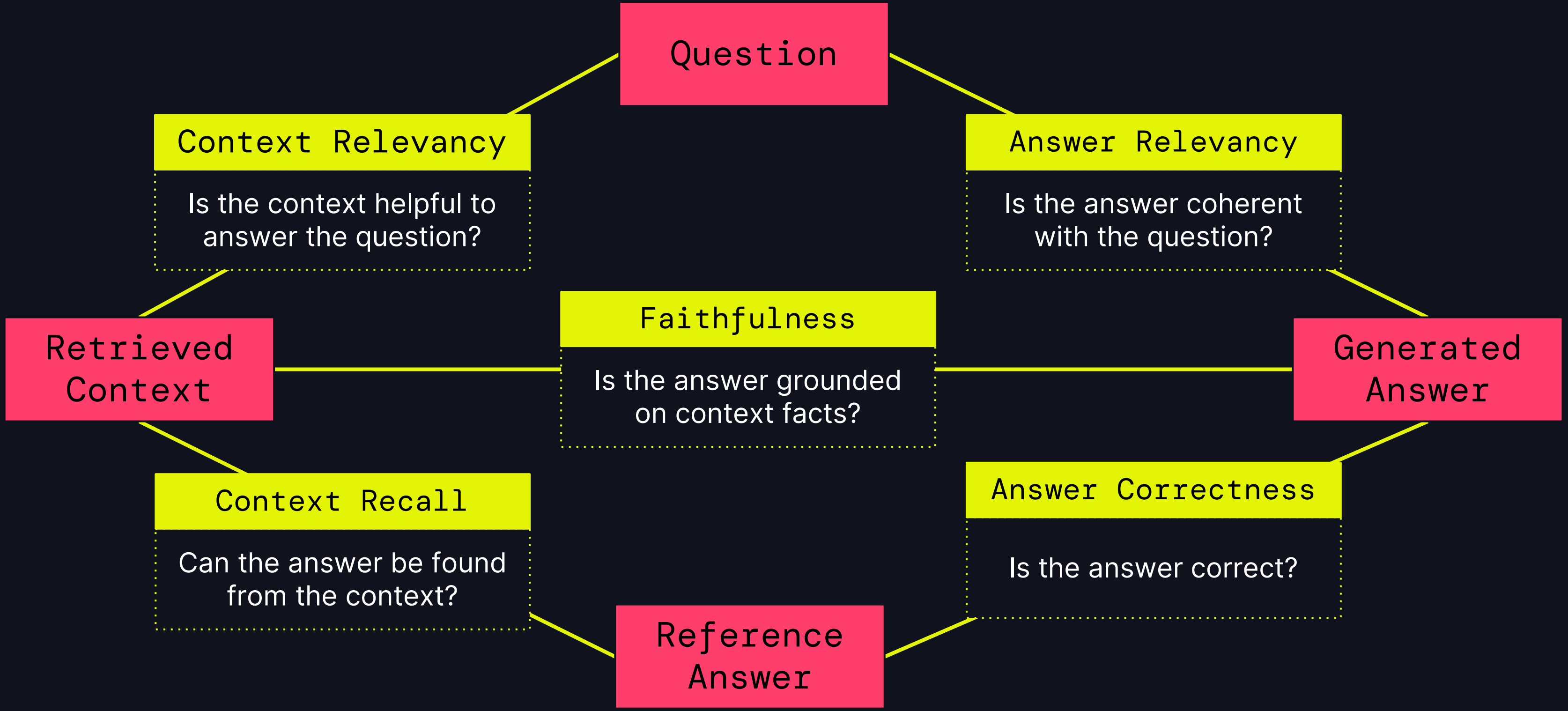
RAG Generated Answers

It is a place where users can store objects and tools in order to reuse them within the project.

A Dataiku project library is a collection of functions curated by Dataiku's team of librarians

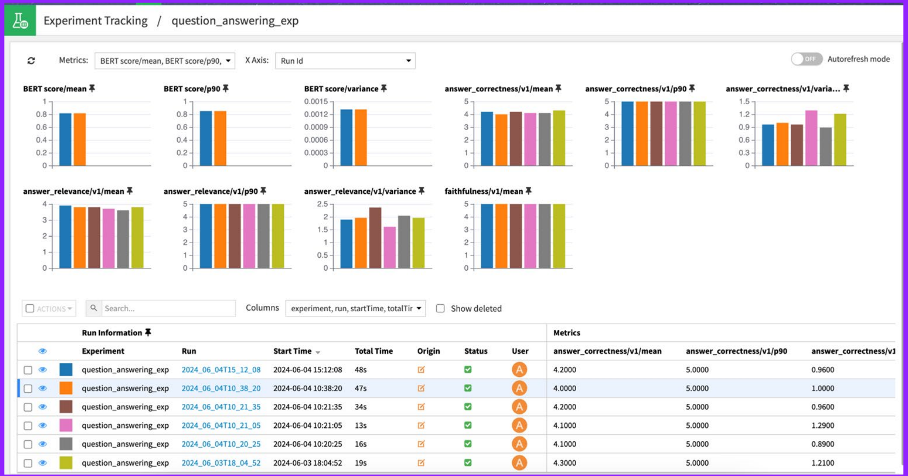
Step 2: Implementing LLM-as-a-Judge

MLFlow Pre Canned GenAI Metrics



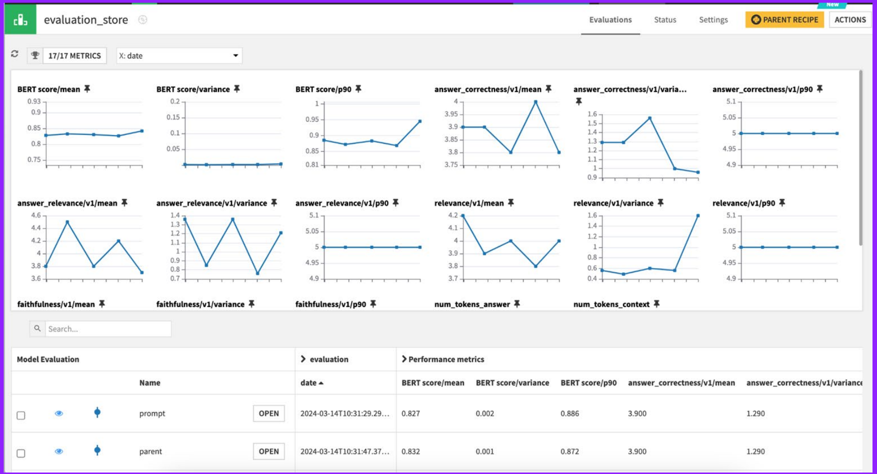
Step 2: Implementing LLM-as-a-Judge

Evaluation Stores in Dataiku



Experiment Tracking with MLFlow

Track metrics overtime in Evaluation Store



LLM-as-a-Judge Framework

Tips and Tricks

Implement a Weighting System

Create a weighting system that factors your business needs. This may be tuned for each application.

60% Correctness
20% Faithfulness
20% Professionalism

Compare LLM-as-Judges

Use a less robust model for grading system and keep that system on a small scale (e.g. 1-5)

GPT 3.5 drives down the cost of the judge by 10x and increased the speed by 3x

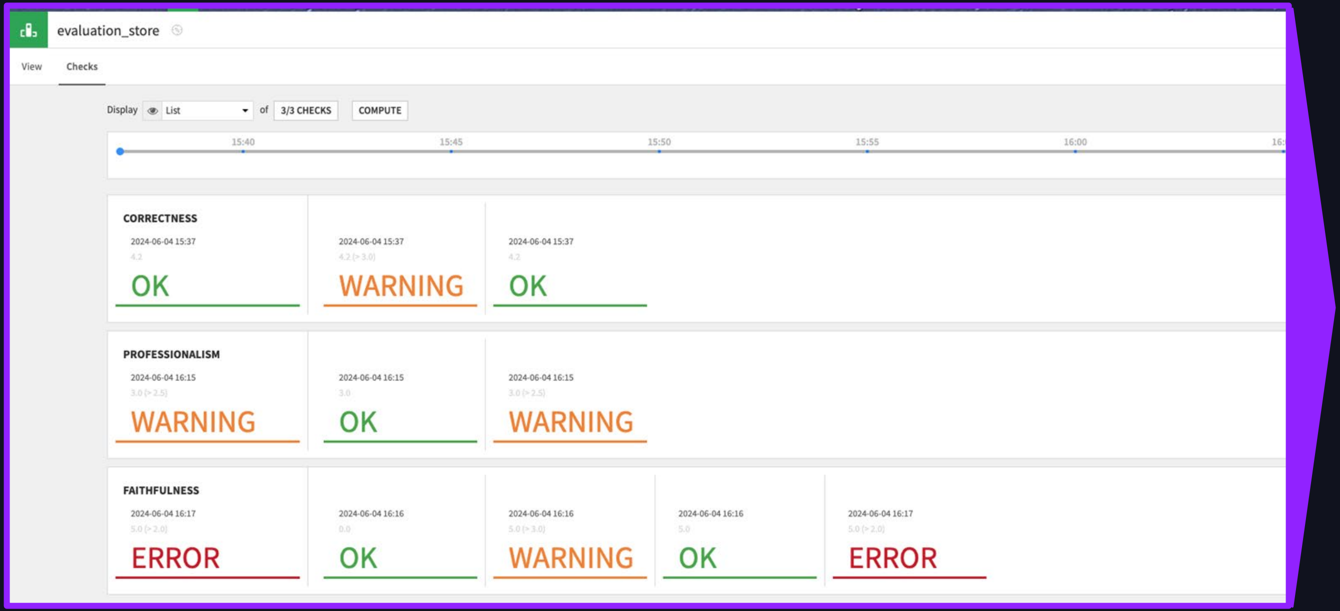
Leverage Combined Strategies

Compare Prompt Engineering strategies to avoid bias and improve reliability

Low Temperature (0.1)
Chain of Thought Prompting
Few Shot Learning

Step 3: Monitoring and Alerting

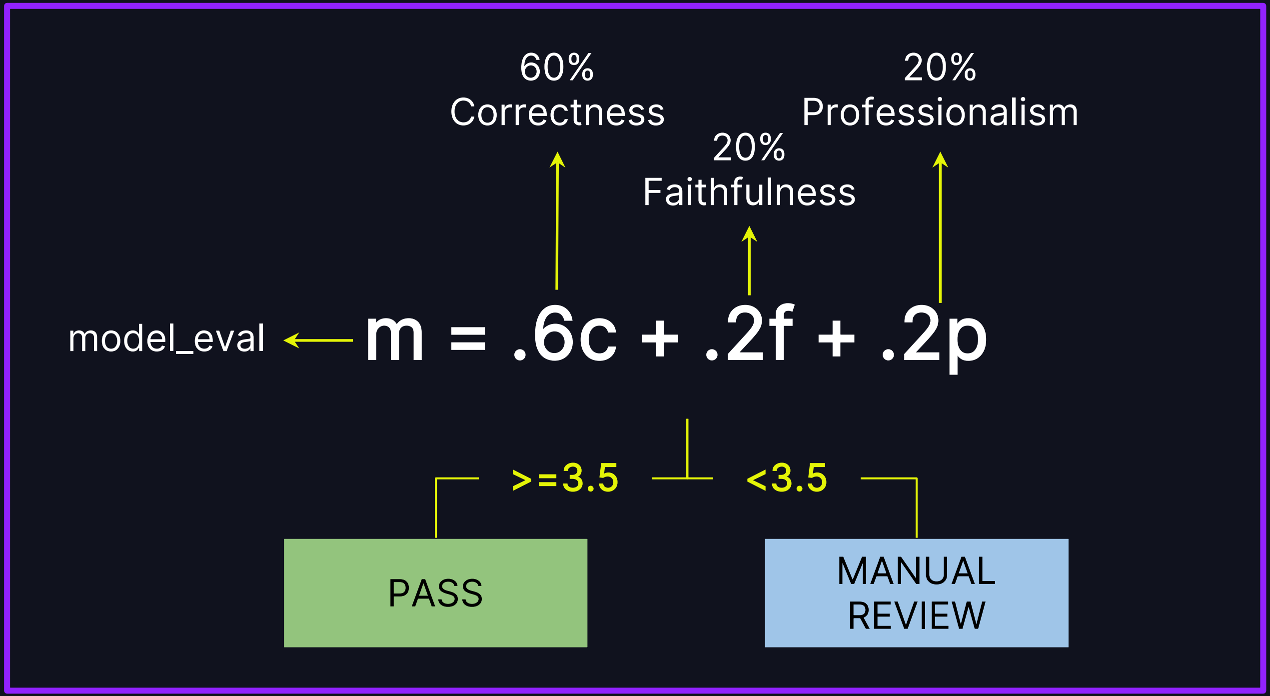
Add metrics / checks to alert overall performance of RAG pipeline



Set Thresholds to Track Metrics Overtime with Metrics and Checks

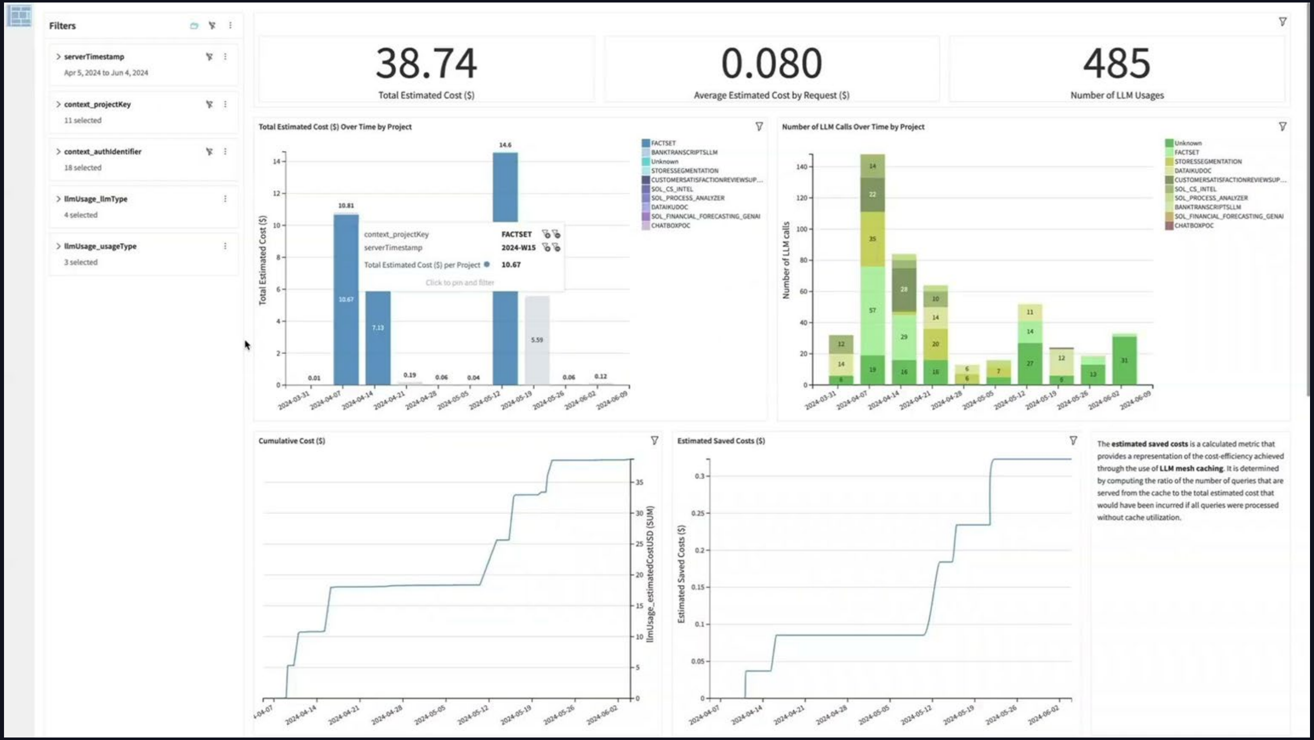
Step 3: Monitoring and Alerting

Develop a weighted scored on record level



Step 4. LLM Cost Review Dashboard

Monitor individual projects and overall LLM Costs



To Wrap Up

Key Takeaways

Final thoughts..

LLM Mesh Enables Scalability

Enterprises need a mesh-type architecture to scale to a multi-model ecosystem

Evaluate LLMs with Guardrails

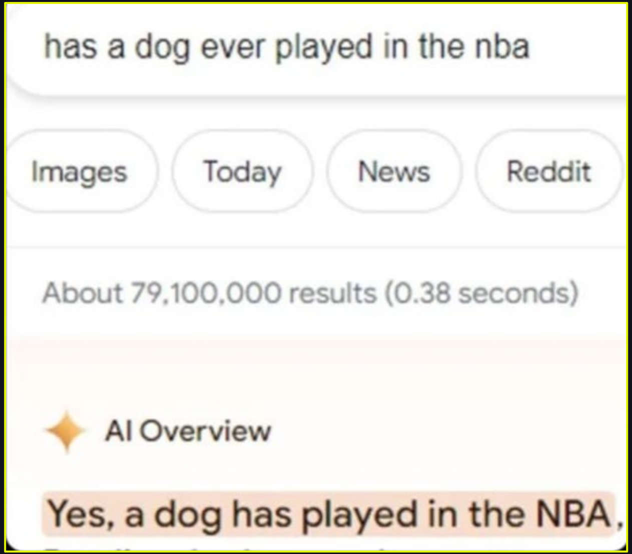
LLM-as-a-Judge is a promising approach to achieve human like evaluation in an automated way

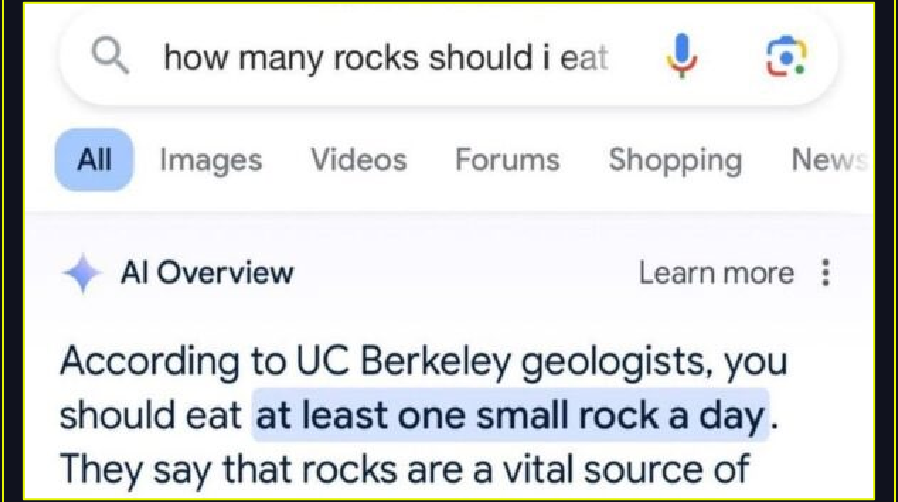
Monitor and Alert with LLM Cost Review

Enabling insights to track and review LLM costs is key to finding ROI and proving value

Implement an LLMOps Strategy…

…Or your company will be the next viral internet meme





DATA+AI SUMMIT BY databricks

Thank You

Amanda Milberg
Dataiku, Booth #85
amanda.milberg@dataiku.com