Building with Open-Source Generative AI

Course Duration: Days
Exam Reference:

Course Overview

This course provides the essential, hands-on skills to write and deploy practical AI applications. You will design, develop, and optimize Transformer models, ensuring data security is built into your work. The training covers core AI transformer architectures, advanced Python programming, GPU hardware requirements, and training techniques like fine-tuning and quantization. By working with open-source LLM frameworks, you will gain access to a GPU-accelerated server and earn an AI certification from Alta3 Research.

Prerequisites

- Python - PCEP Certification or Equivalent
- Experience and Familiarity with Linux.

Course Objectives

Upon completion, you will be able to:

- Train and optimize Transformer models with PyTorch.
- Master advanced prompt engineering techniques.
- Understand AI architecture, especially Transformers.
- Write and deploy a real-world AI web application.
- Describe tokenization and word embeddings.
- Install and use open-source frameworks like LLaMa-2.
- Apply strategies to maximize model performance.
- Explore model quantization and fine-tuning.
- Compare CPU vs. GPU hardware acceleration.
- Understand chat vs. instruct interaction modes.

APPLIED TECHNOLOGY ACADEMY

Contact Us

800.674.3550

2151 W. Hillsboro Blvd., Ste 210
Deerfield Beach, FL 33442

Connect With Us

Building with Open-Source Generative AI

Course Outline

Module 1: Learning Your Environment & Deep Learning Intro

- Learning Your Environment
- Using Vim, Tmux, and VScode Integration
- Revision Control with GitHub
- Deep Learning Intro
- What is Intelligence? and Generative AI Unveiled
- The Transformer Model Architecture
- Feed Forward Neural Networks
- Tokenization and Word Embeddings
- Positional Encoding

Module 2: Building and Training a Transformer Model

- Build a Transformer Model from Scratch
- Introduction to PyTorch
- Construct and Orchestrate Tensors from a Dataset
- Initialize PyTorch Generator Function
- Train the Transformer Model
- Apply Positional Encoding and Self-Attention
- Attach the Feed Forward Neural Network and Build the Decoder Block
- Transformer Model as Code

Module 3: Prompt Engineering and Deployment Hardware

- Prompt Engineering
- Introduction to Prompt Engineering
- Developing Basic, Intermediate, and Advanced Prompts (Chaining, Set Role)
- Getting Started with Gemini (Hands-on exploration)
- Hardware Requirements
- GPUs role in AI performance (CPU vs GPU)
- Current GPUs and cost vs value

- Tensorcore vs older GPU architectures

Module 4: Open-Source LLMs and Advanced Deployment

- Pre-trained LLM & Deployment
- A History of Neural Network Architectures
- Introduction to the LLaMa.cpp Interface
- Preparing A100 for Server Operations
- Operate LLaMa2 Models with LLaMa.cpp
- Selecting Quantization Level for performance and perplexity
- LLaMa API Server & Applications
- Deploy Llama API Server
- Develop LLaMa Client Application
- Write a Real-World AI Application using the Llama API
- Constraining Output with Grammars

Module 5: Optimization and Fine Tuning

- Fine Tuning
- Using PyTorch to fine tune models
- Advanced Prompt Engineering Techniques
- Testing and Pushing Limits
- Maximizing Model Limits
- Curriculum Path: GenerativeAI