Retrieval-Augmented Generation (RAG)

Module 1: Introduction to RAG

- What is RAG?
- Why use RAG over standard LLMs?
- Use cases of RAG (chatbots, enterprise search, legal, healthcare, customer support)
- Limitations of pure LLMs: hallucinations, lack of up-to-date info
- Core architecture: Retriever + Generator

Module 2: Deep Dive into Components

- Retriever
  - Sparse vs Dense Retrieval
  - Traditional Retrieval (TF-IDF, BM25)
  - Dense Vector Retrieval (using Sentence Transformers, BERT, etc.)
  - Vector databases overview: Pinecone, FAISS, Weaviate, Chroma
- Generator
  - Using Pretrained LLMs (OpenAI GPT, Anthropic Claude, HuggingFace models)
  - Prompt Engineering for retrieval output
  - Handling context window limits

Module 3: Embeddings and Vector Search

- What are embeddings?
- How to generate embeddings (OpenAI, Hugging Face, Cohere)
- Storing and indexing embeddings
- Similarity measures (Cosine, Euclidean, Dot product)
- Chunking and splitting strategies for documents

Module 4: Building RAG Pipeline

- Step-by-step architecture:
- Data ingestion
- Chunking
- Embedding generation
- Vector storage
- Retrieval
- Prompt assembly
- Generation
- Tools: LangChain, LlamaIndex, Haystack

Module 5: Enhancing RAG Performance

- Ranking retrieved documents
- Reranking with cross-encoders (e.g., BGE Reranker, Cohere Rerank)
- Summarization of retrieved content before generation

- Compression and filtering techniques
- Memory in RAG (short-term and long-term)

Module 6: Evaluation and Testing

- Accuracy vs Relevance vs Helpfulness
- Human-in-the-loop feedback
- BLEU, ROUGE, Precision@K
- Hallucination detection
- A/B Testing for retrieval/generation quality

Module 7: RAG in Real-world Applications

- Knowledge management bots
- Legal document assistants
- Customer support automation
- Academic question answering systems
- Personal AI knowledge base

Module 8: Advanced Topics

- Multi-hop Retrieval
- Hybrid retrieval (combining dense + sparse)
- Integrating structured + unstructured data
- RAG vs Fine-tuning: When to choose what?
- Streaming RAG with LangChain

Module 9: Deployment and Optimization

- Deploying RAG on cloud (AWS, Azure, GCP)
- Serverless GenAI workflows
- API Integration with front-end apps
- Cost optimization strategies (embedding batching, caching)
- Securing RAG systems (data privacy, access control)

==========

END

==========