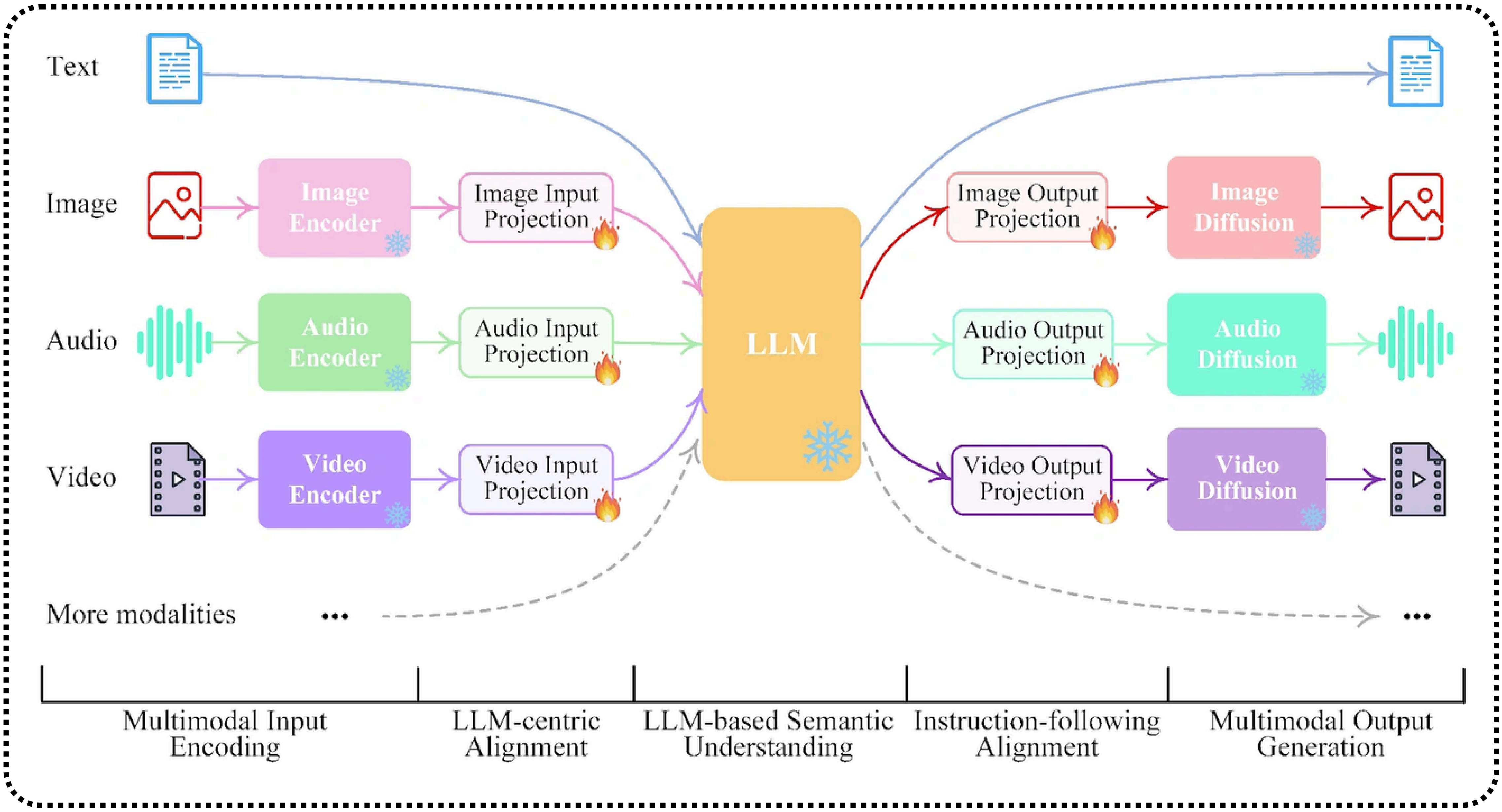
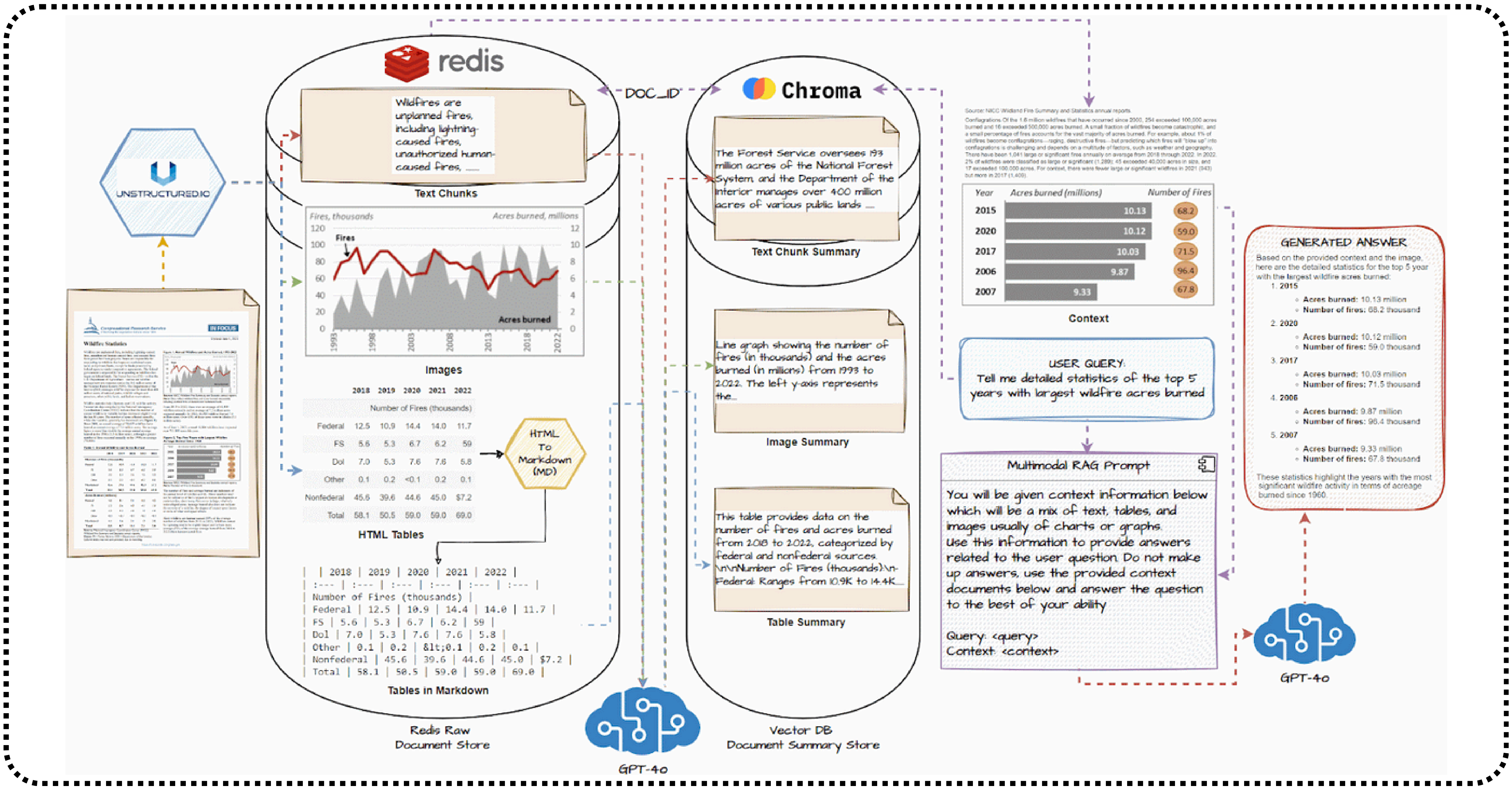
Analytics Vidhya

A Guide to Multimodal RAG System





Wildfires are unplanned fires, including lightning-caused fires, unauthorized human-caused fires, ___

Text Chunks

Images

	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	59
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	$7.2
Total	58.1	50.5	59.0	59.0	69.0

HTML Tables

```
|  | 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 | 59 |
| DoI | 7.0 | 5.3 | 7.6 | 7.6 | 5.8 |
| Other | 0.1 | 0.2 | &lt;0.1 | 0.2 | 0.1 |
| Nonfederal | 45.6 | 39.6 | 44.6 | 45.0 | $7.2 |
| Total | 58.1 | 50.5 | 59.0 | 59.0 | 69.0 |
```

Tables in Markdown

The Forest Service oversees 193 million acres of the National Forest System, and the Department of the Interior manages over 400 million acres of various public lands ...

Text Chunk Summary

Line graph showing the number of fires (in thousands) and the acres burned (in millions) from 1993 to 2022. The left y-axis represents the...

Image Summary

This table provides data on the number of fires and acres burned from 2018 to 2022, categorized by federal and nonfederal sources. \n\nNumber of Fires (thousands):\n- Federal: Ranges from 10.9K to 14.4K...

Table Summary

Source: NIFC Wildland Fire Summary and Statistics annual reports.

Conflagrations Of the 1.8 million wildfires that have occurred since 2000, 254 exceeded 100,000 acres burned and 16 exceeded 500,000 acres burned. A small fraction of wildfires become catastrophic, and a small percentage of fires accounts for the vast majority of acres burned. For example, about 1% of wildfires become conflagrations—raging, destructive fires—but predicting which fires will "blow up" into conflagrations is challenging and depends on a multitude of factors, such as weather and geography. There have been 1,041 large or significant fires annually on average from 2018 through 2022. In 2022, 2% of wildfires were classified as large or significant (1,289); 45 exceeded 40,000 acres in size, and 17 exceeded 100,000 acres. For context, there were fewer large or significant wildfires in 2021 (943) but more in 2017 (1,400).

Year	Acres burned (millions)	Number of Fires
2015	10.13	68.2
2020	10.12	59.0
2017	10.03	71.5
2006	9.87	96.4
2007	9.33	67.8

Context

USER QUERY:
Tell me detailed statistics of the top 5 years with largest wildfire acres burned

Multimodal RAG Prompt

You will be given context information below which will be a mix of text, tables, and images usually of charts or graphs.
Use this information to provide answers related to the user question. Do not make up answers, use the provided context documents below and answer the question to the best of your ability

Query: <query>
Context: <context>

GENERATED ANSWER

Based on the provided context and the image, here are the detailed statistics for the top 5 year with the largest wildfire acres burned:

1. **2015**
   - **Acres burned:** 10.13 million
   - **Number of fires:** 68.2 thousand
2. **2020**
   - **Acres burned:** 10.12 million
   - **Number of fires:** 59.0 thousand
3. **2017**
   - **Acres burned:** 10.03 million
   - **Number of fires:** 71.5 thousand
4. **2006**
   - **Acres burned:** 9.87 million
   - **Number of fires:** 96.4 thousand
5. **2007**
   - **Acres burned:** 9.33 million
   - **Number of fires:** 67.8 thousand

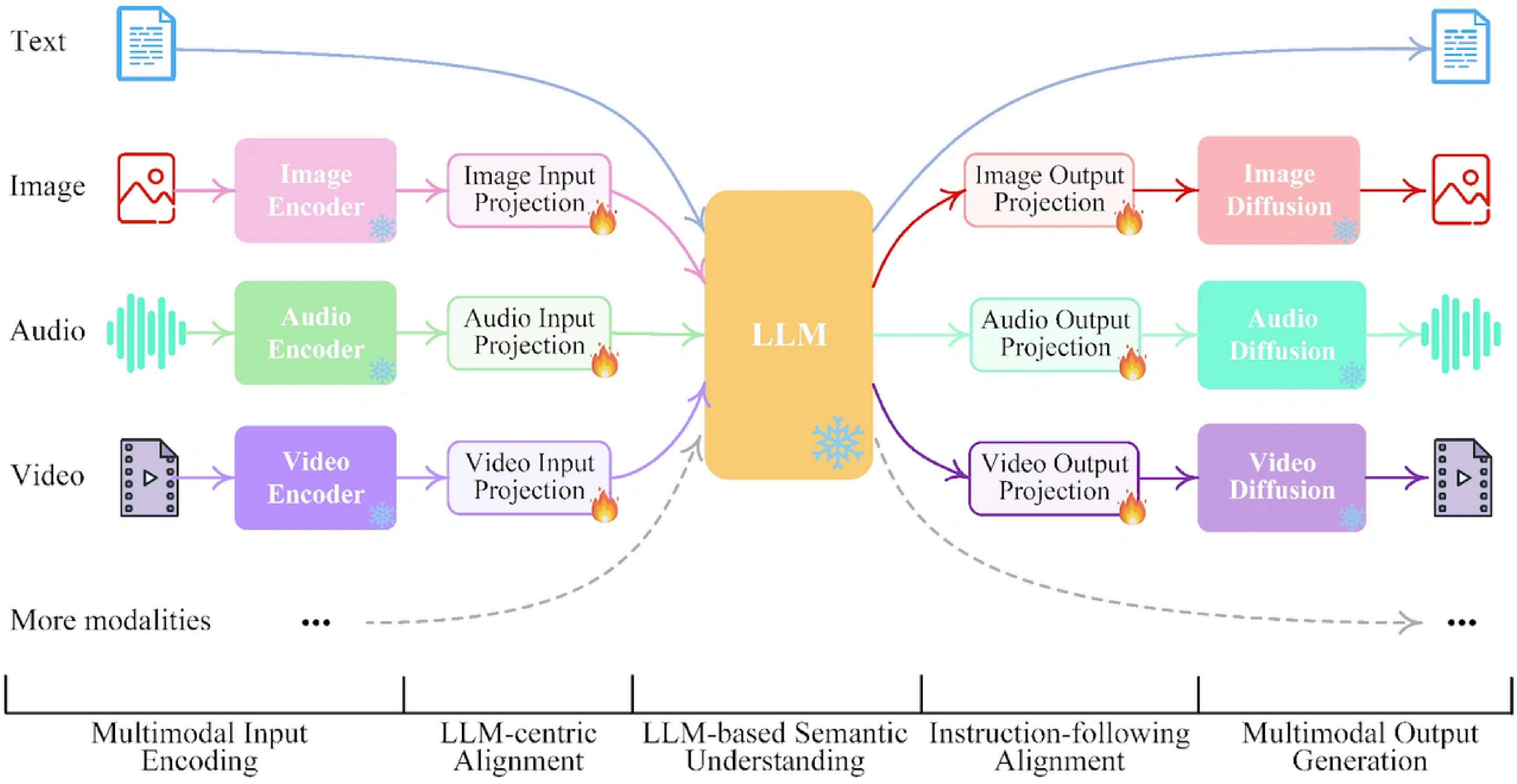
These statistics highlight the years with the most significant wildfire activity in terms of acreage burned since 1960.

Analytics Vidhya

What is a Multimodal Large Language Model?

- Multimodal Large Language Models (LLMs) are essentially transformer-based LLMs that have been pre-trained and fine-tuned on multimodal data to analyze and understand various data formats, including text, images, tables, audio, and video.

- A true multimodal model ideally should be able not just to understand mixed data formats but also generate the same as shown in the following workflow illustration of NExT-GPT, published as a paper, **NExT-GPT: Any-to-Any Multimodal Large Language Model**



From the paper on NExT-GPT, any true multimodal model would typically have the following stages:

- Multimodal Encoding Stage. Leveraging existing well-established models to encode inputs of various modalities.

- LLM Understanding and Reasoning Stage. An LLM is used as the core agent of NExT-GPT. Technically, they employ the **Vicuna LLM** which takes as input the representations from different modalities and carries out semantic understanding and reasoning over the inputs. It outputs 1) the textual responses directly and 2) signal tokens of each modality that serve as instructions to dictate the decoding layers whether to generate multimodal content and what content to produce if yes.

- Multimodal Generation Stage. Receiving the multimodal signals with specific instructions from LLM (if any), the Transformer-based output projection layers map the signal token representations into the ones that are understandable to following multimodal decoders.

However, most current Multimodal LLMs available for practical use are one-sided, which means they can understand mixed data formats but only generate text responses. The most popular commercial multimodal models are as follows:

- **GPT-4V & GPT-4o (OpenAI)**: GPT-4o can understand text, images, audio, and video, although audio and video analysis are still not open to the public.

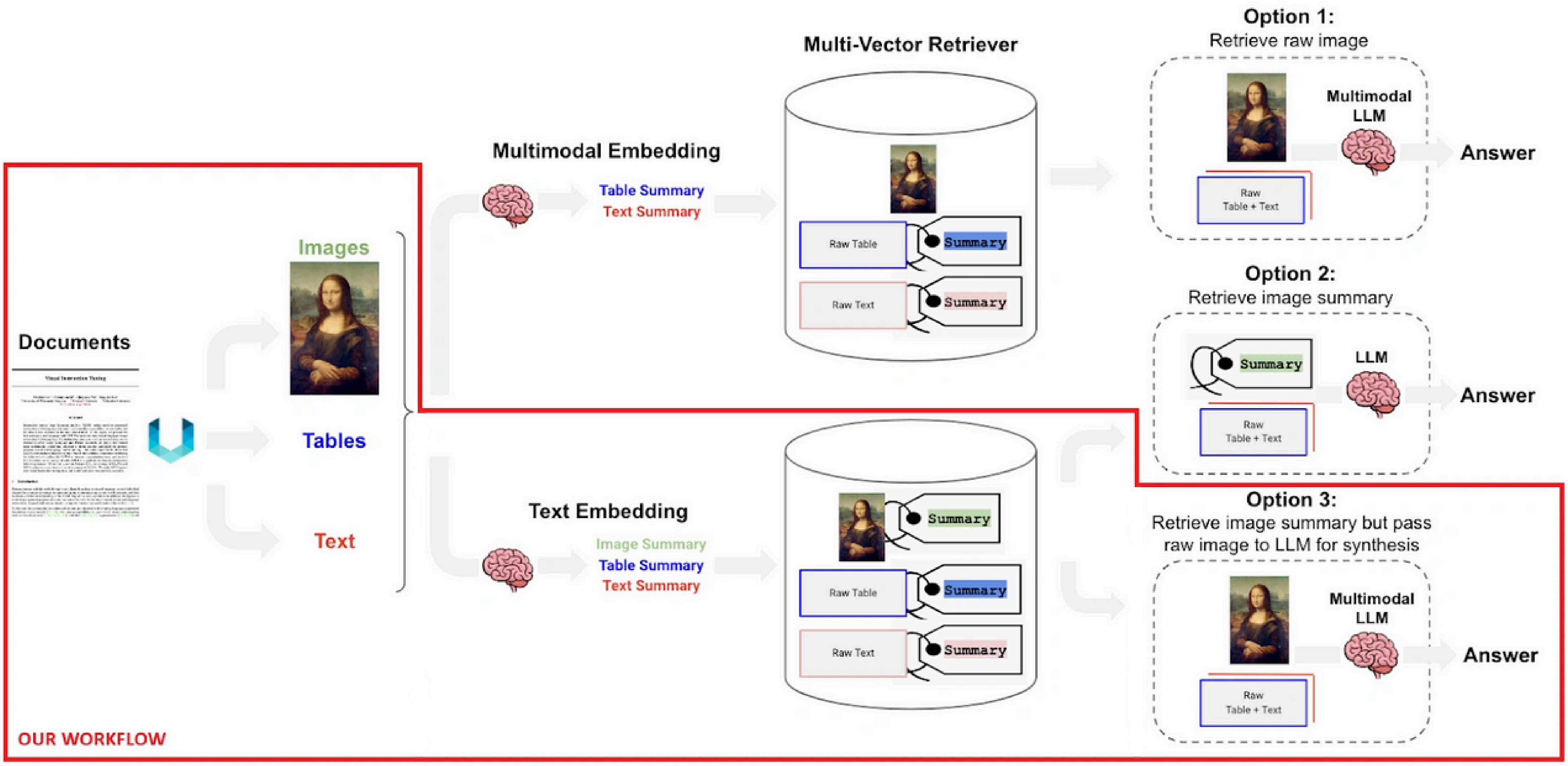
- **Gemini (Google)**: A multimodal LLM from Google with true multimodal capabilities where it can understand text, audio, video, and images.

- **Claude (Anthropic)**: A highly capable commercial LLM that includes multimodal capabilities in its latest versions, such as handling text and image inputs.

For our Multimodal RAG System, we will leverage **GPT-4o**, one of the most powerful multimodal models currently available.

Multimodal RAG System Workflow

- In this section, we will explore potential ways to build the architecture and workflow of a multimodal RAG system. The following figure illustrates potential approaches in detail and highlights the one we will use in this guide.



End-to-End Workflow

Multimodal RAG Systems can be implemented in various ways, the above figure illustrates three possible workflows as recommended in the LangChain blog, this include:

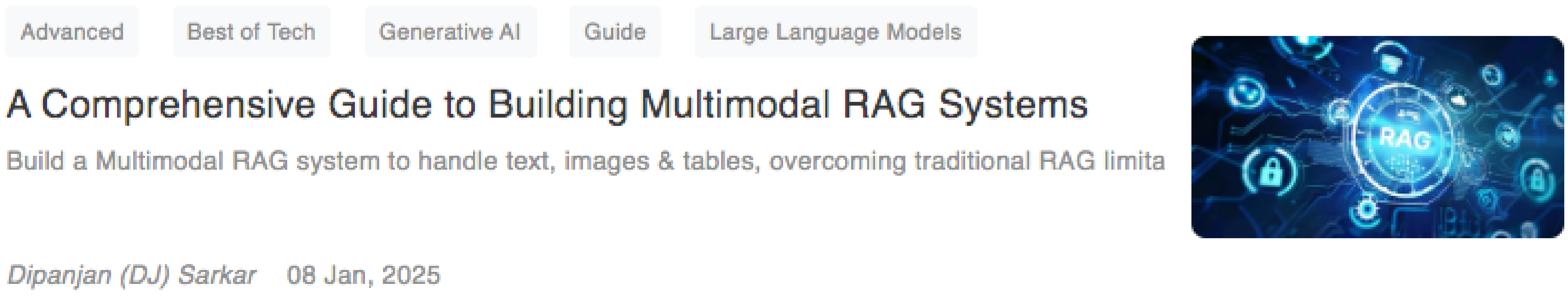
- Option 1: Use multimodal embeddings (such as CLIP) to embed images and text together. Retrieve either using similarity search, but simply link to images in a docstore. Pass raw images and text chunks to a multimodal LLM for synthesis.

- 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. This option is sensible if we don't want to use multimodal embeddings, which don't work well when working with images that are more charts and visuals. However, we can also use multimodal embedding models here to embed images and summary descriptions together if necessary.

There are limitations in Option 1 as we cannot use images, which are charts and visuals, which is often the case with a lot of documents. The reason is that multimodal embedding models can't often encode granular information like numbers in these visual images and compress them into meaningful embedding. Option 2 is severely limited because we do not end up using images at all in this system even if it might contain valuable information and it is not truly a multimodal RAG system.

For more information, visit this article



Advanced | Best of Tech | Generative AI | Guide | Large Language Models

A Comprehensive Guide to Building Multimodal RAG Systems

Build a Multimodal RAG system to handle text, images & tables, overcoming traditional RAG limita

Dipanjan (DJ) Sarkar 08 Jan, 2025