Advance NLP & Generative AI (By Ratnesh Kumar Singh)

NLP with Generative AI

A comprehensive understanding of Natural Language Processing (NLP) through Generative AI, focusing on key concepts like language modeling and text generation. You will gain skills in analyzing, generating, and manipulating textual data using advanced neural architectures such as transformers. By exploring state-of-the-art techniques and tools, you'll be empowered to develop innovative applications that harness the transformative potential of generative AI in language interactions.

Objectives

- Build a strong understanding of NLP, including key concepts, techniques, and challenges associated with generative AI.
- Implement and work with state-of-the-art models such as GPT, BERT, LLAMA, and Mistral for effective text generation and manipulation.
- Utilize RAG techniques to improve contextual accuracy in text generation by integrating external knowledge sources.
- Master prompt engineering and explore the creation of multimodal applications that combine text, images, and audio.
- Investigate the integration of text, image, and audio data for multimodal applications.
- Learn to evaluate and fine-tune language models and deploy NLP applications using Hugging Face and LangChain.

Module 1 – Brief Overview of Classical NLP & Introduction of GenAI

This section gives a brief overview of classical NLP and introduces GenAI. It explains differences between ANN, CNN, and RNN (LSTM, GRU), encoder–decoder architectures, attention mechanisms, and transformer architecture. It also discusses Generative AI, its applications, and ethical and social implications.

Topics

- Neural Network Architectures – ANN, CNN, RNN
- Generative AI Fundamentals – definition and significance
- Key Components of GenAI and Applications
- Ethical Considerations in AI
- Encoder–Decoder Structures

Module 2 – Understanding and Implementing the Transformer Architecture

This section covers self-attention, multi-head attention, masked attention, cross-attention, and transformer encoder–decoder architecture, with hands-on PyTorch implementation.

Topics

- Self-Attention Mechanism Overview
- Geometric Insights into Self-Attention
- Transformer Architecture Components
- Implementing Transformers in PyTorch

Module 3 – Fundamentals of Large Language Models (LLMs)

This section introduces foundation models, BERT, GPT, and LLAMA architectures, and their training processes.

Topics

- Foundation Models Overview
- BERT Model Architecture
- GPT Model Architecture
- LLAMA Model Architecture
- Training Foundation Models

Module 4 – Word and Sentence Embedding

This section explains embeddings, methods, models, and evaluation techniques.

Topics

- Introduction to Embeddings
- Common Embedding Methods
- Word vs. Sentence Embedding
- Embedding Models (Word2Vec, BERT)
- Evaluation of Embeddings

Module 5 – Mastering Hugging Face for NLP and Beyond

Focuses on Hugging Face ecosystem, fine-tuning, deployment, and multimedia models.

Topics

- Introduction to Hugging Face
- Hugging Face API and Inference
- Fine-Tuning Large Language Models

- Model Deployment and Sharing
- NLP with Transformers
- Multimedia Models

Module 6 – Overview of Major AI APIs

Covers OpenAI, Google Gemini, and Anthropic Claude APIs.

Topics

- Introduction to OpenAI APIs
- Setting Up OpenAI Account
- OpenAI Pricing and Models (GPT-3.5, GPT-4)
- Introduction to Google Gemini
- Anthropic Claude Overview

Module 7 – Fine-Tuning for Specialized AI Applications

Explains transfer learning, fine-tuning techniques, RLHF, quantization, and cost analysis.

Topics

- Transfer Learning vs Fine-Tuning
- End-to-End Fine-Tuning Roadmap
- Types of Fine-Tuning Techniques
- Advanced Strategies (DPO, PPO, RLHF)
- Model Quantization (4-bit, 8-bit, 1-bit)
- Fine-Tuning Open-Source Models

Module 8 – Guide to Vector Databases for AI Applications

Introduces vector databases, indexing, similarity search, and tools.

Topics

- Introduction to Vector Databases
- Comparison with SQL & NoSQL
- Data Storage and Architecture
- Types of Vector Databases
- Indexing Methods for Vector Search

- Similarity Search Algorithms (Annoy)
- ChromaDB, FAISS, Qdrant, Pinecone, LanceDB

Module 9 – Retrieval Augmented Generation (RAG)

Covers RAG concepts, pipelines, hybrid search, and multimodal RAG.

Topics

- Introduction to RAG
- End-to-End RAG Pipeline
- Implementing RAG with LangChain
- Hybrid Search and Reranking
- Multimodal RAG Applications
- RAG with Knowledge Graphs

Module 10 – Comprehensive Guide to LangChain

Explains LangChain components, tools, agents, deployment, and monitoring.

Topics

- Introduction to LangChain
- Data & API Connectors
- LangChain Tools & Toolkits
- Prompt Templating and Chains
- Synthetic Data & Memory Management
- AI Agents, LangServe & LangSmith

Module 11 – Overview of LlamaIndex

Comparison with LangChain and RAG implementation.

Topics

- LlamaIndex vs LangChain
- Data Loader & Web Scraper
- RAG with LlamaIndex
- Multimodal Applications
- Agents in LlamaIndex
- Llama Hub Overview

Module 12 – AI Agents

Covers agent frameworks, LangGraph, and multi-agent systems.

Topics

- Introduction to AI Agents
- LangChain Agent Framework
- ReAct, Structured Output, Self-Ask Agents
- LangGraph Introduction
- Agentic RAG with LangGraph
- Multi-Agent Systems

Module 13 – LLM-Based App on Local Infrastructure

Focuses on running LLMs locally.

Topics

- Introduction to Ollama
- Setting Up Llama CPP
- Using LM Studio
- Hugging Face Model Downloader

Module 14 – LLMOps: Optimizing LLM-Powered Applications

Explains MLOps challenges, deployment, and cloud platforms.

Topics

- Challenges in LLM App Development
- Open-Source LLM Deployment
- MLOps Tools (ZenML, MLflow, Prefect)
- Web Frameworks (Flask, FastAPI)
- Cloud Platforms for LLM Deployment

Module 15 – End-to-End Project

Complete GenAI application lifecycle.

Projects

- Trading Bot (Multi-AI Agent System)

    - **Market data collection**
    - **Trend analysis**
    - **Risk monitoring**
    - **AWS deployment**
- **Customer Support Chatbot**
    - **Conversational flow with RAG**
    - **Bot training**
    - **UI design**
    - **CI/CD deployment**
    - **Monitoring & optimization**

##

Tech Stack used in the Advance NLP & Generative AI (By Ratnesh Kumar Singh)

◇ Tech Stack

- **Languages:** Python
- **Deep Learning:** PyTorch, Transformers
- **LLMs:** GPT-3.5/4, BERT, LLaMA, Mistral
- **Fine-Tuning:** SFT, Instruction Tuning, RLHF (PPO, DPO), Quantization
- **Frameworks:** Hugging Face, LangChain, LlamaIndex, LangGraph
- **Vector Databases:** FAISS, ChromaDB, Qdrant, Pinecone, LanceDB
- **RAG:** Hybrid Search, Re-ranking, Knowledge-Graph RAG, Multimodal RAG
- **AI Agents:** ReAct, Structured Output Agents, Multi-Agent Systems
- **Local LLMs:** Ollama, LLaMA.cpp, LM Studio
- **MLOps & Deployment:** FastAPI, Flask, MLflow, ZenML, Prefect
- **Cloud:** AWS, Azure , GCP
- **Monitoring:** LangSmith, Evaluation & Cost Optimization

🧠 AI / ML & NLP

- **Python**
- **PyTorch**
- **Transformers (Attention, Encoder–Decoder)**
- **Classical NLP Techniques**
- **Embeddings** (Word & Sentence)

🤖 Large Language Models (LLMs)

- **GPT-3.5 / GPT-4**
- **BERT**
- **LLAMA**
- **Mistral**
- **Open-Source Foundation Models**

🔧 Fine-Tuning & Optimization

- **Supervised Fine-Tuning (SFT)**
- **Instruction Fine-Tuning**
- **RLHF (PPO, DPO)**
- **Model Quantization** (4-bit, 8-bit, 1-bit)
- **PEFT / LoRA (conceptual)**

📚 Frameworks & Libraries

- **Hugging Face** (Transformers, Hub, Spaces)
- **LangChain**
- **LlamaIndex**
- **LangGraph**
- **CrewAI**
- **AutoGen**

🔍 Vector Databases & Search

- **FAISS**
- **ChromaDB**

- **Qdrant**
- **Pinecone**
- **LanceDB**
- **Annoy (Approximate Nearest Neighbor)**
- **Hybrid / Semantic / Multilingual Search**

🔄 Retrieval-Augmented Generation (RAG)

- **RAG Pipelines**
- **Hybrid Search & Re-ranking**
- **Knowledge Graph-based RAG**
- **Multimodal RAG** (Text, Documents, Video)

🌐 AI APIs & Platforms

- **OpenAI API**
- **Google Gemini**
- **Anthropic Claude**

🧩 AI Agents

- **ReAct Agent**
- **Structured Output Agent**
- **Self-Ask with Search**
- **Multi-Agent Systems**
- **Agentic RAG**

🖥️ Local LLM & Inference Tools

- **Ollama**
- **LLaMA.cpp**
- **LM Studio**
- **Hugging Face Model Downloader**

🚀 Deployment, MLOps & Backend

- **FastAPI**
- **Flask**
- **Docker (implied for deployment)**
- **MLflow**
- **ZenML**
- **Prefect**
- **CI/CD Pipelines**

☁️ Cloud & Infrastructure

- **AWS** (Deployment for projects)
- **Cloud platforms for LLM hosting & scaling**

🧪 Monitoring & Evaluation

- **LangSmith**
- **Model Evaluation Techniques**
- **Cost & Performance Optimization**
