Building with Large Language Models

Leveraging LLMs in the Age of Generative AI

Your Learning Transformation

By the end of this course, you will be able to **build** and **deploy** complex AI applications that produce **high-quality** outputs using the latest in open-source LLM technology.

Learning Outcomes

By the end of this workshop you will:

1. Learn best-practices of prompt engineering and fine-tuning
2. Learn how to fine-tune and instruct-tune LLMs using data-centric approaches
3. Create a ChatGPT-like interface for your own data using indexing and chaining
4. Build and deploy an end-to-end Generative AI application each week!

Project-Based Learning

In this workshop, you will deploy three applications in three weeks, including your own unique LLM capstone project in Week 3. You'll leverage the most powerful LLMs to build and deploy personalized applications in your domain and for your use cases. You'll also fine-tune LLMs to improve classic ML algorithm performance, teach LLMs new structure, and build a question-answering tool for your documents.

Technical Requirements and Prerequisites

This workshop is intended for software engineers, data scientists, ML engineers, or others who have strong Python skills and have working experience deploying applications. Before the workshop, you should:

- Create an account on Hugging Face and be familiar with Hugging Face Datasets, Models, and Spaces. We recommend checking out our Zero to Deployment Adventure Guide from our recent Building Generative AI Applications Workshop, and checking out the Hugging Face Course for more information.
- Create a Google Colab Pro account, which is what we will be using for demos during the workshop.

Detailed Schedule

Module	Topics	Build Activities
Week 1 Fine-Tuning and Deploying LLM Applications	**Introduction and Networking** • Program Overview • A Brief History of GPT Models **Prompt Engineering** • Prompt Engineering Best Practices • Zero-Shot vs. Few-Shot Learning **LLM Fine-Tuning** • Fine-Tuning Input-Output Schema • Instruction-Tuning **Building with LLMs** • The Role of LLM Developer APIs • Leading 2023 LLM Models • The Hugging Face Stack (Datasets, Models, Spaces)	• **Keeping up with LLMs!** • **Building with gpt-3.5-turbo and OpenAI's API** • Interactive End-to-End BLOOMZ Data Curation, Fine-Tuning, and Deployment Demo on Hugging Face • **Homework: Fine-Tuning BLOOMZ with Google Colab** • **Homework: Capstone Project Commitments**
Week 2 Indexing & Document Question Answering	**Technologies for "Building ChatGPT For Your Data"** • *Indexing*, or Structuring Documents for LLMs to Interact with Them • *Vector Databases*, or Search and Retrieval using Vector Embeddings • *Chaining*, or Building Complex AI Applications with LLMs • *Agents*, or Building with Chains that depend on user inputs • Emerging Tools on the Market (e.g., LangChain, LlamaIndex, HayStack) **Applications Considerations** • From One Document to Many • From Qualitative QA to Quantitative • Chatbots versus Virtual Assistants	• **Answering Questions from a Single Document** • Interactive End-to-End Document Question Answering Deployment Demo on Hugging Face • **Homework: Indexing and Querying Multiple Documents using Indexing with LangChain and Open AI's GPT 3.5** • **Homework: Capstone Project MVPs (Dataset, Model Card, Space)**
Week 3 Validating LLM Application	**How Do I Know If My Output is Good?** • Challenges with Assessing Quality • Early LLM Metrics (Perplexity,	• Interactive Demos of how to leverage Stanford's HELM and Eluether AI's Test

Outputs and Demo Day!	Burstiness, etc.) • Emerging Metrics & Model Card Best-Practices **Benchmark Frameworks & Test Harnesses** • A Brief History of NLP Benchmarks • LLM Trade Offs: Size and Capability vs. Inference Cost • Stanford's Holistic Evaluation of Large Language Models (HELM) • Eleuther AI's Test Harness	Harness • **PROJECT: Curate Data, Build a Fine-Tuned AI Application. Deploy on Hugging Face; include a framework for measuring quality of outputs**

* Activities in **Bold Pink** will be working sessions during the class!

Register for Building with LLMs

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 who have 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.