Building RAG Agents with LLMs (NVIDIA)

Duration: 16 hours

Agents powered by large language models (LLMs) have shown great retrieval capability for using tools, looking at documents, and plan their approaches. This course will show you how to deploy an agent system in practice with the flexibility to scale up your system to meet the demands of users and customers.

Course Prerequisites:

- Introductory deep learning knowledge, with comfort with PyTorch and transfer learning preferred.
- Intermediate Python experience, including object-oriented programming and libraries.

About this Course

The evolution and adoption of large language models (LLMs) have been nothing short of revolutionary, with retrieval-based systems at the forefront of this technological leap. These models are not just tools for automation; they are partners in enhancing productivity, capable of holding informed conversations by interacting with a vast array of tools and documents. This course is designed for those eager to explore the potential of these systems, focusing on practical deployment and the efficient implementation required to manage the considerable demands of both users and deep learning models. As we delve into the intricacies of LLMs, participants will gain insights into advanced orchestration techniques that include internal reasoning, dialog management, and effective tooling strategies.

Learning Objectives

- Compose an LLM system that can interact predictably with a user by leveraging internal and external reasoning components.
- Design a dialog management and document reasoning system that maintains state and coerces information into structured formats.
- Leverage embedding models for efficient similarity queries for content retrieval and dialog guardrailing.
- Implement, modularize, and evaluate a RAG agent that can answer questions about the research papers in its dataset without any fine-tuning.

Topics Covered

The workshop includes topics such as LLM Inference Interfaces, Pipeline Design with LangChain, Gradio, and LangServe, Dialog Management with Running States, Working

with Documents, Embeddings for Semantic Similarity and Guardrailing, and Vector Stores for RAG Agents. Each of these sections is designed to equip participants with the knowledge and skills necessary to develop and deploy advanced LLM systems effectively.

Course Outline

- Introduction to the workshop and setting up the environment.
- Exploration of LLM inference interfaces and microservices.
- Designing LLM pipelines using LangChain, Gradio, and LangServe.
- Managing dialog states and integrating knowledge extraction.
- Strategies for working with long-form documents.
- Utilizing embeddings for semantic similarity and guardrailing.
- Implementing vector stores for efficient document retrieval.
- Evaluation, assessment, and certification.