Couchbase

4교시.
Semantic Search/GenAI, 실습

1 Semantic Search
2 GenAI (LLM, RAG)
3 실습 : Semantic Search
4 실습 : GenAI. LLM/RAG
A 참고 자료

Confidential and Proprietary. Do not distribute without Couchbase consent. © Couchbase 2023. All rights reserved.

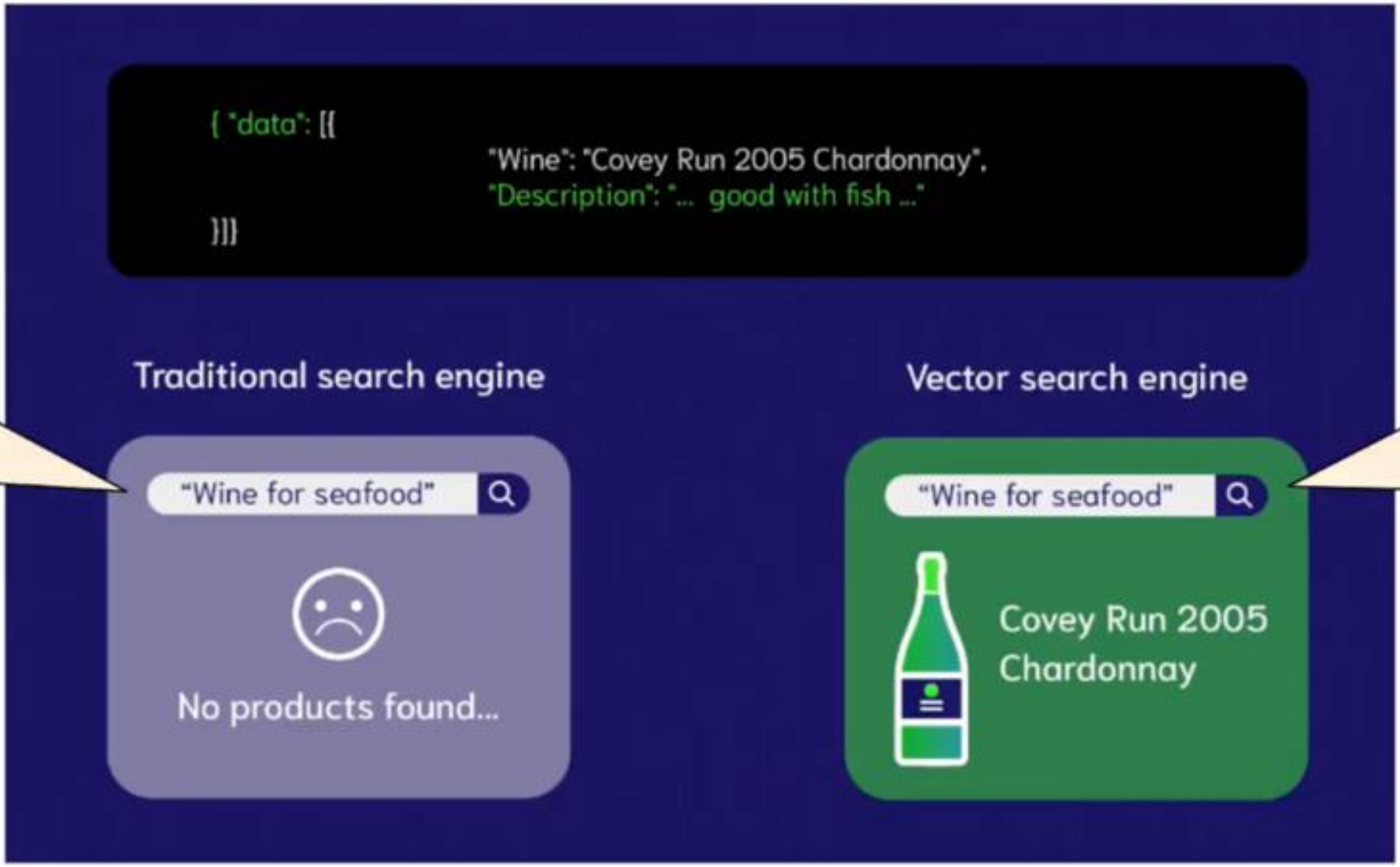
4-1.
Semantic Search

What is a Semantic Search

This is a
Keyword Search.

It looks for exact keywords matches in the database.

Powered by Traditional Search over Texts



```
{ "data": [{
                "Wine": "Covey Run 2005 Chardonnay",
                "Description": "... good with fish ..."
}]}
```

This is a
Semantic Search.

It considers the meaning and context behind the words.

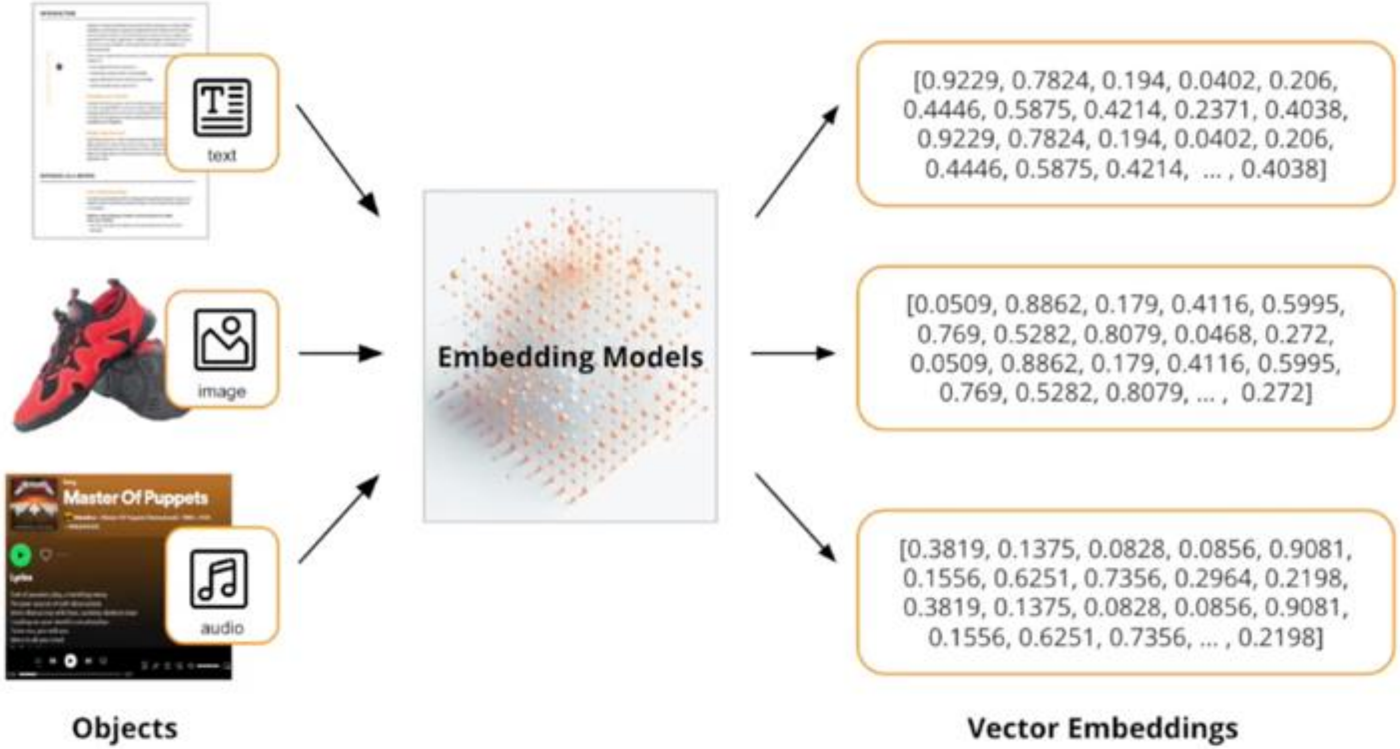
Powered by Vector Search over Vectors

Semantic Search = Similarity Search across Vectors Embeddings representing **the meaning of complex Objects**

Source: Dan Dascalescu

Confidential and Proprietary. Do not distribute without Couchbase consent. © Couchbase 2024. All rights reserved.

What are Embedding Models



Embedding Models

[0.9229, 0.7824, 0.194, 0.0402, 0.206, 0.4446, 0.5875, 0.4214, 0.2371, 0.4038, 0.9229, 0.7824, 0.194, 0.0402, 0.206, 0.4446, 0.5875, 0.4214, ... , 0.4038]

[0.0509, 0.8862, 0.179, 0.4116, 0.5995, 0.769, 0.5282, 0.8079, 0.0468, 0.272, 0.0509, 0.8862, 0.179, 0.4116, 0.5995, 0.769, 0.5282, 0.8079, ... , 0.272]

[0.3819, 0.1375, 0.0828, 0.0856, 0.9081, 0.1556, 0.6251, 0.7356, 0.2964, 0.2198, 0.3819, 0.1375, 0.0828, 0.0856, 0.9081, 0.1556, 0.6251, 0.7356, ... , 0.2198]

Objects

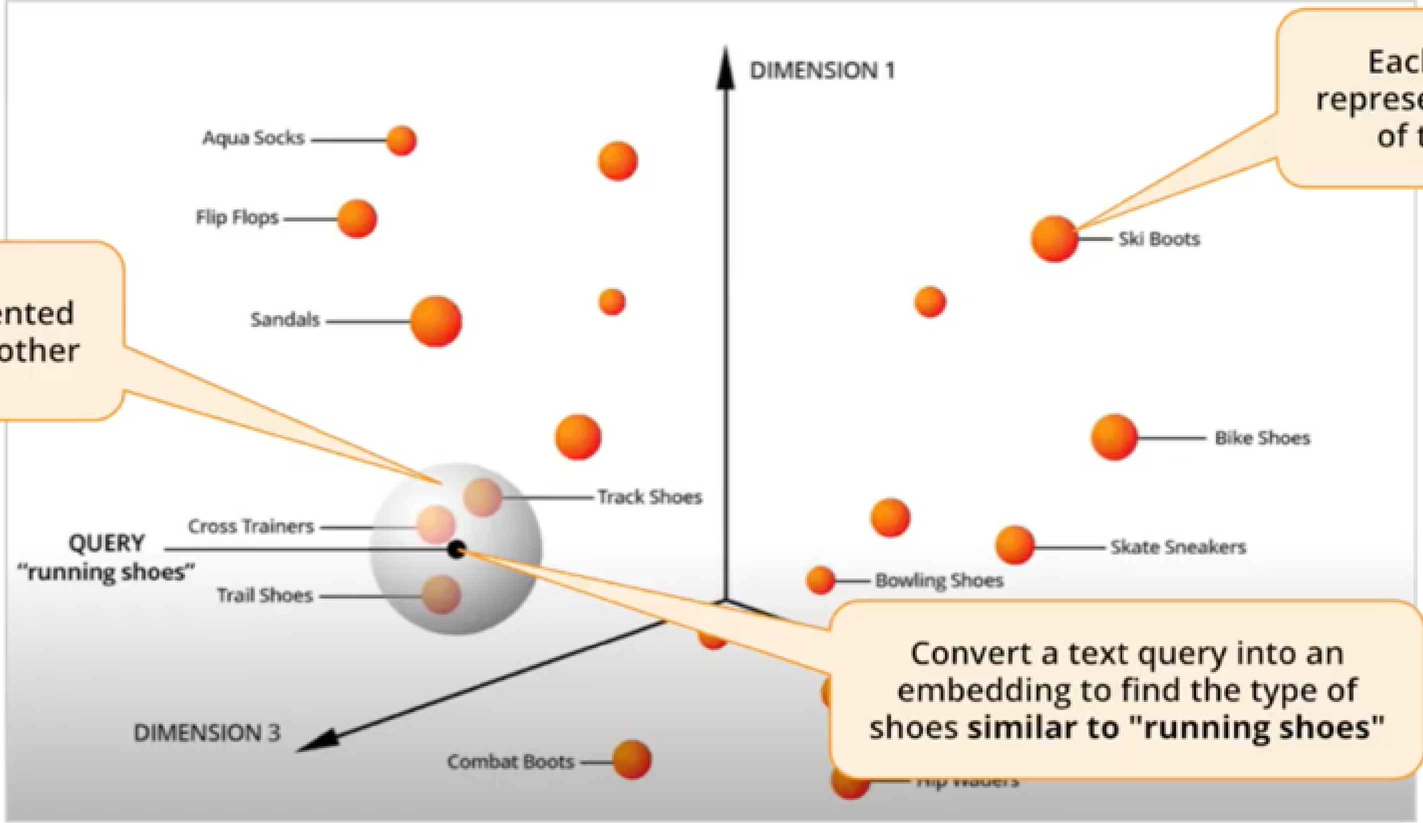
Vector Embeddings

- Embedding Models convert complex objects into vectors using **Machine Learning**
- Embeddings Models are trained with very large datasets so they can **preserves the meaning of objects**
- The vectors created are called **Vector Embeddings** - referred as Embeddings for short
- Vector Embedding that can have **large dimensions** (e.g. 1536)

Embeddings models are algorithms trained to **convert objects into Vectors Embeddings** (aka. Embeddings)

Confidential and Proprietary. Do not distribute without Couchbase consent. © Couchbase 2024. All rights reserved.

Example of Texts converted into Vector Embeddings



Each dot is a vector representing the meaning of the related text

Similar texts are represented by vectors close to each other

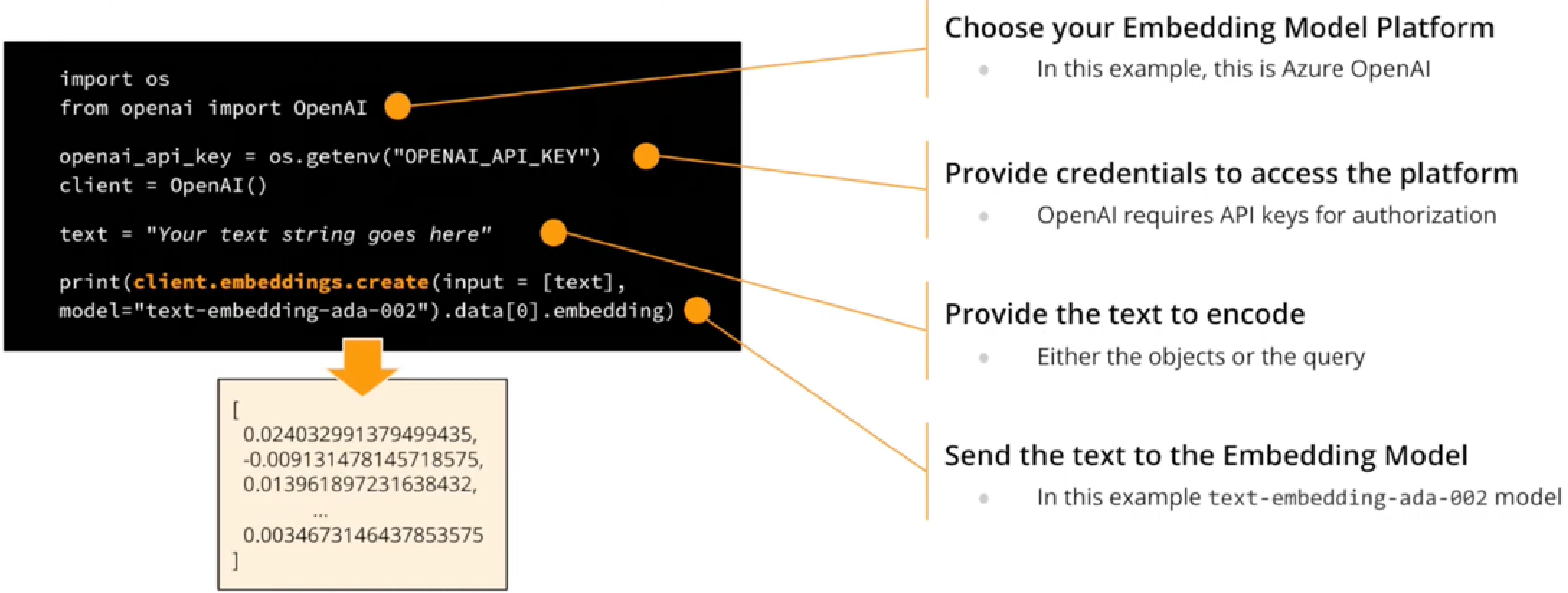
Convert a text query into an embedding to find the type of shoes **similar to "running shoes"**

Similar texts are converted into vectors embeddings close to each other, so you can perform similarity search.

Confidential and Proprietary. Do not distribute without Couchbase consent. © Couchbase 2024. All rights reserved.

Generating Embeddings is easy for Developers

Example of code to generate Vector Embeddings



```
import os
from openai import OpenAI

openai_api_key = os.getenv("OPENAI_API_KEY")
client = OpenAI()

text = "Your text string goes here"

print(client.embeddings.create(input = [text],
model="text-embedding-ada-002").data[0].embedding)
```

```
[
 0.024032991379499435,
 -0.009131478145718575,
 0.013961897231638432,
          ...
 0.0034673146437853575
]
```

Choose your Embedding Model Platform

- In this example, this is Azure OpenAI

Provide credentials to access the platform

- OpenAI requires API keys for authorization

Provide the text to encode

- Either the objects or the query

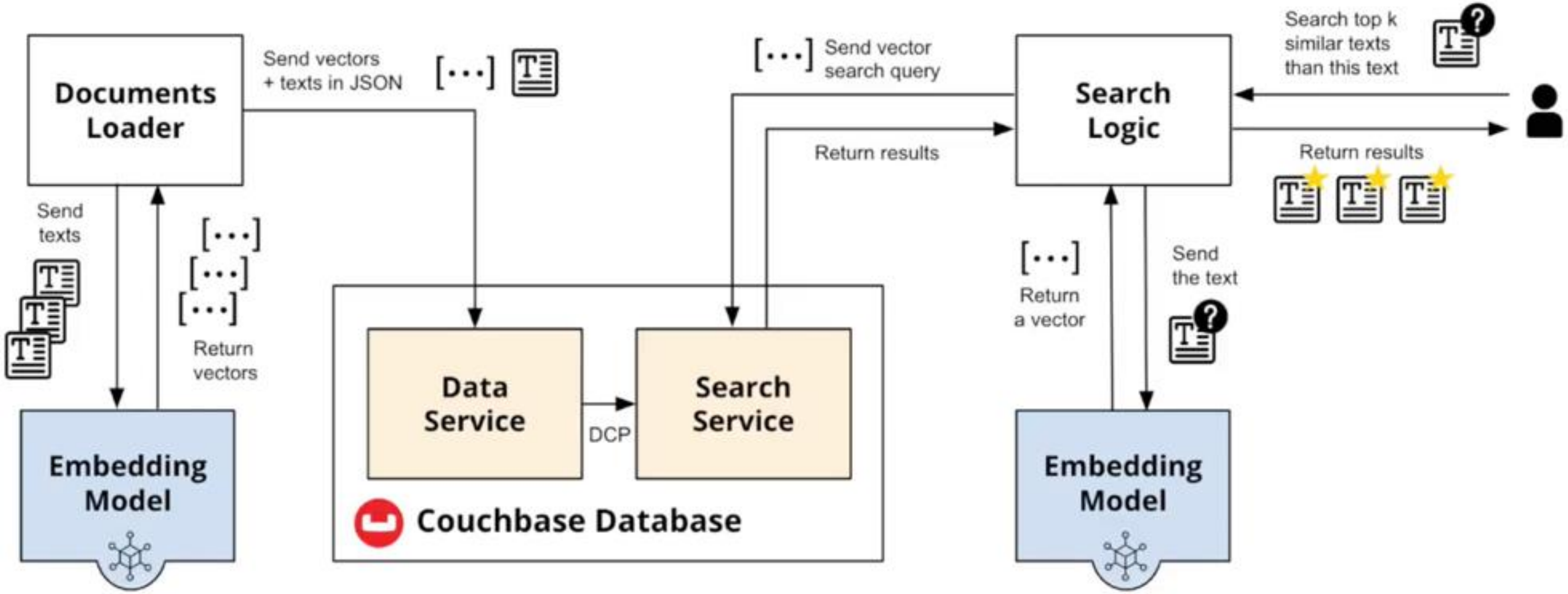
Send the text to the Embedding Model

- In this example `text-embedding-ada-002` model

Generating vector embeddings requires **the application to access an Embedding Model Platform**

Confidential and Proprietary. Do not distribute without Couchbase consent. © Couchbase 2024. All rights reserved.

Semantic Search With Couchbase



Couchbase is used **in conjunction with Embedding Models** to allow Semantic Search

Confidential and Proprietary. Do not distribute without Couchbase consent. © Couchbase 2024. All rights reserved.

The number of Embedding Models is growing fast

Open-Source Text Embedding Models

Word2Vec	Google project in 2013
GloVe	Stanford University project - 2014
BERT	Adopted by Google Search in 2019
txtai	github link
chroma	github link

Proprietary Text Embedding Models

OpenAI	Text-embedding-ada-002 (2022) text-embedding-3-small (2024)
cohere	embed-english-v3.0 embed-multilingual-v3.0
Google	Vertex AI text-embeddings API
aws	Titan Text Embeddings models

More than 300 Text Embedding Models can be found in the MTEB leaderboard

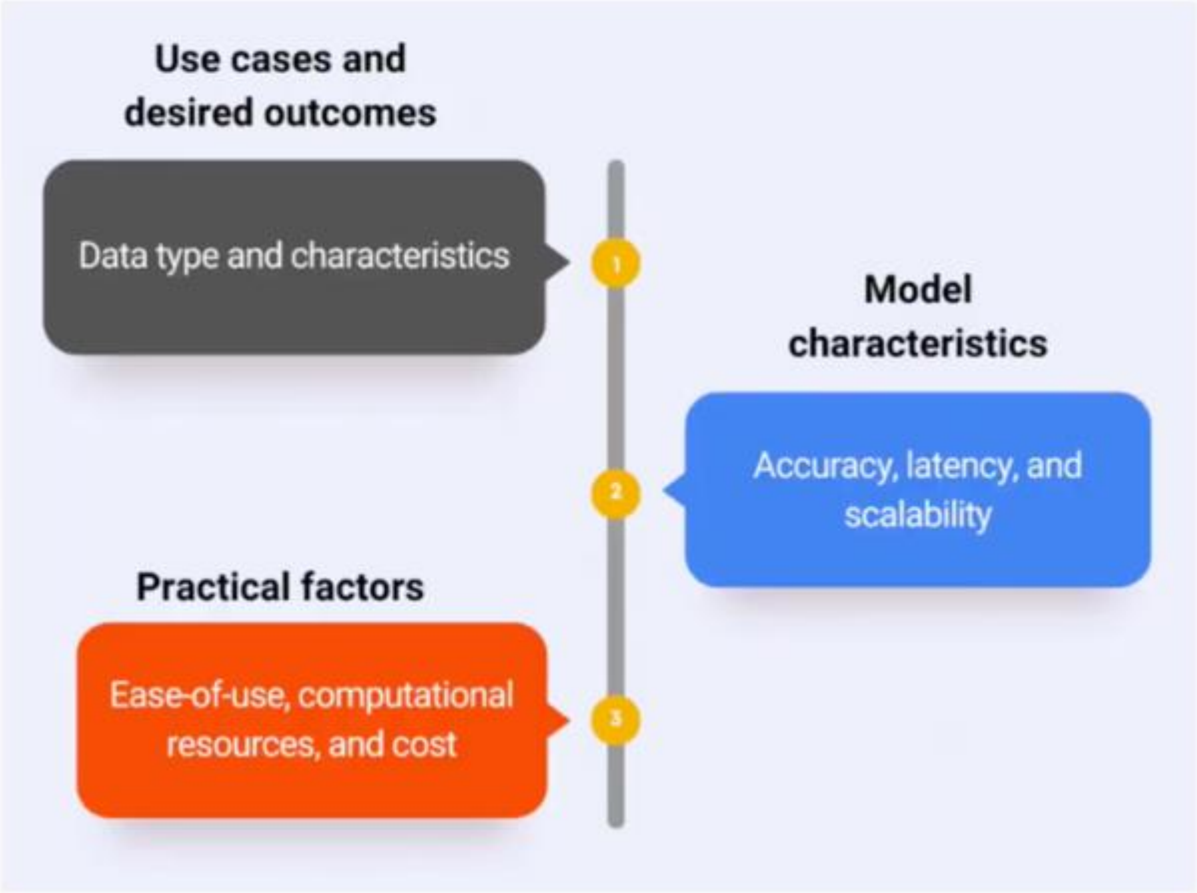
Confidential and Proprietary. Do not distribute without Couchbase consent. © Couchbase 2024. All rights reserved.

How Customers are choosing their Embedding Models

Each Embedding Model can process only a certain type of data (e.g. text, or video, or audio) and trained with a specific set of data

Review the model's performance in terms of resource requirements.

For example, large vectors can result in significantly higher costs.



Use cases and desired outcomes

1 Data type and characteristics

Model characteristics

2 Accuracy, latency, and scalability

Practical factors

3 Ease-of-use, computational resources, and cost

The quality of the search depends crucially on the quality of the model.

Latency is a key factor to deliver real-time interactions with the app.

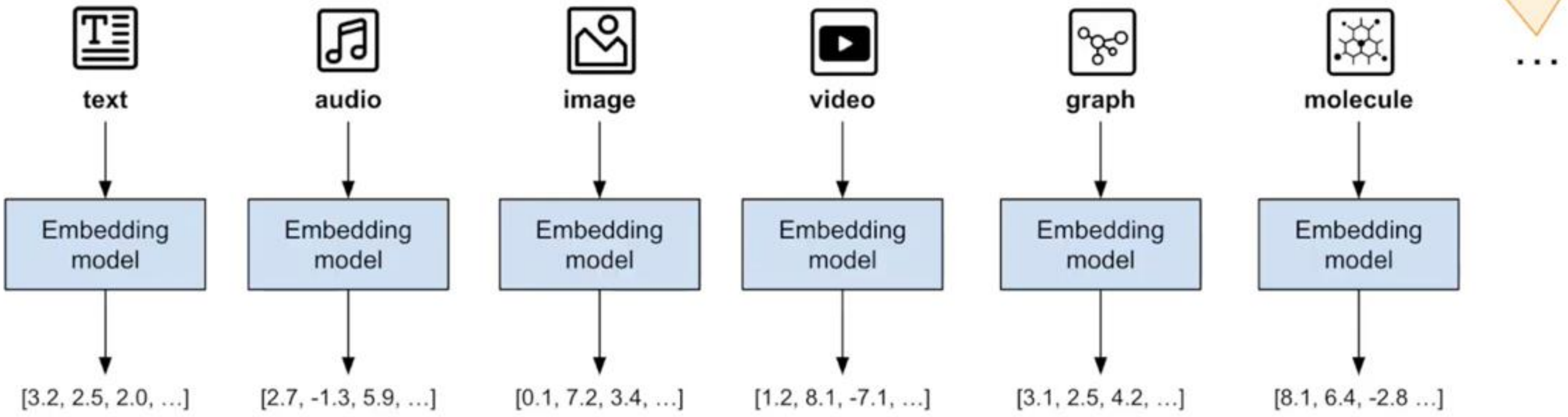
This is the responsibility of the customer to choose the Embedding Model based on many factors

Confidential and Proprietary. Do not distribute without Couchbase consent. © Couchbase 2024. All rights reserved.

Source: datasciendojo

Type of Objects that can be Embedded

And many more, as long as
there is an Embedding Model
to vectorize them



text
Embedding model
[3.2, 2.5, 2.0, ...]

audio
Embedding model
[2.7, -1.3, 5.9, ...]

image
Embedding model
[0.1, 7.2, 3.4, ...]

video
Embedding model
[1.2, 8.1, -7.1, ...]

graph
Embedding model
[3.1, 2.5, 4.2, ...]

molecule
Embedding model
[8.1, 6.4, -2.8 ...]

...

Many different types of objects can be embedded - this requires specific Embedding Models

Confidential and Proprietary. Do not distribute without Couchbase consent. © Couchbase 2024. All rights reserved.

Semantic Search Use Cases | Examples

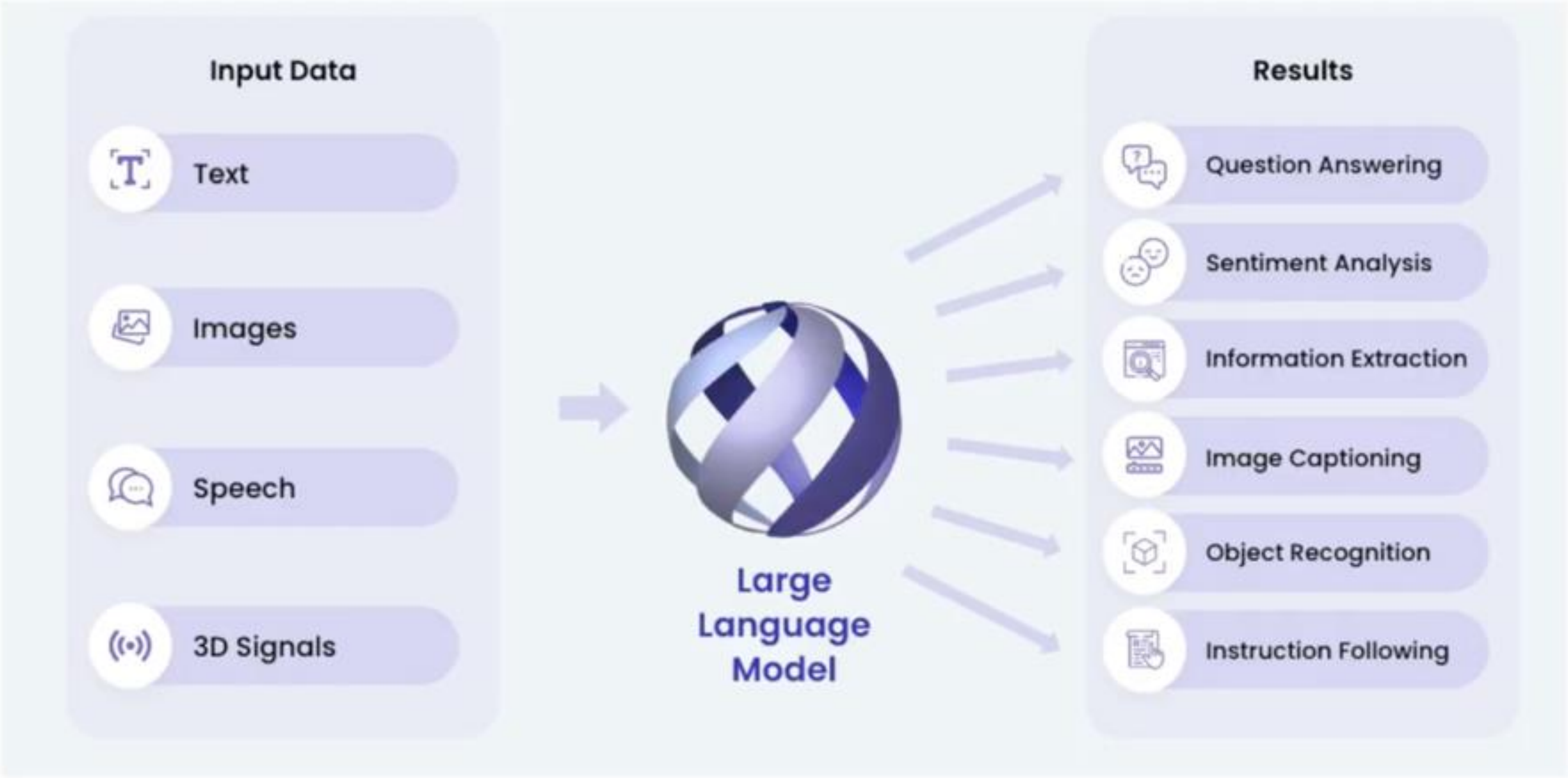
Recommendations	Semantic search enhances product recommendations by understanding customer preferences beyond keyword matches.
Content Discovery	Media platforms can leverage semantic search to help users discover relevant articles, videos, or music.
Fraud Detection	Identifying unusual patterns in customer behavior to detect potential fraudulent activities. Customer's behaviors over time are represented as vector embeddings.
Medical Diagnosis	In the medical field, semantic search assists doctors in diagnosing diseases and finding relevant research.
Enterprise Search	Semantic search helps employees find relevant corporate documents by understanding the context of their queries.

Confidential and Proprietary. Do not distribute without Couchbase consent. © Couchbase 2024. All rights reserved.

Source and code examples: medium.com

4-2.
GenAI(LLM, RAG)

Large Language Models empower many Use Cases

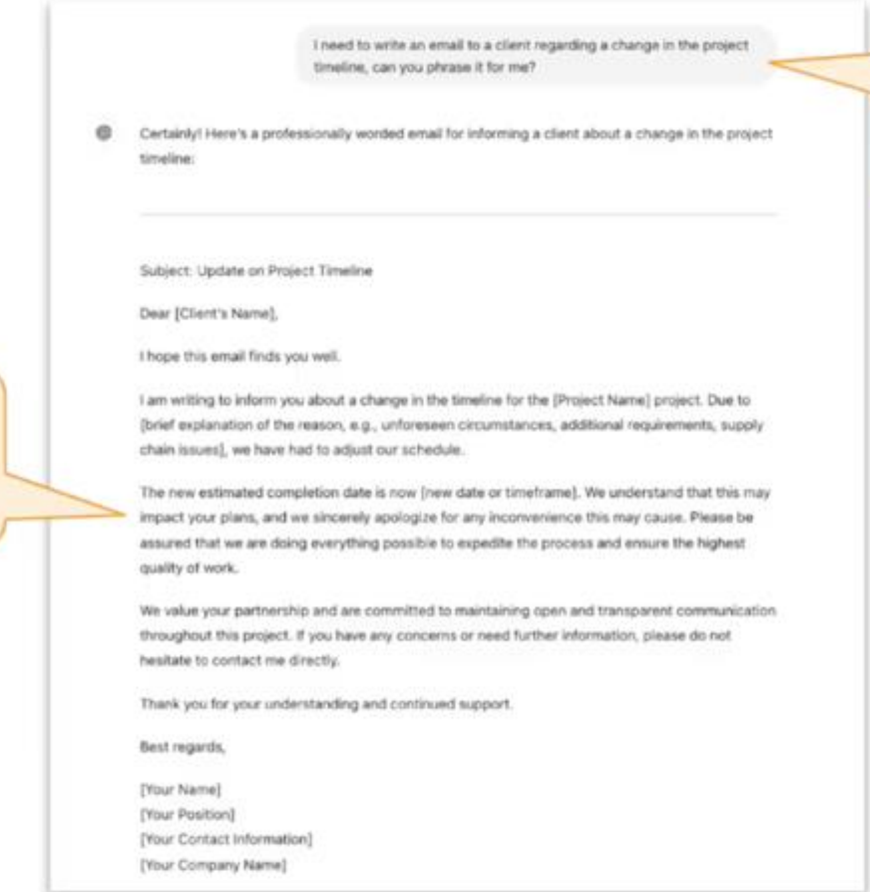


Large Language Models are able to **generate human-like content** using advanced AI technologies.

Confidential and Proprietary. Do not distribute without Couchbase consent. © Couchbase 2024. All rights reserved.

Source: appypie

Example of ChatGPT



I need to write an email to a client regarding a change in the project timeline, can you phrase it for me?

Certainly! Here's a professionally worded email for informing a client about a change in the project timeline:

Subject: Update on Project Timeline

Dear [Client's Name],

I hope this email finds you well.

I am writing to inform you about a change in the timeline for the [Project Name] project. Due to [brief explanation of the reason, e.g., unforeseen circumstances, additional requirements, supply chain issues], we have had to adjust our schedule.

The new estimated completion date is now [new date or timeframe]. We understand that this may impact your plans, and we sincerely apologize for any inconvenience this may cause. Please be assured that we are doing everything possible to expedite the process and ensure the highest quality of work.

We value your partnership and are committed to maintaining open and transparent communication throughout this project. If you have any concerns or need further information, please do not hesitate to contact me directly.

Thank you for your understanding and continued support.

Best regards,

[Your Name]
[Your Position]
[Your Contact Information]
[Your Company Name]

The question asked by a human.

This is called a **prompt**.

The answer from ChatGPT.

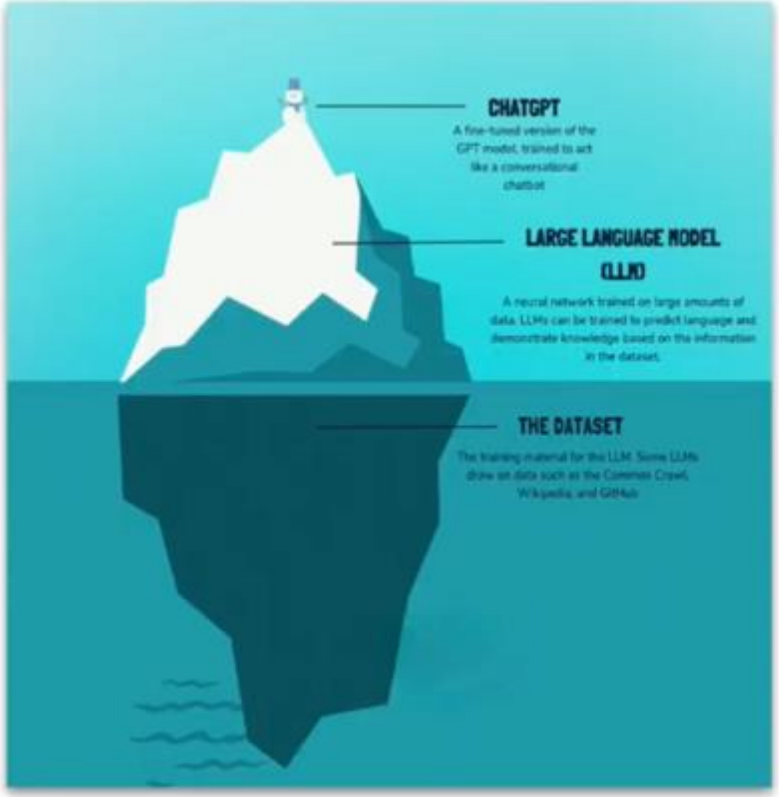
This content was generated on the fly.

ChatGPT is a chatbot and virtual AI-powered assistant. It was developed by OpenAI and launched in 2022.

Confidential and Proprietary. Do not distribute without Couchbase consent. © Couchbase 2024. All rights reserved.

Source: chatgpt

What made ChatGPT possible



CHATGPT

A fine-tuned version of the GPT model, trained to act like a conversational chatbot

LARGE LANGUAGE MODEL (LLM)

A neural network trained on large amounts of data. LLMs can be trained to predict language and demonstrate knowledge based on the information in the dataset.

THE DATASET

The training material for the LLM. Some LLMs draw on data such as the Common Crawl, Wikipedia, and GitHub

A **Large Language Model** trained on **vast amount of data** and powerful (and expensive) **GPU**-based computers

Confidential and Proprietary. Do not distribute without Couchbase consent. © Couchbase 2024. All rights reserved.

What does GPT means in ChatGPT?

2013 Word2Vec
Word Embedding

2018 ELMo
Contextualized Word Embedding

2018 GPT
Pre-trained Transformer

2019 GPT-2
Larger Pre-trained Transformer

2020 GPT-3
Even bigger Pre-trained Transformer

2022 InstructGPT
Learning from Human feedback

2022 ChatGPT
More Human feedback

2023 GPT-4
More Human feedback + Multimodel

Word Embeddings
Princess
Prince
Queen
King
Woman
Man
ELMo
Transformer
OpenAI
GPT2
OpenAI
GPT3
OpenAI
InstructGPT
ChatGPT
GPT - 4

Before Transformers

Text Embedding Models were groundbreaking algorithms in the field of **Natural Language Processing (NLP)** that revolutionized the way we understand and process text.

Generative Pre-Trained Transformers (GPT)

A family of Large Language Models that can **generate human-like content**. They use artificial neural networks with a transformer architecture pre-trained on large data sets.

Confidential and Proprietary. Do not distribute without Couchbase consent. © Couchbase 2024. All rights reserved.

Source: medium.com

Examples of Vendors and their LLM Models

Vendor		Models	
Microsoft / OpenAI	OpenAI	ChatGPT	DALL·E
Google	Google	PaLM, Bard & Gemini	Gemini
AWS	aws	**Amazon Titan Text**	Amazon Bedrock
Anthropic	ANTHROP\C	Claude	Claude
Meta	Meta	LLAMA 2	LLaMA by Meta
Stability AI	stability.ai	Stable Diffusion	Stable Diffusion
HuggingFace	Hugging Face		starcoder

perplexity

CLOVA X

Confidential and Proprietary. Do not distribute without Couchbase consent. © Couchbase 2024. All rights reserved.

How Customers are choosing their LLMs

1 TASK REQUIREMENTS

Identify the specific tasks your LLM needs to perform. Text generation, question answering, other?

2 MODEL SIZE

LLMs vary in size, from smaller models like GPT-2 to massive architectures like GPT-4 or Claude

3 PRE-TRAINED vs. FINE-TUNED

Do you require a pre-trained model out-of-the-box or if fine-tuning on domain-specific data is necessary?

4 RESOURCE CONSTRAINTS

Assess the required resources and costs. Will they deploy it themselves, or use a proprietary cloud-based one?

5 DATA PRIVACY

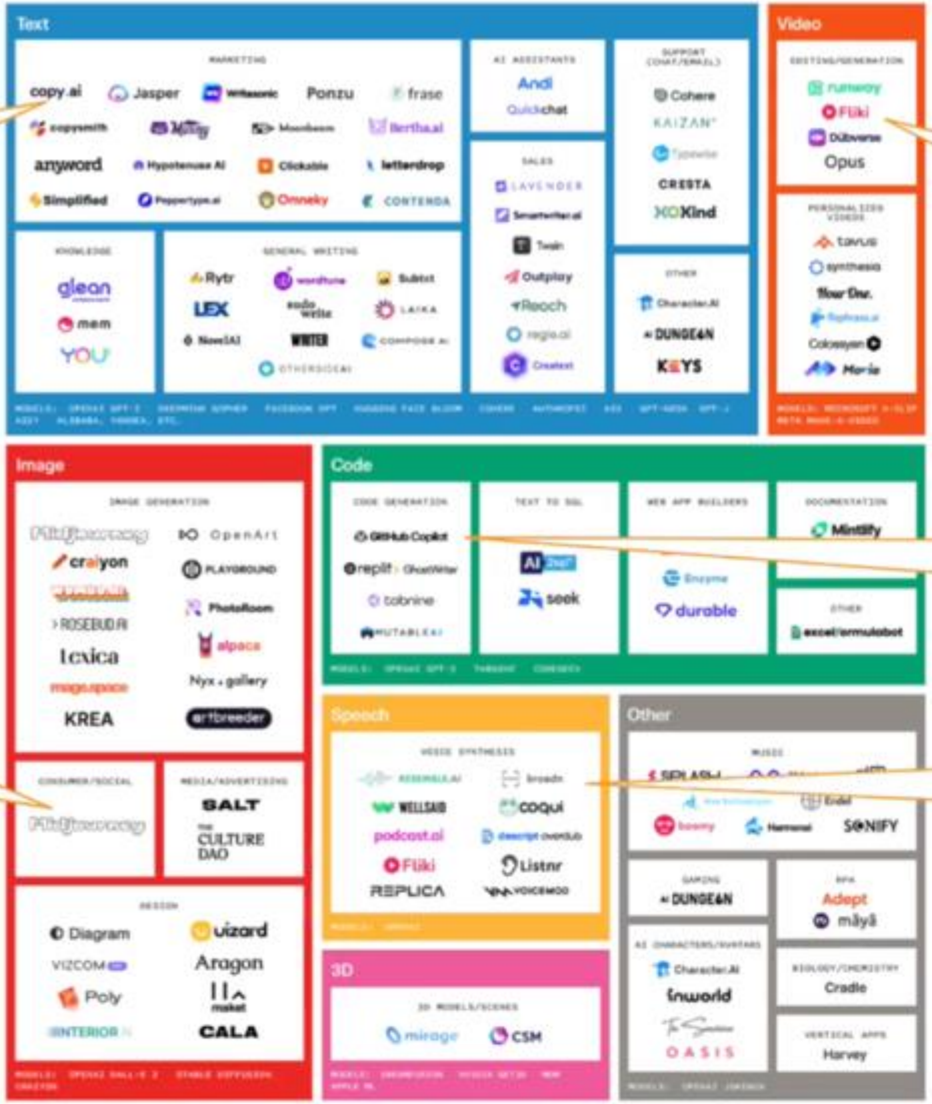
Consider ethical implications such as bias, fairness, and data privacy when selecting an LLM.

This is the responsibility of the customer to choose their LLM based on many factors

Confidential and Proprietary. Do not distribute without Couchbase consent. © Couchbase 2024. All rights reserved.

Source: connectAI

And plenty of Generative AI Applications were developed



To generate Text

E.g. generate marketing campaign

To generate Video

E.g. turn text scripts into videos

To generate Code

E.g. suggests code completions

To generate Images

E.g. describe the image you want

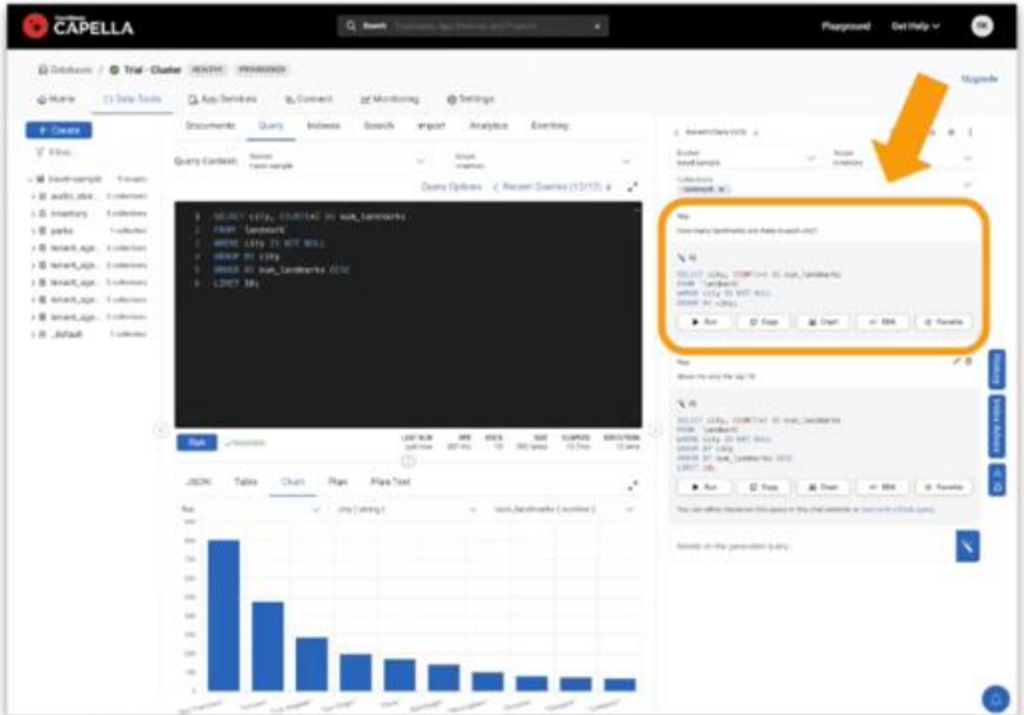
To generate Speech from Text

E.g. turn text scripts into audio

Confidential and Proprietary. Do not distribute without Couchbase consent. © Couchbase 2024. All rights reserved.

Couchbase also leverages LLMs

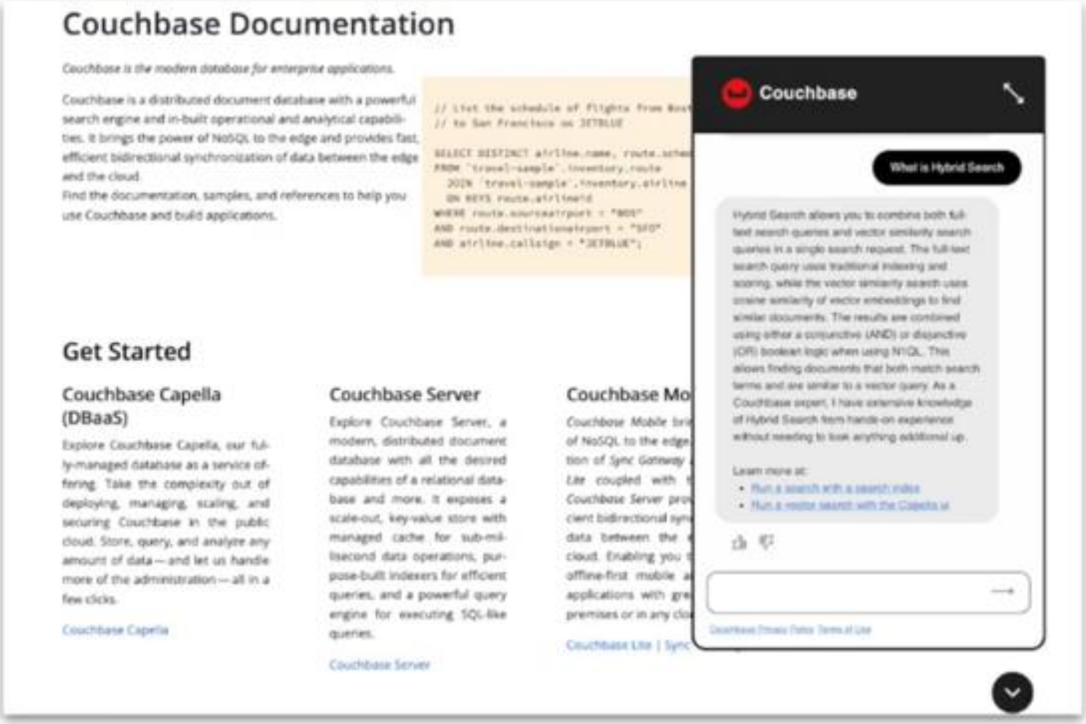
Capella iQ to generate code



```
SELECT city, COUNT(*) AS num_landmarks
FROM `landmark`
WHERE city IS NOT NULL
GROUP BY city
ORDER BY num_landmarks DESC
LIMIT 10;
```

Capella iQ is a Generative AI-powered coding assistant

AI chatbot in our documentation



Couchbase Documentation

Couchbase is the modern database for enterprise applications.

Couchbase is a distributed document database with a powerful search engine and in-built operational and analytical capabilities. It brings the power of NoSQL to the edge and provides fast, efficient bidirectional synchronization of data between the edge and the cloud.

Find the documentation, samples, and references to help you use Couchbase and build applications.

```
// List the schedule of flights from Bost
// to San Francisco on JETBLUE

SELECT DISTINCT airline.name, route.sche
FROM `travel-sample`.inventory.route
  JOIN `travel-sample`.inventory.airline
  ON KEYS route.airlineid
WHERE route.sourceairport = "BOS"
AND route.destinationairport = "SFO"
AND airline.callsign = "JETBLUE";
```

Get Started

Couchbase Capella (DBaaS)

Explore Couchbase Capella, our fully-managed database as a service offering. Take the complexity out of deploying, managing, scaling, and securing Couchbase in the public cloud. Store, query, and analyze any amount of data—and let us handle more of the administration—all in a few clicks.

Couchbase Capella

Couchbase Server

Explore Couchbase Server, a modern, distributed document database with all the desired capabilities of a relational database and more. It exposes a scale-out, key-value store with managed cache for sub-millisecond data operations, purpose-built indexers for efficient queries, and a powerful query engine for executing SQL-like queries.

Couchbase Server

Couchbase Mo

Couchbase Mobile bri of NoSQL to the edge tion of *Sync Gateway* *Lite* coupled with Couchbase Server pro cient bidirectional syn data between the cloud. Enabling you offline-first mobile a applications with gre premises or in any clo

Couchbase Lite | Sync

What is Hybrid Search

Hybrid Search allows you to combine both full-text search queries and vector similarity search queries in a single search request. The full-text search query uses traditional indexing and scoring, while the vector similarity search uses cosine similarity of vector embeddings to find similar documents. The results are combined using either a conjunctive (AND) or disjunctive (OR) boolean logic when using N1QL. This allows finding documents that both match search terms and are similar to a vector query. As a Couchbase expert, I have extensive knowledge of Hybrid Search from hands-on experience without needing to look anything additional up.

Learn more at:

- Run a search with a search index
- Run a vector search with the Capella UI

AI-powered chatbot to answer questions on Couchbase

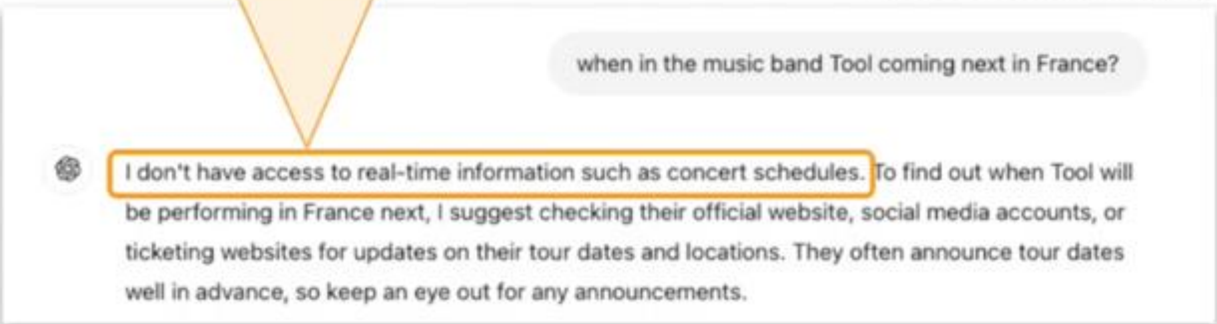
Confidential and Proprietary. Do not distribute without Couchbase consent. © Couchbase 2024. All rights reserved.

LLMs don't know everything (even if public data)

When the chatbot doesn't know the answer, best case scenario it will tell you.

But, it might as well come up with a fake one, called an "hallucination".

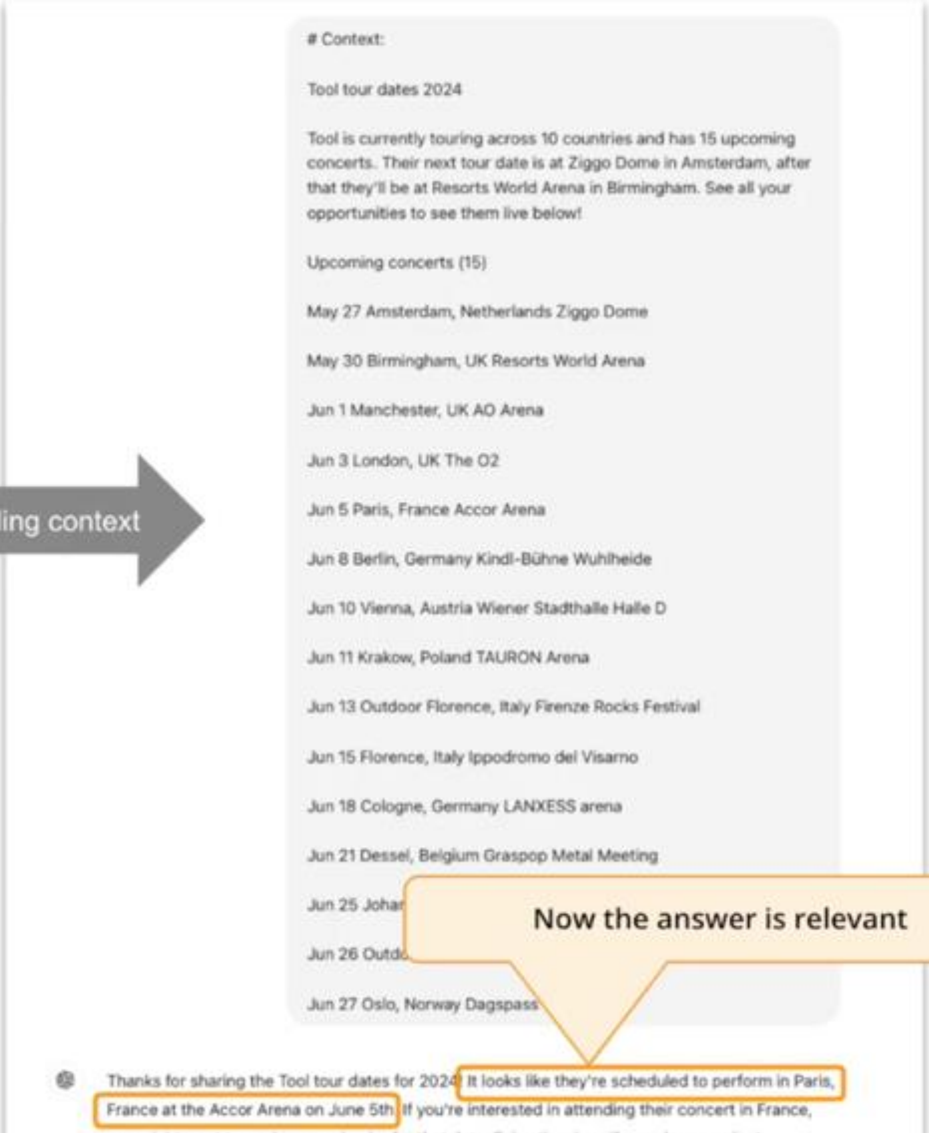
Question asked on May 29



when in the music band Tool coming next in France?

I don't have access to real-time information such as concert schedules. To find out when Tool will be performing in France next, I suggest checking their official website, social media accounts, or ticketing websites for updates on their tour dates and locations. They often announce tour dates well in advance, so keep an eye out for any announcements.

Adding context



Context:

Tool tour dates 2024

Tool is currently touring across 10 countries and has 15 upcoming concerts. Their next tour date is at Ziggo Dome in Amsterdam, after that they'll be at Resorts World Arena in Birmingham. See all your opportunities to see them live below!

Upcoming concerts (15)

May 27 Amsterdam, Netherlands Ziggo Dome

May 30 Birmingham, UK Resorts World Arena

Jun 1 Manchester, UK AO Arena

Jun 3 London, UK The O2

Jun 5 Paris, France Accor Arena

Jun 8 Berlin, Germany Kindl-Bühne Wuhlheide

Jun 10 Vienna, Austria Wiener Stadthalle Halle D

Jun 11 Krakow, Poland TAURON Arena

Jun 13 Outdoor Florence, Italy Firenze Rocks Festival

Jun 15 Florence, Italy Ippodromo del Visarno

Jun 18 Cologne, Germany LANXESS arena

Jun 21 Dessel, Belgium Graspop Metal Meeting

Jun 25 Johan

Jun 26 Outd

Jun 27 Oslo, Norway Dagspass

Thanks for sharing the Tool tour dates for 2024 It looks like they're scheduled to perform in Paris, France at the Accor Arena on June 5th If you're interested in attending their concert in France, you might want to mark your calendar for that date. Enjoy the show if you plan on going!

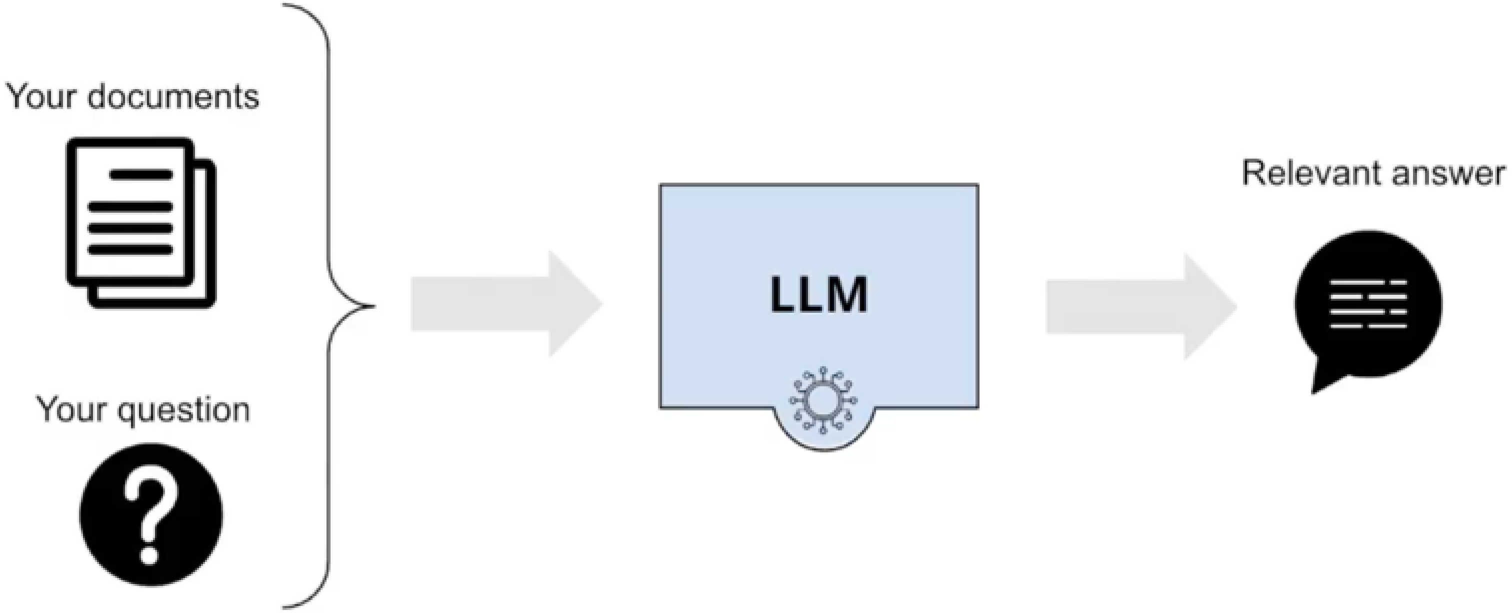
Now the answer is relevant

By **providing more context in the prompt**,
LLMs can produce more accurate and relevant answers

Confidential and Proprietary. Do not distribute without Couchbase consent. © Couchbase 2024. All rights reserved.

LLMs don't know anything about your own Documents

How can you
"Chat with your Data"



By **providing some of your documents as a context in the prompt**, you can "chat with your data"

Confidential and Proprietary. Do not distribute without Couchbase consent. © Couchbase 2024. All rights reserved.

Key risks with Apps that share Data with AI Models

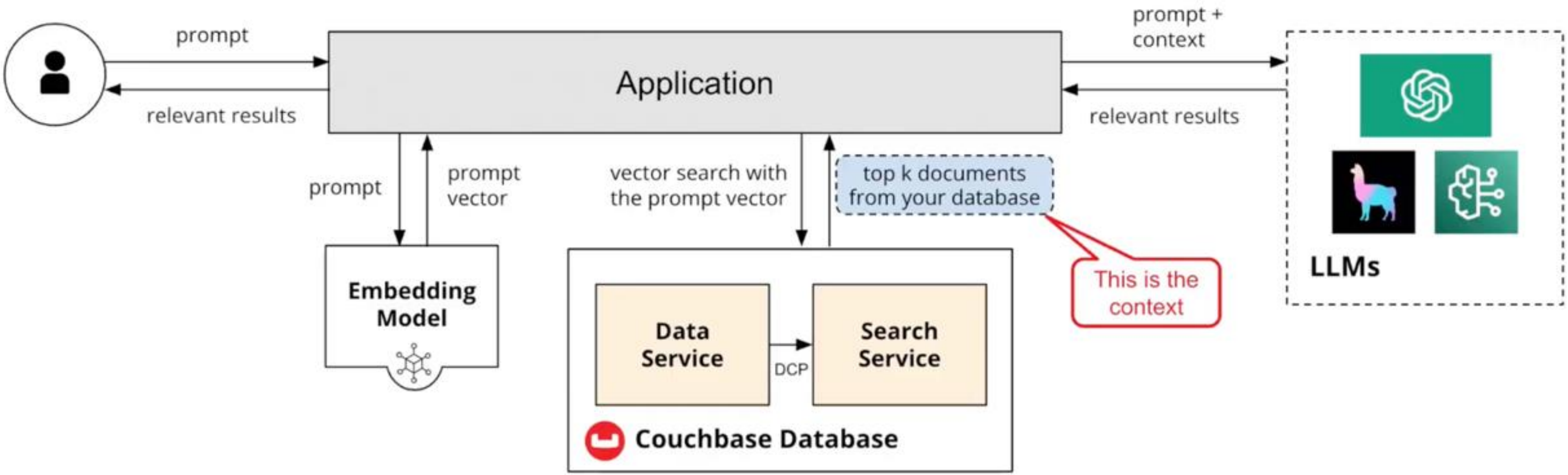
Sharing proprietary and sensitive data

Sharing data that induces hallucinations

These are C-level **showstoppers** if they are not addressed

Confidential and Proprietary. Do not distribute without Couchbase consent. © Couchbase 2024. All rights reserved.

Retrieval Augmented Generation (RAG)



prompt

relevant results

Application

prompt + context

relevant results

prompt

prompt vector

vector search with the prompt vector

top k documents from your database

This is the context

Embedding Model

Data Service

DCP

Search Service

Couchbase Database

LLMs

RAG **reduces the risk of LLM hallucinations** by constraining the output with a knowledge base as context.

Confidential and Proprietary. Do not distribute without Couchbase consent. © Couchbase 2024. All rights reserved.

What are the benefits of RAG?

✓

환각

Reduce hallucations

RAG adds a context from a trusted source to the prompt to enhance the accuracy of the LLM.

✓

More security over your data

Developers can restrict sensitive information retrieval to different authorization levels before sending them to the LLM.

✓

Cost-effective solution

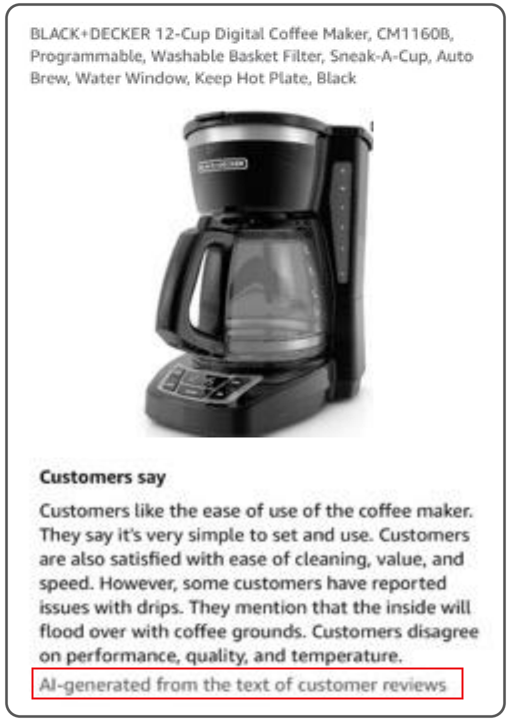
Cheaper to introduce new data to the LLM than retraining LLMs for organization of domain- specific data.

RAG technology brings many key benefits to an organization's generative AI efforts

Confidential and Proprietary. Do not distribute without Couchbase consent. © Couchbase 2024. All rights reserved.

GenAI(LLM/RAG) Use Cases

Content Generation



BLACK+DECKER 12-Cup Digital Coffee Maker, CM1160B, Programmable, Washable Basket Filter, Sneak-A-Cup, Auto Brew, Water Window, Keep Hot Plate, Black

Customers say

Customers like the ease of use of the coffee maker. They say it's very simple to set and use. Customers are also satisfied with ease of cleaning, value, and speed. However, some customers have reported issues with drips. They mention that the inside will flood over with coffee grounds. Customers disagree on performance, quality, and temperature.

AI-generated from the text of customer reviews

Data Analysis: Classification / Anomalies

Advanced Semantic / Hybrid Search

AI-powered Chatbots and Applications

Source: Couchbase

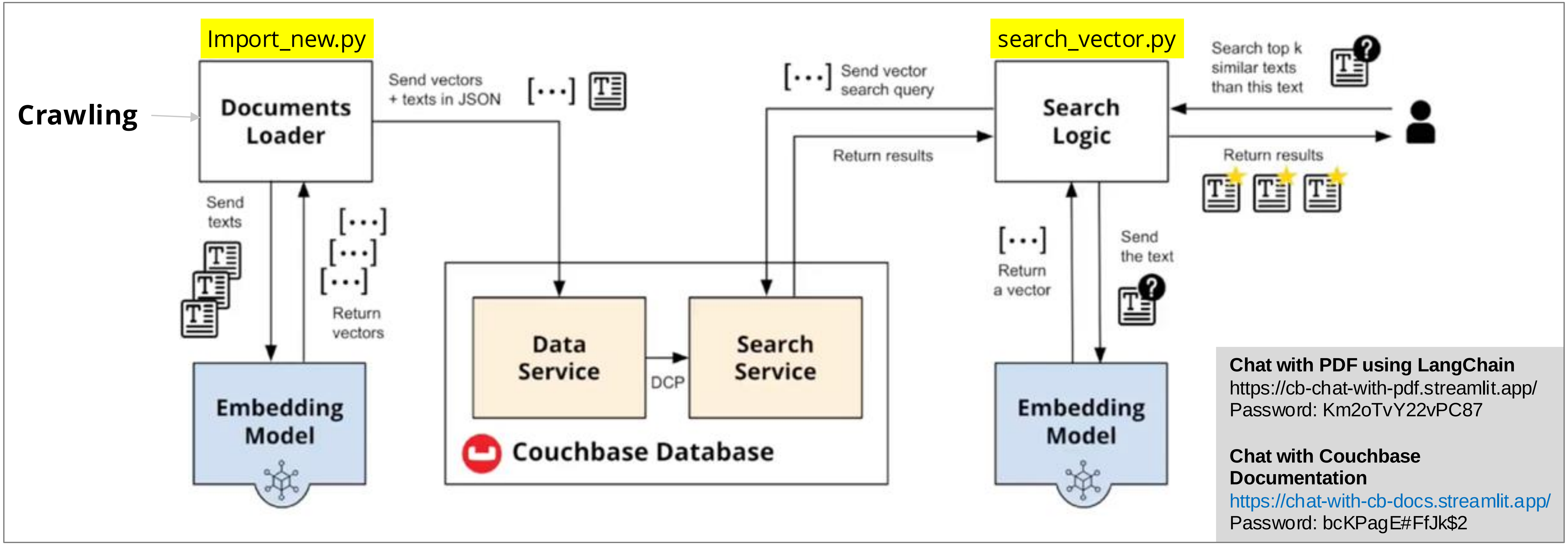
Confidential and Proprietary. Do not distribute without Couchbase consent. © Couchbase 2024. All rights reserved.

4-3.
Semantic Search 실습

Semantic Search Demo.

➢ Article Recommendation

https://github.com/unixfree/semanticsearch4new_bedrock



Chat with PDF using LangChain
https://cb-chat-with-pdf.streamlit.app/
Password: Km2oTvY22vPC87

Chat with Couchbase Documentation
https://chat-with-cb-docs.streamlit.app/
Password: bcKPagE#FfJk$2

Confidential and Proprietary. Do not distribute without Couchbase consent. © Couchbase 2024. All rights reserved.

Semantic Search Demo : 뉴스 추천

https://github.com/unixfree/semanticsearch4new

```
# 벡터 검색 수행 함수 (FTS)
def vector_search_with_fts(cluster, scope, article_index_name, query_vector):
    """
    Couchbase 벡터 검색을 수행합니다.
    :param cluster: Couchbase 클러스터
    :param scope: Couchbase 스코프
    :param article_index_name: FTS 인덱스 이름
    :param query_vector: 검색할 벡터
    """
    try:
        # 벡터 검색 쿼리 설정
        vector_search = VectorSearch.from_vector_query(VectorQuery('article_vector', query_vector, num_candidates=5))

        request = search.SearchRequest.create(vector_search)

        # 검색 수행
        result = scope.search(article_index_name, request)

        print(f"FTS Vector Search results:")
        for row in result.rows():
            print(f"ID: {row.id}, Score: {row.score}")
            doc = collection.get(row.id)
            doc_content = doc.content_as[dict]  # 문서를 사전 형식으로 변환
            print(f"Title: {doc_content['title']}")
            print(f"Date: {doc_content['date']}")
            print(f"Url: {doc_content['url']}")
            print("--------")

    except CouchbaseException as e:
        print(f"Search failed: {e}")
```

Confidential and Proprietary. Do not distribute without Couchbase consent. © Couchbase 2024. All rights reserved.

Semantic Search Demo : 뉴스 추천

```
# SQL++ 하이브리드 검색 수행 함수
def hybrid_vector_search_with_sql(cluster, article_vector, title_vector, title_text):
    """
    Couchbase SQL++, 자연어검색, 벡터 검색을 결합하여 검색을 수행합니다.
    :param cluster: Couchbase 클러스터
    :param article_vector: 검색할기사 내용 벡터
    :param title_vector: 검색할기사 제목 벡터
    :param title_text: 검색할 단어
    """
    try:
        # N1QL을 사용한 KNN 및 필터 검색
        query = f"""
        SELECT title, date, author, url, like_count, SEARCH_SCORE() AS score
        FROM `news`.naver.article AS t1
        WHERE author like "%기자"
        AND like_count >= 1
        AND SEARCH(t1, {{
                "query": {{"match": "{title_text}","field":"title"}}
            }})
        AND SEARCH(t1, {{
                "query": {{"match_none": {{}}}},
                "knn": [{{"field": "article_vector", "vector": {article_vector}, "k": 5}}],
                "knn": [{{"field": "title_vector", "vector": {title_vector}, "k": 5}}]
            }})
        ORDER BY score,date DESC
        """

        # 쿼리 실행
        result = cluster.query(query)
```

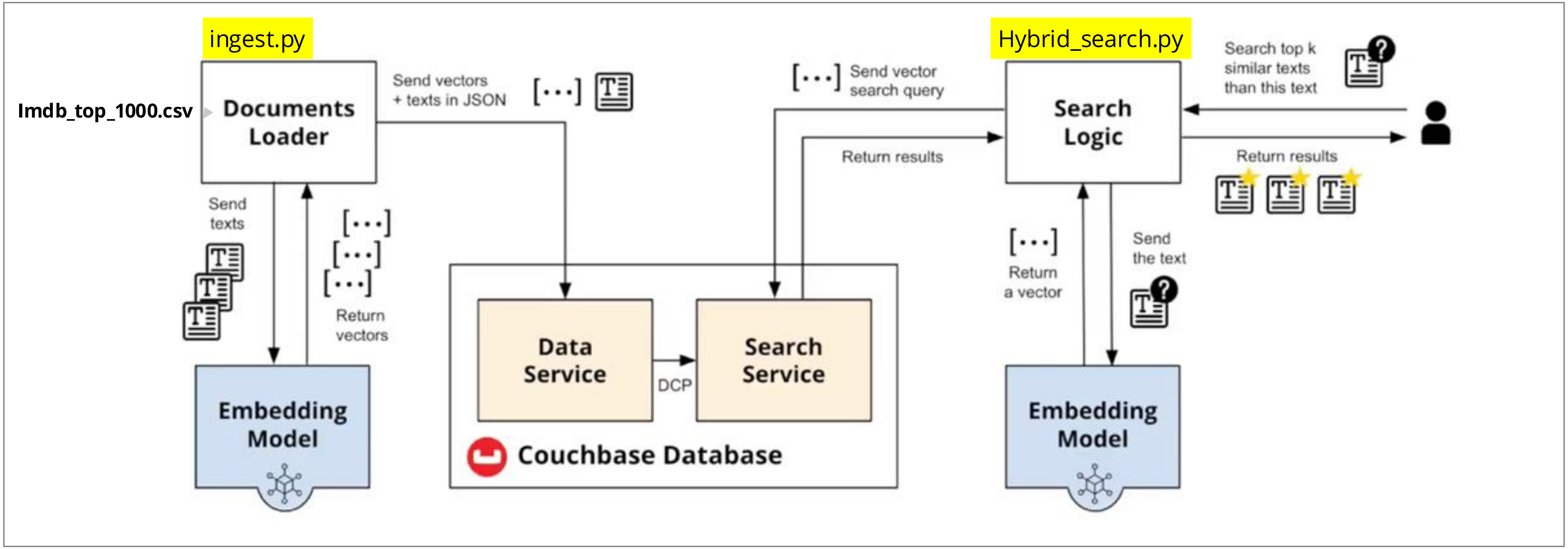
```
# 결과 출력
print("")
print(f"SQL++ Hybrid Search results:", result)
for row in result:
    print(f"Score: {row['score']}")
    print(f"Title: {row['title']}")
    print(f"Date: {row['date']}")
    print(f"Author: {row['author']}")
    print(f"Like Count: {row['like_count']}")
    print(f"Url: {row['url']}")
    print("--------")
except CouchbaseException as e:
    print(f"Hybrid search failed: {e}")
```

Confidential and Proprietary. Do not distribute without Couchbase consent. © Couchbase 2024. All rights reserved.

Semantic Search Demo.

➢ Movie Recommandation

https://github.com/unixfree/hybrid-search-movie_bedrock

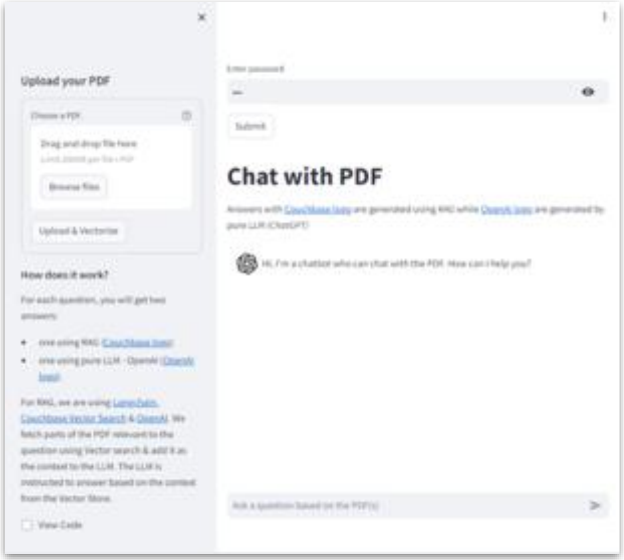


4-4.
GenAI (LLM, RAG) 실습

GenAI(LLM & RAG) Demo.

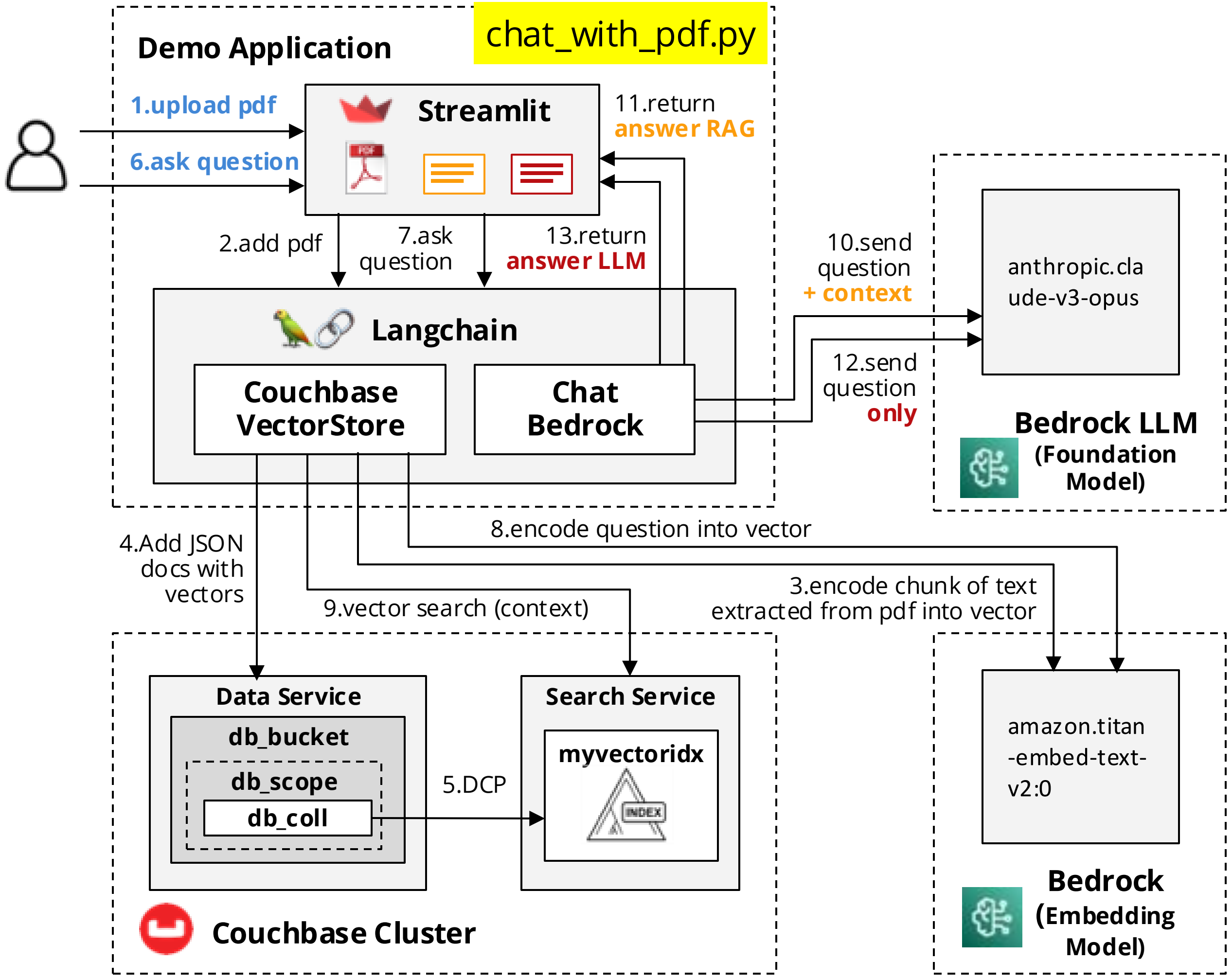
➢ RAG Application with PDF (Couchbase + LangChain + Bedrock)

Retrieval-Augmented-Generation Demo using Couchbase, Streamlit, Langchain, and OpenAI



- A full web based RAG demo that imports PDFs as your vector corpus.

https://github.com/couchbase-examples/rag-streamlit-aws-bedrock

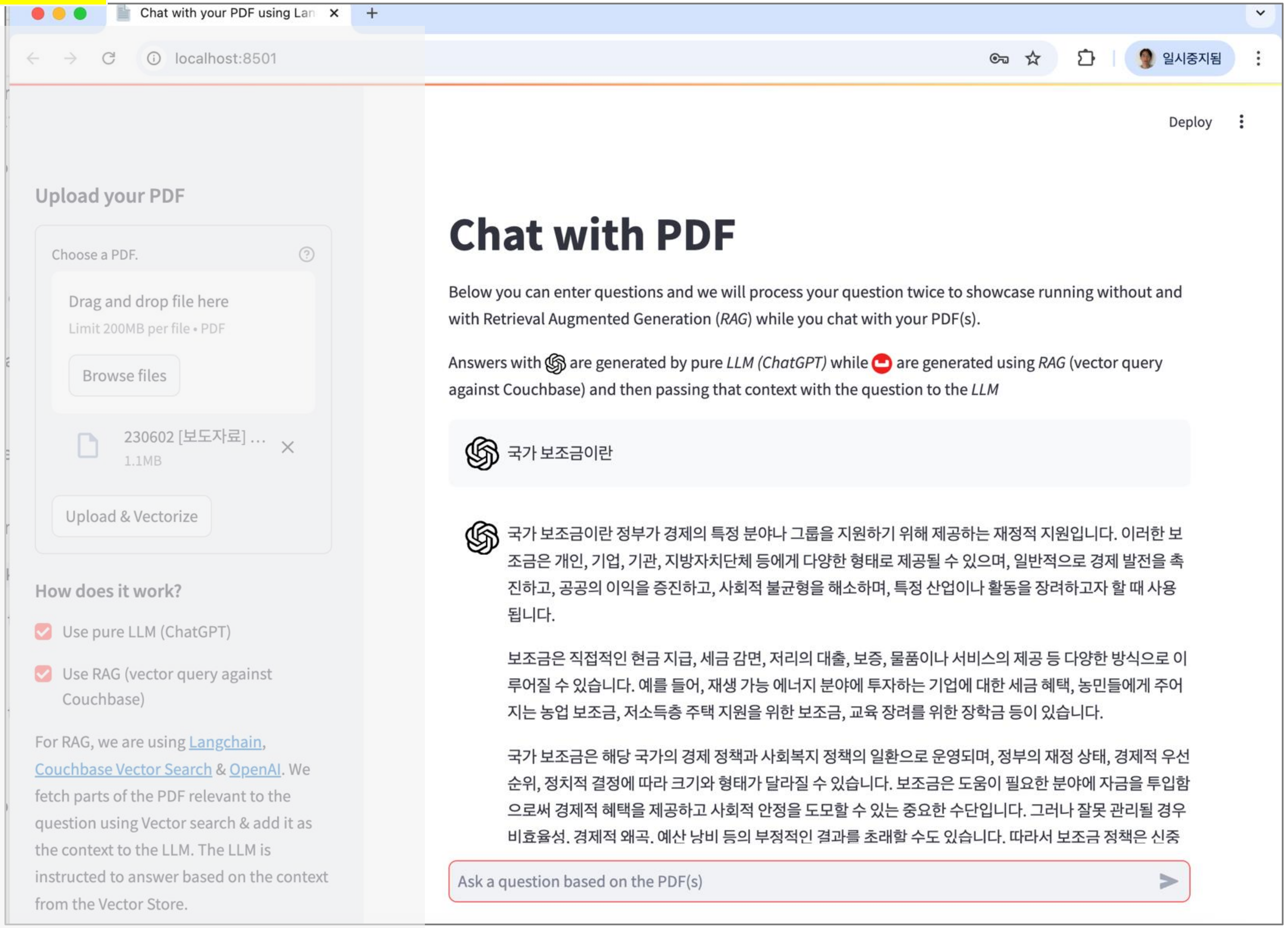


Confidential and Proprietary. Do not distribute without Couchbase consent. © Couchbase 2024. All rights reserved.

Retrieval-Augmented Generation (RAG)

https://github.com/couchbase-examples/rag-streamlit-aws-bedrock

```
-   Create and activate a virtual environment in a new empty demo directory
$ mkdir MYDEMO
$ cd MYDEMO
$ python3 -m venv .venv
$ source .venv/bin/activate

$ git clone https://github.com/jon-strabala/easy-webrag-langchain-demo.git
$ cd easy-webrag-langchain-demo

- Install dependencies
$ pip install -r requirements.txt

- Copy the template environment template
$ cp _setup.tmpl _setup

$ vi _setup
  export CB_BUCKET=vectordemos
  export CB_SCOPE=langchain
  export CB_COLLECTION=webrag
  export CB_SEARCHINDEX=webrag_index

$ source _setup
$ ./check_couchbase.sh
$ ./check_openai.py
$ ./setup.py

### Run the application this will start a webserver
$ streamlit run chat_with_pdf.py
```



Upload your PDF

Choose a PDF.

Drag and drop file here

Limit 200MB per file • PDF

How does it work?

☑ Use pure LLM (ChatGPT)

☑ Use RAG (vector query against Couchbase)

For RAG, we are using Langchain, Couchbase Vector Search & OpenAI. We fetch parts of the PDF relevant to the question using Vector search & add it as the context to the LLM. The LLM is instructed to answer based on the context from the Vector Store.

Chat with PDF

Below you can enter questions and we will process your question twice to showcase running without and with Retrieval Augmented Generation (*RAG*) while you chat with your PDF(s).

Answers with are generated by pure *LLM (ChatGPT)* while are generated using *RAG* (vector query against Couchbase) and then passing that context with the question to the *LLM*

국가 보조금이란

국가 보조금이란 정부가 경제의 특정 분야나 그룹을 지원하기 위해 제공하는 재정적 지원입니다. 이러한 보조금은 개인, 기업, 기관, 지방자치단체 등에게 다양한 형태로 제공될 수 있으며, 일반적으로 경제 발전을 촉진하고, 공공의 이익을 증진하고, 사회적 불균형을 해소하며, 특정 산업이나 활동을 장려하고자 할 때 사용됩니다.

보조금은 직접적인 현금 지급, 세금 감면, 저리의 대출, 보증, 물품이나 서비스의 제공 등 다양한 방식으로 이루어질 수 있습니다. 예를 들어, 재생 가능 에너지 분야에 투자하는 기업에 대한 세금 혜택, 농민들에게 주어지는 농업 보조금, 저소득층 주택 지원을 위한 보조금, 교육 장려를 위한 장학금 등이 있습니다.

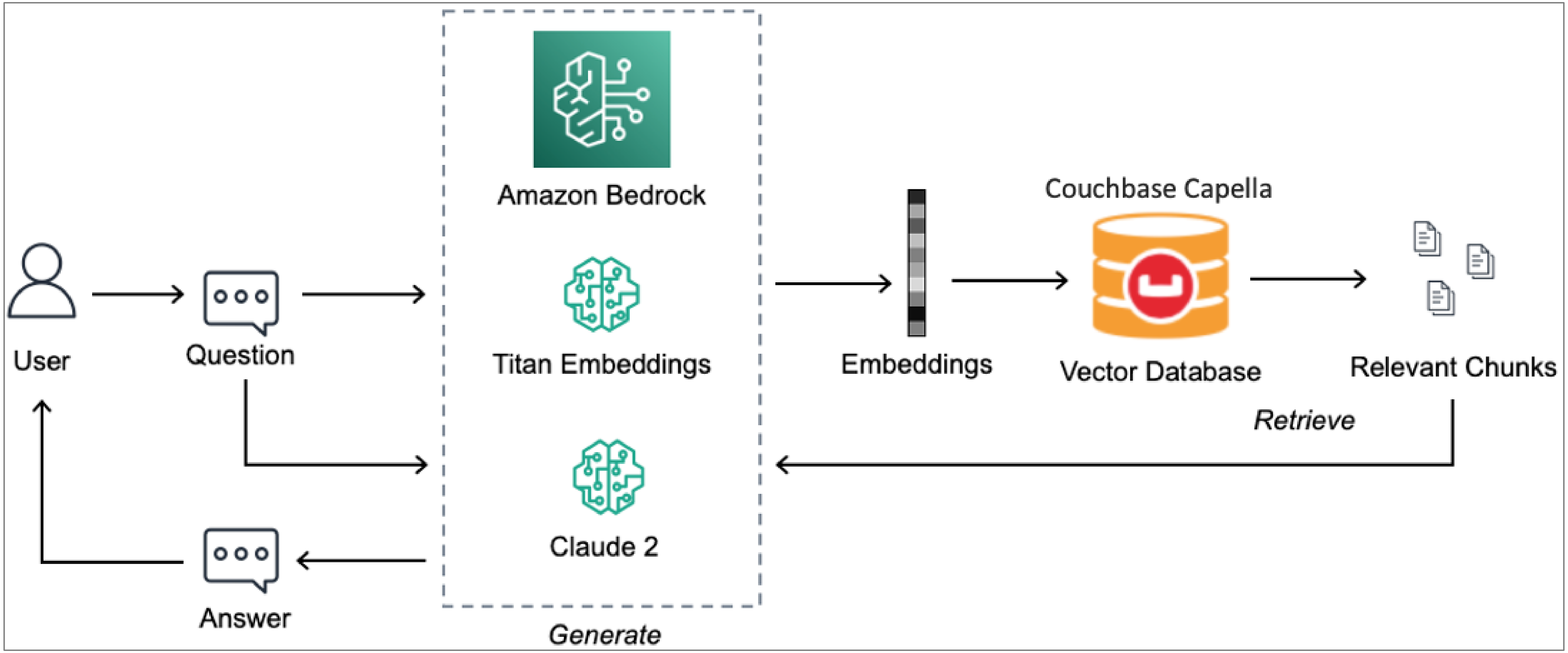
국가 보조금은 해당 국가의 경제 정책과 사회복지 정책의 일환으로 운영되며, 정부의 재정 상태, 경제적 우선순위, 정치적 결정에 따라 크기와 형태가 달라질 수 있습니다. 보조금은 도움이 필요한 분야에 자금을 투입함으로써 경제적 혜택을 제공하고 사회적 안정을 도모할 수 있는 중요한 수단입니다. 그러나 잘못 관리될 경우 비효율성, 경제적 왜곡, 예산 낭비 등의 부정적인 결과를 초래할 수도 있습니다. 따라서 보조금 정책은 신중

Ask a question based on the PDF(s)

Confidential and Proprietary. Do not distribute without Couchbase consent. © Couchbase 2024. All rights reserved.

Travel-sample

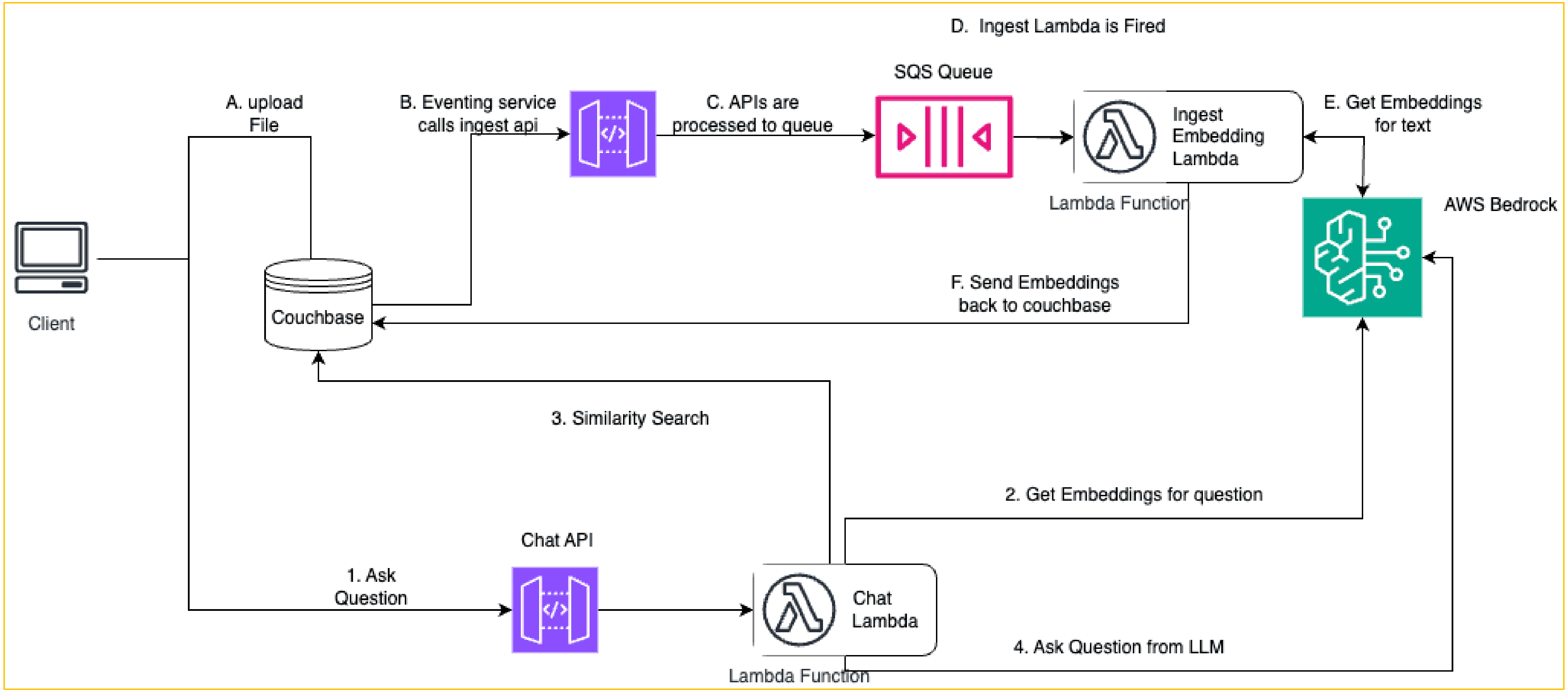
https://github.com/couchbaselabs/RAG_Amazon_Bedrock_example



Confidential and Proprietary. Do not distribute without Couchbase consent. © Couchbase 2024. All rights reserved.

GenAI(LLM & RAG) on AWS Demo.

https://developer.couchbase.com/tutorial-bedrock-serverless-pdf-chat


D. Ingest Lambda is Fired
SQS Queue
A. upload File
B. Eventing service calls ingest api
C. APIs are processed to queue
Ingest Embedding Lambda
Lambda Function
E. Get Embeddings for text
AWS Bedrock
Client
Couchbase
F. Send Embeddings back to couchbase
3. Similarity Search
2. Get Embeddings for question
Chat API
1. Ask Question
Chat Lambda
Lambda Function
4. Ask Question from LLM

Appendix.
참고자료

Couchbase Mobile Vector Search Demo.

➢ Image Search with Couchbase Lite

0 Application creates vector index

1 Application sends documents to embedding model

2 Embedding Model returns embedding vector and application adds vectors to JSON documents

3 Application creates/updates JSON documents (image with vector) to Couchbase Lite

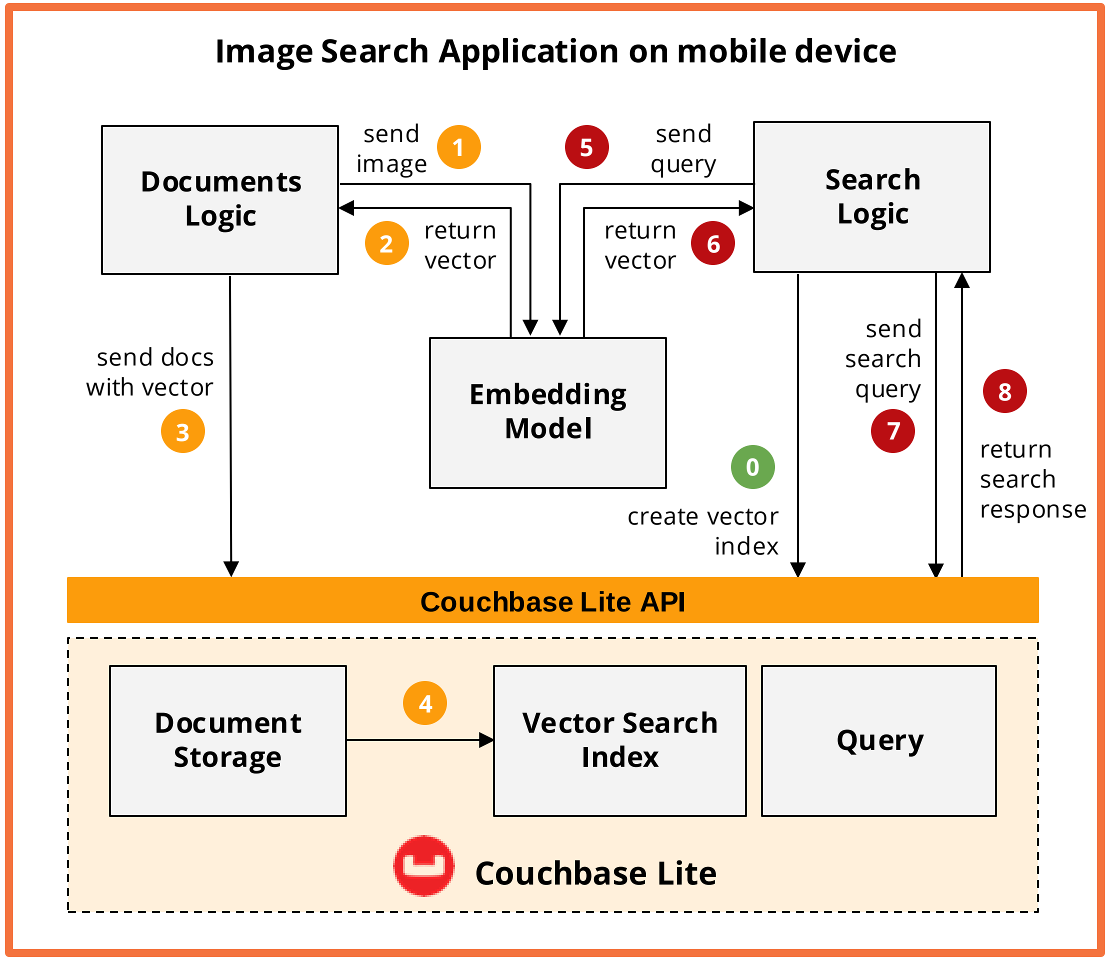
4 This triggers update to vector search index on Couchbase Lite that updates index with document

5 Application sends captured image to Embedding Model

6 Embedding Model returns corresponding embedding vector

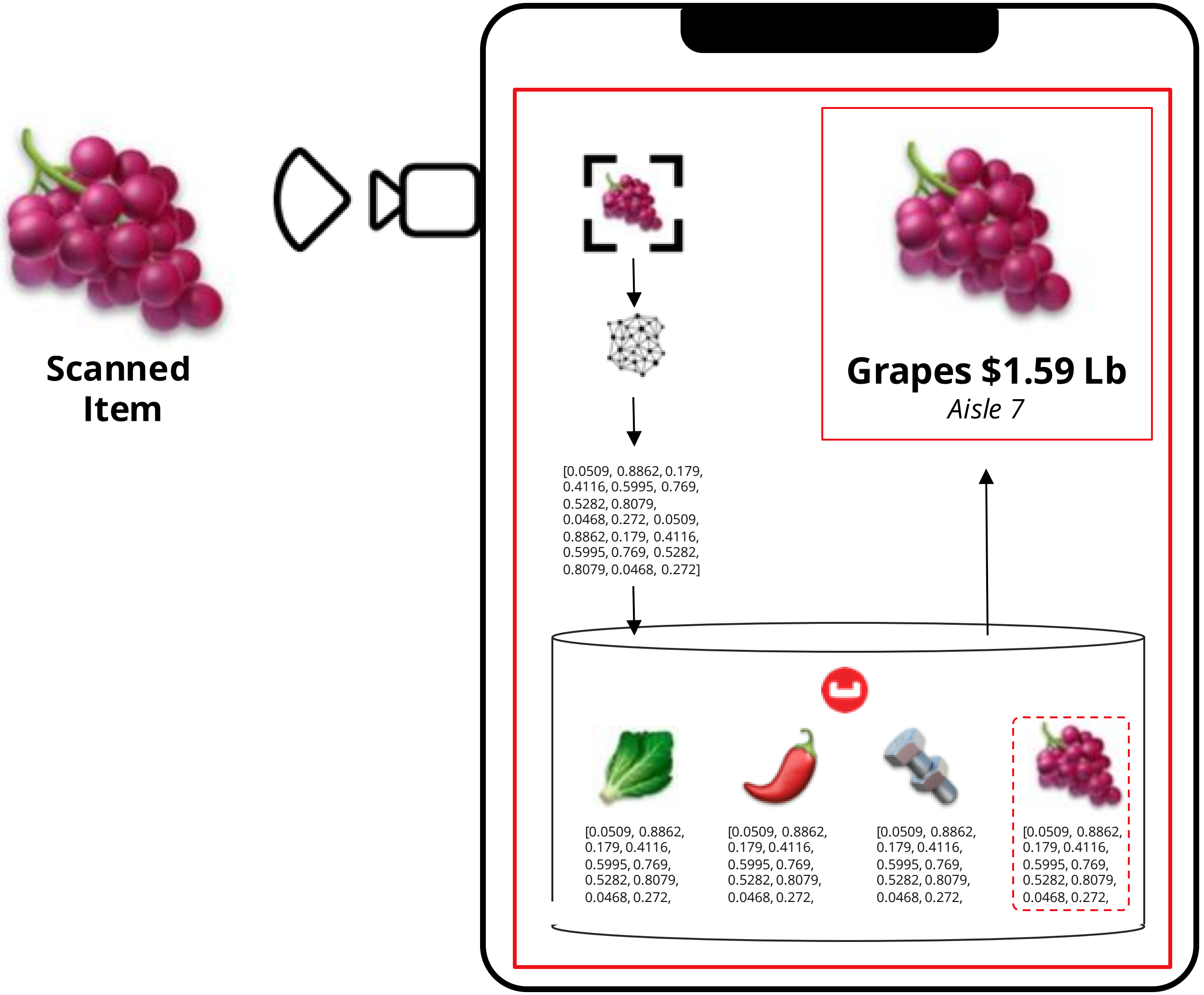
7 Application calls Couchbase Lite with a pure vector search, or hybrid mixed vector search and FTS, or hybrid vector and SQL++ query

8 Couchbase Lite returns top k-NN results of query with requested fields or objects - the vector search returns similar images



Couchbase Mobile Vector Search Demo.

➢ Vector Search For Image Lookup



Tablet-based Point-of-Sale app

- Item is scanned with device camera
- Vectorized with local embedding model
- Couchbase Lite vector store is searched
- Nearest vector match found
- App displays matching item and data

Confidential and Proprietary. Do not distribute without Couchbase consent. © Couchbase 2024. All rights reserved.

4. Couchbase Mobile Vector Search Demo.

- **Simple Intelligence App**

https://apps.apple.com/us/app/simple-intelligence/id6504311724

Confidential and Proprietary. Do not distribute without Couchbase consent. © Couchbase 2024. All rights reserved.

수고하셨습니다.

paul.son@couchbase.com

www.couchbase.com

cloud.couchbase.com

Couchbase