<epam>

Multimodal Retrieval-Augmented Generation (RAG) Implementation

OCTOBER 2024

<epam>

Ashwini Kumar
Lead Software Engineer

Vipul Gupta
Director, Technology Solutions

Sanchit Balchandani
Solution Architect

Contents

Why RAG?

- LLMs are limited to training data - often outdated
- No built-in fact-checking mechanism
- Can't access current information
- More cost-effective than retraining

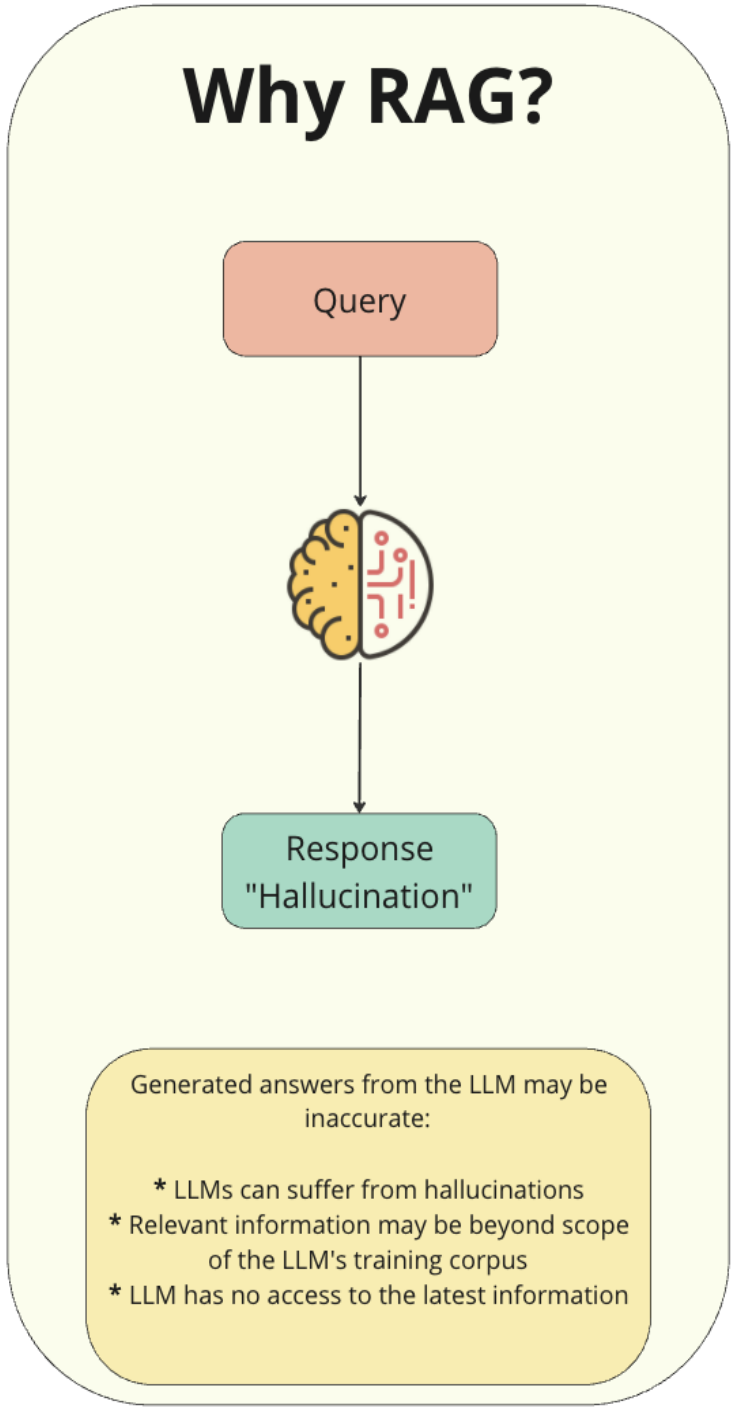


Image credits: https://miro.com/app/board/uXjVLbpeOd4=/

What is RAG?

RAG is a technique that combines Large Language Models with external knowledge retrieval to generate more accurate and contextual responses

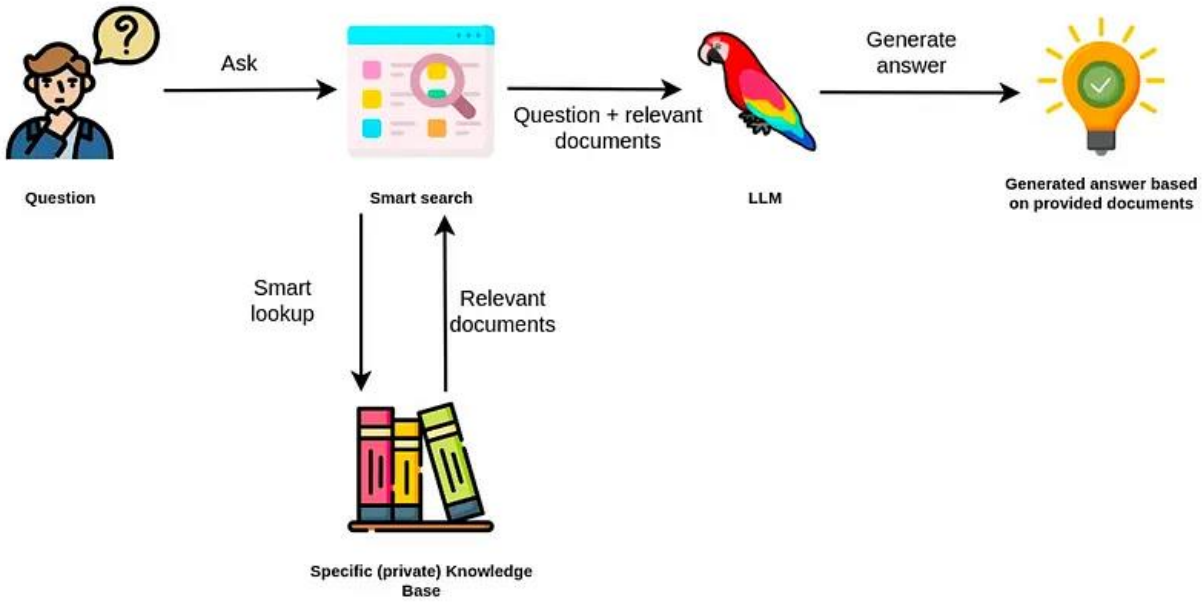
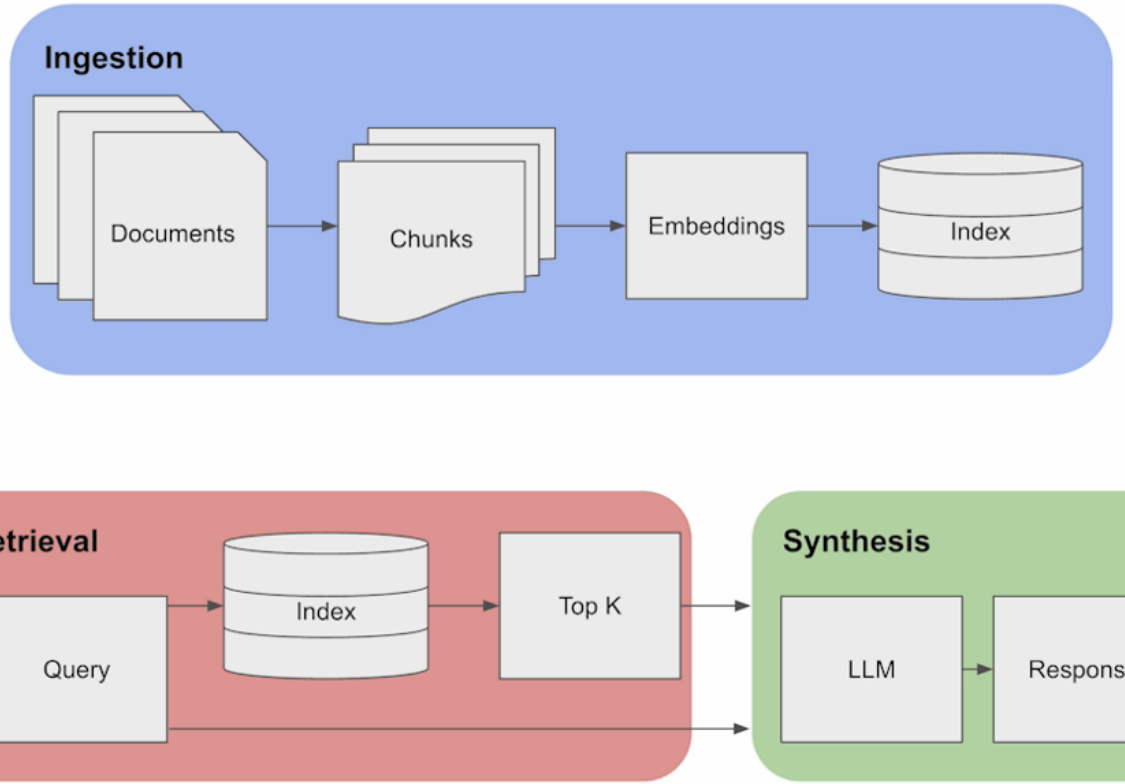


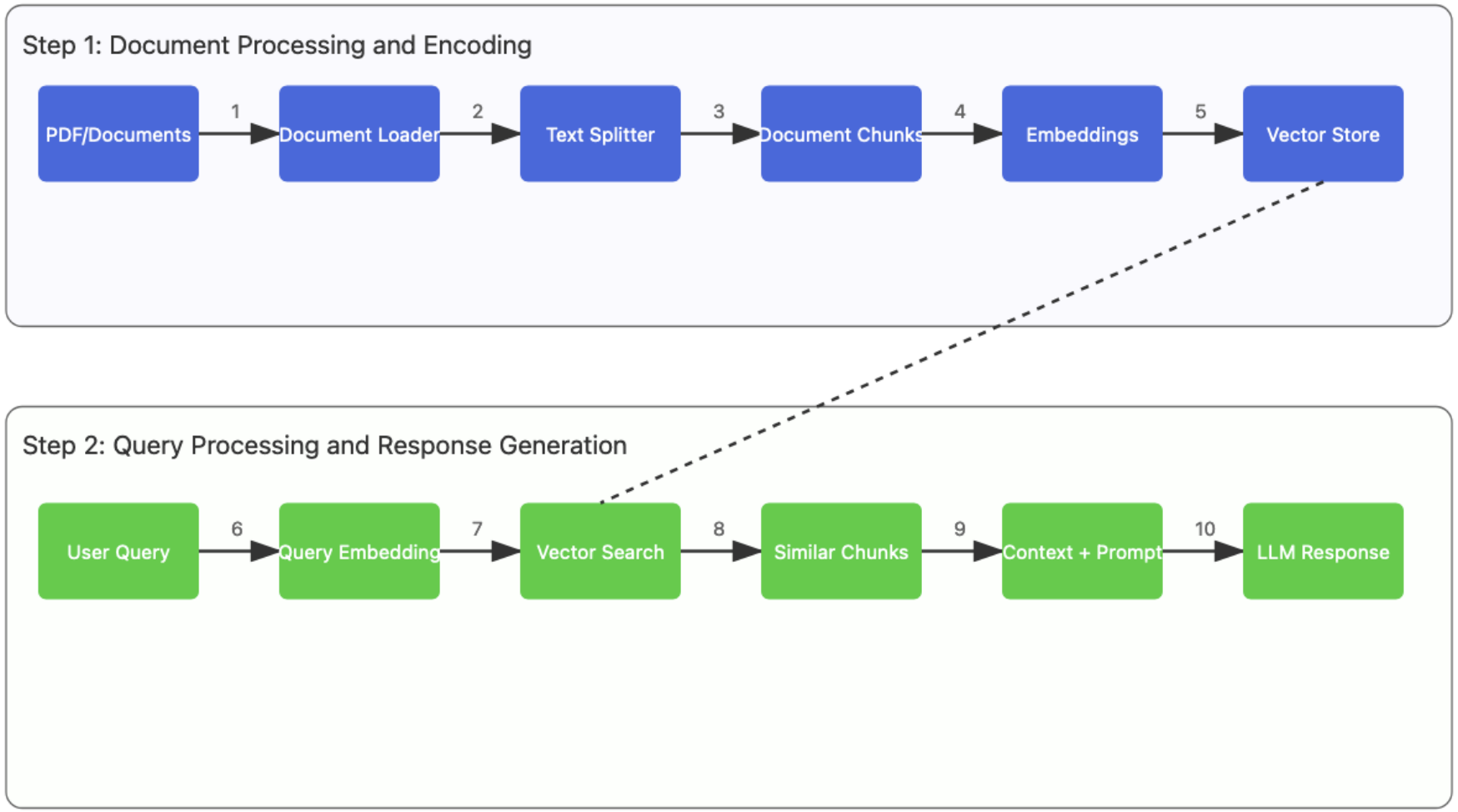
Image credits: https://neo4j.com/blog/what-is-retrieval-augmented-generation-rag/

RAG Key Components

- **Ingestion**: Document processing, embedding generation, and storage
- **Retrieval**: Semantic search and context selection
- **Synthesis**: Combining retrieved context with LLM capabilities

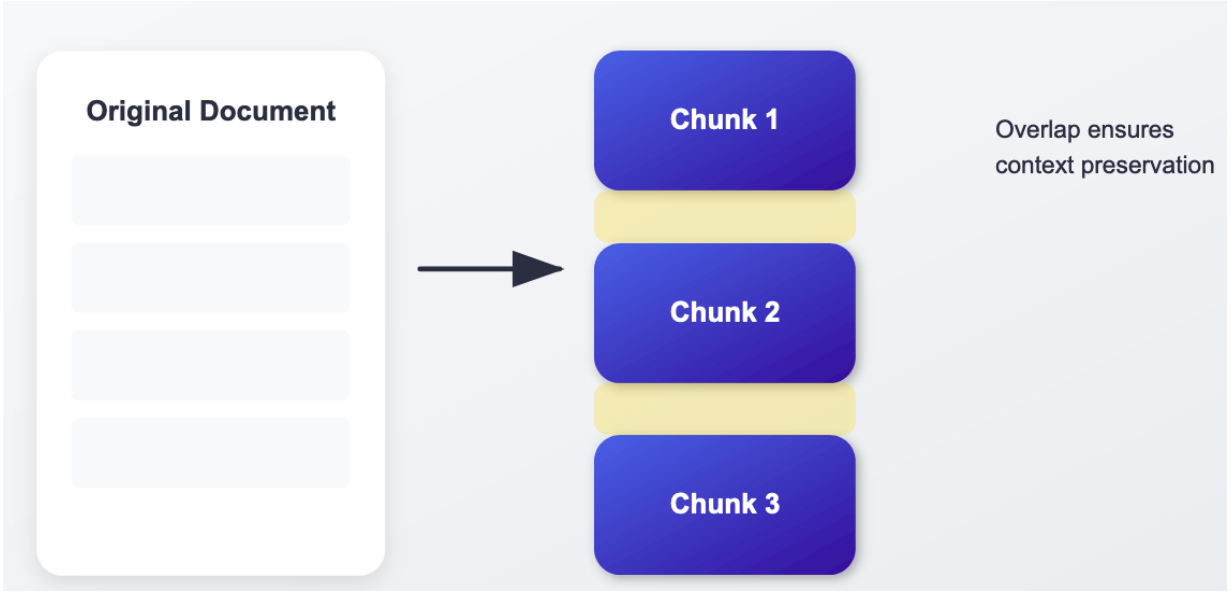


RAG Architecture: Detailed Flow



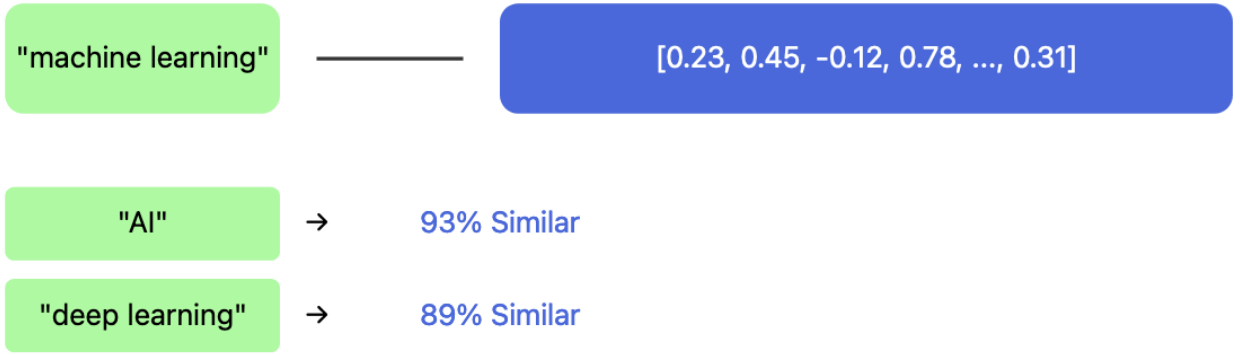
Document Chunking

- Splits large documents into manageable pieces
- Chunk size: 512-1024 tokens recommended
- 10-20% overlap prevents context loss
- Preserves semantic meaning
- Enables efficient processing & retrieval



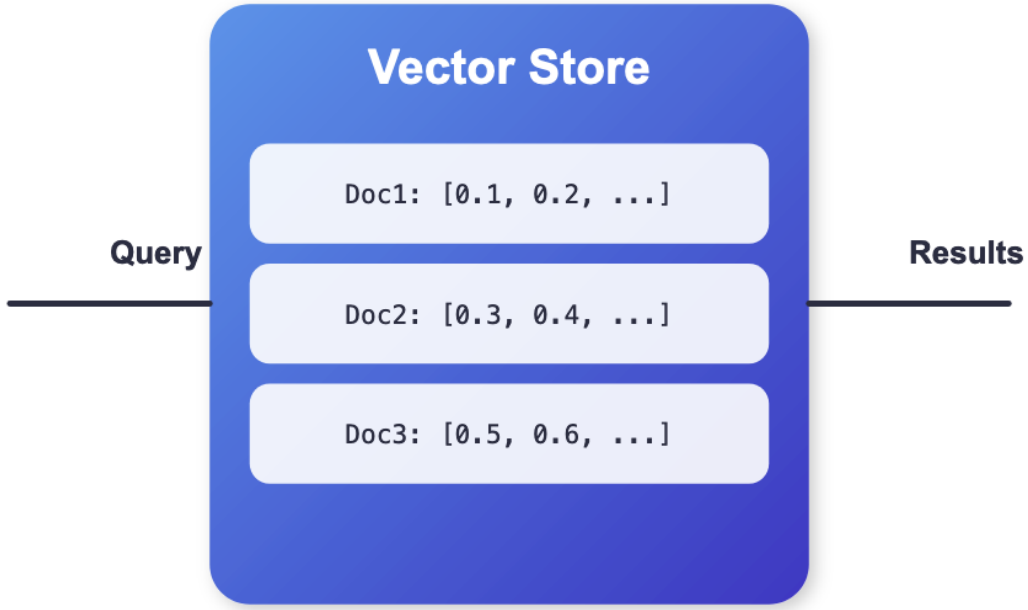
Text Embeddings

- Converts text to numerical vectors
- Captures semantic meaning
- Similar meanings = Similar vectors
- Powers semantic search



Vector Database

- Database for embedding vectors
- Enables similarity search
- Stores document metadata
- Fast retrieval capabilities
- Scalable architecture



Benefits of RAG

Enhanced Accuracy

- Reduces hallucinations
- Verifiable responses
- Fact-based answers

Dynamic Knowledge

- Up-to-date information
- Custom knowledge base
- Domain expertise

Operational Benefits

- Cost-effective scaling
- Easy maintenance
- Quick deployment

Hands On Demo Use Case

- **Purpose:** Developing a RAG-based application for EPAM Financial Data of year 2023. (https://s202.q4cdn.com/436759741/files/doc_financials/2023/q4/Exhibit99_Q4_2023-c.pdf).
- **Data Format:** The data is encapsulated in PDFs containing sophisticated mixes of text, images, and tables.

Demo

bit.ly/mm-rag

Limitations of Traditional RAG

- No access to real-time updates.
- The system is as good as the data you have in your vector database
- Traditional RAG only uses text data for both retrieval and generation phase
- Traditional models only uses text content to generate answer.
- Cannot process multiple types of data like images or tables.

What is Multimodal Data?

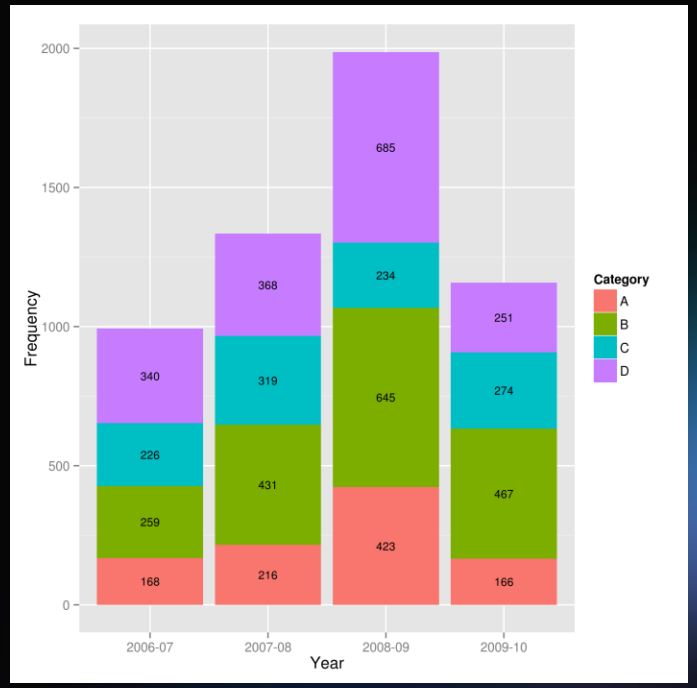
Multimodal data refers to data collected from various types of input and output channels between a human and a computer, encompassing different sensory and interaction modalities such as text, speech, vision, gestures, touch, audio, and biometrics. This data type allows for richer and more intuitive human-computer interaction by integrating diverse forms of information into a cohesive system.

Examples of Multimodal Data

TEXT

das Hammer-Amboßgelenk. Hermann macht hiergegen geltend, daß eine Unsymmetrie in den Schwingungsamplituden des Trommelfelles oder des Hammer-Amboßgelenkes erst bei sehr großen Amplituden eintreten könnte, während man die Kombinationstöne doch auch bei sehr geringer Intensität der Primärtöne vorzüglich höre; ferner höre man Kombinationstöne auch dann, wenn beide Ohren fest mit Watte verstopft sind, obwohl hierdurch gerade das Trommelfell in seinen Schwingungen stark behindert würde. Preyer[1]) gibt allerdings an, daß bei zugestopften Ohren Kombinationstöne nicht zu hören seien. Jedoch konnte ich sie in den meisten Fällen auch vorzüglich hören. Sie schienen mir beim Zustopfen der Ohren um so mehr geschwächt zu werden, je höher die Primärtöne sind. In einigen Fällen konnte ich hier Kombinationstöne überhaupt nicht hören, obwohl sie, wenn die Ohren nicht verstopft waren, ganz besonders stark zu hören waren. Vielleicht ist die sonst kaum verständliche Angabe Preyers damit zu erklären, daß er sehr hohe Primärtöne zu seinen Versuchen gewählt hat.

IMAGES



TABLES

Product Id	Weight	Size Range	Style	measure_1
310	15	60-62 CM	U	$472,743.87
311	13.77	42-46 CM	U	$395,360.20
312	14.13	48-52 CM	U	$474,150.84
313	14.42	48-52 CM	U	$424,906.69
314	14.68	54-58 CM	U	$415,057.86
320	19.79	54-58 CM	U	$5,433.09
321	19.79	54-58 CM	U	$16,295.59
322	19.9	60-62 CM	U	$4,861.18
323	19.9	60-62 CM	U	$10,666.20
324	20	60-62 CM	U	$4,575.23
325	20	60-62 CM	U	$17,480.72
326	18.77	42-46 CM	U	$6,290.94
327	18.77	42-46 CM	U	$14,814.17
328	19.13	48-52 CM	U	$7,434.75
329	19.13	48-52 CM	U	$18,369.57
330	19.42	48-52 CM	U	$5,719.04

What is Multimodal LLM?

A Multimodal Large Language Model (LLM) is an advanced AI model that processes and understands various types of data, such as text, images, tables, audio, and video. It uses a combination of encoding, reasoning, and generation stages to not only interpret mixed data formats but also generate similar outputs, enabling more comprehensive and context-aware interactions across different modalities.

Commercial Multimodal LLMs

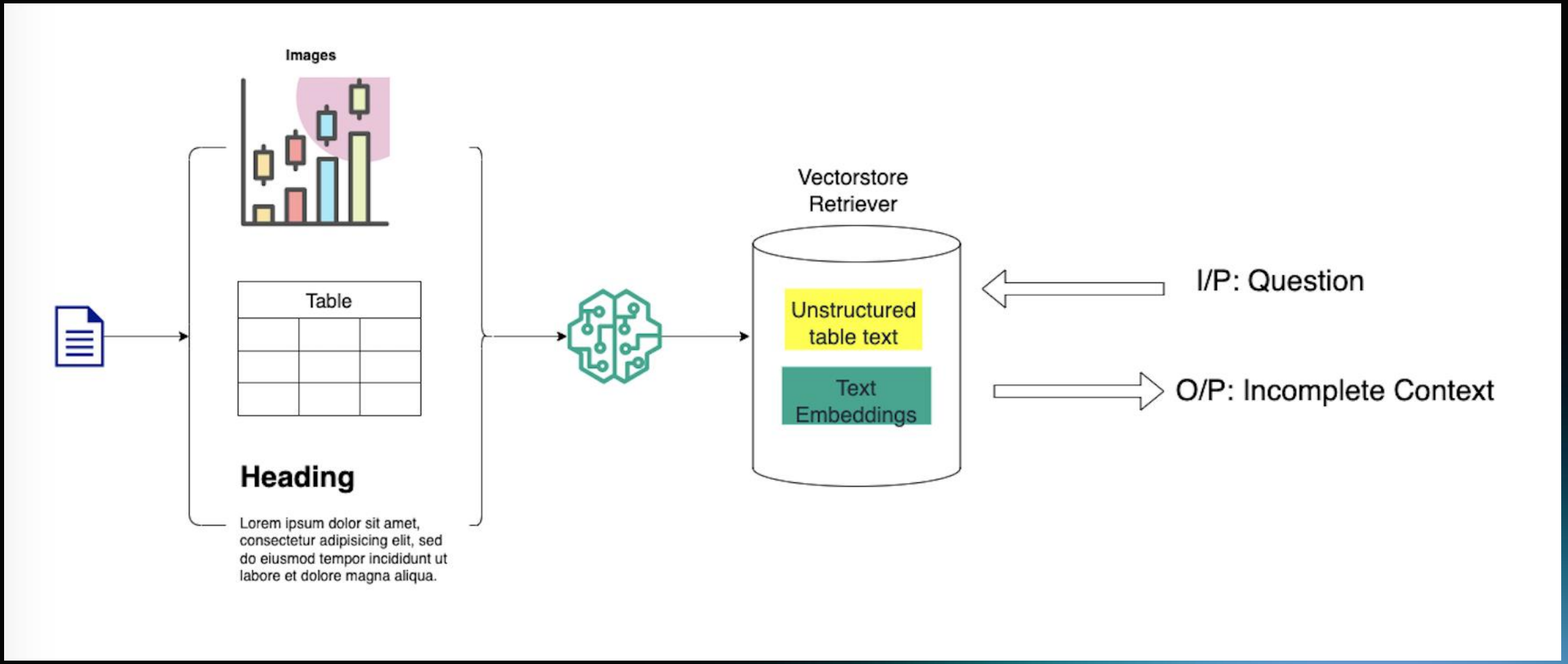
- GPT-4V & GPT-4o (OpenAI):
  - **Capabilities:** Understands text, images, and currently limited access to audio and video.
  - **Notes:** Audio and video analysis are not publicly available yet.
- Gemini (Google):
  - **Capabilities:** True multimodal understanding of text, audio, video, and images.
  - **Notes:** Robust integration across all media formats.
- Claude (Anthropic):
  - **Capabilities:** Handles text and image inputs.
  - **Notes:** Known for high performance in handling dual modalities.

Open-Source Multimodal LLMs

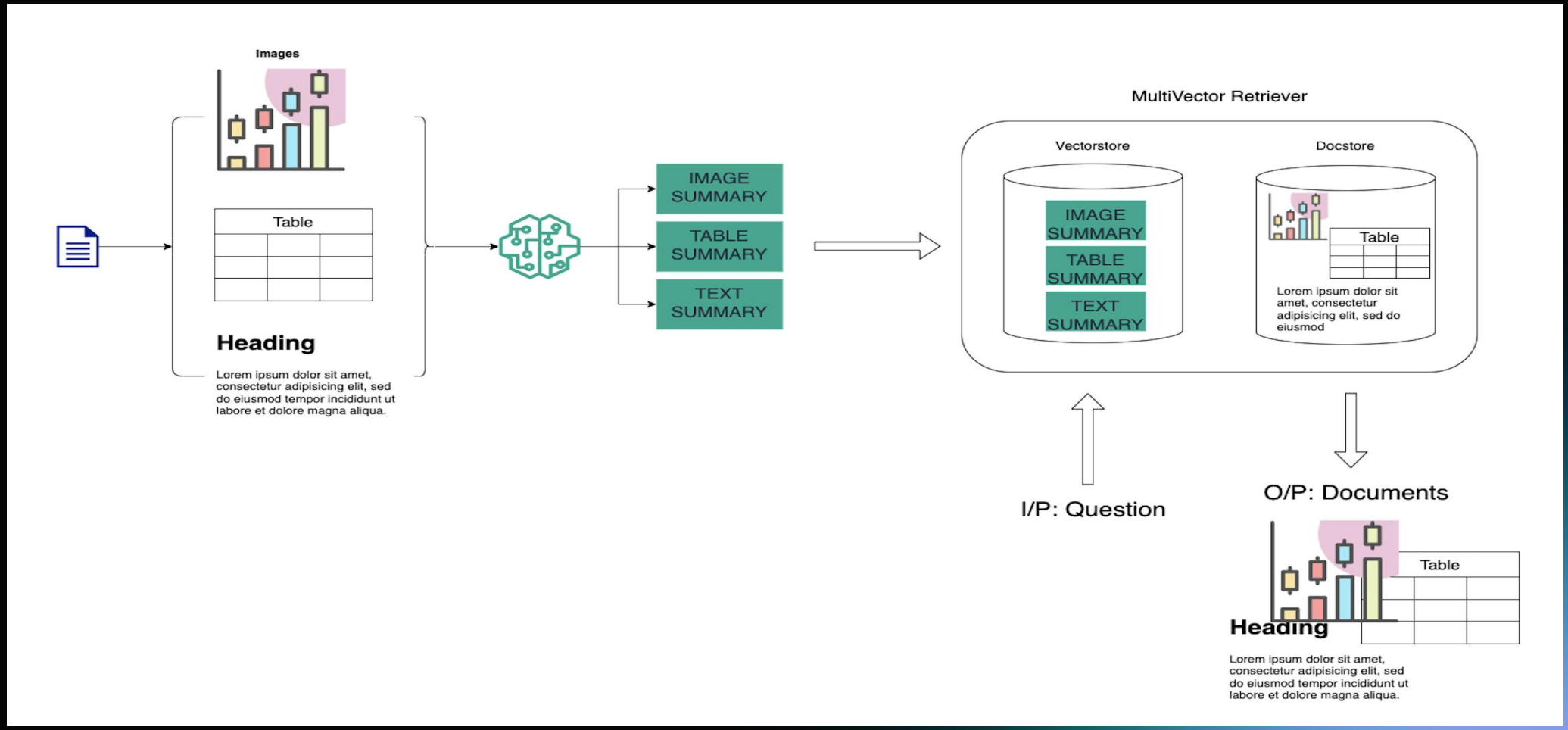
- LLaVA-NeXT:
  - **Capabilities: Works with text, images, and video.**
  - **Notes: An open-source improvement on the LLaVa model.**
- PaliGemma (Google):
  - **Capabilities: Integrates image and text processing, suitable for OCR, object detection, and VQA.**
  - **Notes: A vision-language model focusing on visual and textual interactions.**
- Pixtral 12B (Mistral AI):
  - **Capabilities: Processes images and text.**
  - **Notes: Built on Mistral's Nemo architecture, excels in image captioning and object recognition.**

Challenges

- Textual Data Processing: Traditional RAG systems are proficient in embedding text.
- Multimodal Data Hurdles: A significant struggle arises in interpreting and integrating visual (charts) and tabular data.
- Correctness of Data: Omitting image insights from context will generate incorrect or incomplete response


Images
Table
Heading
Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.
Vectorstore Retriever
Unstructured table text
Text Embeddings
I/P: Question
O/P: Incomplete Context

Multimodal Architecture



Hands-On

- **Part 1:** Data Loading
- **Part 2:** Summarising Multimodal Data
- **Part 3:** Create Multi-Vector Retriever
- **Part 4:** Integrate Chat Model
- **Part 5:** Demo & Testing

Conclusion & Key Takeaways

- **Traditional RAG Limitations**: Explored the challenges these systems face with multimodal data.
- **Understanding Multimodal Concepts**: Defined multimodal data and introduced Multimodal Large Language Models (LLMs).
- **Innovative System Architecture**: Discussed the detailed architecture and functionalities of a Multimodal RAG system utilizing GPT-4o.
- **Practical Application**: Implemented the system using LangChain, demonstrating its practical application with the EPAM financial data for the year 2023.

Q & A