LLMOps:

The Art of Implementation

Course Description

This workshop is a comprehensive three-week course in LLMOps and operationalizing LLM-based applications. You will focus on building LLM applications, evaluating their performance, monitoring and incorporating feedback, and diving into considerations for operationalizing LLMs at scale. We will cover advanced RAG techniques, and review both cloud-based and on-prem solutions.

Learning Outcomes

By the end of this workshop, you will:

1. Understand the LLMOps lifecycle and how it differs from MLOps and work operationalizing multiple LLM applications end-to-end
2. Build with advanced RAG techniques and implement monitoring and feedback loops for continuous improvement
3. Build an evaluation pipeline using industry-standard LLM metrics
4. Build a multi-modal application as a cloud API service and learn about productionizing at scale

Technical Requirements and Prerequisites

This workshop is intended for machine learning engineers, data engineers, devops engineers, and software engineers who have experience working with LLMs and building LLM-based applications.

Before the workshop, you should:

- Create an account on Hugging Face and be familiar with Hugging Face Datasets, Models, and Spaces. We recommend the Hugging Face Course for more information.
- Create a Google Colab Pro account which is the GPU compute environment that will be used for demos during the workshop.

Detailed Schedule

Module	Topics	Build Activities
Week 1 Operationalizing LLM-based Applications	• Review of LLMOps lifecycle, including differentiating it from MLOps • Master prompt engineering techniques including prompt management and experiment tracking • Enhance LLM generation with RAG and fine-tuning, including deploying models and evaluating performance • Host and deploy an LLM on your own infrastructure, including using techniques like PEFT, LORA, fine-tuning, knowledge distillation and quantization, while considering GPU infrastructure and cost trade-offs • Practical techniques for evaluating LLM applications, optimizing performance, and fine-tuning for specific tasks	• **Project:** Build an LLM application that classifies legal examples using Legal Bench • **Tools/Frameworks Used:** Langchain, Hugging Face, FastAPI, Streamlit, Google Colab, RAG, Quantization, Weights and Biases
Week 2 Building Production-Ready Advanced RAG Applications	• Generate question-answering validation data manually and/or synthetically with GPT-4 • Explore advanced RAG enhancements, such as improved embedding models, optimal chunk sizes, re-rankers, query expansion, and more... • Implement experiment tracking to document and analyze experiments • Develop an operational dashboard to monitor key metrics, including API request logs, response latency, retrieved-context documents, and model responses	• **Project:** Use a dataset of 2023 NeurIPS papers to create a chat interface with advanced RAG pipeline to help a research team synthesize work ideas • **Tools/Frameworks Used:** Langchain, Hugging Face, transformers, TRL (Transformer Reinforcement Learning), Weight & Biases, Langsmith, Prometheus, Grafana, FastAPI, Streamlit

	• Implement a lightweight user feedback mechanism • Analyze feedback data to identify and address underperforming queries, and improve a model with techniques like Kahneman-Tversky Optimization (KTO) as an alternative to traditional RLHF or DPO	
Week 3 Beyond Text: Operationalizing at Scale of Multi Modal Models	• Review LLM capabilities beyond text generation, including architecture and limitations of fusing text with images, audio and other modalities • Design multi-modal prompts • Deploy at scale with cloud-managed LLM services, taking into consideration cost, performance, and control • Deploy LLM services as scalable APIs through orchestrated pipelines, version management, and automated testing • Review safety and security needs • monitor and address harmful biases and utilize explainability capabilities	• **Project:** Build a multi-modal content generation application • **Tools / Frameworks used:** Cloud, containers, APIs, CI/CD tools

Register For LLMOps: The Art of Implementation

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.