About this Learning Path

By the end of this learning path, you'll have mastered the tools and techniques needed to design, build, and deploy impactful RAG-based systems. From document processing to enhancing web interactions and personalizing professional engagements, this journey equips you to confidently tackle real-world challenges and drive innovation with the latest technologies in your field.

What is RAG?

Retrieval-Augmented Generation (RAG) is a powerful technique that enhances the capabilities of LLMs by integrating external data sources into their reasoning process. While LLMs excel at reasoning across a broad range of topics, their knowledge is limited to the publicly available data they were trained on, up to a specific cut-off date. RAG addresses this limitation by enabling AI applications to reason about private or newly introduced data. This is achieved by dynamically retrieving relevant information and inserting it into the model's input prompt, effectively "augmenting" the model's knowledge.

RAG is particularly useful for building sophisticated question-answering (Q&A) applications and other interactive tools, such as chatbots, that can answer queries based on specific source data. By combining retrieval mechanisms with generative AI capabilities, RAG allows you to leverage the strengths of LLMs while ensuring that the responses are accurate, timely, and contextually relevant.

RAG Architecture

A typical RAG system is built on two key components:

1. **Indexing**: This offline process involves preparing the data for retrieval by loading, splitting, and storing documents into an indexed format, often using a VectorStore and embeddings models.
2. **Retrieval and Generation**: At runtime, the system retrieves relevant indexed data in response to user queries and integrates it into prompts for LLMs to generate accurate, informed, and contextually aware outputs.

This architecture makes RAG systems indispensable for applications requiring dynamic reasoning with private or real-time data.

Learning Path Overview

Begin your journey with **"Summarize Private Documents Using RAG, LangChain, and LLMs"**, where you'll learn to split and embed private documents for efficient processing. This foundational project introduces secure document summarization using advanced LLMs and demonstrates how to create a chatbot capable of retrieving key information while maintaining data privacy.

Progress to **"RAG with Granite 3: Build a Retrieval Agent Using LlamaIndex"** to deepen your expertise in loading, indexing, and retrieving data from diverse sources like PDFs, HTML, and text files. You'll develop an AI assistant that excels in delivering precise insights, making it invaluable for applications like scientific research and professional analyses.

Next, explore real-world applications in **"Build a Grounded Q/A Agent with Granite 3, LangChain, and RAG"**, where you'll configure LangChain and IBM watsonx Granite LLMs to create a retrieval-augmented pipeline. In just 30 minutes, you'll build a question-answering agent capable of delivering accurate and context-aware responses tailored to specific queries.

Expand your skills further with **"Build a RAG System for Web Data with LangChain and Llama 3.1."** This project focuses on real-time web data retrieval and analysis, enabling you to create dynamic, context-aware interactions using Llama on watsonx.ai.

Proceed to video content processing with **"AI-Powered YouTube Summarizer, Q&A Tool with RAG, LangChain"** You'll extract video transcripts, generate summaries, and build interactive Q&A systems. This project leverages FAISS for efficient segment retrieval and LLMs for advanced NLP, offering powerful tools for saving time and enhancing engagement.

Finally, cap your learning path with **"Build an AI Icebreaker Bot with IBM Granite 3.0 & LlamaIndex."** This project integrates ProxyCurl API to extract LinkedIn profile data and uses LlamaIndex to build a vector database. You'll develop a conversational bot capable of generating tailored icebreakers, perfect for networking events and professional interactions.

☆