International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 06 | Jun 2026 www.irjet.net p-ISSN: 2395-0072

CONTENT SUMMARIZER AND DOCUMENT GPT USING RAG

[1]Shivam Kumar Singh, [2]Ashish Kumar, [3]Dr. Meenakshi Sharma

[123]Department of Computer Science and Engineering Galgotias University Greater Noida,India

---

ABSTRACT-*The exponential growth of digital textual data across domains such as education, healthcare, law, and scientific research has created significant challenges in efficient information extraction, summarization, and document-centric question answering. Traditional document processing techniques rely heavily on keyword-based retrieval and extractive summarization, which lack semantic understanding and struggle with long, complex documents. Although recent advances in Large Language Models have demonstrated impressive capabilities in natural language understanding and generation, their standalone deployment is constrained by hallucination, lack of transparency, and poor grounding in domain-specific or private document collections. Furthermore, fine-tuning large models on continuously evolving datasets is computationally expensive and impractical. To address these limitations, this paper proposes a Retrieval-Augmented Generation based framework for content summarization and document-grounded conversational interaction. By integrating semantic retrieval with generative modeling, the proposed system produces accurate, context-aware summaries and reliable answers while maintaining strong alignment with source documents.*

Keywords-Retrieval-Augmented Generation (RAG), Document Summarization, Document GPT, Large Language Models (LLMs), Semantic Search, Vector Database, Natural Language Processing.

1. INTRODUCTION

The rapid digitization of information in the modern era has resulted in an overwhelming volume of unstructured textual data. Across academic, medical, legal, and industrial domains, organizations generate and store vast collections of documents on a daily basis. While this abundance of information offers unprecedented opportunities for knowledge discovery, it simultaneously creates significant challenges related to information overload. Extracting relevant insights from lengthy documents, identifying key concepts, and answering user-specific queries in a timely manner have become increasingly complex tasks. Traditional document processing systems have attempted to address these challenges through information retrieval techniques such as keyword matching, inverted indexing, and ranking algorithms. While these methods are computationally efficient, they largely depend on surface-level textual features and fail to capture deeper semantic relationships. Similarly, early summarization systems rely on extractive approaches that select important sentences directly from the source text. Although extractive summaries preserve original wording, they often lack coherence, logical flow, and contextual understanding. The emergence of Large Language Models has transformed natural language processing by enabling machines to hallucination. Furthermore, LLMs lack inherent access to private or domain-specific document repositories unless explicitly trained or fine-tuned, which introduces scalability and cost challenges. These limitations highlight the need for hybrid approaches that combine the strengths of retrieval-based systems with the generative capabilities of LLMs. Retrieval-Augmented Generation has emerged as a promising paradigm to address this need by grounding generated responses in retrieved document evidence. This paper explores the design and implementation of a RAG-based content summarizer and document-grounded conversational system aimed at improving accuracy, transparency, and usability.

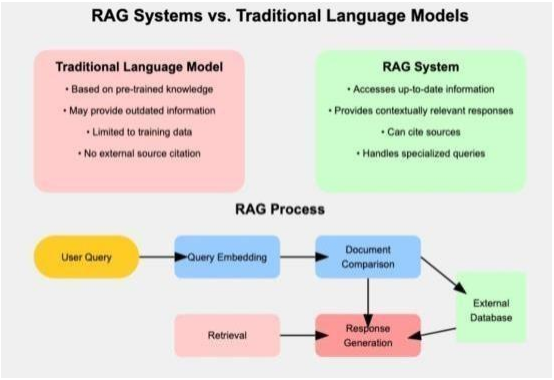

Fig 1 Information Overload in Traditional Document Processing System

2. LITERATURE REVIEW

Evidently, document summarization and document centric question answering are some of the most popular areas of research in natural language processing during the last few decades. The initial studies in this field were mainly extraction summarization methods, which determine and pick out significant sentences or phrases right out of the source text. Sentence importance was typically estimated using such statistical techniques as term frequency-inverse document frequency, sentence position heuristics and lexical similarity measures. Although such methods are computationally efficient and easy to implement, they do not have semantic content and are subject to giving fragmented, redundant and poorly structured summaries. Later developments offered machine learning-based summarization methods, such as supervised and unsupervised models that are trained on labeled data to learn the importance of sentences. Graph based systems like Text Rank and LexRank represented the documents as graphs with sentences as nodes and relationship of similarity represented by edges. These strategies made the summary coherence somewhat better, but they were essentially extractive and failed to produce new sentences and paraphrase meaningfully. The invention of neural network models was a technical breakthrough in the direction of abstractive summarization. Attention-based sequence-to-sequence models allowed systems to acquire the semantic understanding of text and produce more fluent summaries. Although the readability was enhanced, the initial neural models were not quite capable of handling long texts and factual errors were common. Transformer-based architectures went a step further to contribute to better performance through parallel processing and better contextual modeling. Designed models like encoder-decoder transformers had shown good performance on benchmark summarization data, but were constrained by limited input length.

SOTA performance in a diverse set of natural language processing tasks such as summarization, question answering and conversational systems has recently been demonstrated by large language models trained on massive corpora. Such models use the self-attention processes to extract long- range dependencies and produce very fluent text. Nevertheless, various reports have indicated that the members of the LLMs often produce hallucinated messages, especially in cases where query questions necessitate strict factual basis. This is a serious limitation when it comes to their use in sensitive fields like healthcare, law and even scientific research where traceability and accuracy are paramount.

In order to reduce hallucination and enhance factual accuracy, scholars have sought ways of basing language models on external body of knowledge. Knowledge-based techniques augmented with knowledge or retrieval-based techniques use either structured or unstructured data in the generation process. One of them has become Retrieval-Augmented Generation. The RAG-based systems access the desired document fragments in an external corpus and condition the output of the language effective in factual correctness and less hallucinatory than standalone generative models.

Various studies have used RAG-based architectures to open-domain question answering, presenting better answers and evidence grounding. RAG has been investigated by other studies to be applied to domain-specific problems, including analysis of biomedical literature and search of enterprise documents. These papers demonstrate the usefulness of dense vector retrieval in conjunction with generative modeling. Still, they also outline the challenges that have no solutions based on the best chunking strategy, the relevance of retrievals, latency, and scalability.

Even in light of these developments, relatively small amount of research has been conducted to create unified RAG-based systems which provide both content summarization, and interactive document-grounded conversational interactions within a single system. Furthermore, the available literature tends to test systems based on limited criteria without the context of document complexity in the real world and constantly changing data. These shortcomings drive the current study that seeks to come up with an all-encompassing RAG-based content summarizer and document-guided conversations system that can overcome the weaknesses of the previous methods.

Table 1: Literature Comparison

Study	Method	Strength
TF-IDF based	Extractive	Simple
Neural Seq2Seq	Abstractive	Fluent
LLMs	Generative	Context aware
Proposed RAG	Hybrid	Grounded

3. PROBLEM STATEMENT

The exponential growth of digital textual data across domains such as education, healthcare, law, and research has made efficient information extraction, summarization, and document-centric question answering a critical challenge. Most existing document processing systems rely on traditional information retrieval or extractive summarization techniques, which are limited in their ability to understand deep semantic long documents, and adapt to diverse user queries. Recent advances in Large Language Models (LLMs) have significantly improved natural language understanding and generation capabilities. However, LLMs used in isolation suffer from major limitations, including hallucination, lack of transparency, and inability to reliably ground responses in domain-specific or private documents. These limitations reduce their applicability in high-stakes or research- oriented environments. Where factual accuracy and traceability to source documents are essential.

Furthermore, directly fine-tuning LLMs on large or continuously evolving document collections is computationally expensive, time- consuming, and often impractical. This creates a gap between the powerful generative capabilities of LLMs and the need for accurate, document- grounded summarization and question answering systems.

Retrieval-Augmented Generation (RAG) has emerged as a promising paradigm to bridge this gap by combining information retrieval techniques with generative language models. By retrieving relevant document segments and conditioning the generation process on them, RAG-based systems can significantly reduce hallucinations and improve factual correctness. However, designing an efficient and scalable RAG-based architecture for content summarization and interactive document querying remains an open research problem, particularly with respect to chunking strategies, retrieval relevance, response coherence, and system performance. Therefore, this research addresses the problem of developing an effective Content Summarizer and Document-Grounded Conversational System using Retrieval-Augmented Generation, capable of producing concise summaries, accurate, context-aware responses while maintaining strong alignment with the original source documents.

Table2.Key Challenges and RAG-based Solution

Challenge	Existing Systems	RAG-based Solution
Hallucination	High	Reduced
Long documents	Poor handling	Chunking
Scalability	Retraining	Vector DB
Traceability	Missing	Document grounding

4. RETRIEVAL-AUGMENTED GENERATION OVERVIEW

Retrieval-Augmented Generation is a hybrid framework that combines information retrieval with text generation. Instead of relying solely on the internal knowledge of an LLM, a RAG-based system retrieves relevant documents or document fragments and incorporates them into the generation process.

The retrieval component ensures that only relevant and contextually appropriate information is passed to the generative model. The generation component then produces responses based on this retrieved evidence, reducing the likelihood of hallucination and improving factual accuracy. RAG systems are particularly well-suited for document-centric tasks, as they allow models to scale to large document collections while maintaining strong alignment with source material.

4.1 Proposed System Architecture

The proposed content summarizer and document-grounded conversational system consists of several interconnected modules.

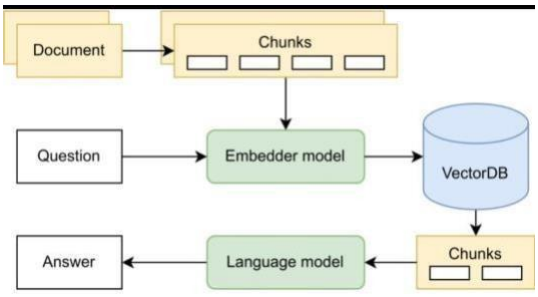

Fig 2.Architecture of the Proposed RAG- based Content Summarizer

4.2 Document Ingestion and Preprocessing

Documents are first collected from various sources and converted into a standardized textual format. Preprocessing steps include noise removal, normalization, tokenization, and segmentation. Documents are divided into smaller, semantically coherent chunks to ensure efficient retrieval and manageable context lengths.

4.3 Semantic Embedding and Indexing

Each document chunk is transformed into a dense vector representation using semantic embedding techniques. These embeddings capture contextual meaning and enable similarity-based retrieval. The vectors are stored in a vector database optimized for fast nearest-neighbor search.

4.4 Retrieval Module

When a user submits a query or summarization request, the system retrieves the most relevant document chunks based onsemantic similarity. This retrieval process ensures that only pertinent information is considered during generation stored in a vector database optimized for fast nearest-neighbor search.

4.5 When a user submits a query or summarization request, the system retrieves the most relevant document chunks based onsemantic similarity. This retrieval process ensures that only pertinent information is considered during generation.

4.6Generative Module

The retrieved document chunks are provided as context to a language model, which generates summaries or answers constrained by the retrieved content. This grounding mechanism significantly reduces hallucinations and improves response reliability.

4.7 Output Generation and Validation
4.8
The final output is presented to the user along with optional references to source documents, enhancing transparency and trustworthiness.

Table 3.Functional Modules of the Proposed System

Module	Description
Ingestion	Document Upload
Preprocessing	Cleaning
Chunking	Context Segmentation
Embedding	Semantic Vectors
Retrieval	Similarity Search

5. PROPOSED METHODOLOGY

The proposed methodology is meant to generate an effective, scaled, and dependable content summarization and document-based conversational service through Retrieval-Augmented Generation (RAG) paradigm. The systematicity of the methodology is a modular one that involves document preprocessing, semantic retrieval, and controlled text generation to guarantee the factual accuracy and contextuality. The collection and ingestion of documents will occur according to the definitions of the documents and their digitization.

5.1 Document Collection and Standardization

The initial phase of the methodology is the gathering of documents of various sources in the form of PDFs, text files, and research articles as well as legal documents and research in the domain-specific reports. The documents cannot be in the same structure, format and quality thus they are transformed into a standardized textual form. This is done to allow compatibility with the downstream processing modules as well as to allow the same compliance with the processing of various types of documents.

5.2 Text Preprocessing

After ingestion, preprocessing of the documents is done to increase the text quality and lessen noise. This involves elimination of meaningless symbols, text normalization, stop words and sentence division. Preprocessing is important to enhance retrieval accuracy as it is important to make sure that semantic embeddings reflect meaningful content and not formatting artifacts.

5.3 Document Chunking Strategy
To get around input length constraints of language models, long documents are decomposed into small, semantically coherent documents to enhance retrieval granularity. The logical divisions upon which chunking is conducted include paragraphs or sections, and overlap is regulated to maintain contextual coherence. An efficient approach to chunking makes sure the key information is not lost or divided, which directly relates to the relevance of retrieval and generation of information.

5.4 Semantic Embedding Generation.

Semantic embedding techniques are used to convert each piece of documention dense representation as a vector. Such embeddings

encode sentence context and semantic lexicon relations. In contrast to the classical cigarette features of a keyword, a dense embedding to keyword allows similarity matching of similarities of meaning, not the direct lexical overlap of query and document words, which means that the system can find a related content even when the query is expressed in different words than the document itself.

6. DISCUSSION AND ANALYSIS

The suggested RAG-based solution points to a number of drawbacks of document processing and standalone language models solutions. The system balances the weaknesses of each of the two paradigms by separating retrieval and generation, which enhances the advantages of each paradigm. The hallucination has been reduced, which is one of the greatest benefits realized. Because the generative model derives on retrieved document evidence but not internal parametric knowledge on its own, the chances of producing unsubstantiated or artificial information is significantly low. This renders the system more factual and document- centric.

There is also an improvement in the flexibility of the system. The proposed method can also implement new or updated documents by merely updating the vector database unlike the fine-tuned language models that need retraining every time that document collections vary. This feature is very cost effective in both calculation and maintenance. On the scalability front, the scalability of the system is enabled by the use of semantic embeddings and vector indexing that enables the system to scale to large documents repositories. Nevertheless, the size of a chunk, the quality of embedding, and the threshold of retrieval are some of the factors that affect performance. Poor chunking can cause context loss whereas excessive chunking can cause the model to exceed its input capacity or include extraneous information.

Another relevant factor is latency. Retrieval also introduces a new step to the inference pipeline; however, the system as a whole is not impractical thanks to effective indexing and similarity search facilities. More optimization may be provided by using caching techniques and the approximation nearest-neighbor search algorithm.

Comprehensively, the discussion suggests systems based on the RAG yield a fair trade-off between precision, scalability and usability, and, therefore, are optimal in the document-centric applications.

Table 4.Qualitative Performance Evaluation

Aspect	**Observation**
Hallucination	Significantly reduced
Scalability	High
Adaptability	Dynamic updates
Latency	Moderate

7. APPLICATIONS

The suggested content summarizer and document-based conversational system is applicable in a wide range of areas:

7.1Research and Academic Areas.

The system would be useful to the researchers in summarizing lengthy research papers, extracting key contributions, and responding to literature-related questions. This saves the time and time taken in conducting the literature review and improves the productivity of the research.

7.2 Healthcare

The system has the ability to extract patient records, clinical reports, and medical literature in a summarized format in healthcare settings. Clinical decision support is an area where document-grounded responses are especially useful, and accuracy and traceability are crucial.

7.3 Legal Sector

The system allows legal professionals to analyze contracts, case laws and regulations. Efficiency and minimizing possibility of misinterpretation are enhanced by the capability of retrieving and referring to the particular parts of documents.

7.4 Organization

The system can be deployed by organizations to control internal documentation, policies and technical manuals. The interaction between documents and queries allows employees to get brief summaries, enhancing access to the knowledge.

7.5 Education

The system may be used in learning institutions to assist the students and teachers in in summarizing textbooks, lecture notes, and study resources and to reply to course-related questions.

8. CONCLUSION AND FUTURE WORK

This study provided an in-depth procedure of creating a Retrieval-Augmented Generation based content summarizer and a document- based conversational system. The combination of semantic retrieval with controlled generative modeling makes the proposed approach successful in overcoming the shortcomings of conventional document processing systems and standalone Large Language Models.

Table 5. Summary of Contributions and Future Directions

Aspect	Description
Contribution	RAG-based unified system
Key Benifit	Grounded generation
Limitation	Retrieval latency
Future Work	Citations

9. REFERENCES

1) P. Lewis, E. Perez, A. Piktus, F. Petroni,V. Karpukhin, N. Goyal, H. Küttler, M. Lewis,W. Yih,T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," arXiv preprint arXiv:2005.11401, 2020.

2) T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray,

3) B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language Models are Few-Shot Learners," arXiv preprint arXiv:2005.14165, 2020.

4) S. Zhang, S. Roller, S. Goyal, and J. Pineau, "EfficientlySummarizing Text with a Retrieval-Augmented Generation Model," arXiv preprint arXiv:2105.04623, 2021.

5) L. Wang, Y. Li, S. Zhang, and S. Feng, "Enhancing Extractive Text Summarization with Topic-Aware Graph Neural Networks," in Proc. of the 2019 Conference on Empirical MethodsinNatural Language Processing (EMNLP), 2019, pp. 3343.

6) Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "A Survey on Neural Network-Based Summarization Methods," arXiv preprint arXiv:1909.03186, 2019.

7) D. Bahri, H. Ji, M. Mazouchi, and Y. Choi, "Bridging the Gap between Relevance Matching and Semantic Matching for Document Summarization," in Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020,pp. 3002-3012.

8) Y. Liu, J. Gu, Z. Zhang, X. Wang, and J. Tang, "Hierarchical Transformers for Multi-Document Summarization," arXiv preprint arXiv:2004.06176, 2020.

9) Q. Cao, T. Shen, T. Xie, F. Wei, and T. Liu, "Pre-trained Models for Natural Language Processing: A Survey," Science China Technological Sciences, vol. 63, no. 10, pp. 1872-1897, 2020.

ACKNOWLEDGMENT

Authors would like to thank Dr. Meenakshi Sharma,Professor, Department of Computer Science and Engineering, Galgotias University, for her valued guidance, encouragement, and sustained support for the development of this project. They are also grateful to the faculties of School of Computing Science and Engineering for providing the necessary technical infrastructures and favorable academic ambiance. The authors are also thankful to peers and mentors for constructive suggestions and feedback, which led to significant enhancement in the quality and performance of the proposed system.