Module 8 of 10

Multimodal RAG

Grid University

This module aims to teach you how to effectively integrate and work with diverse data modalities within a RAG system. You will learn the fundamental approaches to handling multimodal data, enabling more effective retrieval and storage, and ultimately implementing a more comprehensive RAG solution.

Key learning areas:

- **Approaches to Multimodal RAG Implementation:** Delve into the primary architectural strategies for building multimodal RAG systems, including:
  - **Shared Vector Space:** Mapping all modalities into a common embedding space for unified retrieval.
  - **Single Grounded Modality (Conversion to Text):** Transcribing or captioning non-textual data into text before retrieval, simplifying the pipeline but potentially losing nuance.
  - **ColPali approach:** treat an entire document page as an image and directly analyse and understand not only the text but also crucial visual features such as layouts, tables, charts, figures, and even font styles and colours
- **Leveraging Multimodal LLMs for Generation:** Understand how to integrate and utilize Large Multimodal Models (LMMs) that can directly process and generate responses from a combination of text, images, and other modalities, enhancing the final output grounded in diverse contexts.

By the end of this module, you will be able to add multimodal data sources to your RAG system and effectively choose the optimal strategy for working with them, enabling your RAG solution to understand and respond to queries involving various data types.

To complete this module, you need to finish the course listed below and review the reading materials.

Building Multimodal RAG systems

With the help of these materials you'll dive into the core architectural approaches for building Multimodal RAG systems. You'll learn the distinct strategies for integrating diverse data types using a shared vector space, converting modalities to text, employing separate retrieval pipelines, or combining these in hybrid approaches:

Embedding multimodal data for similarity search using transformers, datasets and FAISS
Baseline approach of transforming everything into a single modality.
READ

Unlocking the Power of Multimodal Embeddings
Multimodal embeddings (Cohere models).
READ

Multimodal Retrieval-Augmented Generation (RAG) with Document Retrieval (ColPali) and Vision Language Models (VLMs)
ColPali approach.
READ

LLMs and Multimodal Retrieval in RAG Systems

Learn how LLMs can directly process diverse retrieved content (text, images, etc.) to generate more nuanced and contextually rich answers.

Image understanding
Gemini models are built to be multimodal from the ground up, unlocking a wide range of image processing and computer vision tasks including but not limited to image captioning, classification, and visual question answering without having to train specialized ML models.
READ

Video understanding
Gemini models can process videos, enabling many frontier developer use cases that would have historically required domain specific models.
READ

Additional readings:

Agentic RAG for PDFs with mixed data
RAG on multimodal PDF document.
READ