Generative AI Engineering Program

Master LLMs, Fine-Tuning, RAG, Open-Source Models & Production AI Systems
(Duration – 140 Hrs)

Program Objective

To train students to:

- Build **AI applications from scratch**
- Work with **both APIs and open-source LLMs**
- Customize models using **fine-tuning & embeddings**
- Design **production-ready AI systems**

Course Structure

Module 1: Python for AI Development (Fast-Track)

This module ensures everyone can comfortably build AI systems in Python.

Topics

- Python refresher (syntax, functions, classes)
- Virtual environments (venv, poetry, pip)
- Working with APIs and JSON
- Async programming (async/await)
- Python for data handling (pandas, numpy)
- REST APIs using FastAPI
- Introduction to Uvicorn (ASGI server for FastAPI)

Hands-on

- Build a simple API in FastAPI
- Call an LLM API from Python
- Async request handling

Module 2: Foundations of LLMs (How AI Really Works)

Topics:

- AI vs ML vs Deep Learning vs Generative AI
- Neural networks basics
- Evolution of AI → LLMs
- Tokens, embeddings, context
- Transformer architecture (intuitive)
- How LLMs generate responses
- Limitations (hallucination, bias, context limits)

Module 3: Advanced Prompt Engineering

Topics:

- Role-based prompting
- Few-shot & chain-of-thought prompting
- Structured output (JSON/schema control)
- Prompt chaining
- Prompt debugging & optimization

Hands-on:

- Build structured AI outputs (resume, reports, APIs)

Module 4: LLM APIs & Application Layer

Topics:

- API-based LLMs (OpenAI & alternatives)
- Open Source LLMs (Hugging Face)
- Model selection strategies
- GPU vs CPU considerations
- Parameter tuning (temperature, tokens, etc.)
- Streaming responses
- Tokenizers & pipelines

Hands-on:

- AI chatbot
- Text summarizer

- Run a local LLM chatbot

Module 5: Fine-Tuning & Model Customization

Topics:

- What is fine-tuning
- Instruction tuning
- LoRA / PEFT techniques
- Dataset preparation
- Evaluation of fine-tuned models
- Fine-tuning vs RAG decision

Project:

- Domain-specific chatbot (custom-trained)

Module 6: Embeddings & Vector Databases (Deep Dive)

Goal: Build scalable AI memory

Topics:

- Embedding models
- Semantic similarity
- Vector search concepts
- FAISS / vector DB tools
- Indexing strategies

Module 7: Advanced RAG Systems (Production-Level)

Topics:

- RAG architecture (deep dive)
- Chunking strategies (naive vs smart chunking)
- Context window optimization
- Re-ranking techniques
- Hybrid search (keyword + semantic)

Project:

- Enterprise-level document Q&A system

Module 8: Evaluation & Optimization of AI Systems

Advanced industry skill

Topics:

- Evaluating LLM outputs
- Reducing hallucinations
- Latency optimization
- Cost optimization
- Prompt vs fine-tune vs RAG decisions

Module 9: AI System Design & Architecture

Engineer-level thinking

Topics:

- Designing AI applications end-to-end
- Architecture patterns for GenAI apps
- Choosing:
  - API vs local model
  - RAG vs fine-tuning
- Case studies of real AI systems

Module 10: Deployment & MLOps for GenAI

Goal: Production readiness

Topics:

- Deploying AI applications
- Model serving basics
- API deployment
- Monitoring & logging
- Versioning models

Module 11: AI Safety, Security & Governance

Topics:

- Prompt injection attacks
- Data leakage risks
- Bias & fairness
- Responsible AI practices

Capstone Project

End-to-End AI Application (with deployment)

Final Outcome

Students will be able to:

- Build **LLM-powered applications from scratch**
- Customize models using **fine-tuning & embeddings**
- Design **production-grade AI systems**
- Work as:
  - Generative AI Engineer
  - LLM Engineer
  - AI Backend Developer

"Become a Generative AI Engineer – Go Beyond APIs to Build, Customize & Deploy Real AI Systems"
