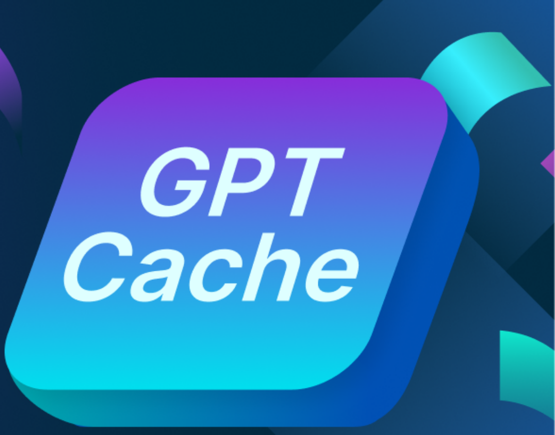
zilliz

GPTCache: An Open-Source Semantic Cache for LLM Applications Enabling Faster Answers and Cost Savings



Key challenges in building LLM-based applications

1. Spiking costs with unnecessary API calls for semantically identical questions that the LLM has already answered, which will waste your money and resources.
2. Poor performance and scalability with high response latency. Additionally, LLM services enforce rate limits, restricting the number of API calls your applications can make to the server within a given timeframe.

What is GPTCache?

GPTCache is an open-source semantic cache designed to improve the efficiency and speed of GPT-based applications by storing and retrieving the responses generated by language models. GPTCache allows users to customize the cache to their specific requirements, offering a range of choices for embedding, similarity assessment, storage location, and eviction policies. Furthermore, GPTCache supports both the OpenAI ChatGPT interface and the Langchain interface, with plans to support more interfaces in the coming months.

GPTCache benefits

Drastic cost reduction in LLM API calls

LLMs charge for each API call. GPTCache helps developers cache LLM responses semantically for similar and repeatedly asked questions, reducing API costs to zero if the cache is hit.

Improved scalability

Caching LLM responses reduces the load on the LLM service, improving your app scalability and preventing bottlenecks while handling growing requests.

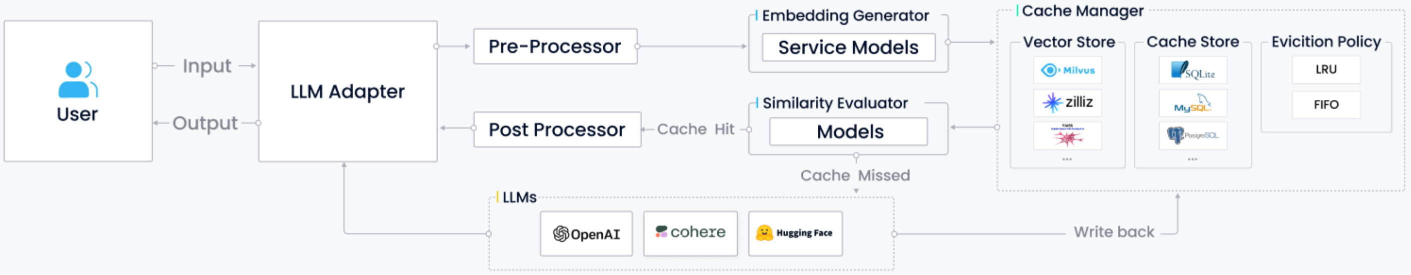
faster in responses

GPTCache can significantly reduce response time for LLM applications. During ChatGPT's peak times, responses can take up to several seconds. However, with GPTCache, applications can retrieve previously requested answers in less than 100 milliseconds.

Better availability

LLM services often set rate limits, restricting the times your app can access the server within a specific timeframe. GPTCache reduces the overall number of API calls and enables your app to scale quickly to handle an increasing volume of queries.

GPTCache Architecture


User
Input
Output
LLM Adapter
Pre-Processor
Post Processor
Cache Hit
Embedding Generator
Service Models
Similarity Evaluator
Models
Cache Missed
LLMs
OpenAI
cohere
Hugging Face
Write back
Cache Manager
Vector Store
Milvus
zilliz
FAISS
...
Cache Store
SQLite
MySQL
PostgreSQL
...
Evicition Policy
LRU
FIFO