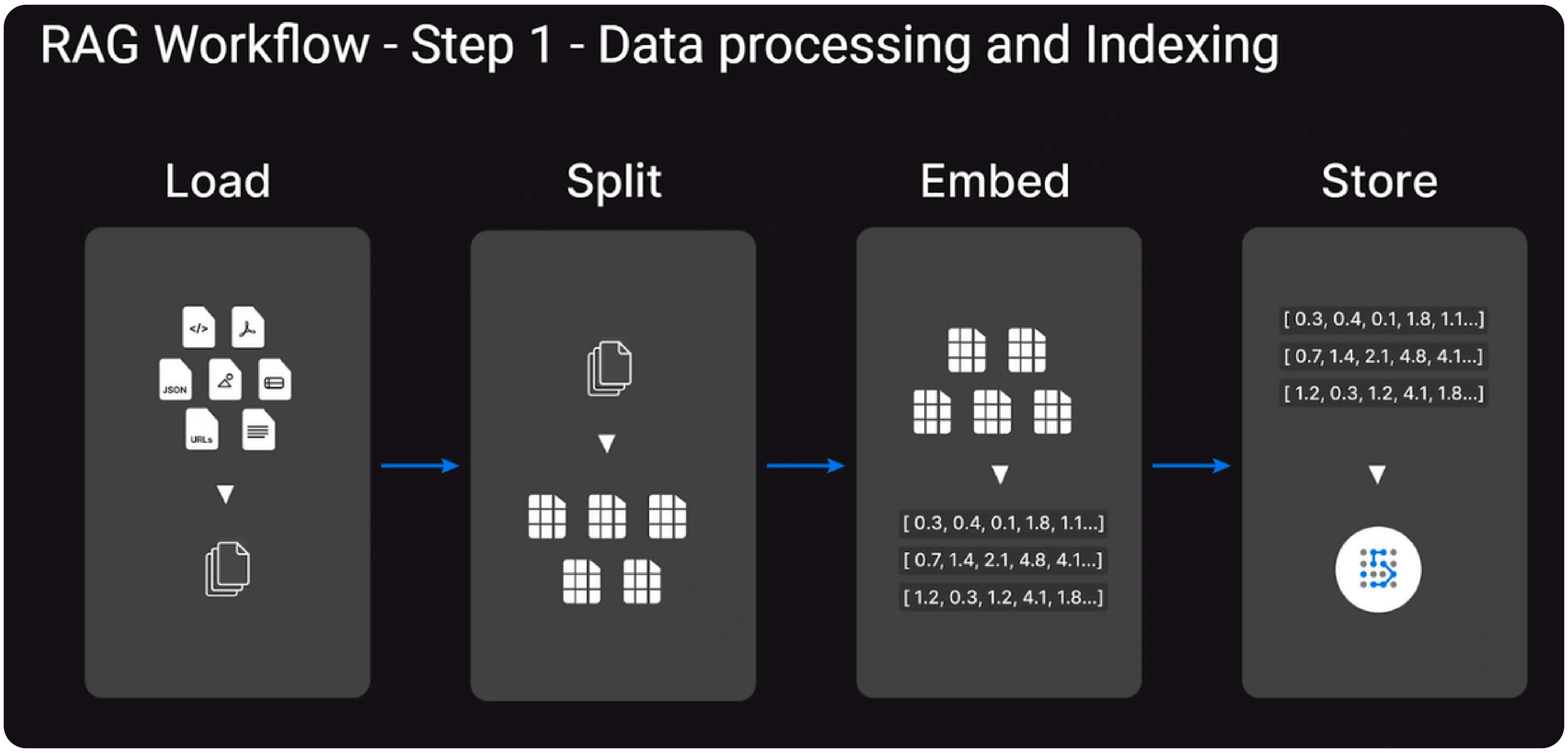
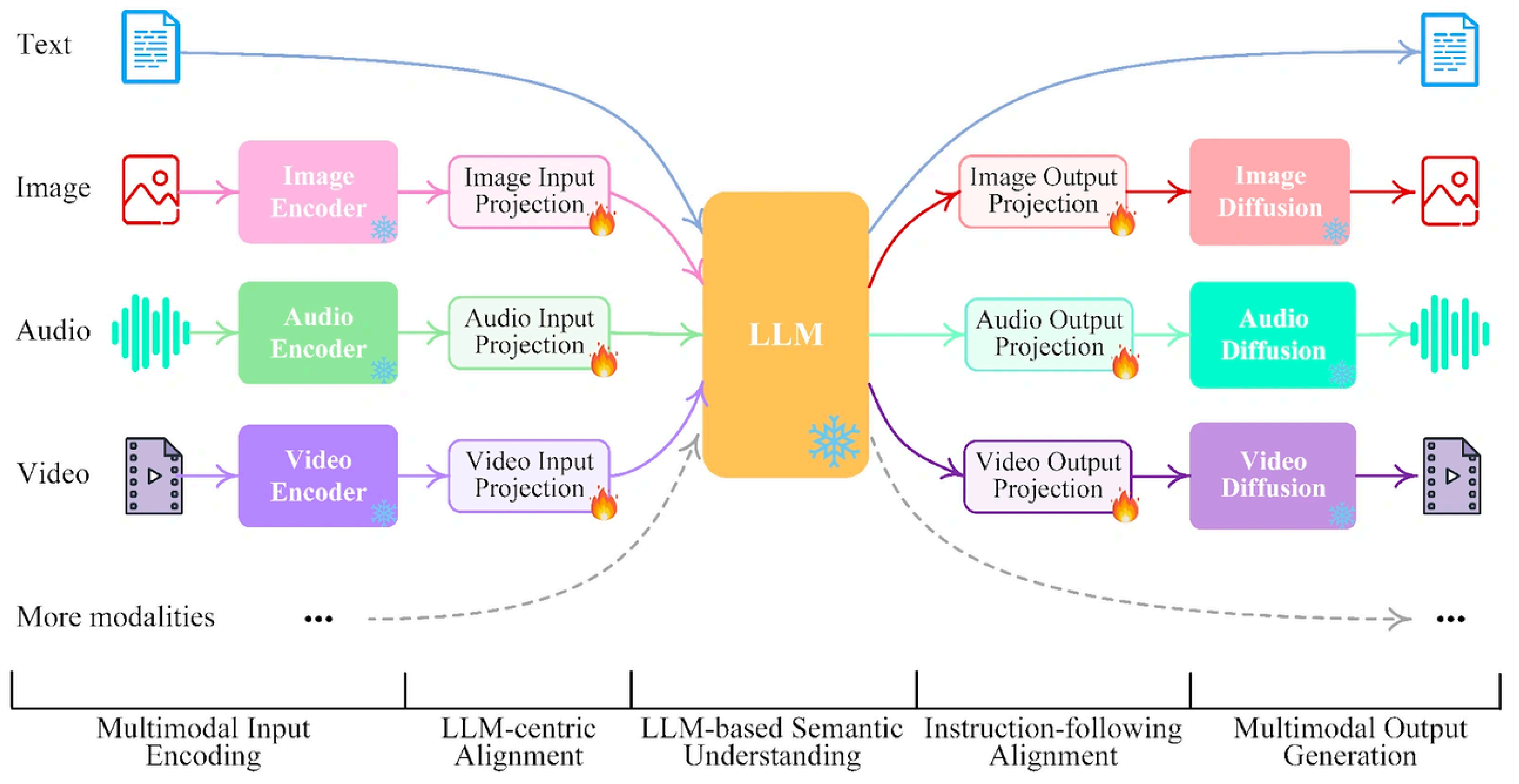
Analytics Vidhya

Mastering RAG

A Guide to Multimodal Models





So you've heard the buzzword "**multimodal**" thrown around, right? Maybe in the context of ChatGPT that can see images, or models that understand both text and video. But what does multimodal actually mean—and why is it such a big deal?

Let's break it down.

What Does "Multimodal" Even Mean?

In plain terms:

Multimodal AI = An AI system that can process and generate more than one type of data.

That includes stuff like:

- Text (what most LLMs do)
- Images
- Audio
- Video
- Structured data (e.g., tables, graphs)
- Sensor data (e.g., from robotics or IoT)

In contrast, **unimodal** models only handle one kind—like GPT-3, which just does text.

Why Multimodal Is a Game-Changer

Here’s why this matters: humans are multimodal by default. When you:

- Look at a map
- Listen to music
- Read captions on a meme
- Watch a video lecture with slides

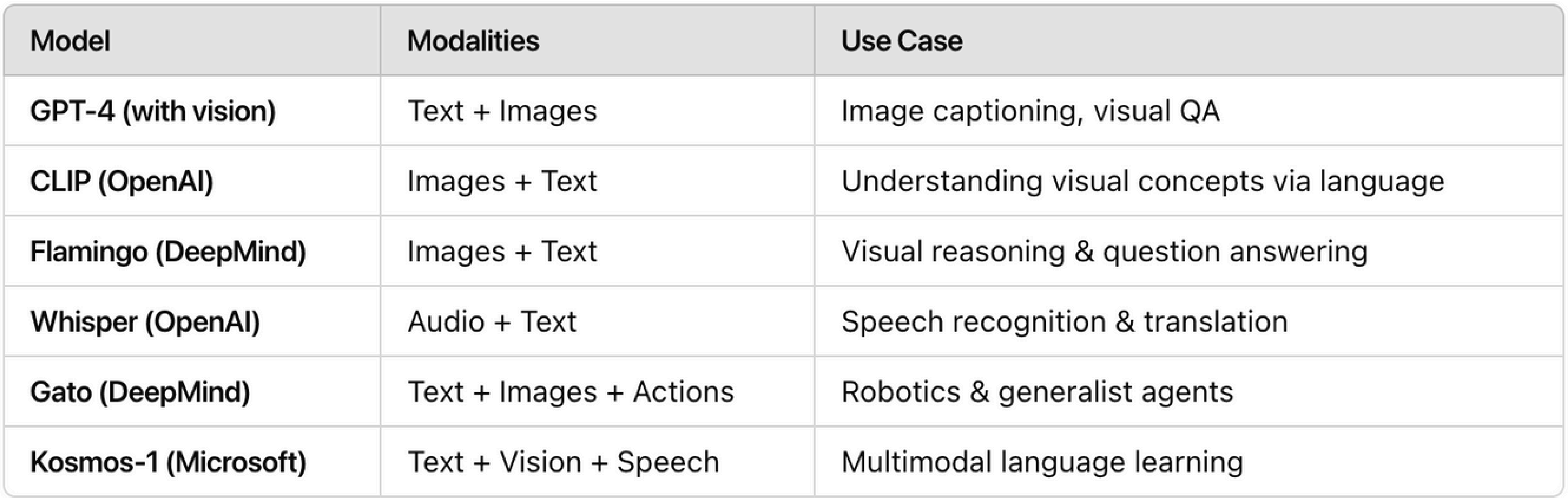
...you’re processing multiple streams of data at once.
If we want AI to act more like humans (or at least understand our world better), it needs to do the same.

Multimodal models unlock stuff like:

- Reading and describing images
- Answering questions about videos
- Translating speech into text (or vice versa)
- Interpreting graphs and generating reports
- Understanding human emotion through voice, expression, and words together

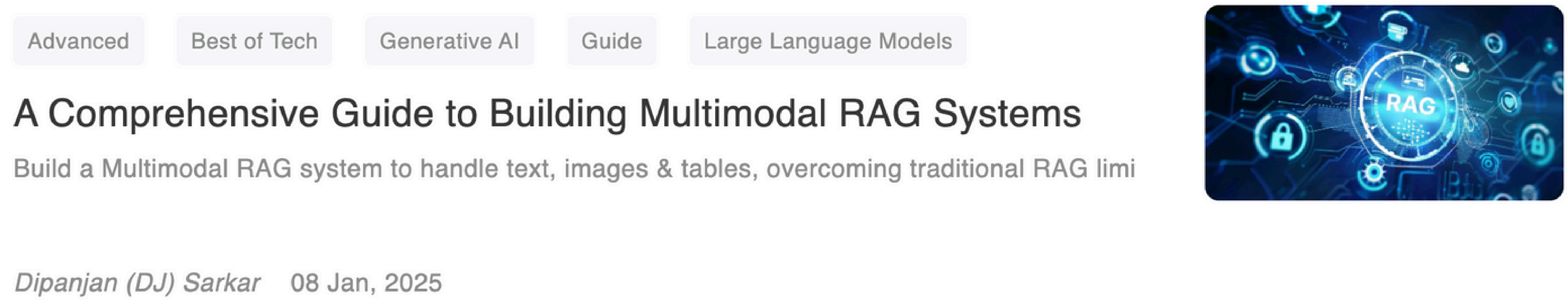
Examples of Multimodal Systems

Let’s get specific. Here are some high-profile multimodal systems:



Model	Modalities	Use Case
GPT-4 (with vision)	Text + Images	Image captioning, visual QA
CLIP (OpenAI)	Images + Text	Understanding visual concepts via language
Flamingo (DeepMind)	Images + Text	Visual reasoning & question answering
Whisper (OpenAI)	Audio + Text	Speech recognition & translation
Gato (DeepMind)	Text + Images + Actions	Robotics & generalist agents
Kosmos-1 (Microsoft)	Text + Vision + Speech	Multimodal language learning

For more deeper understanding, kindly 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 limi

Dipanjan (DJ) Sarkar 08 Jan, 2025