GENERATIVE AI RAMP-UP

Following courses should be covered to get and understanding of the basics that are necessary to build Gen AI applications.

Understanding LLMs

- Intro to LLMs
- Generative AI with LLMs

Prompt Engineering

- PE basics

Vector Databases

- Building apps with vector dbs

BUILDING RAG APPLICATIONS

Retrieval **A**ugmented **G**eneration (RAG) applications are the most popular type of applications in the LLM space. LLMs are typically trained on vast amounts of general information and don't have domain/task specific information. RAG applications augment LLM capabilities with domain specific data. One such use case is outlined in this blog post. Starters may use the same use case using the sample dataset from here.

Setup

Using a local BERT model, a vector database and a local LLM would be ideal to test multiple iterations of your RAG application without having to pay for embedding generation, storage or LLM inference. Below outlines how to setup both components on your local machine

Generating Embeddings & storing them locally

Ingesting documents & generating embeddings of your data is the first step in a RAG solution. Then you need a vector store.

For data/document ingestion use either the document loaders from Langchain or from LlamaIndex.

To generate embeddings, use GPT4All, preferably nomic model. Langchain also has abstractions for GPT4All. Then use Chroma as a local vector store.

Now you have the embeddings in the vector store, next is to run the LLM locally.

Running LLM locally

Install ollama and run llama3, or phi3 if the system is resource constrained. Use ChatOllama to interact with the model programmatically.

Chatbot

There are multiple choices for building chatbot interface. Choose one of below

Gradio

Streamlit

Nicegui

ADVANCED RAG

The final result in a RAG system is only as good as the documents retrieved corresponding to the query. New techniques emerge continuously to improve the accuracy of RAG systems. Enhance the Naïve RAG implemented above with below

1. Implement various techniques outlined in this series (Part 5 onwards).
    a. Series as a single tutorial
2. Couple techniques in this short course.
3. Implement RAG Agents.
4. Finally implement RAFT. More info.