GenAI using Ollama - Course Outline

MODULE 1

- Overview of Ollama
- Key Features and Functional Capabilities of Ollama
- Purpose and Need for Ollama in the AI Ecosystem
- Advantages and Competitive Edge of Ollama
- Prerequisites for Installing and Using Ollama
- How Ollama actually works [where models are stored in our device , how they are used, Ollama ecosystem etc]
- Step-by-Step Guide to Ollama Installation

MODULE 2

- Local Model Architecture and Storage (Manifest, Blobs, and Configuration Files)
- Different types of models in Ollama
- Command-Line Interface (CLI) Prompting and Execution
- All CLI commands.
- Managing and Running Multiple Models
- Introduction to Multimodal Ollama
- Experimentation with Parameters
- Performance Tuning using CLI

MODULE 3

- Utilizing the Ollama Library for Programmatic Interaction
- Understanding the `generate` and `chat` Commands
- Exploring Options and Configurable Parameters
- Streaming vs. Non-Streaming Response Handling
- Core Ollama Commands (`list`, `pull`, `show`) via the Ollama Library
- Introduction to the Modelfile Concept
- Syntax and Structural Components of a Modelfile
- Building and Running Custom Models in Ollama

MODULE 4

- Overview of the Ollama REST API
- Understanding Ollama's Backend Architecture and API Workflow
- How Ollama Processes Requests Internally through REST Endpoints
- REST API Documentation and Key Endpoints Overview
- Implementing API Calls for Text Generation and Chat Functions
- Using requests module to interact with Ollama.
- Introduction to Tool Calling and Function

MODULE 5

- Utilizing Ollama for Vector Search–Based Applications
- Fundamentals of Embeddings and Retrieval-Augmented Generation (RAG)
- Understanding Embeddings and Retrieval-Augmented Generation (RAG) with Ollama
- Generating Embeddings using Ollama’s Embedding Models and the `/api/embeddings` Endpoint
- Managing and Storing Embeddings in Vector Databases (FAISS, ChromaDB, Weaviate)
- Connecting Ollama with External Data Sources (Documents, PDFs, and CSVs)
- Integration of Ollama with LangChain for RAG and Workflow Automation
- Minor Project: Building a Local Retrieval-Augmented Generation (RAG) Application

MODULE 6

- Executing and Managing Multiple Models Concurrently in Ollama
- Running multiple models in Ollama
- Implementing Model Chaining (Using One Model’s Output as Another’s Input)
- Understanding GPU offloading
- Monitoring System Utilization and Model Runtime Efficiency
- Introduction to Ollama Cloud Models
- Hosted API services of Ollama
- Getting models from the hugging face library using Ollama.
- Ollama app.

MODULE 7

- Project Overview: Building an AI-Powered Brand Monitoring System using Ollama
- Designing and Creating a Database for Storing Reddit Brand Mentions Over Time
- Automating Data Collection from Reddit using API Integration
- Structuring and Managing Brand Mention Data for Efficient Analysis
- Integrating Ollama Models for Sentiment Analysis and Context Understanding
- Building an Interactive Dashboard using python for Visualizing Brand Trends and Insights
- Containerizing the Entire Application and Database using Docker
- Creating and Managing Docker Images for Seamless Deployment

MODULE 8

- LM Studio introduction and installation.
- Running model through LM studio app.
- Model customization and model tuning.
- Multimodal input.
- Developer tab of LM studio.
- Running models using LM studio with the help of python library.
- Difference between LM studio and Ollama.
- LM studio Integration with Hugging face library.