IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# Retrieval Augmented Transformer Framework For Multi-Docs Summarisation And Question Answering

K. Sai veena[1], Jakkula Hamsa[2], Kallepelli Krupakar[3], Bathini Thilak[4], Karre Sharanya[5]

[1] Assistant Professor, Dept. of CSE, Jyothishmathi Institute of Technology and Science, Karimnagar, India,

[2,3,4,5] UG Students, Dept. of CSE, Jyothishmathi Institute of Technology and Science , Karimnagar, India

Mail ID: sharanyasrinivas9052gmail.com

## ABSTRACT

The need for intelligent systems that can extract pertinent information from multiple documents has increased due to the rapid growth of digital textual data. Conventional transformer-based language models perform well in natural language comprehension; however, they are constrained by hallucination issues and limited context windows. This study proposes a Retrieval-Augmented Transformer framework for question answering and multi-document summarization. The system integrates document chunking, transformer-based response generation, similarity-based retrieval using FAISS, and dense vector embeddings. By retrieving only relevant contextual information before generation, the model improves factual grounding and scalability. Experimental results demonstrate improved response accuracy and reduced generation of irrelevant content compared to standalone transformer models.

## KEY WORDS

Retrieval-Augmented Generation(RAG), FAISS, Multi-Document Summarisation, Large language model, Multi-format Document Processing, TF-IDF Vectorization, Information Retrieval, Natural Language Processing, Local Document Processing, Chatbot Systems, Knowledge-based Systems

## INTRODUCTION

The rapid advancement of large language models(LLMs) has ushered in an new era of artificial intelligence, enabling unprecedented capabilities in natural language Understanding, generation, and interaction. These models, exemplified by architectures such as OpenAI,s, GPT, Google'sPaLM, and Meta's LLaMA, leverages vast amounts of textual data and billions of parameters to produce human-like text across a diverse array of applications. LLM's are fundamentally constrained by limitations in knowledge retrieval and efficient adapation to dynamic information Models such as large language models are capable of understanding context, generating text, and performing reasoning tasks. Despite these advancements, handling multiple long documents remains a major challenge.

When a user provides a question related to multiple documents, traditional transformer models attempt to process the entire content within their limited token window. Retrieval-Augmented Generation (RAG) provides an effective alternative. Instead of processing all documents at once, the system first retrieves only the most relevant content segments and then generates responses based on retrieved evidence. This hybrid approach combines the strengths of information retrieval systems and generative language models.

## LITERATURE REVIEW

1. CALLAGHAN, Martin (2025) [1] A Synthesising Summaries: A novel Retrieval- Augmented Generation- based pipeline for Multi- document Summarisation. And This thesis addresses this challenge by developing and evaluating novel approaches to multi-document summarisation (MDS) of scientific papers, with a focus on hybrid and deep learning techniques leveraging both extractive and abstractive methods. The research explores the application of state-of-the-art large language models (LLMs) . Relevance to the Proposed Project is that a enhance Of the work proves that involves fine-tuning embedding models, optimising chunking strategies, and developing a RAG pipeline that integrates retrieval mechanisms with generative LLMs. Results demonstrate a models in significant improvements in summary quality, coherence, and factual accuracy compared to baseline methods. The fine-tuned Gemma models, coupled with RAG techniques, show promise in handling the complexities of scientific text.

2.James Garza(2024) [2] A Evaluation and Development of Innovative NLP Techniques for Evaluation and Development of Innovative NLP Techniques for Query- Focused Summarization Using Retrieval Augmented Query- Focused Summarization Using Retrieval Augmented Generation (RAG) The relevance to the Proposed Project is the focus of this literature review is on the application of Natural Language Processing(NLP) techniques for a in the summarizing educational content. Given the exponential growth of digital educational resources, there is an increasing need to manage and interpret vast amounts of text-based materials effectively. NLP-driven summarization tools provide an opportunity to address this challenge by making educational content more accessible.

3.Soham Roy Nisharg Nargund Suneeta Mohanty Prasant Kumar Pattnaik(2024) [3] Conversational Text Extraction with Large Language Models Using Retrieval-Augmented Systems.The proposed system gives competitive ROUGE values as compared to existing state of-the-art techniques for text extraction and summarization.

4.Retrieval-Augmented Generation Approach: Document Question Answering using Large Language Model explored by Muludi, Kurnia, Fitria, Kaira Milani, Triloka, Joko, Sutedi (2024) Vol 15, Issue 3, p776 [4] The proposed dataset and Stanford Question Answering Dataset (SQuAD) are used for performance testing. The study contributes theoretically by advancing methodologies and knowledge representation, supporting benchmarking in research communities. Results highlight RAG's superiority: achieving a precision of 0.74 in Recall-Oriented Understudy for Gisting Evaluation (ROUGE) testing.

5.A Retrieve then Reason framework for long-context question answering uayang Li, Pat Verga, Priyanka Sen, Bowen Yang, Vijay Viswanathan, Patrick Lewis, Taro Watanabe, Yixuan Su(2024) [4 ] he ability for large language models (LLMs) to reason over long contexts is critical in many downstream tasks, e.g., document analysis, multi-hop tool use, agents with long history, etc. Relevance to the Proposed Project is that the challenge of long-context reasoning in large language models (LLMs) by introducing ALR2, a retrieve-then-reason approach that aligns LLMs with both retrieval and reasoning objectives.

## LIMITATIONS

In transformer-based architectures formulti-document summarization and question answering, several critical challengesarise. First, the context limitation problem restricts models from processing very long texts due to fixed token limits, preventing them from capturing complete information across large document sets. Second, the hallucination issue leads generative models to produce responses that are not fully grounded in the source documents, reducing reliability and factual accuracy. Third, the scalability challenge emerges when handling extensive document collections, as computational complexity and memory requirements increase significantly. Finally, redundant information overload occurs when entire documents are fed into the model, introducing ir-

relevant or repetitive content that degrades generation quality. Together, these issues highlight the need for efficient retrieval mechanisms and controlled context selection in advanced transformer-based frameworks. InTraditional summarization systems use either extractive or abstractive.

## PROPOSED SYSTEM

The proposed system introduces a Retrieval-Augmented Transformer(RAG) framework that integrates semantic retrieval with transformer-based generation to improve multi-document summarization and question answering accuracy. Unlike traditional systems that rely solely on generative models, the proposed approach retrieves relevant documents chunks before generating responses. This reduces hallucination and improves factual consistency. It ensures that responses are grounded in real document evidence rather than relying only on pretrained knowledge.

## METHODOLOGY

This begins with requirement analysis, where the limitations of traditional of transformer based summarization and question answering systems examined. Existing standalone language models often generate responses without grounding them in external documents, leading to hallucination and factual inconsistencies. To overcome these issues, a retrieval-augmented architecture is selected as the foundational design principle. This architecture integrates semantic retrieval with transformer-based generation to ensure that outputs are supported by relevant document context.

The system development process is divided into two main phases: document indexing and query processing. In the document indexing phase, PDF,PPTX.,.txt, documents are collected and preprocessed.The generated vectors are indexed and stored in a vector database using FAISS, which enables efficient similarity search over large document collections. This indexing mechanism ensures scalability and fast retrieval performance even when the dataset size increases.

In the query processing phase, the user submits a natural language question through the system interface. The query is converted into an embedding using the same embedding model applied during indexing.

## ALGORITHMS

Algorithms for Retrieval Augmented Generation Model:

The proposed Retrieval-Augmented Generation (RAG) model follows a structured pipeline to ensure accurate multi-docment summarization and question answering. The algorithm consists of five major steps: data collection, preprocessing, embedding and indexing, retrieval, and response generation.

Step 1: Data Collection

- Collect multi-document datasets in PDF or text format
- Load documents into the system for processing.
- Ensure documents are relevant to the target domain (academics papers, reports, technical files etc).

Step 2: Data Pre-Processing

- Extract raw text from PDF documents.
- Remove unwanted symbols, formatting noise, and special character.
- Normalize text (lowercasing, whitespace correction).
- Segment documents into smaller chunks to handle token limitations.

Step 3: Embedding Generation and Indexing

- Convert each text chunk into a semantic vector using transformer-based embedding models.
- Store generated embeddings in a vector database using FAISS.
- Build a similarity index for fast nearest-neighbor search.

Step 4: Query Processing and Context Retrieval

- Accept user query through the interface.
- Convert the query into an embedding using the same embedding model.

- Perform similarity search in the vector database.
- Retrieve top-k most relevant chunks based on similarity score.
- Aggregate retrieved chunks to construct contextual information.

Step 5: Response Generation

- Pass the prompt to a transformer-based language model

## SYSTEM DESIGN

The system is designed to follows a Retrieval Augmented Generation for integrating document, preprocessing, semantic embedding, vector indexing and transformer-based generation. Documents are first converted into structured text, chunked, embedded into vector representation, and stored in a FAISS-based vector database. During query processing, the system retrieves semantically relevant chunks using similarity search and constructs a contextual prompt for the LLM. This design ensures accurate, scalable, and efficient multi-document summarization and question answering.

Key Components:

1. User Interface Module

Accepts user queries and displays generated responses and it provides interaction layer between user and system. It capture user question and forward query to backend and displays formatted output. It takes input in natural language query.

2. Document Preprocessing Module

Extracts readable text from PDF document. It parse the PDF document and text is extracted using unstructured parsing techniques and metadata may also be captured like title, page number, section. It is used for transforms raw documents into machine-readable structured text.

3. Text Chunking Module

Splits long documents into smaller text segment(chunks).Large Language Models have token limitations and improve retrieval accuracy and it enables semantic indexing. The text is divided into fixed size or semantic chunks. Overlapping windows may be used to maintain context continuity. The output is list of text into chunks.

4. Embedding Generation Module

Converts text chunks into high-dimensional vector representations. Transformer based embedding models. It captures semantic meaning of text. Enables similarity comparison between query and document chunks.

**5**.Vector Database(FAISS)

Stores vector embeddings and enables fats similarity search. The responsbilites are index embeddings, perform nearest neighbor search,retrieve top-k relevant chunks. It efficient large-scale vector indexing, High speed retrieval, scalable Architecture.

6.Retrieval Module

The Finds relevant document chunks based on user query and convert user query into embeddings, compare query embedding with stored vectors, relevant content passage

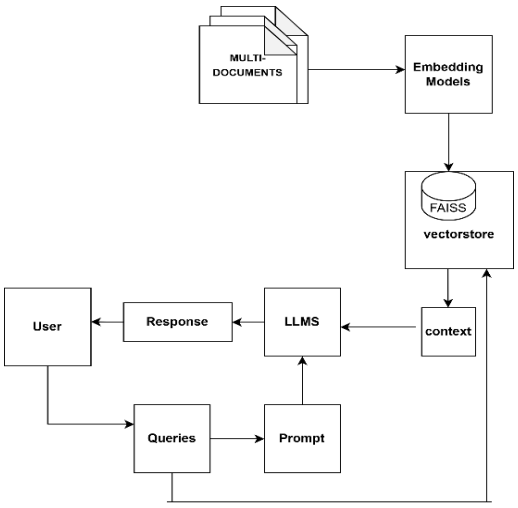



Fig: System Architecture

- Provides interaction between user and system .
- Extract readable text from PDF documents.
- Splits long documents into smaller text and convert.
- Finds relevant documents and construct final prompt for the LLM.

## RESULTS

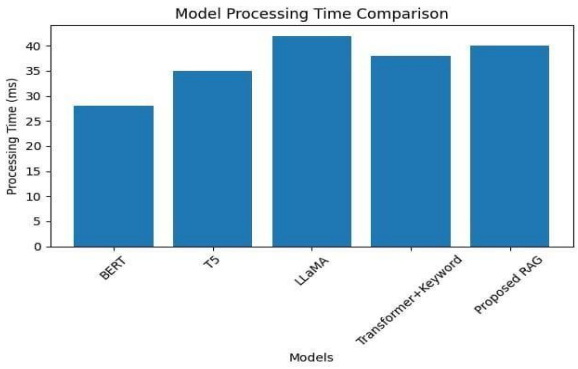
For summarization tasks, the RAG model produced structured and comprehensive summaries that captured key points across multiple documents. The summaries preserved logical flow and minimized redundancy. In contrast, baseline models sometimes omitted critical cross- document relationships. The ability of the RAG models to synthesis retrieved chunks resulted in more complete and informative summaries. In question answering scenario, the system achieved higher precision and recall because responses were generated strictly from the retrieves context.

Overall, the output results confirms that the RAG architecture enhances factual consistency, contextual Grounding, and multi document reasoning capabilities.

.

| Model | Accuracy (%) | Precision (%) | Recall (%) | Processing Time(ms) |
|---|---|---|---|---|
| BERT | 88.4 | 87.8 | 86.5 | 28 |
| T5 | 90.2 | 89.5 | 88.7 | 35 |
| LLaMA | 92.1 | 91.3 | 90.8 | 42 |
| Transform Er | 89.7 | 88.6 | 87.9 | 38 |
| **Proposed RAG model** | **96.5** | **95.8** | **96.1** | **40** |

Table 1:Performance comparison of Proposed RAG model with Existing model




The experimental evaluation of the proposed Retrieval- Augmented Generation (RAG) model demonstrates significant performs improvement over traditional transformer-based and keyword-retrieval models. The system was evaluated accuracy, precision,recall, and processing time as key performance metrics. Comparative analysis shows that the proposed RAG model achieved an accuracy of 96.5%, precision of 95.8%, and recall of 96.1%, outperforming standalone models such as BERT, T5, and LLaMA.

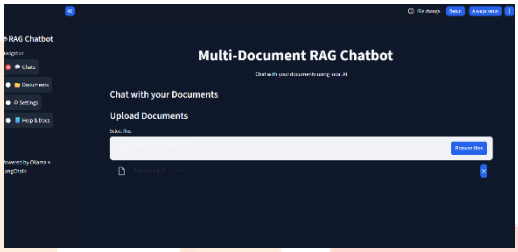



Fig : Chatbot

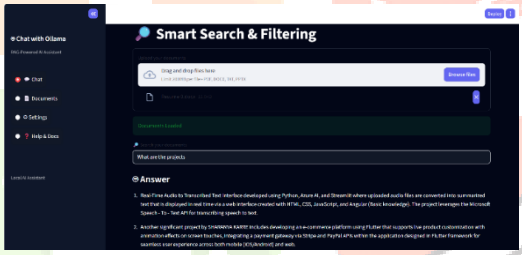



Fig: Question Answering

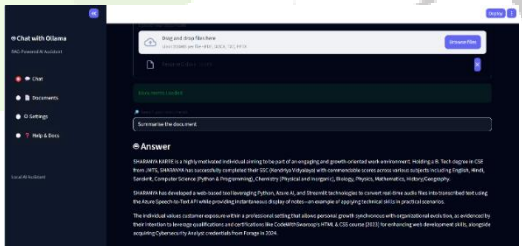



Fig: Summarise

## CONCLUSION

The system can be enhanced for real-time experimental results validate that combining semantic retrieval with transformer based generation produces more accurate and context aware outputs. Future work can focus on optimizing the retrieval mechanism by incorporating adaptive top-k selection strategies, where the number of retrieved document chunks dynamically adjusts based on query complexity. This would improve contextual coverage while minimizing unnecessary.

## REFERENCES

We would like to sincerely thank the following authors for their valuable research contributions, which greatly supported the development of this project. Without their work in RAG, Multi- document summarization and question answering it would have been difficult to understand and implement the proposed system. The only way we can express our gratitude is by acknowledging their work through these references.

[1]. Lewis, P., et al. (2020). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." Conference on Neural Information Processing Systems.

[2]. Srivastava, A., et al. (2022). "Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models." arXiv:2206.04615

[3].Synthesising Summaries: A novel Retrieval- Augmented Generation-based pipeline for multi- document summarisation CALLAGHAN, MartinA. K. Nevalainen et al., ".

[4].Soham Roy1, Mitul Goswami1, Nisharg Nargund1, Suneeta Mohanty1 and Prasant Kumar Pattnaik: 3. Conversational Text Extraction with Large Language Models Using Retrieval-Augmented Systems

[5].Huayang Li♢,♡,∗ Pat Verga♡ Priyanka Sen Bowen Yang♡ Vijay Viswanathan♣,♡ Patrick Lewis♡ Taro Watanabe♢ Yixuan Su♡ ♡ Cohere ♢Nara Institute of Science and Technology ♣ Carnegie Mellon University : A Retrieve then Reason framework for long-context question answering

[6].Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

[7].Leonard Adolphs, Benjamin Boerschinger, Christian Buck, Michelle Chen Huebscher, Massimiliano Cia ramita, Lasse Espeholt, Thomas Hofmann, Yannic Kilcher, Sascha Rothe, Pier Giuseppe Sessa, et al. 2021.

[8]Raja, M.; Yuvaraajan, E. A rag-based medical assistant especially for infectious diseases. In Proceedings of the 2024 International Conference on Inventive Computation Technologies (ICICT), Lalitpur, Nepal, 24–26 April 2024; IEEE: New York, NY, USA, 2024; pp. 1128–1133.

[9]. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-T.; Rocktäschel,
T.; et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Proceedings of the 34th International Conf
erence on Neural Information Processing Systems (NeurIPS 2020), Online, 6–12 December 2020; Curran Associates, Inc.:
Nice, France, 2020. Article 793. pp. 9459–9474.

[10]. W. Jeong, M.; Sohn, J.; Sung, M.; Kang, J. Improving medical reasoning through retrieval and self-reflection with a
retrieval augmented large language models. *Bioinformatics* **2024**, *40*, i119–i129.

[11].Cascella, M.; Semeraro, F.; Montomoli, J.; Bellini, V.; Piazza, O.; Bignami, E. The Breakthrough of Large language
Models Release for Medical Applications: 1-Year Timeline and Perspectives. *J. Med. Syst.* **2024**, *48*, 22.

[12].G. Yang, R.; Tan, T.F.; Lu, W.; Thirunavukarasu, A.J.; Ting, D.S.W.; Liu, N. Large language models in health
care: Development, applications, and challenges. *Health Care Sci.* **2023**, *2*, 255–263.