

Introduction to Retrieval Augmented Generation (**RAG**) and Implementing with Databricks

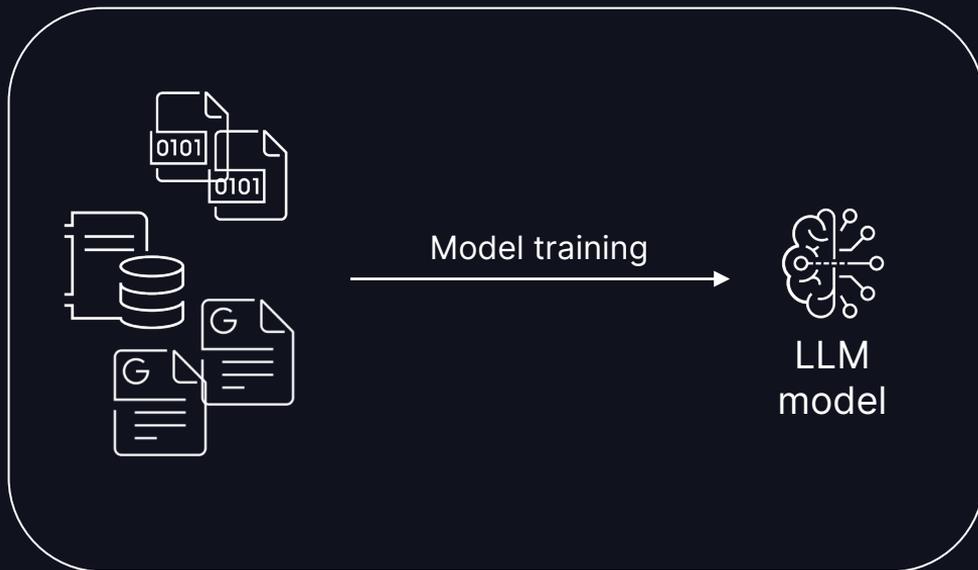
Ina Koleva, Product @ Databricks
June 2024



Agenda

1. Working with LLMs: The **good** and the **bad**
2. Talk to *my* data
3. RAG
4. When to use RAG, when **not** to?
5. **Best practices** for RAG
6. **DEMO**

How do LLMs work?



- **Gives a direct answer**
The user does not need to aggregate multiple search results
- **Great at text generation**
The model produces text with the correct language structure
- **Knowledge from training data only**
Only knows of facts contained in the data. May not be up to date.

LLM responses may sound *believable*,
but they are *not always factually correct*.



What comes after “thrice”?

After “thrice” the next term used to indicate numerical repetition is “four times” or “quadrice”.

The sequence of repetition terms is

- Once
- Twice
- Thrice
- Four times (or **Quadrice**)
- Five times (or **Quintrice**)
- Six times (or **Sextrice**)
- Seven times (or **Septrice**)
- Eight times (or **Octrice**)



Are you sure? Oxford dictionary says you're wrong

You're correct, and I apologise for the confusion in my previous response. The terms like “quadrice”, “quintrice”, and others I mentioned do not exist in standard English usage.





How many years ago was Roland Garros founded?

Roland Garros was founded **133 years ago**, in 1891.



Roland Garros, the stadium that hosts the French Open tennis tournament, was founded in 1928. This means that Roland Garros was founded **96 years ago**, in 1928.



Roland Garros, also known as the French Open, was founded in 1891. Therefore, as of 2023, Roland Garros was founded **132 years ago**.



Roland Garros, the premier clay court tennis championship, was founded **131 years ago**. It is named after the French aviator Roland Garros, who was a pioneer in aviation. The tournament takes place annually in Paris, France, and is one of the four Grand Slam tournaments in tennis.



Screenshots taken on 3rd June 2024



Searched 4 sites

The winner of the 2024 Roland Garros tournament is Aryna Sabalenka, defeating Aryna Sabalenka in the final. This win marks another significant milestone for Sabalenka, as she becomes the first player to win the French Open in women's tennis.



Searched 5 sites

The final match of the 2024 Roland Garros tennis tournament was held on June 8, 2024. The match was between Aryna Sabalenka and Aryna Sabalenka. Sabalenka won the match in three sets, 6-3, 6-3, 6-3.

The French Open

May 26–Jun 9, 2024

MEN'S SINGLES WOMEN'S SINGLES MEN'S DOUBLES WOMEN'S DOUBLES MIXED DOUBLES

	Thu 30 May	Fri 31 May	Sat 1 Jun	Sun 2 Jun	Today 3 Jun	Tue 4 Jun	Wed 5 Jun	Thu 6 Jun	Sat 8 Jun
Fourth round · Philippe-Chatrier 22 E. Navarro 2 A. Sabalenka				Final > 2 3 6 6 <	Fourth round · Court Suzanne Lenglen E. Avanesyan 12 J. Paolini			Final > 6 0 1 4 6 6 <	
Video highlights · 1 hour ago 					Video highlights · 1 hour ago 				
Fourth round · Philippe-Chatrier 15 E. Svitolina 4 E. Rybakina				Final > 4 3 6 6 <	Fourth round · Court Suzanne Lenglen 16:30 * V. Gracheva M. Andreeva				
Video highlights · 3 hours ago 					Video highlights · 3 hours ago 				

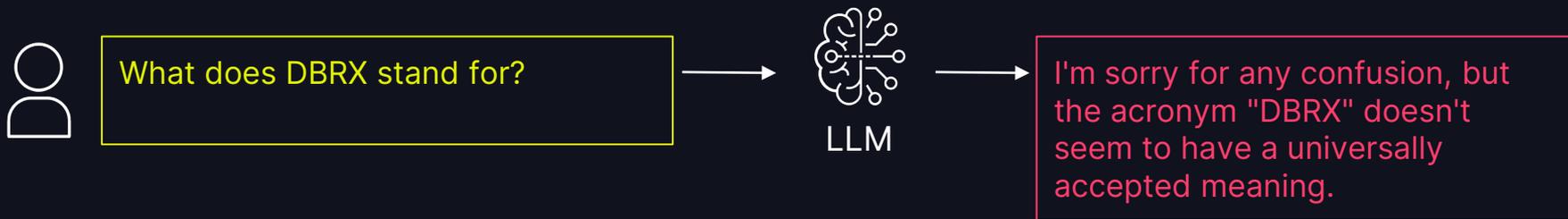
All times are in Central European Time Feedback

* Estimated time



Challenges

- **May not be correct**
LLMs can lie very confidently
- **No source**
The model does not provide referenceable materials
- **Not up to date**
The model gave a response which is no longer valid
- **Not based on my data**
The LLM wasn't trained on my enterprise dataset



Talk to *my* data: Use Cases



Internal Q&A bot

Must answer question on the basis of confidential internal data.



Customer Support Bot

Must have context about the Customer to provide personalised support.



Content/Email generation

Marketing content needs to be based on the Enterprise product offerings.



Summarisation

Must have access to the internal knowledgebase (meeting notes, emails, Customer 360, etc.)



Talk to your data

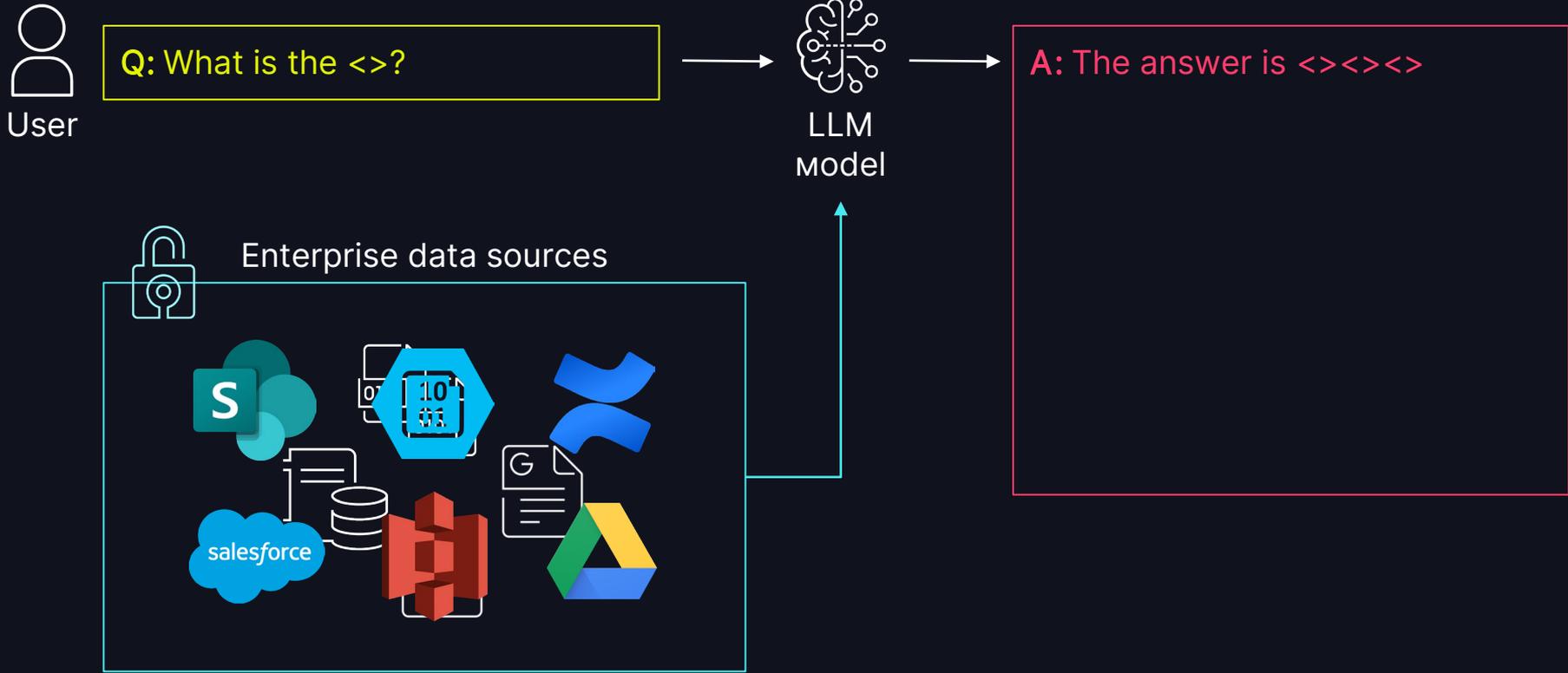
Discover aggregative trends, find information, summarise key insights, output reports, etc.



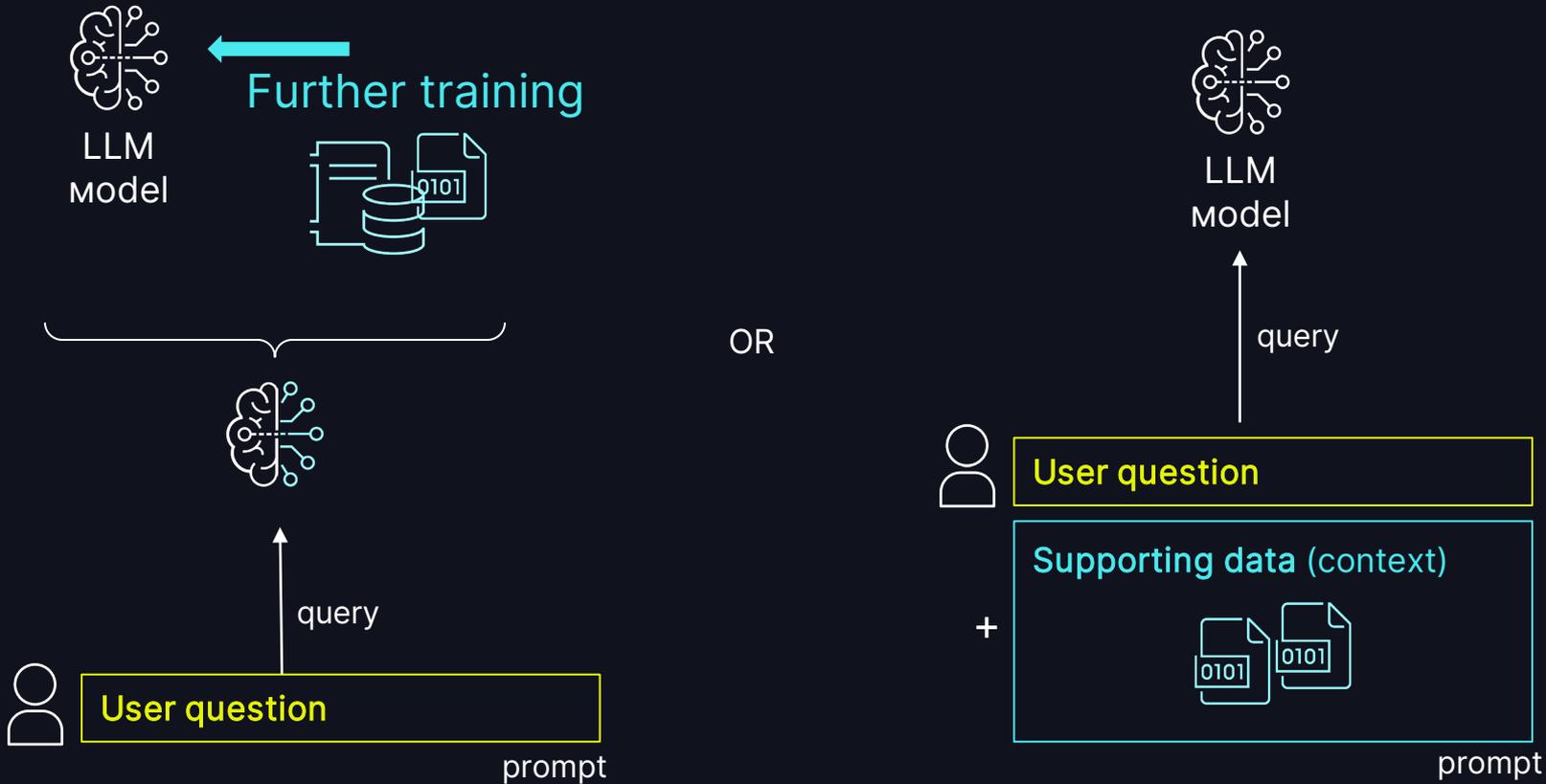
Educational tools

Personalised ed-experience should have context on the learner's progress, available resources, and be up to date.

Talk to *my* data: Methods



Talk to *my* data: Methods



The following is an excerpt from an article on the Eiffel Tower:

"The Eiffel Tower is a wrought-iron lattice tower on the Champ de Mars in Paris, France. It is named after the engineer Gustave Eiffel, whose company designed and built the tower. Locally nicknamed "La dame de fer" (French for "Iron Lady"), it was constructed from 1887 to 1889 as the centerpiece of the 1889 World's Fair.

The tower is 324 meters (1,063 ft) tall, about the same height as an 81-story building, and the tallest man-made structure in Paris. Its base is square, measuring 125 meters (410 ft) on each side. During its construction, the Eiffel Tower surpassed the Washington Monument to become the tallest man-made structure in the world, a title it held for 41 years until the Chrysler Building in New York City was finished in 1930."

Given the context information above, please answer the following question:

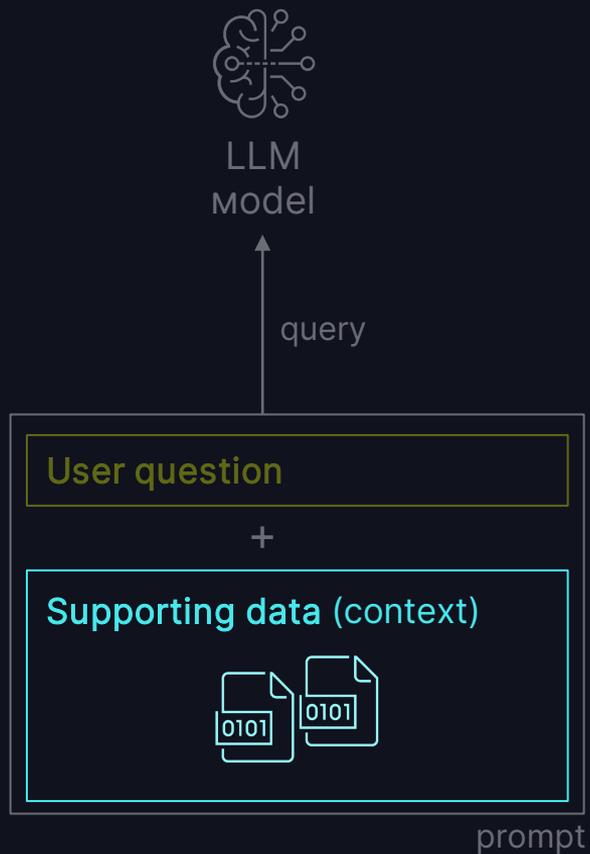
"What was the purpose of constructing the Eiffel Tower, and how long did it take to build?"

prompt

query



LLM
model



Needs to be **real-time**

I need to construct the prompt in real time since it's based on the user's question.

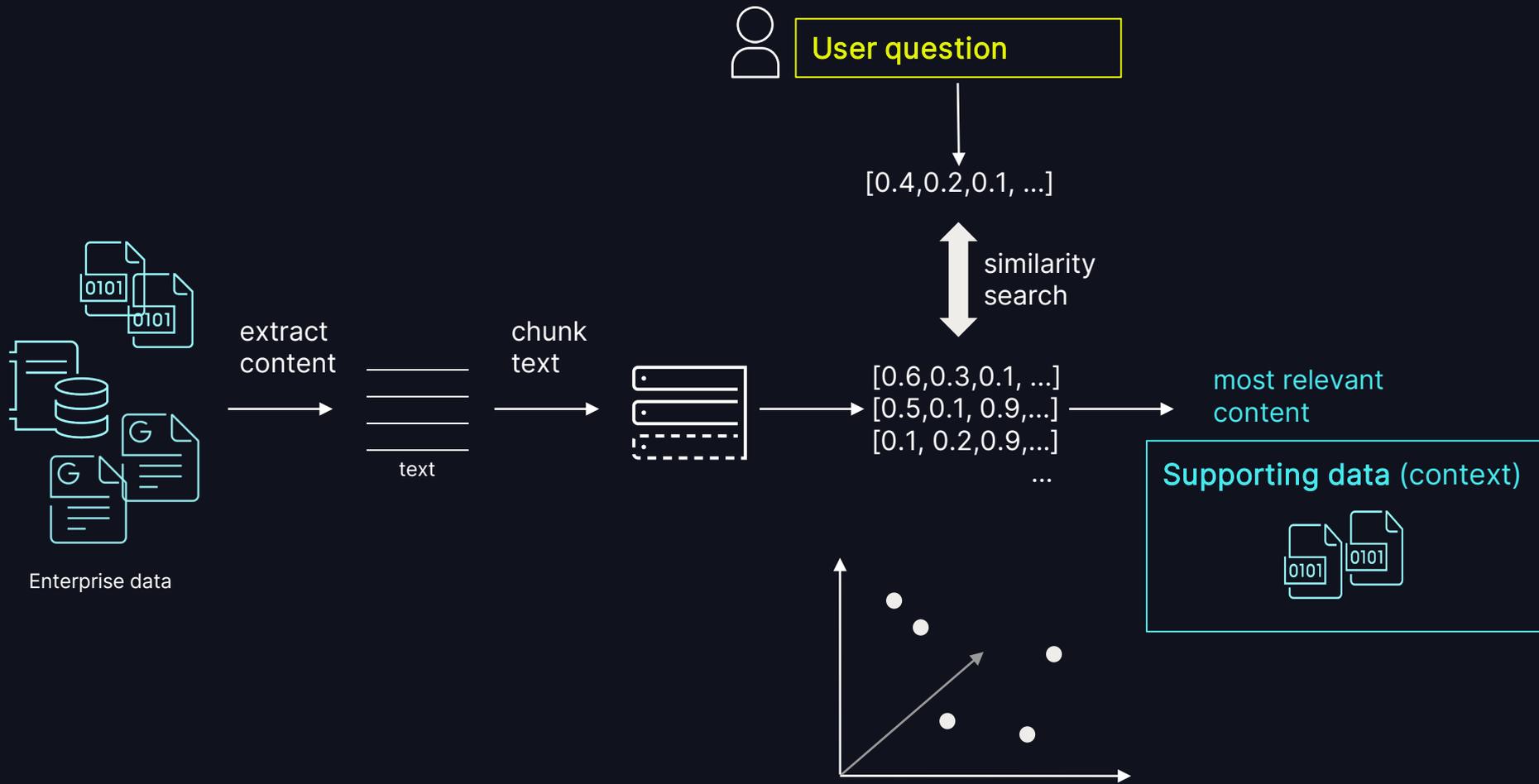
Needs to be **relevant content**

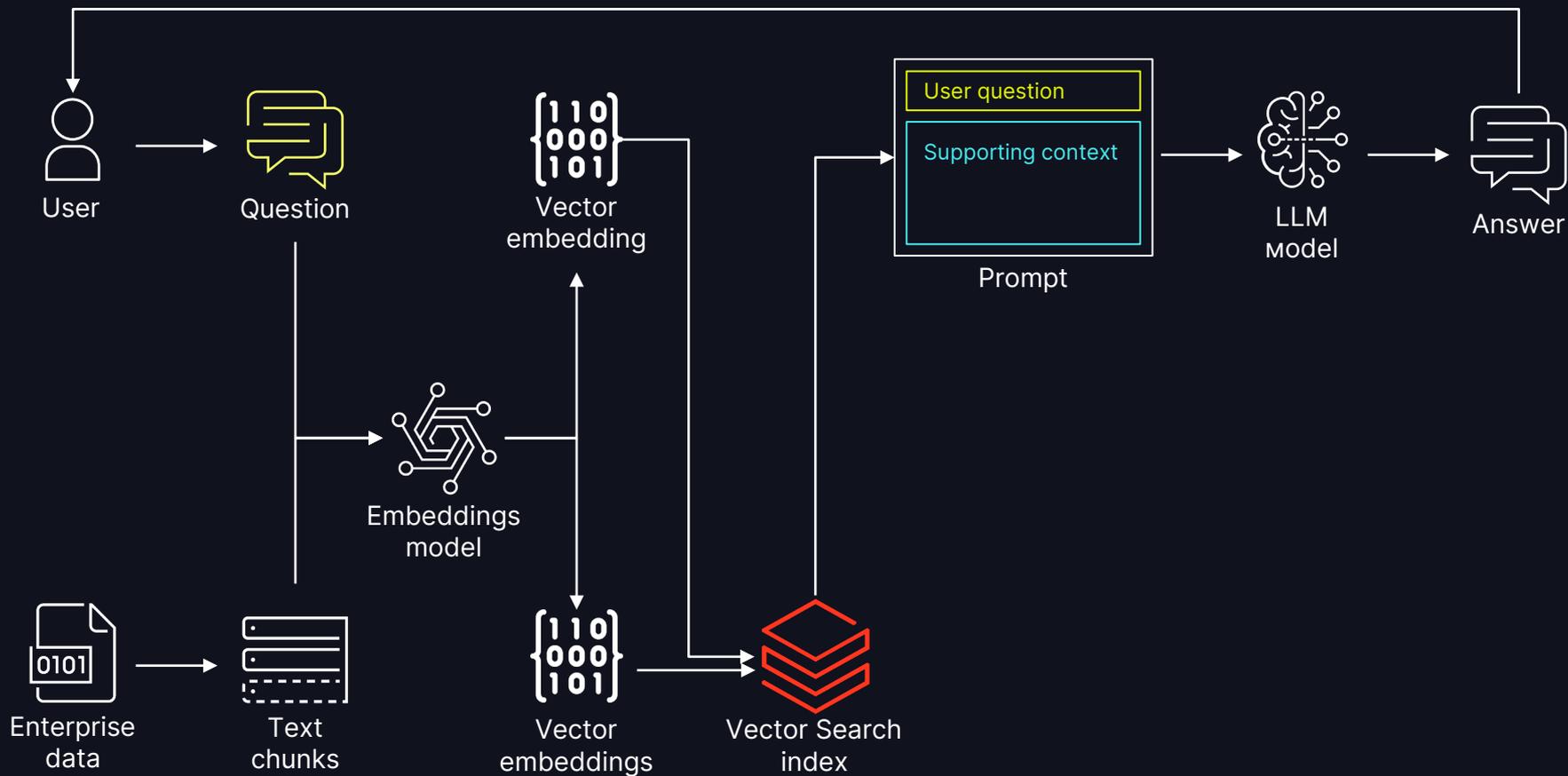
The context must contain the information that is needed to answer the user's question.

Extracted from **enterprise data**

The content needs to come from the organisation's knowledge base.







RAG

=

Retrieval Augmented Generation



Challenge

Customer Support operations had challenges bringing new agents up to speed amidst typical agent churn

Solution

Built an agent-assist experience using  Databricks.

Used their content from product manuals, YouTube videos, and support cases.

Deployed a RAG solution to help agents find answers to customer queries.

Impact

Enhance efficiencies

Elevate customer support experiences and time to get agents up-to-speed



What does DBRX stand for?

without RAG



I'm sorry for any confusion, but the acronym "DBRX" **doesn't seem to have a universally accepted meaning.**

Acronyms can stand for different things in different contexts, and without a specific context, it's challenging to provide an accurate meaning.

If "DBRX" is related to a specific field or industry, providing that context would help in determining the correct meaning.

with RAG

DBRX stands for Databricks.

DBRX is the name of the fine-grained sparse mixture-of-experts (MoE) model which was open-sourced by Databricks in March 2024.

The model is a transformer-based decoder-only large language model (LLM).



Advantages of RAG



Allows you to leverage Enterprise data
RAG is a compound System.



Groundedness
Less likely to hallucinate



No model modifications
It is a simple and computationally light approach



Freshness of data
Can continuously incorporate new up-to-date data in the bot

When **not** to do RAG

Ask yourself...

Do you have the **datasets**,
and are they good quality?

yes

Amazing. Next step is to ensure the datasets are securely governed, and accessible for downstream usage.

no

Consider data marketplaces, like Databricks Marketplace
Curate a high-quality collection of datasets by removing incorrect content and performing ETL to clean the data

Does the **user query** change
the whole process?

yes

The real-time input from the user is what drives value proposition for this use case. RAG is recommended!
E.g. Q&A bots, Customer Support

no

You have a use case, which might need to leverage enterprise data, but does not necessarily require user's input
E.g. **Meeting notes summarisation, Offline operations**

When **not** to do RAG

Ask yourself...

Do you need to perform a specific task?
(summarisation, code completion...)

yes

You want to perform a *specific* task which leverages enterprise data

- 1 - Prompt Engineering
- 2 - Create a RAG app (+ Prompt)
- 2 - Fine-tune the model for the task (+ RAG)
- 3 - Train a fully custom model (+ fine-tune for task, + RAG)

Do you have **latency** and **cost** constraints?

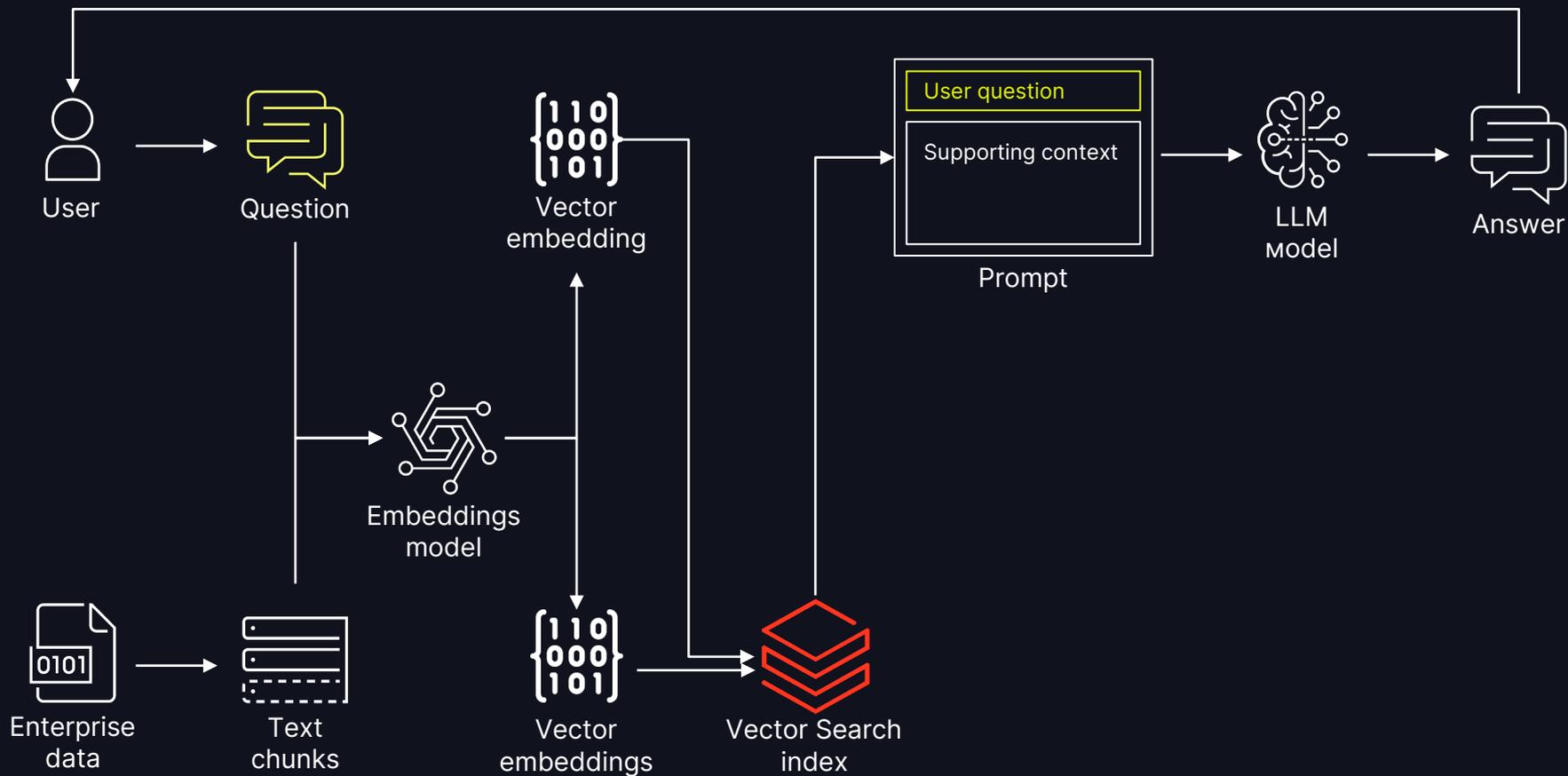
yes

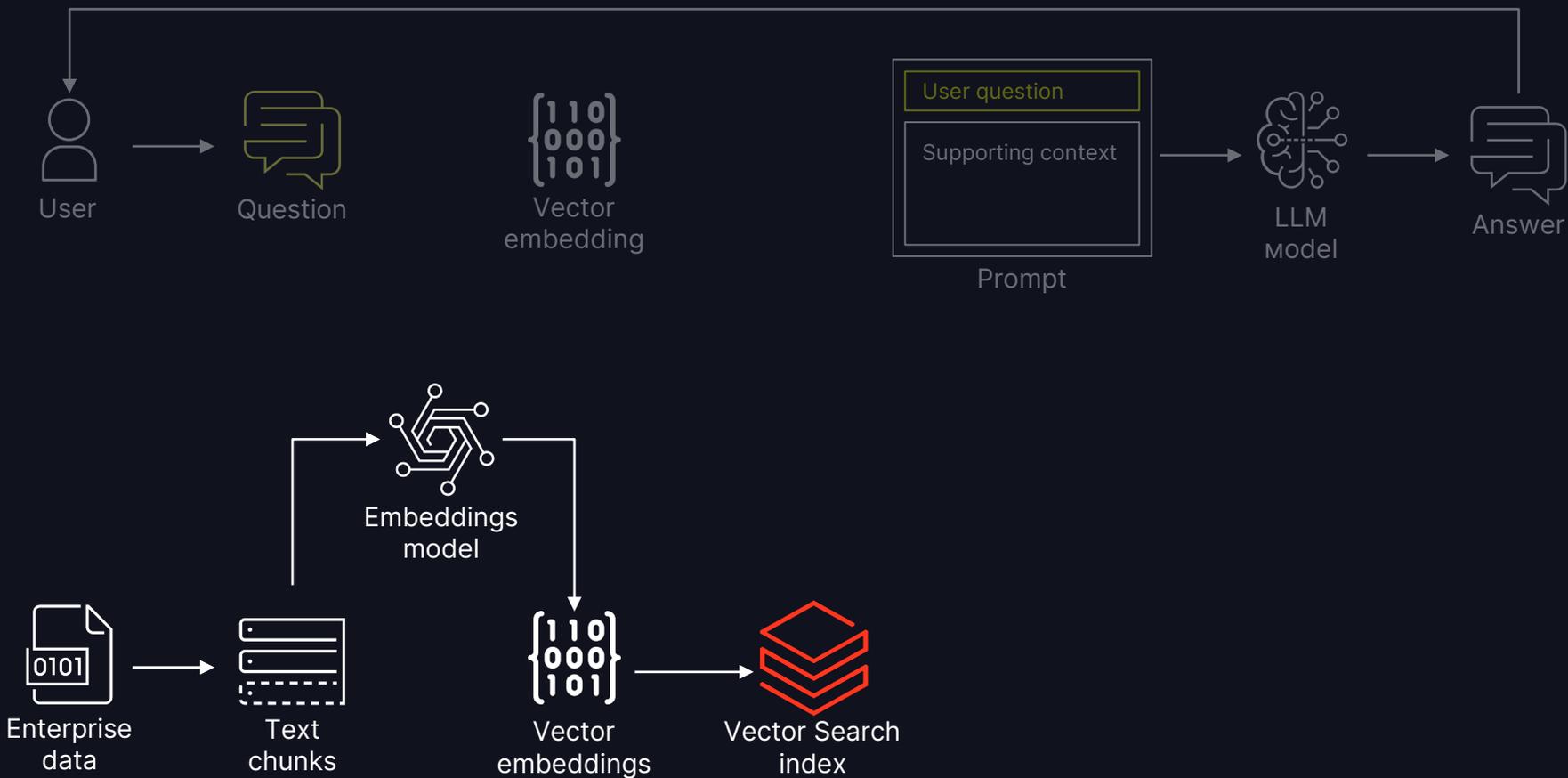
RAG can drive token count up, and thus costs.

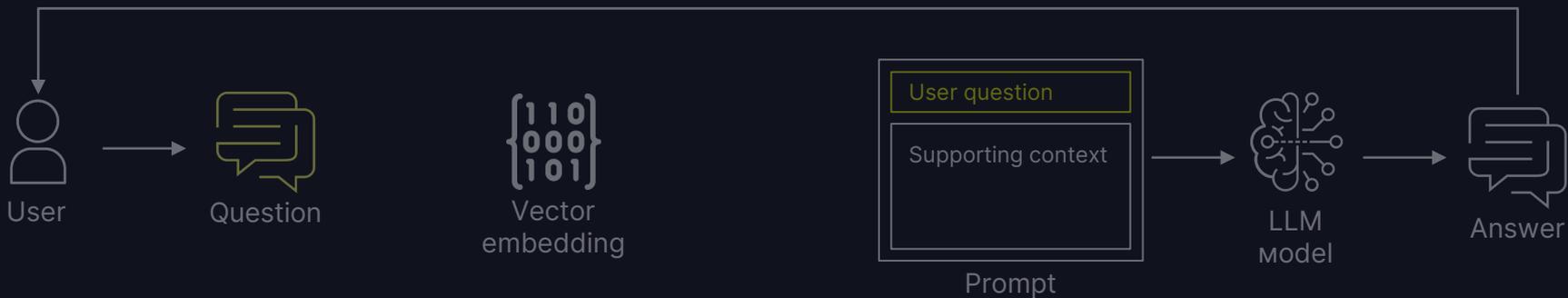
- 1 - Weigh those costs against how useful it is for the quality of the answer

RAG can drive latency up, mainly retrieval and generation on larger context length

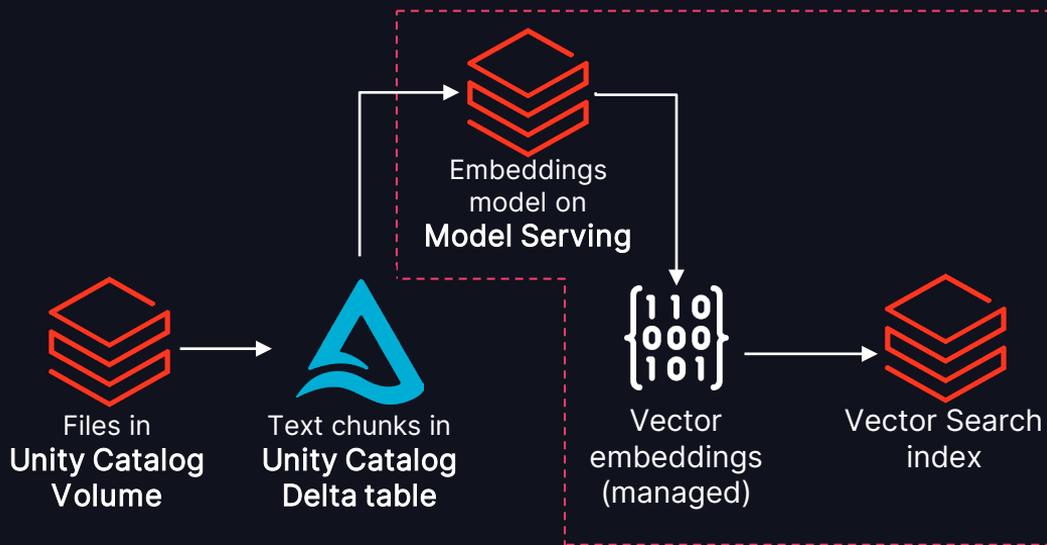
- 1 - Evaluate how long each chain step take
- 2 - Consider fine-tuning the model







Mosaic AI Vector Search



DEMO

1/ ENTERPRISE DATA use in VECTOR SEARCH

Embed video of the demo as backup

Best Practices - Chunking strategy

You should consider...



or



What are your data sources?

Short-form

Long-form

Tweets, instant messages,
papers,
Short articles, FAQs
manuals



or

Books, Research



Lengthy articles,

*What questions are users
going to ask?*

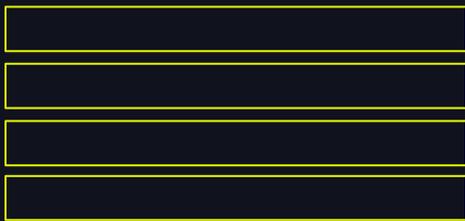
**Precise Questions
Questions**

Factual queries, information
retrieval, precise questions

Complex Open

Open-ended questions,
need to synthesize

Best Practices - Chunking size



or



Small chunks

(e.g. 128-256 tokens)

- **Precise and Factual Queries:** Small chunks increase the likelihood of retrieving sentences that directly address the query.
- **Short-Form Content:** tweets, instant messages, or short articles
- **Limited Context Dependency:** If the queries do not require extensive context, smaller chunks can suffice

Big chunks

(e.g. 512-1024 tokens)

- **Complex and Open-Ended Queries:** That requires synthesizing information from multiple sources and need the necessary context
- **Long-Form Content:** larger chunk sizes can better preserve the semantic coherence
- **High Context Dependency:** If the queries require understanding long-range dependencies or extensive context

Best Practices - Keyword or Semantic



*Best Italian restaurants
in San Francisco*

Keyword search

*"Best Pizza and Italian Restaurants in the Bay Area"
"The best Italian restaurant is in New York, not San Francisco"*

Semantic search

*"best pizza places in San Francisco"
"top-rated Italian eateries in the San Francisco area"*

Best Practices - Keyword or Semantic

Keyword search

Matches queries with documents containing the exact keywords or phrases

○ **Queries include the entities** : product names, technical terms, i.e. phrases that need to be matched exactly.

○ **Queries requiring verbatim matches**: In cases where the exact phrasing/wording is *crucial* (e.g. legal documents or code snippets)

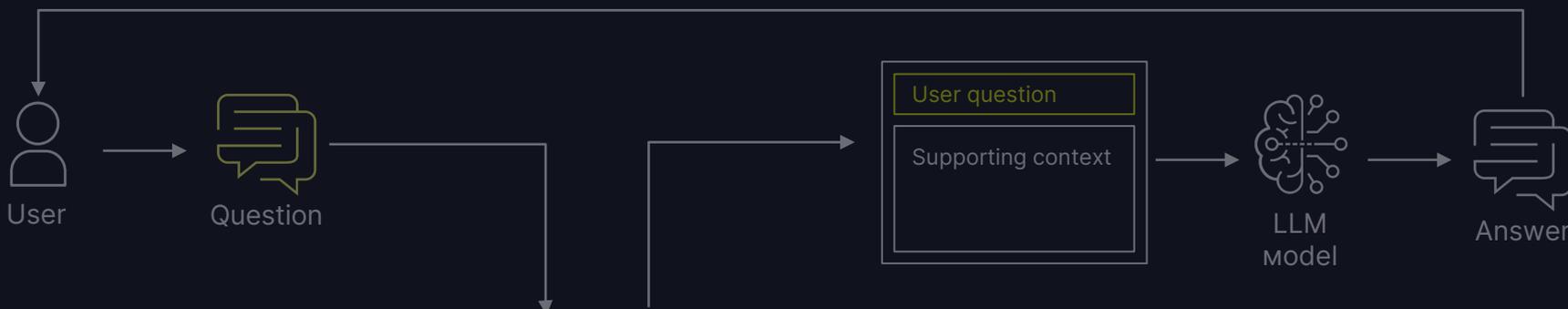
Semantic search

Understand the meaning of queries to return results, even if the documents don't contain the exact keywords

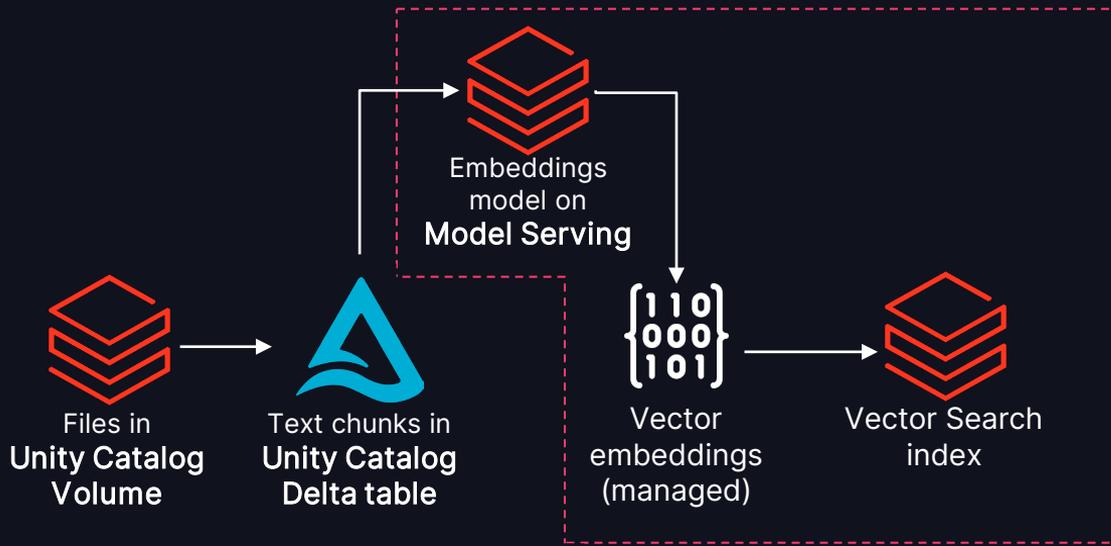
○ **Open-ended or conversational queries**: users are likely to express their information needs in natural language

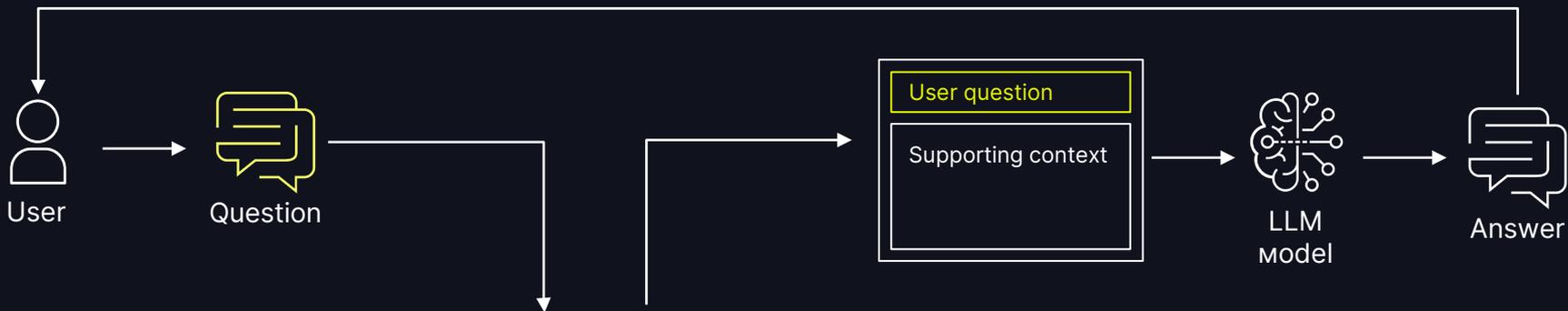
○ **Queries in specialized domains**: semantic search can leverage domain-specific knowledge to understand the context

○ **Multilingual queries**: Semantic search can be particularly useful for handling queries in multiple languages

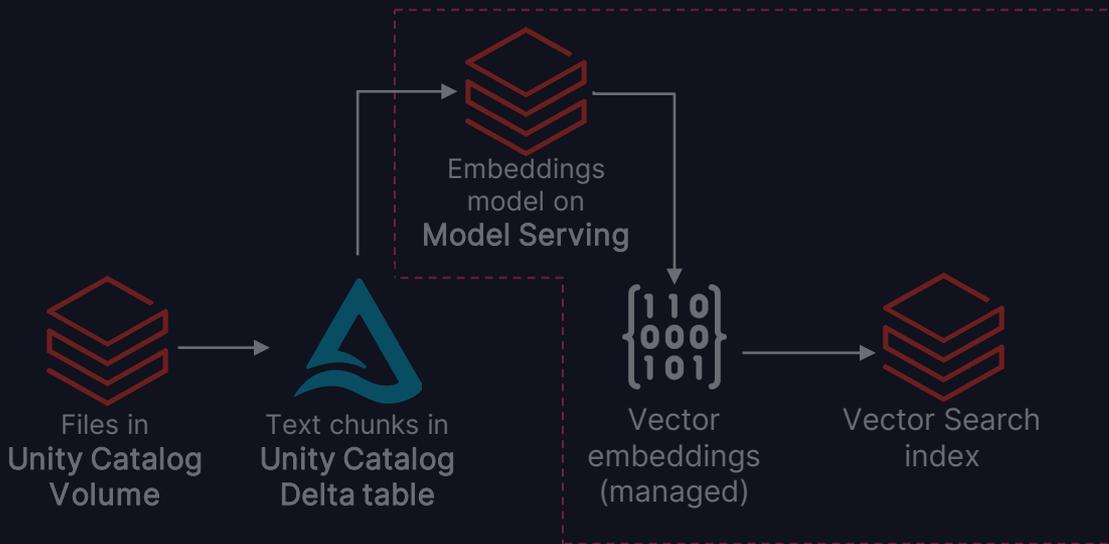


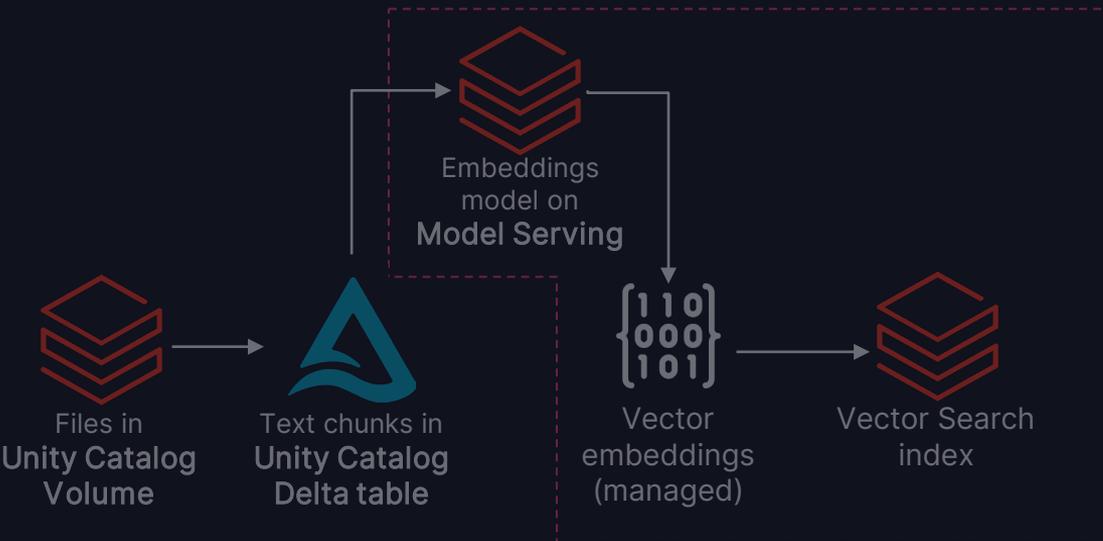
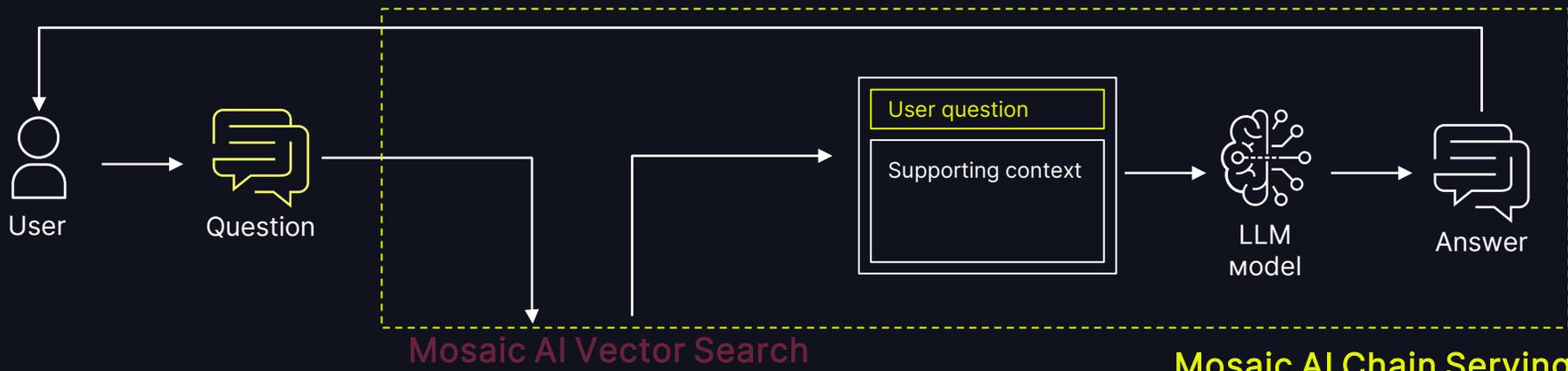
Mosaic AI Vector Search





Mosaic AI Vector Search





Mosaic AI Chain Serving

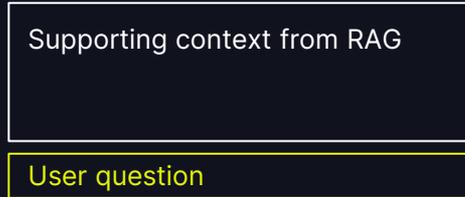


DEMO

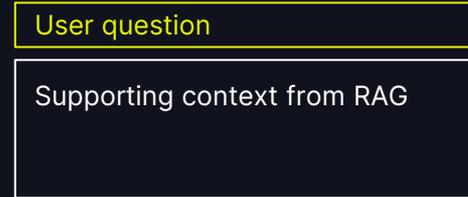
2/ use CHAIN SERVING
for app interaction

Embed video of the demo as backup

Best Practices - Prompt Structure



or



^ Better!

Giving the context first helps the LLM better comprehend the question and formulate a relevant response based on the provided context

^ May confuse the LLM

If the question comes first without any context, the LLM may struggle to understand the intent or background, potentially leading to irrelevant or incoherent outputs

Best Practices - Prompt Templates



Why is this order optimal?

- **System prompt** first - ensures the model understands the task and the requirements.
- **Supporting context from RAG** immediately after the system prompt allows the model to ground its understanding.
- **Few-shot examples** after the context helps the model learn the desired response structure and style, informed by the provided context.
- Keeping the **user query at the end** focuses the model's attention on generating a response leveraging the preceding context and examples.

Empower users via Templates

Prompt templates help ensure **consistency** and **quality** across LLM interactions by providing a standardisation.

They allow non-expert users to leverage the full capabilities of LLMs effectively.

Reusable prompt templates save time and effort.

/ Going to (real) PRODUCTION

But, creating high quality,
accurate ~~RAGs~~ is hard



Systems

Multi-step
chains

RAG
Applications

Agent
& Tools

Fine-tuned
systems

Multi-model
usage

...

“Compound AI systems will likely be the best way to maximize AI results in the future”

Evaluating & improving

Models

GPT
Llama3
DBRX
Mixtral
Gemini
...



Evaluating & improving

Systems

RAG
Agents
Tools
Chains
Fine-tuned systems
...

But, creating high quality, accurate **Systems** is **hard**

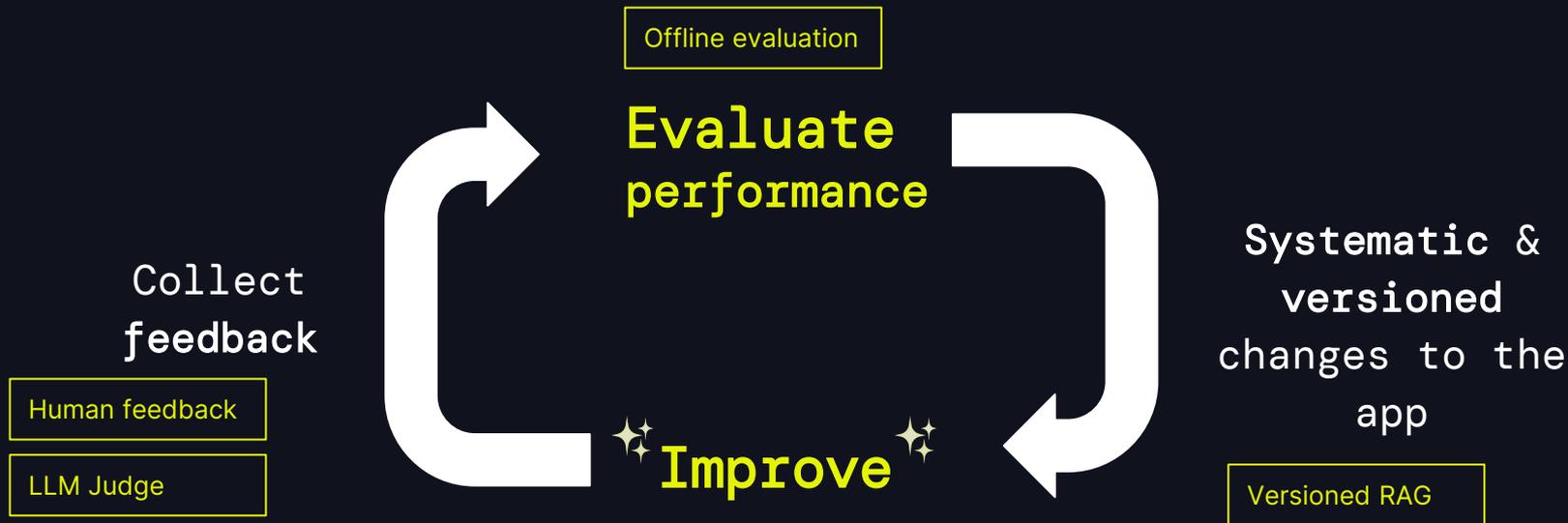
- ✘ Difficult to **measure and evaluate accuracy**
- ✘ Difficult to **collect enough feedback**
- ✘ Difficult to **improve accuracy/reduce hallucination**
- ✘ Difficult to enforce **guardrails** in the system

✨ Improve ✨
your RAG
app

You have to ...
Evaluate
your RAG
app

You have to ...
Have enough
feedback on
the RAG app

"If you can't measure it, you cannot improve it"



Summary

- **RAG**
RAG is a proven method for leveraging your *enterprise data* in GenAI apps
- **Use in combination with other techniques**
You can *combine* RAG with Prompt Engineering, Agents, Fine-tuning, Custom models, etc
- **Understand the RAG architecture**
RAG consists of an *offline* process to hydrate your knowledgebase, and an *online* process to retrieve the data in real-time
- **Getting started with RAG**
How to build a RAG app on Databricks without prior knowledge
- **Compound Systems**
RAG is a compound System.
- **Evaluation & Quality**
To improve your RAG applications, you need to be able to evaluate them.



Call to Action



Guide to building
RAG apps



RAG QuickStart
Solution (e2e demo)



DAIS sessions
on RAG

QR code leading to
RAG docs page

QR code leading to
DB demos for RAG

“Deep Dive into Building production
quality Gen AI Applications with [RAG
Studio]”

Advanced, 12th June 11:20am

“How to evaluate your GenAI
application quality”

Intermediate

“Using Vector Search for Efficient
Retrieval on Databricks”

Beginner



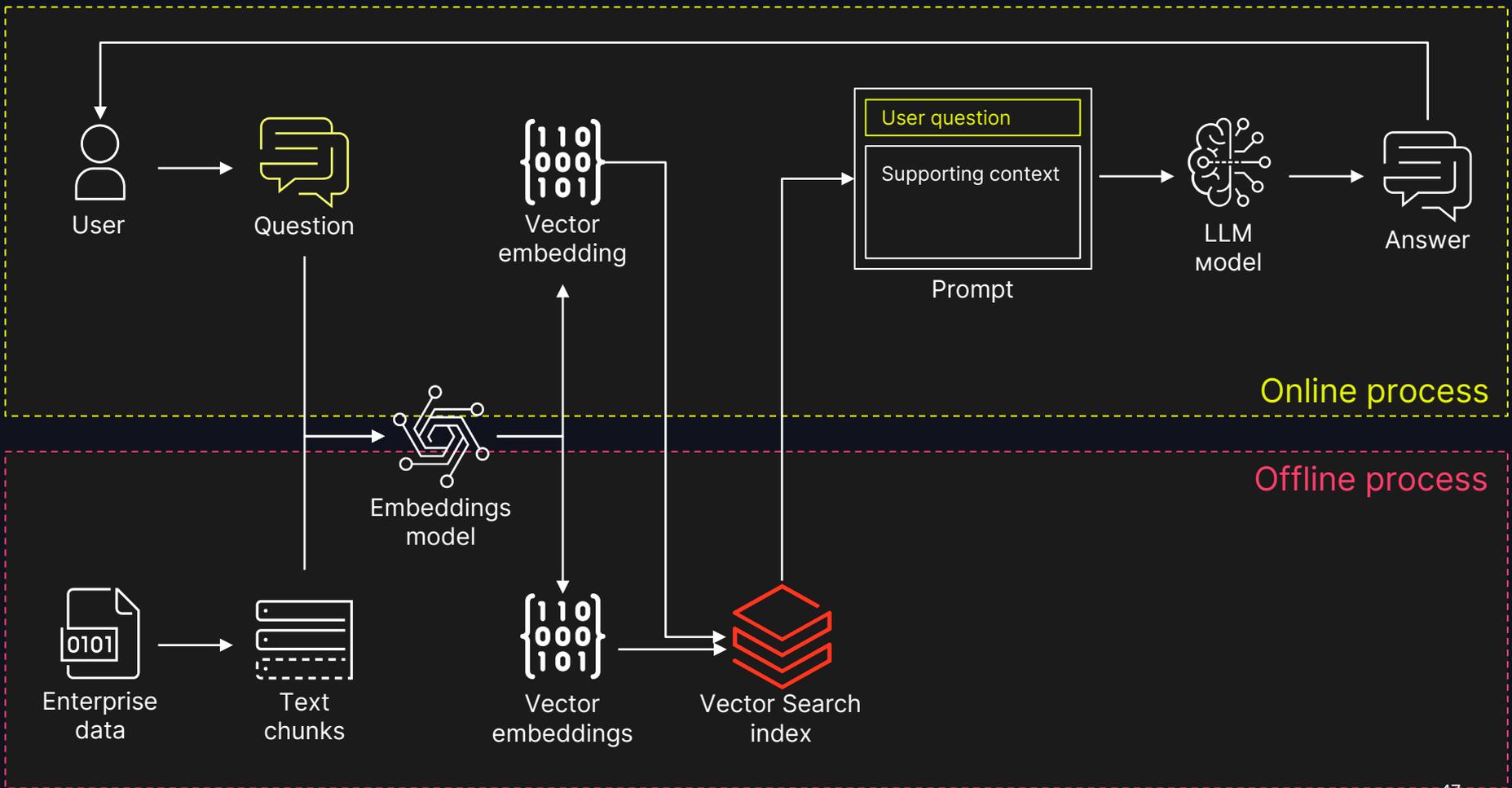
DATA+AI SUMMIT

Thank you!

 www.linkedin.com/in/inakoleva/

Appendix





Online process

Offline process

