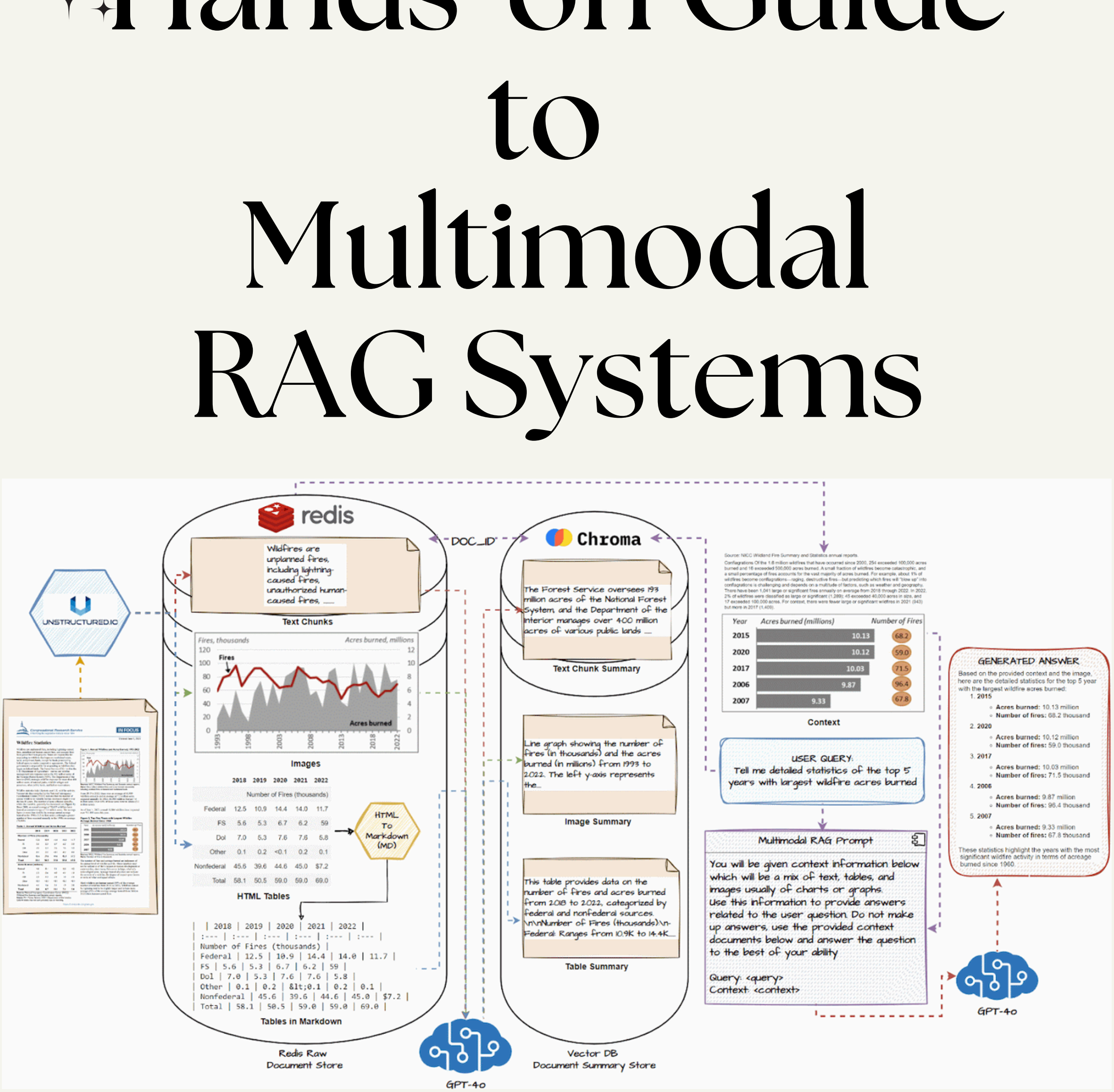
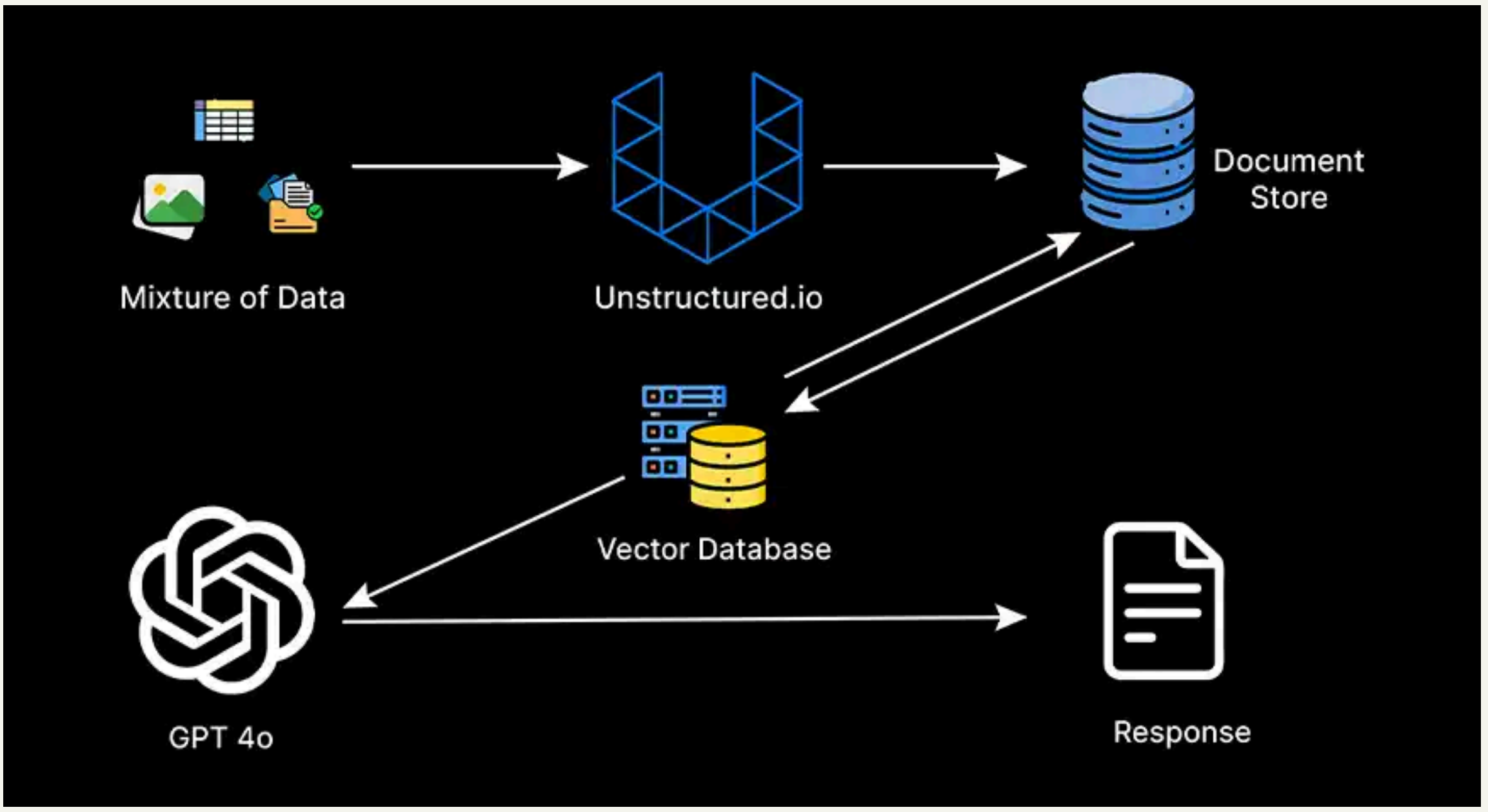
Hands-on Guide to Multimodal RAG Systems



Dipanjan (DJ)

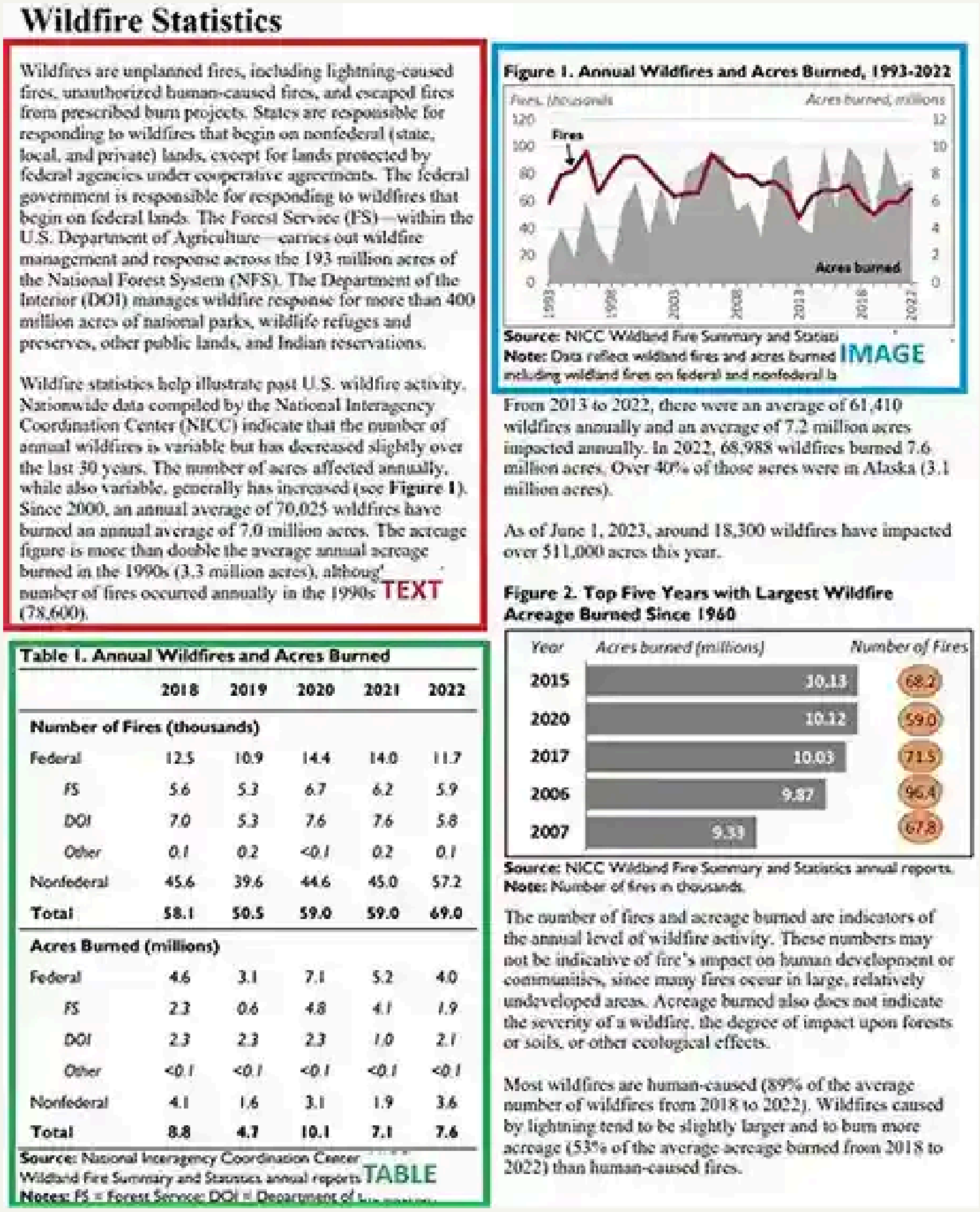
Multimodal RAG System



- **Traditional RAG systems are constrained to text data, making them ineffective for multimodal data, which includes text, images, tables, and more**
- **These systems integrate multimodal data processing (text, images, tables) and utilize multimodal LLMs, like GPT-4o, to provide more contextual and accurate answers.**
- **The guide provides a detailed guide on building a Multimodal RAG system with LangChain, integrating intelligent document loaders, vector databases, and multi-vector retrievers.**

Source: A Comprehensive Guide to Building Multimodal RAG Systems

Multimodal Datasets



Wildfire Statistics

Wildfires are unplanned fires, including lightning-caused fires, unauthorized human-caused fires, and escaped fires from prescribed burn projects. States are responsible for responding to wildfires that begin on nonfederal (state, local, and private) lands, except for lands protected by federal agencies under cooperative agreements. The federal government is responsible for responding to wildfires that begin on federal lands. The Forest Service (FS)—within the U.S. Department of Agriculture—carries out wildfire management and response across the 193 million acres of the National Forest System (NFS). The Department of the Interior (DOI) manages wildfire response for more than 400 million acres of national parks, wildlife refuges and preserves, other public lands, and Indian reservations.

Wildfire statistics help illustrate past U.S. wildfire activity. Nationwide data compiled by the National Interagency Coordination Center (NICC) indicate that the number of annual wildfires is variable but has decreased slightly over the last 30 years. The number of acres affected annually, while also variable, generally has increased (see **Figure 1**). Since 2000, an annual average of 70,025 wildfires have burned an annual average of 7.0 million acres. The acreage figure is more than double the average annual acreage burned in the 1990s (3.3 million acres), although number of fires occurred annually in the 1990s (78,600).

TEXT

Figure 1. Annual Wildfires and Acres Burned, 1993-2022

Source: NICC Wildland Fire Summary and Statisti
Note: Data reflect wildland fires and acres burned including wildland fires on federal and nonfederal la

IMAGE

From 2013 to 2022, there were an average of 61,410 wildfires annually and an average of 7.2 million acres impacted annually. In 2022, 68,988 wildfires burned 7.6 million acres. Over 40% of those acres were in Alaska (3.1 million acres).

As of June 1, 2023, around 18,300 wildfires have impacted over 511,000 acres this year.

Table 1. Annual Wildfires and Acres Burned

	2018	2019	2020	2021	2022
Number of Fires (thousands)					
Federal	12.5	10.9	14.4	14.0	11.7
FS	*5.6*	*5.3*	*6.7*	*6.2*	*5.9*
DOI	*7.0*	*5.3*	*7.6*	*7.6*	*5.8*
Other	*0.1*	*0.2*	*<0.1*	*0.2*	*0.1*
Nonfederal	45.6	39.6	44.6	45.0	57.2
Total	**58.1**	**50.5**	**59.0**	**59.0**	**69.0**
Acres Burned (millions)					
Federal	4.6	3.1	7.1	5.2	4.0
FS	*2.3*	*0.6*	*4.8*	*4.1*	*1.9*
DOI	*2.3*	*2.3*	*2.3*	*1.0*	*2.1*
Other	*<0.1*	*<0.1*	*<0.1*	*<0.1*	*<0.1*
Nonfederal	4.1	1.6	3.1	1.9	3.6
Total	**8.8**	**4.7**	**10.1**	**7.1**	**7.6**

Source: National Interagency Coordination Center Wildland Fire Summary and Statistics annual reports.
Notes: FS = Forest Service; DOI = Department of

TABLE

Figure 2. Top Five Years with Largest Wildfire Acreage Burned Since 1960

Source: NICC Wildland Fire Summary and Statistics annual reports.
Note: Number of fires in thousands.

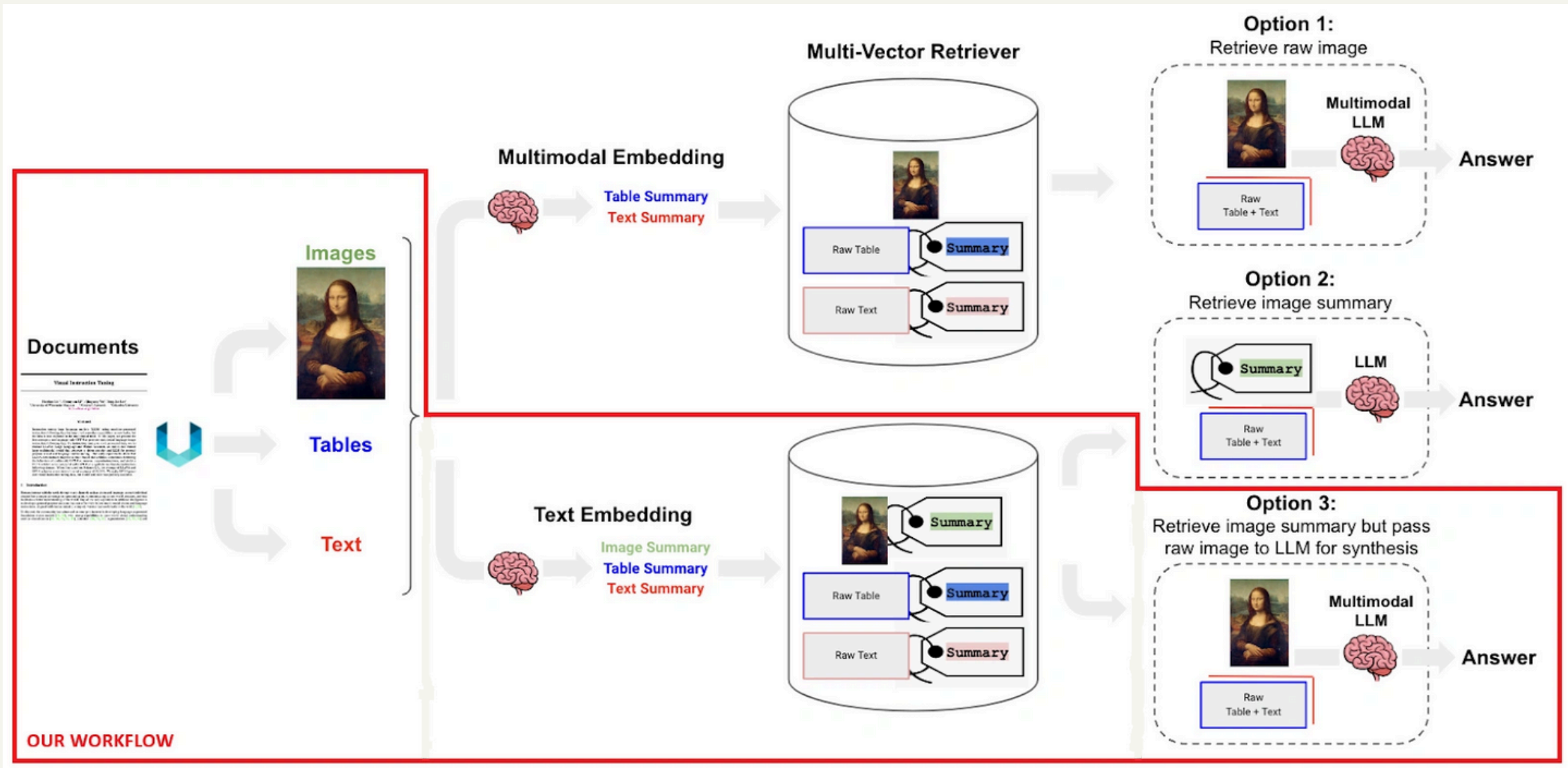
The number of fires and acreage burned are indicators of the annual level of wildfire activity. These numbers may not be indicative of fire's impact on human development or communities, since many fires occur in large, relatively undeveloped areas. Acreage burned also does not indicate the severity of a wildfire, the degree of impact upon forests or soils, or other ecological effects.

Most wildfires are human-caused (89% of the average number of wildfires from 2018 to 2022). Wildfires caused by lightning tend to be slightly larger and to burn more acreage (53% of the average acreage burned from 2018 to 2022) than human-caused fires.

- **Multimodal data consists of a mixture of text, tables, images, graphs and optionally audio and video**
- **The idea is to detect, parse and extract these different elements separately and then generate downstream artifacts like embeddings**

Source: A Comprehensive Guide to Building Multimodal RAG Systems

Multimodal RAG Workflow

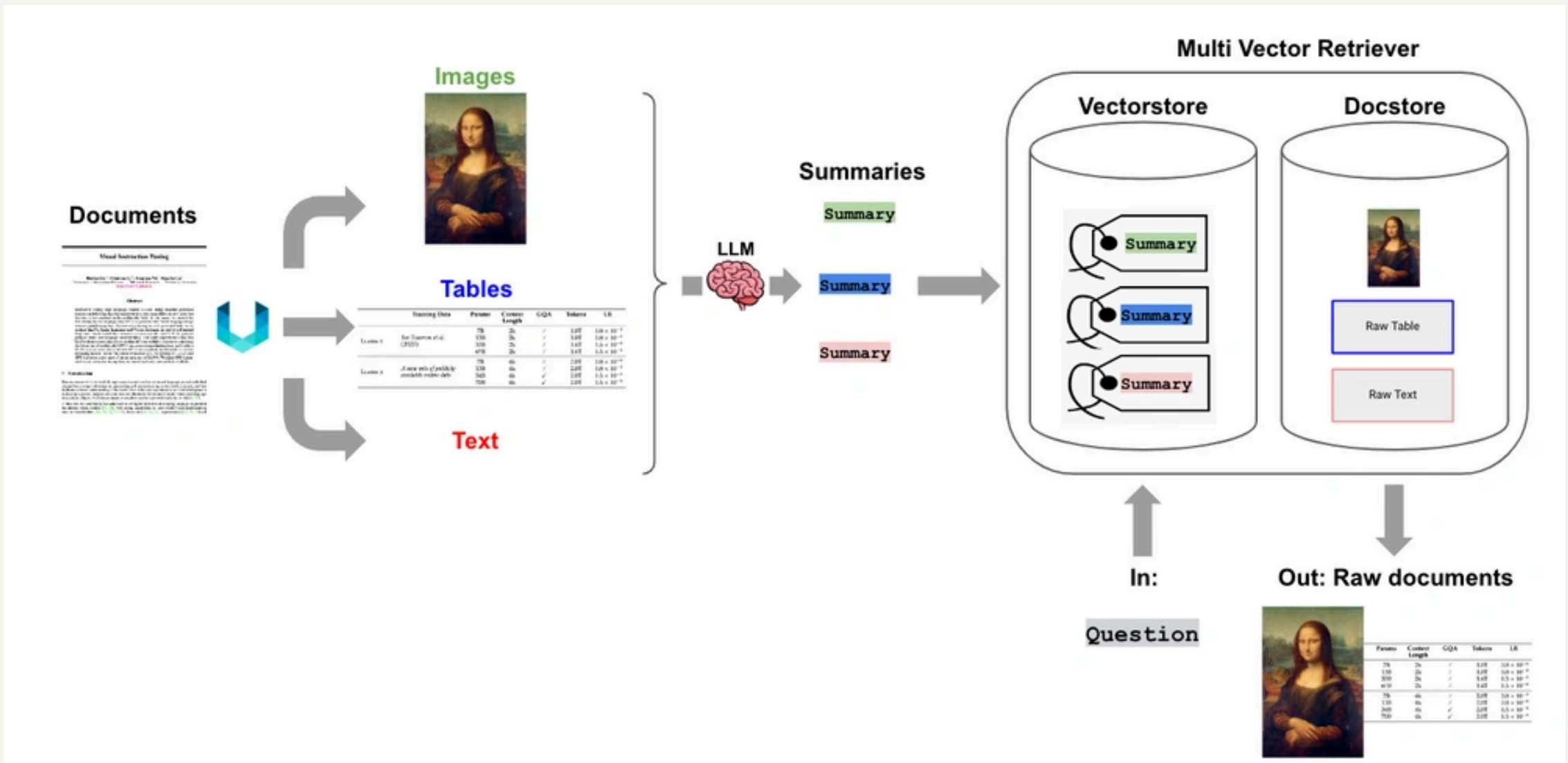


- **Option 1:** Use multimodal embeddings (such as CLIP) to embed images and text together. Retrieve either using similarity search, but simply link to images in a docstore. Pass raw images and text chunks to a multimodal LLM for synthesis.
- **Option 2:** Use a multimodal LLM (such as GPT-4o, GPT4-V, LLaVA) to produce text summaries from images. Embed and retrieve text summaries using a text embedding model. Again, reference raw text chunks or tables from a docstore for answer synthesis by a regular LLM; in this case, we exclude images from the docstore.
- **Option 3:** Use a multimodal LLM (such as GPT-4o, GPT4-V, LLaVA) to produce text, table and image summaries (text chunk summaries are optional). Embed and retrieved text, table, and image summaries with reference to the raw elements, as we did above in option 1. Again, raw images, tables, and text chunks will be passed to a multimodal LLM for answer synthesis.

Option 3 is the best especially if you have charts as images else you can also generate multimodal embeddings from text and image combinations

Source: A Comprehensive Guide to Building Multimodal RAG Systems

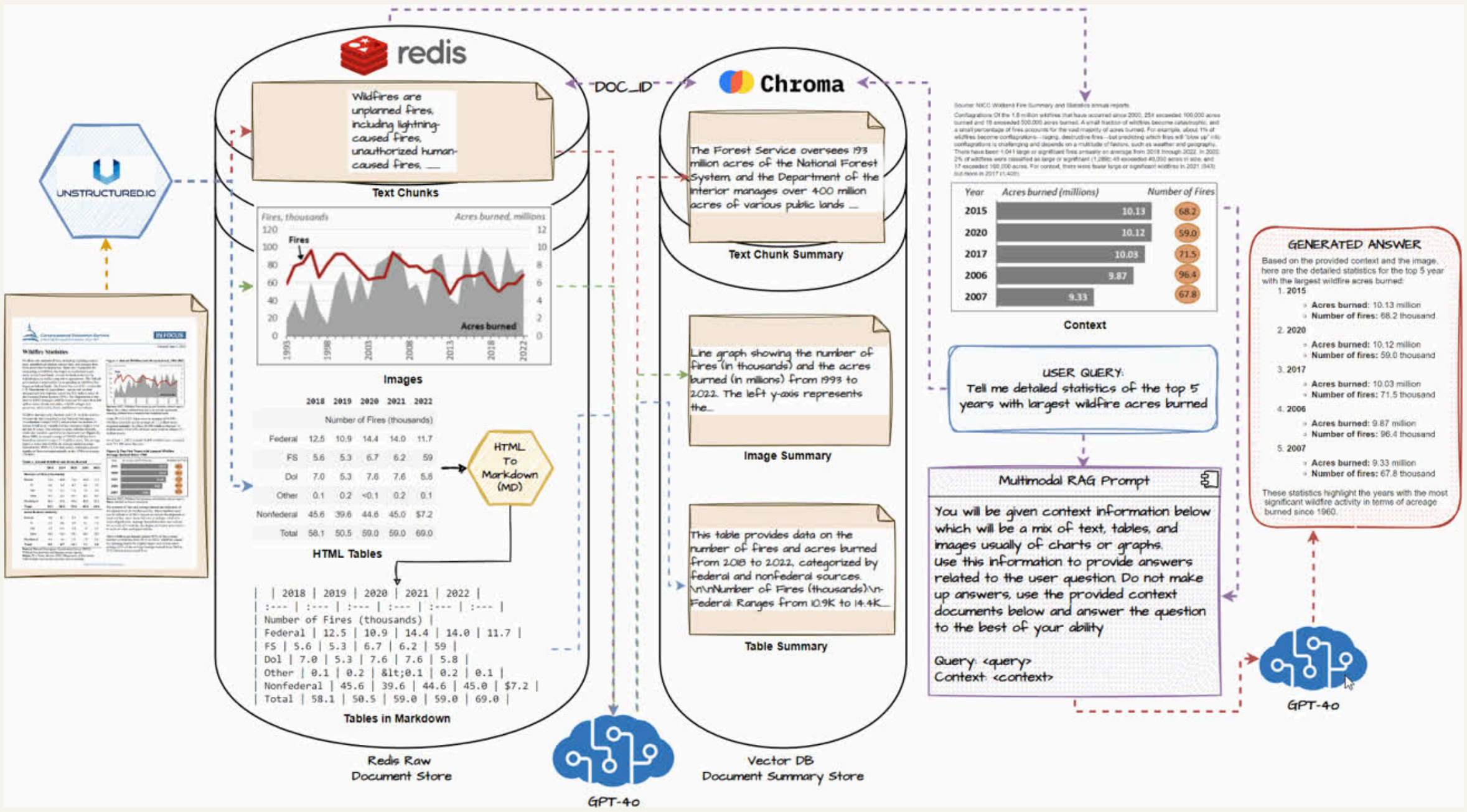
MultiVector Retriever Workflow



- **We will first use a document parsing tool like Unstructured to extract the text, table and image elements separately**
- **Then we will pass each extracted element into an LLM and generate a detailed text summary as depicted above.**
- **Next we will store the summaries and their embeddings into a vector database by using any popular embedder model like OpenAI Embedders. We will also store the corresponding raw document element (text, table, image) for each summary in a document store, which can be any database platform like Redis.**
- **The multi-vector retriever links each summary and its embedding to the original document's raw element (text, table, image) using a common document identifier (doc_id).**
- **Now, when a user question comes in, first, the multi-vector retriever retrieves the relevant summaries, which are similar to the question and then using the common doc_ids, the original text, table and image elements are returned back which are further passed on to the RAG system's LLM as the context to answer the user question.**

Source: A Comprehensive Guide to Building Multimodal RAG Systems

Multimodal RAG Architecture



- **Load all documents and use a document loader like unstructured.io to extract text chunks, image, and tables.**
- **If necessary, convert HTML tables to markdown; they are often very effective with LLMs**
- **Pass each text chunk, image, and table into a multimodal LLM like GPT-4o and get a detailed summary.**
- **Store summaries in a vector DB and the raw document pieces in a document DB like Redis**
- **Connect the two databases with a common document_id using a multi-vector retriever to identify which summary maps to which raw document piece.**
- **Connect this multi-vector retrieval system with a multimodal LLM like GPT-4o.**
- **Query the system, and based on similar summaries to the query, get the raw document pieces, including tables and images, as the context.**
- **Using the above context, generate a response using the multimodal LLM for the question.**

Source: A Comprehensive Guide to Building Multimodal RAG Systems

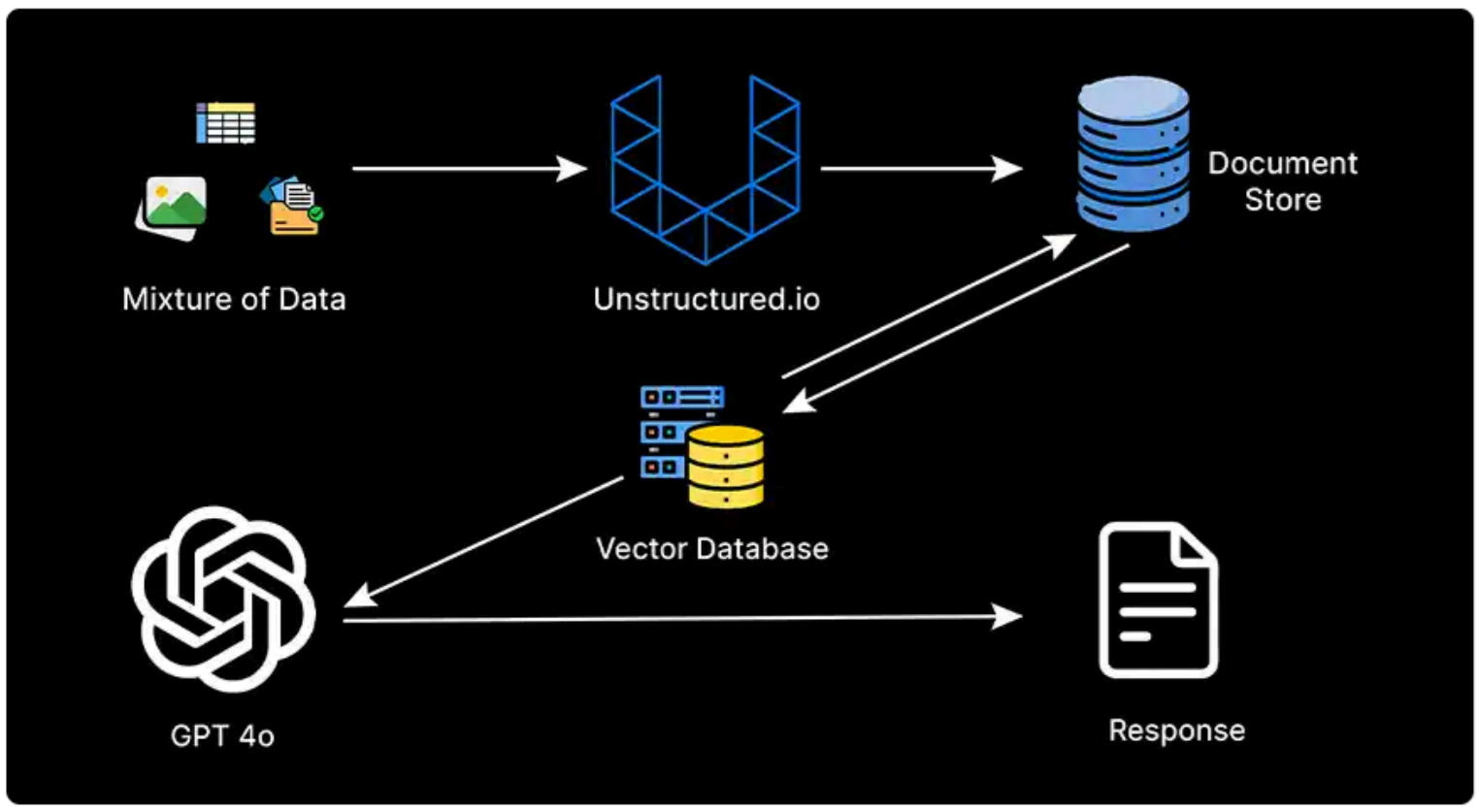
Hands-on Guide

A Comprehensive Guide to Building Multimodal RAG Systems

D Dipanjan Sarkar
Last Updated : 30 Sep, 2024

Introduction

Retrieval Augmented Generation systems, better known as RAG systems, have become the de-facto standard for building intelligent AI assistants answering questions on custom enterprise data without the hassles of expensive **fine-tuning of large language models (LLMs)**. One of the key advantages of RAG systems is you can easily integrate your own data and augment your LLM's intelligence, and give more contextual answers to your questions. However, the key limitation of most RAG systems is that it works well only on text data. However, a lot of real-world data is multimodal in nature, which means a mixture of text, images, tables, and more. In this comprehensive hands-on guide, we will look at building a Multimodal RAG System that can handle mixed data formats using intelligent data transformations and multimodal LLMs.



Overview

CHECK OUT THE HANDS-ON GUIDE HERE