Evaluating RAG Pipeline in Multimodal LLM-based Question Answering Systems

2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS) | 979-8-3315-3242-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICACRS62842.2024.10841620

Madhuri Barochiya
Computer Science & Engineering Department
Devang Patel Institute of Advance Technology and Research, Charotar University of Science and Technology (CHARUSAT)
Changa, Anand, India
22dcs005@charusat.edu.in

Pratishtha Makhijani
Computer Science & Engineering Department
Devang Patel Institute of Advance Technology and Research, Charotar University of Science and Technology (CHARUSAT)
Changa, Anand, India
22dcs039@charusat.edu.in

Hetul Niteshbhai Patel
Computer Science & Engineering Department
Devang Patel Institute of Advance Technology and Research, Charotar University of Science and Technology (CHARUSAT)
Changa, Anand, India
hetulpatel1516@gmail.com

Parth Goel
Computer Science & Engineering Department
Devang Patel Institute of Advance Technology and Research, Charotar University of Science and Technology (CHARUSAT)
Changa, Anand, India
parthgoel.ce@charusat.ac.in

Bankim Patel
Computer Science & Engineering Department
Devang Patel Institute of Advance Technology and Research, Charotar University of Science and Technology (CHARUSAT)
Changa, Anand, India
bankimpatel.dcs@charusat.ac.in

***Abstract*— Recent improvements have increased the adoption of Multimodal Large Language Models (MLLMs) over traditional uni-modal systems since they can manage and combine information from several types of data. However, current multimodal systems have certain limitations, particularly when it comes to retrieving relevant information for domain-specific queries. This study investigates how Retrieval-Augmented Generation (RAG) approaches can help multimodal LLMs provide more contextually accurate answers from external data sources for Question and Answering (Q&A) systems. The research compares two state-of-the-art models—Google's Gemini-1.0-Pro and OpenAI's GPT-4o-mini for a Q&A system that utilizes Multimodal RAG architecture. Additionally, the study incorporates several embedding models, including Text-embedding-ada-002-v2 and embedding-001, to further improve retrieval performance. In the absence of any standardized criterion, the study used a custom-made multimodal dataset and evaluated the system using six metrics defined by human analysis. The results revealed that using RAG with multimodal LLMs greatly improves performance in Q&A tasks, with GPT-4o-mini slightly outperforming Gemini-1.0-Pro by 5%. These findings promote advancements for research in the field of Multimodal-RAG systems and highlight their potential for much better information retrieval in specialized areas.**

Keywords— Multimodal llms; Retrieval Augmented Generation (Rag); Multimodal-Rag (Mu-Rag), Question answering systems, GenAI (Generative AI).

I. INTRODUCTION

Recent advancements in Retrieval-Augmented Generation (RAG) have significantly enhanced the capabilities of Large Language Models (LLMs) by improving their accuracy and reliability, particularly in tasks requiring access to large external knowledge bases. RAG methods utilize both parametric and non-parametric memory to perform better on knowledge-intensive NLP tasks by generating more factual and context-aware responses [1]. The work by Lewis et al. (2020) demonstrates the effectiveness of these models in combining retrieval mechanisms with generation capabilities to produce specific, diverse, and accurate content [1]. However, these models predominantly focus on text-based inputs, which limits their applicability in real-world scenarios that involve diverse data forms such as images, tables, and videos.

Multimodal Large Language Models (MM-LLMs) extend the capabilities of traditional LLMs by incorporating visual, textual, and sometimes auditory data, enabling more context-aware responses. This is particularly important as the world's information is inherently multimodal, and combining modalities allows for deeper understanding and richer contextualization [2]. However, even with their advanced input processing, current MM-LLMs often fall short when generating integrated content across these modalities. Studies such as demonstrate the importance of precise retrieval mechanisms in ensuring the effectiveness of such models in open-domain question-answering tasks [3].

To address these limitations, Multimodal Retrieval-Augmented Generation (Mu-RAG) systems have been proposed. These systems improve upon base multimodal LLMs by not only retrieving relevant information from parametric memory but also from external, non-parameterized sources like databases and knowledge repositories [4]. This expansion of the data pool enhances the model's accuracy and relevance in generating responses to complex, knowledge-intensive questions [2]. The study carried out in [5] further emphasize the potential of advanced RAG techniques in optimizing model performance through the integration of external knowledge sources, which aligns with the goals of developing scalable Mu-RAG frameworks [5].

The research proposes an enhanced Mu-RAG framework that aims to address these challenges. By integrating

Authorized licensed use limited to: DePaul University. Downloaded on July 02,2025 at 04:38:39 UTC from IEEE Xplore. Restrictions apply.

retrieval-augmented multimodal models, the research aims to enable base MM-LLMs to reference relevant text, images, and tables from non-parameterized data sources. This combination improves the accuracy and context-awareness of generated content, supporting a more efficient approach to handling the complexities of multimodal tasks.

This study assesses the performance of two state-of-the-art MM-LLMs—ChatGPT by OpenAI [6] and Gemini by Google—within a RAG-based question-answering system designed to handle PDFs containing multimodal data. The models are evaluated based on how well they can retrieve, integrate, and generate responses from multimodal information.

The contributions of this research are threefold: first, the study proposes a scalable Mu-RAG framework capable of handling the growing complexity of multimodal tasks; second, this study conducts a comparative analysis of ChatGPT and Gemini to evaluate their effectiveness in retrieving and integrating multimodal data; and third, it explores the broader implications and future potential of Mu-RAG systems across various knowledge-intensive domains.

Here is the outline of the main contributions:

1. The question answer system is implemented with multi model data which are images, tables and text.
2. A comparative evaluation of two state-of-the-art multimodal large language models (MM-LLMs), ChatGPT by OpenAI and Gemini by Google, is conducted within a RAG for question answer system.
3. The analysis is focused on how effectively these models retrieve, integrate, and generate responses from multimodal data, such as PDFs.

Section II describes related work from the previous studies. The methodology is outlined in Section III. Section IV discusses results obtained along with a list of hyper-parameters, description of dataset, and the evaluation metrics used. Section V finally summarizes the conclusions drawn from this research and suggests possible further directions

II. Literature Review

The increased advancement in the field of Generative AI research and innovation is largely due to the creation of large language models (LLMs). This advancement began with fundamental language models and progressed to the development of large neural networks like GPT-3.5 and GPT-4, etc. The most recent advancement comes in the form of Multimodal LLMs (MLLMs). These models are versatile across domains and can generate efficient, intelligent solutions. They excel at understanding visual cues in chatbot conversations, leading to more context-aware responses. MLLMs can handle multi-modal inputs and provide a diverse range of content due to large-scale training on enormous multi-modal datasets. In comparison, traditional chatbots often rely on speech and textual techniques to facilitate human-machine interactions, as seen in various voice-based assistants [7].

Preliminary research on multimodal LLMs identified numerous ways of combining textual and visual knowledge, as demonstrated by BLIP-2 projects BLIP-2 (Li et al., 2023e) [8], LLaVA [9], and MiniGPT4 (Zhu et al., 2023a) [10]. Furthermore, research has explored transformer-based multimodal self-supervised learning using raw text, videos, and audio [11]. Flamingo [12] introduced a cross-attention method to connect image encoders with LLMs. Building on this, BLIP-2 [7] improved performance by using a lightweight Querying Transformer. This advancement led to significant improvements, surpassing Flamingo80B's performance by 8.7% on zero-shot VQAv2 while requiring far fewer parameters [13]. The literature contains numerous examples of core multimodal LLMs developed by aligning well-trained encoders from different modalities with the textual feature space of LLMs to process other modal inputs.

A growing body of research has been exploring the potential of Multimodal LLMs to handle open-ended question-answering (Q/A) tasks. A notable advancement in this field is BLIVA [14], which enhances InstructBLIP by adding a Visual Assistant. The model integrates query embeddings and maps patch embeddings directly into the LLM [14]. Enhancing the flexibility of MM-LLM, the study demonstrates its remarkable capability in solving tasks across various domains, including healthcare. In [15], the authors examined the effectiveness of MM-LLM in addressing challenges in the healthcare sector and proposed a framework called HeLM (Health Large Language Model for Multimodal Understanding), which learns to encode complex data modalities like images, etc.

In a major advance, Multimodal Large Language Models (MM-LLMs) have showcased transformative capabilities in language comprehension and output generation. However, they still face fundamental challenges, such as outdated internal knowledge and hallucinations. With the remarkable potential of Retrieval-Augmented Generation (RAG) in offering current and useful auxiliary information through non-parametric data sources, Multimodal LLMs combined with RAG (MuRAG) have been developed to utilize external and credible knowledge bases.

However, most studies based on the experimental analysis of the existing multimodal-LLMs suggested that the conventional Multimodal RAG processes the images and text separately for the generation of results. Research in 2011 [16] focused on experiments using two datasets. These experiments combined image and text data processing to generate context-aware answers using an external non-parametric index [16]. Although this approach of Multimodal Large Language Models – RAG (Mu-RAG) have contributed immensely in understanding and interpreting visual cues to provide more context-aware responses, there is a considerable void in the literature regarding a comprehensive review of this approach [17].

This study proposes a generalized solution on achieving effective retrieval and generation of results from documents with semantically connected multimodal data elements like text, tables and images.

III. Methodology

This research sets up a Retrieval-Augmented Generation (RAG)-based question-answering system to compare the performance of two prominent models, OpenAI's GPT-4o-mini [18] and Gemini-1.0-Pro by Google [19], in handling

Authorized licensed use limited to: DePaul University. Downloaded on July 02,2025 at 04:38:39 UTC from IEEE Xplore. Restrictions apply.

PDFs that include multimodal data such as text, images, and tables. These models were selected for their advanced multimodal processing capabilities and distinct architectural approaches. This selection offers valuable insights for both academic research and industry applications.

The overall workflow the system is presented in figure 1. The system processes these PDFs by embedding text and generating summaries for images and tables. Both models GPT-4o-mini for OpenAI and Gemini-1.0-Pro for Google—generate responses based on these embeddings. These embeddings are stored in ChromaDB [20] , a vector database that allows for fast retrieval in response to user queries. The performance of both models is compared based on the quality of their outputs in identical tasks.

A. Retrieval Component

The first stage of the process is Document Preprocessing, where the content of the PDF is divided into three key categories: text, images, and tables. For the text data, the content is segmented into manageable parts to ensure efficient retrieval later. Images are extracted and summarized in a way that captures their essential ideas, for instance, summarizing graphs and key visual elements. Tables, on the other hand, are summarized to highlight the most critical information and trends. The raw content extraction from PDFs is acheived using the Unstructured package from the LangChain framework. This package efficiently separates text, images, and tables while maintaining their semantic relationships. By processing these three categories in a structured manner, the system ensures uniform handling of text, image summaries, and table summaries before converting them into embeddings.

Following preprocessing, the second stage is Embedding Generation. In this step, embeddings for the text, image summaries, and table summaries are produced. The method for generating these embeddings depends on the model in use. For OpenAI, the system uses the Text-Embedding-Ada-002 model to convert the processed text chunks, image summaries, and table summaries into embeddings. The Text-Embedding-Ada-002 model was specifically chosen for its p superior performance in handling semantic data, outperforming alternatives like SBERT and T5 in multimodal document retrieval accuracy. GPT-4o-mini then uses these embeddings to generate responses. For Gemini-1.0-Pro, the system employs Gemini's Embedding-001 to embed the same data types, which is then used by the Gemini-1.0-Pro LLM to produce answers. This step enables a direct comparison between how OpenAI and Gemini handle identical multimodal data.

In the third stage, Vector Storage comes into play, where all embeddings (text, image summaries, and table summaries) are stored in ChromaDB. ChromaDB indexes these embeddings to facilitate quick and efficient retrieval during the query-answering phase. The database also supports Similarity Search, where a user's query, once embedded, triggers ChromaDB to search for relevant document embeddings. The retrieval process pulls content related to the user's query, covering all three data types—text, image summaries, and table summaries

The fourth step, Query Embedding Generation, occurs when a user submits a query. Depending on the model in use, either GPT-4o-mini or Gemini-1.0-Pro generates an embedding for the query. For OpenAI's model, GPT-40-mini processes the query embedding, while for Gemini, the embedding is generated using Gemini-1.0-Pro. Once the query is embedded, the system searches ChromaDB for document embeddings that closely match the query, retrieving the most relevant text, image summaries, and table summaries.

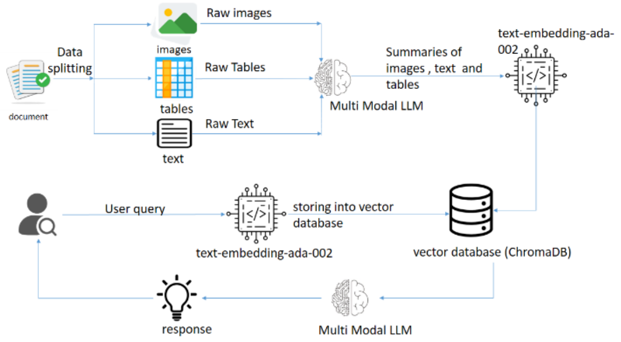

Fig. 1. Outline of a Multi-Modal Retrieval-Augmented Generation (RAG) system architecture

The final step in this component is the Retrieval Phase, where the system uses a vector similarity search in ChromaDB to find the most pertinent embeddings based on the query. For OpenAI, the system retrieves embeddings processed by the Text-Embedding-Ada-002 model, while for Gemini-1.0-Pro, it retrieves embeddings processed by Gemini's Embedding-001. These embeddings are then provided to the respective language models for generating responses.

B. Summarization of Images and Tables

Once the retrieval process is complete, the system summarizes non-textual elements like images and tables.

A streamlined approach is implemented to process PDFs in the system. A single LLM (either GPT-4o-mini or Gemini-1.0-Pro) is utilized for summarization task. Initially, summaries of the content are created. For images, details and features are captured by the LLM and then the summaries for the same are generated. For tables, brief summaries highlighting key trends and important numbers are produced.

First, for Image Summaries, the system creates embeddings that capture the key points of visual content such as graphs and charts. For OpenAI, the Text-Embedding-Ada-002 model is used to generate embeddings from these image summaries, while Gemini-1.0-Pro uses its Embedding-001 for the same task. Similarly, Table Summaries are generated by summarizing the extracted tables to focus on essential data points, trends, and comparisons. These summaries are then embedded using Text-Embedding-Ada-002 for OpenAI and Embedding-001 for Gemini-1.0-Pro. This ensures that even non-textual elements are incorporated into the question-answering process, making it comprehensive and efficient.

Authorized licensed use limited to: DePaul University. Downloaded on July 02,2025 at 04:38:39 UTC from IEEE Xplore. Restrictions apply.

C. *Query Processing and Answer Generation*

Following the retrieval of embeddings from ChromaDB, the system moves into the Context Retrieval phase. Here, the pertinent embeddings, including text, image summaries, and table summaries, are combined to create a contextual foundation for generating a response. For OpenAI, GPT-4o-mini utilizes the embeddings produced by Text-Embedding-Ada-002 to generate the response. Similarly, for Gemini-1.0-Pro, the embeddings from Gemini's Embedding-001 are used to generate the final answer. In the final stage, Answer Generation, the language models generate responses based on the retrieved context. GPT-4o-mini for OpenAI and Gemini-1.0-Pro both use the embedded content, which includes text, image summaries, and table summaries, to produce the final response. This ensures that the models generate answers that are not only accurate but also contextualized using data from all available modalities.

D. *Summary of Retrieval and Answer Process*

The entire process begins with PDF Content Extraction, where the text, images, and tables are separated and summarized. Next, Embedding Generation and Storage ensures that these components—text, image summaries, and table summaries—are embedded using either Text-Embedding-Ada-002 for OpenAI or Embedding-001 for Gemini-1.0-Pro and stored in ChromaDB for easy retrieval. In the Query Embedding and Retrieval phase, user queries are embedded and processed to retrieve the relevant data from ChromaDB. Finally, in the Response Generation phase, the models (GPT-4o-mini for OpenAI or Gemini-1.0-Pro) use the retrieved embeddings to generate responses that include text, image summaries, and table summaries.

By comparing the performance of OpenAI and Gemini-1.0-Pro in this RAG-based system, the research provides insights into how each model handles multimodal data (text, images, and tables) and evaluates their respective capabilities in generating accurate and contextually relevant responses.

IV. Results Analysis

A. *Description Of The Dataset*

The proposed Multimodal RAG-based QA method was thoroughly evaluated on a carefully self-curated multimodal dataset. As shown in table 1, the dataset is segregated into four important categories: question, answers, context, and ground truth. Unlike conventional text-based ground truth, this dataset contains images and tables in addition to unlabeled text data, question pairs, replies, and relevant background information. This combination of textual and visual input is essential for the efficient training and assessment of the multimodal RAG system.

B. *Experimental Setup Used*

The experiments were conducted using Google Colab, which provides a cloud-based environment for machine learning tasks. The Colab instance was equipped with an NVIDIA Tesla T4 GPU, featuring 2560 CUDA cores and 16 GB of GDDR6 memory, crucial for handling large language models efficiently. The system also had 12 GB of RAM . The operating system utilized for the experiment was Windows 11, which provided a reliable and efficient environment for testing models. Important libraries used in the proposed system include ChromaDB for optimized similarity search, the transformers library for GPT-4o-mini, and the Google Gemini API, all of which are essential components for developing a Multimodal RAG-based Question-Answering system.

TABLE I. EXAMPLE FROM UTILIZED DATASET

<table>
<tr><td><b>Question</b></td><td>Can you explain the table with title : PRECISIONS USING LSTM AND DIFFERENT CLASSIFIERS ?</td></tr>
<tr><td><b>RAG Answer</b></td><td>The table provided shows the precision scores achieved using an LSTM network combined with different classifiers for a specific task. The table shows that the LSTM network achieved very high precision scores (above 99%) when combined with most of the classifiers. The highest precision was achieved with the SVM classifier (99.9632%), followed closely by the kNN classifier (99.9604%). The other classifiers also performed well, with precision scores ranging from 99.96% to 99.88%.\n\nNote:*\n\nThe table only shows precision scores and doesn't provide information about other important metrics like recall, F1-score, or accuracy. It's important to consider these metrics as well to get a complete picture of the model's performance.</td></tr>
<tr><td><b>Context</b></td><td>'The table provides a comparison of the precision rates of various network models including LSTM Classifier, Support Vector Machine, k-Nearest Neighbors, XGBoost, Gradient Boosting, Random Forest, and Decision Tree. The precision rates range from 99.8814% to 99.9632%.\n\nConclusion: The LSTM Classifier model has the highest precision rate at 99.9632%, while the Decision Tree and Random Forest models have the lowest precision rate at 99.8814%.',</td></tr>
<tr><td><b>Reference Standard</b></td><td>The table used as a reference for multimodal Q/A
<table>
<tr><th><i>Depp Learning Model</i></th><th><i>Classifier Used</i></th><th><i>Precision Achieved</i></th></tr>
<tr><td>RNN</td><td>SVM</td><td>92.911%</td></tr>
<tr><td>RNN</td><td>Random Forest</td><td>97.830%</td></tr>
<tr><td>RNN</td><td>---</td><td>96.952%</td></tr>
<tr><td>RNN</td><td>KNN</td><td>94.874%</td></tr>
</table></td></tr>
</table>

C. *Evaluation of Multimodal-RAG System*

There is considerable amount of literature regarding the methods devised for accurate evaluation of conventional text-based RAG systems, one of the most widely being known RAGAS framework proposed by [21]. Here, the key challenge is that it fails to provide a more nuanced evaluation for Multi-Modal RAG systems. These current frameworks are largely designed for unimodal systems, with much emphasis on textual data sources. However, Multimodal RAG systems need more refined evaluation metrics that considers the semantic interaction among multiple modalities (e.g., image, video, etc.)

Authorized licensed use limited to: DePaul University. Downloaded on July 02,2025 at 04:38:39 UTC from IEEE Xplore. Restrictions apply.

TABLE II. HYPERPARAMETERS LIST

Parameter Name	Description	Value
Maximum Concurrency	Maximum concurrency defines the limit on how many tasks or operations can be executed concurrently, i.e., at the same time, during the execution of a function or process.	5
Top K relevant searches	The most relevant document chunks are extracted based on their contextual similarity to the user query vector.	10
QA Prompt template	It is an input format that combines a user's query with retrieved context information to assist the model in creating a relevant and correct response.	"You are an educational assistant that responds to questions by retrieving the information from the given summaries. The answers generated should be concise and analytical. Make sure to provide answer in a complete sentence that includes comprehensive and relevant background context."
Embedding model	It is a model that converts data (e.g., text, images) into dense, low-dimensional vectors embeddings to capture the semantic relationships	text-embedding-ada-002-v2 , embedding -001
Multimodal LLMs	The models which process multiple modalities of data, such as text, images, and audio, to generate more context-aware responses or perform tasks across different types of data.	GPT-4o-mini, Gemini-1.0-pro
Vector Database	A database that stores data as vectors to enable fast and efficient similarity search.	Chroma DB

which the present frameworks fail to achieve.

Recognizing the shortcomings of the present frameworks, this study utilized a more qualitative approach, using human analysis to evaluate the efficiency of the Mu-RAG system. Human evaluators were entrusted with examining the coherence, relevance, and accuracy of the outputs generated by the proposed system, providing a more contextually aware evaluation that existing metrics cannot completely capture.

The evaluation metrics (coherence, relevance, and accuracy) were assessed using a 5-point scale by independent domain experts with expertise in the domain of Artificial Intelligence and Natural Language Processing. Each metric was evaluated on a scale of 1 (lowest) to 5 (highest), with clearly defined rubrics for scoring. For coherence, evaluators assessed the logical flow of responses. Relevance was measured by how well the response addressed the query asked by the user. Accuracy was evaluated based on the factual correctness of the generated content. The final percentage scores were calculated using a normalization formula: (average score - 1) × 25, which converts the 5-point scale to a 0-100% range. This standardization process was important while evaluation to minimize individual bias.

D. Hyperparameters

Hyperparameters play a critical aspect in determining the performance, accuracy and coherence of Multimodal RAG based Question-Answering Systems.

The proposed Multimodal-RAG system is implemented using six different hyper-parameters for both retrieval and generation tasks, as shown in Table 2. Since the proposed study deals with Multimodal data, maximum concurrency is an important hyper-parameter as it directly affects several aspects of the system's performance, including training efficiency, resource consumption, and model quality.

Through considerable testing, the best concurrency value was determined to be 5 that ensured optimal balance between the generated output of the system and resource utilization. Highlighting the importance and after thorough evaluation, it was found that the best outcomes for this architecture were achieved with a maximum concurrency value of 5.

E. Result Discussions:

Addressing the void in the literature regarding a bias-free generic framework to evaluate Multimodal RAG systems, this study implemented human analysis for detailed examination of the Mu-RAG system.

Human analysis was performed to investigate the interplay of coherence, relevance, and correctness in outputs that included text, tables, and images—an area where the existing automated frameworks may fall short of while capturing complex semantic knowledge.

Table 3 shows the evaluation of the two state-of-art multimodal language models: ChatGPT-4o-mini and Gemini-1.0-pro, across different modalities—text, tables, and images. This assessment was based on three key factors: coherence, relevance, and accuracy of the generated outputs.

- Coherence: This measures how logically consistent and well-structured the generated output is.
- Relevance: Relevance assesses how well the generated output corresponds to the query or task being addressed.
- Accuracy: Accuracy refers to how contextually-aware and concise the generated content is.

Authorized licensed use limited to: DePaul University. Downloaded on July 02,2025 at 04:38:39 UTC from IEEE Xplore. Restrictions apply.

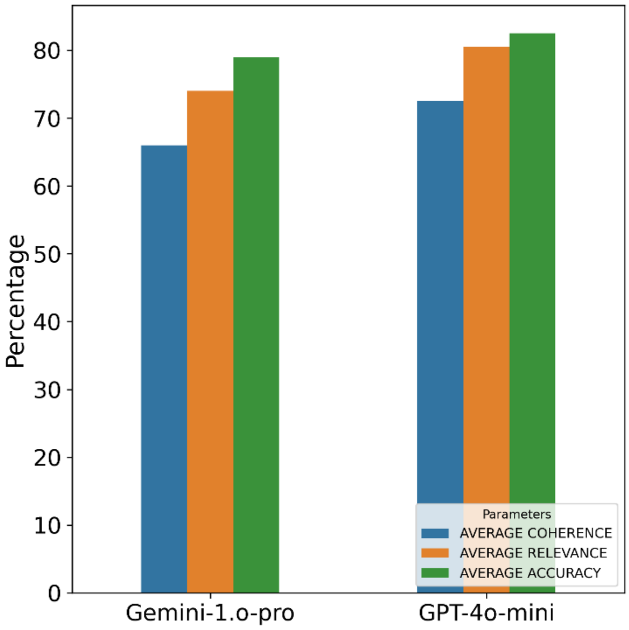

Fig. 2. Human Evaluation Based Performance Analysis of Mu-RAG framework

Figure 2. Demonstrates the analysis of Mu-RAG LLMs, Gemini-1.0-pro and GPT-4o-mini across three parameters notably- Average coherence, Average relevance and Average Accuracy based on Human Analysis. This analysis revealed that GPT-4o-mini (72.5%) maintained contextual accuracy and logical consistency in comparison to Gemini-1.o-pro (66%). The Relevancy rate for GPT-4o-mini (80.5%) was strikingly high indicating that the later outperforms Gemini-1.o-pro (74%) in generating responses that are contextually relevant to the user query. Interestingly, there was no significant difference identified between the two models suggesting that GPT-4o-mini (82.5%) was marginally capable than Gemini-1.0-pro (79%) for fact based retrieval tasks.

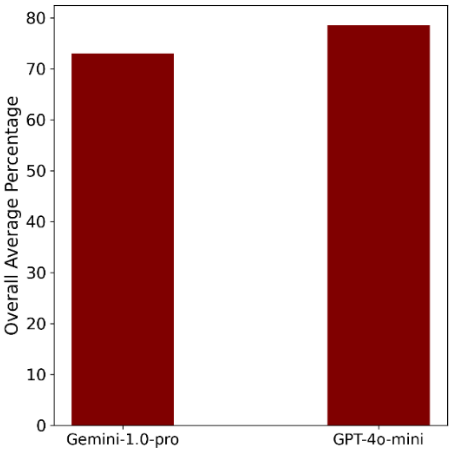

Fig. 3. Overall Performance of Multimodal LLMs

Figure 3 summarizes the overall performance of each Multimodal LLM through an in depth human evaluation based on the key factors- coherence, accuracy and relevancy. The graph illustrates a comparative performance analysis on the basis of the average results achieved by Gemini-1.0-pro and GPT-4o-mini across the proposed three metrics. The overall performance average achieved by Gemini-1.0-pro and GPT-4o-mini are 75% and 80%, respectively. Interestingly, both the models demonstrate considerably good performance, with GPT-4o-mini marginally outperforming Gemini-1.o-pro for Multi-Modal documents. This slight difference between the performance scores depict that both the models are effective in processing Multi-Modal data but GPT-4o-mini demonstrates superior performance most likely due to advanced fine-tuning.

TABLE III. EVALUATION RESULTS FOR MU-RAG FRAMEWORK FOR UNSTRUCTURED TEXT, IMAGES & TABLES

Evaluation Parameter	Unstructured Text		Images and Tables	
	Gemini-1.0-pro	GPT-4o-mini	Gemini-1.0-pro	GPT-4o-mini
Coherence	72%	75%	60%	70%
Relevance	79%	82%	69%	79%
Accuracy	78%	80%	80%	85%

V. Conclusion and Future Scope

This study concludes that Multimodal systems represent the paradigm of the future as they exhibit exceptional understanding in comparison to conventional text-based framework across various modalities of data. Furthermore, the incorporation of Retrieval Augmented Generation (RAG) with MM-LLM will allow the models to curate systems that can handle various cross-modality queries related to real-world applications. This study utilized Human evaluation to draw comprehensive analysis between Gemini-1.0-pro and GPT-4o-mini for Multimodal Based Information Retrieval systems. The evaluation highlighted the overall performance of each Multi-Modal LLM, with GPT-4o-mini marginally outperforming Gemini-1.0-pro across three factors i.e., coherence, accuracy and relevance. This study also addresses the significant gap in the literature regarding the development of bias-free and generic evaluation metrics for Mu-RAG framework, proposing future work in this realm of retrieval augmented tasks. This research faces limitations related to dataset customization and the challenges of processing multiple data types simultaneously. The human evaluation process provided useful insights with some personal bias due to its reliance on human judgment. Future work should focus on developing better evaluation metrics and improving how the system handles interactions between text, images, and tables. This would make the system more practical for real-world applications while reducing computational overhead.

Acknowledgement

This research is financially supported by Charotar University of Science and Technology (CHARUSAT).

References

[1] P. Lewis, E. Perez, A. Piktus, et al., “Retrieval-augmented generation for knowledge-intensive nlp tasks,” Advances in Neural Information Processing Systems, vol. 33, pp. 9459–9474, 2020.

[2] W. Yu, “Retrieval-augmented generation across heterogeneous knowledge,” in Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies: student research workshop, 2022, pp. 52–58.

[3] V. Karpukhin, B. O˘guz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, “Dense passage retrieval for open-domain question answering,” arXiv preprint arXiv:2004.04906, 2020.

Authorized licensed use limited to: DePaul University. Downloaded on July 02,2025 at 04:38:39 UTC from IEEE Xplore. Restrictions apply.

[4] S. Siriwardhana, R. Weerasekera, E. Wen, T. Kaluarachchi, R. Rana, and S. Nanayakkara, "Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering,"Transactions of the Association for Computational Linguistics, vol. 11,pp. 1–17, 2023.

[5] J. Huang, W. Ping, P. Xu, M. Shoeybi, K. C.-C. Chang, and B. Catanzaro, "Raven: In-context learning with retrieval augmented encoder-decoder language models," arXiv preprint arXiv:2308.07922, 2023.

[6] T. B. Brown, "Language models are few-shot learners," arXiv preprint arXiv:2005.14165, 2020.

[7] P. Goel and A. Ganatra, "A survey on chatbot: Futuristic conversational agent for user interaction," in 2021 3rd international conference on signal processing and communication (ICPSC). IEEE, 2021, pp. 736– 740.

[8] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in International conference on machine learning. PMLR,2023, pp. 19 730–19 742.

[9] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao, "Llava-med: Training a large language-and-vision assistant for biomedicine in one day," Advances in Neural Information Processing Systems, vol. 36, 2024.

[10] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," arXiv preprint arXiv:2304.10592, 2023.

[11] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong, "Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text," Advances in Neural Information Processing Systems, vol. 34, pp. 24 206–24 221, 2021.

[12] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc,A. Mensch, K. Millican, M. Reynolds et al., "Flamingo: a visual language model for few-shot learning," Advances in neural information processing systems, vol. 35, pp. 23 716–23 736, 2022.

[13] S. Wu, H. Fei, L. Qu, W. Ji, and T.-S. Chua, "Next-gpt: Any-to-any multimodal llm," arXiv preprint arXiv:2309.05519, 2023.

[14] W. Hu, Y. Xu, Y. Li, W. Li, Z. Chen, and Z. Tu, "Bliva: A simple multimodal llm for better handling of text-rich visual questions," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 3, 2024, pp. 2256–2264.

[15] A. Belyaeva, J. Cosentino, F. Hormozdiari, K. Eswaran, S. Shetty,G. Corrado, A. Carroll, C. Y. McLean, and N. A. Furlotte, "Multimodal llms for health grounded in individual-specific data," in Workshop on Machine Learning for Multimodal Healthcare Data. Springer, 2023,pp. 86–102.

[16] W. Chen, H. Hu, X. Chen, P. Verga, and W. W. Cohen, "Murag: Multimodal retrieval-augmented generator for open question answering over images and text," arXiv preprint arXiv:2210.02928, 2022.

[17] "GPT-4O Mini: Advancing cost-efficient intelligence," Openai, https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence (accessed Aug. 27, 2024).

[18] S. Pichai, "Introducing Gemini: Our largest and most capable AI model," Google, https://blog.google/technology/ai/google-gemini-ai/ (accessed Aug. 27, 2024).

[19] S. Tong, Z. Liu, Y. Zhai, Y. Ma, Y. LeCun, and S. Xie, "Eyes wide shut? exploring the visual shortcomings of multimodal llms," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 9568–9578.

[20] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "Ragas: Automated evaluation of retrieval augmented generation," arXiv preprint arXiv:2309.15217, 2023.

Authorized licensed use limited to: DePaul University. Downloaded on July 02,2025 at 04:38:39 UTC from IEEE Xplore. Restrictions apply.