Training and Deploying Open-Source LLM's on Bare Metal

Price: 2595

Duration: 3 days

Delivery Methods: Virtual

Overview

About this Course

Learn how to write practical AI applications via hands-on labs. You will design and develop Transformer models, ensuring data security in your work. The course covers AI transformer architectures, Python programming, hardware requirements, training techniques, and AI tasks like classification and regression. It includes hands-on exercises with open-source LLM frameworks, advanced topics like fine-tuning and quantization, and offers AI certification from Alta3 Research. Ideal for Python Developers, DevSecOps Engineers, and Managers or Directors, the course requires basic Python skills and provides access to a GPU-accelerated server for practical experience.

Audience Profile

- Project Managers
- Architects
- Developers
- Data Acquisition Specialists

At Course Completion

- Train and optimize Transformer models with PyTorch.
- Master advanced prompt engineering.
- Understand AI architecture, especially Transformers.
- Write a real-world AI web application.

- Describe tokenization and word embeddings.

- Install and use 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.

Outline

Deep Learning Intro

- Lecture: What is Intelligence?

- Lecture: Generative AI Unveiled

- Lecture: The Transformer Model

- Lecture: Feed Forward Neural Networks

- Lecture + Lab: Tokenization

- Lecture + Lab: Word embeddings

- Lecture + Lab: Positional Encoding

Build a Transformer Model from Scratch

- Lecture: PyTorch

- Lecture + Lab: Construct a Tensor from a Dataset

- Lecture + Lab: Orchestrate Tensors in Blocks and Batches

- Lecture + Lab: Initialize PyTorch Generator Function

- Lecture + Lab: Train the Transformer Model

- Lecture + Lab: Apply Positional Encoding and Self-Attention

- Lecture + Lab: Attach the Feed Forward Neural Network

- Lecture + Lab: Build the Decoder Block

- Lecture + Lab: Transformer Model as Code

Prompt Engineering

- Lecture: Introduction to Prompt Engineering

- Lecture + Lab: Getting Started with Gemini

- Lecture + Lab: Developing Basic Prompts

- Lecture + Lab: Intermediate Prompts: Define Task/Inputs/Outputs/Constraints/Style

- Lecture + Lab: Advanced Prompts: Chaining, Set Role, Feedback, Examples

Hardware requirements

- Lecture: GPUs role in AI performance (CPU vs GPU)

- Lecture: Current GPUs and cost vs value

- Lecture: Tensorcore vs older GPU architectures

Pre-trained LLM

- Lecture: A History of Neural Network Architectures
- Lecture: Introduction to the LLaMa.cpp Interface
- Lecture: Preparing A100 for Server Operations
- Lecture + Lab: Operate LLaMa2 Models with LLaMa.cpp
- Lecture + Lab: Selecting Quantization Level to Meet Performance and Perplexity Requirements
- Lecture: Running the llama.cpp Package
- Lecture + Lab: Llama interactive mode
- Lecture + Lab: Persistent Context with Llama
- Lecture + Lab: Constraining Output with Grammars
- Lecture + Lab: Deploy Llama API Server
- Lecture + Lab: Develop LLaMa Client Application
- Lecture + Lab: Write a Real-World AI Application using the Llama API

Fine Tuning

- Lecture + Lab: Using PyTorch to fine tune models
- Lecture + Lab: Advanced Prompt Engineering Techniques

Testing and Pushing Limits

- Lecture + Lab: Maximizing Model Limits

Prerequisites

- Python or Equivalent Experience
- Familiarity with Linux

Course Schedule

Date	Time	Price	Options

Why Professional Choose TOPTALENT?

Dedicated Texas-Based Support

Get assistance every step of the way from our **Texas-based team**, ensuring your training experience is hassle-free and aligned with your goals.

3000+ Curated Professional Courses

Access an extensive portfolio of over 3000 courses across IT, Business Application and Leadership – Designed to meet evolving Industry demands

95% Client Approval Rating

Trusted by professionals nationwide our 95% approval rating reflects consistent quality, measurable impact and exceptional service.

Certified Industry Instructor

Learn from professionaly certified experts with real world experience and a proven commitment to learner success.

For questions

call:

[(469) 721-6100](tel:4697216100)

Email:

info@toptalentlearning.com

[Find More Training](#)

FAQ

What if I have to reschedule my class due to conflict?

Ten (10) business days' notice is required to reschedule a class with no additional fees. Notify TOPTALENT LEARNING as soon as possible at 469-721-6100 or by written notification to info@toptalentlearning.com to avoid rescheduling penalties.

How do I enroll for this class?

Please contact our team at 469-721-6100; we will gladly guide you through the online purchasing process.

What happens once I purchase a class?

You will receive a receipt and an enrollment confirmation sent to the email you submitted at purchase. Your enrollment email will have instructions on how to access the class. Any additional questions our team is here to support you. Please call us at 469-721-6100.

What is your late policy?

If a student is 15 minutes late, they risk losing their seat to a standby student. If a student is 30 minutes late or more, they will need to reschedule. A no-show fee will apply. Retakes are enrolled on a stand-by basis. The student must supply previously issued courseware. Additional fees may apply.

What happens when I finish my class?

You will receive a 'Certificate of Completion' once you complete the class. If you purchased an exam voucher for the class, a team member from TOPTALENT LEARNING will reach out to discuss your readiness for the voucher and make arrangements to send it.