Embedding

Monday, 22 May 2023 11:32 AM

https://blog.devgenius.io/so-you-want-to-build-an-ai-application-that-utilizes-llm-lets-talk-about-embedding-and-semantic-166acfc013a6

https://towardsdatascience.com/getting-started-with-langchain-a-beginners-guide-to-building-llm-powered-applications-95fc8898732c

Chunking in LLM:
https://www.pinecone.io/learn/chunking-strategies/?utm_content=244745025&utm_medium=social&utm_source=twitter&hss_channel=tw-1287624141001109504

Vector DB:
https://dutchengineer.substack.com/p/vector-databases?sd=pf
https://dutchengineer.substack.com/p/vector-databases?utm_source=twitter&sd=pf

Need :
https://blog.devgenius.io/so-you-want-to-build-an-ai-application-that-utilizes-llm-lets-talk-about-embedding-and-semantic-166acfc013a6

https://github.com/pgvector/pgvector

https://ann-benchmarks.com/index.html

https://blog.devgenius.io/so-you-want-to-build-an-ai-application-that-utilizes-llm-lets-talk-about-data-pre-processing-7fc7cf871d08

https://ann-benchmarks.com/index.html

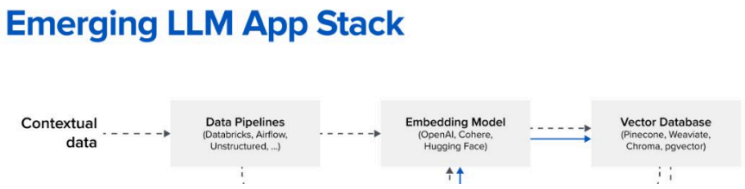
https://engineering.fb.com/2017/03/29/data-infrastructure/faiss-a-library-for-efficient-similaritv-search/

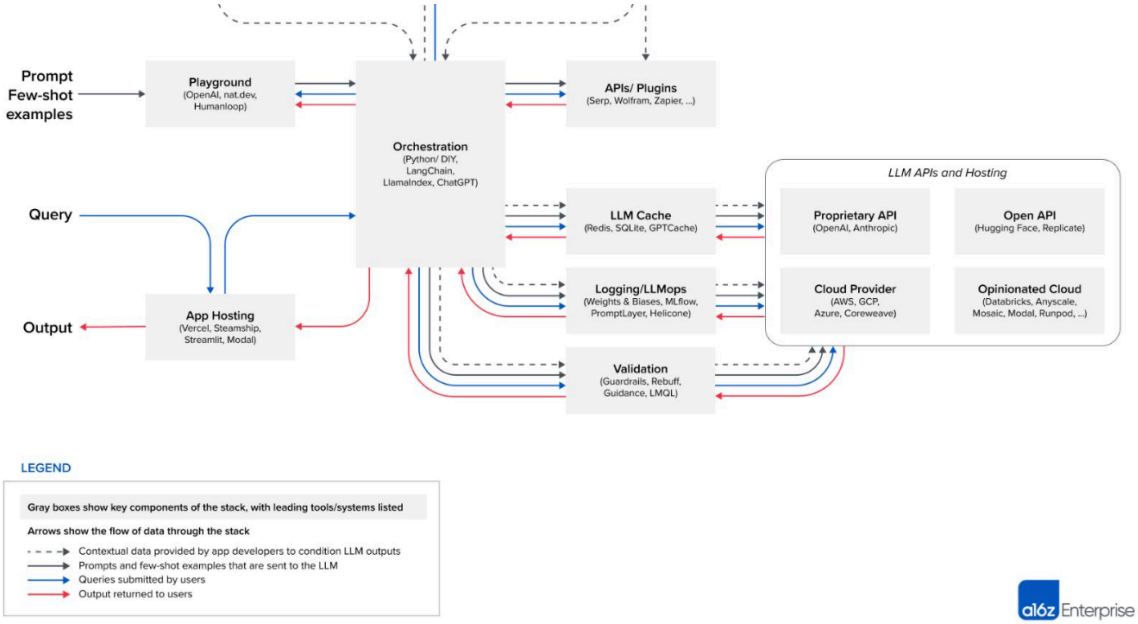
https://towardsdatascience.com/getting-started-with-langchain-a-beginners-guide-to-building-llm-powered-applications-95fc8898732c

https://towardsdatascience.com/10-exciting-project-ideas-using-large-language-models-llms-for-your-portfolio-970b7ab4cf9e

https://medium.com/the-generator/31-ai-prompts-better-than-rewrite-b3268dfe1fa9

> When you feed information into these APIs, they're billed based on tokens, both for input (the prompt) and output. The longer the prompt, the higher the cost. So, if you put explicit detail and examples into the prompt, you'll likely achieve better model performance, but it'll cost more.
>
> For a quick example, a simple task might require a prompt of about 300-1000 tokens. However, if you incorporate additional context like documents or internet-sourced information, the prompt can shoot up to a whopping 10k tokens!
>
> While prompt engineering is a cost-effective way to get a model up and running quickly, the real cost of Large Language Model Operations (LLMOps) lies in inference. If you use GPT-4 with 10k tokens in input and 200 tokens in output, it'll cost $0.624 per prediction.
>
> In contrast, GPT-3.5-turbo with 4k tokens for both input and output will cost $0.004 per prediction or $4 per 1k predictions. On a small scale, this might seem affordable, but let's put it in perspective. Imagine a scenario like DoorDash in 2021, making 10 billion predictions a day. At $0.004 per prediction, the cost soars to a mind-boggling $40 million per day!
>
> The dance of managing costs while striving for better performance in using LLM APIs is delicate but crucial. It's essential to understand the balance to optimize your models effectively without breaking the bank. 🏦⚖️





https://www.anyscale.com/blog/continuous-batching-llm-inference?trk=feed_main-feed-card_feed-article-content

Tokenizer:
https://sachindharashivkar.substack.com/p/how-to-train-large-language-models-6ad?r=6juth&utm_campaign=post&utm_medium=web