KRISHNAIK ACADEMY

Advance NLP Generative AI Bootcamp

NLP with Generative AI

This course offers 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.

Learning 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 to leverage the capabilities of generative AI.
- Investigate the integration of text, image, and audio data to create multimodal applications that leverage generative AI across different media types.
- Learn how to evaluate and fine-tune language models for specific tasks and gain practical experience deploying NLP applications using frameworks like Hugging Face and LangChain.

Module 1

Breif Overview of Classical NLP & Introduction of GenAI

This section gives a brief overview of classical NLP and introduces GenAI. It explains the differences between various types of neural networks like ANN, CNN, and RNN, including LSTM and GRU. The section also covers encoder-decoder architectures that use attention mechanisms and introduces the transformer architecture, which has changed the way we handle NLP tasks. Generative AI is defined, highlighting its importance and core components, along with its many applications and use cases. Finally, it discusses the ethical and social implications of Generative AI, emphasizing the need for responsible use in technology.

Topics	
Neural Network Architectures Overview	Explore differences in ANN, CNN, RNN.
Generative AI Fundamentals	Understand generative AI's definition and significance.
Key Components of GenAI and Applications	Identify core elements and practical uses.
Ethical Considerations in AI	Discuss social implications of generative AI.
Ethical Considerations in AI	Learn about encoder-decoder structures.

Module 2

Understanding and Implementing the Transformer Architecture

This section covers the understanding and implementation of the transformer architecture. It begins with self-attention and multi-headed self-attention, explaining their roles in processing text. The geometric intuition of self-attention is discussed, along with masked self-attention for language modeling and cross-attention for comparing sequences. The architecture of transformer encoders and decoders is detailed, highlighting their functions. Finally, it concludes with a hands-on implementation of the transformer architecture from scratch in PyTorch, enabling practical application of the concepts learned.

Topics	
Self-Attention Mechanism Overview	Understand self-attention and multi-head concepts.
Geometric Insights into Self-Attention	Explore geometric intuition behind self-attention.
Transformer Architecture Components	Learn about encoder and decoder structures.
Implementing Transformers in PyTorch	Build transformer architecture from scratch.

Module 3

Fundamentals of Large Language Models (LLMs)

This section introduces the fundamentals of Large Language Models (LLMs), focusing on foundation models and their importance in NLP. It examines the architectures of key models like BERT, GPT, and LLAMA, emphasizing their distinct features. Additionally, the section covers the training processes for these models, highlighting techniques used to enhance their performance across various language tasks, providing learners with a comprehensive understanding of the concepts and architectures behind modern LLMs.

Topics	
Foundation Models Overview	Understand the concept of foundation models.
BERT Model Architecture	Explore architecture and functionalities of BERT.
GPT Model Architecture	Learn about the structure of GPT models.
LLAMA Model Architecture	Examine the design and features of LLAMA.
Training Foundation Models	Understand techniques for training these models.

Module 4

Word and Sentence Embedding

This section covers the concept of embeddings, starting with an introduction to their significance in NLP. It distinguishes between word embeddings and sentence embeddings, exploring common methods used to create them. The section reviews various embedding models, including Word2Vec, BERT, and Sentence Transformers, highlighting their unique capabilities. Additionally, it addresses the evaluation of embeddings, providing insights into assessing their effectiveness for different language processing tasks, and establishing their relevance in NLP applications.

Topics	
Introduction to Embeddings	Learn what embeddings are in NLP.
Common Embedding Methods	Explore various techniques for embeddings.
Word vs. Sentence Embedding	Differentiate between word and sentence embeddings.
Embedding Models Overview	Understand models like Word2Vec and BERT.
Evaluation of Embeddings	Learn methods to assess embedding quality.

Module 5

Mastering Hugging Face for NLP and Beyond

This section introduces Hugging Face, covering its API and the process of model inference. It emphasizes fine-tuning large language models and deploying them using Hugging Face Hub and Spaces. The curriculum includes using the Transformers library for NLP, leveraging pipelines for streamlined inference, and exploring applications in image, video, and audio models, providing a comprehensive understanding of Hugging Face's capabilities.

Topics	
Introduction to Hugging Face	Discover the Hugging Face ecosystem and tools.
Hugging Face API and Inference	Learn to utilize the API for model inference.
Fine-Tuning Large Language Models	Customize and optimize LLMs with Transformers.
Model Deployment and Sharing	Push models to the Hub and deploy in Spaces.
NLP and Hugging Face Transformers	Leverage Transformers for effective natural language processing.
Multimedia Models with Hugging Face	Work with image, video, and audio models.

Module 6

Overview of Major AI APIs: Setup and Key Features

This section provides an overview of major AI APIs, starting with an introduction to OpenAI APIs, including account setup and pricing details. It covers various OpenAI models like GPT-3.5 and GPT-4. The section also introduces Google Gemini, including its setup, pricing, and model variations. Lastly, it discusses Anthropic Claude, detailing its key setup and available models, offering a comprehensive understanding of these influential AI tools.

Topics	
Introduction to OpenAI APIs	Overview of OpenAI's API capabilities.
Setting Up OpenAI Account	Steps to create and configure an account.
OpenAI Pricing and Models	Understand pricing and review GPT-3.5, GPT-4.
Introduction to Google Gemini	Overview of Google Gemini AI model features.
Anthropic Claude Overview	Introduction to Claude and its capabilities.

Module 7

Fine-Tuning for Specialized AI Applications

This section covers the essentials of fine-tuning AI models for specialized applications, distinguishing between transfer learning and fine-tuning. It outlines various techniques, including supervised fine-tuning, instruction fine-tuning, and advanced methods like DPO and PPO. Key considerations include model quantization strategies and cost analysis. Practical applications are explored through fine-tuning popular models like BERT and open-source options such as LLAMA and Mistral, including specifics on fine-tuning OpenAI models like GPT-3.5 and GPT-4.

Topics	
Transfer Learning vs. Fine-Tuning	Distinguishing transfer learning from fine-tuning.
Complete End-to-End Finetuning Roadmap	Comprehensive guide for effective fine-tuning.
Types of Fine-Tuning Techniques	Explore supervised, instruction, and RLHF techniques.
Advanced Fine-Tuning Strategies	Discuss DPO, PPO, and RLHF comparisons.
Model Quantization Techniques	Overview of 4-bit, 8-bit, and 1-bit.
Fine-Tuning Open-Source Models	Techniques for fine-tuning various models.

Module 8

Guide to Vector Databases for AI Applications

This guide introduces vector databases, emphasizing their differences from SQL and NoSQL databases. It outlines their capabilities, architecture, and types, including in-memory, local disk, and cloud-based options. The section discusses efficient indexing methods for vector searches and explores similarity search techniques, including Annoy for approximate nearest neighbors. It highlights multilingual and semantic search functionalities and provides insights on sparse, dense, and hybrid search types. Additionally, it covers setup and query operations using popular vector databases like Chroma DB, Faiss, and Pinecone, as well as integrating with NoSQL databases like MongoDB and Cassandra.

Topics	
Introduction to Vector Databases	Overview of databases for vector data.
Comparison with SQL and NoSQL	Contrast vector databases with traditional options.
Data Storage and Architecture	Structure and design of vector databases.
Types of Vector Databases	Explore in-memory, local, and cloud options.
Indexing Methods for Vector Search	Efficient techniques for fast vector retrieval.
Search Algorithms for Vector Similarity	Annoy (Approximate Nearest Neighbor Oh Yeah)
Setup and Query Operations	Practical use of popular vector databases.(Chroma DB, Faiss, Quadrant, Pinecone, LanceDB)

Module 9

Retrieval Augument Generation(RAG)

This overview covers Retrieval-Augmented Generation (RAG), starting with its foundational concepts and the end-to-end pipeline. It discusses implementing RAG using LangChain, vector databases, and large language models (LLMs). The guide includes techniques for hybrid search and reranking, along with various retrieval methods and memory integration within RAG systems. It explores multimodal applications, such as video processing and document parsing, and highlights the integration of knowledge graphs to enhance retrieval capabilities.

Topics	
Introduction to RAG	Overview of retrieval-augmented generation concepts.
End-to-End RAG Pipeline	Steps involved in RAG implementation.
Implementing RAG with Tools	Using LangChain and vector databases effectively.
Hybrid Search and Reranking	Enhancing search quality through multiple methods.
Multimodal RAG Applications	Integrating video and document processing techniques.
RAG with Knowledge Graphs	Utilizing knowledge graphs for enriched retrieval.

Module 10

Comprehensive Guide to LangChain

This comprehensive guide to LangChain introduces its core components and functionalities. It covers data connectors, API connectors, and chat models, along with various tools and toolkits available within LangChain. The guide delves into prompt templating, the creation of chains, and the use of LangChain LCEL and runnables. Key features such as synthetic data generation, memory management, and AI agents are explored, alongside LangSmith for model monitoring and LangServe for seamless model deployment in applications.

Topics	
Introduction to LangChain	Overview of LangChain's capabilities and features.
Data and API Connectors	Integrating data sources and API connections seamlessly.
LangChain Tools and Toolkit	Available tools for enhancing workflow efficiency.
Prompt Templating and Chains	Creating and managing prompt templates and chains.
Synthetic Data Generation and Memory Management	Techniques for generating synthetic data and managing memory.
AI Agents and Model Monitoring & Deployment	Deploying models using LangServe and monitoring with LangSmith.

Module 11

Overview of LlamaIndex

This overview of LlamaIndex compares it with LangChain, highlighting its unique features. It discusses the LlamaIndex Data Loader and Web Scraper, showcasing their roles in data handling. The guide covers Retrieval-Augmented Generation (RAG) capabilities with LlamaIndex and explores multimodal applications. Additionally, it introduces the concept of agents within LlamaIndex and provides insights into Llama Hub, emphasizing its functionalities and how it enhances the LlamaIndex ecosystem for developers and researchers.

Topics	
LlamaIndex vs. LangChain	Comparison of LlamaIndex and LangChain functionalities.
Data Loader and Web Scraper	Tools for loading data and web scraping efficiently.
Retrieval-Augmented Generation (RAG)	Implementing RAG techniques with LlamaIndex.
Multimodal Applications	Exploring various multimodal capabilities of LlamaIndex.
Agents in LlamaIndex	Utilizing agents for enhanced task automation.
Llama Hub Overview	Features and functionalities of the Llama Hub.

Module 12

AI Agents

This guide introduces AI agents and their integration within the LangChain framework, detailing types such as the ReAct Agent, Structured Output Agent, and Self-Ask with Search Agent. It provides an overview of LangGraph and its components, focusing on the development of multi-agent systems. Additionally, the guide explores Retrieval-Augmented Generation (RAG) with LangGraph, including Corrective RAG, Agentic RAG, and Self-RAG. Finally, it highlights CrewAI and Autogen for built-in agent capabilities, enhancing the functionality of AI applications.

Topics	
Introduction to AI Agents	Overview of what AI agents are and their roles.
LangChain Agent Framework	Exploring how AI agents function within LangChain.
ReAct, Structured Output,Self-Ask with Search Agent	Explanation of ReAct, structured output agents, and self-ask techniques for enhanced search capabilities.
LangGraph Introduction	Overview of LangGraph and its significance in AI systems.
RAG with LangGraph	Implementing Agentic Retrieval-Augmented Generation (RAG) techniques using LangGraph.
Multi-Agent Systems	Implementing multi-agent systems using LangGraph.

Module 13

LLM-Based App on Local Infrastructure

This guide focuses on building LLM-based applications on local infrastructure, featuring Ollama for model management, and providing instructions for setting up LLama CPP and LM Studio. It also covers the Hugging Face Model Downloader to facilitate easy access to various pre-trained models for enhanced development and experimentation.

Topics	
Introduction to Ollama	Overview of Ollama and its features for LLM management.
Setting Up Llama CPP	Guidelines for configuring Llama CPP for local model execution.
Utilizing LM Studio	Exploring LM Studio for building and testing LLM applications.
Hugging Face Model Downloader	Instructions for downloading and integrating models from Hugging Face.

Module 14

LLMOps: Optimizing LLM-Powered Applications

This guide explores the challenges of developing LLM-powered applications and emphasizes the importance of effective MLOps practices. It covers open-source deployment strategies and highlights essential tools for managing LLMs, including various web frameworks tailored for LLM applications. Additionally, it reviews cloud platforms that facilitate the deployment and scaling of large language models, ensuring optimal performance and accessibility in real-world scenarios.

Topics	
Challenges in LLM Application Development	Identifying obstacles in building LLM-powered applications.
Open Source LLM Deployment	Exploring the advantages and methods for deploying open-source LLMs.
MLOps Tools for Deployment	Overview of essential MLOps tools like zenml, mlflow, prefect etc to streamline LLM deployment processes.
Web Frameworks for LLM Applications	Discussing various web frameworks like flask, fastapi suitable for LLM integration.
Cloud Platforms for LLM Deployment	Comparing cloud platforms that facilitate efficient LLM deployment.

Module 15

End to End Project

The End-to-End Project in your curriculum focuses on the complete development and deployment lifecycle of Generative AI (GenAI) applications powered by large language models (LLMs). It covers essential aspects such as data collection, preprocessing, RAG and fine-tuning LLMs to optimize performance for specific tasks. Additionally, the project integrates MLOps principles for efficient deployment, monitoring, and maintenance of LLM-powered applications, ensuring scalability and adaptability to meet evolving business requirements in the AI landscape.

Topics	
Developing a Trading Bot Using a MultiAI-Agent System	• Define roles for specialized agents. • Gather real-time market data effectively. • Analyze trends and generate insights. • Monitor risks and protect investments. • Test strategies and refine performance. • Deploy it over the AWS
Develop a Custmor Support Chatbot for any Domain	• Design Conversational Flow with RAG • Build and Train the Bot • User Interface Design • Testing and Iteration • Deployment with CI-CD • Monitor and Optimize