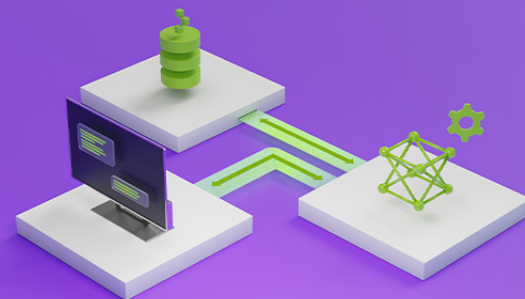




Building RAG Agents With LLMs



Workshop Overview

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.

In this course, you'll design retrieval-augmented generation (RAG) systems and bundle them into deliverable formats. Along the way, you'll learn advanced LLM composition techniques for internal reasoning, dialog management, and tooling.

Learning Objectives

By participating in this workshop, you'll learn how to:

- > 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

Overview

Duration	8 hours
Price	Contact us for pricing.
Prerequisites	Introductory deep learning, with comfort with PyTorch and transfer learning preferred. Content covered by DLI's Getting Started With Deep Learning or Fundamentals of Deep Learning courses or similar experience is sufficient. Intermediate Python experience, including object-oriented programming and libraries. Content covered by Python Tutorial (w3schools.com) or similar experience is sufficient.
Tools, libraries, and frameworks	Python, LangChain, NVIDIA AI Foundation Endpoints, FAISS, Gradio, LangServe, FastAPI
Assessment type	A RAG-enabled chatbot that can answer questions from a pool of research papers.
Certificate	Upon successful completion of the assessment, participants will receive an NVIDIA DLI certificate to recognize their subject matter competency and support professional career growth.
Hardware requirements	Desktop or laptop computer capable of running the latest version of Chrome or Firefox. Each participant will be provided with dedicated access to a fully configured, GPU-accelerated workstation in the cloud.
Language	English

Workshop Outline

Introduction

(15 minutes)

Meet the instructor.

- > Create an account at courses.nvidia.com/join and catalog.ngc.nvidia.com.
- > Learn about the course objectives and get comfortable with the format.

LLM Inference Interfaces

(60 minutes)

Explore the course environment, microservices, and LLM inferencing options:

- > Get comfortable with the course environment and learn about microservices for software compartmentalization and resource delivery.
- > Discuss LLM service options for inference use cases, including local and scalable deployment strategies and value propositions.
- > Get comfortable with remotely accessible access points like GPT4 and NGC™-hosted NVIDIA AI Foundation models and endpoints.

Break (15 minutes)

Pipeline Design With LangChain, Gradio, and LangServe

(60 minutes)

Orchestrate LLM endpoints into pipelines using open-source frameworks:

- > Learn how to use LangChain to chain multiple LLM-enabled modules using the functional LangChain Expression Language (LCEL) syntax.
- > Formalize internal and external reasoning and modularize them into runnables.
- > Use LangServe to interact with a Gradio frontend by sending an LLM chain over a port interface.

Break (60 minutes)

Dialog Management With Running States

(45 minutes)

Develop running logic systems to remember information and guide dialog:

- > Learn about running state logic to retain state as your chain runs.
- > Leverage knowledge extraction via slot filling to keep a smart knowledge base.
- > Integrate a dialog managing chatbot to coerce the user for credentials, retrieve info from a database interface, and maintain dialog state.

Working With Documents

(45 minutes)

Learn about how to work with longer-form documents that exceed context limits:

- > Learn about document chunking, reduction, and refinement strategies.
- > Use the same LLM chaining skills to build systems that summarize research papers by exporting a while-loop-enabled runnable.

Embeddings for Semantic Similarity and Guardrailing

(60 minutes)

Explore the use of embedding models for vector-enabled semantic reasoning:

- > Formalize encoder-vs-decoder benefits and understand how embedding logic works.
- > Use vector representations to reason about passage meanings and similarity.
- > Design a guardrailing system that leverages a custom-build input rail to answer a question or kindly refuse.

Break (15 minutes)

Vector Stores for RAG Agents

(60 minutes)

Integrate vector stores to help agent systems retrieve and reason with documents:

- > Formalize vector stores as structures that help automate vector reasoning logic.
- > Incorporate vector stores into retrieval-augmented generation pipelines that reason about conversation history and preprocessed document pools.

Evaluation, Assessment, and Q&A

(60 minutes)

Evaluate your RAG system using LLM-as-a-judge evaluation chains:

- > Review key learnings and motivate evaluation as a natural progression.
- > Launch your retriever component into the frontend to run the assessment.
- > **Finish the course and complete the assessment to earn a certificate.**

Why Choose NVIDIA Deep Learning Institute for Hands-On Training?

- > Access workshops from anywhere with just your desktop/laptop and an internet connection. Each participant will have access to a fully configured, GPU-accelerated server in the cloud.
- > Obtain hands-on experience with the most widely used, industry-standard software, tools, and frameworks.
- > Learn to build deep learning and accelerated computing applications for industries, such as healthcare, robotics, manufacturing, accelerated computing, and more.
- > Gain real-world experience through content designed in collaboration with industry leaders, such as the Children's Hospital of Los Angeles, Mayo Clinic, and PwC.
- > Earn an NVIDIA DLI certificate to demonstrate your subject matter competency and support your career growth.

Ready to Get Started?

For the latest DLI workshops and trainings, visit
www.nvidia.com/dli

For questions, contact us at nvdl@nvidia.com

© 2024 NVIDIA Corporation and affiliates. All rights reserved. NVIDIA, the NVIDIA logo, and NGC are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. All other trademarks and copyrights are the property of their respective owners. 3141100. FEB24

