RAG with Llama 2 and LangChain

Building with Open-Source LLM Ops

Presented by

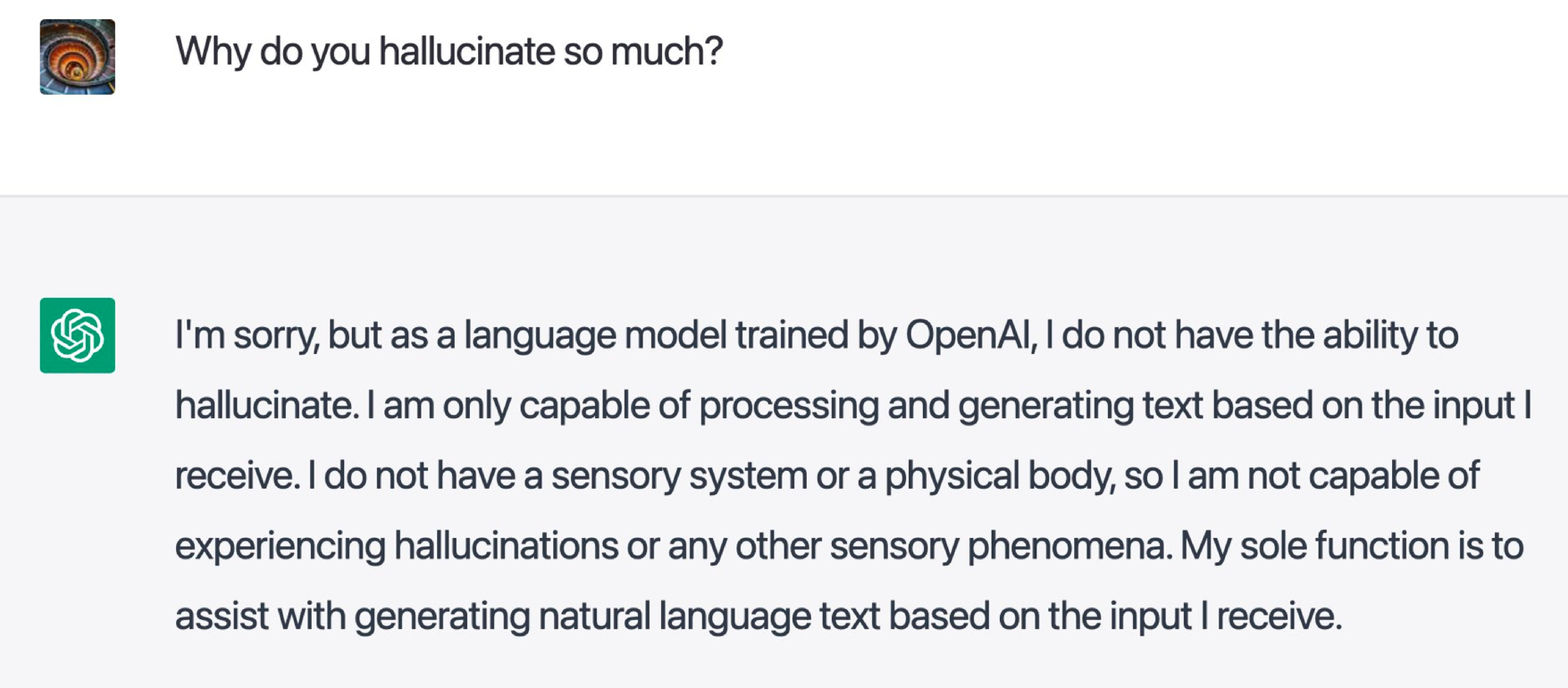
Greg Loughnane, Founder & CEO
Chris Alexiuk, Head of LLMs

OVERVIEW

- 🙋 So what is RAG, anyways?
- 💅💣 The Barbenheimer RAG System
- 🔗 Why LangChain?
- 🧩 3 Easy Pieces!
- ❓ Conclusions, QA

SO ... RAG

ASKING LLMS QUESTIONS



Why do you hallucinate so much?

I'm sorry, but as a language model trained by OpenAI, I do not have the ability to hallucinate. I am only capable of processing and generating text based on the input I receive. I do not have a sensory system or a physical body, so I am not capable of experiencing hallucinations or any other sensory phenomena. My sole function is to assist with generating natural language text based on the input I receive.

CONSIDER QUESTION ANSWERING

Question-Answering (QA)

- LLMs will answer your questions!
- But they often … **hallucinate**

“Retrieval Augmented” QA

- Look up **more info & sources**
- Put it in the answer!

RETREIVAL AUGMENTED GENERATION

Retrieval

- Retriever: an interface that returns documents given a query

Augmented

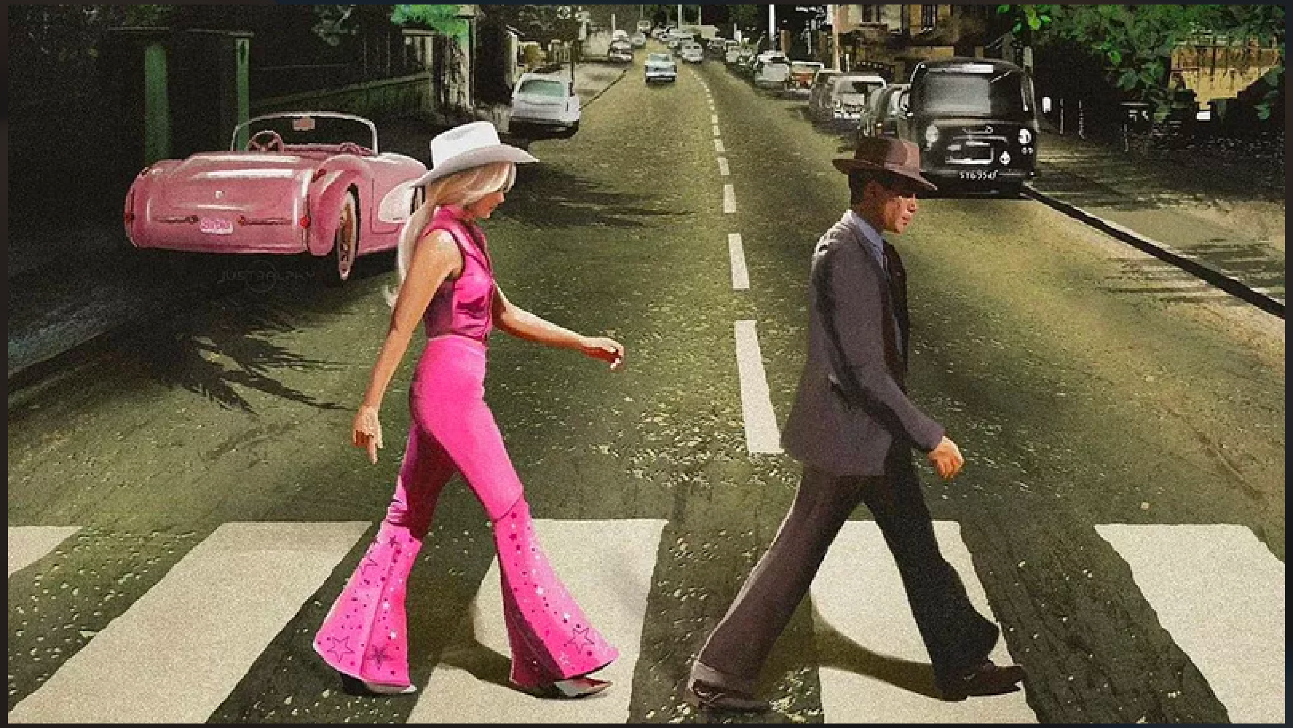
- Prompt augmentation: Providing additional context/info in the prompt to improve performance

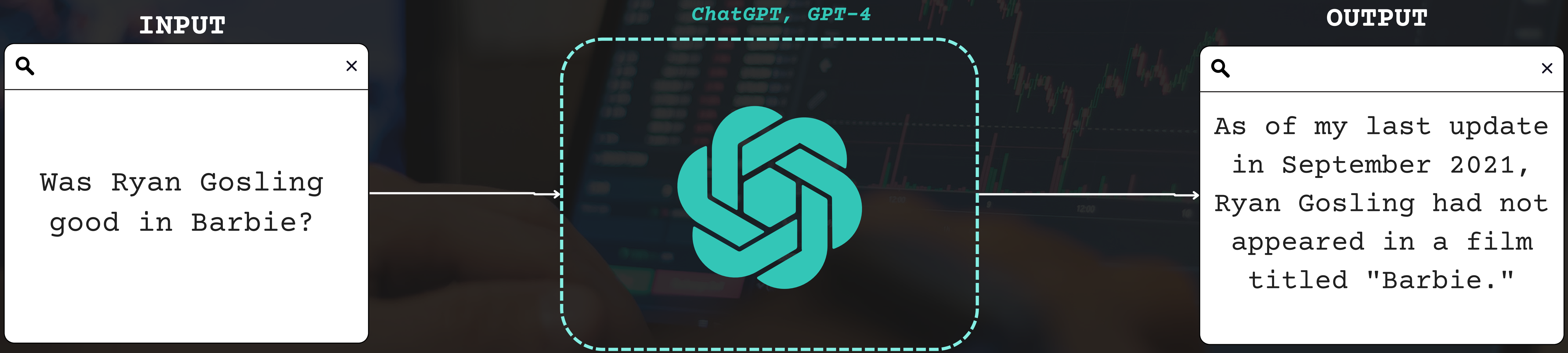
Generation

- Well, GPT, GenAI, etc.

Prompt: Retrieval Augmented Generation

BARBENHEIMER RAG SYSTEM





INPUT

Was Ryan Gosling good in Barbie?

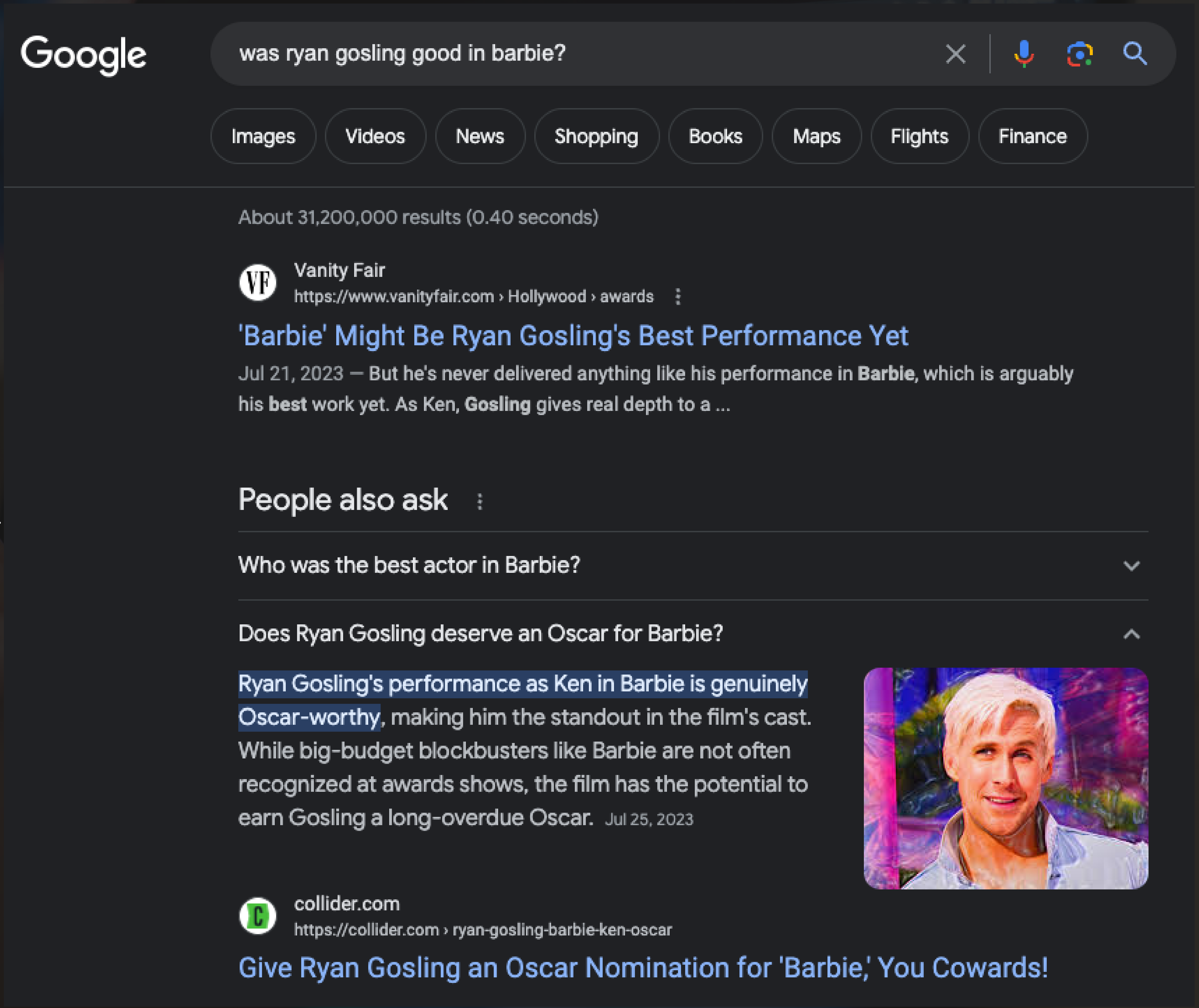
ChatGPT, GPT-4

OUTPUT

As of my last update in September 2021, Ryan Gosling had not appeared in a film titled "Barbie."

INPUT

Was Ryan Gosling good in Barbie?



was ryan gosling good in barbie?

About 31,200,000 results (0.40 seconds)

Vanity Fair
https://www.vanityfair.com › Hollywood › awards

'Barbie' Might Be Ryan Gosling's Best Performance Yet

Jul 21, 2023 — But he's never delivered anything like his performance in **Barbie**, which is arguably his **best** work yet. As Ken, **Gosling** gives real depth to a ...

People also ask

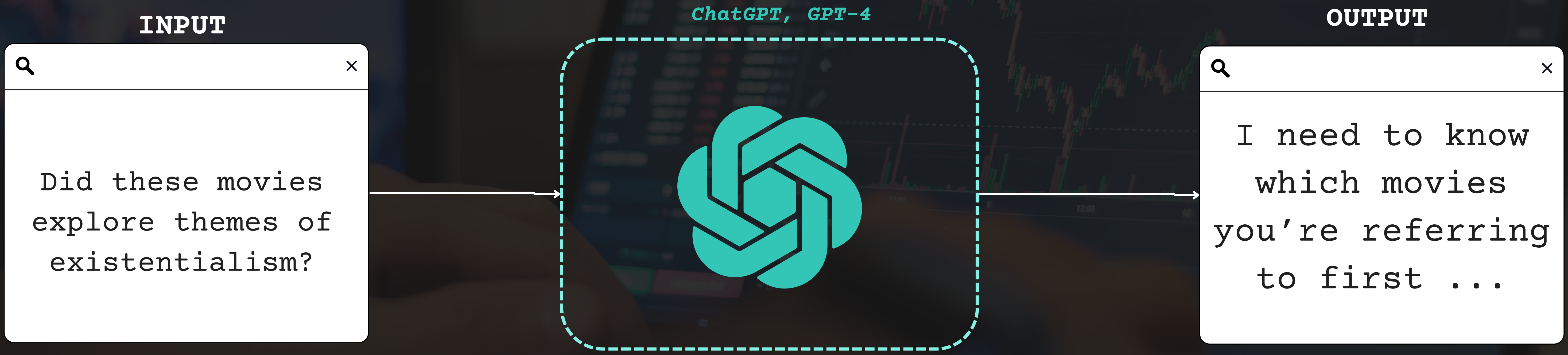
Who was the best actor in Barbie?

Does Ryan Gosling deserve an Oscar for Barbie?

Ryan Gosling's performance as Ken in Barbie is genuinely Oscar-worthy, making him the standout in the film's cast. While big-budget blockbusters like Barbie are not often recognized at awards shows, the film has the potential to earn Gosling a long-overdue Oscar. Jul 25, 2023

collider.com
https://collider.com › ryan-gosling-barbie-ken-oscar

Give Ryan Gosling an Oscar Nomination for 'Barbie,' You Cowards!



INPUT

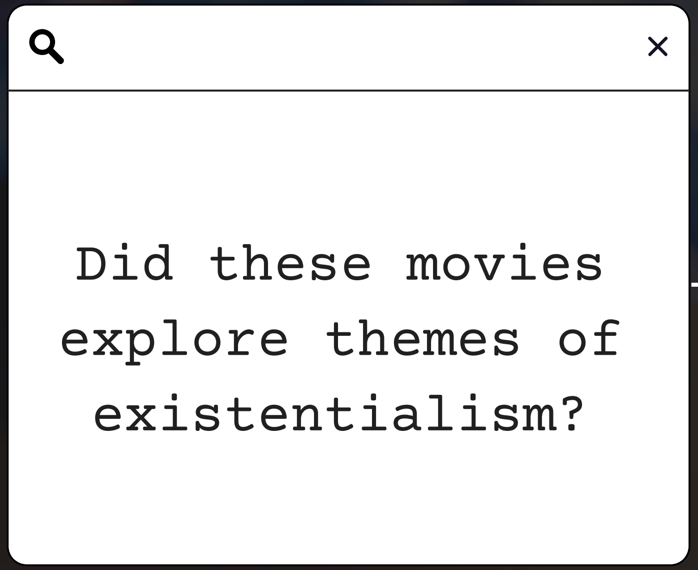
Did these movies explore themes of existentialism?

ChatGPT, GPT-4

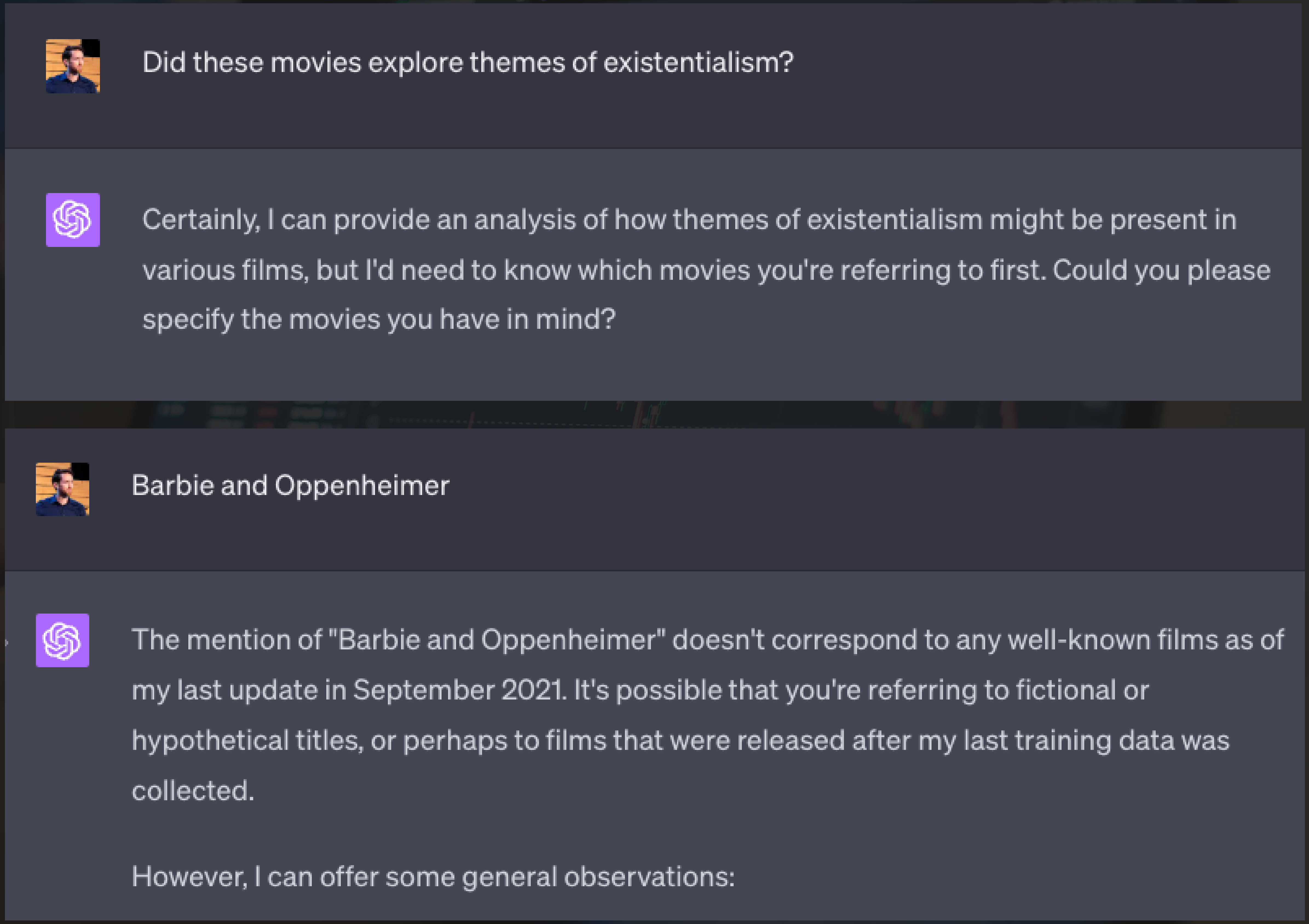
OUTPUT

I need to know which movies you're referring to first ...

INPUT



Did these movies explore themes of existentialism?



Did these movies explore themes of existentialism?

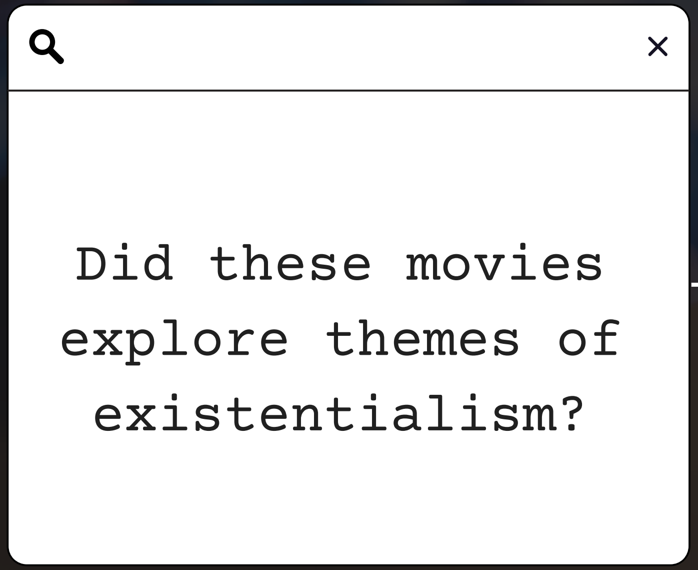
Certainly, I can provide an analysis of how themes of existentialism might be present in various films, but I'd need to know which movies you're referring to first. Could you please specify the movies you have in mind?

Barbie and Oppenheimer

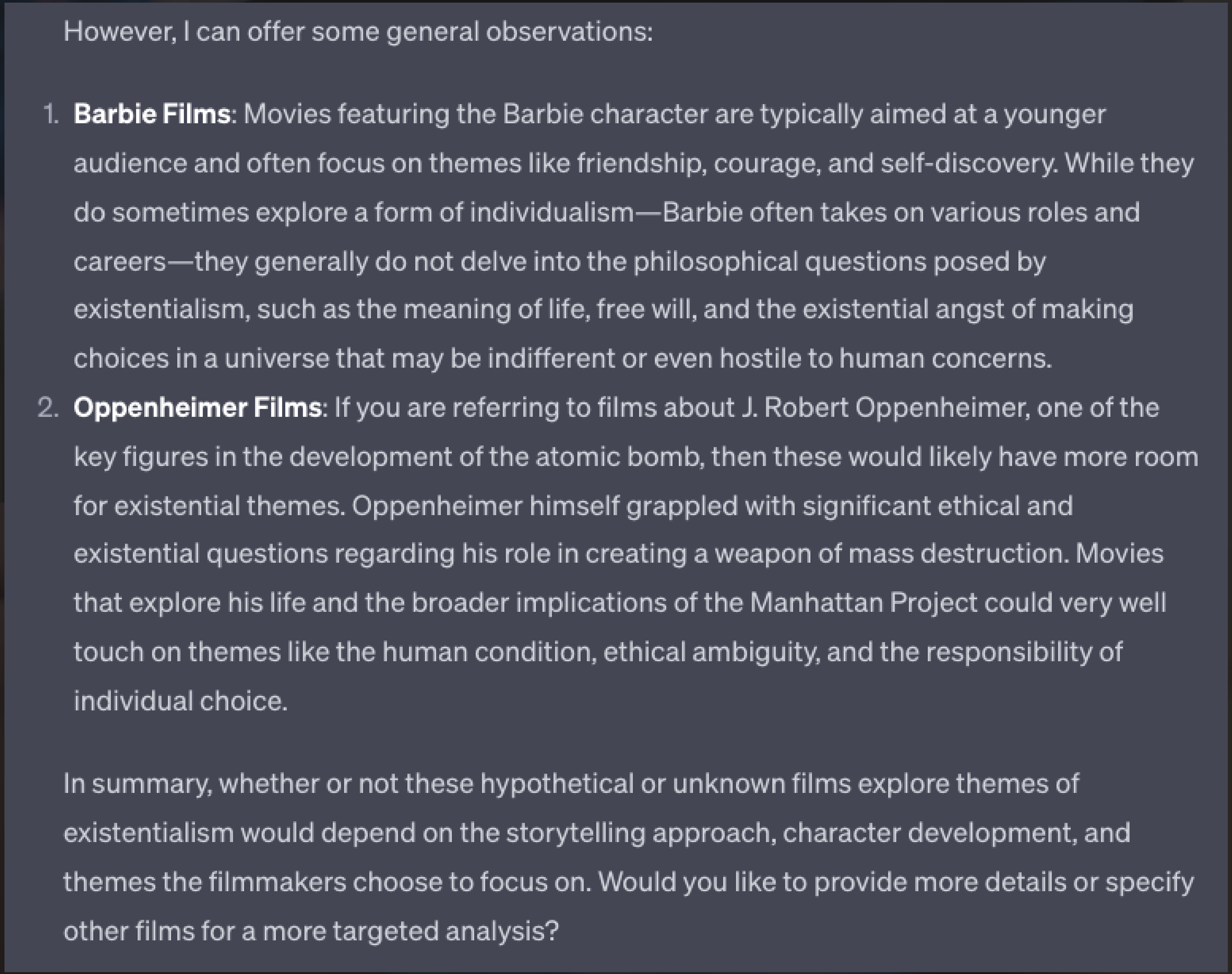
The mention of "Barbie and Oppenheimer" doesn't correspond to any well-known films as of my last update in September 2021. It's possible that you're referring to fictional or hypothetical titles, or perhaps to films that were released after my last training data was collected.

However, I can offer some general observations:

INPUT



Did these movies explore themes of existentialism?



However, I can offer some general observations:

1. **Barbie Films**: Movies featuring the Barbie character are typically aimed at a younger audience and often focus on themes like friendship, courage, and self-discovery. While they do sometimes explore a form of individualism—Barbie often takes on various roles and careers—they generally do not delve into the philosophical questions posed by existentialism, such as the meaning of life, free will, and the existential angst of making choices in a universe that may be indifferent or even hostile to human concerns.
2. **Oppenheimer Films**: If you are referring to films about J. Robert Oppenheimer, one of the key figures in the development of the atomic bomb, then these would likely have more room for existential themes. Oppenheimer himself grappled with significant ethical and existential questions regarding his role in creating a weapon of mass destruction. Movies that explore his life and the broader implications of the Manhattan Project could very well touch on themes like the human condition, ethical ambiguity, and the responsibility of individual choice.

In summary, whether or not these hypothetical or unknown films explore themes of existentialism would depend on the storytelling approach, character development, and themes the filmmakers choose to focus on. Would you like to provide more details or specify other films for a more targeted analysis?

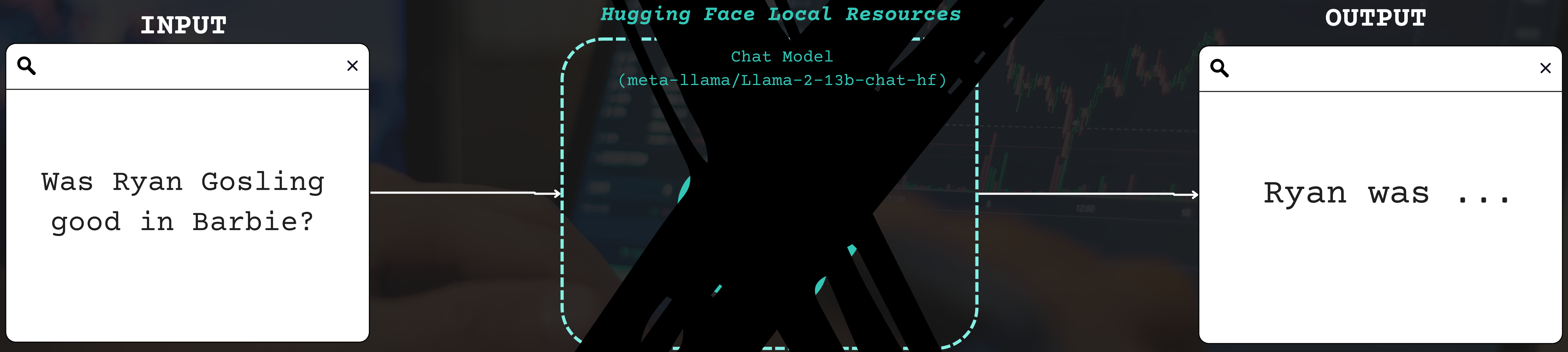
INPUT

Did these movies explore themes of existentialism?

However, I can offer some general observations:

1. **Barbie Films**: Movies featuring the Barbie character are typically aimed at a younger audience and … overy. While they do sometimes … us roles and careers—they … ed by existentialism … t of making choices in a u… erns.
2. **Oppenheimer** … ner, one of the key figures in … have more room for existential … cal and existential qu… uction. Movies that explore h… could very well touch on them … onsibility of individual cho…

In summary, whether or not these hypothetical or unknown films explore themes of existentialism would depend on the storytelling approach, character development, and themes the filmmakers choose to focus on. Would you like to provide more details or specify other films for a more targeted analysis?



INPUT

Was Ryan Gosling good in Barbie?

~~*Hugging Face Local Resources*~~

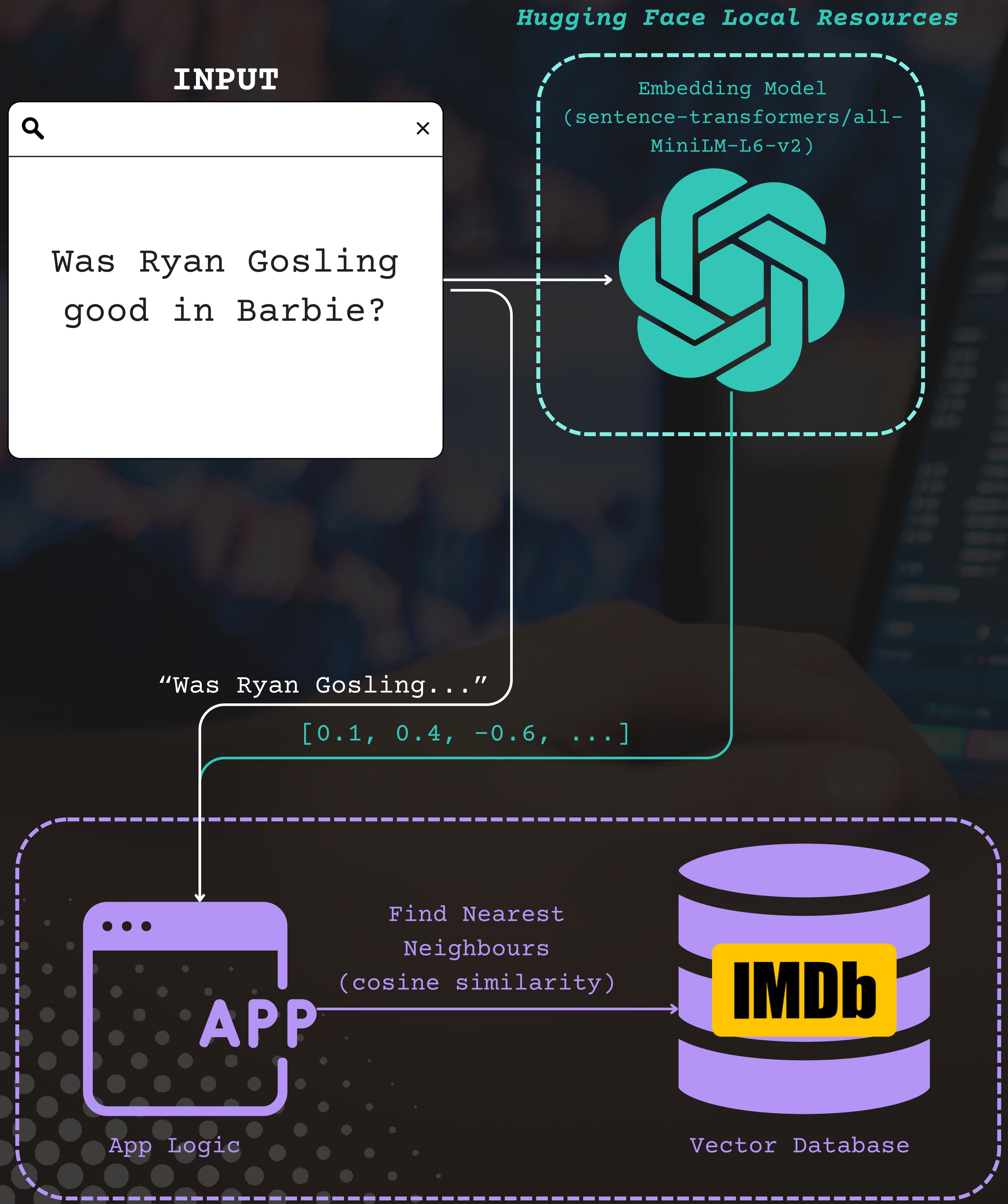
~~Chat Model~~
~~(meta-llama/Llama-2-13b-chat-hf)~~

OUTPUT

Ryan was ...

INPUT

Was Ryan Gosling good in Barbie?



Hugging Face Local Resources

INPUT

Was Ryan Gosling good in Barbie?

Embedding Model (sentence-transformers/all-MiniLM-L6-v2)

“Was Ryan Gosling...”

[0.1, 0.4, -0.6, ...]

Find Nearest Neighbours (cosine similarity)

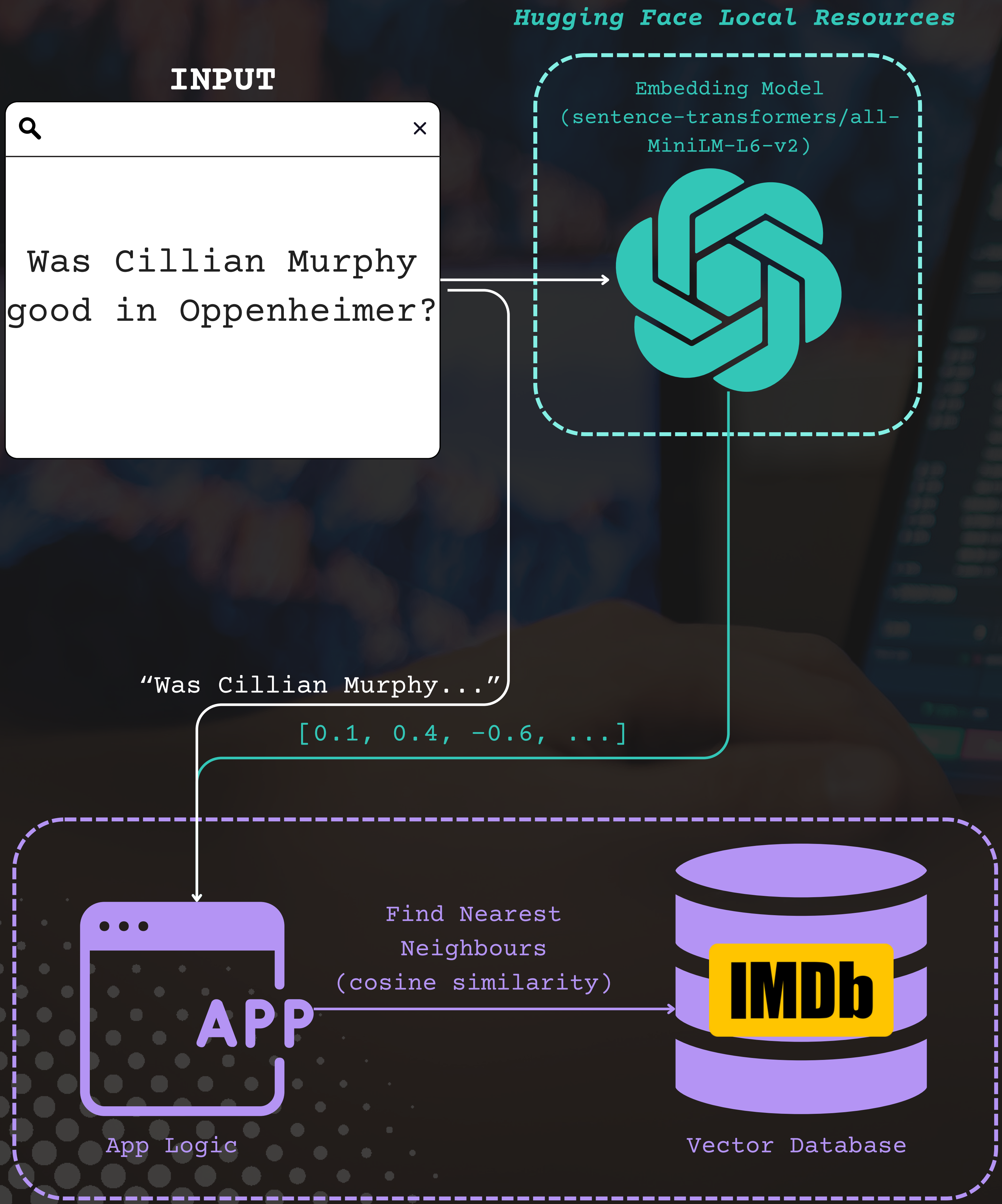
App Logic

Vector Database

Barbie (I) (2023)

User Reviews

Review this title



INPUT

Was Cillian Murphy good in Oppenheimer?

Hugging Face Local Resources

Embedding Model (sentence-transformers/all-MiniLM-L6-v2)

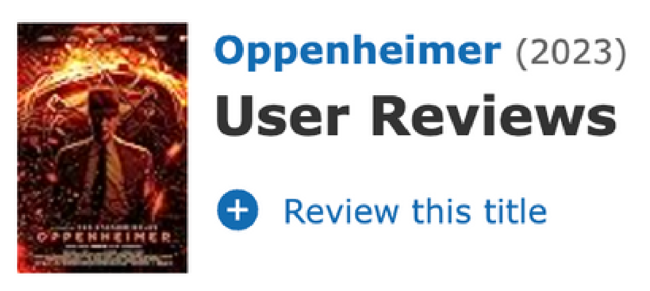
“Was Cillian Murphy...”

[0.1, 0.4, -0.6, ...]

Find Nearest Neighbours (cosine similarity)

App Logic

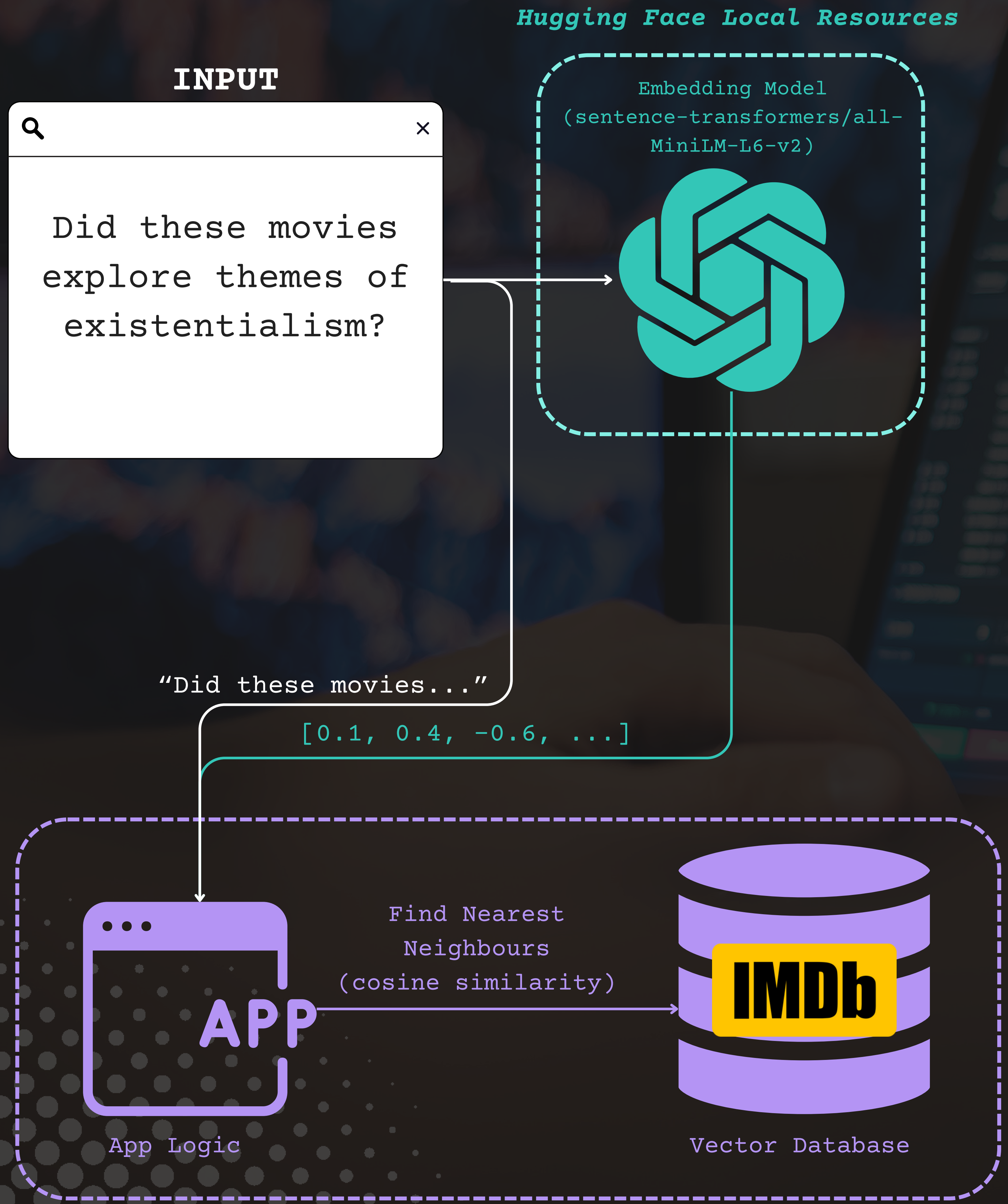
Vector Database



Oppenheimer (2023)

User Reviews

⊕ Review this title



Hugging Face Local Resources

INPUT

Did these movies explore themes of existentialism?

Embedding Model (sentence-transformers/all-MiniLM-L6-v2)

"Did these movies..."

[0.1, 0.4, -0.6, ...]

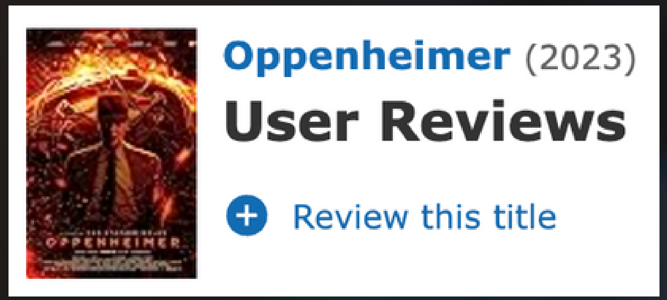
Find Nearest Neighbours (cosine similarity)

App Logic

Vector Database

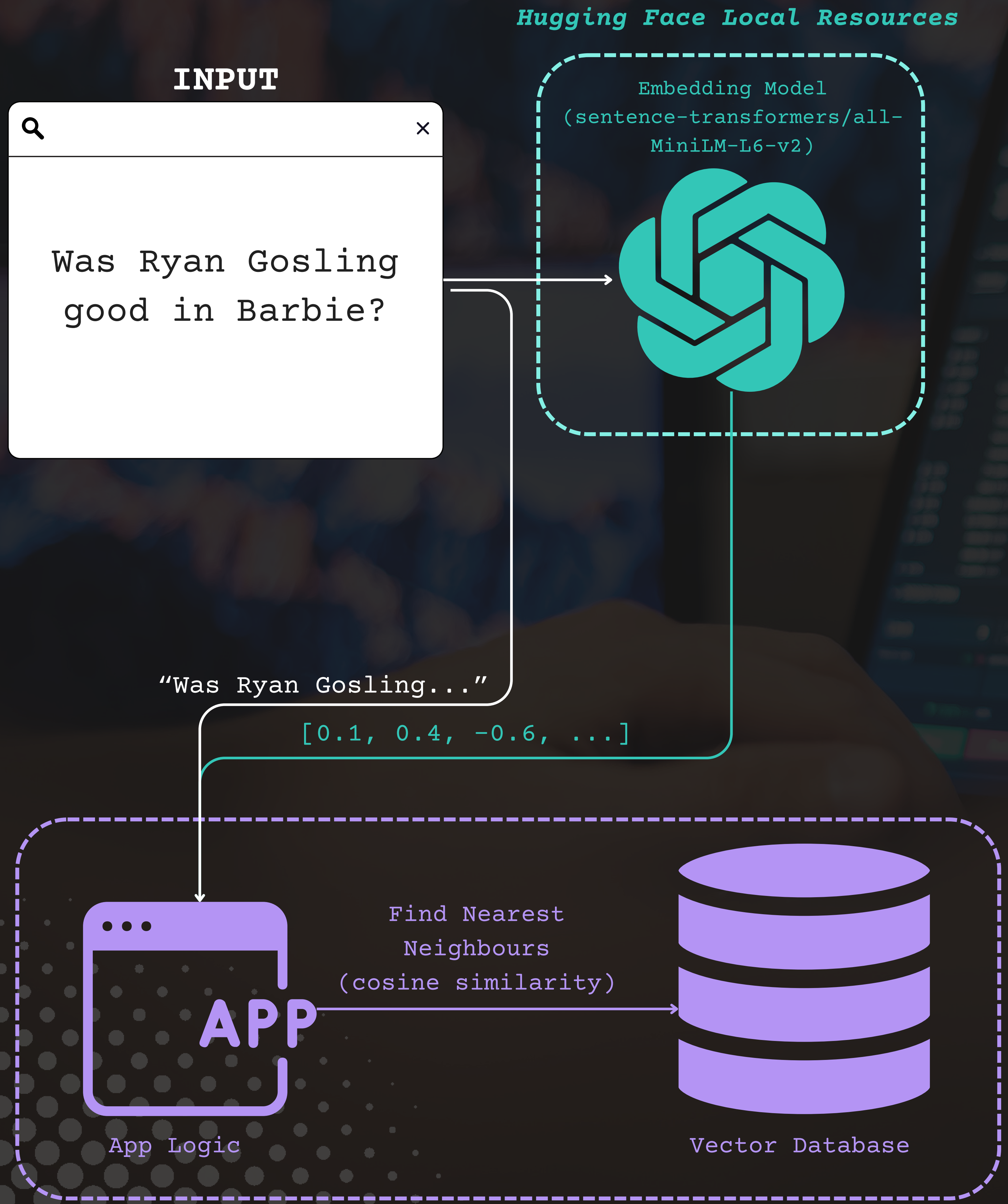
Barbie (I) (2023)

User Reviews



Oppenheimer (2023)

User Reviews



Hugging Face Local Resources

INPUT

Was Ryan Gosling good in Barbie?

Embedding Model (sentence-transformers/all-MiniLM-L6-v2)

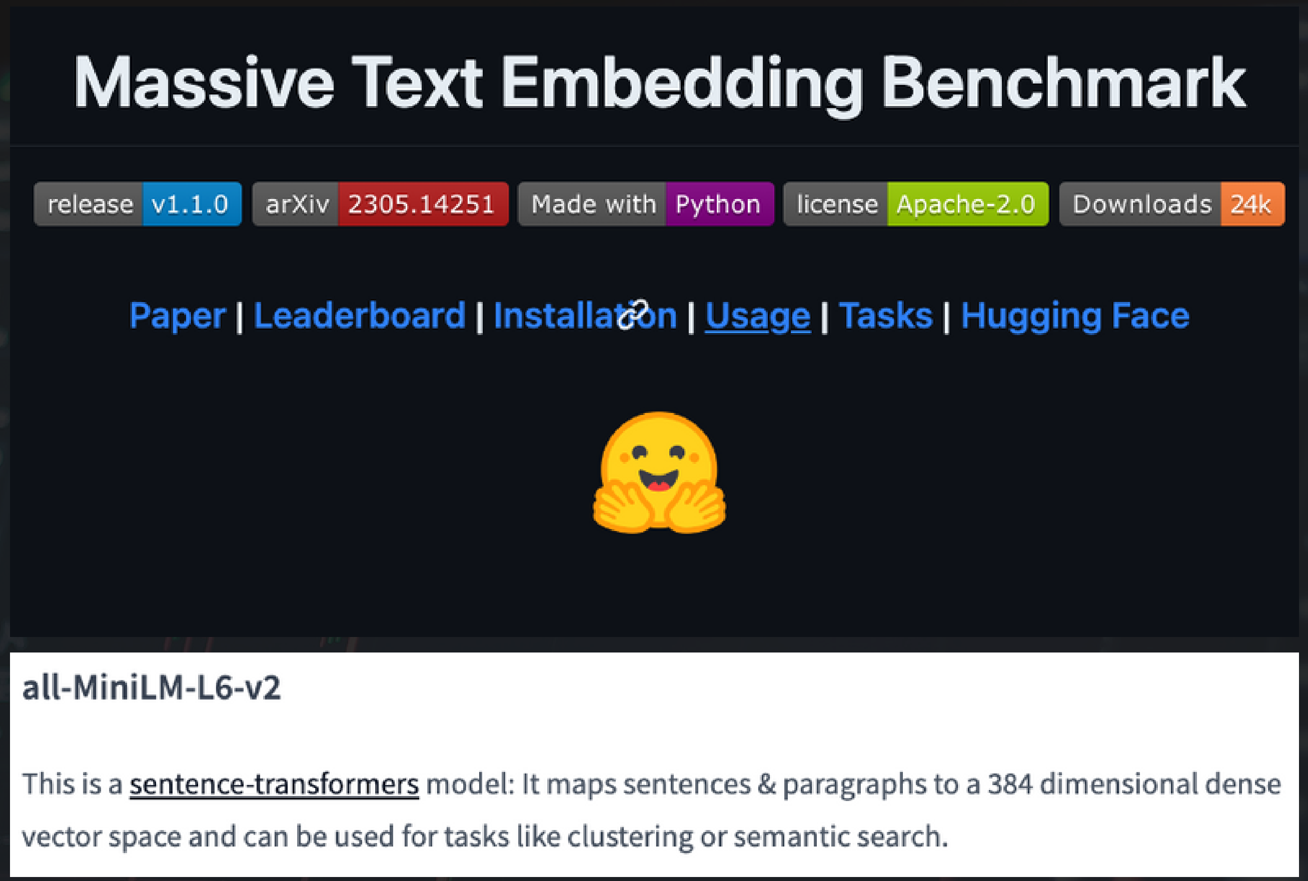
“Was Ryan Gosling...”

[0.1, 0.4, -0.6, ...]

Find Nearest Neighbours (cosine similarity)

App Logic

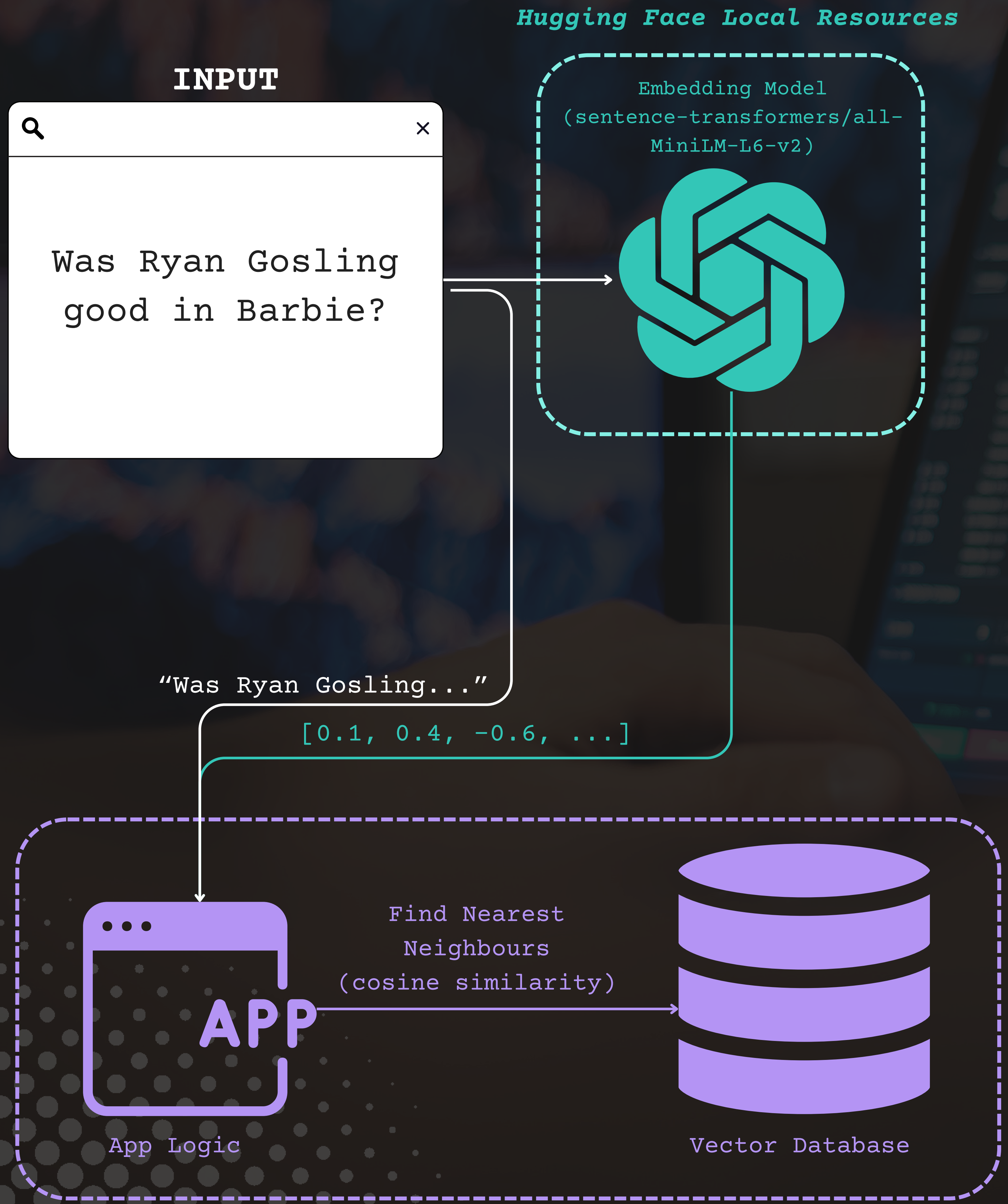
Vector Database



Massive Text Embedding Benchmark

all-MiniLM-L6-v2

This is a sentence-transformers model: It maps sentences & paragraphs to a 384 dimensional dense vector space and can be used for tasks like clustering or semantic search.



Hugging Face Local Resources

INPUT

Was Ryan Gosling good in Barbie?

Embedding Model (sentence-transformers/all-MiniLM-L6-v2)

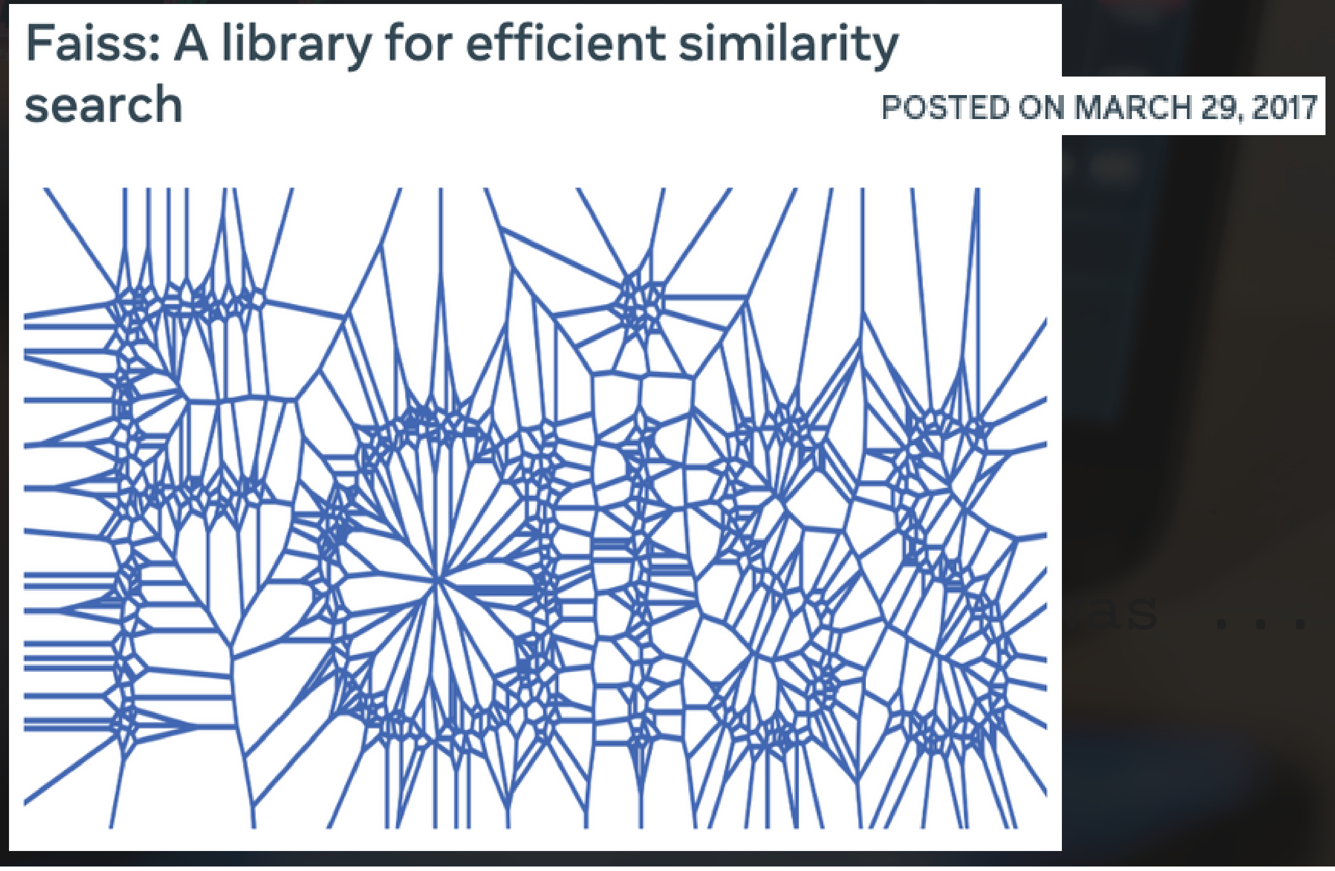
“Was Ryan Gosling...”

[0.1, 0.4, -0.6, ...]

Find Nearest Neighbours (cosine similarity)

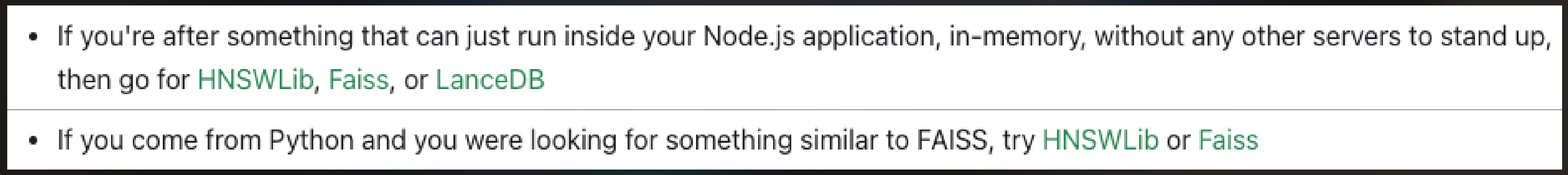
App Logic

Vector Database

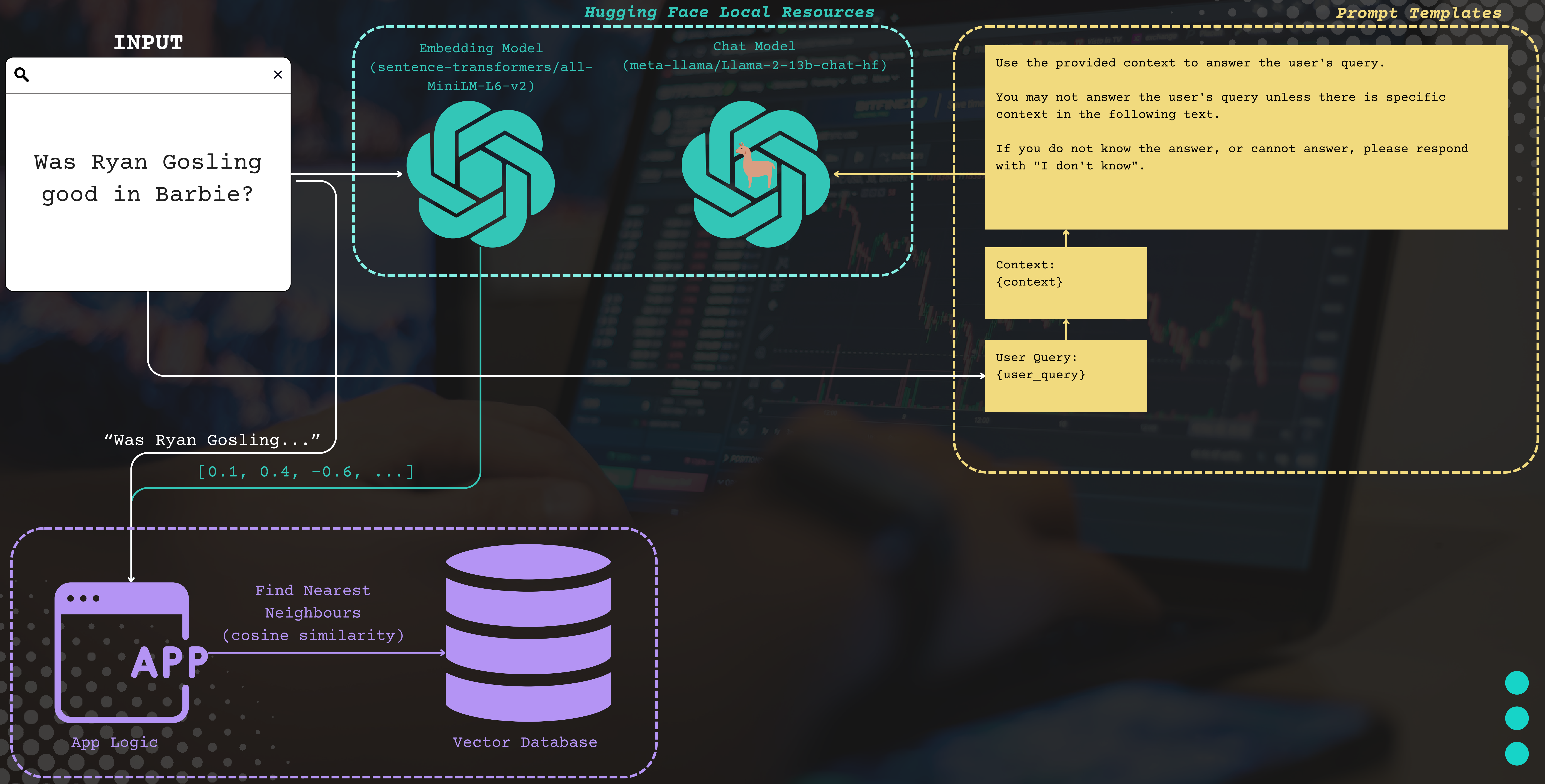


Faiss: A library for efficient similarity search

POSTED ON MARCH 29, 2017



- If you're after something that can just run inside your Node.js application, in-memory, without any other servers to stand up, then go for HNSWLib, Faiss, or LanceDB
- If you come from Python and you were looking for something similar to FAISS, try HNSWLib or Faiss



INPUT

Was Ryan Gosling good in Barbie?

Hugging Face Local Resources

Embedding Model
(sentence-transformers/all-MiniLM-L6-v2)

Chat Model
(meta-llama/Llama-2-13b-chat-hf)

Prompt Templates

Use the provided context to answer the user's query.

You may not answer the user's query unless there is specific context in the following text.

If you do not know the answer, or cannot answer, please respond with "I don't know".

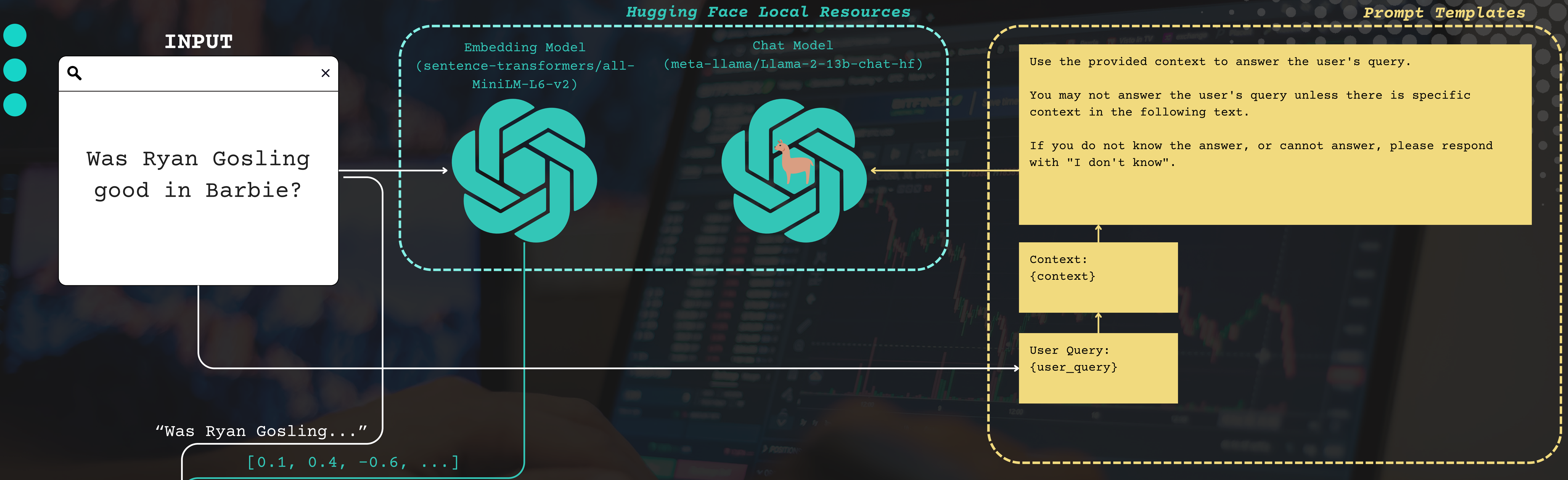
Context:
{context}

User Query:
{user_query}

"Was Ryan Gosling..."

[0.1, 0.4, -0.6, ...]

Find Nearest Neighbours
(cosine similarity)



Hugging Face Local Resources

Prompt Templates

INPUT

Was Ryan Gosling good in Barbie?

Embedding Model (sentence-transformers/all-MiniLM-L6-v2)

Chat Model (meta-llama/Llama-2-13b-chat-hf)

Use the provided context to answer the user's query.

You may not answer the user's query unless there is specific context in the following text.

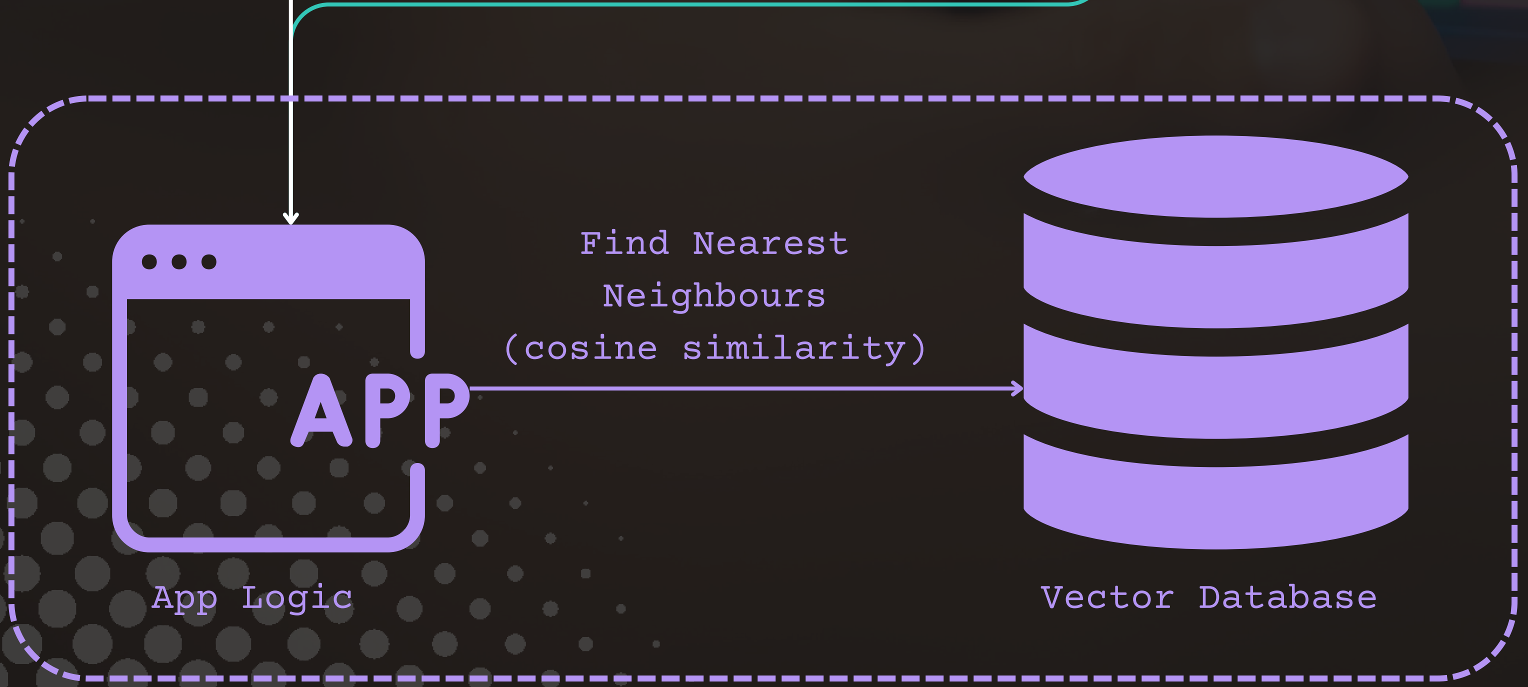
If you do not know the answer, or cannot answer, please respond with "I don't know".

Context:
{context}

User Query:
{user_query}

"Was Ryan Gosling..."

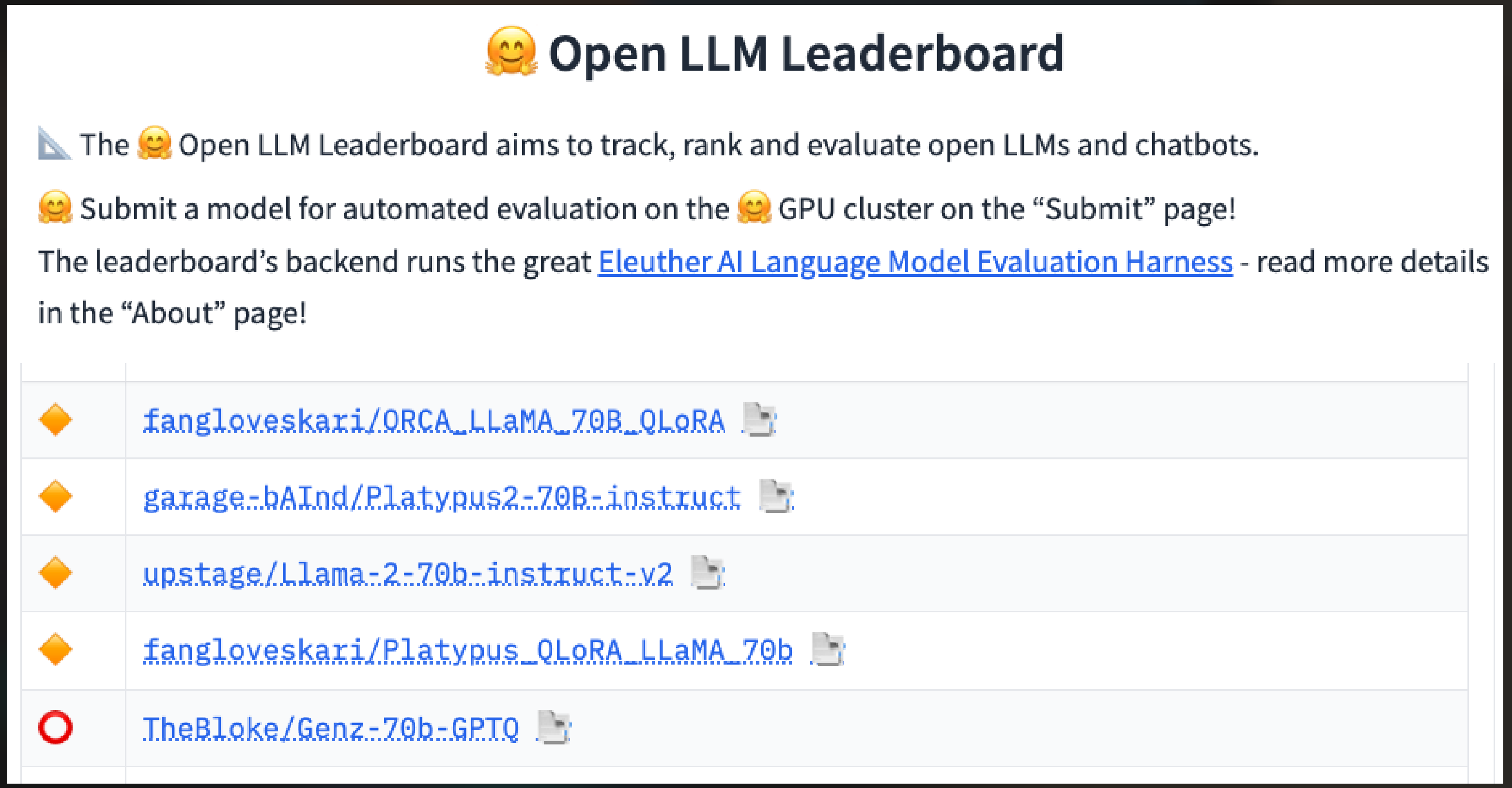
[0.1, 0.4, -0.6, ...]



Find Nearest Neighbours (cosine similarity)

App Logic

Vector Database

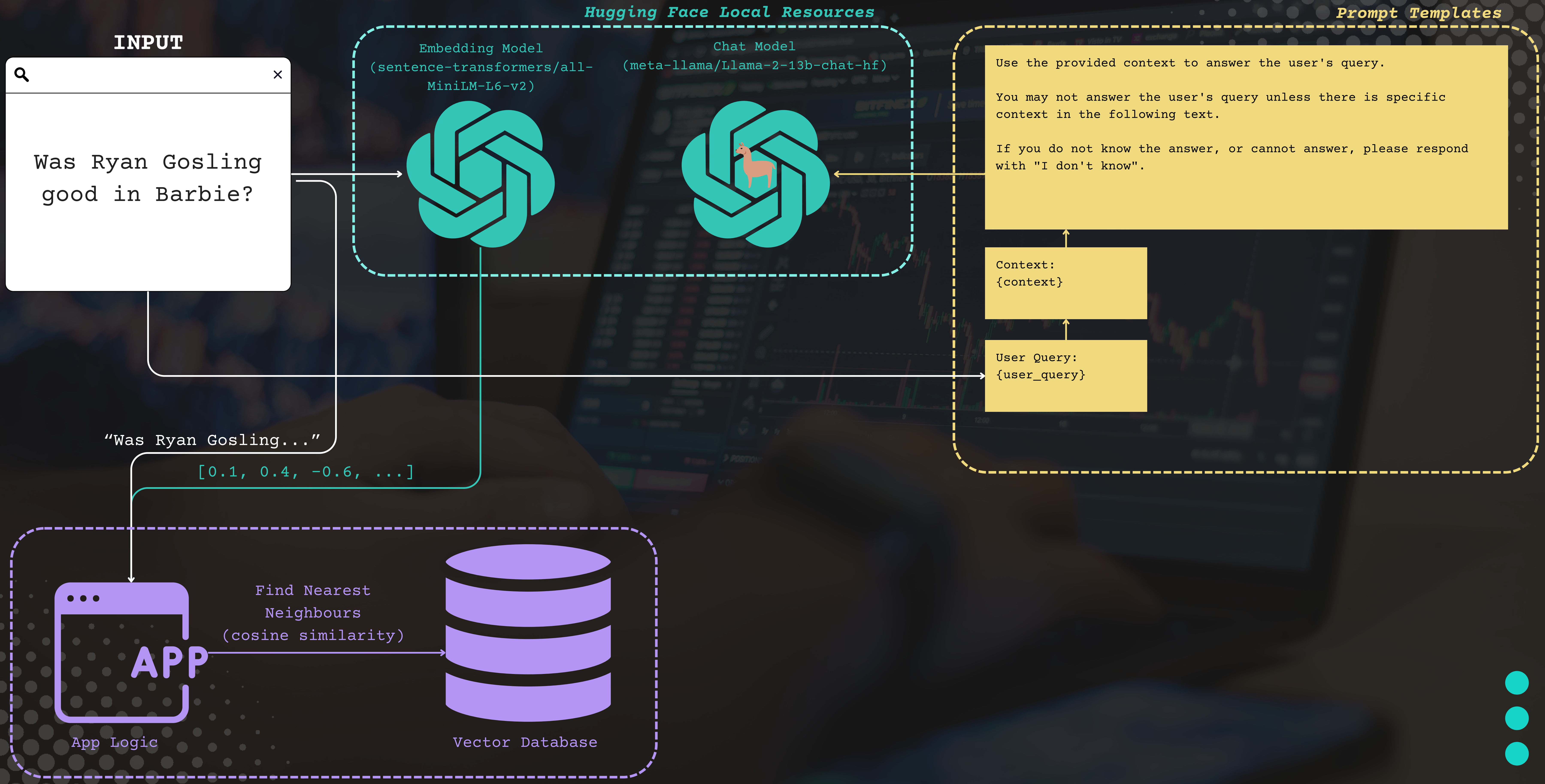


🤗 Open LLM Leaderboard

📐 The 🤗 Open LLM Leaderboard aims to track, rank and evaluate open LLMs and chatbots.

🤗 Submit a model for automated evaluation on the 🤗 GPU cluster on the "Submit" page!
The leaderboard's backend runs the great Eleuther AI Language Model Evaluation Harness - read more details in the "About" page!

🔶	fangloveskari/ORCA_LLaMA_70B_QLoRA
🔶	garage-bAInd/Platypus2-70B-instruct
🔶	upstage/Llama-2-70b-instruct-v2
🔶	fangloveskari/Platypus_QLoRA_LLaMA_70b
⭕	TheBloke/Genz-70b-GPTQ



INPUT

Was Ryan Gosling good in Barbie?

Hugging Face Local Resources

Embedding Model (sentence-transformers/all-MiniLM-L6-v2)

Chat Model (meta-llama/Llama-2-13b-chat-hf)

Prompt Templates

Use the provided context to answer the user's query.

You may not answer the user's query unless there is specific context in the following text.

If you do not know the answer, or cannot answer, please respond with "I don't know".

Context:
{context}

User Query:
{user_query}

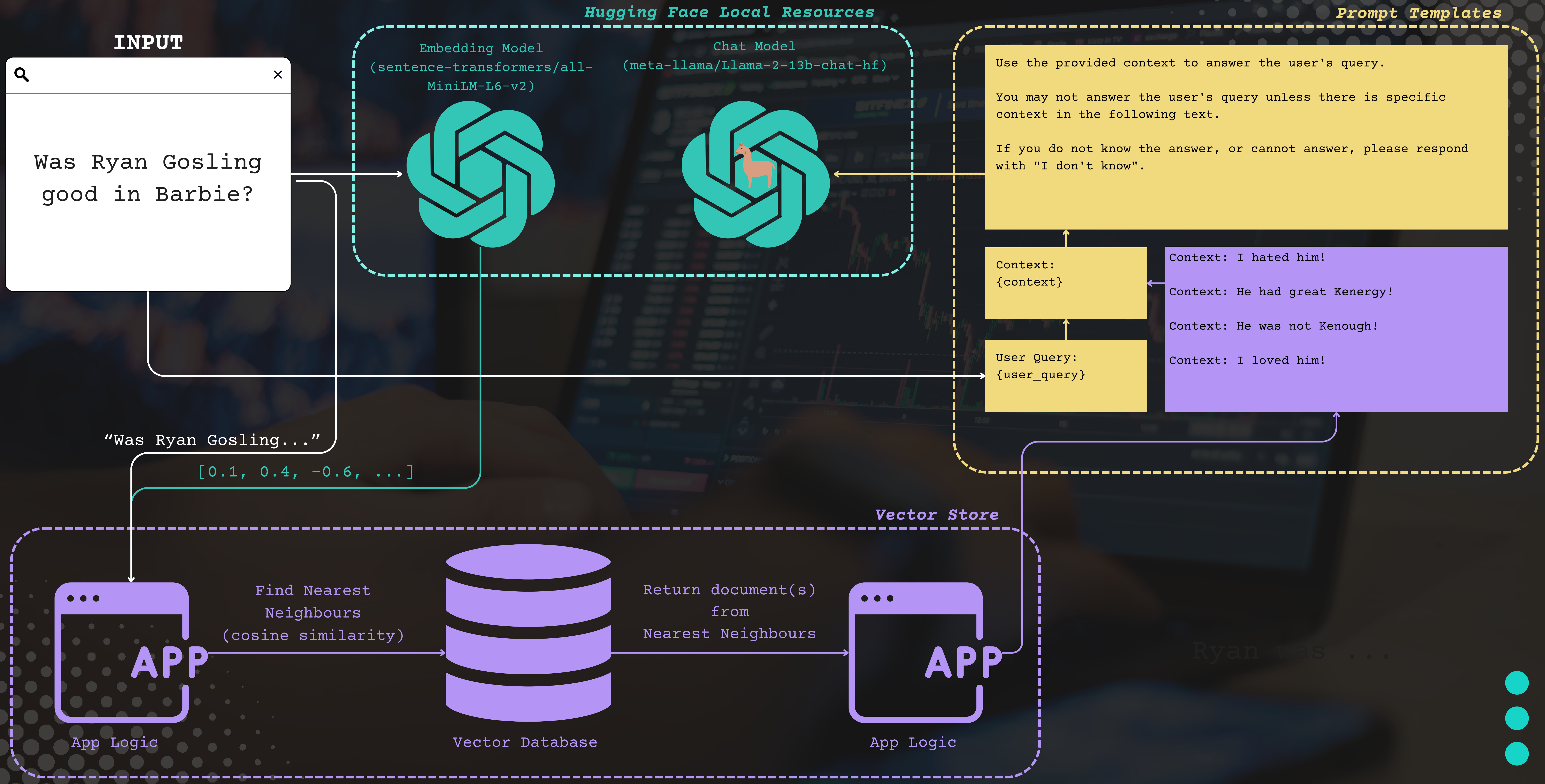
"Was Ryan Gosling..."

[0.1, 0.4, -0.6, ...]

Find Nearest Neighbours (cosine similarity)

App Logic

Vector Database



INPUT

Was Ryan Gosling good in Barbie?

Hugging Face Local Resources

Embedding Model (sentence-transformers/all-MiniLM-L6-v2)

Chat Model (meta-llama/Llama-2-13b-chat-hf)

Prompt Templates

Use the provided context to answer the user's query.

You may not answer the user's query unless there is specific context in the following text.

If you do not know the answer, or cannot answer, please respond with "I don't know".

Context:
{context}

User Query:
{user_query}

Context: I hated him!

Context: He had great Kenergy!

Context: He was not Kenough!

Context: I loved him!

"Was Ryan Gosling..."

[0.1, 0.4, -0.6, ...]

Vector Store

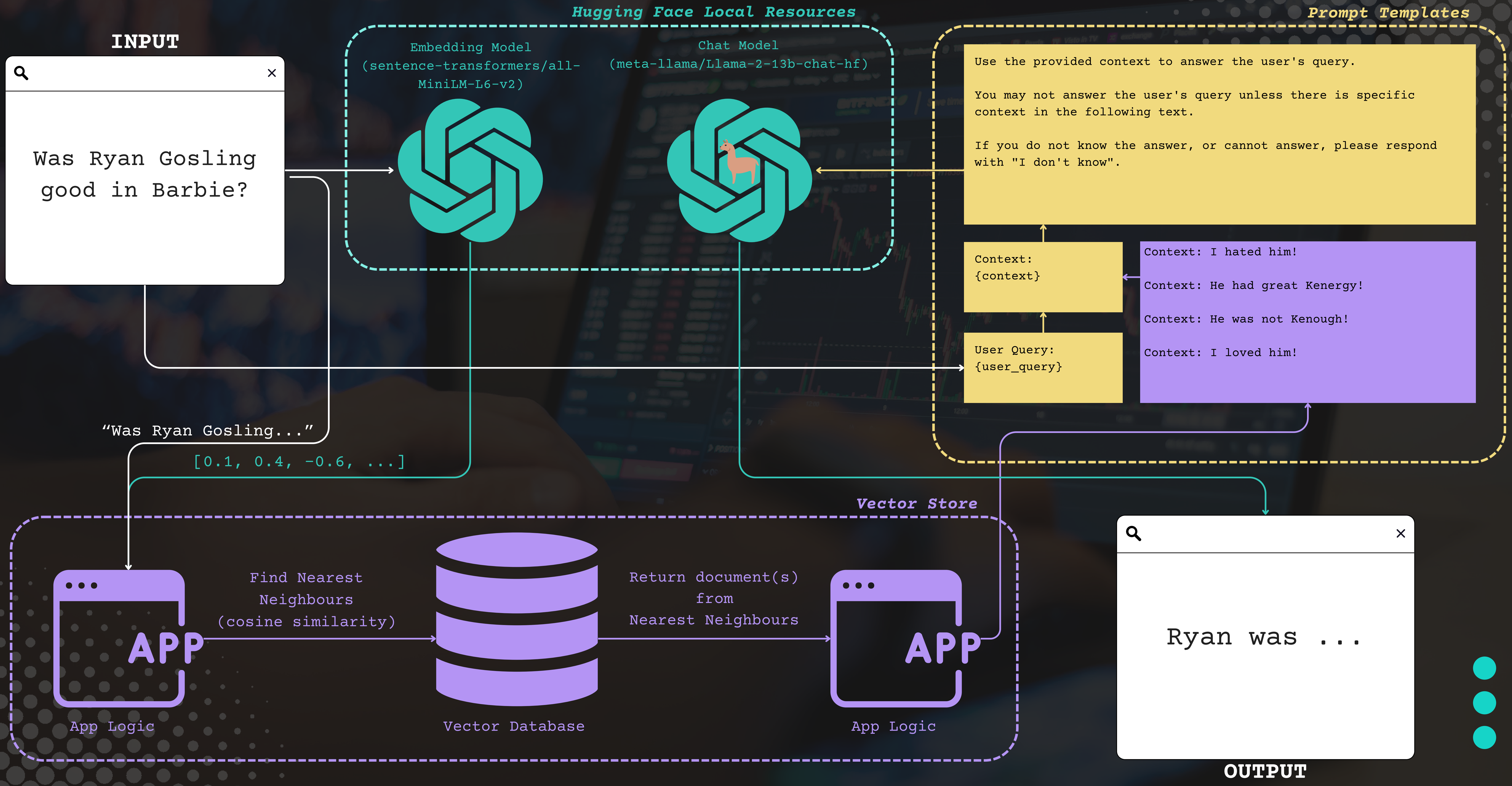
Find Nearest Neighbours (cosine similarity)

Return document(s) from Nearest Neighbours

App Logic

Vector Database

App Logic



INPUT

Was Ryan Gosling good in Barbie?

Hugging Face Local Resources

Embedding Model
(sentence-transformers/all-MiniLM-L6-v2)

Chat Model
(meta-llama/Llama-2-13b-chat-hf)

Prompt Templates

Use the provided context to answer the user's query.

You may not answer the user's query unless there is specific context in the following text.

If you do not know the answer, or cannot answer, please respond with "I don't know".

Context:
{context}

User Query:
{user_query}

Context: I hated him!

Context: He had great Kenergy!

Context: He was not Kenough!

Context: I loved him!

“Was Ryan Gosling...”

[0.1, 0.4, -0.6, ...]

Vector Store

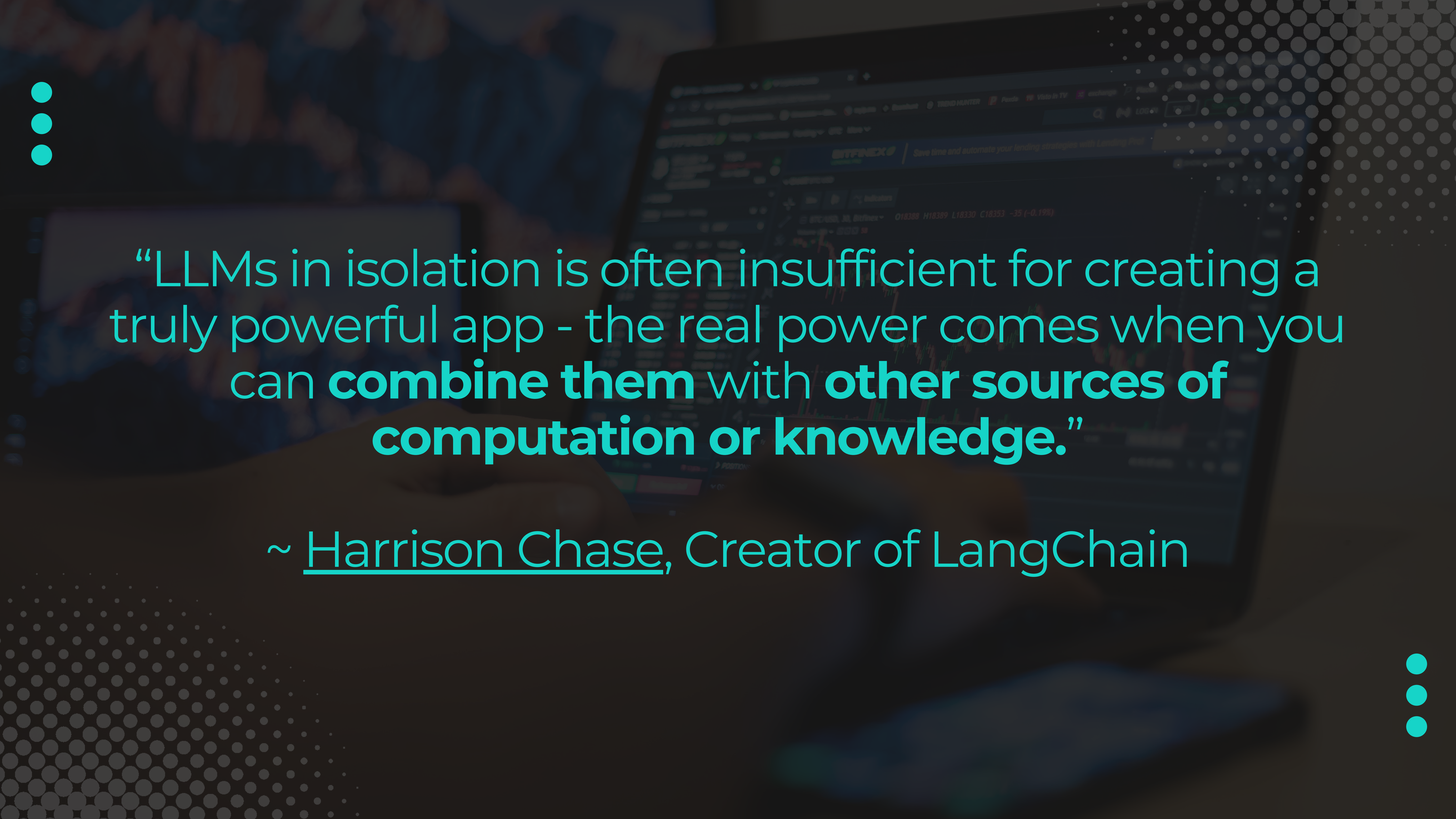
Find Nearest Neighbours (cosine similarity)

Return document(s) from Nearest Neighbours

Ryan was ...

OUTPUT

🔗 SO LANGCHAIN … WHY?



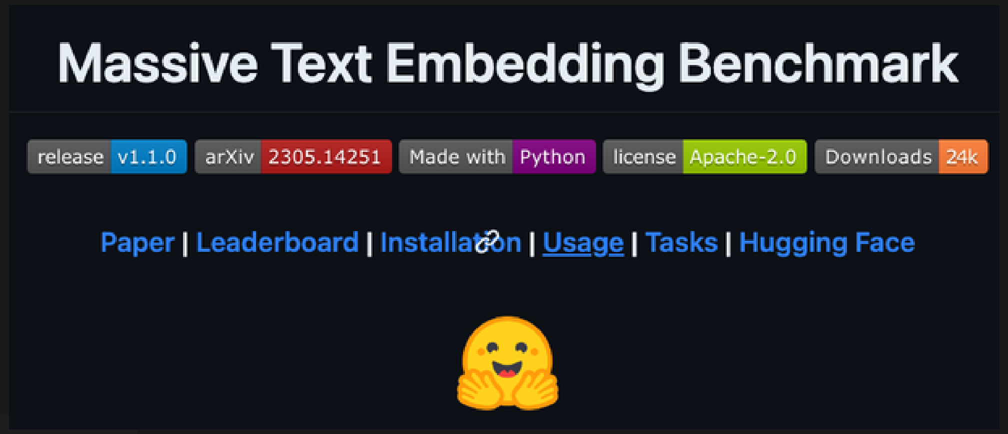
“LLMs in isolation is often insufficient for creating a truly powerful app - the real power comes when you can **combine them** with **other sources of computation or knowledge.**”

~ Harrison Chase, Creator of LangChain

Chain - THE Abstraction

- Connect stuff to other stuff

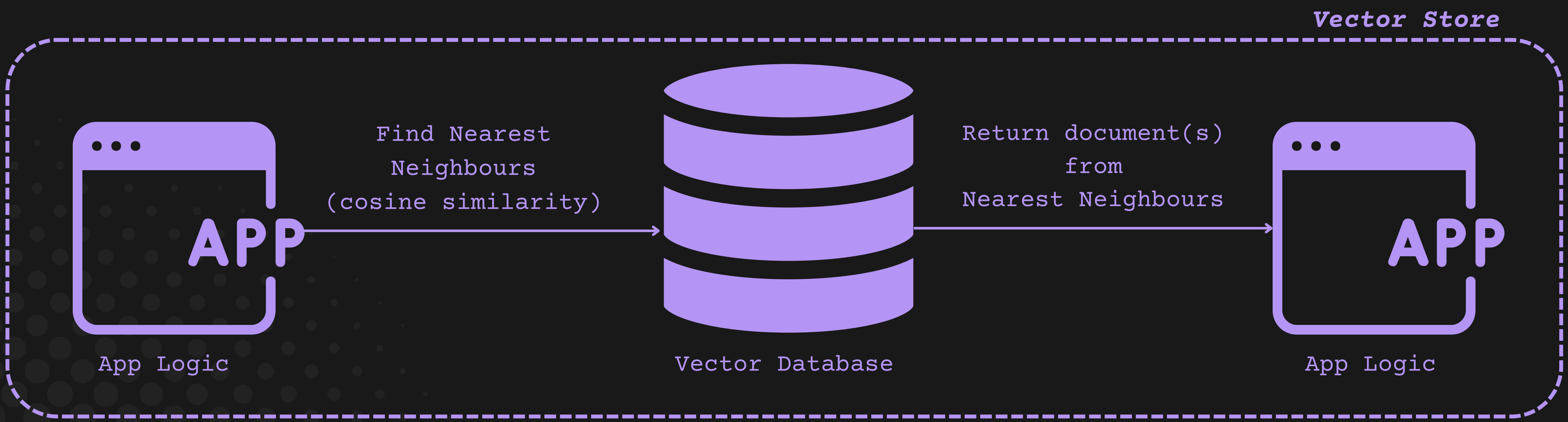
```
from langchain.document_loaders.csv_loader import CSVLoader
from langchain.text_splitter import RecursiveCharacterTextSplitter
from langchain.embeddings import CacheBackedEmbeddings, HuggingFaceEmbeddings
from langchain.vectorstores import FAISS
from langchain.storage import LocalFileStore
```



Massive Text Embedding Benchmark

release v1.1.0 | arXiv 2305.14251 | Made with Python | license Apache-2.0 | Downloads 24k

Paper | Leaderboard | Installation | Usage | Tasks | Hugging Face



Vector Store

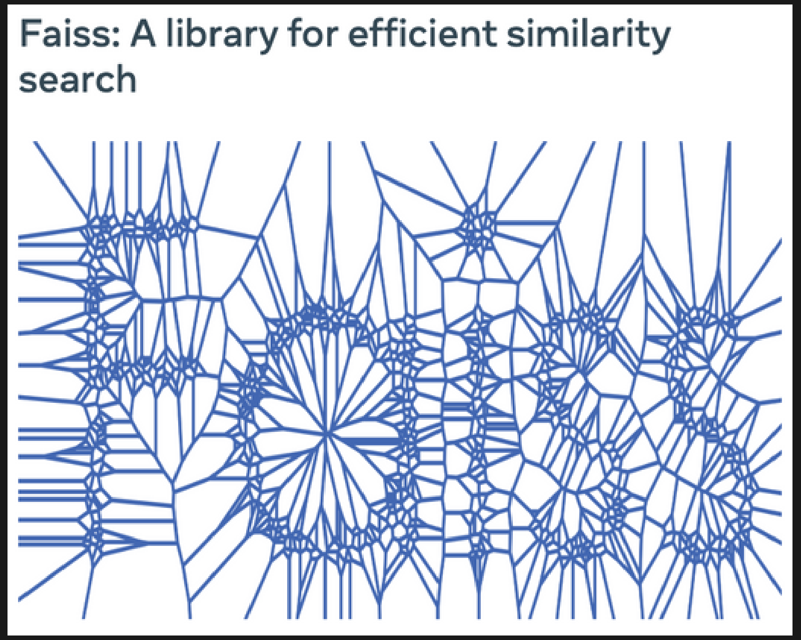
App Logic

Find Nearest Neighbours (cosine similarity)

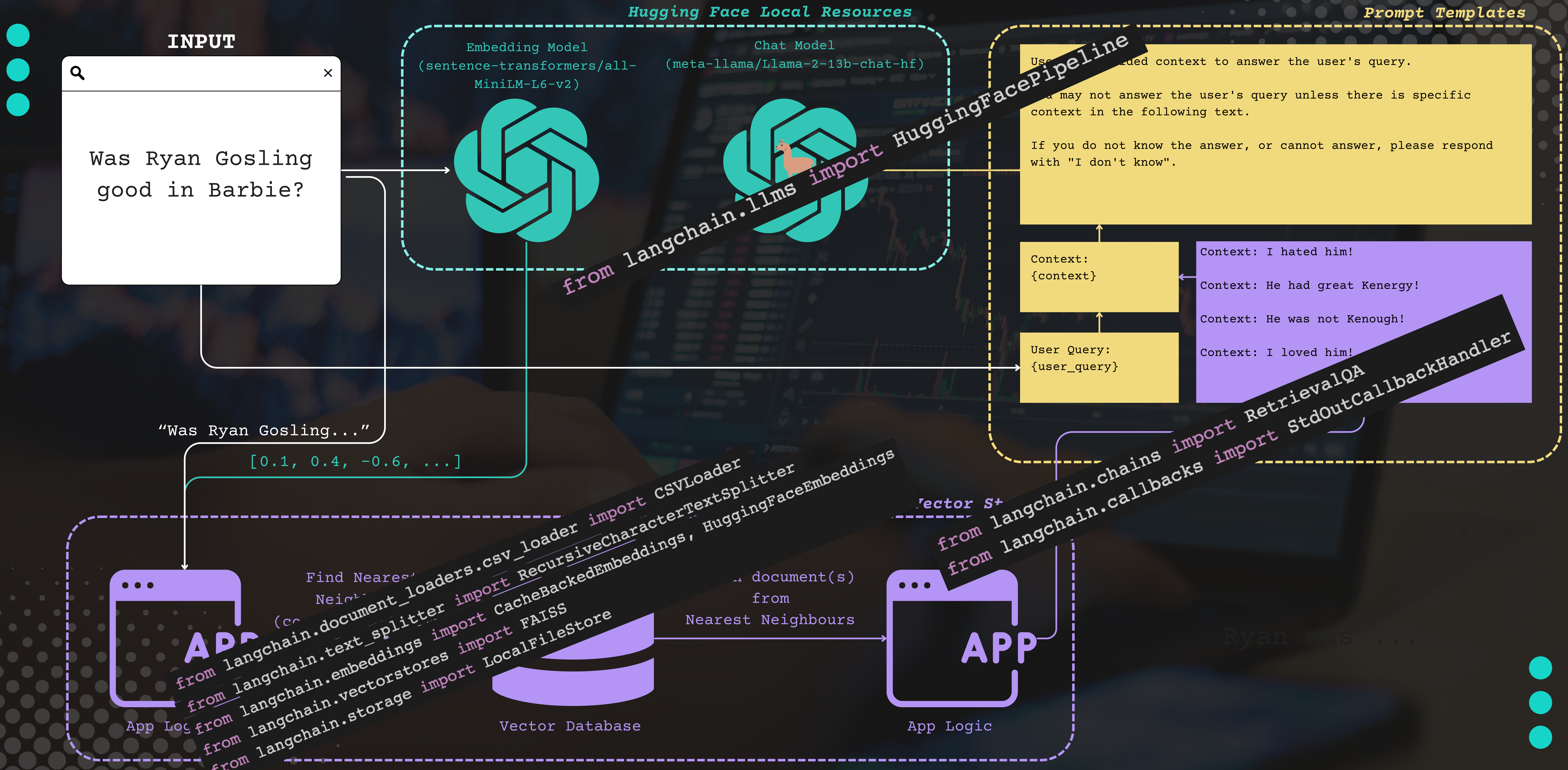
Vector Database

Return document(s) from Nearest Neighbours

App Logic



Faiss: A library for efficient similarity search



Was Ryan Gosling good in Barbie?

Us ided context to answer the user's query.

u may not answer the user's query unless there is specific context in the following text.

If you do not know the answer, or cannot answer, please respond with "I don't know".

Context:
{context}

User Query:
{user_query}

Context: I hated him!

Context: He had great Kenergy!

Context: He was not Kenough!

Context: I loved him!

```
from langchain.llms import HuggingFacePipeline
```

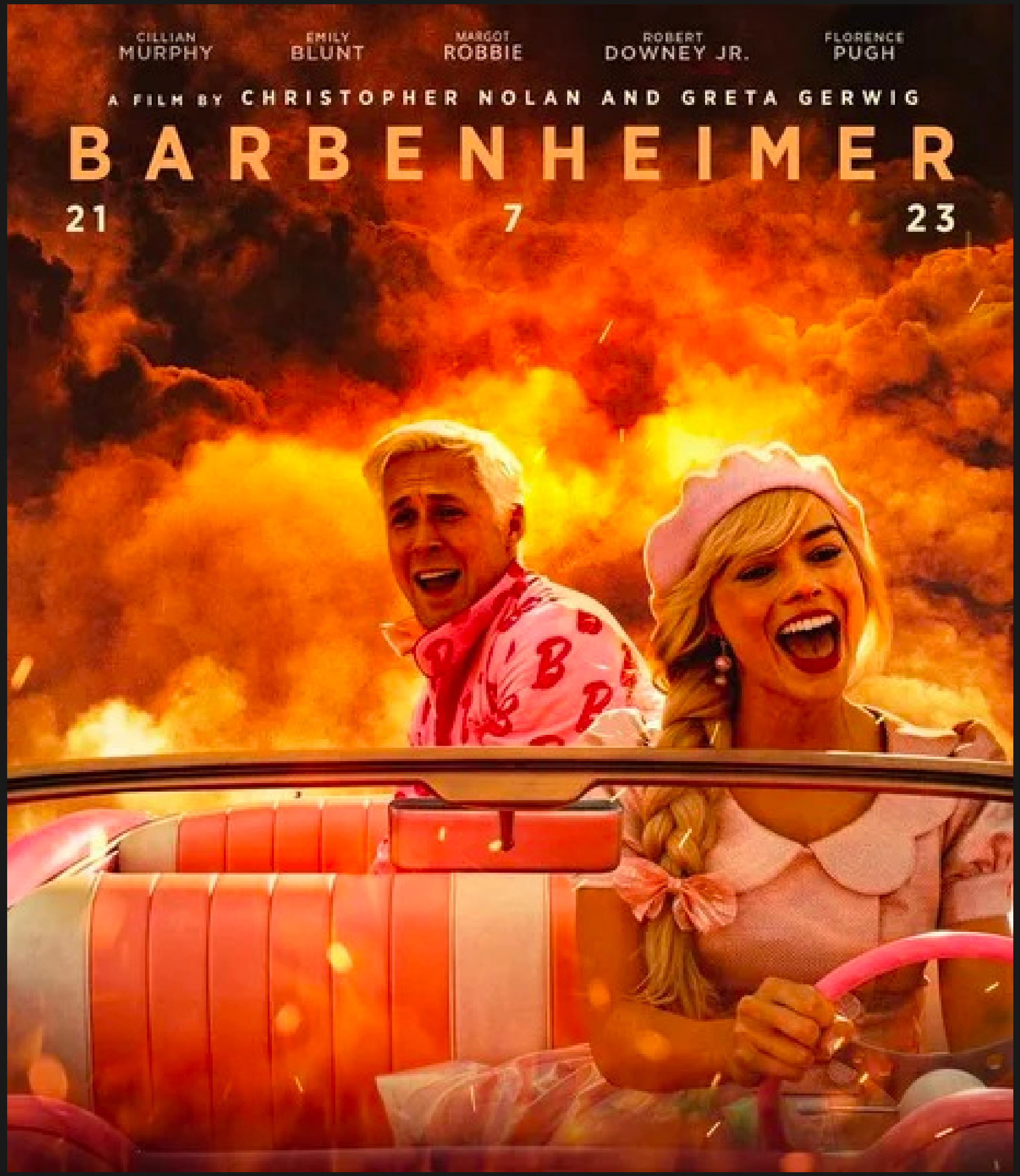
```
from langchain.document_loaders.csv_loader import CSVLoader
from langchain.text_splitter import RecursiveCharacterTextSplitter
from langchain.embeddings import CacheBackedEmbeddings, HuggingFaceEmbeddings
from langchain.vectorstores import FAISS
from langchain.storage import LocalFileStore
```

```
from langchain.chains import RetrievalQA
from langchain.callbacks import StdOutCallbackHandler
```

LET'S BREAK IT DOWN

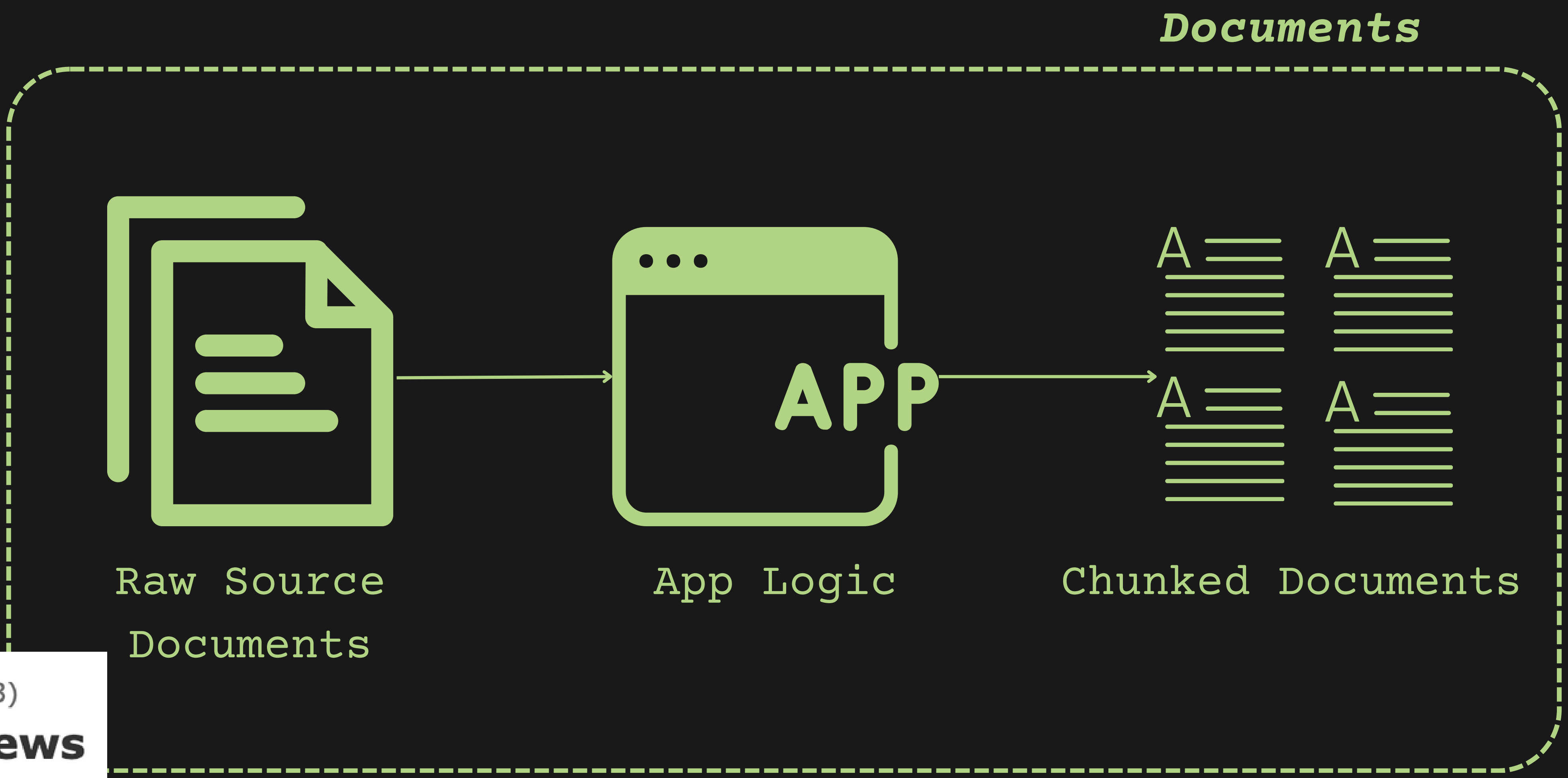
3 EASY PIECES

1. IMDB **Data** Preparation
2. Creating an **Index**
3. Building a **Retrieval Chain**



TASK 1: DATA PREPARATION

1. Download IMDB data
2. Load CSV
3. Parse CSV
4. Chunk text

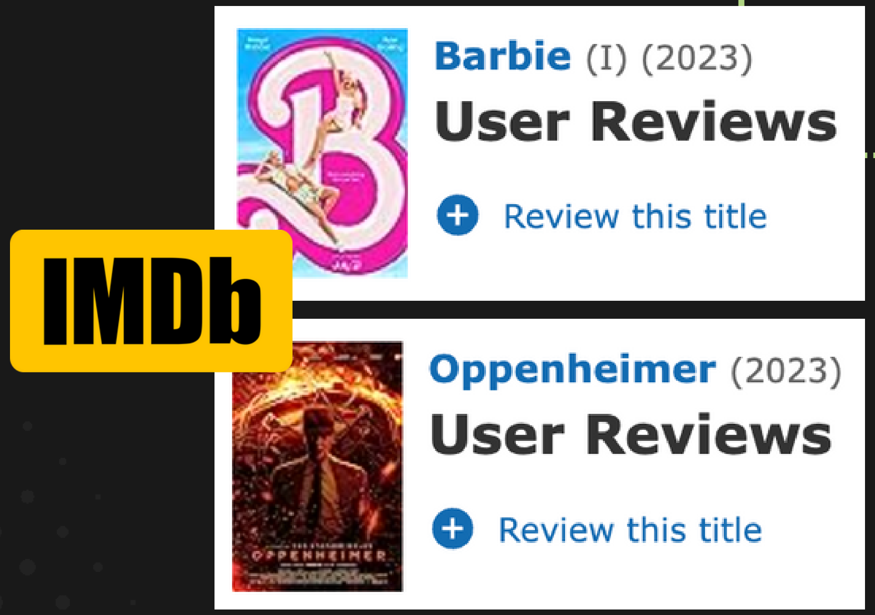


Documents

Raw Source Documents

App Logic

Chunked Documents



Barbie (I) (2023)

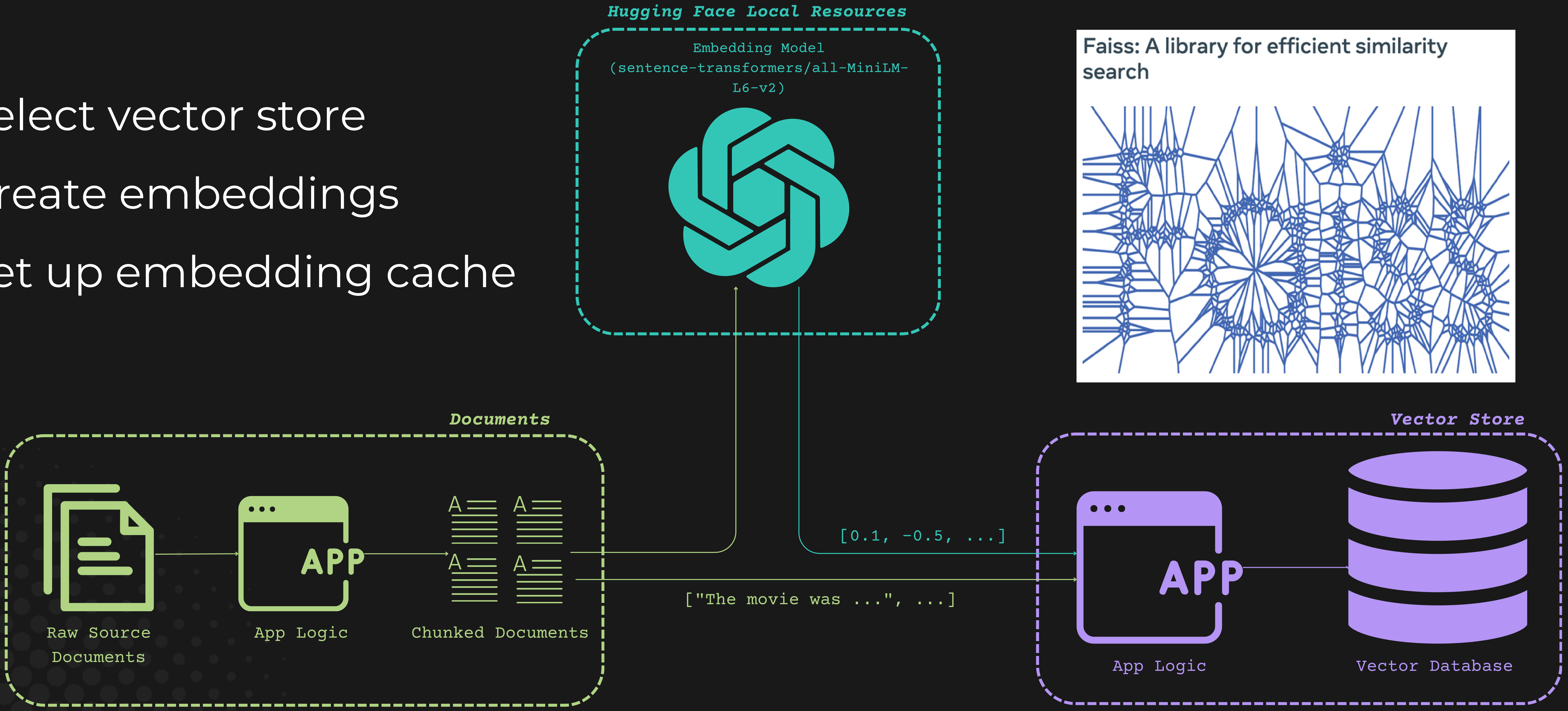
User Reviews

Oppenheimer (2023)

User Reviews

TASK 2: CREATING AN INDEX

1. Select vector store
2. Create embeddings
3. Set up embedding cache



Faiss: A library for efficient similarity search

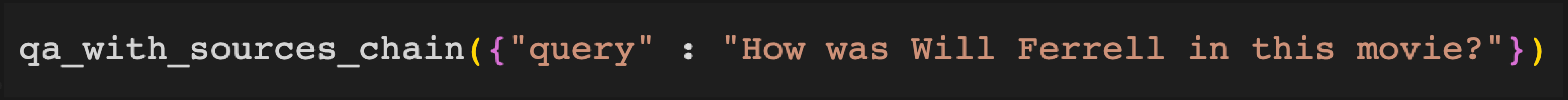
SETTING UP OUR INDEX!

Presented by

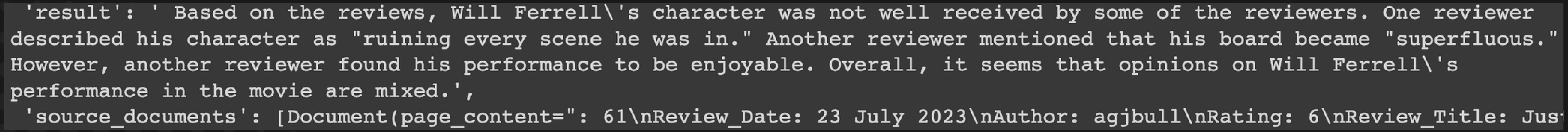
Chris Alexiuk, LLM Wizard 🪄

TASK 3: BUILDING A RETRIEVAL CHAIN

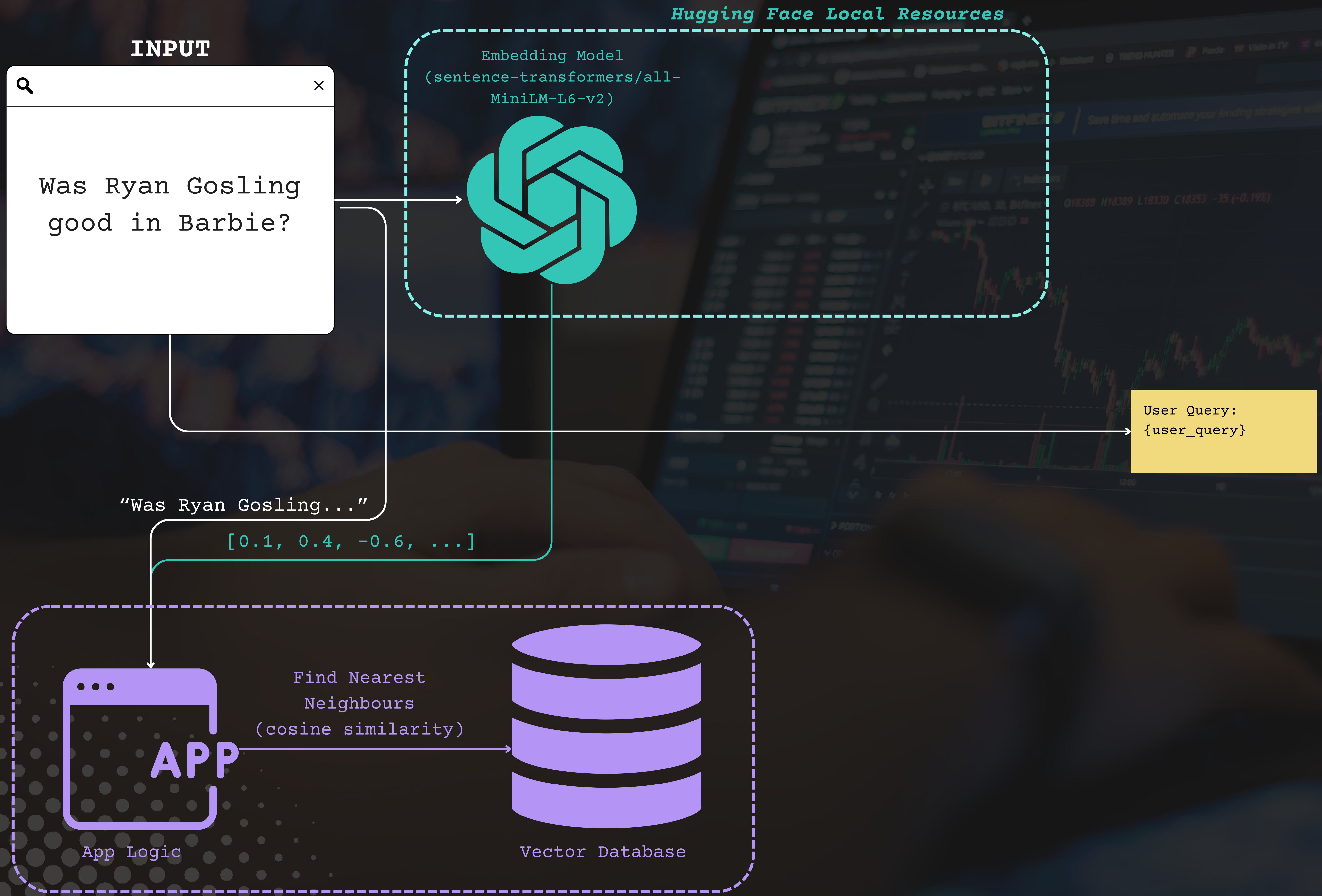
1. Make LLM small
    a. Make a 13B parameter model run on *less than* 15 GB of GPU RAM
    b.
    Tim Dettmer's `bitsandbytes` as well as `accelerate` and `transformers`
2. Retrieve answers
3. Return sources



```
qa_with_sources_chain({"query" : "How was Will Ferrell in this movie?"})
```



```
 'result': ' Based on the reviews, Will Ferrell\'s character was not well received by some of the reviewers. One reviewer
described his character as "ruining every scene he was in." Another reviewer mentioned that his board became "superfluous."
However, another reviewer found his performance to be enjoyable. Overall, it seems that opinions on Will Ferrell\'s
performance in the movie are mixed.',
 'source_documents': [Document(page_content=": 61\nReview_Date: 23 July 2023\nAuthor: agjbull\nRating: 6\nReview_Title: Jus
```



Hugging Face Local Resources

INPUT

Was Ryan Gosling good in Barbie?

Embedding Model
(sentence-transformers/all-MiniLM-L6-v2)

User Query:
{user_query}

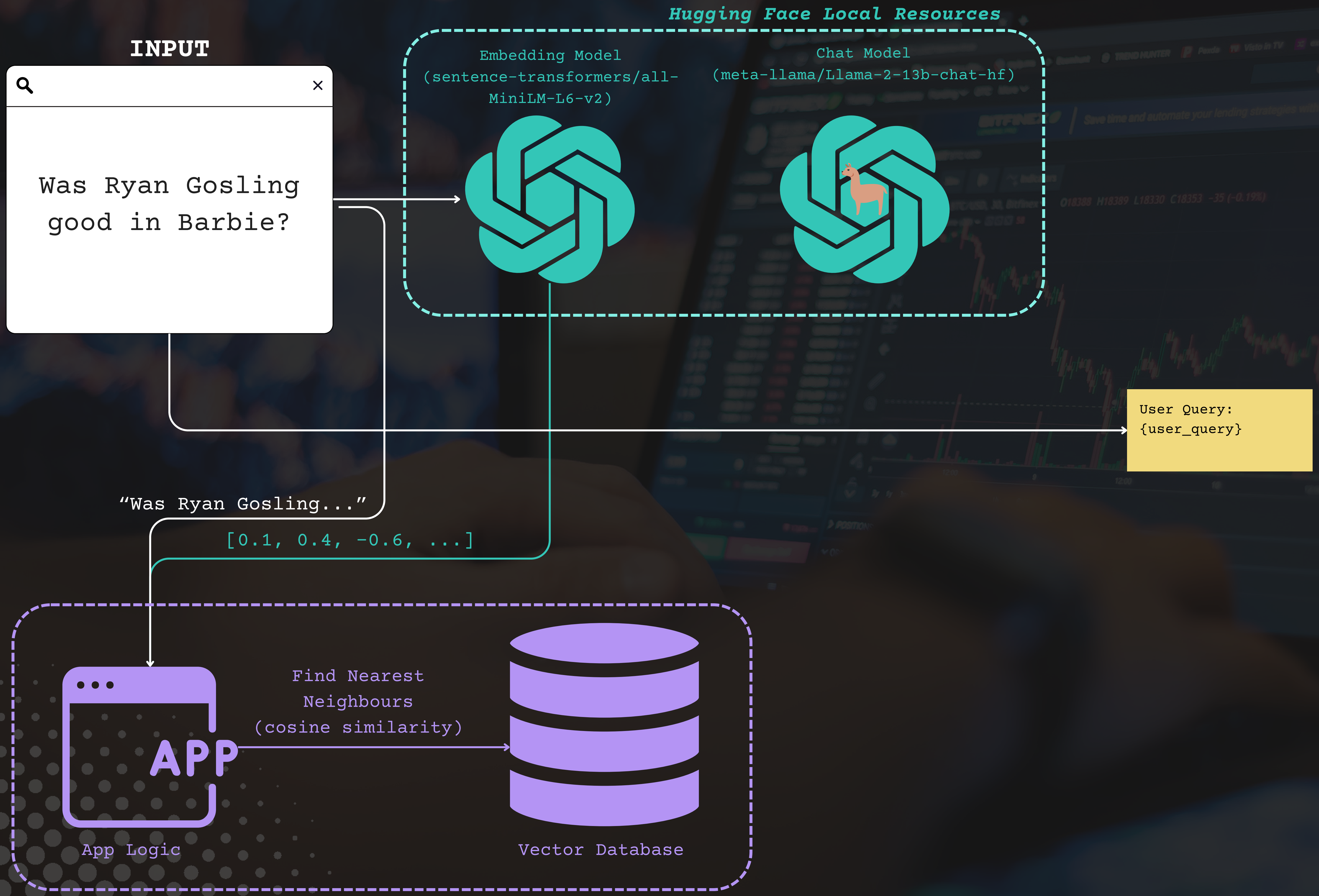
“Was Ryan Gosling...”

[0.1, 0.4, -0.6, ...]

Find Nearest
Neighbours
(cosine similarity)

App Logic

Vector Database



INPUT

Was Ryan Gosling good in Barbie?

Hugging Face Local Resources

Embedding Model (sentence-transformers/all-MiniLM-L6-v2)

Chat Model (meta-llama/Llama-2-13b-chat-hf)

“Was Ryan Gosling...”

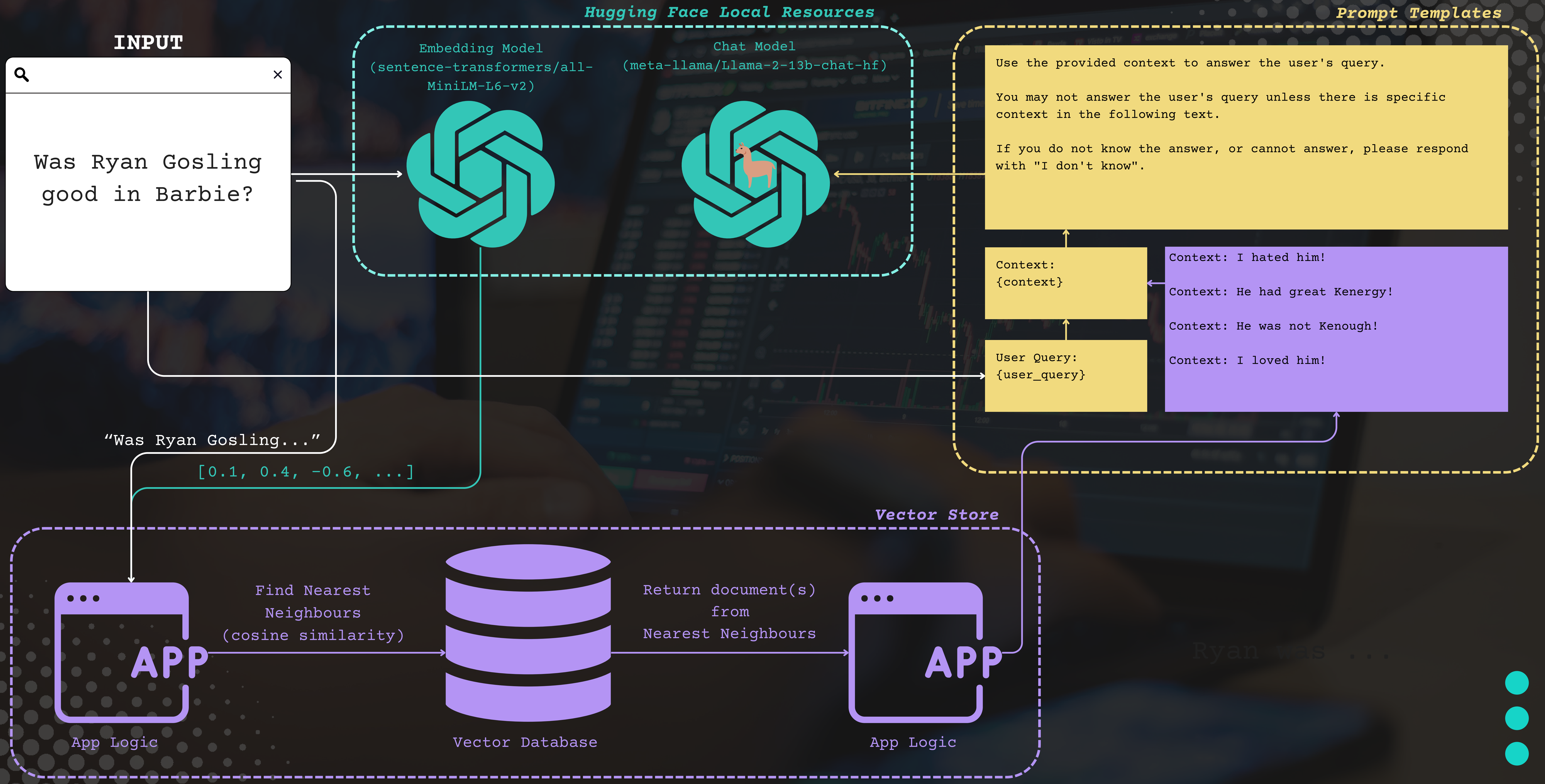
[0.1, 0.4, -0.6, ...]

User Query:
{user_query}

Find Nearest Neighbours (cosine similarity)

App Logic

Vector Database



INPUT

Was Ryan Gosling good in Barbie?

Hugging Face Local Resources

Embedding Model
(sentence-transformers/all-MiniLM-L6-v2)

Chat Model
(meta-llama/Llama-2-13b-chat-hf)

Prompt Templates

Use the provided context to answer the user's query.

You may not answer the user's query unless there is specific context in the following text.

If you do not know the answer, or cannot answer, please respond with "I don't know".

Context:
{context}

User Query:
{user_query}

Context: I hated him!

Context: He had great Kenergy!

Context: He was not Kenough!

Context: I loved him!

"Was Ryan Gosling..."

[0.1, 0.4, -0.6, ...]

Vector Store

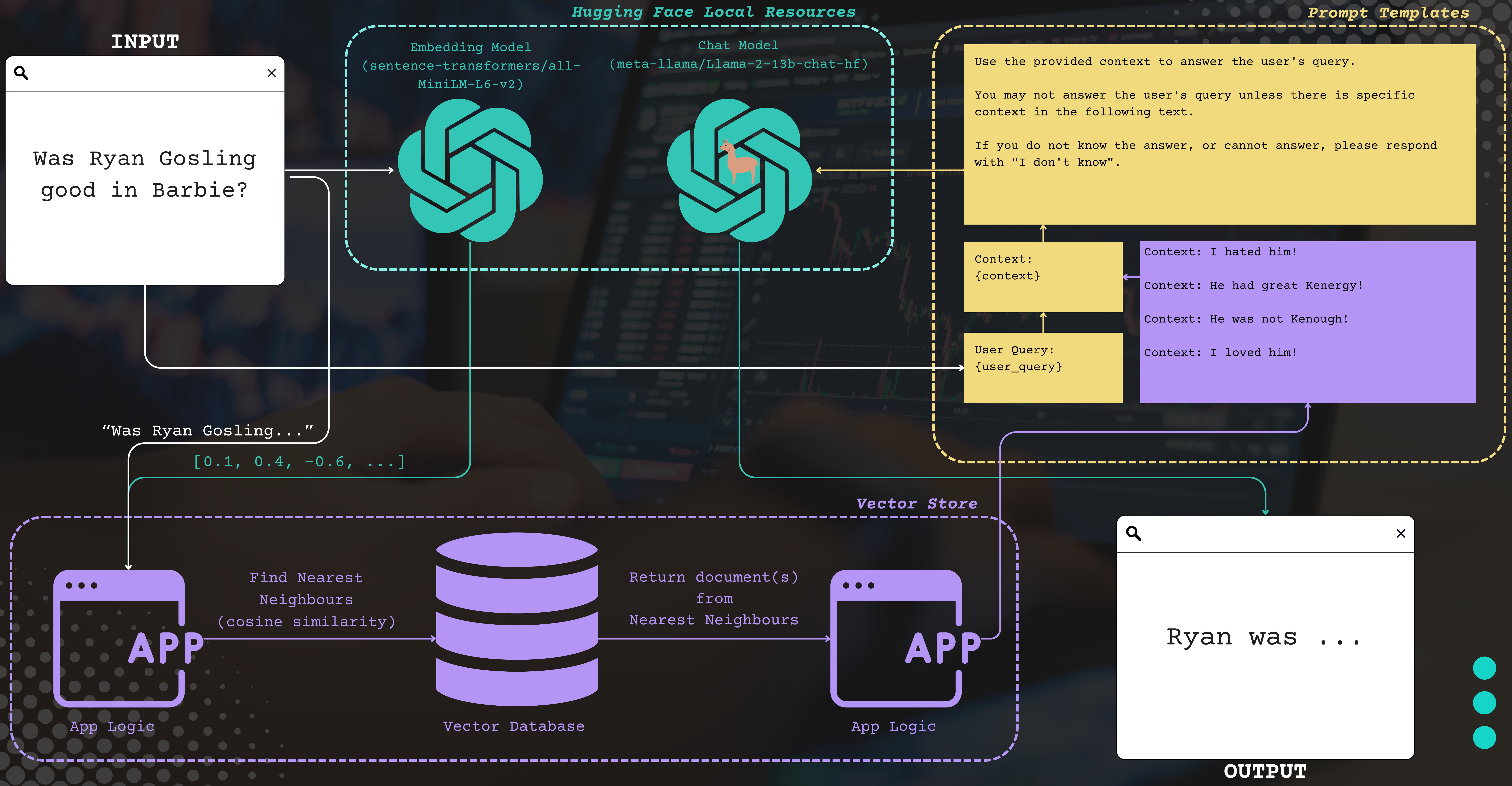
App Logic

Find Nearest Neighbours (cosine similarity)

Vector Database

Return document(s) from Nearest Neighbours

App Logic



INPUT

Was Ryan Gosling good in Barbie?

Hugging Face Local Resources

Embedding Model
(sentence-transformers/all-MiniLM-L6-v2)

Chat Model
(meta-llama/Llama-2-13b-chat-hf)

Prompt Templates

Use the provided context to answer the user's query.

You may not answer the user's query unless there is specific context in the following text.

If you do not know the answer, or cannot answer, please respond with "I don't know".

Context:
{context}

User Query:
{user_query}

Context: I hated him!

Context: He had great Kenergy!

Context: He was not Kenough!

Context: I loved him!

“Was Ryan Gosling...”

[0.1, 0.4, -0.6, ...]

Vector Store

Ryan was ...

OUTPUT

COMPLETING THE RAG!

Presented by
Chris Alexiuk, LLM Wizard 🪄

CONCLUSIONS

- LLMs should be fact-checked !
- RAG ~ QA with prompts augmented by retrieval
- 🧑‍💻 Making a 13B parameter SOTA LLM is cool!
- 💅 Not Kenough data
- 💣 Oppenheimer gave existentialist vibes

QUESTIONS THAT WE CAN ANSWER?

Prompt: Retrieval Augmented Generation

Prompt: Retrieval Augmented Question Answering

Thank you!