GOTC 2023

全球开源技术峰会

THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE

OPEN SOURCE，INTO THE FUTURE

「AI is Everywhere」专场

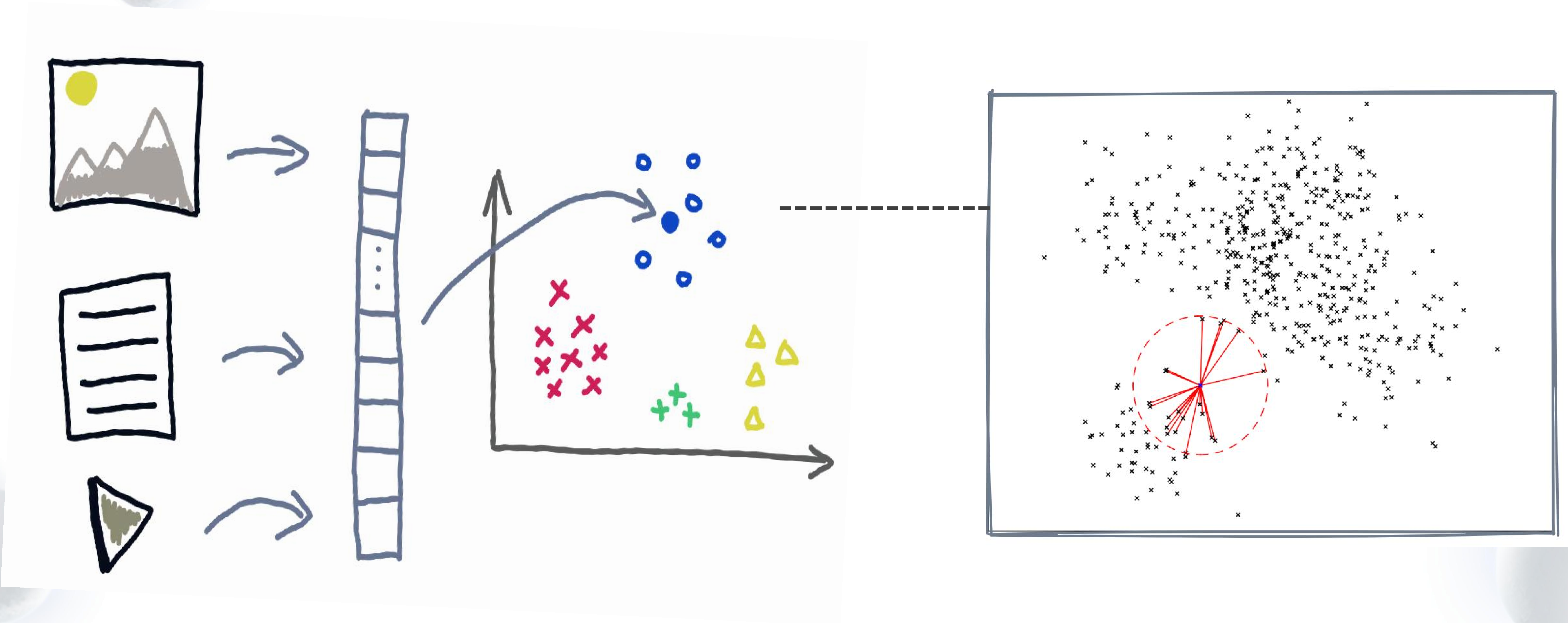
向量数据库：面向 AIGC 的海量记忆体

李晨 2023年5月28日

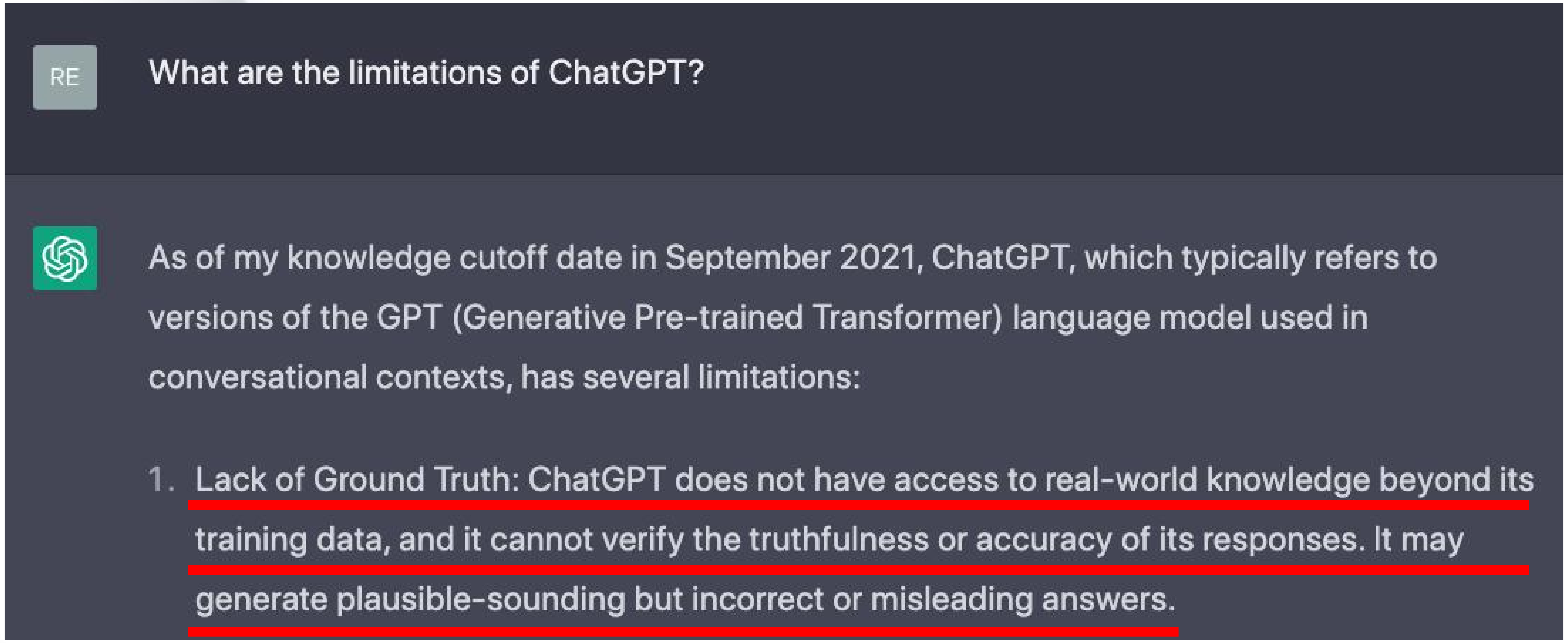
zilliz

Building Memory & Cache for AIGC with VectorDB

Embedding and Vector Search



LLM Limitations



RE

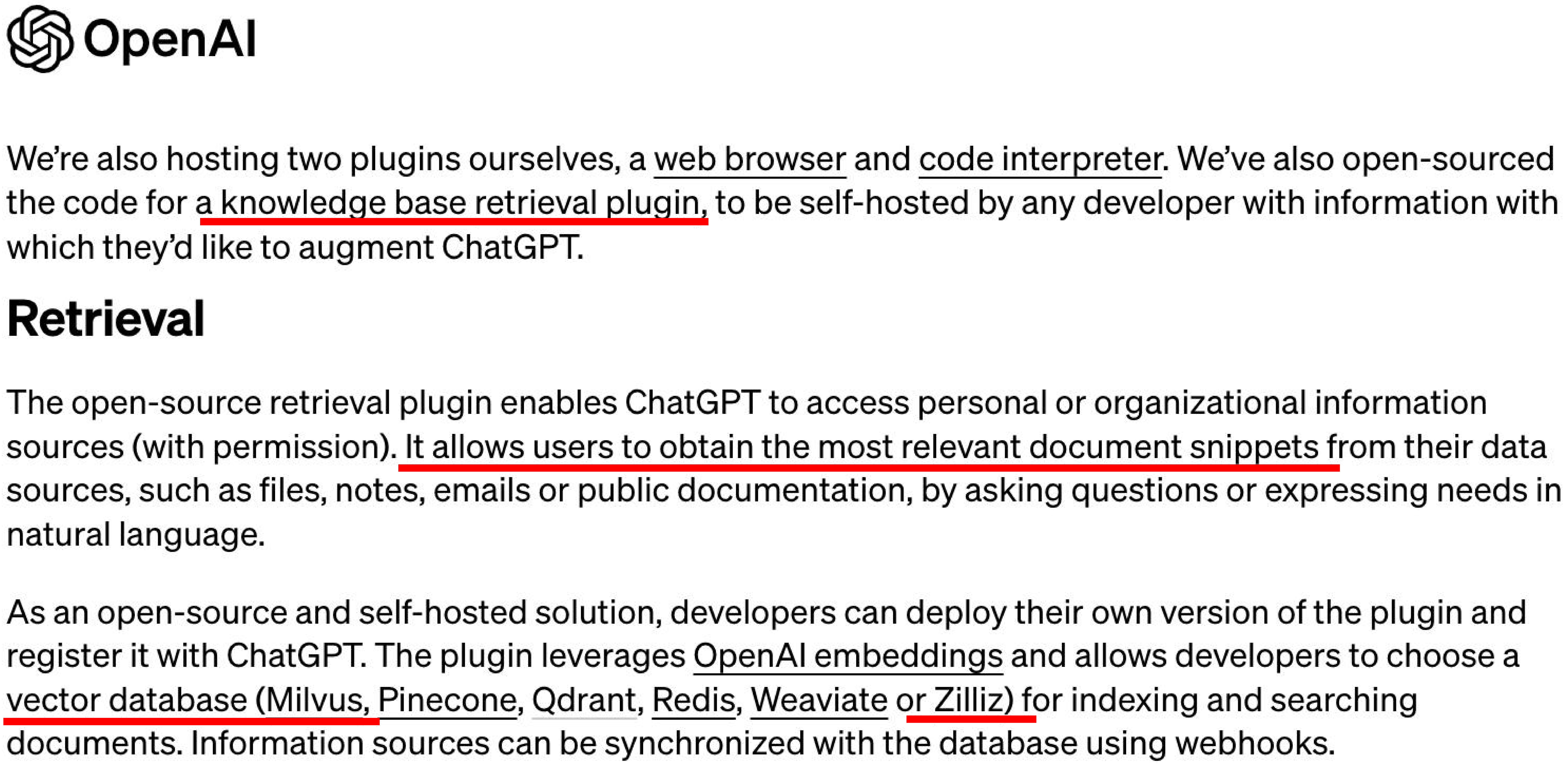
What are the limitations of ChatGPT?

As of my knowledge cutoff date in September 2021, ChatGPT, which typically refers to versions of the GPT (Generative Pre-trained Transformer) language model used in conversational contexts, has several limitations:

1. Lack of Ground Truth: ChatGPT does not have access to real-world knowledge beyond its training data, and it cannot verify the truthfulness or accuracy of its responses. It may generate plausible-sounding but incorrect or misleading answers.

- Domain–specific knowledge WRONG!
- Real–time informatio WRONG!

Knowledge Retrieval for LLMs



OpenAI

We’re also hosting two plugins ourselves, a web browser and code interpreter. We’ve also open-sourced the code for a knowledge base retrieval plugin, to be self-hosted by any developer with information with which they’d like to augment ChatGPT.

Retrieval

The open-source retrieval plugin enables ChatGPT to access personal or organizational information sources (with permission). It allows users to obtain the most relevant document snippets from their data sources, such as files, notes, emails or public documentation, by asking questions or expressing needs in natural language.

As an open-source and self-hosted solution, developers can deploy their own version of the plugin and register it with ChatGPT. The plugin leverages OpenAI embeddings and allows developers to choose a vector database (Milvus, Pinecone, Qdrant, Redis, Weaviate or Zilliz) for indexing and searching documents. Information sources can be synchronized with the database using webhooks.

LLM + VectorDB + Prompt

LLM

Analysis and reasoning unit
& natural language adapter

VectorDB

Long–term memory
& knowledgebase

Prompt

Business logic
modeling

Our Partners: Zilliz x AI x Cloud

zilliz

NVIDIA

Hugging Face

文心一言

MINIMAX

aws

Google Cloud

金山云

阿里云

towhee

https://towhee.io/

Pipeline & Ecosystem Integration

LLM + VectorDB + Prompt chained together

Chat with
Open Source Software

Enhanced ChatGPT with documentation, issues, blog posts, community Q&A as knowledge bases. Built for every community and developer.

Ask a question about GitHub projects...

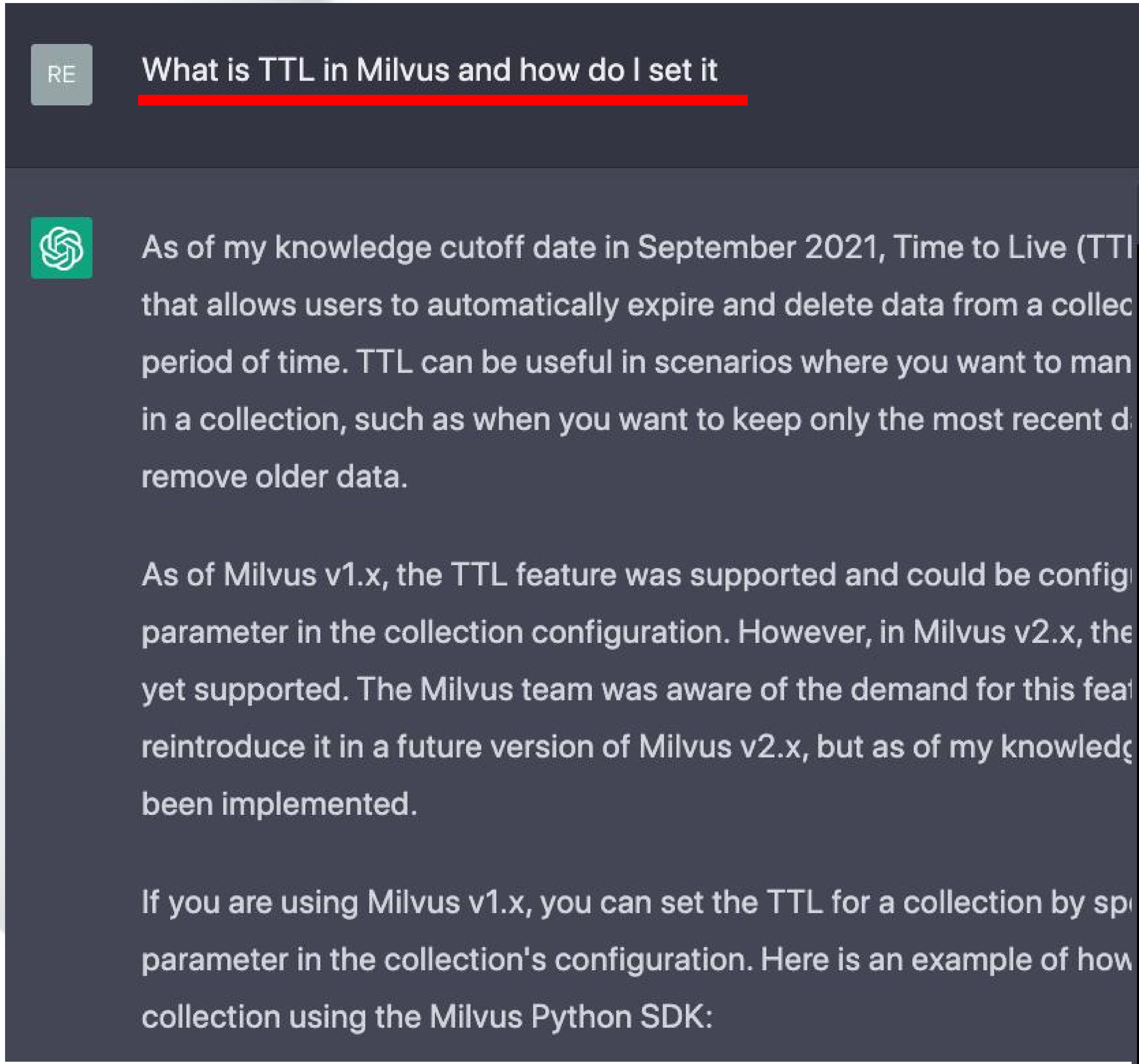
- Where can I find the new stable diffusion models? →
- Where can I find a list of recommended books on deep learning? →
- Can PyTorch Image Models be used for image classification? →
- What are the basics of DeFi development? Tell me all you know. →

Choose one of your favorite projects to chat with

© 2023 osschat. All rights reserved.Terms of service

Powered by Zilliz vector database cloud service and OpenAI's ChatGPT

Answer from ChatGPT

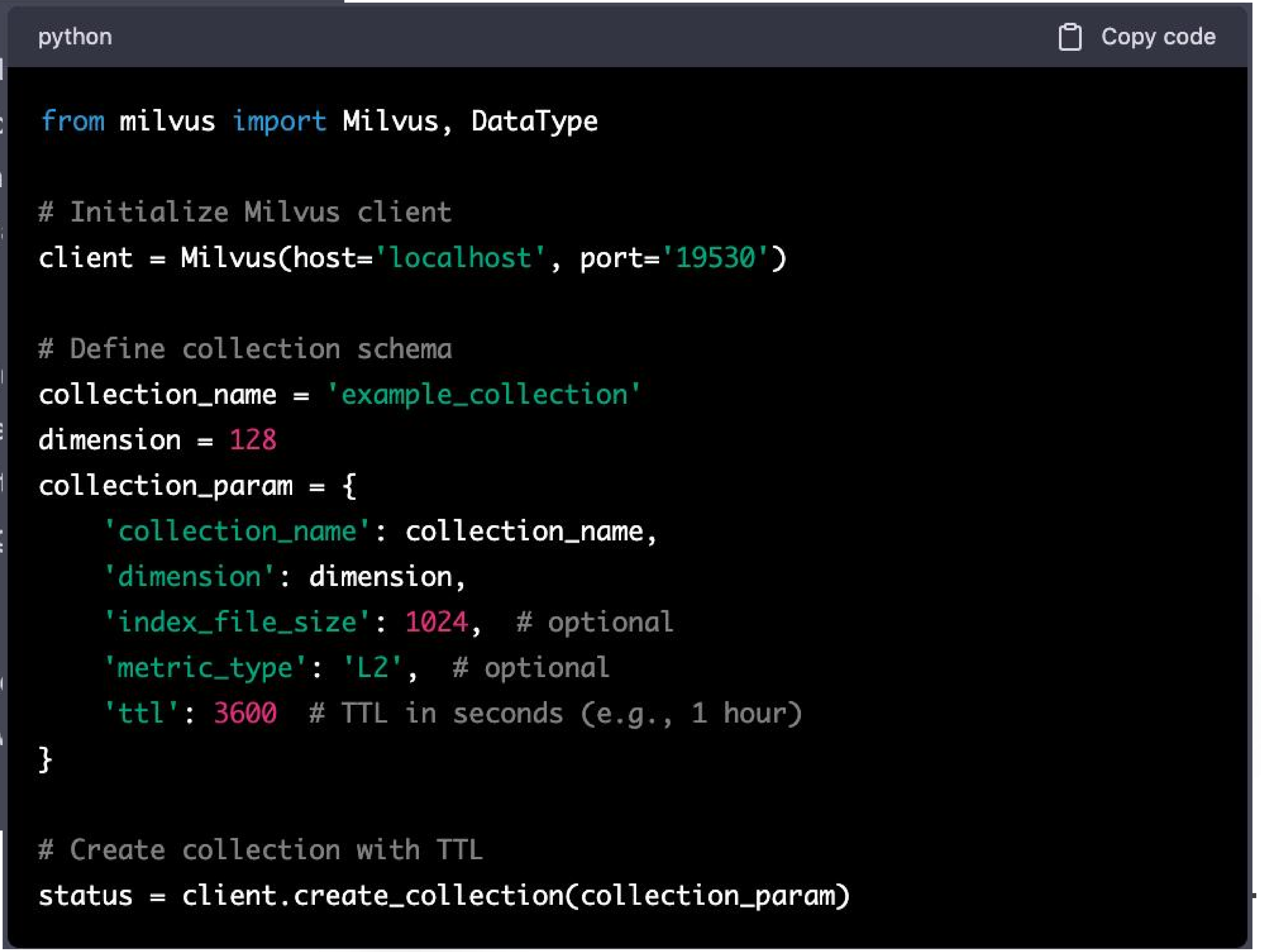


What is TTL in Milvus and how do I set it

As of my knowledge cutoff date in September 2021, Time to Live (TTI
that allows users to automatically expire and delete data from a collec
period of time. TTL can be useful in scenarios where you want to man
in a collection, such as when you want to keep only the most recent d
remove older data.

As of Milvus v1.x, the TTL feature was supported and could be config
parameter in the collection configuration. However, in Milvus v2.x, the
yet supported. The Milvus team was aware of the demand for this feat
reintroduce it in a future version of Milvus v2.x, but as of my knowledg
been implemented.

If you are using Milvus v1.x, you can set the TTL for a collection by sp
parameter in the collection's configuration. Here is an example of how
collection using the Milvus Python SDK:



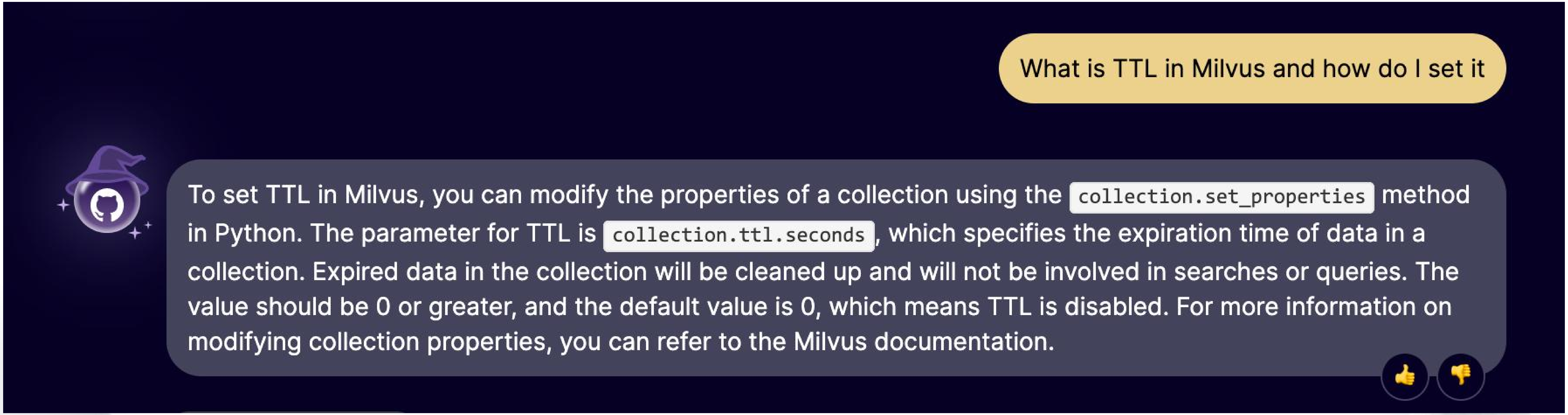
```python
from milvus import Milvus, DataType

# Initialize Milvus client
client = Milvus(host='localhost', port='19530')

# Define collection schema
collection_name = 'example_collection'
dimension = 128
collection_param = {
    'collection_name': collection_name,
    'dimension': dimension,
    'index_file_size': 1024,  # optional
    'metric_type': 'L2',  # optional
    'ttl': 3600  # TTL in seconds (e.g., 1 hour)
}

# Create collection with TTL
status = client.create_collection(collection_param)
```

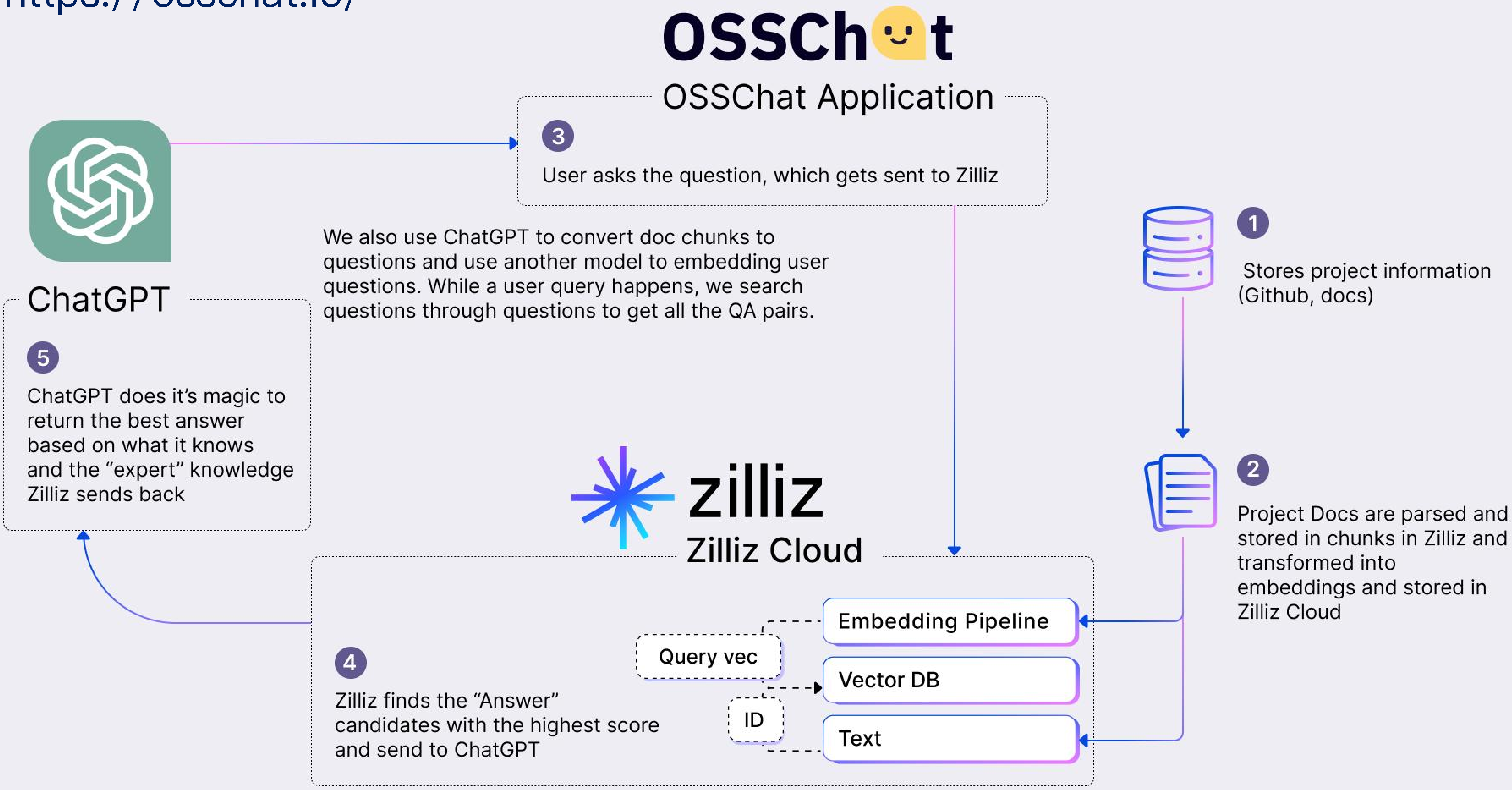
Answer from OSSChat



What is TTL in Milvus and how do I set it

To set TTL in Milvus, you can modify the properties of a collection using the `collection.set_properties` method in Python. The parameter for TTL is `collection.ttl.seconds`, which specifies the expiration time of data in a collection. Expired data in the collection will be cleaned up and will not be involved in searches or queries. The value should be 0 or greater, and the default value is 0, which means TTL is disabled. For more information on modifying collection properties, you can refer to the Milvus documentation.

https://osschat.io/



OSSChat

OSSChat Application

3

User asks the question, which gets sent to Zilliz

ChatGPT

5

ChatGPT does it’s magic to return the best answer based on what it knows and the “expert” knowledge Zilliz sends back

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 the 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

zilliz

Zilliz Cloud

4

Zilliz finds the “Answer” candidates with the highest score and send to ChatGPT

Embedding Pipeline

Query vec

Vector DB

ID

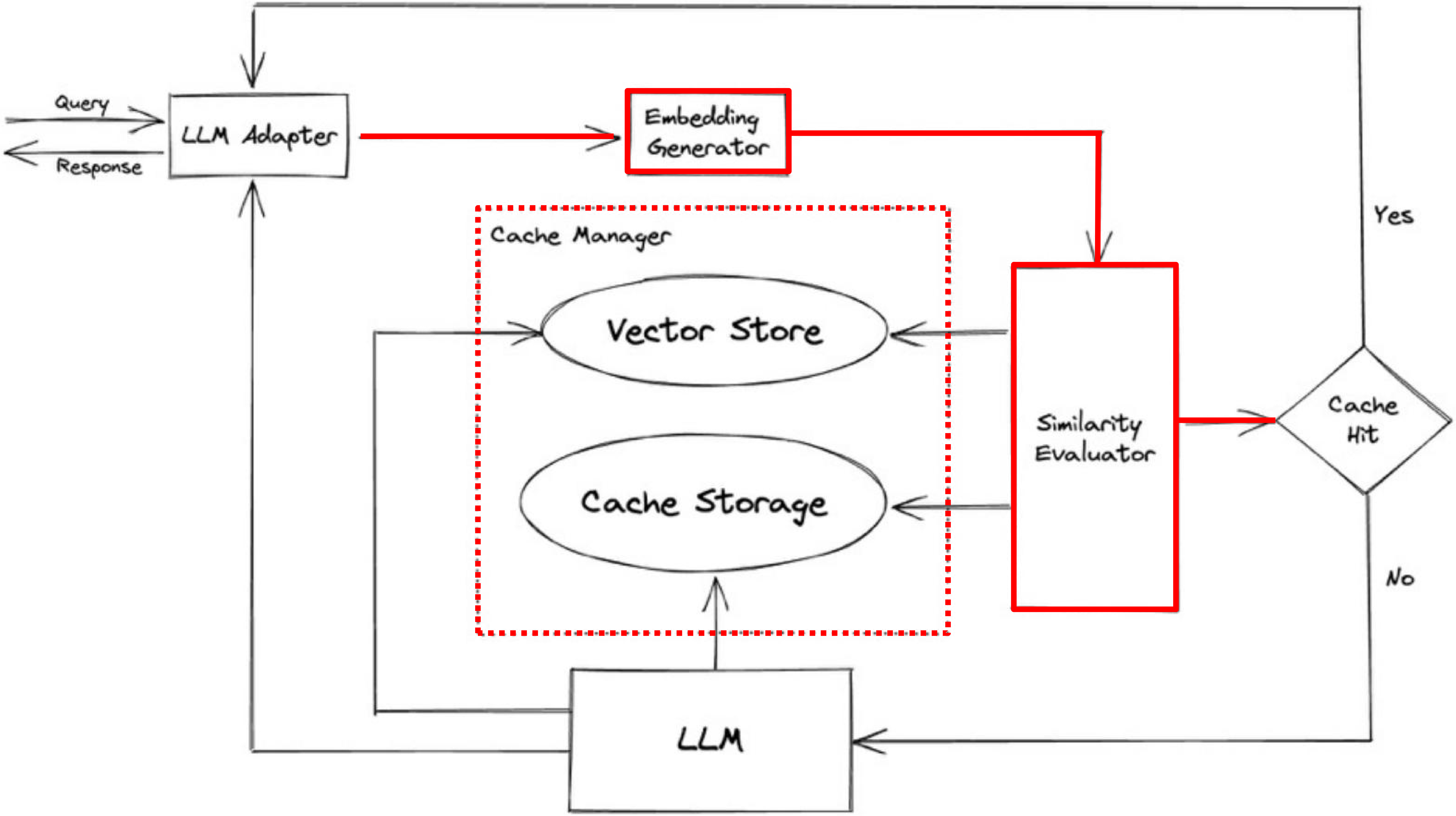
Text

GPTCache :

A Library for Creating Semantic Cache for LLM Queries

Slash Your LLM API Costs by 10x �
Boost Speed by 100x

Architecture



GOTC

Github

公众号

扫码并回复“技术交流”
加入交流群

THANKS

全球开源技术峰会

THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE