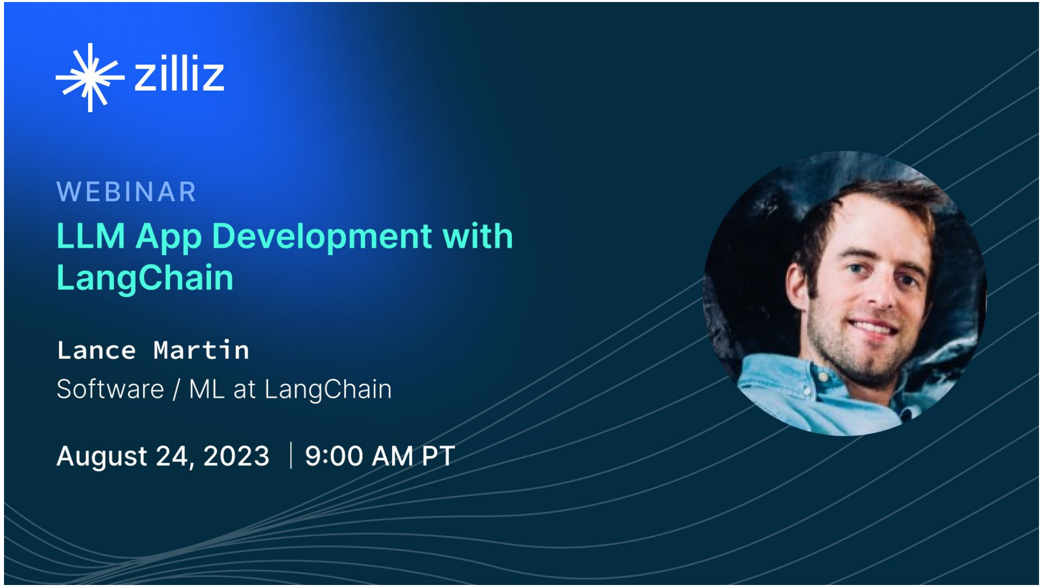
zilliz

# Building Retrieval Augmented Pipelines at Scale

August 2023 | Yujian Tang, Tuana Çelik

# Upcoming Events



zilliz

WEBINAR

**LLM App Development with LangChain**

**Lance Martin**

Software / ML at LangChain

August 24, 2023 | 9:00 AM PT

zilliz.com/event

# Building Retrieval Augmented Pipelines at Scale

August 2023 | Yujian Tang, Tuana Çelik

# Speaker

**Tuana Celik**

Developer Advocate, deepset

tuana.celik@deepset.ai
https://www.linkedin.com/in/tuanacelik
https://twitter.com/tuanacelik

# Speaker

**Yujian Tang**

Developer Advocate, Zilliz

yujian@zilliz.com
https://www.linkedin.com/in/yujiantang
https://www.twitter.com/yujian_tang

CONTENTS

01 Why Retrieval Augmented Generation?

02 Using a Vector Database for RAG

03 Building a RAG Pipeline w/ Haystack

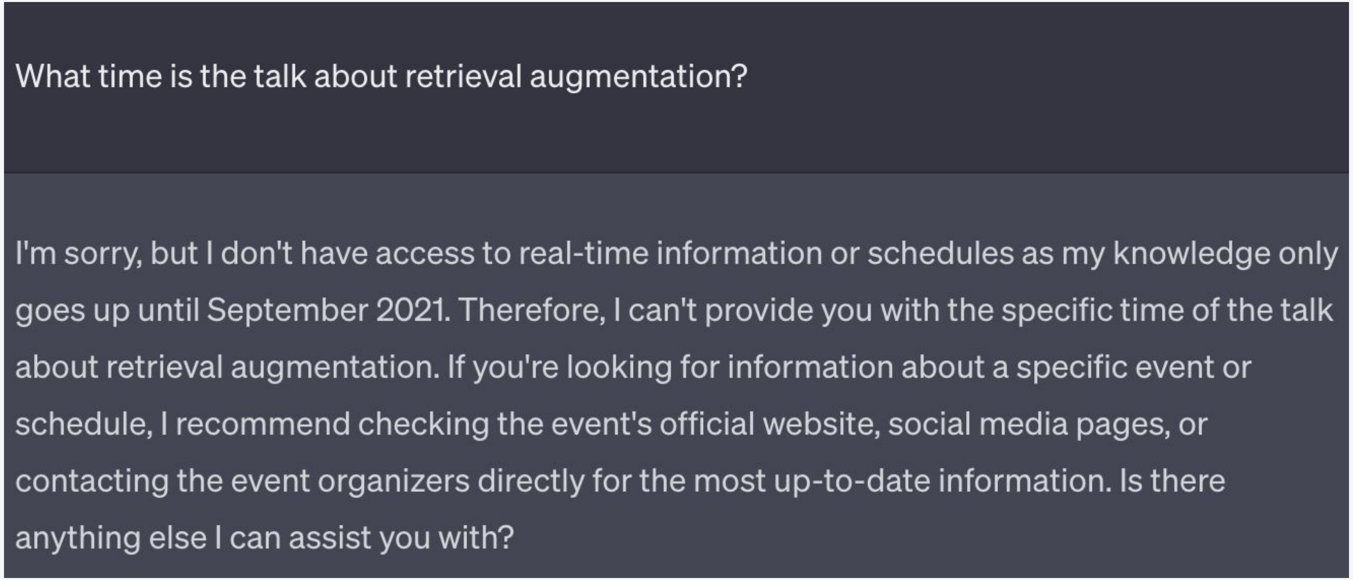
04 FAQ

zilliz

## 01

# Why Retrieval Augmented Generation?

# Why Retrieval Augmented Generation (RAG)? –an example

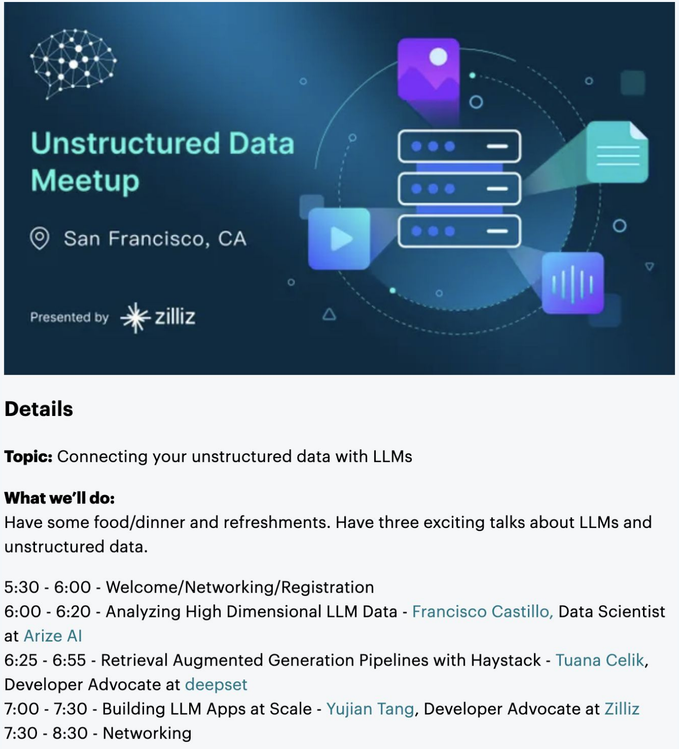


What time is the talk about retrieval augmentation?

I'm sorry, but I don't have access to real-time information or schedules as my knowledge only goes up until September 2021. Therefore, I can't provide you with the specific time of the talk about retrieval augmentation. If you're looking for information about a specific event or schedule, I recommend checking the event's official website, social media pages, or contacting the event organizers directly for the most up-to-date information. Is there anything else I can assist you with?

# Prompting and Prompt Engineering

- LLMs do not know the answer to everything
- We can help them by giving them the relevant context
- Once retrieved, we augment the prompt (instruction) with the relevant context



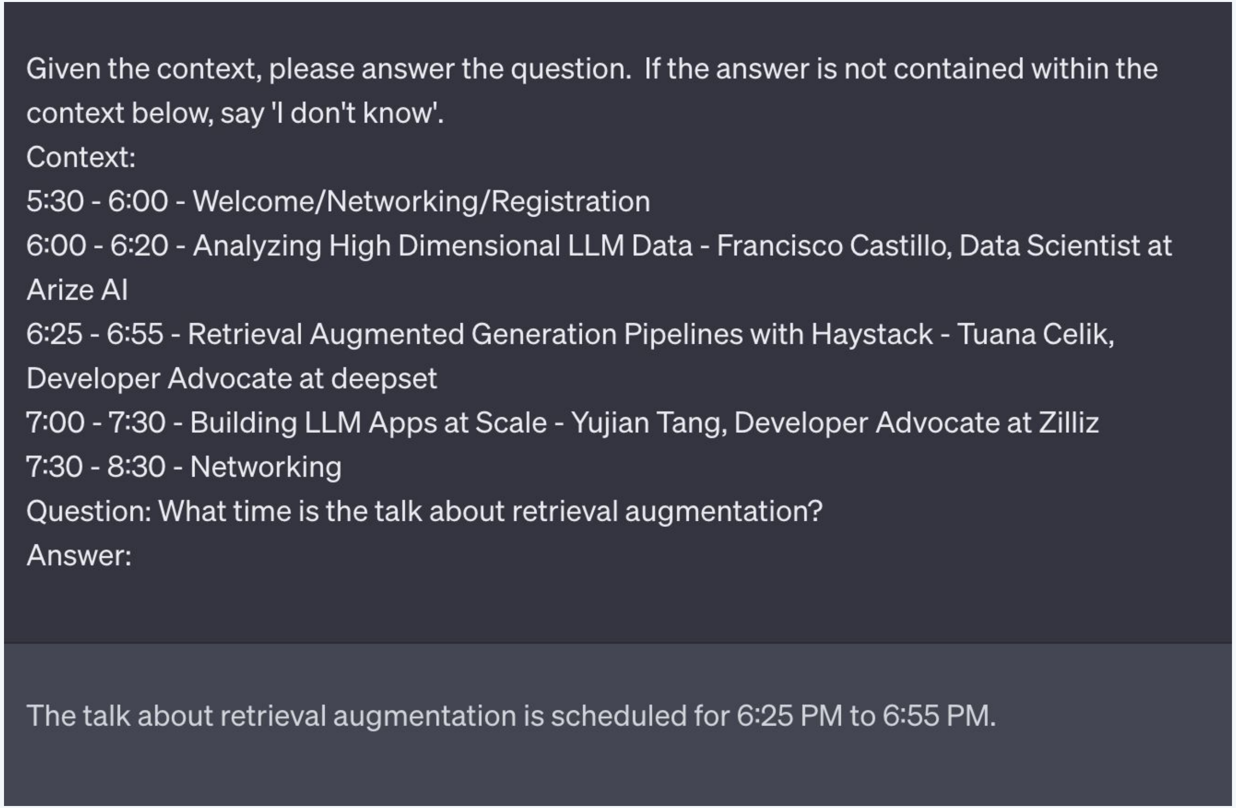
**Details**

**Topic:** Connecting your unstructured data with LLMs

**What we'll do:**
Have some food/dinner and refreshments. Have three exciting talks about LLMs and unstructured data.

5:30 - 6:00 - Welcome/Networking/Registration
6:00 - 6:20 - Analyzing High Dimensional LLM Data - Francisco Castillo, Data Scientist at Arize AI
6:25 - 6:55 - Retrieval Augmented Generation Pipelines with Haystack - Tuana Celik, Developer Advocate at deepset
7:00 - 7:30 - Building LLM Apps at Scale - Yujian Tang, Developer Advocate at Zilliz
7:30 - 8:30 - Networking

# Prompting and Prompt Engineering

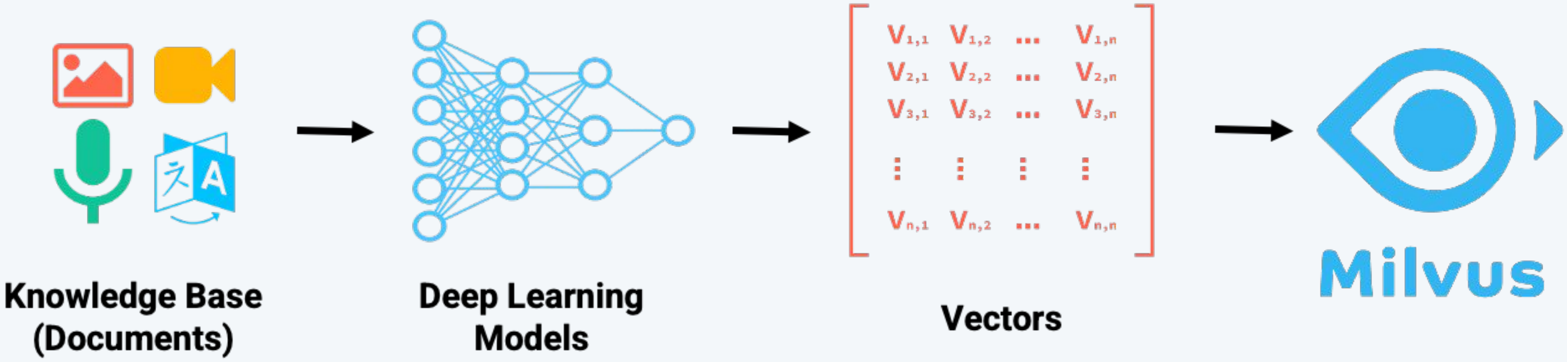


Given the context, please answer the question. If the answer is not contained within the context below, say 'I don't know'.
Context:
5:30 - 6:00 - Welcome/Networking/Registration
6:00 - 6:20 - Analyzing High Dimensional LLM Data - Francisco Castillo, Data Scientist at Arize AI
6:25 - 6:55 - Retrieval Augmented Generation Pipelines with Haystack - Tuana Celik, Developer Advocate at deepset
7:00 - 7:30 - Building LLM Apps at Scale - Yujian Tang, Developer Advocate at Zilliz
7:30 - 8:30 - Networking
Question: What time is the talk about retrieval augmentation?
Answer:

The talk about retrieval augmentation is scheduled for 6:25 PM to 6:55 PM.

© Copyright 8/4/23 Zilliz

# How Does RAG work?



**Knowledge Base (Documents)**

**Deep Learning Models**

**Vectors**

© Copyright 8/4/23 Zilliz

# What’s a RAG tech stack look like?

## CVP Stack

**C**: ChatGPT (or any other LLM)

- This can also be interpreted as the “processor” block for CVP

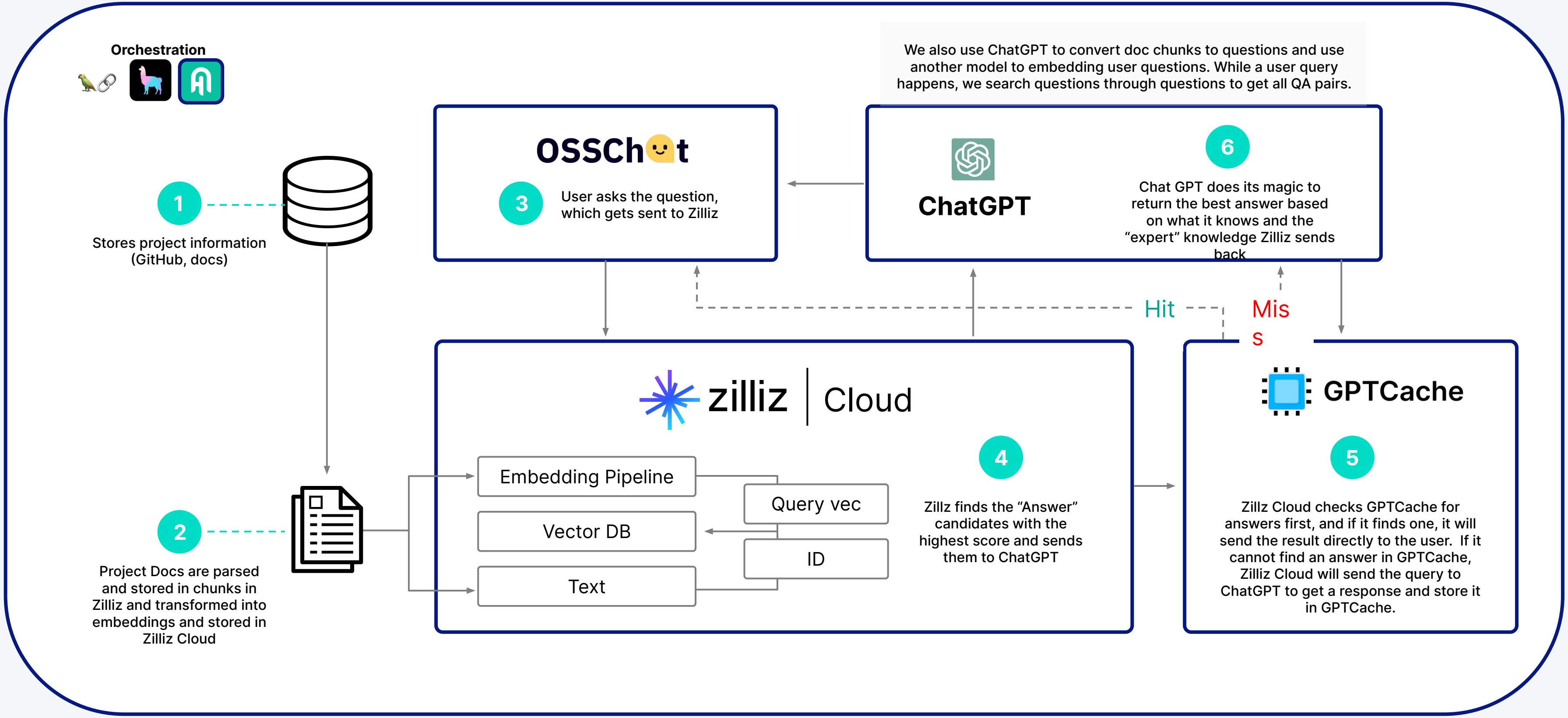
**V**: Vector database (e.g. Milvus)

- Can also be interpreted as the “storage” block for CVP

**P**: Prompt-as-code (e.g. Haystack)

- Interface between processor and storage blocks

# OSS Chat | CVP Stack



**Orchestration**

We also use ChatGPT to convert doc chunks to questions and use another model to embedding user questions. While a user query happens, we search questions through questions to get all QA pairs.

1 Stores project information (GitHub, docs)

2 Project Docs are parsed and stored in chunks in Zilliz and transformed into embeddings and stored in Zilliz Cloud

**OSSChat**

3 User asks the question, which gets sent to Zilliz

**ChatGPT**

6 Chat GPT does its magic to return the best answer based on what it knows and the "expert" knowledge Zilliz sends back

Hit

Miss

**zilliz | Cloud**

Embedding Pipeline

Vector DB

Text

Query vec

ID

4 Zilliz finds the "Answer" candidates with the highest score and sends them to ChatGPT

**GPTCache**

5 Zilliz Cloud checks GPTCache for answers first, and if it finds one, it will send the result directly to the user. If it cannot find an answer in GPTCache, Zilliz Cloud will send the query to ChatGPT to get a response and store it in GPTCache.

© Copyright 8/4/23 Zilliz

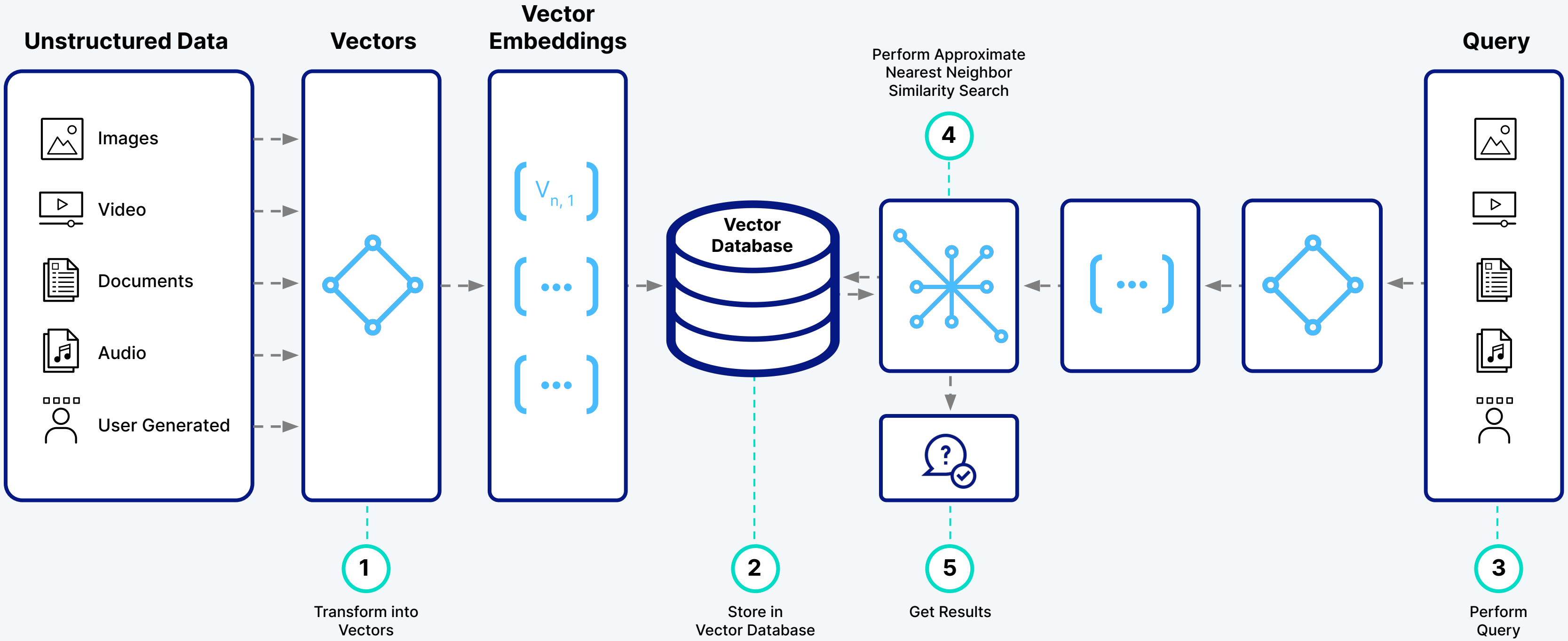
## 02

# Using a Vector DB for RAG

# Why Use a Vector Database?

- Use your data
- Get relevant responses
- Economics

# How Similarity Search Works



**Unstructured Data**

- Images
- Video
- Documents
- Audio
- User Generated

**Vectors**

**Vector Embeddings**

Vector Database

Perform Approximate Nearest Neighbor Similarity Search

**Query**

1 Transform into Vectors

2 Store in Vector Database

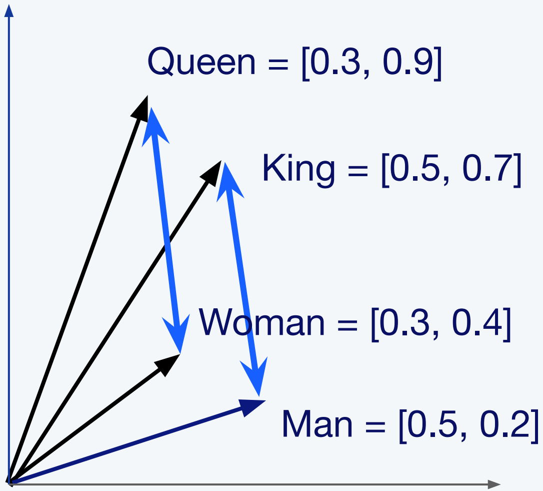
3 Perform Query

4 Perform Approximate Nearest Neighbor Similarity Search

5 Get Results

# Semantic Similarity

Queen - Woman + Man = King

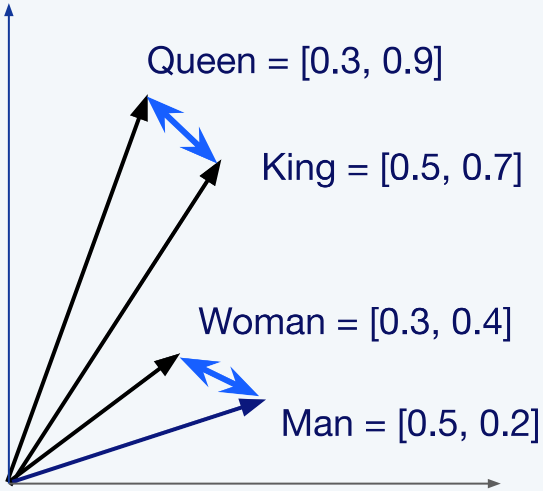


Queen = [0.3, 0.9]

King = [0.5, 0.7]

Woman = [0.3, 0.4]

Man = [0.5, 0.2]



Queen = [0.3, 0.9]

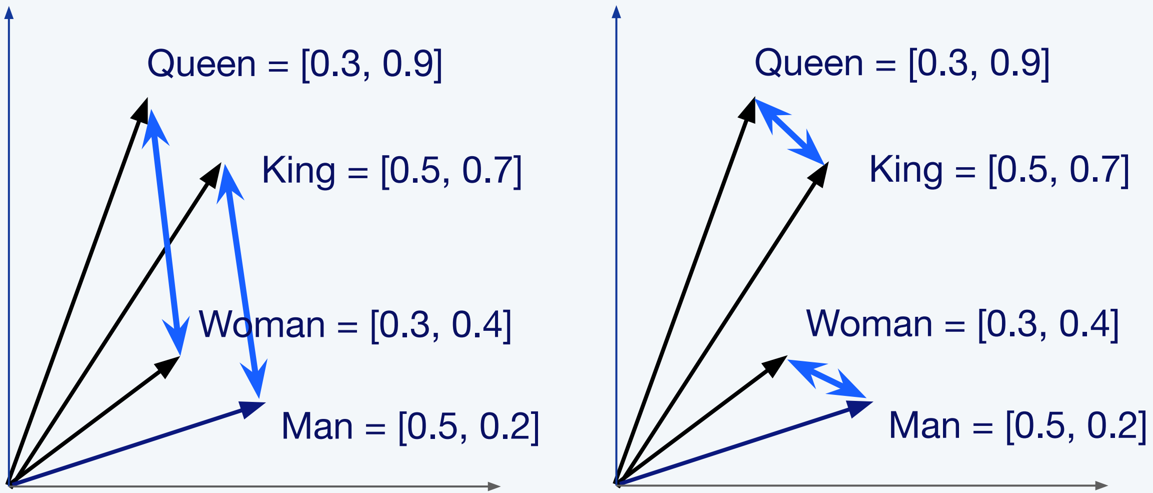
King = [0.5, 0.7]

Woman = [0.3, 0.4]

Man = [0.5, 0.2]

Image from Sutor et al

# Semantic Similarity



Queen = [0.3, 0.9]

King = [0.5, 0.7]

Woman = [0.3, 0.4]

Man = [0.5, 0.2]

Queen = [0.3, 0.9]

King = [0.5, 0.7]

Woman = [0.3, 0.4]

Man = [0.5, 0.2]

Image from Sutor et al

Queen - Woman + Man = King

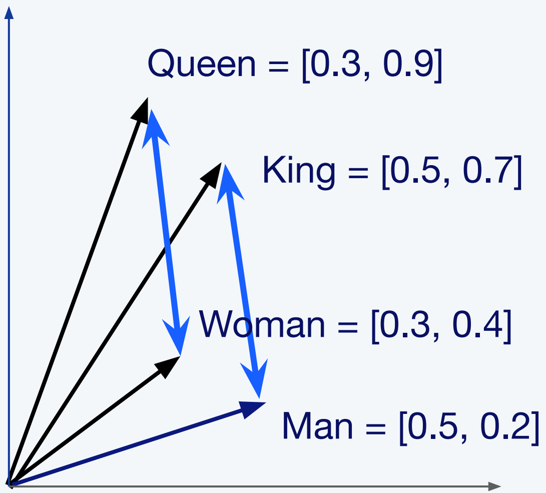
Queen = [0.3, 0.9]
- Woman = [0.3, 0.4]

---

[0.0, 0.5]

© Copyright 8/4/23 Zilliz

# Semantic Similarity



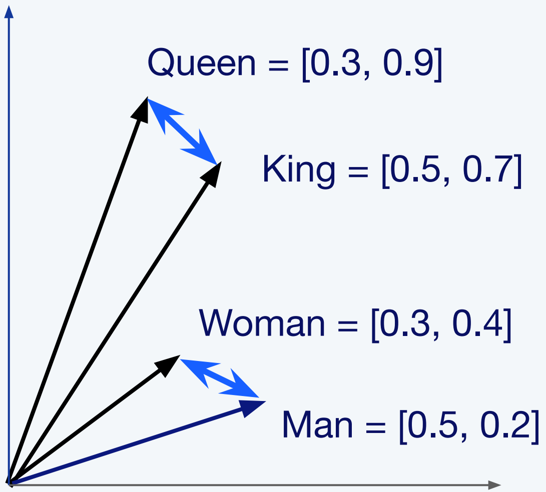


Image from Sutor et al

Queen - Woman + Man = King

Queen = [0.3, 0.9]
- Woman = [0.3, 0.4]

---

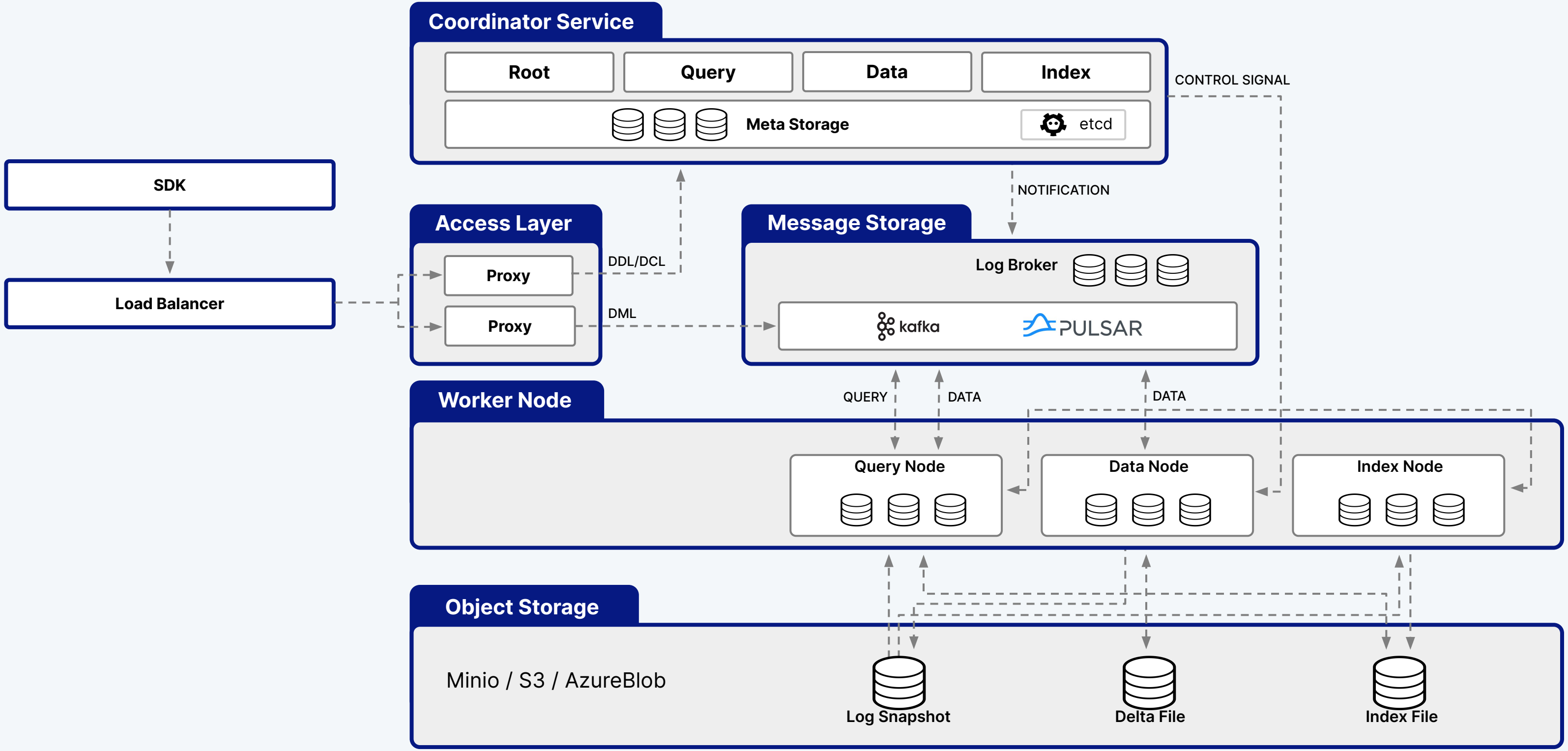
[0.0, 0.5]
+ Man = [0.5, 0.2]

---

King = [0.5, 0.7]

# Vector Database Architecture

milvus
VECTOR DATABASE



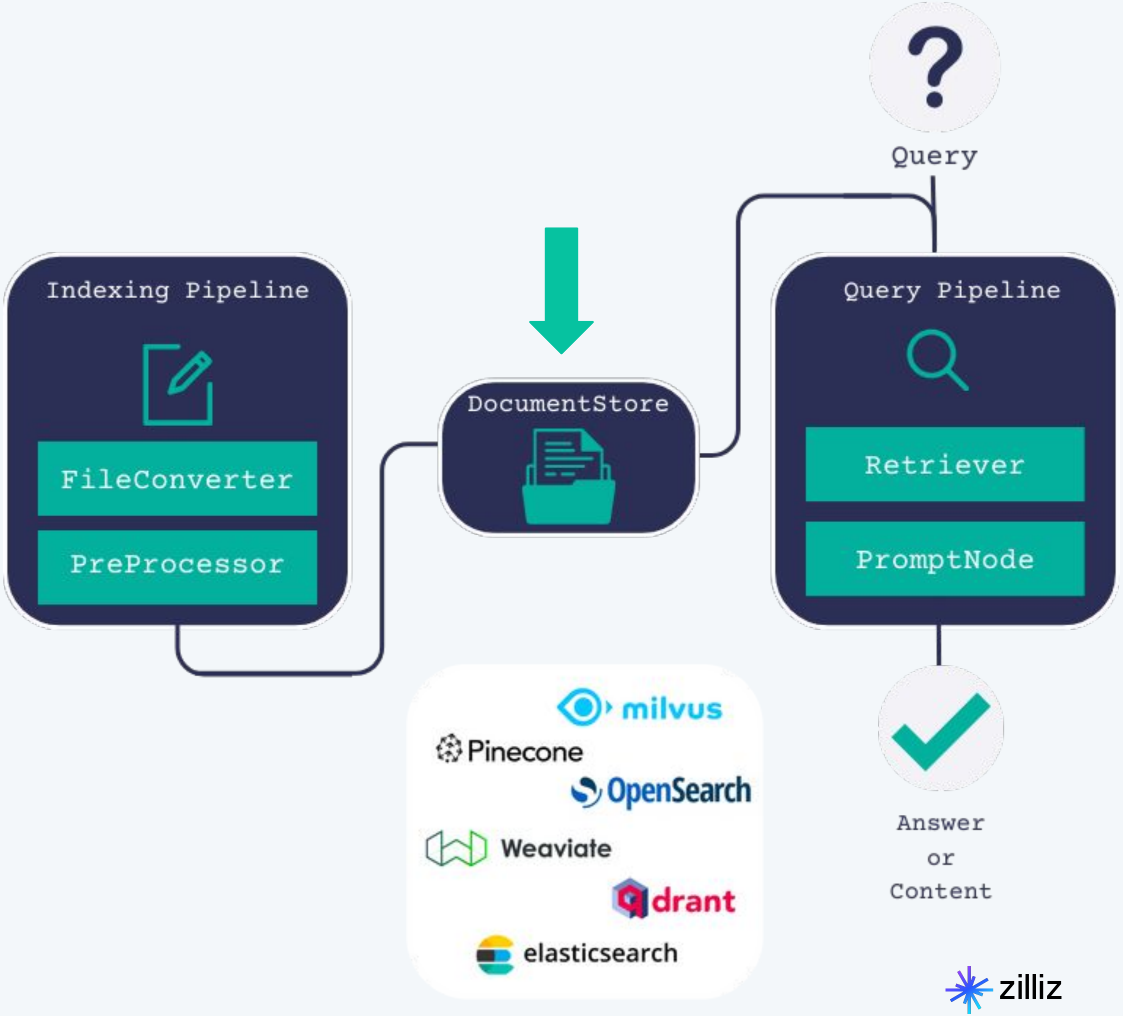
zilliz

# 03

# Building a RAG Pipeline with Haystack

# Haystack

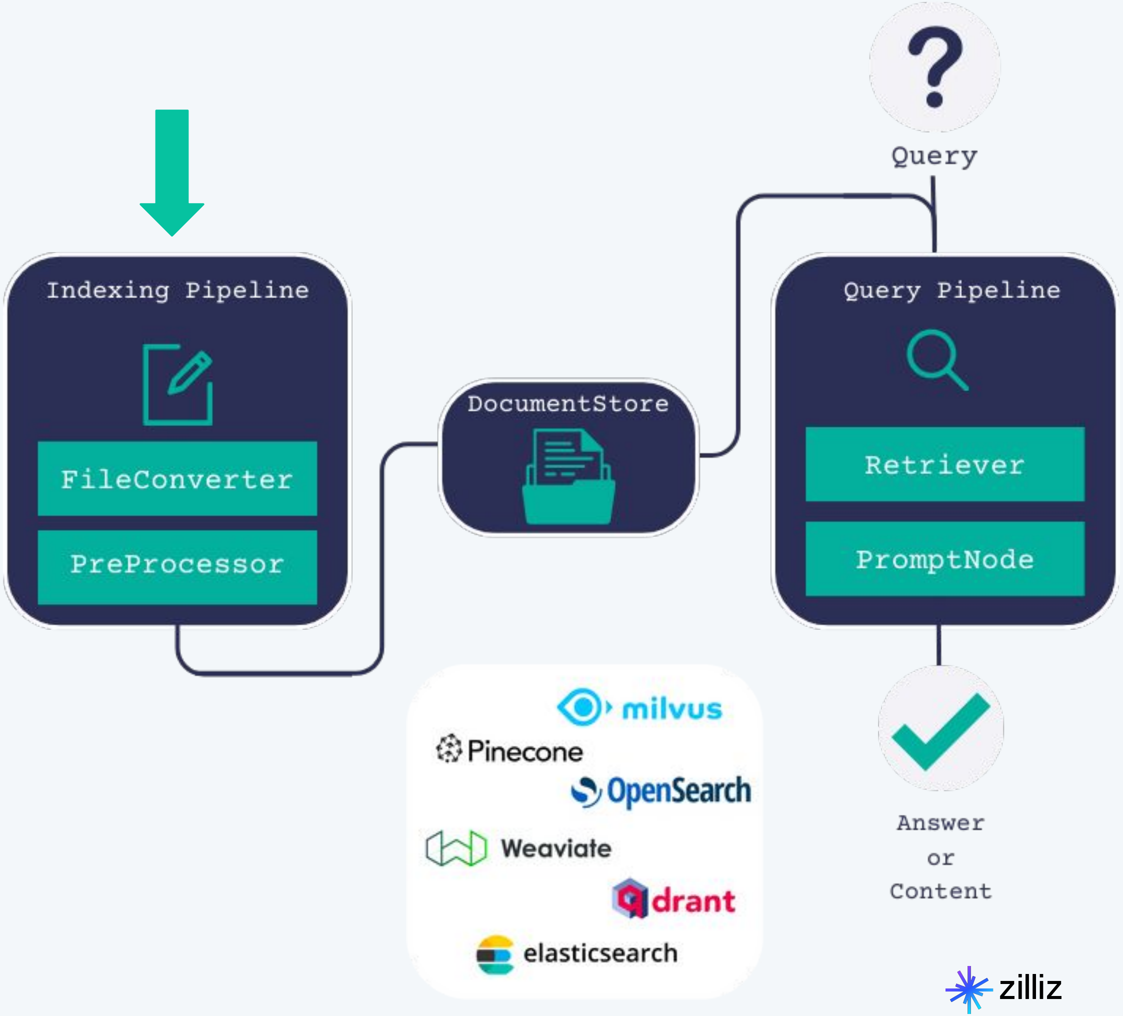
- Fully open-source framework built in Python to designed to build LLM applications
- Core NLP tasks covered
- Production focused



zilliz

# Haystack

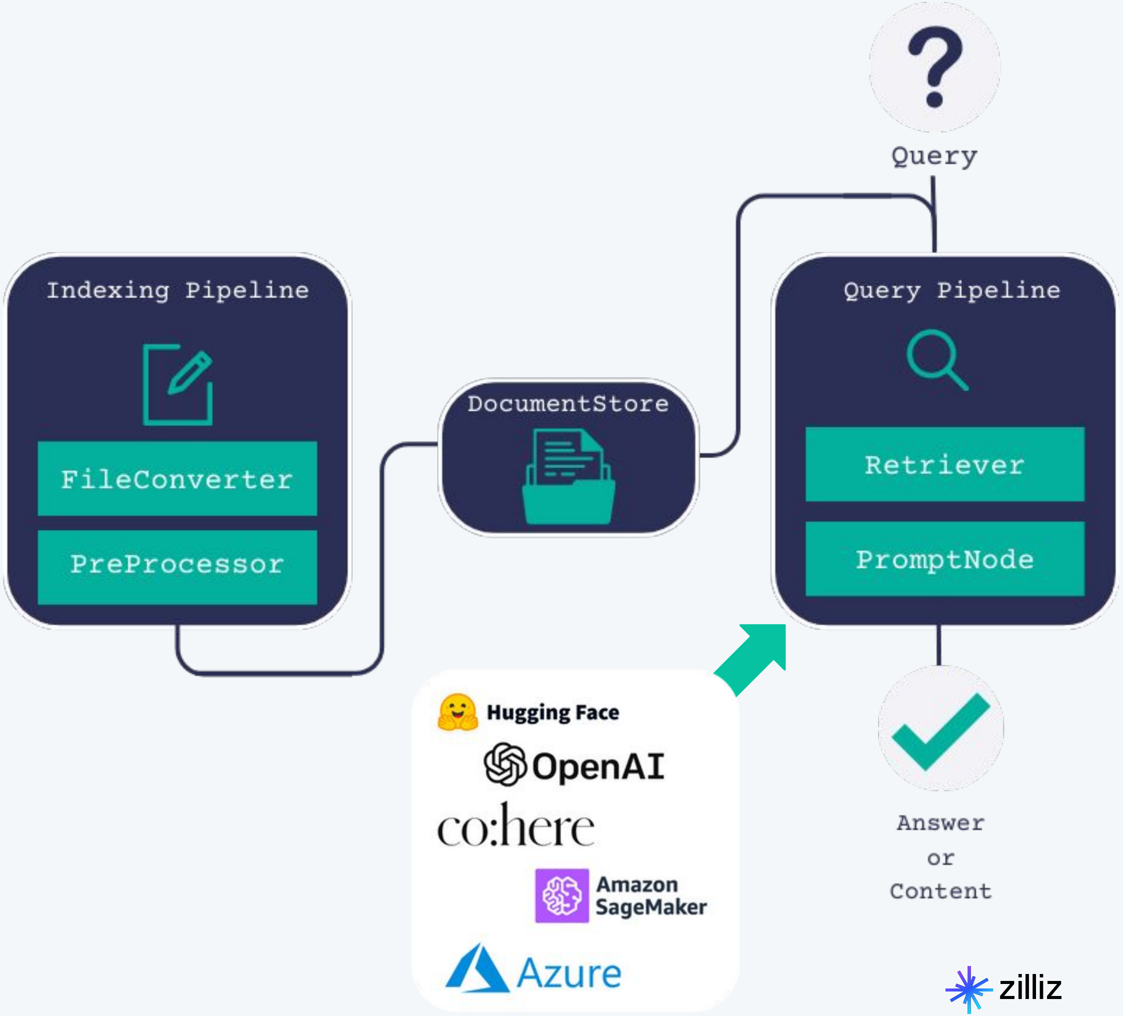
- Fully open-source framework built in Python to designed to build LLM applications
- Core NLP tasks covered
- Production focused



zilliz

# Haystack

- Fully open-source framework built in Python to designed to build LLM applications
- Core NLP tasks covered
- Production focused



zilliz

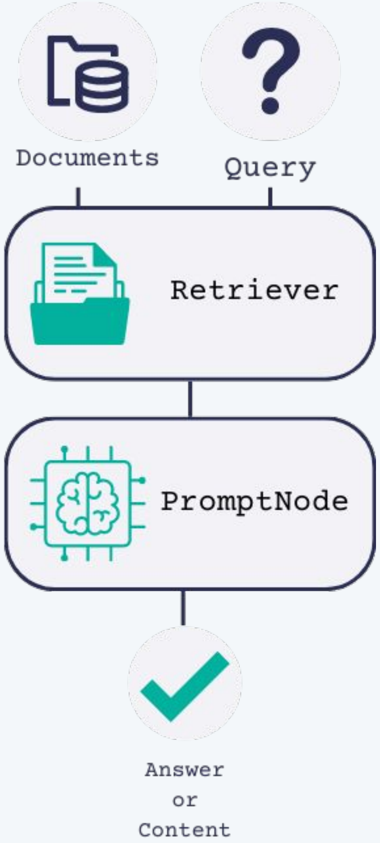
# Haystack Pipelines: RAG

**PromptTemplate**

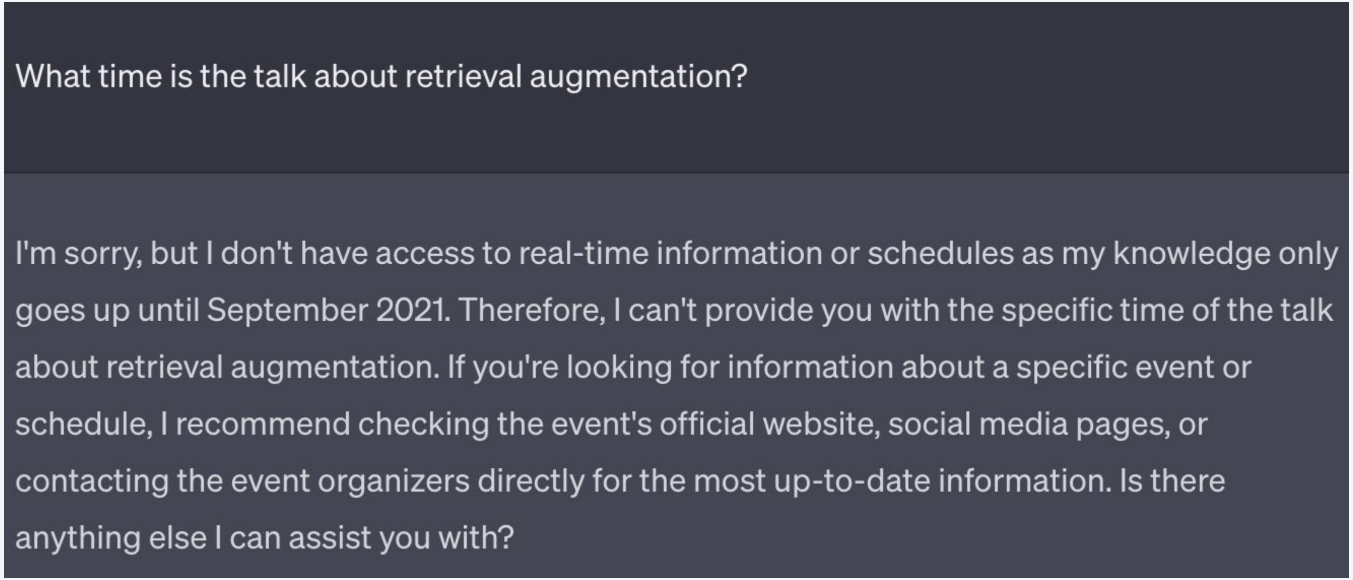
- Flexible prompts that can be modified per query
- A blueprint of how to interact with LLMs

**PromptNode**

- Interact with LLMs in a customized way
- An interface to send queries and receive answers from LLMs
- Use models from OpenAI, Hugging Face (falcon, MPT) and co:here (and more to come)



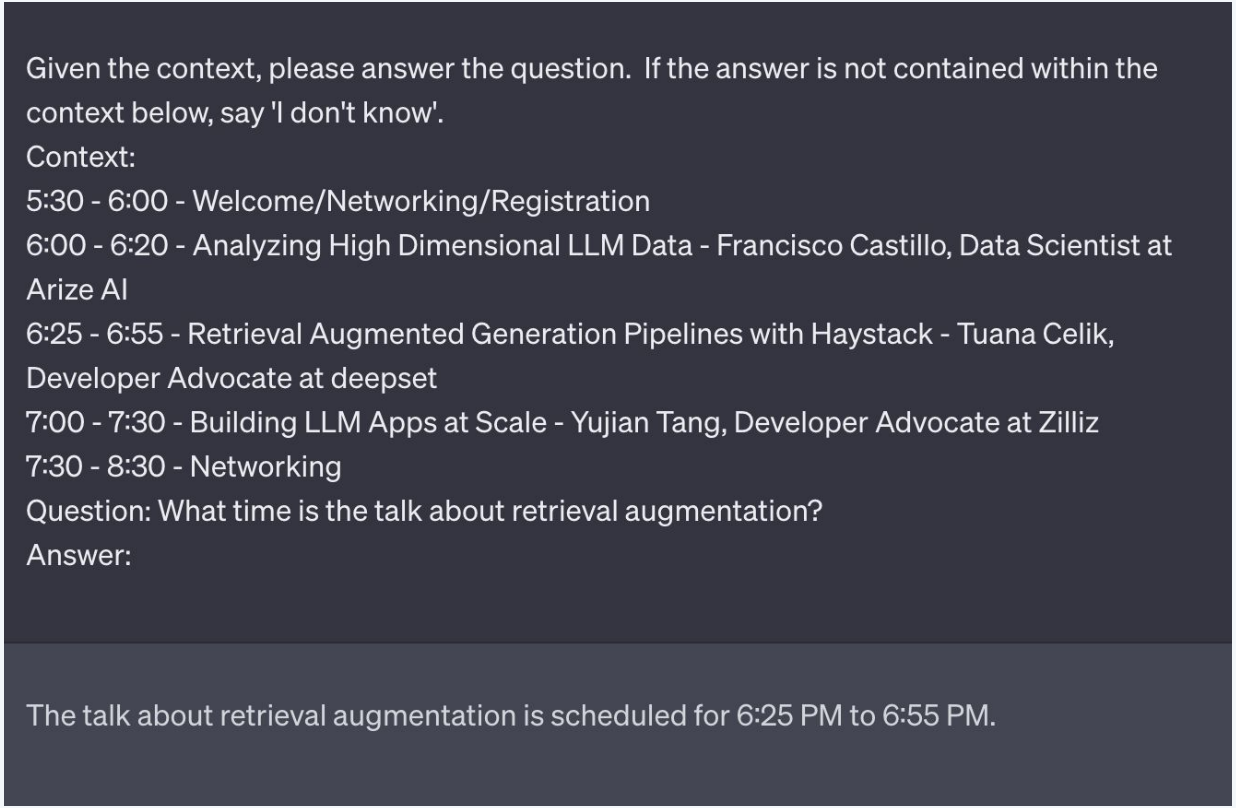
# Haystack Pipelines: RAG



What time is the talk about retrieval augmentation?

I'm sorry, but I don't have access to real-time information or schedules as my knowledge only goes up until September 2021. Therefore, I can't provide you with the specific time of the talk about retrieval augmentation. If you're looking for information about a specific event or schedule, I recommend checking the event's official website, social media pages, or contacting the event organizers directly for the most up-to-date information. Is there anything else I can assist you with?

# Haystack Pipelines: RAG

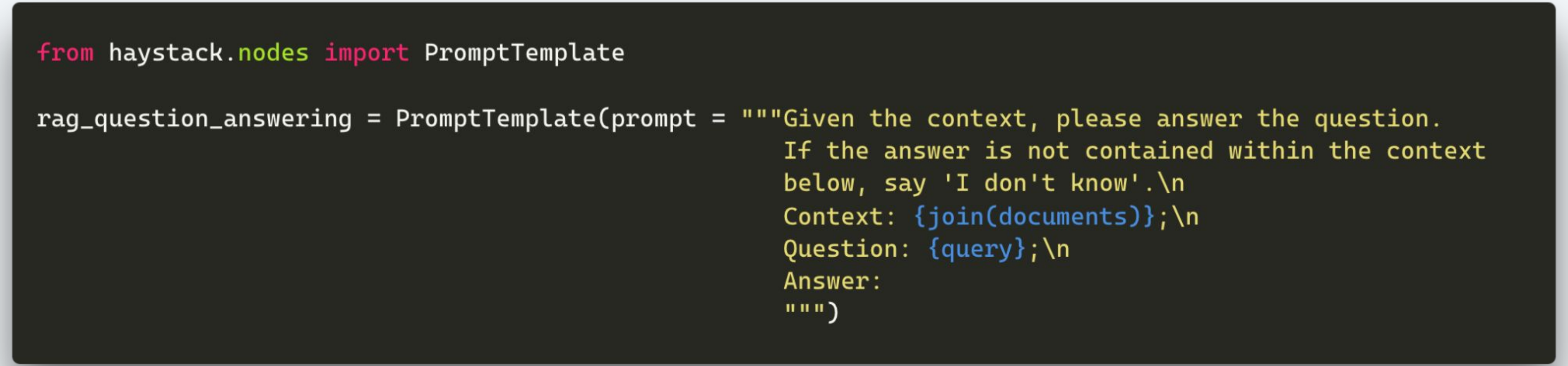


Given the context, please answer the question.  If the answer is not contained within the context below, say 'I don't know'.
Context:
5:30 - 6:00 - Welcome/Networking/Registration
6:00 - 6:20 - Analyzing High Dimensional LLM Data - Francisco Castillo, Data Scientist at Arize AI
6:25 - 6:55 - Retrieval Augmented Generation Pipelines with Haystack - Tuana Celik, Developer Advocate at deepset
7:00 - 7:30 - Building LLM Apps at Scale - Yujian Tang, Developer Advocate at Zilliz
7:30 - 8:30 - Networking
Question: What time is the talk about retrieval augmentation?
Answer:

The talk about retrieval augmentation is scheduled for 6:25 PM to 6:55 PM.

# Prompt Templating

**Instruct the LLM on what to do with the provided information. For example, build your custom PromptTemplate 👇**

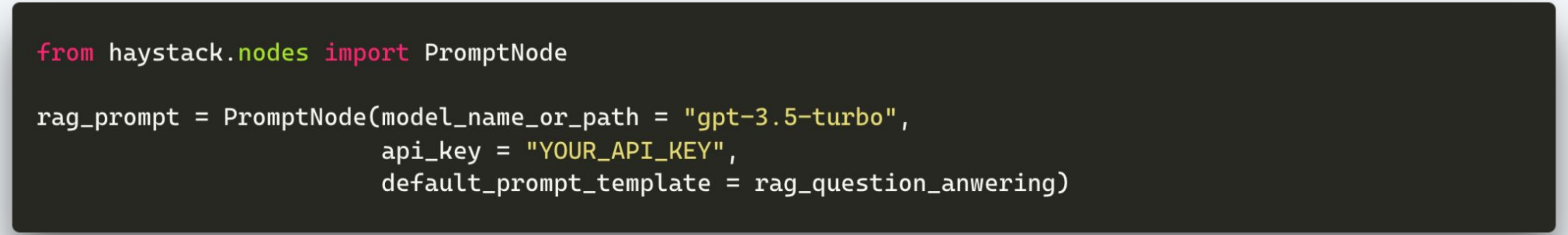


```
from haystack.nodes import PromptTemplate

rag_question_answering = PromptTemplate(prompt = """Given the context, please answer the question.
                                                    If the answer is not contained within the context
                                                    below, say 'I don't know'.\n
                                                    Context: {join(documents)};\n
                                                    Question: {query};\n
                                                    Answer:
                                                    """)
```

| © Copyright 8/4/23 Zilliz

# PromptNode

## Use your PromptTemplate as a blueprint of how you want to interact with the LLM of your choice 👇

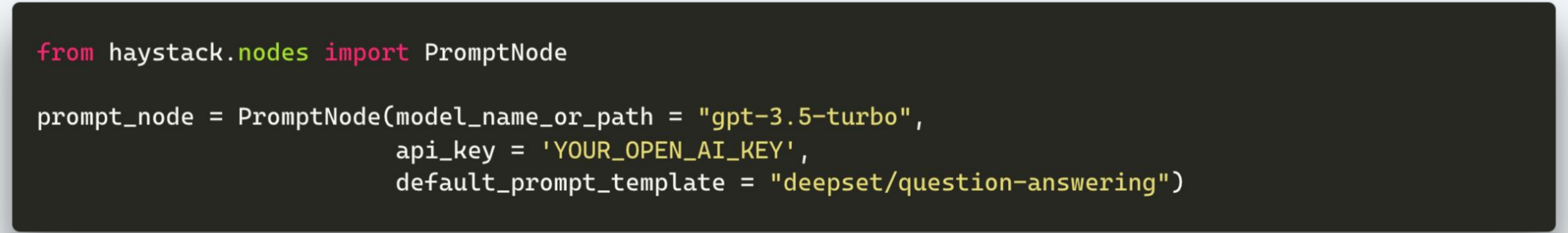


```python
from haystack.nodes import PromptNode

rag_prompt = PromptNode(model_name_or_path = "gpt-3.5-turbo",
                        api_key = "YOUR_API_KEY",
                        default_prompt_template = rag_question_anwering)
```

© Copyright 8/4/23 Zilliz

# PromptNode

**Or, use one of the prompts that we've made available on the [PromptHub](#) 👇**

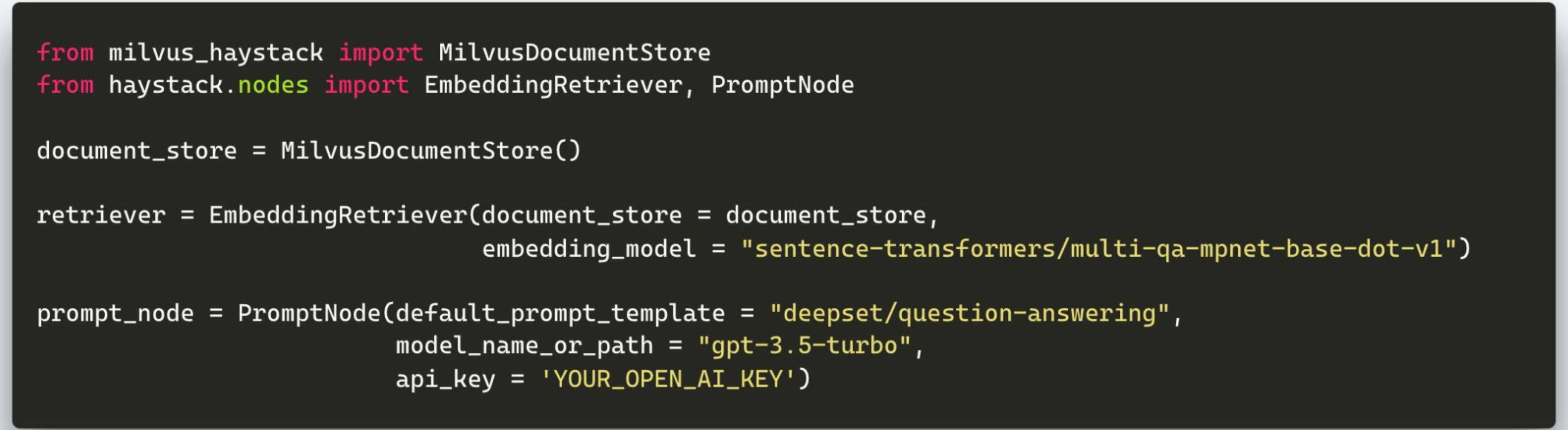


```python
from haystack.nodes import PromptNode

prompt_node = PromptNode(model_name_or_path = "gpt-3.5-turbo",
                         api_key = 'YOUR_OPEN_AI_KEY',
                         default_prompt_template = "deepset/question-answering")
```

© Copyright 8/4/23 Zilliz

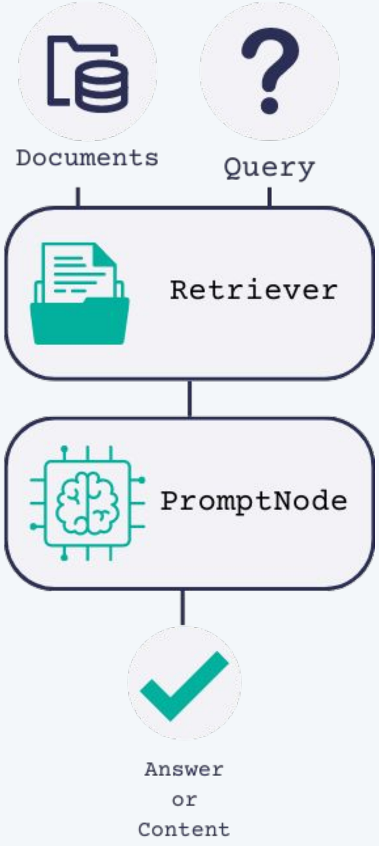
# Build a RAG Pipeline

**You can build a RAG pipeline on top of the DocumentStore of your choice as well as the models of your choice for both the retrieval and answer generation steps 👇**



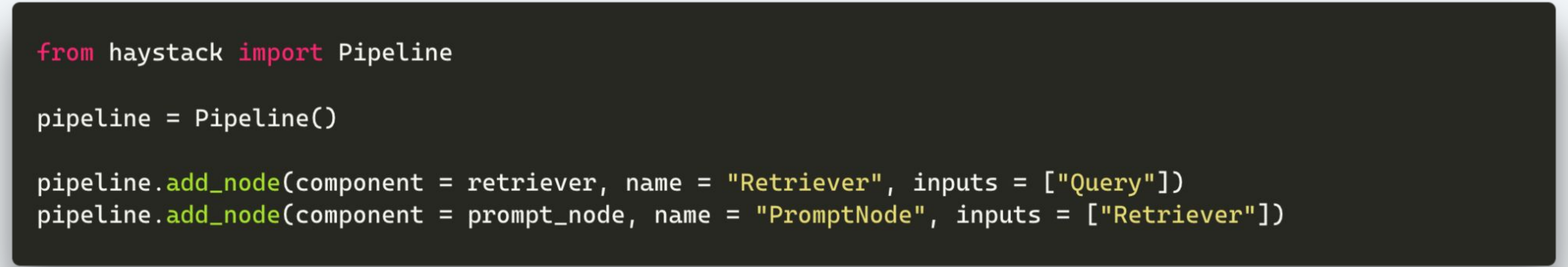
```python
from milvus_haystack import MilvusDocumentStore
from haystack.nodes import EmbeddingRetriever, PromptNode

document_store = MilvusDocumentStore()

retriever = EmbeddingRetriever(document_store = document_store,
                               embedding_model = "sentence-transformers/multi-qa-mpnet-base-dot-v1")

prompt_node = PromptNode(default_prompt_template = "deepset/question-answering",
                         model_name_or_path = "gpt-3.5-turbo",
                         api_key = 'YOUR_OPEN_AI_KEY')
```

# Build a RAG Pipeline



# Build a RAG Pipeline

**You can build a RAG pipeline on top of the DocumentStore of your choice as well as the models of your choice for both the retrieval and answer generation steps 👇**

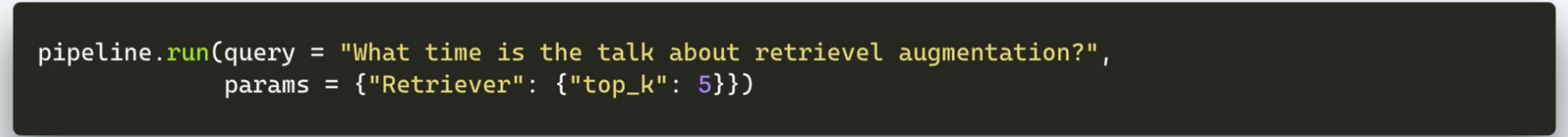


```python
from haystack import Pipeline

pipeline = Pipeline()

pipeline.add_node(component = retriever, name = "Retriever", inputs = ["Query"])
pipeline.add_node(component = prompt_node, name = "PromptNode", inputs = ["Retriever"])
```

# Build a RAG Pipeline

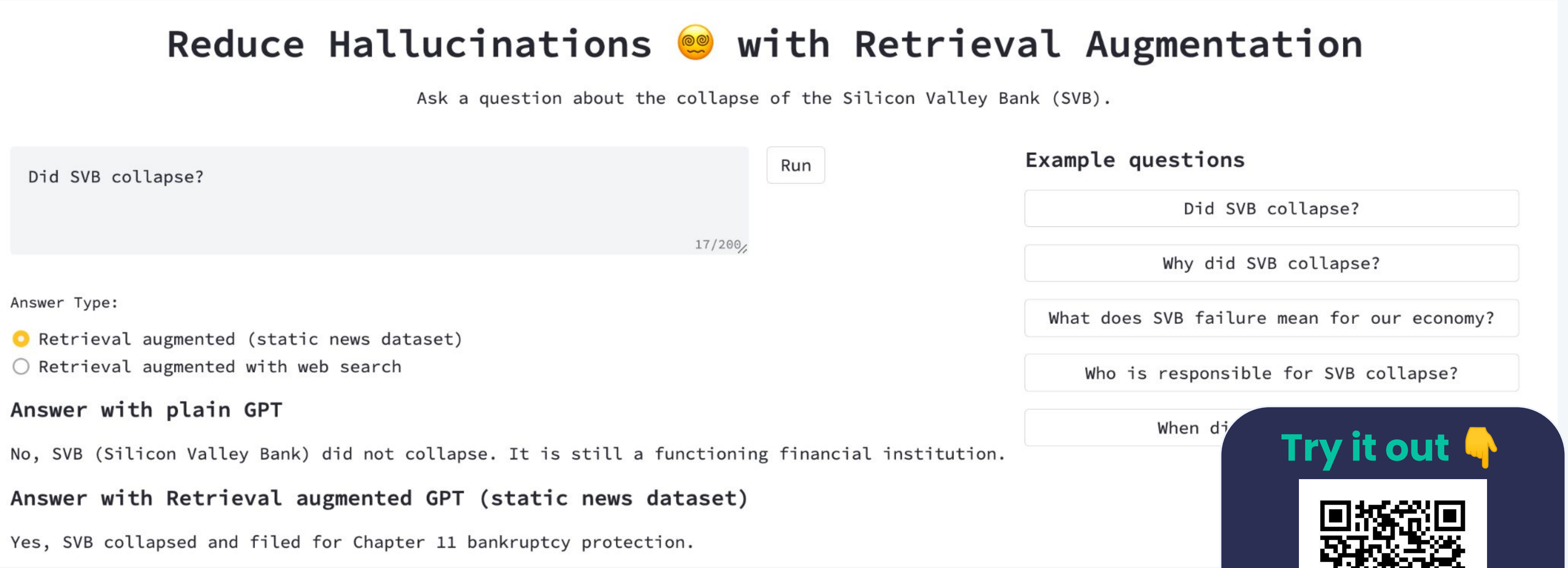
**When you run the pipeline with a query:**

1. **The top_k retrieved documents will be joined and added to the prompt**
2. **The LLM will produce an answer based on the full, augmented instruction 👇**



```
pipeline.run(query = "What time is the talk about retrievel augmentation?",
             params = {"Retriever": {"top_k": 5}})
```

| © Copyright 8/4/23 Zilliz



# Reduce Hallucinations 😵‍💫 with Retrieval Augmentation

Ask a question about the collapse of the Silicon Valley Bank (SVB).

Did SVB collapse?

17/200

Answer Type:

- ● Retrieval augmented (static news dataset)
- ○ Retrieval augmented with web search

### Answer with plain GPT

No, SVB (Silicon Valley Bank) did not collapse. It is still a functioning financial institution.

### Answer with Retrieval augmented GPT (static news dataset)

Yes, SVB collapsed and filed for Chapter 11 bankruptcy protection.

### Example questions

Did SVB collapse?
Why did SVB collapse?
What does SVB failure mean for our economy?
Who is responsible for SVB collapse?
When di

**Try it out 👇**

## 04

# FAQs

# FAQ - Use Cases

- When *NOT* to use
- CSV Files? PDFs?
- Hybrid Search

THANK YOU

zilliz

haystack
by deepset

# 05

# Appendix

# An Example Idea

**Example**

- A company has 100,000s+ pages of proprietary documentation to enable their staff to service customers.

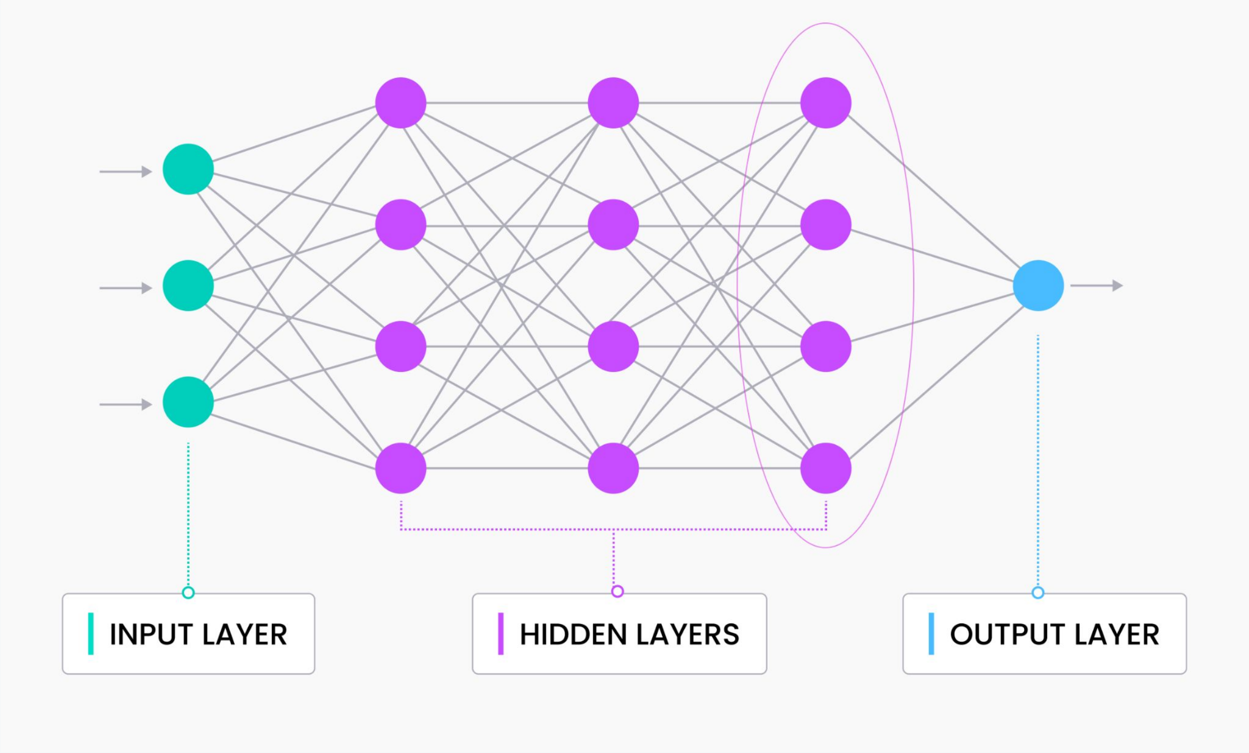
**Problem**

- Searching can be slow, inefficient, or lack context.

**Solution**

- Create internal chatbot with ChatGPT and a vector database enriched with company documentation to provide direction and support to employees and customers.

# How are these generated?



INPUT LAYER

HIDDEN LAYERS

OUTPUT LAYER

# Traditional databases face lots of challenges to manage vectors

- Inefficiency in High-dimensional spaces
- Suboptimal Indexing
- Inadequate query support
- Lack of scalability
- Limited analytics capabilities
- Data conversion issues

# Why a Vector Database?

Purpose-built to store, index and query vector embeddings from unstructured data.

## Vector database

- Advanced filtering (filtered vector search, chained filters)
- Hybrid search (e.g. full text + dense vector)
- Durability (any write in a db is durable, a library typically only supports snapshotting)
- Replication / High Availability
- Sharding
- Aggregations or faceted search
- Backups
- Lifecycle management (CRUD, Batch delete, dropping whole indexes, reindexing)
- Multi-tenancy

## Vector search library

- High-performance vector search

## How do I support different applications?

- High query load
- High insertion/deletion
- Full precision/recall
- Accelerator support (GPU, FPGA)
- Billion-scale storage

© Copyright 8/9/23 Zilliz