LLM (Large Language Models) Complete Guide

1. What is an LLM?

LLM stands for Large Language Model. It is an AI model trained on massive amounts of text to understand and generate human-like language. GPT, Gemini, Claude, LLaMA are types of LLMs.

2. What is GPT?

GPT stands for Generative Pre-trained Transformer. It is OpenAI’s specific LLM family. All GPTs are LLMs, but not all LLMs are GPTs.

3. When to Use an LLM (GPT)?

- Text generation

- Summarization

- Coding assistance

- Translation & rewriting

- Chatbots & automation

- RAG applications

- Reasoning tasks

4. RAG (Retrieval Augmented Generation)

RAG allows GPT to use YOUR documents by:

1. Loading files

2. Chunking text

3. Creating embeddings

4. Storing vectors in FAISS/Chroma

5. Retrieving relevant chunks

6. Asking GPT using the retrieved context

5. Tools Used in RAG:

LangChain:

A powerful framework for building LLM apps, pipelines, retrieval, agents, and more.

LlamaIndex:

A simple framework focused mainly on RAG. Easier and cleaner for beginners.

FAISS:

A fast vector database created by Meta. Good for local, high-performance search.

ChromaDB:

A beginner-friendly vector store. Easy to set up and perfect for small/medium RAG apps.

6. Recommended Setup for Beginners:

- Framework: LlamaIndex or LangChain
- Vector DB: ChromaDB (easy) or FAISS (fast)
- Embeddings: text-embedding-3-small (OpenAI)
- LLM: GPT-4o-mini or GPT-4o

This PDF summarizes the core ideas needed to understand LLMs, GPT, and RAG systems.