LLM Apps

LlamaIndex

Intro

- In its first version, LlamaIndex is a framework less generalist than LangChain. It aims to do fewer things but do them better. It specializes in generating possibilities for professional RAG applications.

- It is still a very young and small company. It still needs to find its strategic vision and business model. It has a great dependency on the evolution of chatGPT.

- Although the first version of LlamaIndex was not very user-friendly, they are now striving to make a second improved version in that regard.

Quickstart to LlamaIndex

- Load private document
- Create vector database
- Ask questions to the private document
- Save the vector database

Customization options

- parse into smaller chunks
- use a different vector store
- retrieve more context when I query
- use a different LLM
- use a different response mode
- stream the response back

Use cases

- QA
- Chatbot
- Agent
- Structured Data Extraction
- Multimodal

Optimizing

- Advanced Retrieval Strategies
- Evaluation
- Building performant RAG applications for production

Other

- LlamaPacks and Create-llama = LangChain templates
- Very recent, still in beta
- Very interesting: create-llama allows you to create a Vercel app!
- Very interesting: open-source end-to-end project (SEC Insights)
  - llamaindex + react/nextjs (vercel) + fastAPI + render + AWS
  - environment setup: localStack + docker
  - monitoring: sentry
  - load testing: loader.io
  - web: https://www.secinsights.ai/
  - code: https://github.com/run-llama/sec-insights