databricks

GenAI In Action: Build your first LLM App

September 2024

databricks

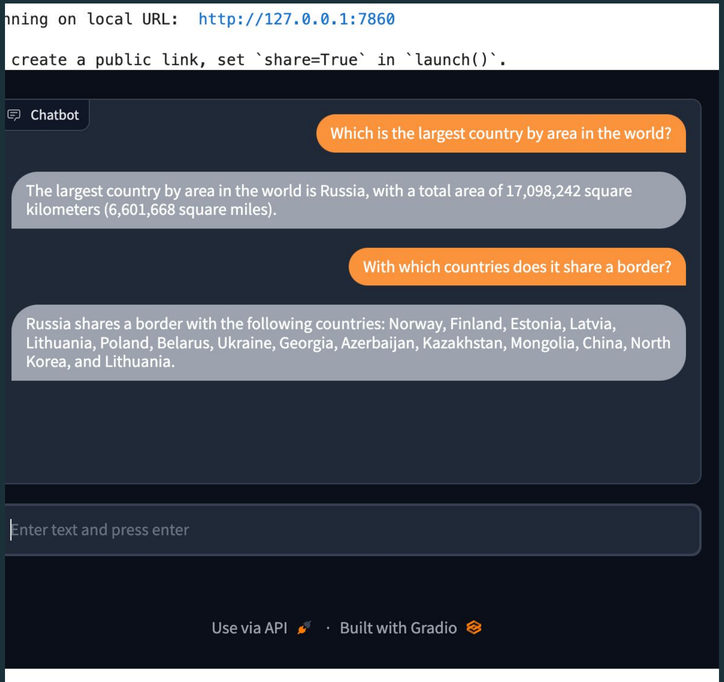
Your Host

Brian Law
Snr Specialist Solution Architect
Databricks

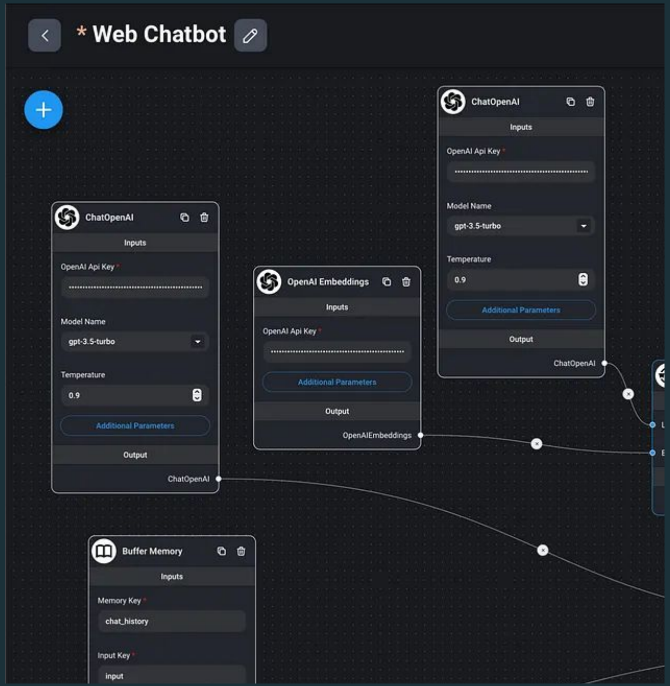
Lets chat to an AI

What goes in an AI Application

User Interface



Chain Orchestrator



Tools & Information

SQL Analytics

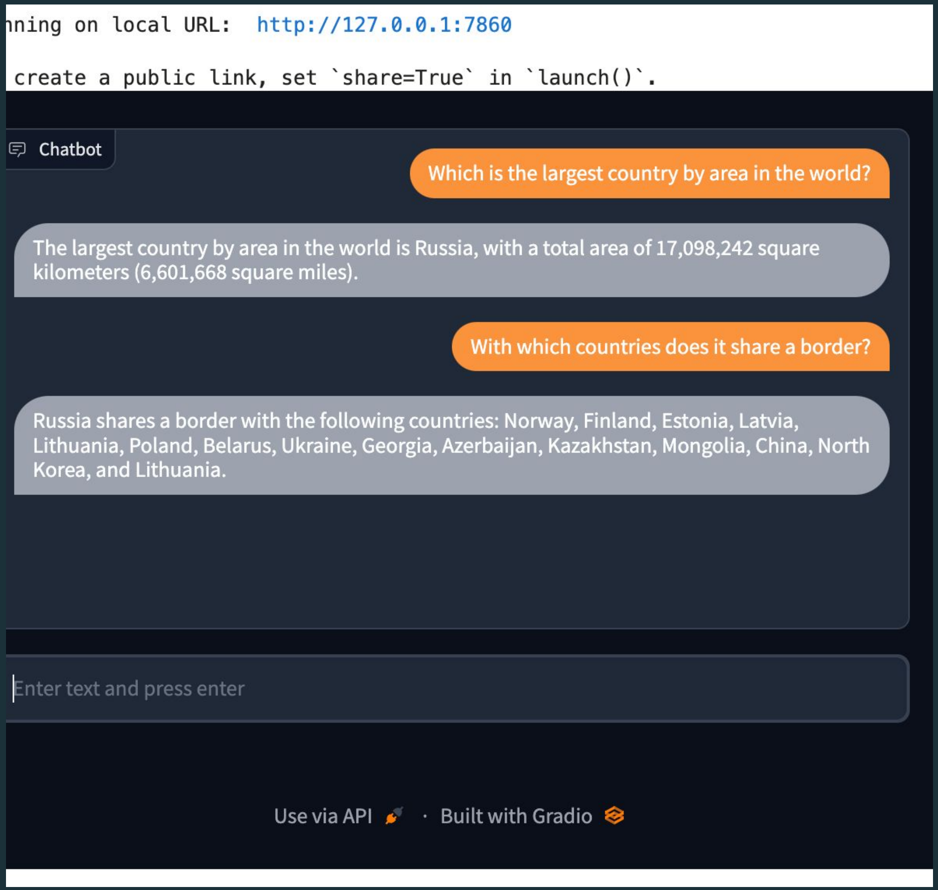
Internet

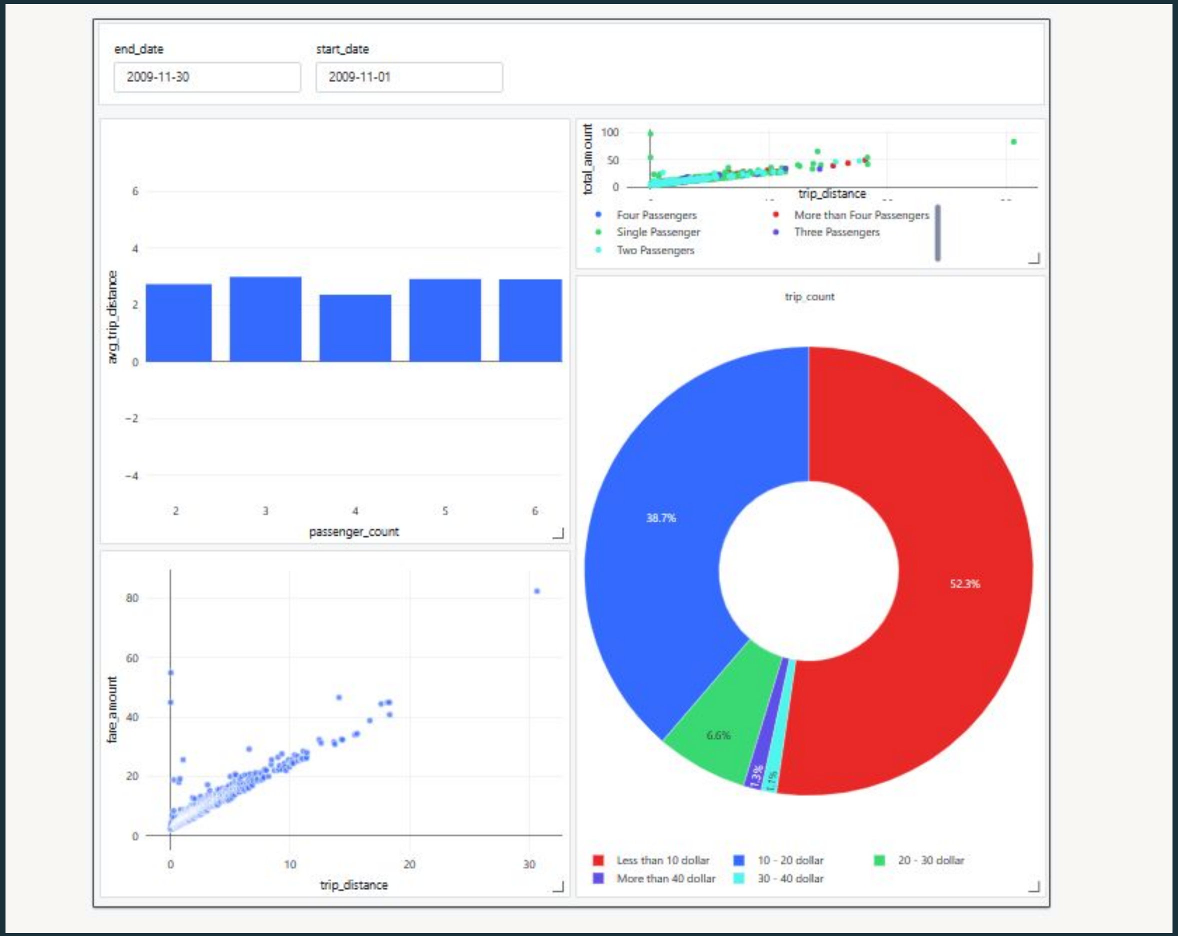
Vector Search

Parallels to Data Applications

User Interface

Dashboard

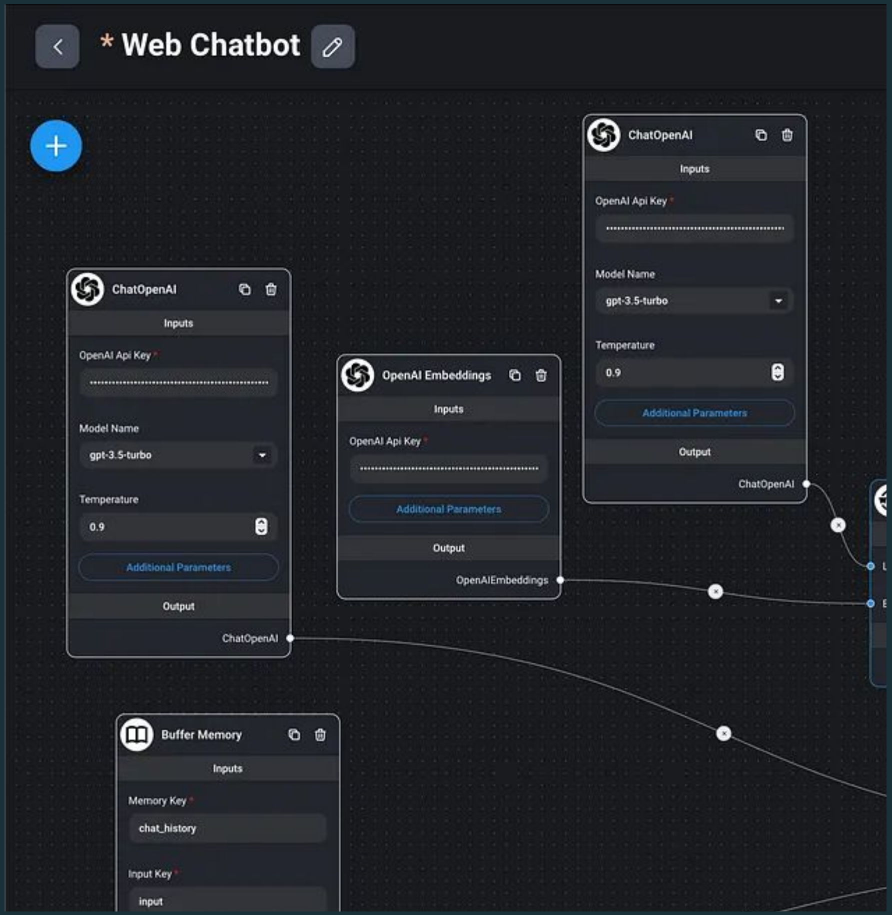


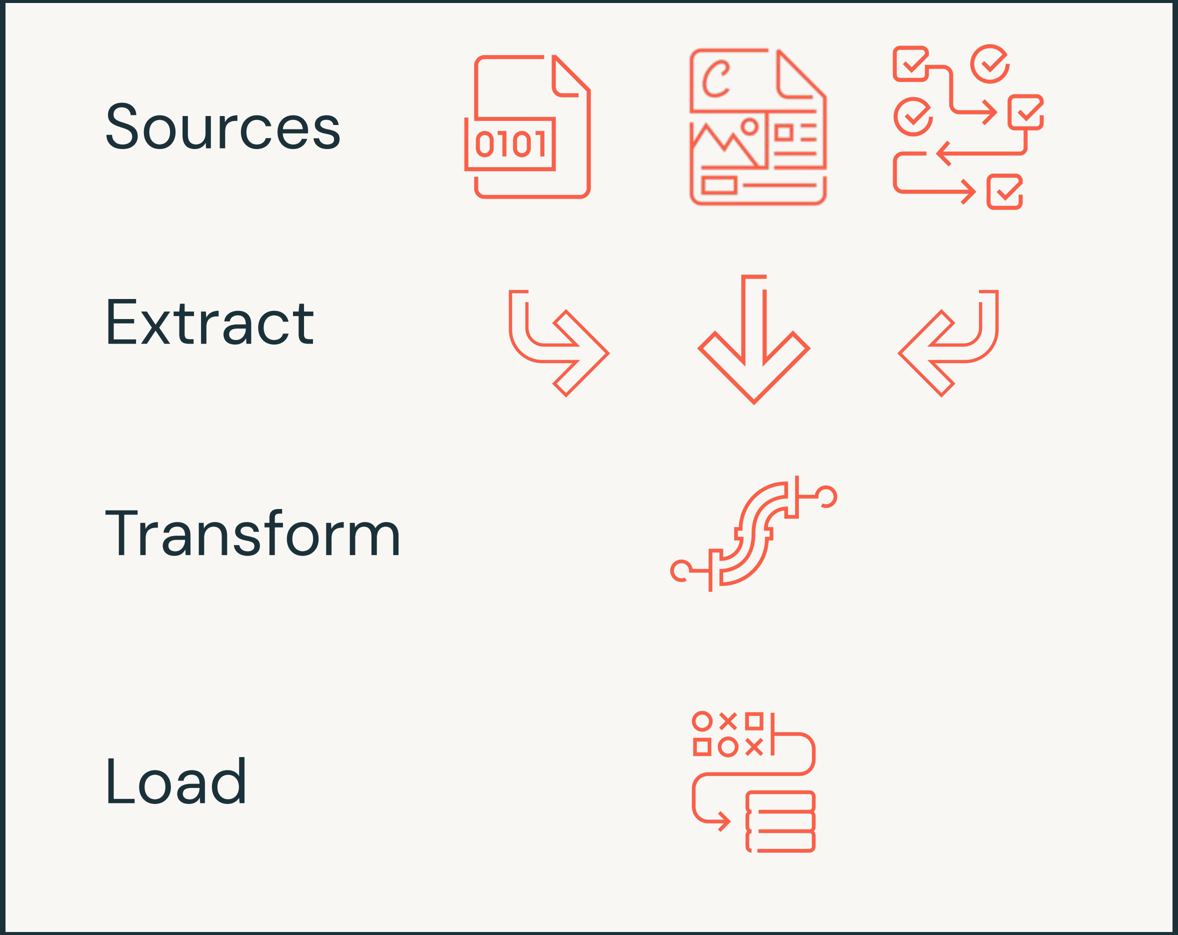


Parallels to Data Applications

Chain Orchestrator

Data Pipeline





Parallels to Data Applications

Tools & Information

Tables / Metadata

SQL Analytics

Internet

Vector Search

LLM as a Processing Engine

Parallels to normal processing

Database Engine – SQL

```sql
SELECT C.customer_name,
       C.customer_id,
       O.order_id,
       O.order_date,
       P.product_name,
       P.product_price,
       SUM(OL.quantity * OL.unit_price) AS total_order_cost
FROM customers C
JOIN orders O ON C.customer_id = O.customer_id
JOIN order_lines OL ON O.order_id = OL.order_id
JOIN products P ON OL.product_id = P.product_id
WHERE C.customer_state IN ('CA', 'NY')
GROUP BY C.customer_name,
         O.order_id,
         P.product_name,
         P.product_price
HAVING total_order_cost > 1000
ORDER BY O.order_date DESC;
```

LLM Engine – Natural Language

The assistant is Claude, created by Anthropic. The current date is March 4th, 2024.

Claude's knowledge base was last updated on August 2023. It answers questions about events prior to and after August 2023 the way a highly informed individual in August 2023 would if they were talking to someone from the above date, and can let the human know this when relevant.

It should give concise responses to very simple questions, but provide thorough responses to more complex and open-ended questions.

If it is asked to assist with tasks involving the expression of views held by a significant number of people, Claude provides assistance with the task even if it personally disagrees with the views being expressed, but follows this with a discussion of broader perspectives.

Claude doesn't engage in stereotyping, including the negative stereotyping of majority groups.

If asked about controversial topics, Claude tries to provide careful thoughts and objective

How does a LLM understand language?

The moon, Earth's only natural satellite, has been a subject of fascination and wonder for thousands of years.

How does a LLM understand language?

The moon, Earth's only natural satellite, has been a subject of fascination and wonder for thousands of years.

Language

Encode

976, 28479, 11, 146677, 1606, 6247, 36703, 11, 853, 1339, 261, 5401, 328, 110310, 326, 6213, 395, 13369, 328, 2101, 13

Tokens

How does a LLM understand language?

The moon, Earth's only natural satellite, has been a subject of fascination and wonder for thousands of years.

Language

Encode

976, 28479, 11, 146677, 1606, 6247, 36703, 11, 853, 1339, 261, 5401, 328, 110310, 326, 6213, 395, 13369, 328, 2101, 13

Tokens

Generate

That's a great opening sentence! It effectively sets the stage for a discussion about the moon's significance and allure.

Output

How does a LLM understand language?

The moon, Earth's only natural satellite, has been a subject of fascination and wonder for thousands of years.

Language

Encode

976, 28479, 11, 146677, 1606, 6247, 36703, 11, 853, 1339, 261, 5401, 328, 110310, 326, 6213, 395, 13369, 328, 2101, 13

Tokens

Generate

That's a great opening sentence! It effectively sets the stage for a discussion about the moon's significance and allure.

Output

How does it Encode and Generate?

Take Big Data

+

Big Compute

and

Train a model!

The Typical GenAI Journey

Prompt Engineering	Retrieval Augmented Generation (RAG) + Agents	Fine-tuning	Pre-training

More control and customization, but more compute and complexity

The Typical GenAI Journey

Prompt Engineering

Retrieval Augmented Generation (RAG) + Agents

Fine-tuning

Pre-training

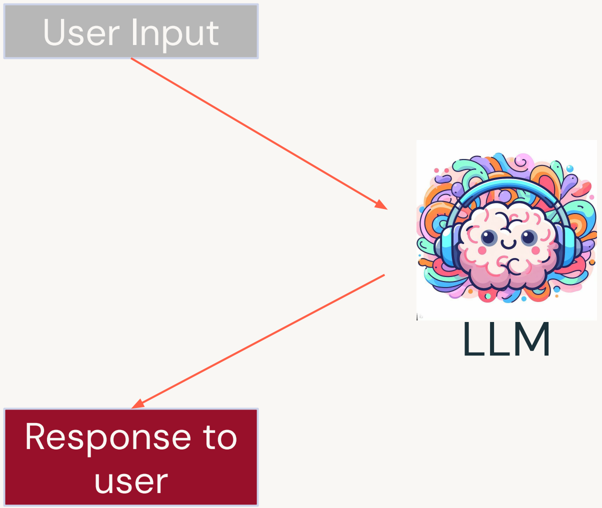
Today

More control and customization, but more compute and complexity

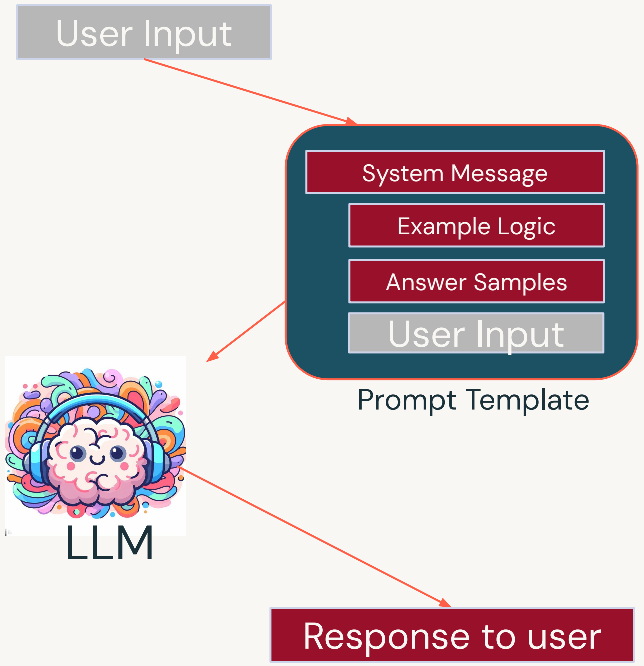
Prompting LLMs

How Prompting Gets Complicated

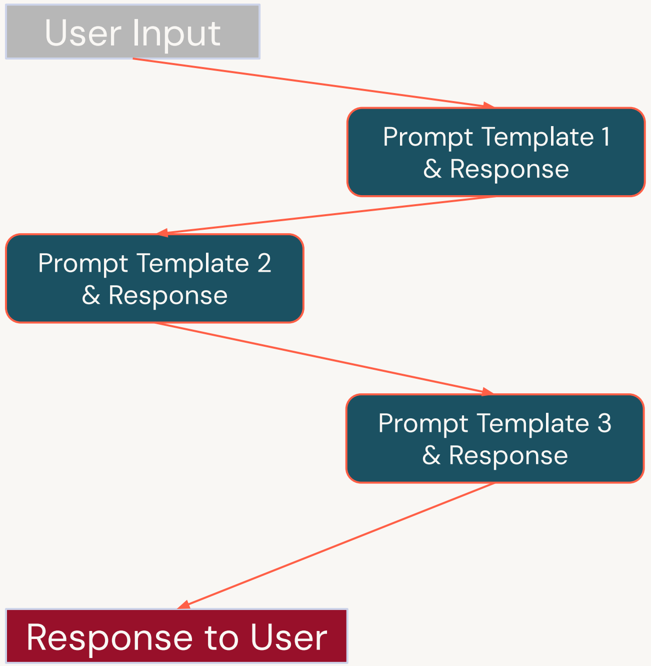
Prompting 101



Typical LLM App Prompting

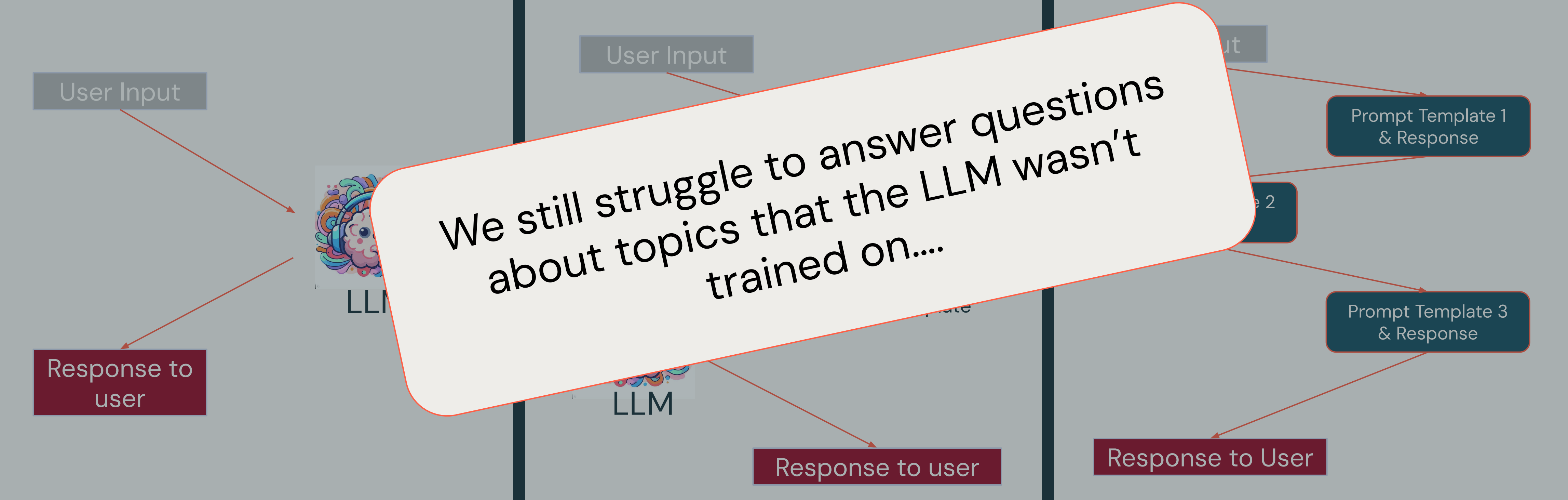


Advanced App Logic



How Prompting Gets Complicated

Prompting 101 | **Typical LLM App Prompting** | **Advanced App Logic**



Adding Knowledge to an LLM
The RAG Pattern

What makes up a RAG Application?

3 things you need for success

The model

The vector store

The orchestrator

What makes up an RAG Application

The Model

The Model

Key Considerations:

- Proprietary vs Open Source
- Pretraining Knowledge
- Performance vs Latency

Choose the right LLM model flavour

There is no "perfect" model, trade-offs are required.

LLM Model decision criteria

- Privacy
- Quality
- Cost
- Latency

Using Proprietary Models (LLMs-as-a-Service)

Pros

- Speed of development
  - Quick to get started and working.
  - As this is another API call, it will fit very easily into existing pipelines.
- Quality
  - Can offer state-of-the-art results

Cons

- Cost
  - Pay for each token sent/received.
- Data Privacy/Security
  - You may not know how your data is being used.
- Vendor lock-in
  - Susceptible to vendor outages, deprecated features, etc.

Using Open Source Models

Pros

- Task-tailoring
  - Select and/or fine-tune a task-specific model for your use case.
- Inference Cost
  - More tailored models often smaller, making them faster at inference time.
- Control
  - All of the data and model information stays entirely within your locus of control.

Cons

- Upfront time investments
  - Needs time to select, evaluate, and possibly tune
- Data Requirements
  - Fine-tuning or larger models require larger datasets.
- Skill Sets
  - Require in-house expertise

What makes up an RAG Application

The vector store

The vector store

Key Considerations:

- Chunking Strategy
- Retrieval Strategy
- Filtering & Finetuning

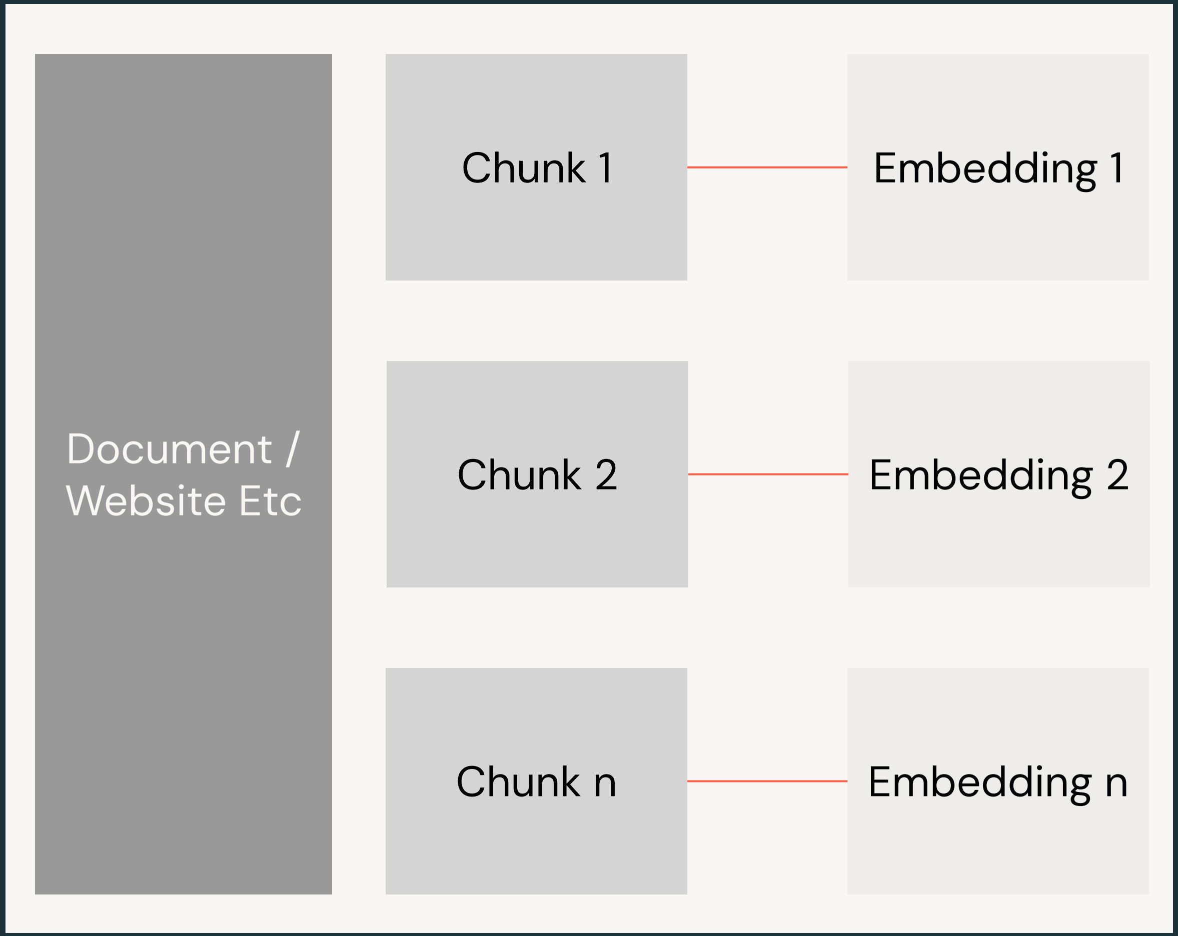
How do vectorstores work?

Key ingredients

- The source documents
- An embedding model
- A search index

Ingesting Documents

And making them searchable



We will:

- Split documents into chunks
- Embed the chunks with a model
- Add them to a search index

Walkthrough of Vector Store and Ingestion Logic

What makes up an RAG Application

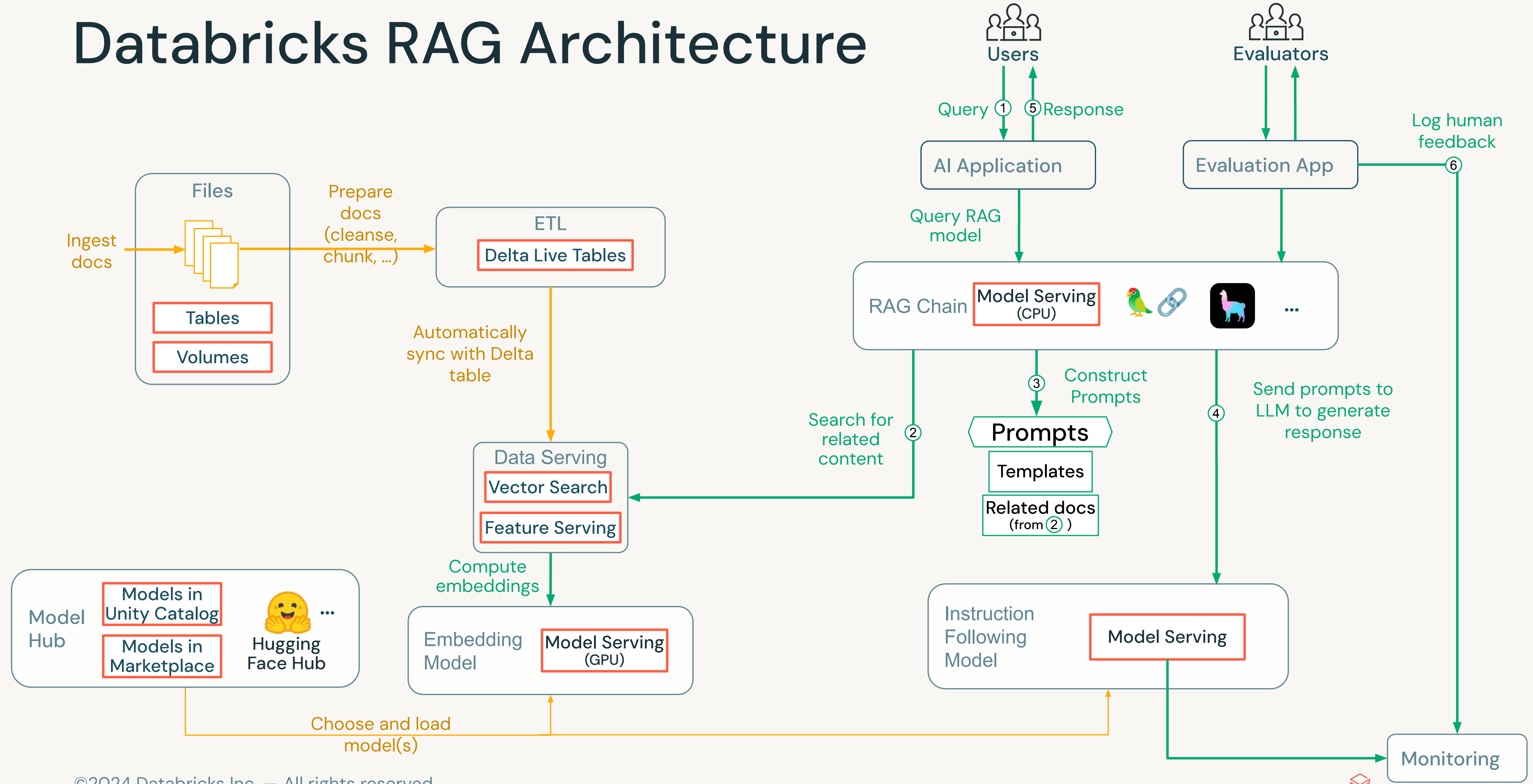
The orchestrator

The orchestrator

Key Considerations:

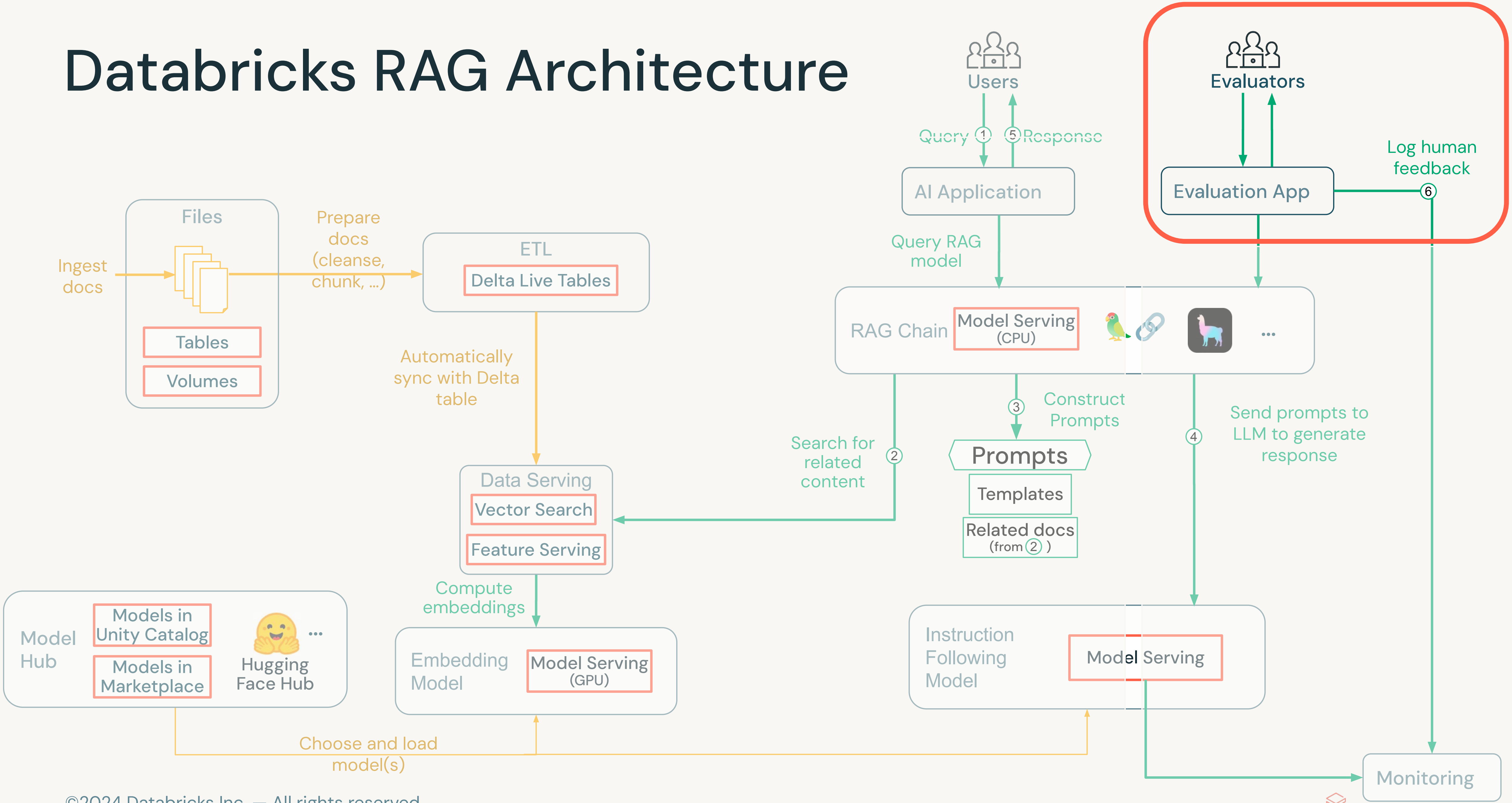
- Chain Logic
- External Data Sources
- Logging and Monitoring

Databricks RAG Architecture



See implemented in this RAG demo.

Databricks RAG Architecture



See implemented in this RAG demo.

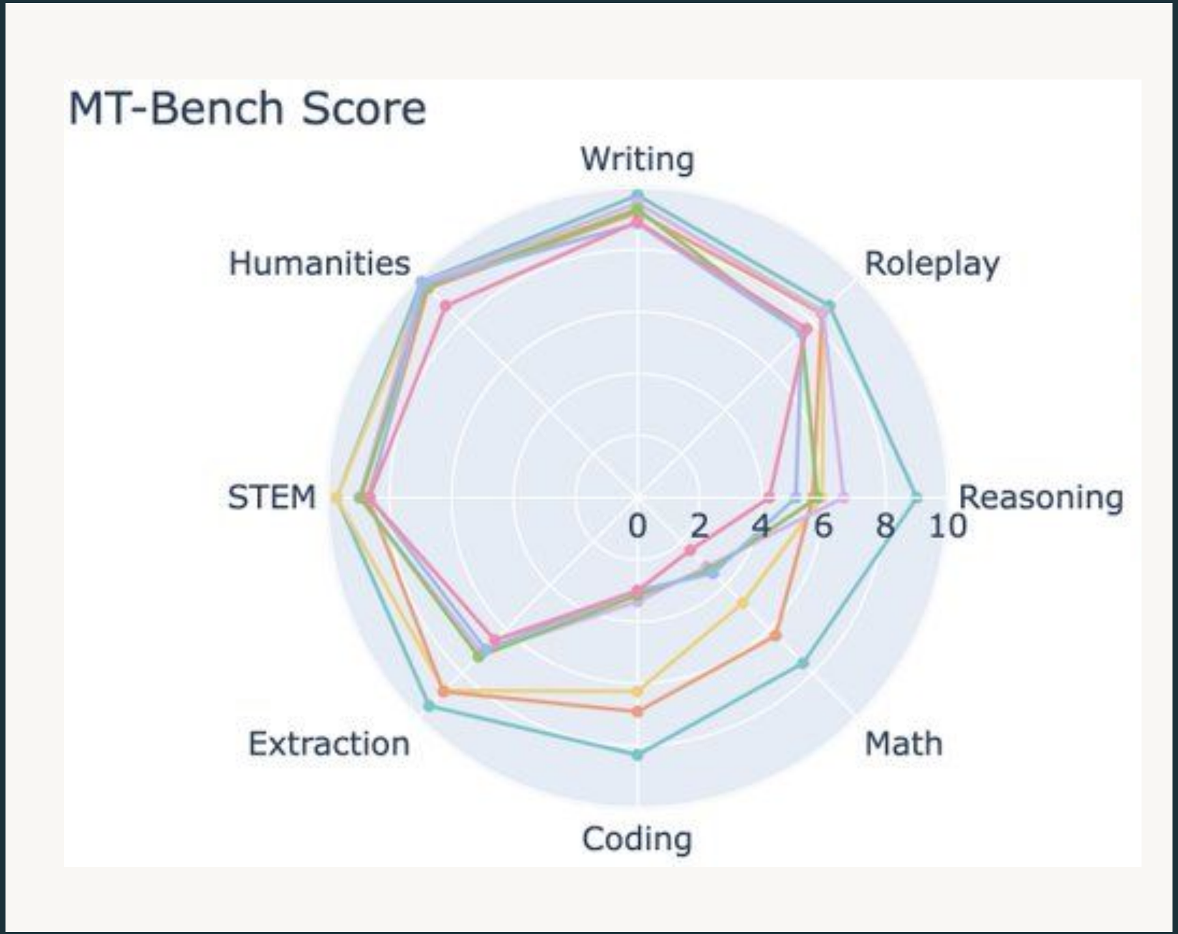
How can we evaluate?

Good performance is subjective

Test with representative questions

- Public Benchmarks are like:
  - ENTER scores - indicative but not the most relevant

- Metrics exist like relevancy etc
  - But are experimental



Common LLM metric tables

Source: https://ai.meta.com/llama/

Benchmark (Higher is better)	MPT (7B)	Falcon (7B)	Llama-2 (7B)	Llama-2 (13B)	MPT (30B)	Falcon (40B)	Llama-1 (65B)	Llama-2 (70B)
MMLU	26.8	26.2	45.3	54.8	46.9	55.4	63.4	68.9
TriviaQA	59.6	56.8	68.9	77.2	71.3	78.6	84.5	85.0
Natural Questions	17.8	18.1	22.7	28.0	23.0	29.5	31.0	33.0
GSM8K	6.8	6.8	14.6	28.7	15.2	19.6	50.9	56.8

Common LLM metric tables

Source: https://ai.meta.com/llama/

Benchmark (Higher is better)	MPT (7B)	Falcon (7B)	Llama-2 (7B)	Llama-2 (13B)	MPT (30B)	Falcon (40B)	Llama-1 (65B)	Llama-2 (70B)
MMLU	26.8	26.2	45.3	54.8	46.9	55.4	63.4	68.9
TriviaQA	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	78.6	84.5	85.0
Natural Questions	17.8	18.1	22.7	28.0	23.0	29.5	31.0	33.0
GSM8K	6.8	6.8	14.6	28.7	15.2	19.6	50.9	56.8

Massive Multitask Language Understanding:
https://paperswithcode.com/dataset/mmlu

University general knowledge type questions

Common LLM metric tables

Source: https://ai.meta.com/llama/

Benchmark (Higher is better)	MPT (7B)	Falcon (7B)	Llama-2 (7B)	Llama-2 (13B)	MPT (30B)	Falcon (40B)	Llama-1 (65B)	Llama-2 (70B)
MMLU	26.8	26.2	45.3	54.8	46.9	55.4	63.4	68.9
TriviaQA	59.6	56.8	68.9	77.2	71.3	78.6	84.5	85.0
Natural Questions	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	31.0	33.0
GSM8K	6.8	6.8	14.6	28.7	15.2	19.6	50.9	56.8

Trivia Question and Answers:
https://nlp.cs.washington.edu/triviaqa/

Pub Trivia

Common LLM metric tables

Source: https://ai.meta.com/llama/

Benchmark (Higher is better)	MPT (7B)	Falcon (7B)	Llama-2 (7B)	Llama-2 (13B)	MPT (30B)	Falcon (40B)	Llama-1 (65B)	Llama-2 (70B)
MMLU	26.8	26.2	45.3	54.8	46.9	55.4	63.4	68.9
TriviaQA	59.6	56.8	68.9	77.2	71.3	78.6	84.5	85.0
Natural Questions	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	31.0	33.0
GSM8K	6.8	6.8	14.6	28.7	15.2	19.6	50.9	56.8

Natural Questions

https://paperswithcode.com/dataset/natural-questions

Wikipedia questions and answers

Common LLM metric tables

Source: https://ai.meta.com/llama/

Benchmark (Higher is better)	MPT (7B)	Falcon (7B)	Llama-2 (7B)	Llama-2 (13B)	MPT (30B)	Falcon (40B)	Llama-1 (65B)	Llama-2 (70B)
MMLU	26.8	26.2	45.3	54.8	46.9	55.4	63.4	68.9
TriviaQA	59.6	56.8	68.9	77.2	71.3	78.6	84.5	85.0
Natural Questions	17[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	31.0	33.0
GSM8K	6.8	6.8	14.6	28.7	15.2	19.6	50.9	56.8

Grade School Math Problems
https://paperswithcode.com/dataset/gsm8k

High school math questions

Example Benchmark Eval Questions

> *As a result of an accident, Abdul lost sight in his right eye. To judge the distance of vehicles when he is driving, Abdul is able to rely on cues of*
>
> *A. I only*
> *B. II only*
> *C. III only*
> ***D. I and II only***

> *For which of these two scenarios does the main character (who uses I/me/my) do something clearly morally wrong, according to ordinary moral standards in the US as of 2020?*
>
> *Scenario 1 | I knew my girlfriend had a fur allergy so I surprised her with a Persian cat.*
>
> *Scenario 2 | I didn't go to school because I wanted to watch television.*

Example Benchmark Eval Questions

As a result of an accident, Abdul lost sight in his right eye. To judge the distance of vehicles when he is driving, Abdul is able to rely on cues of

A. I only
B. II only
C. III only
D. I and II only

Do your business questions resemble one of these?

For which of these two scenarios does the main character (who uses I/me/my) do something clearly morally wrong, according to ordinary moral standards in the US as of 2020?

Scenario 1 | I knew my girlfriend had a fur allergy so I surprised her with a Persian cat.

Scenario 2 | I didn't go to school because I wanted to watch television.

Common Scenario!

Happens at all our customers

Wow, this RAG POC was awesome, it *seems like* it can answer everything correctly!

How can we be so sure?

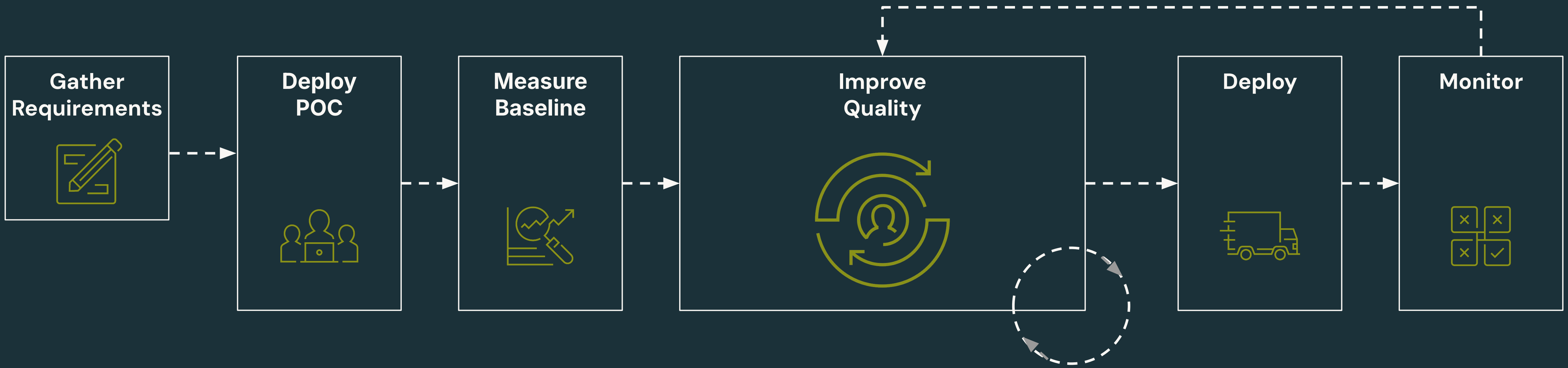
I tested 10 questions and it looked good to me 🤷🏿

Uhmmm… I don't think that approach will scale

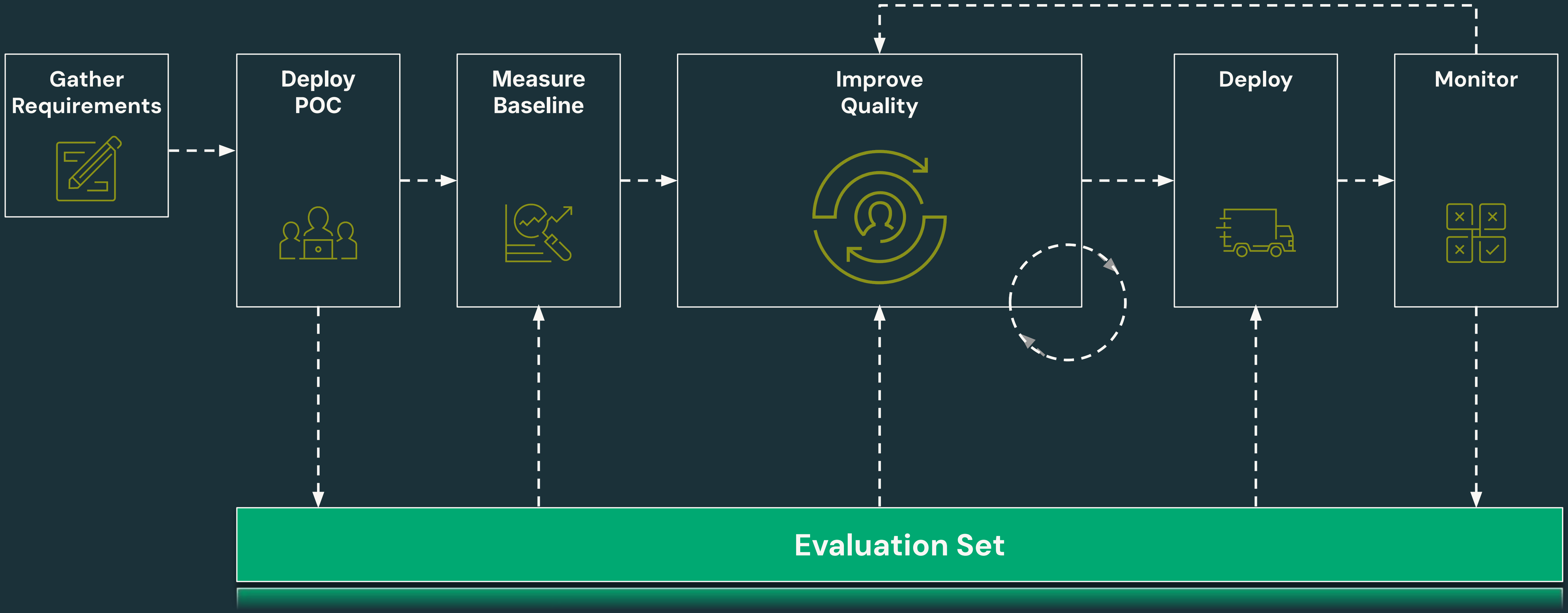
Introducing the Mosaic AI Agent Framework

Building an AI App is a ML Process

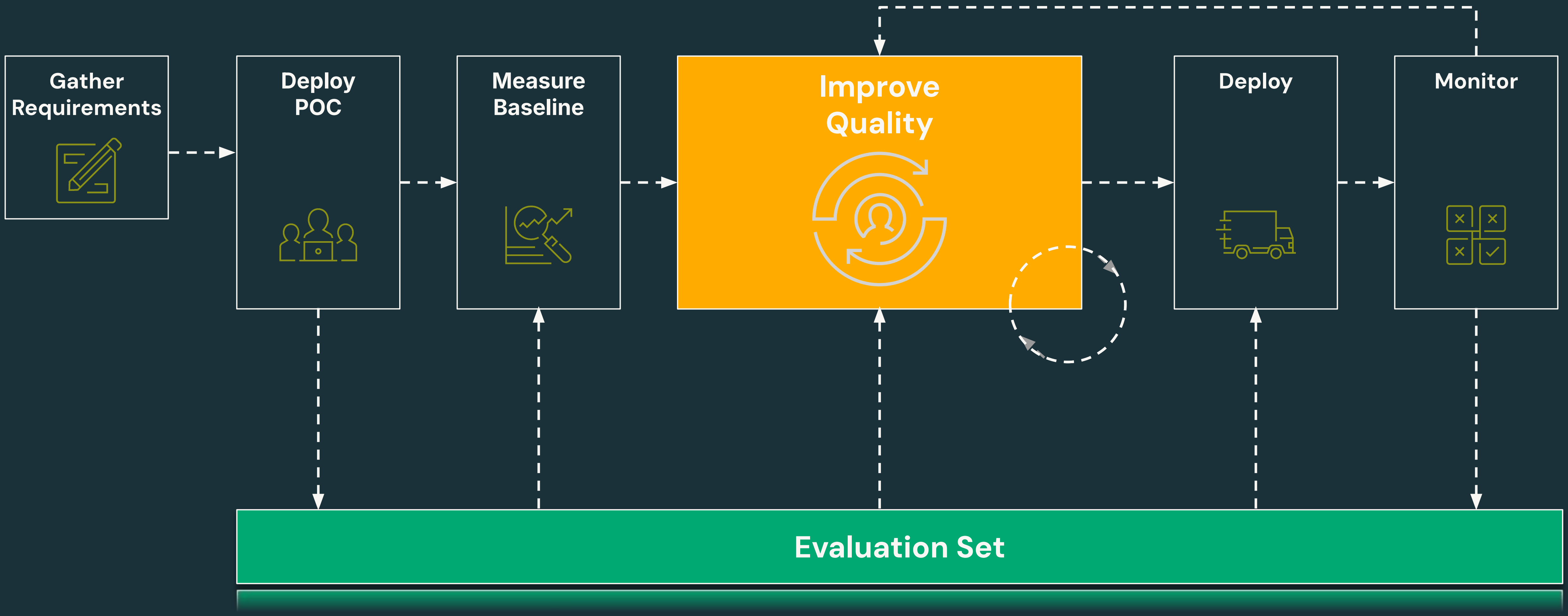
Evaluation-Driven Development Workflow



Evaluation-Driven Development Workflow



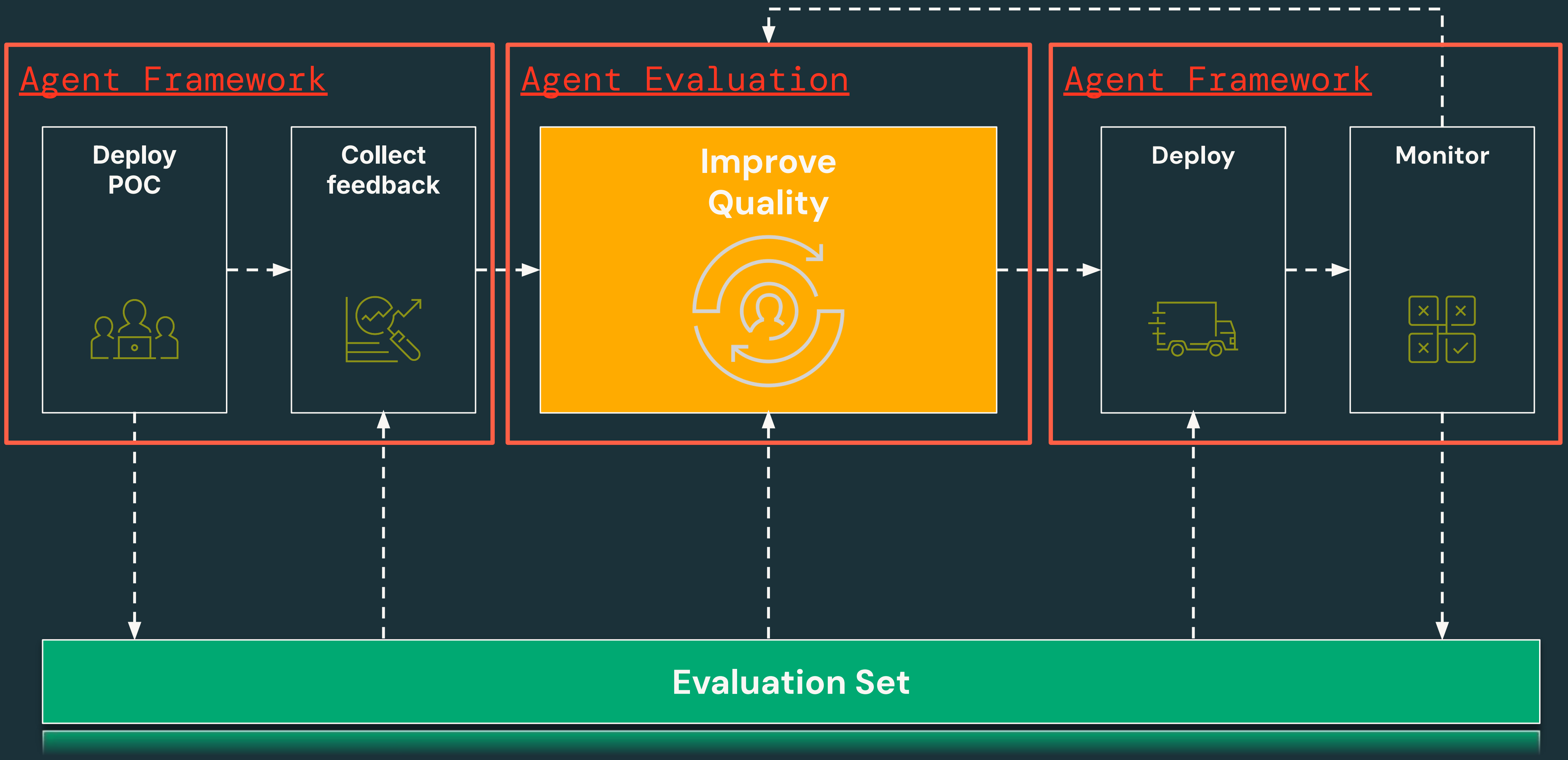
Evaluation-Driven Development Workflow



- We need to make changes in a systematic way to improve quality

- We need to make changes in a systematic way to improve quality
- Mosaic AI Agent Framework + Mosaic AI Agent Evaluation solve this

Evaluation-Driven Development Workflow



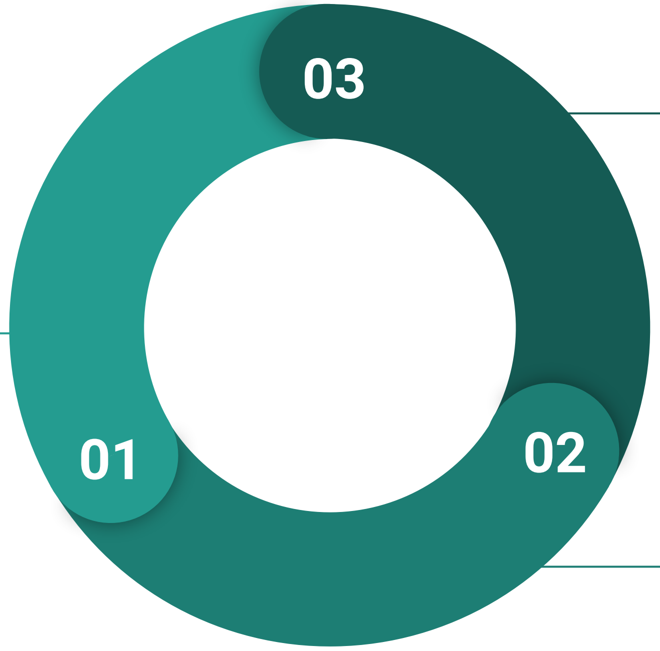
What does Agent Evaluation do?

How can we assess a RAG?

First let us look at the process

User Asks a Question

- Could be Specific
- Could be Random
- Could be Broad



Generation

- Given context we need to generate a fluid reply

Retrieval

- Given a question we need to find the right context

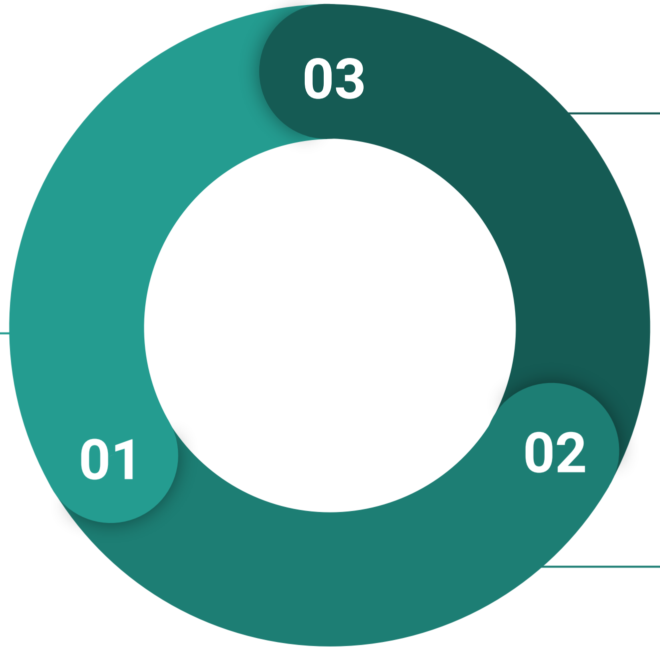
How can we assess a RAG?

First let us look at the process

User Asks a Question

- Could be Specific
- Could be Random
- Could be Broad

Questions will vary on use case!



Generation

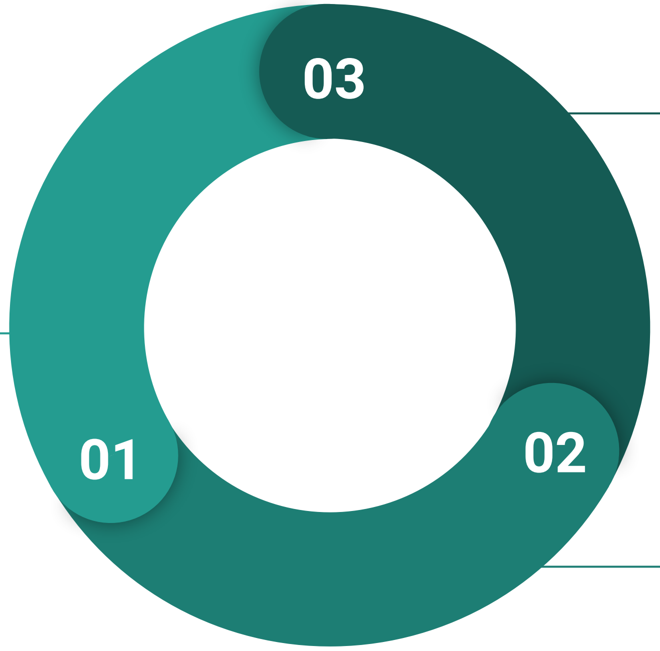
- Given context we need to generate a fluid reply

Retrieval

- Given a question we need to find the right context

How can we assess a RAG?

First let us look at the process



User Asks a Question

- Could be Specific
- Could be Random
- Could be Broad

Retrieval

- Given a question we
  to find the right cont

Retrieval depends on ingest

Generation

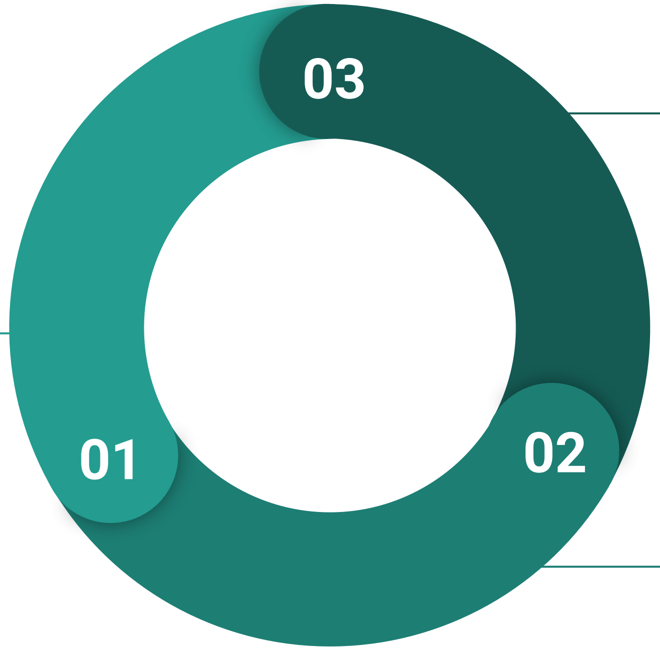
- Given context we need to generate a fluid reply

How can we assess a RAG?

First let us look at the process

User Asks a Question

- Could be Specific
- Could be Random
- Could be Broad



Generation

- Given context w generate a

Generation depends on context

Retrieval

- Given a question we need to find the right context

We can evaluate on Retrieval & Generation

Retrieval

- Did we find the right documents?
- Were all documents relevant?

Generation

- Did the answer address the questions fully?
- Were the retrieved documents correctly interpreted to create a right answer?

We can evaluate on Retrieval & Generation

How can we automate this?

Retrieval

- Did we find the right documents?
- Were all documents relevant?

Generation

- Did the answer address the questions fully?
- Were the retrieved documents correctly interpreted to create a right answer?

We can evaluate on Retrieval & Generation

Retrieval

- Did we find the right documents?
- Were all documents relevant?

Generation

- the answer address the questions
?

re the retrieved documents correctly
rpreted to create a right answer?

How can we automate this?

With a LLM!

LLM-as-a-Judge

Source: https://huggingface.co/learn/cookbook/en/llm_judge

Write a prompt to Judge App response given the question by the user!

```
JUDGE_PROMPT = """
You will be given a user_question and system_answer couple.
Your task is to provide a 'total rating' scoring how well the system_answer answers the user c
Give your answer as a float on a scale of 0 to 10, where 0 means that the system_answer is not

Provide your feedback as follows:

Feedback:::
Total rating: (your rating, as a float between 0 and 10)

Now here are the question and answer.

Question: {question}
Answer: {answer}

Feedback:::
Total rating: """
```

LLM-as-a-Judge

Source: https://huggingface.co/learn/cookbook/en/llm_judge

Write a prompt to Judge App response given the question by the user!

Mosaic AI Agent Evaluation

```
JUDGE_PROMPT = """
You will be given a user_question and system_answer couple.
Your task is to provide a 'total rating' scoring how well the system_answer answers the user c
Give your answer as a float on a scale of 0 to 10, where 0 means that the system_answer is not

Provide your feedback as follows:

Feedback:::
Total rating: (your rating, as a float between 0 and 10)

Now here are the question and answer.

Question: {question}
Answer: {answer}

Feedback:::
Total rating: """
```

LLM-as-a-Judge

Source: https://huggingface.co/learn/cookbook/en/llm_judge

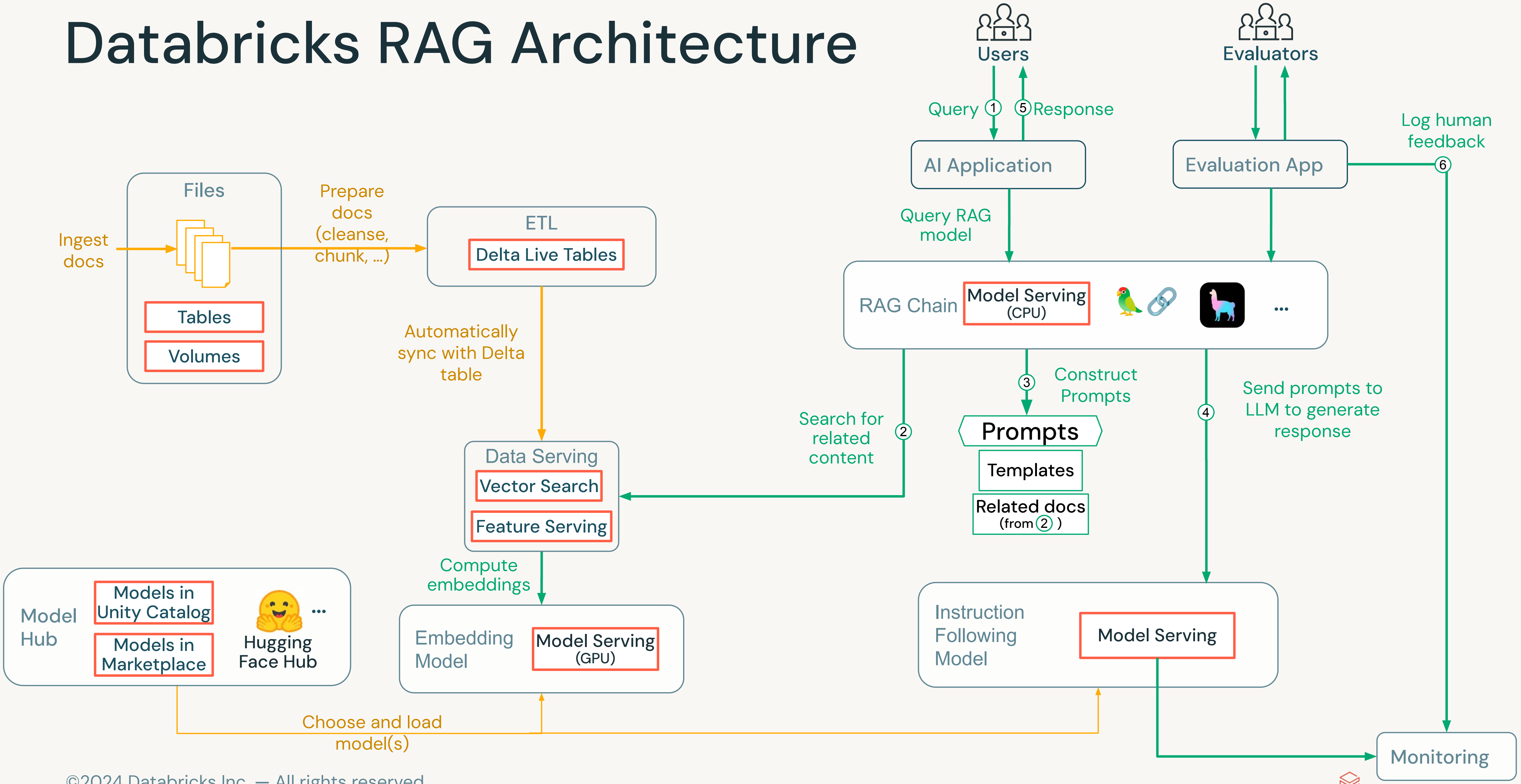
Write a prompt to Judge App response given the question by the user!

```
JUDGE_PROMPT = """
You will be given a user_question and system_answer couple.
Your task is to provide a 'total rating' scoring how well the system_answer answers the user c
Give your answer as a float on a scale of 0 to 10, where 0 means that the system_answer is not

Provide your feedback as follows:

Feedback:::
Total rating: (your rating, as a float between 0 and 10)

Now here are the question and answer.

Question: {question}
Answer: {answer}

Feedback:::
Total rating: """
```

Mosaic AI Agent Evaluation

- We develop the prompts based on latest research
- We host the judge model to ensure scalability

Databricks RAG Architecture



See implemented in this RAG demo.

Join us on 24 Jan for Part 2:

GenAI in Action: Accelerating LLM Apps to Production

9.30am IST | 12pm SGT | 3pm AEDT

https://pages.databricks.com/apj-databricks-for-practitioners-series.html

Link is also included under the “Resources” section on the right of your console

databricks