Skillsoft
global knowledge™

BUILDING RAG AGENTS WITH LLMS

Course Code: 847001

This course will observe how you can deploy an agent system in practice and scale up your system to meet the demands of users and customers.

Agents powered by large language models (LLMs) are quickly gaining popularity as people are finding new capabilities and opportunities to greatly improve their productivity. An especially powerful recent development has been the popularization of retrieval-based LLM systems that can hold informed conversations by using tools, looking at documents, and planning their approaches. These systems are fun to experiment with and offer unprecedented opportunities to make life easier, but they also require many queries to large deep learning models and need to be implemented efficiently. This course will observe how you can deploy an agent system in practice and scale up your system to meet the demands of users and customers. Along the way, you'll learn advanced LLM orchestration techniques for internal reasoning, dialog management, tooling, and retrieval.

What You'll Learn

Our journey begins with an introduction to the workshop, setting the stage for a deep dive into the world of LLM inference interfaces and the strategic use of microservices. We will explore the design of LLM pipelines, leveraging tools such as LangChain, Gradio, and LangServe to create dynamic and efficient systems. The course will guide participants through managing dialog states, integrating knowledge extraction techniques, and employing strategies for handling long-form documents. We will continue with an examination of embeddings for semantic similarity and guardrailing, culminating in the implementation of vector stores for document retrieval. The final phase of the course focuses on the evaluation, assessment, and certification of participants, ensuring a comprehensive understanding of RAG agents and the development of LLM applications.

- 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.