ADDING NEW KNOWLEDGE TO LLMS

Course Code: 847000

This course provides a comprehensive, hands-on guide to the essential techniques for augmenting and customizing LLMs.

This course takes you on a complete journey from raw data to a fine-tuned, optimized model. You will begin by learning how to curate high-quality datasets and generate synthetic data with NVIDIA NeMo Curator. Next, you will dive deep into the crucial process of model evaluation, using benchmarks, LLM-as-a-judge, and the NeMo Evaluator to rigorously assess model performance. With a solid foundation in evaluation, you will then explore a suite of powerful customization techniques, including Continued Pretraining to inject new knowledge, Supervised Fine-Tuning to teach new skills, and Direct Preference Optimization (DPO) to align model behavior with human preferences.

What You'll Learn

- Curate high-quality datasets and generate synthetic data using NVIDIA NeMo Curator.
- Rigorously evaluate LLM performance with benchmarks (MMLU), LLM-as-a-judge, and the NeMo Evaluator.
- Inject new domain-specific knowledge into LLMs using Continued Pretraining (CPT).
- Teach LLMs new skills and align them to specific tasks with Supervised Fine-Tuning (SFT).
- Align model behavior to human preferences for style, tone, and safety using Direct Preference Optimization (DPO).
- Compress and optimize LLMs for efficient deployment using Quantization, Pruning, and Knowledge Distillation with TensorRT-LLM and NeMo.
- Apply end-to-end model customization workflows to solve real-world problems.
- Topics Covered

Prerequisites

- Familiarity with Python programming and Jupyter notebooks.
- Basic understanding of Large Language Models and their applications.
- Conceptual knowledge of deep learning and neural networks.
- Tools, libraries, frameworks used: Python, NVIDIA NeMo, NVIDIA TensorRT-LLM, Docker, MLflow

Visit us at www.globalknowledge.com or call us at 1-866-716-6688.