Building with Open-Source LLMs:

Keep your Data Private

Course Overview

This advanced course is designed for senior machine learning engineers, technical/product leads, and hands-on engineering leaders aiming to leverage open-source or self-hosted LLMs for enterprise applications, especially those with domain-specific requirements and sensitive data. You will dive deep into using the best open-source models for different use cases that leverage prompt engineering, RAG, and fine-tuning methods and work to achieve frontier model performance on your private cloud or local compute infrastructure.

Learning Outcomes

By the end of this course, you will:

1. Develop practical skills and insights by applying different GenAI design patterns (i.e. advanced prompt engineering, RAG, fine-tuning) to build robust solutions leveraging open-source foundation models
2. Learn to adapt open-source pre-trained models to specific use cases or domains, ensuring they excel in specialized tasks leveraging prompt engineering techniques such as Chain-of-Thought (CoT), few-shot learning, automated methods like DSPy, and more.
3. Implement end-to-end RAG pipelines with features such as query decomposition, colbertV2 retrieval, hybrid search, and retrieval-augmented fine-tuning (RAFT)
4. Gain a comprehensive understanding of LLM fine-tuning techniques and practical hands-on experience with LoRA, QLoRA, and best practices for managing model artifacts, effectively evaluating fine-tuned models, and serving multiple fine-tuned LLMs

Recommended Prerequisites

- Proficiency in Python programming
- Solid understanding of machine learning fundamentals
- Familiarity with deep learning concepts
- Intermediate knowledge of Natural Language Processing (NLP) is recommended

Detailed Curriculum Schedule

Topic	Topics
Prompt Engineering & Advanced RAG	• A brief introduction to open-source LLMs, evaluation benchmarks, and choosing the best starting point • Conceptual Introduction to advanced prompt engineering techniques (i.e. CoT, few-shot learning, DSPy, etc.) • **Hands-on activity #1**: Entity extraction from newspaper articles and comparing frontier models to open-source • Best practices for managing and versioning prompts • Overview of advanced RAG techniques to go beyond a proof-of-concept (i.e. hybrid search, query decomposition / re-writing, RAFT, and RAG evaluation best practices) • **Hands-on activity #2**: RAG pipeline with ArXiv research papers
Fine-Tuning & Model Deployment	• Overview of LoRA, QLoRA, when to consider fine-tuning, and guidance for setting LORA fine-tuning parameters • **Hands-on activity #3**: Fine-tuning an entity extraction model and evaluating performance improvement • Highlight best practices for managing model artifacts and reproducible workflows • **Hands-on activity #4**: Efficiently deploying and serving LoRA fine-tuned models • Recap of key takeaways and additional resources to further your learning on building with open-source LLMs.

Register Now!

About FourthBrain

FourthBrain trains engineers, developers, data scientists, and leaders to make an impact in the Artificial Intelligence field, with our flexible, accessible education programs. We are training a new generation of engineers and leaders with more than just technical ability; they have an awareness and mindset of what is needed to succeed with AI. We are part of the AI Fund, founded by Andrew Ng.