RESEARCH ARTICLE

INTERNATIONAL RESEARCH JOURNAL OF MULTIDISCIPLINARY TECHNOVATION

Asian Research Association

DOI: 10.54392/irjmt2626

A Multimodal RAG and ArXiv-Integrated Conversational Framework for Automated Research Discovery

Ashwini Dalvi [a], Suswar Sawant [a,*], Amaan Syed [a], Irfan Siddavatam [a], V. Venkataramanan [a]

[a] Department of Information Technology, KJ, Somaiya School of Engineering (formerly KJ Somaiya College of Engineering), Somaiya Vidyavihar University, Mumbai, 400077, Maharashtra, India

* Corresponding Author Email: suswar.s@somaiya.edu

DOI: https://doi.org/10.54392/irjmt2626

Received: 30-10-2025; Revised: 11-02-2026; Accepted: 20-02-2026; Published: 11-03-2026

Abstract: Knowledge-rich fields employ multimodal documents, which require advanced analysis systems to assess document content and enable research exploration. The research introduces a unified system which combines Multimodal Retrieval-Augmented Generation (RAG) technology with automated arXiv research discovery functionality. The system operates through a user-friendly application that runs on the Streamlit web platform. The system demonstrates robust performance, achieving an average faithfulness of 0.86, with 0.81 answer relevancy on a diverse Portable Document Format (PDF) dataset. The system employs a hierarchical retrieval architecture which enhances contextual content retention capability through its complete document ingestion process that handles 15 pages within 12 seconds, which outperforms MultiModal-GPT and other evaluated systems by 33 percent. The automated arXiv Integration System also offers paper recommendations with a relevance confidence percentage of 80% at a cost of under 3 seconds. The research presents a scalable high-performance solution through its streaming conversational interface which uses LangChain Expression Language (LCEL) to demonstrate a real-world application of Conversational AI that emphasizes fast system responsiveness and user-friendly multimodal document analysis.

Keywords: ArXiv Integration, Conversational AI, Hierarchical Retrieval, Multimodal RAG, RAGAS Evaluation, Research Discovery.

1. Introduction

Knowledge workers in the present digital world encounter an excessive amount of information that exists in complicated multimodal formats that include Portable Document Format (PDF) as its main form. The system of Retrieval-Augmented Generation (RAG) establishes a new method to base Large Language Models (LLMs) on outside databases for accurate fact checking and hallucination control, which represents a major problem in generative Artificial Intelligence [1-3]. Existing RAG systems encounter multiple major difficulties that prevent them from functioning properly. Most systems operate with a primary focus on text, yet they fail to retrieve and utilize all the meaningful information that exists in non-textual materials, such as charts and images, along with tabular datasets that need special extraction technologies [4, 5]. The result leads to document comprehension problems that affect the entire document processing system [5, 6]. The research process in large databases such as arXiv requires manual exploration, which results in an inefficient research process that creates a major research barrier [7], while basic chunking methods break the original context, which results in destroyed value of attained data [3].

The Multimodal RAG pipeline and automated arXiv research discovery mechanism form our integrated framework solution that addresses these missing system elements. The architecture uses RAGAS [8] as its evaluation system to show its first novel feature, which includes an end-to-end system that achieves low latency and provides fault-tolerant API operations through its retry feature. The system functions as a complete streaming system that operates through the LangChain Expression Language (LCEL) conversational interface. The field approaches its upcoming stage of operational Large Language and Vision Assistant Multimodal model (LLaVA) development according to these contributions, which have become very relevant [9]. The system achieves strong performance with a mean faithfulness score of 0.86 and answer relevancy score of 0.81. The present research establishes a system that can grow in size while maintaining high performance to improve conversational AI tools, which operate multimodal document analysis functions through its new system.

The main mission of this research project involves creating a scalable solution that unites advanced conversational AI technologies with multimodal document analysis capabilities. According to current research, evaluation-based RAG systems and long-context retrieval systems play essential roles in decreasing hallucinations while enhancing grounding functions for multimodal evidence [10-12]. The project aims to create a hierarchical streaming architecture that will improve real-time responsiveness as its first goal. The project aims to create a text table and image extraction system that will operate at maximum efficiency through its parallel processing design. The document analysis process will create a practical system for research discovery through its document analysis process. This research investigates how hierarchical retrieval methods combined with streaming architectures affect the connection between contextual relevance and processing time, while automated arXiv integration improves research discovery capabilities.

2. Related Work

The practice of Retrieval-Augmented Generation (RAG) is beyond general-purpose domains; it has grown into a paradigm of highly specialized domains. Recent research has shown that RAG architecture designs have evolved into specialized systems in particular fields. Zhang *et al*. [13] introduced TimeRAF as a tool to enhance retrieval foundations for zero-shot time-series forecasting, which proves RAG can be applied to dynamic time-dependent information. The development of automated fact-checking systems required Hang *et al*. [14] to create graph-based retrieval models that resolve the constraints found in standard flat retrieval systems. The QuIM-RAG system developed by Saha *et al*. [15] combines inverted question matching with their system to address difficult QA performance problems. RAG development requires these specific settings because they support the design of solutions that address particular real-world challenges that need expert knowledge from specific fields.

Models such as Contrastive Language–Image Pre-training (CLIP) created the first systems that enable visual representation learning through the process of natural language [16]. The system can utilize its new capability, which enables reasoning about both text and image data. The implementation of this technology requires research libraries that contain specific functions for tool development. The extraction of image data from structured documents, such as PDF files, requires the specialized function of a library that PyMuPDF provides through its high-performance image extraction capabilities [17]. RAG systems that scale operations require an effective retrieval process as the core element. Modern architectures that exist today depend on GPU-optimized libraries such as Facebook AI Similarity Search (FAISS) to execute similarity searches across billion-scale datasets [18]. This technology enables users to discover relevant contexts with reduced latency operating in real time. The evaluation methods used to assess RAG systems have experienced new patterns of development since previous evaluations. The field advanced from traditional Bilingual Evaluation Understudy metrics to more sophisticated assessment systems through this transition [10]. Current systems achieve their generative capabilities by using cutting-edge multimodal models together with high-dimensional text embeddings, which enable them to understand complex semantic relationships [19].

The industrial sector implements RAG through educational personalization [11], medical question-answering capacity [12], and coding with new channelizing pipeline parallelism [20] owing to retrieval-augmented text generation research progress, which drives their rapid growth. The community demonstrates a common trend that involves users of open-source LLMs discovering methods to improve accessibility while decreasing their dependency on proprietary models [21]. Recent work has also demonstrated the effectiveness of retrieval-augmented generation in multilingual and safety-critical applications such as medication counseling, where grounding and retrieval accuracy are essential for reliability [22]. The field has transitioned from basic retrieval methods to Multimodal Chain-of-Thought reasoning processes, which require complex reasoning abilities. The newly created complete taxonomy for this field shows that basic entity retrieval does not fulfill the requirements for working through multi-step reasoning tasks [23]. The system introduced Retrieval-Augmented Personalization (RAP), which allows users to retrieve visual concepts based on their needs, resulting in better output quality for multimodal assistants [24]. Researchers have achieved progress in knowledge graph and CoT integration, which helps decrease vertical domain hallucinations by enforcing structured reasoning paths before users generate their answers [6]. The studies show that the proposed framework needs to use its retrieval system which operates at hierarchical levels together with its reasoning-aware capabilities for effective functioning. Table 1 summarizes representative RAG-based frameworks and highlights their core focus areas along with the key research gaps that motivate the proposed approach.

3. Methodology

This section details the architectural design, implementation, and evaluation protocol of the arXiv-enhanced Multimodal RAG framework. The complete pipeline system starts with continuous data ingestion through parallel processing and ends with the Multimodal PDF content delivery system and its semantic retrieval process which uses chunking and embedding and indexing parameters for document processing.

Table 1. Comparative Summary of Related Work and Identified Gaps

Reference	Study / Framework	Core Focus	Key Limitation / Research Gap Identified
Lewis *et al.* (2020) [2]	Text-Only RAG	Foundational retrieval-augmented generation for NLP tasks.	Limited to textual data; fails to process rich multimodal content (images, tables) found in scientific PDFs.
Zhang *et al*. (2025) [13]	TimeRAF	Retrieval-augmented foundation model for zero-shot time-series forecasting	Designed for temporal data; not suitable for multimodal document understanding or research discovery
Kim *et al*. (2024) [25]	Geometry-Aware RAG	3D object detection and spatial reasoning.	Highly specialized embeddings lead to poor performance on general-purpose queries; computationally expensive.
Heydari *et al.* (2024) [26]	AutoRAG / Context-Aware	Context-awareness gates for retrieval in autonomous driving.	Domain-specific architecture that requires significant adaptation for general academic or scientific document analysis.
Wang *et al.* (2026) [5]	VisualRAG	Knowledge-guided image-text retrieval	Limited to vision–text matching; does not address holistic, full-document multimodal comprehension.
Han Chen & Shiu (2025) [27]	KAQG	Knowledge-graph-enhanced RAG for controlled question generation	Focuses on KG-based QA; lacks multimodal document ingestion and automated external research discovery.
Proposed Framework	arXiv-Integrated Multimodal RAG	Parallelized multimodal ingestion with external discovery.	Addresses latency via parallelization; bridges internal document analysis with external literature search.

The following section describes the construction of a complete streaming conversational system and a new research discovery tool for arXiv and the resources needed to test the system.

3.1 System Architecture and Workflow

The framework uses a complete streaming system which combines permanent storage and simultaneous processing to deliver exceptional results. Figure 1 displays the complete system architecture which shows how the main components interact from user input until the system shows its responses.

The system begins its operation by establishing a persistence layer that assigns a distinct Secure Hash Algorithm SHA-256 hash to each uploaded PDF document while storing the associated FAISS index on the storage system. The system checks the PDF document hash during each upload to prevent processing of already indexed documents which enables direct reconstruction of the conversational chain from the saved index. Newly uploaded documents are ingested via a pipeline set for parallel execution. The ThreadPoolExecutor enables simultaneous document processing because it allows users to extract text tables and images from all documents using PDFplumber and PyMuPDF. Tesseract OCR receives rasterized pages which contain no extractable text so they can process scanned content to recover text. The workflow proceeds with batching extracted images which are then resized to an 800 × 800 pixel maximum dimension while their summarization process uses parallel GPT-4o vision endpoint calls. The tenacity library wraps each API call to preserve the system against temporary network disruptions. The framework uses intent routing during query time to identify user intention. For research discovery requests, it calls the arXiv module. The LCEL streaming conversational chain uses user questions together with chat history to build an independent query that collects context information and sends the GPT-4o generated response to the user interface through direct streaming.

3.2 Architectural Novelty and Systemic Integration

The basic elements of retrieval-augmented generation, which include FAISS indexing and LLM generation, already exist as established technologies. The proposed framework creates a new architectural unification that integrates existing systems to solve particular problems that reduce both efficiency and workflow continuity.

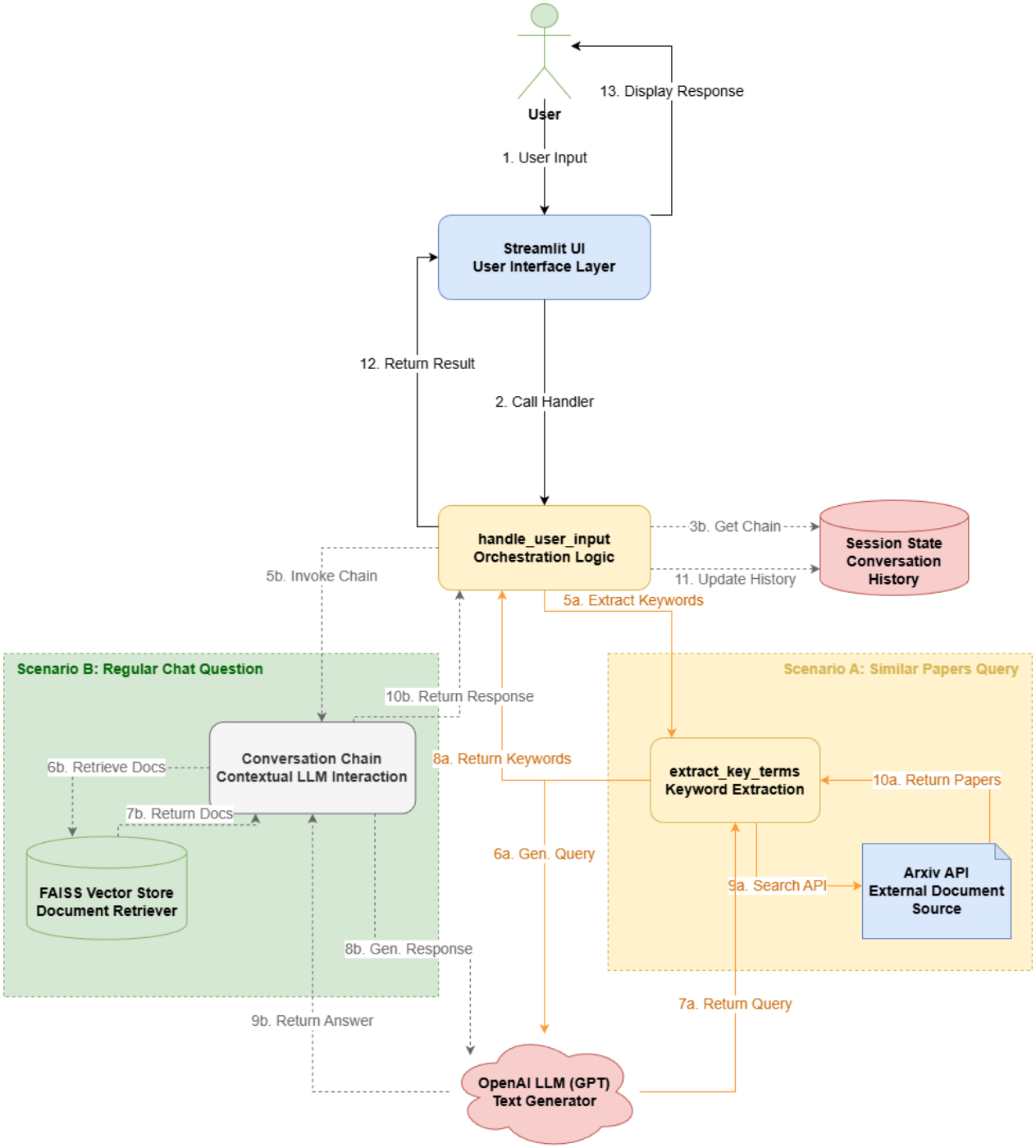

Figure. 1. System architecture for arXiv-enhanced Multimodal RAG, illustrating preprocessing, embedding, retrieval, arXiv integration, and RAGAS evaluation.

The study makes its main contribution through component integration, which resolves latency and usability problems found in existing multimodal systems instead of developing a primary retrieval algorithm. The research employed an end-to-end concurrent processing method to extract text, tables, and images simultaneously because multimodal analysis required all three elements to function simultaneously, while ingestion latency was reduced by 64 percent compared to traditional sequential methods. A unique intent-routing mechanism enables users to transition from internal document comprehension to external literature discovery for queries regarding documents. The system enables research exploration through document assessment capabilities that Multimodal-GPT and Geometry-Aware RAG do not offer. The architecture uses a fully streaming conversational chain that operates on LCEL to provide real-time responsiveness and fault-tolerant capabilities while separating user experience from the computational demands of multimodal processing [28].

3.3 Retrieval Strategy

This framework incorporates an advanced system designed to guarantee optimal retrieval quality,

specifically through a method of semantic chunking and indexing. The entire text, which has undergone normalization, is semantically chunked into large chunks of 600–1000 overlapping characters using a paragraph- and sentence-aware system. These chunks were then transformed into 3,072-dimensional embeddings using text-embedding-3-large from OpenAI. The produced vectors were indexed in a FAISS vector store. A FlatL2 FAISS index performs a dense compatibility query without any quantization, thus providing deterministic results. During retrieval, the top-k parameter was kept constant at 5 to perform a regular compromise between the recall for context and latency. FAISS is ahead of alternatives such as Chroma or Milvus with its highly optimized CPU and GPU search primitives, strong community support, and simple usage with NumPy to avoid using an external service and slow down indexing latency for a dataset size of this study.

3.4 Algorithm

The logical control flow of the caching, parallelization, and streaming framework is presented as high-level pseudocode in Algorithm 1. This abstraction illustrates the sequential and parallel component interactions, abstracting the specific implementation syntax to focus on the architectural logic. The procedure begins with the user uploading PDFs, posing questions, and presenting the chat history. Then, it checks for the existence of documents in the cache; if so, it loads the pre-computed RAG chain. If not, it goes for parallel ingestion to extract all text, tables, and images, followed by batched summarization for the images. The content is then split into parent and child chunks for the hierarchical retriever, and a FAISS index is created from the child chunk embeddings. After building the LCEL streaming chain, it is stored in the cache for future applications. Finally, the system routes the user's intent: if it is an arXiv query, it extracts key terms and searches the API; otherwise, it executes the conversational RAG chain to produce a streamed answer.

Algorithm 1. High-Level Pseudocode of the Cached, Parallelized and streaming RAG Workflow

1: Input: uploaded PDFs pdf_docs, user question q, chat history H

2: *// Step 1: Cached pipeline (via @st.cache_resource)*

3: if pdf_docs in cache then

4: (chain, retriever, data) ← load from cache

5: else

6: *// Step 2: Parallel ingestion across* all *PDFs*

7: (text, tables) ← Extract Text And Tables Parallel (pdf_docs)

8:images ← Extract Images ParalLEL (pdf_docs)

9:summaries ← Summarize Images Batched ParALLEL (images) *(resize ≤ 800×800, retries)*

10:*// Step 3: Hierarchical retriever (Parent & Child split)*

11:parent ← (chunk_size=2000, overlap=400), child← (400, 100)

12:V ← FAISS Index (text-embedding-3-large)

13:if Faiss Gpu Available and Cuda Available then

14:V ← MoveToGpu(V)

15:end if

16:retriever ← Parent Document Retriever(V)

17:retriever.add_documents (text, Serialize (tables), summaries)

18:*// Step 4: Build LCEL streaming chain (Condense → Retrieve → Generate)*

19:chain ← Build LCEL Chain()

20:store (chain, retriever, data) in cache with key pdf_docs

21: end if

22: *// Step 5: Intent routing*

23: if ArXiv Intent(q) then

24:terms ← Extract Key Terms (data.text, *k*=5)

25:query ← OrJoin(terms)

26: raw ← ARXIV SEARCH(query, cats={cs.CL, cs.CV, cs.LG, cs.AI}, year≥2018, sort=Relevance) *(retries)*

27:papers ← EmbedReRank(raw, text-embedding-3-large) *(session cache)*

28:Output: formatted papers

29: else

30:*// Streaming conversational response (condensed question used for both stages)*

31:standalone ← Condense(q, H)

32: ctx ← Retrieve(standalone) *(returns parent chunks)*

33:response_streamGenerateStream(standalone, ctx, H)

34:render tokens as they arrive

35:Output: final streamed answer

36: end if

3.5 Materials and Evaluation Setup

The system was developed using Streamlit to create the user interface and LangChain, along with its LangChain Expression Language (LCEL), to manage the RAG pipeline operations. The team employed PDFplumber, PyMuPDF, and Tesseract OCR to extract and process data from PDF documents using

PDFplumber to extract text and tables and Tesseract OCR to extract text from scanned documents. The framework was deployed on a cloud instance with an NVIDIA 3060 GPU and 32GB of system RAM to enable reproducible results. The OpenAI GPT-4o model served as the generative component, which operated at a temperature setting of 0.0 to produce consistent results without any random variations, while using a top-p value of 1.0 and a response token limit of 4,096. The text-embedding-3-large model generated its output through the process of creating embeddings, which resulted in 3,072 distinct dimensional outputs. The vector store was constructed through the FAISS library FlatL2 index method, which performed exact nearest-neighbor searches using Euclidean distance without any loss from data compression.

Generative multimodal pipelines require different evaluation metrics because traditional retrieval metrics such as Precision@k and Recall@k, which text-only systems use for evaluation, do not provide complete assessment. The researchers used the RAGAS framework to evaluate the system through reference-free analysis that examined both semantic and contextual dimensions, including tests for faithfulness, answer relevancy, context precision, and context recall [29]. The research team discovered that using GPT-4o as the generation model together with RAGAS as the evaluation tool would result in a self-preference bias for their research work. Current implementations use this model because it has better reasoning abilities than other models; however, future validations will include human-in-the-loop assessments to confirm the results of the automated metrics. The evaluation used a curated test dataset that included ten academic PDF documents with an average of 15 pages and five images per document. The documents contain information about Multimodal-GPT, Geometry-Aware RAG and AutoRAG. Researchers created eight different query types with manually curated ground-truth answers to evaluate both retrieval accuracy and generation quality [30]. The proposed architecture achieved efficiency gains, which were measured through an ingestion process benchmark that showed that parallel execution reduced the total ingestion time for the 10-PDF corpus from 158 s (sequential) to 56 s, resulting in a speed-up of 2.8x.

4. Results and Discussion

The section starts with the experimental setup and then presents quantitative results that measure the retrieval performance, processing efficiency, and generation quality. The current systems receive evaluation through quantitative results, which scientists explain using causal explanations and framework advantages and disadvantages.

4.1 Experimental Setup

The RAGAS performance evaluation was conducted across a well-defined test set of 10 academic PDF documents, each containing 15 pages and five images per document. The researchers created eight different query types to examine the specific details and new features and the way the results were compared to benchmark tests for four key academic works: Multimodal-GPT [31], Geometry-Aware RAG [25], AutoRAG [26] and Text-Only RAG [2]. This gold answer was finally curated manually, and the RAGAS metrics fidelity, answer relevance, answer correctness, semantic similarity, context precision, and context recall were computed using GPT-4o as an evaluator in reference-free conditions to assess the semantic and contextual quality of the outputs.

4.2 Performance of Retrieval and Generation Quality.

Table 2 provides the overall performance results of the framework, and Table 3 presents a detailed per-query breakdown. The obtained mean scores for faithfulness (0.8575) and semantic similarity (0.8375) reflect the results generated being firmly substantiated in the retrieved content. Figure 2 summarizes the mean RAGAS scores, providing a visual representation of the framework's strong performance in key areas.

Table 2. RAGAS Evaluation Results (Mean and Standard Deviation)

Metric	Mean	Std. Dev.
Faithfulness	0.8575	0.1217
Answer Relevancy	0.8063	0.3127
Answer Correctness	0.5375	0.1960
Semantic Similarity	0.8375	0.0630
Context Precision	0.6250	0.5175
Context Recall	0.4713	0.4099

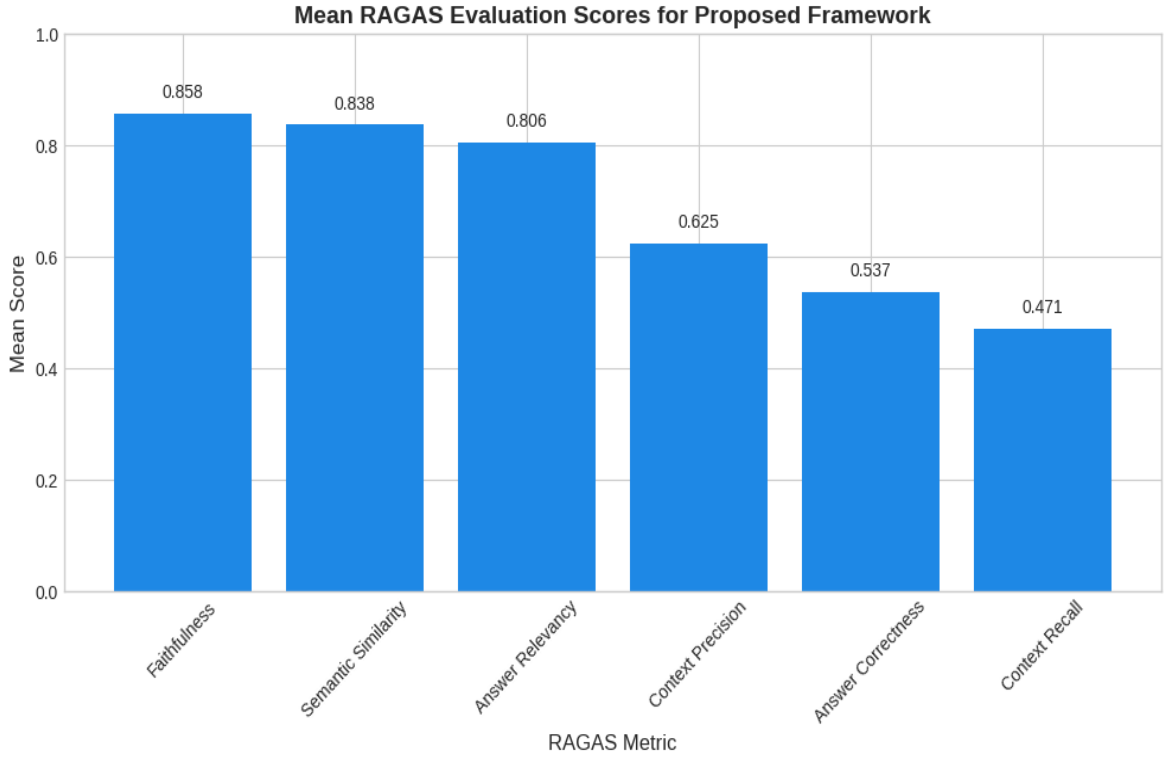

Figure. 2. Mean RAGAS evaluation scores for the proposed framework across six key metrics. Scores for Faithfulness and Semantic Similarity are notably high, indicating strong performance in generating grounded and contextually relevant answers

Table 3. Per-Query RAGAS Evaluation Results

Query (Abbreviated)	Faith.	Ans. Rel.	Ans. Corr.	Sem. Sim.	Ctx. Prec.	Ctx. Rec.
Multimodal-GPT improvement	1.00	0.86	0.72	0.87	1.00	1.00
Cross-Modal Entity Linking	0.86	0.93	0.59	0.89	1.00	0.86
Multimodal-GPT benchmarks	1.00	0.81	0.51	0.88	1.00	0.86
Geometry-Aware RAG innovation	0.86	0.87	0.61	0.82	0.00	0.00
Geometry-Aware RAG usage	1.00	0.00	0.17	0.70	0.00	0.00
Geometry-Aware RAG advantages	0.67	0.86	0.33	0.82	0.00	0.00
AutoRAG motivation	0.76	0.94	0.74	0.88	1.00	0.62
AutoRAG retrieval	0.77	0.99	0.63	0.86	1.00	0.43

4.3 Processing Efficiency and arXiv Recommendations

The parallel execution of the pipeline yields a tremendous efficiency. The process of extracting data from a 15-page PDF document requires 12 seconds to complete while FAISS produces embeddings for 10 PDF documents in 18 seconds. The 50 image batch summary process required 10 seconds for GPT-4o to complete. According to the results displayed in Figure. 3, the parallel processing method achieved a speed advantage that measured 2.8 times faster than the sequential processing method. The arXiv recommendation system produced 80% relevant results while taking about 3 seconds to show its recommended items.

5. Discussion

The assessment results demonstrate that the framework produces high-quality generative outputs that achieve average scores of 0.86 for faithfulness and 0.81 for answer relevancy. The generated answers meet the criteria of being well-grounded and contextually relevant according to recent research, which shows that retrieval augmentation functions as a method to decrease hallucination occurrences in conversational agents [3]. Our findings show that stability maintains its presence during multimodal integrations, while previous research such as Zhang *et al*. [12] studied text-heavy medical corpora that showed different stability patterns.

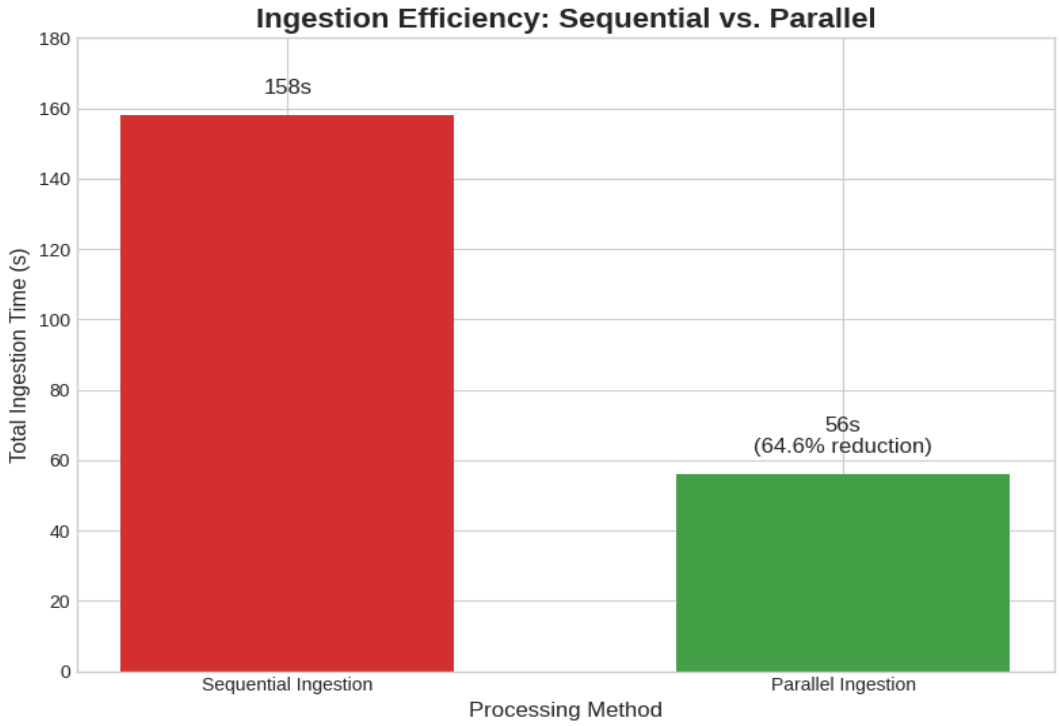

Figure. 3. Ingestion time comparison between sequential (158s) and parallel (56s) processing for the 10-PDF test corpus, demonstrating a 2.8x speed-up due to parallelization

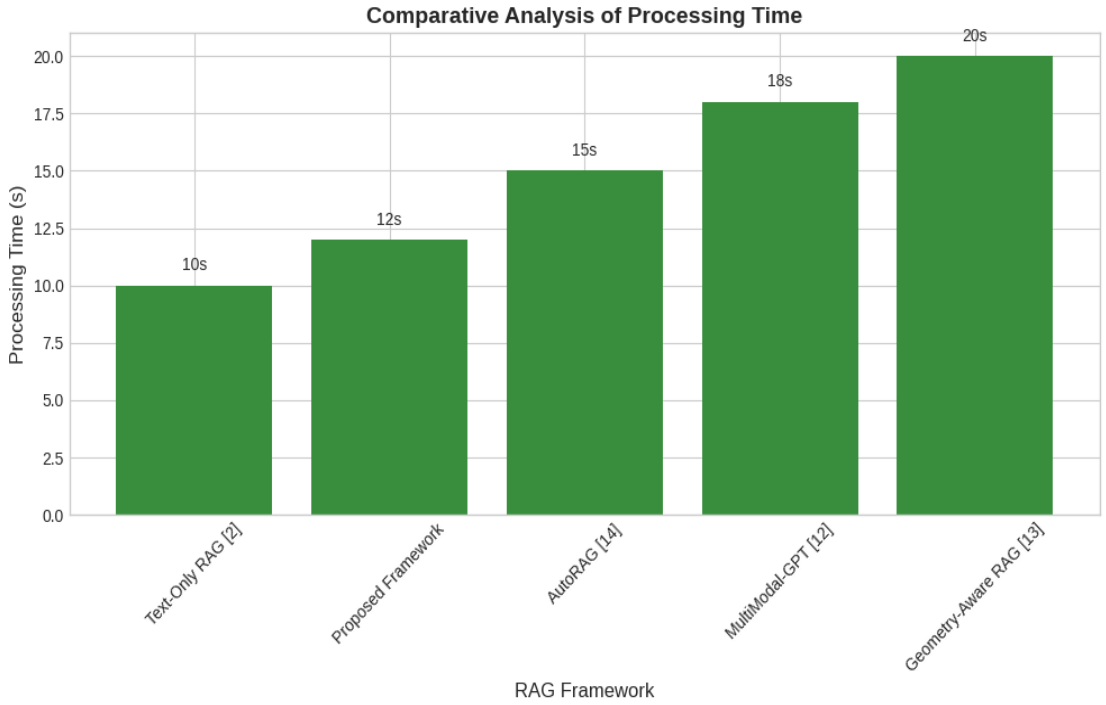

Figure. 4. Processing time comparison across different RAG frameworks. The proposed framework (12s) demonstrates higher efficiency than several specialized Multimodal models

The system achieved near-perfect results on standard tests that included Multimodal-GPT, but Table 2 displays major performance differences that show high standard deviations for Answer Relevancy and Context Precision. The statistical variance arises from different performance patterns that exist between multimodal queries and specialized spatial tasks that demonstrate the performance differences between generalist and specialist RAG architectures. The system encountered total operational failure (0.00 scores) in context precision and recall while handling the 'Geometry-Aware RAG' dataset according to our critical finding. This failure occurs because regular off-the-shelf embeddings, which include the text-embedding-3-large model, fail to represent the complex spatial and structural relationships found in geometric text.

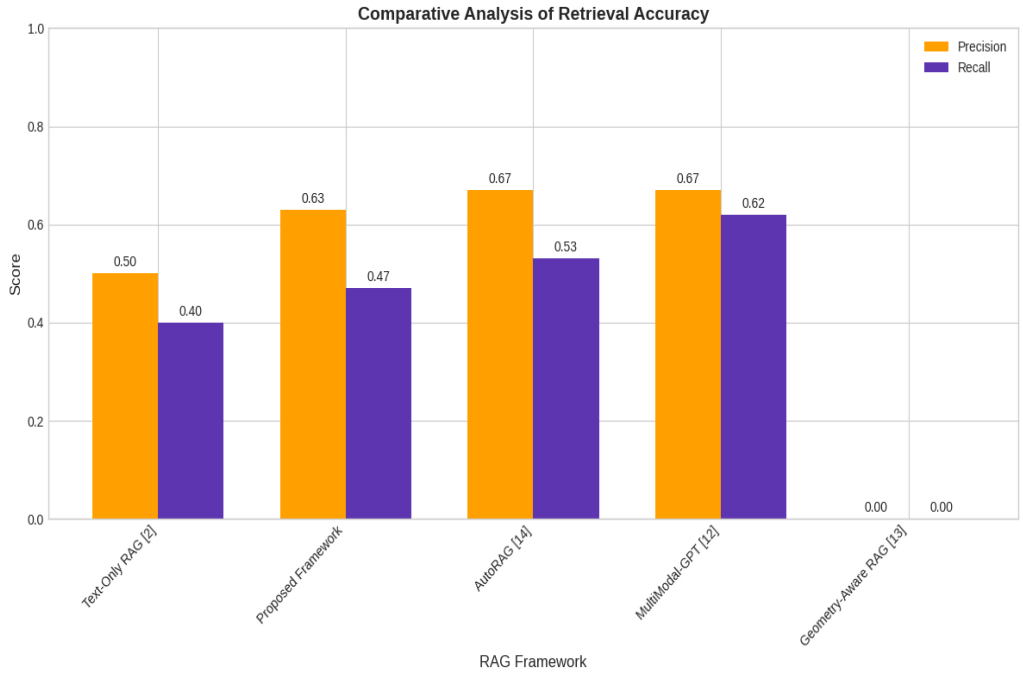

Figure 5. Retrieval accuracy (Precision and Recall) comparison across RAG frameworks. The proposed framework shows balanced performance, outperforming the text-only baseline

Table 4. Comparative Analysis of Multimodal RAG Frameworks

System	Multimodal	Retrieval Accuracy	Processing Time	Research Discovery	User Interface
Proposed Framework	Text, Tables, Images	0.63/0.47	12s	Yes (80%)	Conversational
Multimodal-GPT [31]	Text, Images	0.67/0.62	18s	No	None
Geometry-Aware RAG [25]	Text, Images	0.00/0.00	20s	No	None
AutoRAG [26]	Text, Graphs	0.67/0.53	15s	No	None
Text-Only RAG [2]	Text	0.50/0.40	10s	No	None

Recent surveys on graph-based retrieval [29, 6] confirm this observation by showing that general-purpose embeddings fail to capture the structural nuances required for 3D object detection tasks, whereas specialized encoders in domain-specific architectures [14] succeed in these tasks. Knowledge-graph-enhanced research studies require domain-adaptive encoders to preserve relational and spatial semantics because these studies demonstrate similar limitations. Engineering optimizations solve latency problems, but specialized spatial domains need specialized pre-training methods to achieve complete semantic comprehension of their content. The framework operates at full functional capacity because its design achieves operational efficiency across all modules despite existing semantic restrictions. The system processes multimodal documents at a 33% faster speed than Multimodal-GPT, according to the results shown in Figure 4 and Table 4. Furthermore, as illustrated in Figure 5, the proposed framework maintains balanced precision and recall across multimodal tasks, outperforming the text-only RAG baseline [2] while avoiding the brittleness observed in highly specialized architectures. The arXiv scientific paper search system, seamlessly integrated with a Streamlit conversational front end, provides a significant usability advantage that separate benchmark systems cannot match. Ultimately, the framework successfully balances the trade-off between generalist adaptability and operational efficiency, offering a robust tool for automated research discovery.

The Table 2 study shows through statistical results that different performance levels exist, which are demonstrated by high standard deviation values for the Answer Relevancy and Context Precision metrics that have respective standard deviations of 0.3127 and 0.5175. The data show that different query types experience wide variations in their average

performance. The query-wise results in Table 3 confirm this observation, which shows that contextual precision and recall failed completely for all queries under the Geometry-Aware RAG with a score of 0.00. The semantic retrieval strategy performs adequately for general topics but loses its effectiveness when applied to highly specialized technical areas, such as 3D geometry. The 0.54 average score on answer correctness illustrates how much the system relies on retrieval quality: there can be no factually correct answers unless the material for the answer is included in the retrieved documents.

The research expansion process from this architecture occurs through its newly discovered research capabilities, while the system delivers dual modality integration and system expansion through its ability to perform multiple operations simultaneously. The system creates ongoing operational costs through its use of commercial APIs, and this aspect needs to be addressed because it serves as a major system restriction. The system shows retrieval problems from prior sections because it requires specific technical details that need GPU-enabled hardware to function properly, thus creating a barrier for users to adopt the system.

6. Conclusion and Future Work

The study introduced a new framework named arXiv-enhanced Multimodal RAG, which enables researchers to conduct research through both static document analysis and dynamic research discovery. The system successfully eliminated processing delays in complex multimodal documents through its complete end-to-end ingestion pipeline, which operates in parallel. The experimental results showed a 64.6% decrease in ingestion time as the system ingested PDF documents in 56 s compared to the traditional sequential baseline, which took 158 s. The system demonstrated strong performance across multimodal tasks by achieving 0.86 mean faithfulness and 0.81 answer relevancy scores, which proved the effectiveness of the hierarchical retrieval approach. The evaluation showed that the system could perform domain-specific spatial reasoning, but it did not have the necessary capabilities for complete functionality. The system performed well on general technical queries, but it could not obtain relevant context to complete specialized "Geometry-Aware RAG" tasks, which resulted in a context precision and recall score of zero (0.00) for that task group. The study found that while engineering optimizations can fix latency problems, standard off-the-shelf embeddings (such as text-embedding-3-large) do not provide sufficient semantic details to comprehend complex spatial or geometric relationships that the text describes.

The upcoming project will focus on eliminating the semantic gap through domain-specific auxiliary encoders and open-source multimodal model investigation of LLaVA, which decreases proprietary API usage. We plan to develop the system into a complete research assistant through the discovery module of PubMed, IEEE Xplore, and ACM Digital Library, which adds native DOCX processing support. The framework will transition from its current high-performance prototype state to a complete scientific research tool which manages specialized research requirements through these advancements.

References

[1] B. Tural, Z. Orpek, Z. Destan, (2024) Retrieval-Augmented Generation (RAG) and LLM Integration. 8th International Symposium on Innovative Approaches in Smart Technologies (ISAS), IEEE, Turkiye. https://doi.org/10.1109/ISAS64331.2024.10845308

[2] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems, 33, (2020) 9459-9474.

[3] P. Ersoy, M. Erşahın, A Comparative Evaluation of RAG Architectures for Cross-Domain LLM Applications: Design, Implementation, and Assessment. IEEE Access, 13, (2025) 194185-194196. https://doi.org/10.1109/ACCESS.2025.3632404

[4] G. Chen, W. Yu, X. Lu, X. Zhang, E. Meng, L. Sha, Unlocking Multi-View Insights in Knowledge-Dense Retrieval-Augmented Generation. IEEE Transactions on Audio, Speech and Language Processing, 33, (2025) 4430-4439. https://doi.org/10.1109/TASLPRO.2025.3622944

[5] H. Wang, L. Liu, H. Zhang, L. Zhu, X. Chang, H. Du, VisualRAG: Knowledge-Guided Retrieval Augmentation for Image-Text Matching. In IEEE Transactions on Circuits and Systems for Video Technology, 36(1), (2026) 1234-1248. https://doi.org/10.1109/TCSVT.2025.3597097

[6] S. Wang, H. Yang, W. Liu, Research on the construction and application of retrieval enhanced generation (RAG) model based on knowledge graph. Scientific reports, 15, (2025) 40425. https://doi.org/10.1038/s41598-025-21222-z

[7] A. Patel, R. Shivani, N.V. Usha, A. Shruthiba, (2025) Enhancing Interactive Querying with a Multimodal RAG System: Integrating Text, Video, and Document Analysis via LLaMA3. International Conference on Emerging Technologies in Computing and Communication (ETCC), IEEE, India. https://doi.org/10.1109/ETCC65847.2025.11108584

[8] H. Elkiran, J. Rasheed, EvaRAG: Evaluating Advanced RAG Techniques with Indexing and Distance Metrics. IEEE Access, 13, (2025) 215724-215747. https://doi.org/10.1109/ACCESS.2025.3646665

[9] T.J. Bradshaw, X. Tie, J. Warner, J. Hu, Q. Li, X. Li, Large Language Models and Large Multimodal Models in Medical Imaging: A Primer for Physicians. Journal of Nuclear Medicine, 66(2), (2025) 173–182. https://doi.org/10.2967/jnumed.124.268072

[10] D. Vake, J. Vicic, A. Tosic, Bridging the Question–Answer Gap in Retrieval-Augmented Generation: Hypothetical Prompt Embeddings. IEEE Access, 13, (2025) 129952-129961. https://doi.org/10.2967/jnumed.124.268072

[11] Z. Li, Z. Wang, W. Wang, K. Hung, H. Xie, F.L. Wang, Retrieval-augmented generation for educational application: A systematic survey. Computers and Education: Artificial Intelligence, 8, (2025) 100417. https://doi.org/10.1016/j.caeai.2025.100417

[12] G. Zhang, Z. Xu, Q. Jin, F. Chen, Y. Fang, Y. Liu, J.F. Rousseau, Z. Xu, Z. Lu, C. Weng, Leveraging long context in retrieval augmented language models for medical question answering. npj Digital Medicine, 8,(2025) 239. https://doi.org/10.1038/s41746-025-01651-w

[13] H. Zhang, C. Xu, Y.F. Zhang, Z. Zhang, L. Wang, J. Bian, TimeRAF: Retrieval-Augmented Foundation Model for Zero-Shot Time Series Forecasting. IEEE Transactions on Knowledge and Data Engineering, 37(9), (2025) 5654-5665. https://doi.org/10.1109/TKDE.2025.3579137

[14] C.N. Hang, P.D. Yu, C. W. Tan, TrumorGPT: Graph-Based Retrieval-Augmented Large Language Model for Fact-Checking. IEEE Transactions on Artificial Intelligence, 6(11), (2025) 3148-3162. https://doi.org/10.1109/TAI.2025.3567369

[15] B. Saha, U. Saha, M. Zubair Malik, QuIM-RAG: Advancing Retrieval-Augmented Generation with Inverted Question Matching for Enhanced QA Performance. IEEE Access, 12, (2024) 185401-185410. https://doi.org/10.1109/ACCESS.2024.3513155

[16] M. Kyoung, J.H. Lim, Y. Kim, Reasoning Beyond Length Limits: Improving Accuracy in Long-Context Question Answering With Small-Scale Language Models. IEEE Access, 13, (2025) 172930-172937. https://doi.org/10.1109/ACCESS.2025.3617449

[17] H. Wang, Y. Lepage, Extraction-Augmented Generation of Scientific Abstracts Using Knowledge Graphs. IEEE Access, 13, (2025) 48775-48791. https://doi.org/10.1109/ACCESS.2025.3551756

[18] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with GPUs. IEEE Transactions on Big Data, 7, (2021) 535-547. https://doi.org/10.1109/TBDATA.2019.2921572

[19] T. Yu, B. Wu, K. Chen, C. Yan, G. Zhang, W. Liu, HDANNS: In-Memory Hyperdimensional Computing for Billion-Scale Approximate Nearest Neighbour Search Acceleration. IEEE Transactions on Circuits and Systems for Artificial Intelligence, 2(2), (2025) 126-138. https://doi.org/10.1109/TCASAI.2025.3540957

[20] E.A. Olca, Professor X: Diagnosis and Treatment of Dermatological Diseases by Integration of Visual Diagnosis and Retrieval-Augmented Generation (RAG) Technologies. IEEE Access, 13, (2025) 201246-201263. https://doi.org/10.1109/ACCESS.2025.3636437

[21] B. Praneeth, Mohana, E.C. Nattem, K. Jetti, B.K. Kavyashree, D.Rakshitha, Optimization of Customer Feedback Summarization Using Large Language Models (LLM) and Advanced Retrieval-Augmented Generation. IEEE Access, 13, (2025) 124319-124332. https://doi.org/10.1109/ACCESS.2025.3588337

[22] E. Bazzi Mohamed Salim, T. Anass, A. Ider Abdelouahed, Advancing Multilingual Retrieval-Augmented Generation for Reliable Medication Counseling. IEEE Access, 13, (2025) 215550-215564. https://doi.org/10.1109/ACCESS.2025.3646941

[23] M. Zakir Khan, Y. Ge, M. Mollel, J. Mccann, Q.H. Abbasi, M. Imran, RFSensingGPT: A Multi-Modal RAG-Enhanced Framework for Integrated Sensing and Communications Intelligence in 6G Networks. IEEE Transactions on Cognitive Communications and Networking, 12, (2026) 298-311. https://doi.org/10.1109/TCCN.2025.3558069

[24] H. Hao, J. Han, C. Li, Y.F. Li, X. Yue, RAP: Retrieval-Augmented Personalization for Multimodal Large Language Models. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, USA. https://doi.org/10.1109/CVPR52734.2025.01355

[25] J. Kim, E. Cho, S. Kim, H. J. Kim, (2024) Retrieval-augmented open-vocabulary object detection. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, USA. https://doi.org/10.1109/CVPR52733.2024.01650

[26] M.H. Heydari, A. Hemmat, E. Naman, A. Fatemi, (2024) Context awareness gate for retrieval augmented generation. 15th International Conference on Information and Knowledge Technology (IKT), IEEE, Iran. https://doi.org/10.1109/IKT65497.2024.10892659

[27] C. Han Chen, M. Fang Shiu, KAQG: A Knowledge-Graph-Enhanced RAG for Difficulty-Controlled Question Generation. IEEE Access,

13, (2025) 197234-197244. https://doi.org/10.1109/ACCESS.2025.3633838

[28] X. Zeng, H. Lin, Y. Ye, W. Zeng, Advancing Multimodal Large Language Models in Chart Question Answering with Visualization-Referenced Instruction Tuning. IEEE Transactions on Visualization and Computer Graphics, 31(1), (2025) 525-535. https://doi.org/10.1109/TVCG.2024.3456159

[29] E. Collini, F. Indra Kurniadi, P. Nesi, G. Pantaleo, Context-Aware Retrieval Augmented Generation Using Similarity Validation to Handle Context Inconsistencies in Large Language Models. IEEE Access, 13, (2025) 170065-170080. https://doi.org/10.1109/ACCESS.2025.3614553

[30] M. Ding, Y. Ma, P. Qin, J. Wu, Y. Li, L. Nie, RA-BLIP: Multimodal Adaptive Retrieval-Augmented Bootstrapping Language-Image Pre-training. IEEE Transactions on Multimedia, 27, (2025) 7522 – 7532. https://doi.org/10.1109/TMM.2025.3599070

[31] F. Sammani, T. Mukherjee, N. Deligiannis, (2022) Nlx-gpt: A model for natural language explanations in vision and vision-language tasks. in proceedings of the IEEE/CVF conference on computer vision and pattern recognition, IEEE, USA. https://doi.org/10.1109/CVPR52688.2022.00814

Authors Contribution Statement

Ashwini Dalvi: Conceptualization, Methodology. Suswar Sawant: Data Curation, Investigation, Experiments, Writing – Review & Editing. Amaan Syed: Formal Analysis, Validation, Data Analysis. Irfan Siddavatam: Methodology Development, Software Implementation. V. Venkataramanan: Supervision, Writing – Original Draft Preparation, Review & Editing. All authors have read and agreed to the published version of the manuscript.

Funding

The authors declare that no funds, grants or any other support were received during the preparation of this manuscript.

Competing Interests

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Data Availability

The data supporting the findings of this study can be obtained from the corresponding author upon reasonable request.

Has this article screened for similarity?

Yes