ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may2022

Multimodal RAG Based System to Handle Financial Documents

Rucha Dhage[1]; Dr. Manisha Bharati[2]

[1]Department of Technology SPPU, Pune, India
[2]Department of Technology, SPPU, Pune, India

Publication Date: 2026/06/15

Abstract: Traditional Retrieval-Augmented Generation (RAG) systems are very effective at querying text-based documents, but real-world documents are not just text based, they are complex and contain images, graphs, tables and more. Thus, traditional text only based RAG systems struggle to process such multimodal documents that contain more than just text effectively. This project presents the development of a Multimodal RAG system which is designed to bridge this gap. By Utilizing LangChain, HuggingFace embeddings, ChromaDB, and the LLaVA 1.5 Vision-Language Model (VLM), the system processes documents that contains not just text, but images and tabular data as well and extracts textual and visual elements such as images and graphs, and answers user queries based on both text and visual information. By indexing both textual and visual summaries into a unified vector space, the system retrieves multimodal context and gives accurate, grounded responses while reducing hallucinations related to chart colors and visual data trends.

Keywords: *Retrieval-Augmented Generation (RAG), LangChain, ChromaDB, HuggingFace Embeddings, LLaVA, Multimodal Retrieval.*

How to Cite: Rucha Dhage; Dr. Manisha Bharati (2026) Multimodal RAG Based System to Handle Financial Documents. *International Journal of Innovative Science and Research Technology*, 11(5), 4345-4351. https://doi.org/10.38124/ijisrt/26may2022

I. INTRODUCTION

Large Language Models (LLMs) have greatly improved how organizations search, understand, and use information through natural language. However, LLMs can sometimes give incorrect or outdated answers because they rely only on the information learned during training. To solve this problem, Retrieval-Augmented Generation (RAG) was introduced. RAG systems improve accuracy by searching external knowledge sources and providing relevant, real-time information to the LLM before generating a response.

At the same time, organizational data is not limited to plain text. Important information is often stored in different formats such as financial charts, diagrams, infographics, and tables. This makes modern enterprise data naturally multimodal, meaning it contains both text and visual information.

This research presents an end-to-end Multimodal Retrieval-Augmented Generation system designed to systematically overcome these limitations. By constructing a sophisticated, five-stage pipeline, this project bridges the semantic gap between textual and visual data.

Instead of ignoring images in documents, the proposed system uses a quantized Vision-Language Model to create detailed text descriptions of every visual element during the data ingestion process. These descriptions include important details such as chart axes, labels, and color information. The generated image summaries are then stored together with normal text and tables in a single vector database.

During response generation, the system retrieves all relevant text, table, and image information and combines them using carefully designed prompts. This ensures that the model generates answers only from verified visual and textual evidence, helping reduce incorrect or misleading visual interpretations (visual hallucinations).

The remainder of this thesis is structured as follows: Chapter 2 reviews the existing literature on RAG architectures and Vision-Language Models. Chapter 3 details the proposed system architecture and the structural schema of the pipeline. Chapter 4 provides an in-depth methodology, outlining the mathematical and algorithmic approaches to data preparation, vectorization, and multimodal synthesis. Chapter 5 discusses the results and Finally, Chapter 6 presents the conclusion and directions for future research.

II. LITERATURE REVIEW

Retrieval-Augmented Generation (RAG) has emerged as one of the most effective approaches for improving the factual accuracy and contextual grounding of Large Language Models (LLMs). Traditional LLMs often suffer

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may2022

from hallucination, limited contextual memory, and an inability to access external knowledge dynamically. To address these issues, researchers have increasingly integrated retrieval mechanisms with generative AI models. Recent studies demonstrate that RAG systems significantly improve information retrieval, domain-specific reasoning, and question-answering accuracy across multiple applications such as finance, legal analytics, healthcare, and academic research.

A. Darji et al. proposed a RAG-based financial risk analysis framework integrating GPT-4o, Gemini-1.5-flash, and Llama3.1 for automated audit report analysis. Their system evaluated metrics such as faithfulness, context precision, and answer relevance, with Llama3.1 achieving the highest retrieval and contextual performance. Similarly, S. Mehta et al. introduced "Finalyze," a RAG-based financial document analysis assistant integrating LangChain, FAISS, sentiment analysis, and real-time financial news retrieval. Their framework demonstrated high context relevance and rapid response times, highlighting the growing applicability of RAG systems in financial analytics.

Several researchers have explored the optimization of retrieval mechanisms within RAG architectures. M. Stäbler et al. investigated the impact of chunking strategies and embedding models on retrieval quality in domain-specific RAG systems. Their findings showed that sentence-based chunking with overlapping windows produced superior retrieval accuracy while maintaining computational efficiency. Likewise, A. Çağlayan et al. compared fine-tuned LLMs with code-augmented RAG pipelines for structured financial question answering. Their enhanced RAG architecture outperformed traditional fine-tuned models in factual reasoning and data-grounded responses.

In addition to retrieval optimization, multiple studies focused on reducing hallucinations in generative models. S. AboulEla et al. evaluated Naïve RAG, Graph RAG, and fine-tuned DistilBERT models to study hallucination reduction in knowledge-intensive tasks.

Their experiments revealed that while RAG improves contextual retrieval, fine-tuned models still achieved higher overall accuracy. However, Graph RAG demonstrated improved handling of complex queries through structured knowledge representation. Similarly, F. Yamout and H. A. Hasan proposed GPT-2++, a resource-efficient RAG framework integrating BERT and prompt engineering techniques. Their model substantially improved contextual accuracy and reduced hallucinations while operating under limited computational resources.

Researchers have also explored multimodal extensions of RAG systems. M. Barochiya et al. evaluated multimodal RAG pipelines using GPT-4o-mini and Gemini-1.0-Pro for question answering tasks involving both textual and visual information. Their findings demonstrated that multimodal retrieval significantly improves contextual accuracy compared to traditional text-only retrieval systems. Similarly, A. A. Rao et al. developed ContractIQ, a multimodal RAG-based legal contract analysis framework integrating Chain-of-Thought prompting and Gemini LLMs for clause extraction, compliance checking, and risk analysis.

Although existing literature demonstrates substantial progress in RAG-based architectures, most systems remain primarily text-centric and struggle to interpret visual elements such as graphs, charts, and embedded images within documents. Current approaches often convert only textual information into embeddings, resulting in the loss of important visual context. Furthermore, many systems lack strict grounding mechanisms during answer generation, which increases the risk of hallucinations in visually rich documents. To address these limitations, the proposed research introduces a Multimodal RAG framework capable of jointly processing text, tables, and images. By integrating the LLaVA Vision-Language Model with ChromaDB and semantic embeddings, the system transforms visual content into searchable textual summaries, enabling accurate multimodal retrieval and grounded response generation for complex financial documents.

III. SYSTEM ARCHITECTURE AND DESIGN

The proposed Multimodal Retrieval-Augmented Generation (RAG) system is designed to seamlessly process, index, and synthesize heterogeneous document modalities. Unlike traditional architectures that rely exclusively on textual data flows, this system implements a parallel processing pipeline that treats visual and textual elements as interconnected semantic entities. The overarching architecture is decoupled into five primary processing modules: Document Ingestion, Visual Semantic Translation, Vectorization and Storage, Context Retrieval, and Multimodal Synthesis.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may2022

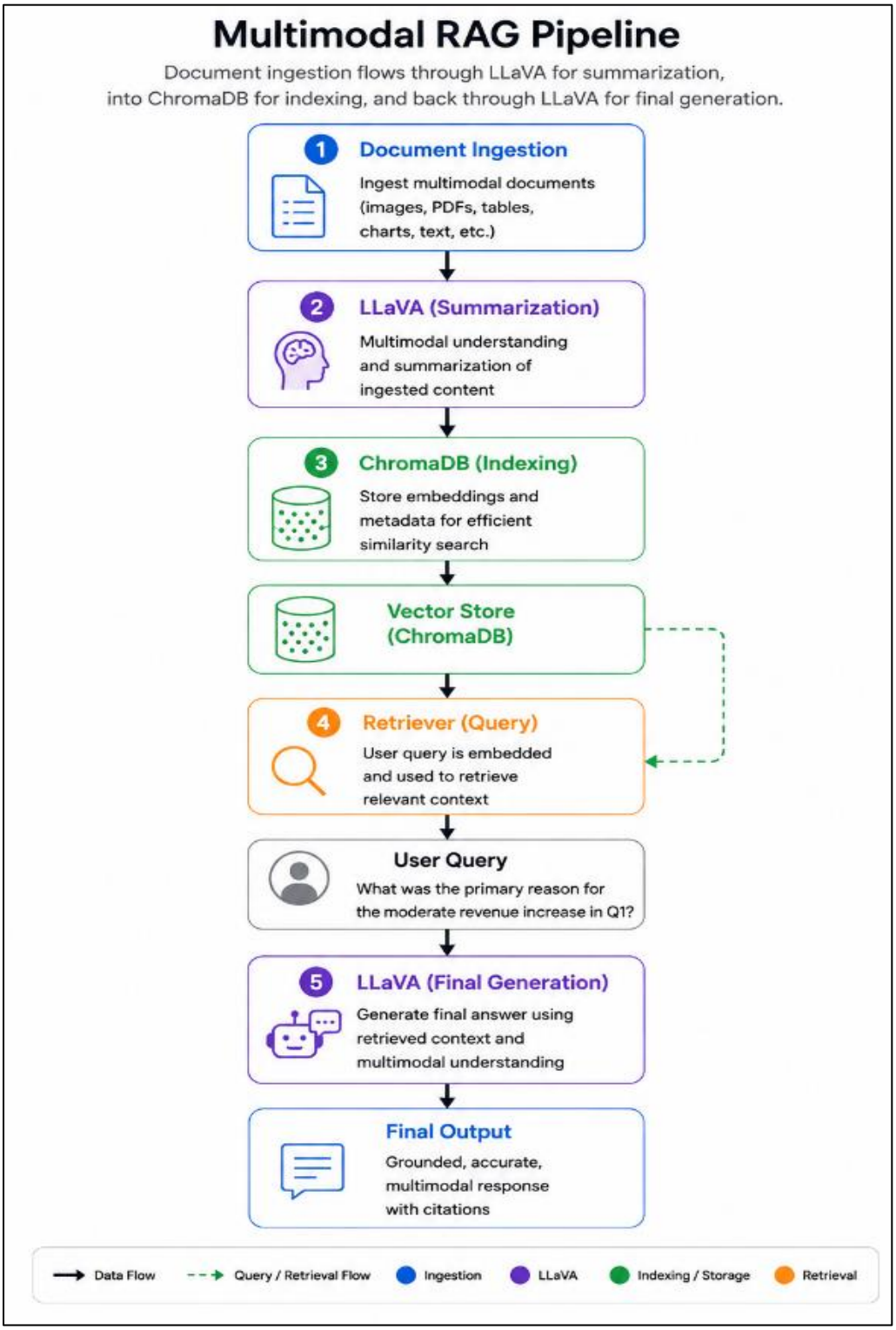

Fig 1 System Architecture

➢ *Document Ingestion*

The pipeline initiates with the ingestion of multimodal documents. This stage processes raw files—such as PDFs—and systematically extracts their diverse underlying components, including plain text, structured tables, visual charts, and embedded images. This comprehensive extraction ensures that all critical data modalities are captured for downstream semantic processing.

➢ *Multimodal Summarization (LLaVA)*

Following ingestion, the extracted components are routed into the Vision-Language Model, specifically LLaVA. In this phase, the system leverages LLaVA's multimodal understanding capabilities to comprehensively analyze the ingested content. The model generates high-fidelity textual summaries and representations of the visual and structural data, effectively translating non-textual elements into a format suitable for vectorization.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may2022

➢ *Indexing and Storage (ChromaDB)*

The generated representations and summaries are subsequently embedded and indexed within a ChromaDB Vector Store. This database acts as the central knowledge repository, storing the high-dimensional embeddings alongside their associated metadata. This vector storage infrastructure is explicitly optimized to facilitate highly efficient similarity searches.

➢ *Query Processing and Retrieval*

The active retrieval flow is triggered when a user submits a natural language question (e.g., *"What was the primary reason for the moderate revenue increase in Q1?"*). The user query is first mathematically embedded into the same vector space as the stored documents. The Retriever then utilizes this embedded query to perform a similarity search against the ChromaDB Vector Store, isolating and extracting the most relevant document chunks to form a highly targeted Retrieved Context.

➢ *Final Generation (LLaVA)*

In the concluding phase, the system routes the Retrieved Context back through the LLaVA model. The VLM synthesizes the retrieved information alongside the user's initial query, utilizing its multimodal understanding to construct a coherent answer. The final output is a grounded, accurate, and multimodal response that relies strictly on the retrieved context to ensure factual reliability.

IV. METHODOLOGY

➢ *Data Ingestion and Structural Parsing*

The foundational step of the proposed methodology involves transforming raw, multimodal PDF documents into a machine-readable format. To achieve this, the system leverages the unstructured library, which applies heuristic-based document partitioning.

Rather than extracting the document as a continuous string of characters, the parsing algorithm identifies logical boundaries using a by_title chunking strategy. The maximum character limit per chunk is strictly set to 4000, with a soft limit of 3800 to ensure context is not severed mid-sentence. During this partition phase, the algorithm explicitly categorizes elements into distinct data structures: CompositeElement (standard text paragraphs), Table (structured tabular data), and Image blocks. Text and tables are retained in active memory, while image blocks are extracted and saved as localized physical files for specialized downstream processing.

➢ *Visual Semantic Extraction via Vision-Language Models*

To incorporate visual data into the text-based vector space, the methodology implements a semantic translation process. The extracted raw images are passed to a 4-bit quantized LLaVA 1.5 (7B) Vision-Language Model.

This model is deployed using HuggingFace's BitsAndBytesConfig, configured with a torch.float16 compute data type to optimize GPU memory footprint during inference. A rigorously engineered extraction prompt is applied to each image, forcing the VLM to perform targeted optical and structural analysis. The model is constrained to explicitly identify and transcribe key chart attributes, including x/y-axis labels, data point values, legend keys, and exact bar or line colors. This methodological step converts previously inaccessible visual data into highly descriptive, mathematically indexable textual summaries.

➢ *Vectorization and High-Dimensional Embedding*

Once the document is parsed and the visual elements are semantically summarized, the data must be mapped into a unified vector space. The methodology utilizes the sentence-transformers/all-MiniLM-L6-v2 embedding model.

This model maps the text chunks, tabular structures, and VLM-generated visual summaries into a 384-dimensional dense vector space. The embedding process captures the deep semantic meaning of the inputs, ensuring that a natural language query regarding a visual trend is mapped into the same dimensional neighborhood as the VLM's textual description of that trend.

➢ *Database Indexing and Similarity Search*

The generated dense vectors are indexed into ChromaDB, an AI-native vector database optimized for multidimensional similarity search. To preserve the relationship between the vectors and their original modalities, an InMemoryStore is utilized as a relational mapping system, linking each generated vector UUID to its raw text payload or physical image file path.

Upon receiving a natural language user query, the system vectorizes the query and performs a k-nearest neighbors search ($k=3$) to retrieve the most contextually relevant document chunks. The retrieval mechanism relies on Cosine Similarity to calculate the distance between the query vector and the document vectors.

The cosine similarity between a query vector A and a document vector B is mathematically defined as the dot product of the vectors divided by the product of their magnitudes:

$$\text{Cosine\_Similarity}(A, B) = \frac{A \cdot B}{\|A\|\|B\|}$$

Vectors with a smaller angular distance (yielding a cosine similarity closer to 1) are ranked highest and retrieved as the context for the final generation phase.

➢ *Retrieval-Augmented Synthesis*

The final methodological step is the multimodal synthesis of the retrieved data. The system dynamically evaluates the metadata of the top k retrieved vectors. If the retrieved vector corresponds to text or a table, the raw string is extracted. Crucially, if the retrieved vector corresponds to an image summary, the system uses the In Memory Store to fetch the original raw image file rather than just the text summary.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may2022

V. RESULTS & DISCUSSION

The implemented Multimodal Retrieval-Augmented Generation (RAG) system success fully processed complex documents containing text, tables, and images, and answered detailed user queries in real time. The system accurately retrieved information from both textual and visual content, allowing it to handle different types of enterprise documents effectively.

To improve accuracy, a two-stage prompt engineering approach was used. First, detailed summaries of images and charts were created during document processing. Then, strict instructions were applied during answer generation so the model only used the retrieved evidence. This helped reduce Vision-Language Model (VLM) hallucinations and enabled the system to correctly identify details such as chart colors, labels, and data points instead of making random guesses.

The system also performed well for text-based queries by accurately retrieving the most relevant text sections from the documents. Finally, integrating the backend pipeline with a Gradio web interface transformed the project into an interactive chatbot, demonstrating that the system is practical, user-friendly, and suitable for real-time deployment.

- *Interactive Chatbot Interface*

The developed Multimodal RAG system was deployed using a Gradio-based web interface to enable real-time interaction between the user and the retrieval pipeline. The interface allows users to ask natural language questions related to textual content, tables, charts, and graphical elements present inside uploaded PDF documents.

The chatbot interface integrates the retrieval and generation pipeline into a conversa tional environment where retrieved multimodal context is passed to the Vision-Language Model for response generation. Example prompts include financial trend analysis and chart interpretation tasks, demonstrating the system's ability to process both textual and visual information simultaneously.

- *Text-Based Query Result*

The first experiment evaluated the system's ability to retrieve and answer information grounded in textual context. The user asked the system: "What was the primary reason for the moderate revenue increase in Q1?" The RAGpipeline successfully retrieved the relevant text chunk from the document and generated the correct response indicating that the increase was caused by the introduction of new product lines.

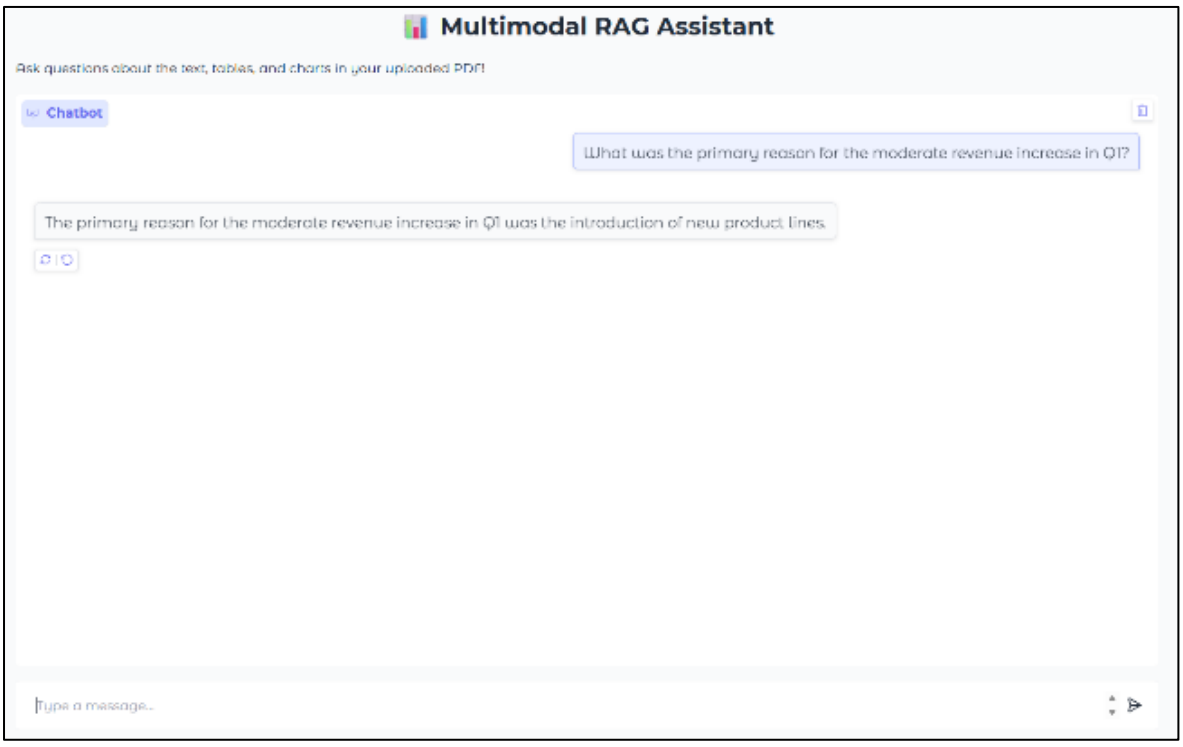

Fig 2 Text Based Answer

This result demonstrates that the semantic retrieval mechanism correctly identified and retrieved the appropriate textual evidence before passing it to the Vision-Language Model for grounded answer generation.

- *Image-Based Query Result*

The second experiment tested the multimodal capability of the system using a visual reasoning query. The user asked: "What color does the bar for Q2 represent?" The system retrieved the relevant chart image, analyzed the visual content through the LLaVA Vision-Language Model, and correctly identified the Q2 bar as green.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may2022

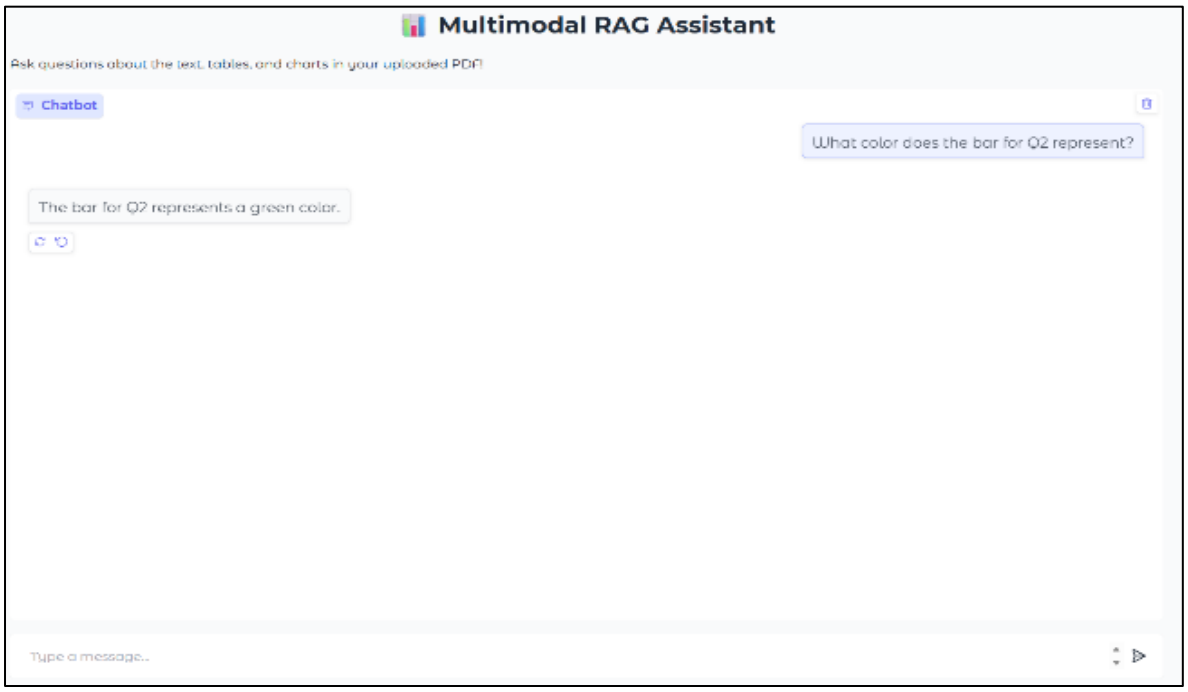

Fig 3 Image Based Result

This experiment validates the core contribution of the proposed Multimodal RAG architecture. Unlike traditional text-only RAG systems, the developed framework successfully interprets visual information such as chart colors and graphical elements. The result also demonstrates the effectiveness of prompt engineering strategies used to reduce hallucinations and enforce visual grounding during answer generation.

VI. CONCLUSION & FUTURE WORK

This project successfully demonstrated the design and implementation of a Multimodal Retrieval-Augmented Generation (RAG) system capable of simultaneously processing and retrieving information from diverse document modalities, including plain text, structured tables, and embedded images. By integrating the LLaVA 1.5 vision-language model with a ChromaDB vector store and the all-MiniLM-L6-v2 sentence embedding model, the system bridges the longstanding gap between textual and visual understanding in document question-answering tasks.

The pipeline follows a coherent and reproducible architecture: raw PDF documents are parsed and decomposed into typed elements using the Unstructured library, images are semantically described with LLaVA-powered summarization, and all content is unified into a single searchable vector index. At inference time, semantically relevant chunks — whether text, table data, or image descriptions — are retrieved and passed back to the vision-language model alongside the original images, enabling grounded and contextually accurate responses.

A key contribution of this work is the application of strict prompt engineering strategies that constrain the model to visually verify chart elements such as axis labels, bar colors, and data values before responding. This approach measurably reduces hallucination — a critical limitation of large language models operating on ambiguous or visually rich content. The resulting system demonstrates that carefully designed retrieval pipelines, when combined with multimodal generative models, can produce reliable, explainable answers from complex, real-world documents without requiring fine-tuning or additional training.

Overall, this project establishes a strong and extensible foundation for intelligent document understanding, with demonstrated capability across the three core modalities present in most professional and academic documents.

The present system handles text, tables, and static images. Future iterations could incorporate video and audio retrieval. LLaVA 1.5 (7B parameters) was selected for its efficiency under constrained GPU resources. Replacing or augmenting it with larger models such as LLaVA-1.6, InternVL2, or GPT-4o would improve visual reasoning accuracy, particularly for complex scientific figures, multi-series charts, and dense infographics where fine-grained color and spatial understanding is essential.

REFERENCES

[1]. Darji, F. Kheni, D. Chodvadia, P. Goel, D. Garg and B. Patel, “Enhancing Financial Risk Analysis using RAG-based Large Language Models,” in 2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS), Pudukkottai, India, 2024, pp. 754–760.

[2]. N. Chinaksorn and D. Wanvarie, “LLM-RAG for Financial Question Answering: A Case Study from SET50,” in 2025 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Fukuoka, Japan, 2025, pp. 952–957.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may2022

[3]. K. Kocot, M. Płonka, K. Hołda, K. Daniec and A. Nawrat, "Advanced Document Processing Using LLM and RAG: AnInnovative Approach to Efficiency and Privacy," in 2025 5th Intelligent Cybersecurity Conference (ICSC), Tampa, FL, USA, 2025, pp. 227–231.

[4]. T. Mitadera et al., "RAG based Question Answering of Accounting Knowledge," in 2025 IEEE 14th Global Conference on Consumer Electronics (GCCE), Osaka, Japan, 2025, pp. 1456–1460.

[5]. M. Stäbler, S. Turnbull, T. Müller, C. Langdon, J. Marx-Goméz and F. Köster, "The Impact of Chunking Strategies on Domain-Specific Information Retrieval in RAG Systems," in 2025 IEEE International Conference on Omni-layer Intelligent Systems (COINS), Madison, WI, USA, 2025, pp. 1–6.

[6]. A. Çağlayan, S. N. Gökçe and D. Ayata, "Structured Financial QA with LLMs: Fine Tuning vs. Code-Augmented Retrieval," in 2025 10th International Conference on Computer Science and Engineering (UBMK), Istanbul, Turkiye, 2025, pp. 539–544.

[7]. S. Mehta, T. Negandhi and S. Ghane, "Finalyze: A RAG-Based Framework for Intelligent Financial Document Analysis," in 2025 5th International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT), MANDYA, India, 2025, pp. 1–5.

[8]. J. Xu, "Enhancing Financial Risk Management with Retrieval-Augmented Large Language Models," in 2025 4th International Conference on Artificial Intelligence, Internet and Digital Economy (ICAID), Guangzhou, China, 2025, pp. 138–141.

[9]. Z. Huang, K. Du, X. Zhang, R. Mao and E. Cambria, "Combining LLM-Generated Knowledge Graphs with RAG for Financial Sentiment Extraction," in 2025 IEEE International Conference on Data Mining Workshops (ICDMW), Washington, DC, USA, 2025, pp. 2056–2063.

[10]. S. AboulEla, P. Zabihitari, N. Ibrahim, M. Afshar and R. Kashef, "Exploring RAG Solutions to Reduce Hallucinations in LLMs," in 2025 IEEE International Systems Conference (SysCon), Montreal, QC, Canada, 2025, pp. 1–8.

[11]. M. Barochiya, P. Makhijani, H. N. Patel, P. Goel and B. Patel, "Evaluating RAG Pipeline in Multimodal LLM-based Question Answering Systems," in 2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS), Pudukkottai, India, 2024, pp. 69–75.

[12]. F. Yamout and H. A. Hasan, "GPT-2++: An Optimized GPT-2 for RAG by Integrating BERT, Prompt Engineering, and Fine-Tuning," in 2025 3rd International Conference on Foundation and Large Language Models (FLLM), Vienna, Austria, 2025, pp. 34–37.

[13]. A. A. Rao, A. R. Revankar, N. Nair, S. Mittal, U. D and U. M. Kumar, "ContractIQ : A Multimodal RAG-Based Agentic System for Intelligent Contract Understanding," in TENCON 2025- 2025 IEEE Region 10 Conference (TENCON), Kota Kinabalu, Malaysia, 2025, pp. 1698–1702.

[14]. O. Keleş and T. Bayraklı, "Llama-2-econ: Enhancing title generation, abstract classification, and academic Q&A in economic research," 2024, accessed Apr. 17, 2025.

[15]. M. Thomas, S. Khot, M. Bhole and N. Shaji, "InvestMate: An Integrated AI-Driven Framework for Personalized Financial Planning and Real-Time Market Analysis," in 2025 International Conference on Computing, Intelligence, and Application (CIACON), Durgapur, India, 2025, pp. 1–7.