GENERATIVE AI MEET UP #3

16th September, 2023 | Binary Labs

Zephania Reuben

Machine Learning Specialist

www.nsoma.me | @nsomazr

Expectation Setting

Building with LLMs

Outline

Scikit-LLM

Custom LLM[GPT]

SCIKIT-LLM

Scikit-Learn

Scikit-Learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms.

* Source: Machine Learning with Scikit Learn

Scikit-Learn Design

Estimator

Transformer

Predictor

SCIKIT-LLM

Scikit-LLM is a Python package that integrates large language models (LLMs)into the scikit-learn framework mainly for text analysis tasks.

.

Some Scikit-LLMs Features

Zero-shot Text Classification

Multi-label Zero-shot Text Classification

Text Translation

Text Summarization

Text Vectorization

Hands On

CUSTOM GPT

#roughly [From Generative AI Meet Up #2]

People(Talents/Labelers)

Data

Computational

Computational Challenge

1 Parameter ~4 Bytes [32 bit float]

1B Parameters ~4 GigaBytes

Only for Model Parameters

175B Parameters ~14000 GigaBytes

(Model parameters, gradients, optimizer states, and temp memory)

Computational Challenge

Single NVIDIA A100 GPU has 80GB Memory Size

Computational Challenge

What is the cost for pre-training GPT-4 with about

1.76 Trillion Parameters *

* source: the-decoder.com

How do we adapt LLMs locally?

Quantization

Multi-GPU Compute
Strategies

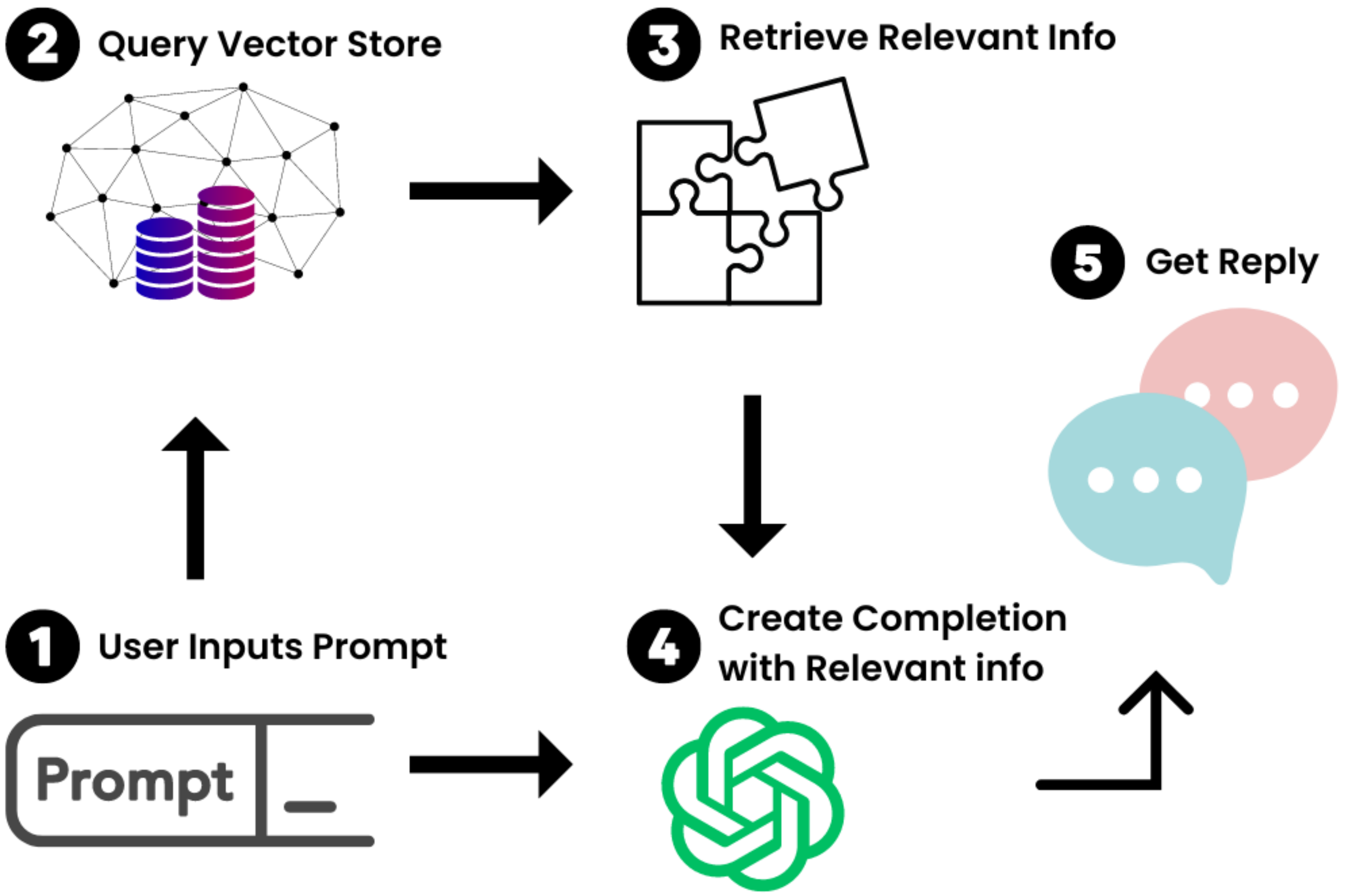
LLMs Fine Tuning

Retrieval Augmented Generation

Why RAG?

Base LLMs (ex. Llama-2-70b,GPT-4 etc.) are only aware of the information that they've been trained on and will fall short when we require them to know information beyond that.

Big Picture



2 Query Vector Store

3 Retrieve Relevant Info

5 Get Reply

1 User Inputs Prompt

4 Create Completion with Relevant info

Tools

LangChain

Llama Index(gpt-index)

OpenAI API

Gradio

LangChain

LangChain is a robust library designed to streamline interaction with large language models (LLMs) providers like OpenAI. It supports other LLM providers as such as Cohere, Bloom, and Huggingface.

LlamaIndex

Llama Index previously known as gpt-index is a data framework which provides a simple, flexible interface to connect LLMs with external data(e.g your private data)

OpenAI API

OpenAI provides APIs to interact and use GPT LLM series in our own applications. To use a GPT model via the OpenAI API, we need to send a request containing the inputs and your API key, and receive a response containing the model’s output.

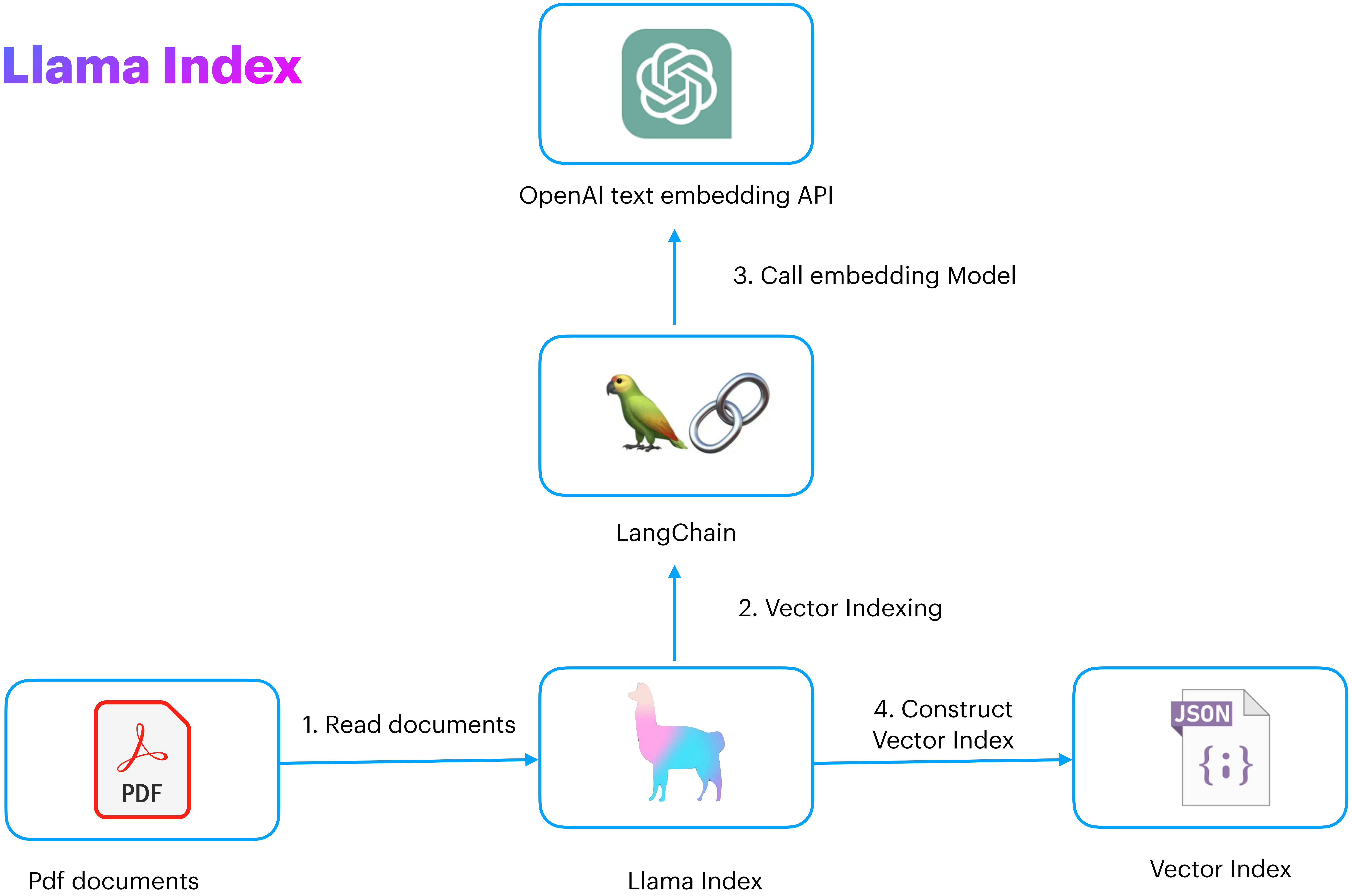
Gradio

Gradio is an open-source Python library that allows developers to quickly create user interfaces for machine learning models. It simplifies the process of building web-based interfaces, interactive applications, and demos for machine learning models, making it accessible to both developers and non-technical users.

Get API Key

http://platform.openai.com/account/api-keys

Creating Llama Index



OpenAI text embedding API

3. Call embedding Model

LangChain

2. Vector Indexing

Pdf documents

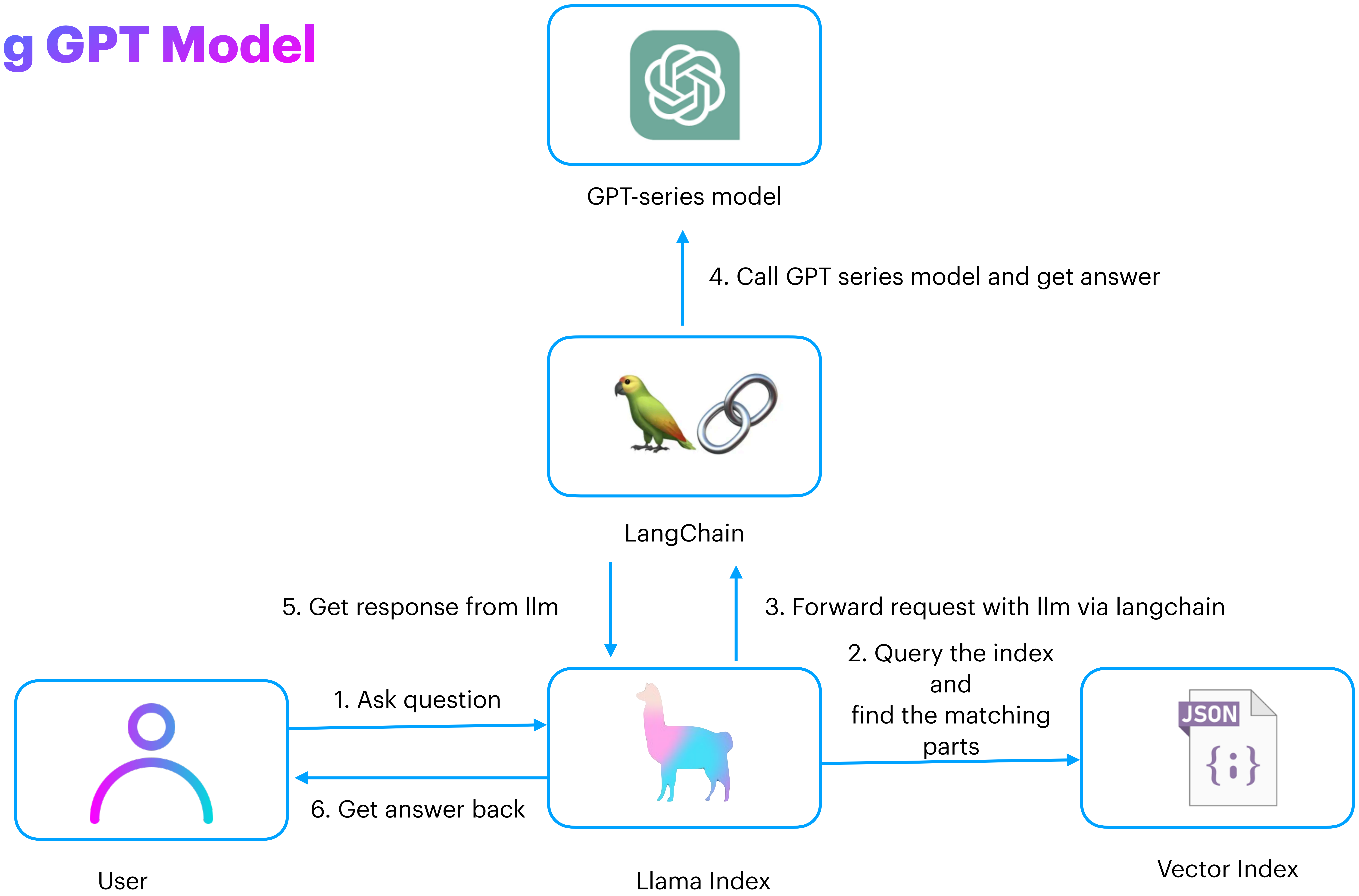
1. Read documents

Llama Index

4. Construct Vector Index

Vector Index

Querying GPT Model



GPT-series model

4. Call GPT series model and get answer

LangChain

5. Get response from llm

3. Forward request with llm via langchain

2. Query the index and find the matching parts

1. Ask question

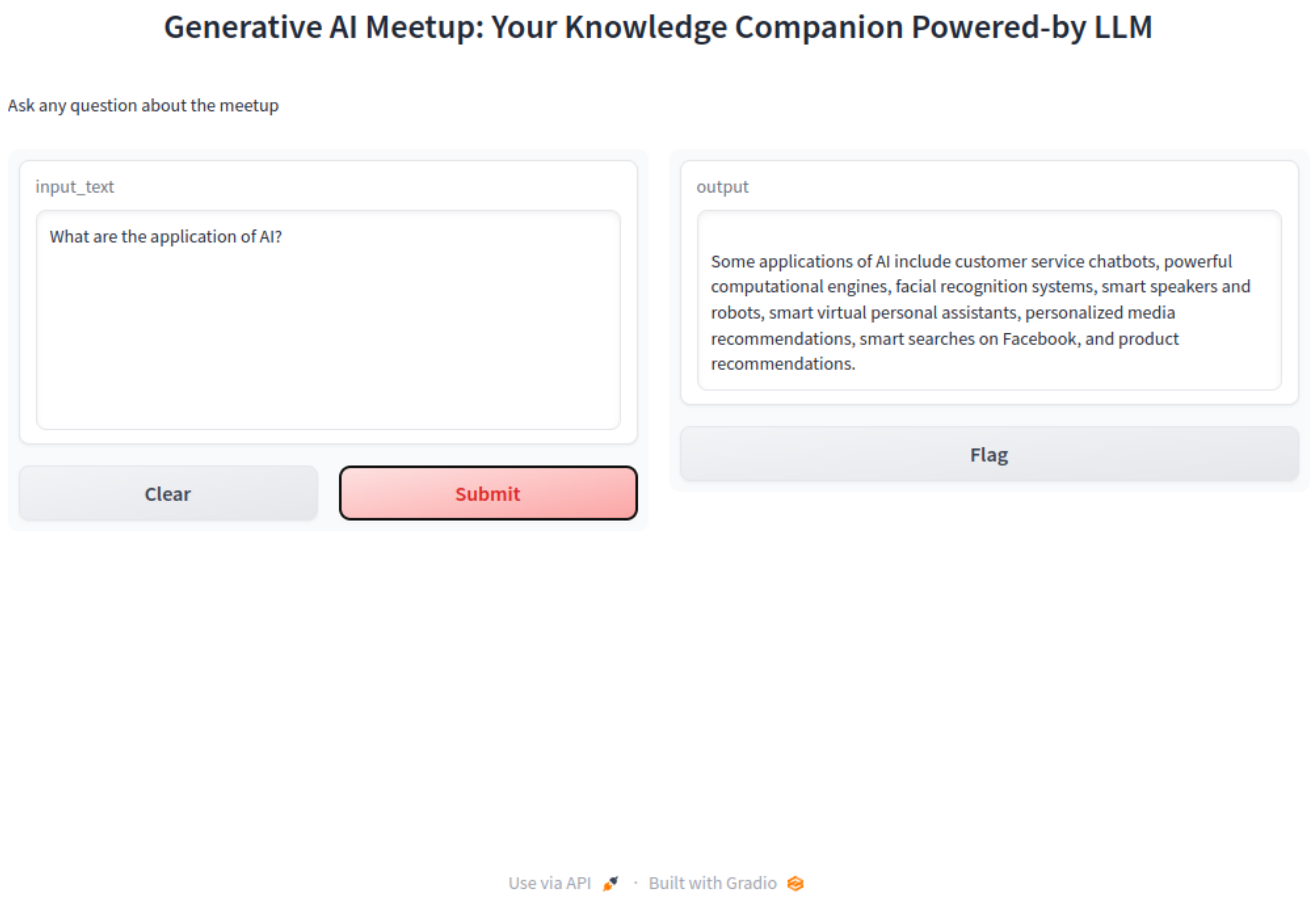
6. Get answer back

User

Llama Index

Vector Index

Building UI



Generative AI Meetup: Your Knowledge Companion Powered-by LLM

Ask any question about the meetup

input_text

What are the application of AI?

output

Some applications of AI include customer service chatbots, powerful computational engines, facial recognition systems, smart speakers and robots, smart virtual personal assistants, personalized media recommendations, smart searches on Facebook, and product recommendations.

Hands On

Next Step

Evaluating Performance

Referencies

1.https://blog.dataiku.com/large-language-model-chatgpt/

2.https://medium.com/llamaindex-blog/build-a-chatgpt-with-your-private-data-using-llamaindex-and-mongodb-b09850eb154c/

Thank You