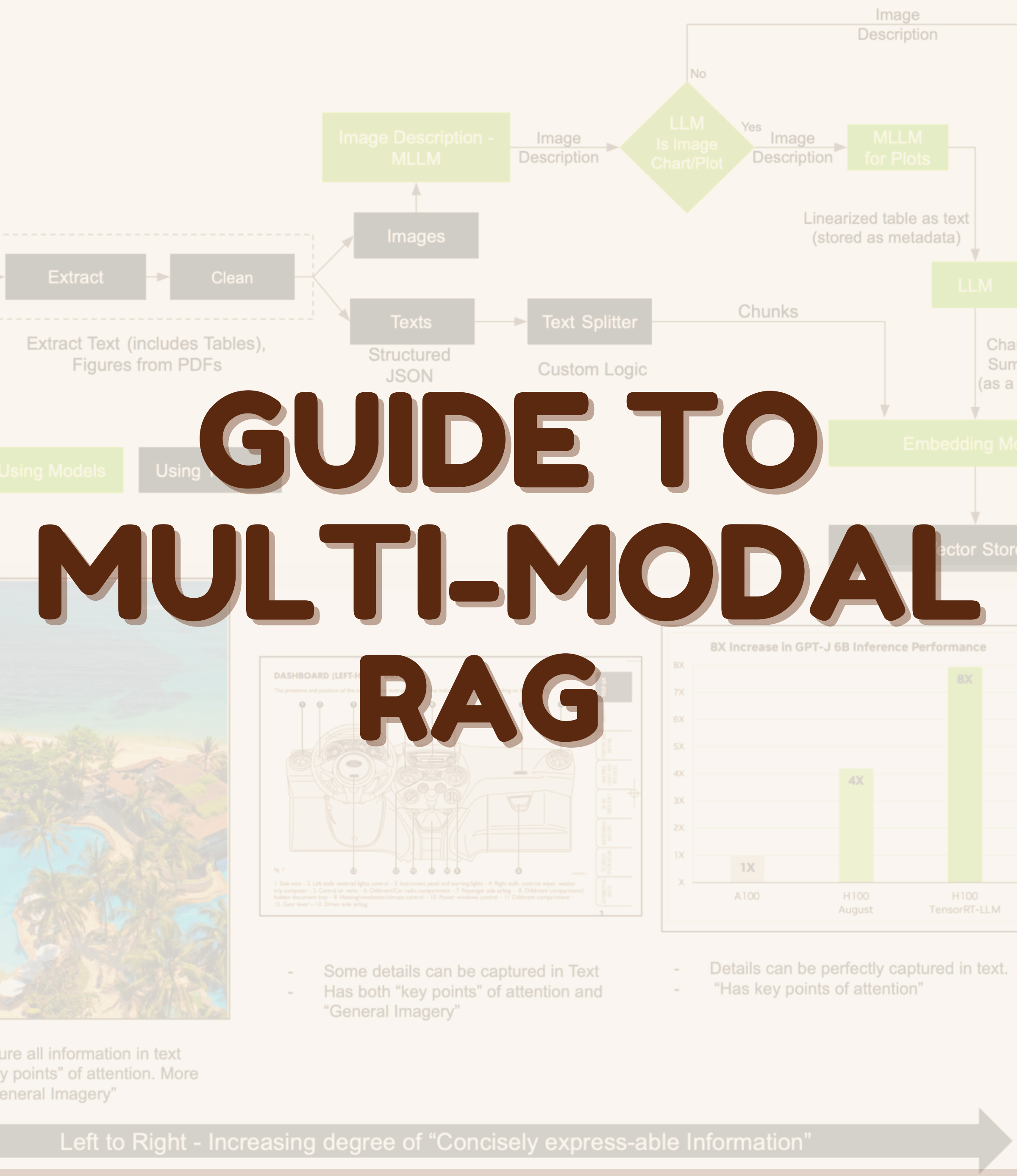
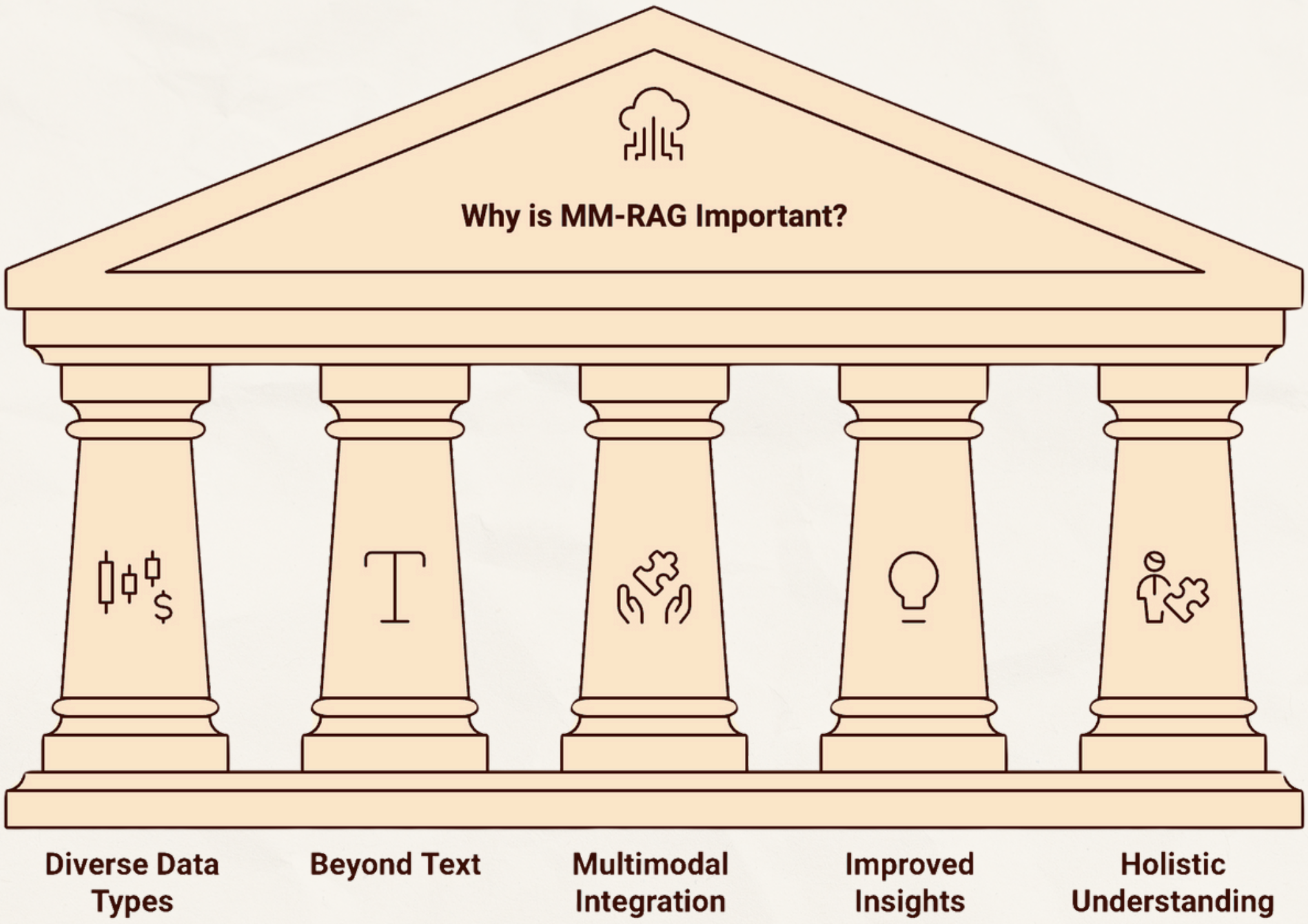
GUIDE TO MULTI-MODAL RAG



Bhavishya Pandit

INTRODUCTION

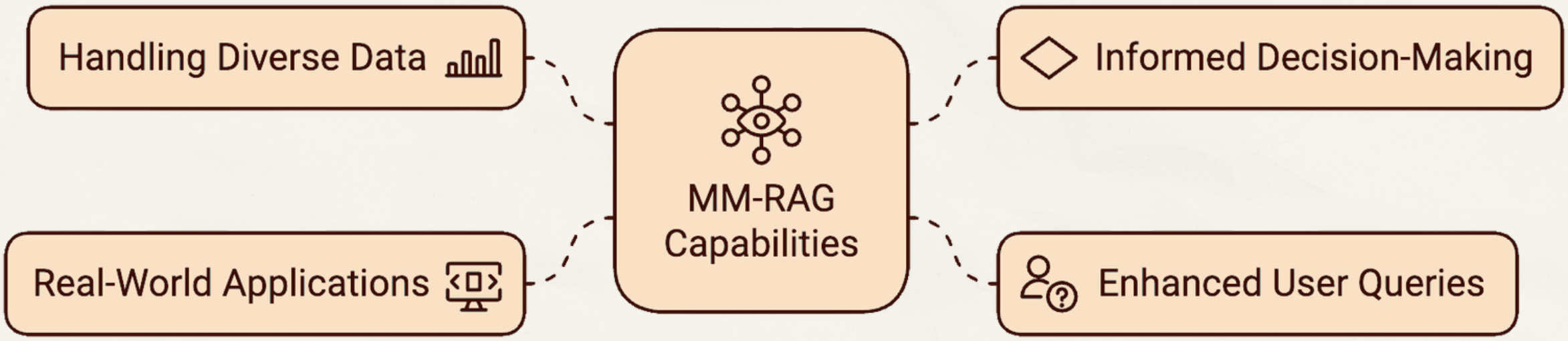


Multimodal Retrieval-Augmented Generation (MM-RAG) is a powerful AI system that enhances the standard Retrieval-Augmented Generation (RAG) framework by incorporating multiple types of data modalities—such as text, images, tables, charts, audio, and more. Instead of limiting the generation process to textual data, MM-RAG can retrieve and understand information from various formats, making it a versatile tool for more complex use cases.

MM-RAG = Text + Images = Better Results

WHY SHOULD YOU USE MM-RAG?

.



Handle Diverse Data

- Wide Range of Formats: processes images, text, diagrams, and audio
- Comprehensive Analysis: Analyzes various data types for a holistic view of queries

Informed and Contextual Decision-Making

- Nuanced Decisions: Multiple data sources lead to more informed choices
- Integrated Insights: Combines visual and textual cues for deeper understanding.

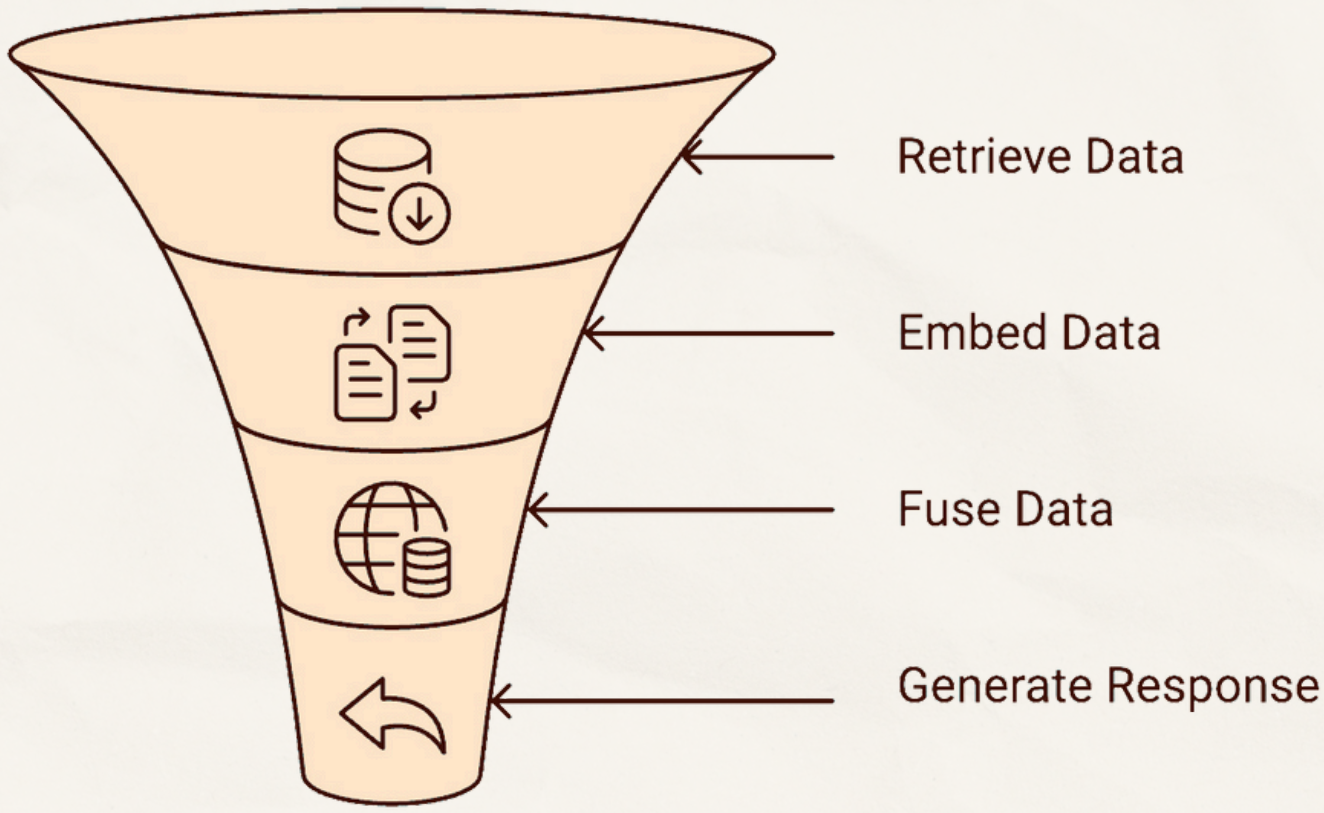
Enhanced User Queries

- Complex Query Handling: Answers intricate questions involving visual and textual information.
- Detailed Responses: Analyzes charts and descriptions to deliver integrated answers

Real-World Applications

- Versatile Usage: Applicable in fields like e-commerce, finance, healthcare, and legal
- Multi-Format Interpretation: Analyzes medical images with records and reviews legal documents with diagrams.

HOW MM-RAG WORKS



Step 1: Retrieving Multimodal Data

MM-RAG retrieves relevant information from textual and non-textual sources (images, diagrams, audio) using specialized retrieval models.

Step 2: Multimodal Embedding and Fusion

Different modalities are converted into a unified format. Images are turned into feature embeddings (using models like CLIP or LLaVA), while text is processed with LLMs. These embeddings are fused to create a cohesive multimodal context.

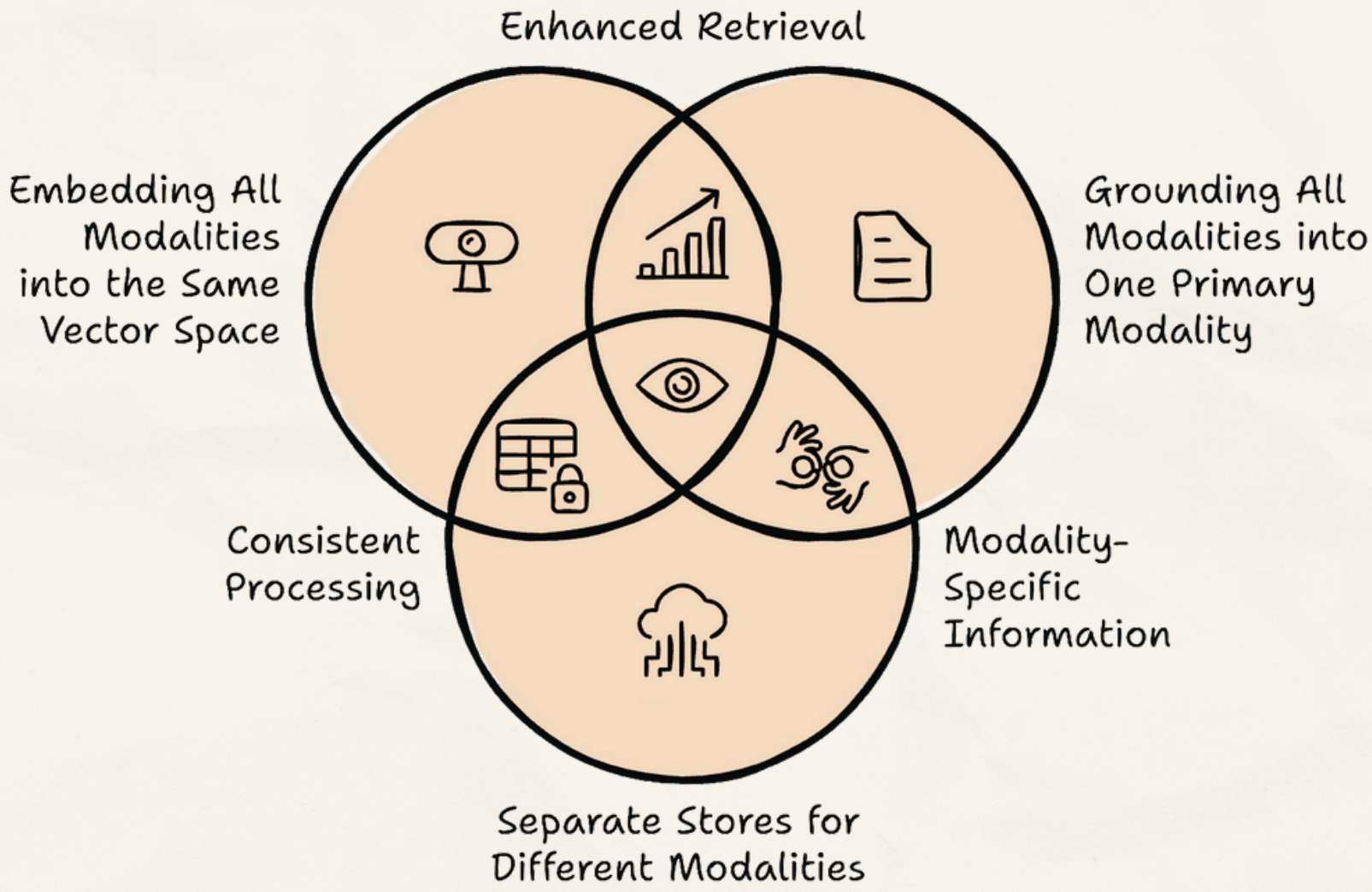
Step 3: Contextual Understanding and Response Generation

The multimodal context is processed through a generation model (e.g., GPT-4) to produce a response incorporating insights from all data formats.

Example Use Case

For a user query comparing two graphics cards, MM-RAG retrieves textual descriptions (specs, reviews) and visual data (performance charts), generating a comprehensive comparison that integrates both text-based and visual insights.

HANDLING MULTIPLE MODALITIES



Embedding All Modalities into the Same Vector Space

- Models like CLIP encode both images and text into a shared vector space, facilitating easier retrieval and generation across modalities.

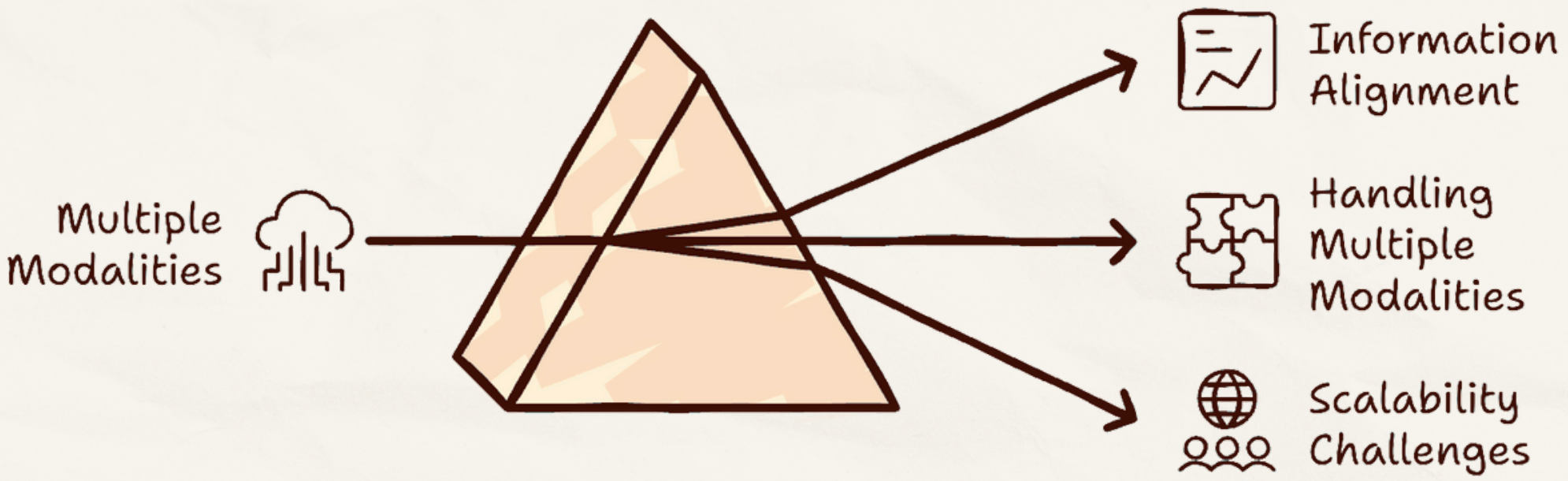
Grounding All Modalities into One Primary Modality

- This method selects a primary modality (often text) and converts other formats (like images) into text descriptions or tags, simplifying retrieval but potentially losing nuances of non-textual data.

Separate Stores for Different Modalities

- Different data types (text, images, charts) are stored in separate databases. A multimodal re-ranking system selects and combines relevant data from each modality, preserving modality-specific information and enhancing accuracy.

CHALLENGES OF MM-RAG



Handling Multiple Modalities

- Each modality (text, images, tables) has unique structures, requiring different processing techniques. For example, text is embedded using LLMs, while images are processed with models like CLIP. Creating a unified system without losing context is challenging.

Information Alignment Across Modalities

- Aligning insights from different formats (e.g., charts and related text) is difficult. MM-RAG must ensure data is accurately interpreted to avoid misalignment and inaccurate outputs.

Scalability and Resource Constraints

- Processing multiple modalities at scale is resource-intensive, as it involves handling large volumes of high-dimensional data, leading to increased computational costs and scalability concerns in production environments.

WHICH USE CASE DID YOU WORK ON USING MM-RAG?

Let me know in the comments

Follow to stay updated on Generative AI

SAVE

LIKE

SHARE