zilliz

GPT Cache

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.

How does GPTCache work?

Simply put, GPTCache stores LLMs' responses in the cache. Therefore, when users make similar queries that LLMs had previously responded to, GPTCache searches and returns the results to the users without the need to call the LLM again. Unlike traditional cache systems such as Redis, GPTCache employs semantic caching, which stores and retrieves data through embeddings. It utilizes embedding algorithms to transform the user queries and LLMs' responses into embeddings and conducts similarity searches on these embeddings using a vector store such as Milvus.

GPTCache comprises six core modules: LLM Adapter, Pre-processor (Context Manager), Embedding Generator, Cache Manager, Similarity Evaluator, and Post-processor.