

A Survey of Retrieval-Augmented Language Models for Knowledge-Intensive Text Applications

Mr. Sachin Manekar¹

¹ Assistant Professor, Department of Computer Sciences and Applications, Mandsaur University, Mandsaur, India

ABSTRACT: Retrieval-Augmented Language Models (RALMs) have emerged as an effective approach for addressing the limitations of conventional language models in knowledge-intensive text applications. These models combine external knowledge retrieval and language generation, enabling them to deliver more relevant, accurate, and contextually appropriate responses and alleviate the need for internal knowledge. This survey reviews the fundamental concepts, architecture, knowledge sources, retrieval techniques, and major types of Retrieval-Augmented Generation (RAG) systems. It explores how the retrieval-based language generation process is affected by textual documents, scientific literature, databases, knowledge bases, enterprise documents and multimodal sources. The survey also covers the use of RAG for knowledge-intensive tasks, such as question answering, reasoning, document analysis, summarization, and information extraction. Particularly, emerging applications in healthcare and education, where reliable and domain-specific knowledge retrieval is a necessity, are given special attention. Besides, the survey stresses on some challenges related to retrieval quality, knowledge freshness, contextual relevance, hallucination, and system scalability. Last, future research directions on enhancing the robustness, efficiency, reliability and domain adaptability of RALMs are discussed.

KEYWORDS: Artificial Intelligence, Retrieval-Augmented Generation (RAG), Knowledge Retrieval, Natural Language Processing (NLP), Text.

1. Introduction

The emerging field of RAG has grown rapidly but has not yet been accompanied by a systematic synthesis that might provide a more general course for the field. This survey endeavors to fill this gap by mapping out the RAG process and charting its evolution and anticipated future paths, with a focus on the integration of RAG within LLMs [1]. This paper considers both technical paradigms and research methods, summarizing three main research paradigms from over 100 RAG studies, and analyzing key technologies in the core stages of "Retrieval," "Generation," and "Augmentation." [2].

One of the most basic data mining techniques is retrieval, which is designed to understand the input query and find relevant information from an external data source. It has been widely applied in many areas including search, question answering and recommender system. A good example of retrieval is the search engine, such as Google, Bing or Baidu, which can filter and retrieve the most relevant web pages or documents that can match a user's query and thus allow

the user to retrieve the desired information effectively. Meanwhile, retrieval models with the help of effective data maintenance in external databases can be a source to provide faithful and timely external knowledge, thus offering critical service in many knowledge intensive activities. The strong capabilities of retrieval techniques have paved the way for its successful integration into advanced generative models in the age of AI-Generated Content (AIGC). In particular, retrieval models have been combined with language models, resulting in a popular technique in the generative AI domain: RAG, designed to improve the quality of the created text content by incorporating retrieved information [3].

NLP is one of the most important fields in AI, and it has been applied to a wide range of tasks, from basic text classification and summarization to more complex tasks like machine translation and question-answering. One of the most notable applications of NLP is NLG, which deals with the creation of natural language from structured or unstructured information [4]. The objective of NLG is to make it possible for machines to produce coherent, relevant and contextually appropriate text, thereby making it easier for people to interact with machines.

With the evolution of AI, the need for more contextually relevant and factually accurate generated content has grown, which has introduced new challenges and innovations in the field of NLG. As AI continues to develop, the demand for generated content that is relevant and factually correct has increased, presenting new challenges and innovations in NLG. NLG models have shown impressive progress in producing fluent and coherent text, particularly sequence-to-sequence models. These models, however, are typically very dependent on the training data, and they may also fail to produce factually accurate or contextually rich text for queries that exceed their training set. This leads to models such as GPT or BERT based text generators being susceptible to hallucinating, which means they can generate plausible but factually incorrect or invented information [5][6].

This paper is organized into five sections: Section II describes Retrieval-Augmented Language Models. Section III explores Applications of Retrieval-Augmented Language Models in Knowledge-Intensive Tasks. Section IV reviews recent literature, and Section V concludes the paper while outlining potential future research directions.

2. Retrieval-Augmented Language Models

RELM incorporates retrieval into the pre-training phase of the language model, guaranteeing that the retriever and generator are optimized for tasks that come later on. A major advancement in REALM is its ability to learn which documents to retrieve in order to enhance the model's performance on certain tasks, such as answering questions or summarizing materials [5]. By updating the generator and the retriever simultaneously during training, REALM optimizes the retrieval process for the generating job. The REALM retriever has been taught to find documents that are useful for producing correct and logical answers, in addition to being related to the question. Consequently, REALM greatly enhances the caliber of produced replies, especially for jobs involving external knowledge.

A. Retrieval Components

The retrieval module is responsible for scanning large repositories and identifying data segments that pertain to the query and its approaches include:

- **Dense Retrieval:** Neural embedding techniques convert texts and queries into vector spaces with many dimensions [7]. Which documents are most relevant are determined by similarity measures like cosine similarity.
- **Sparse Retrieval:** Methods like TF-IDF and BM25 rely on lexical overlaps and phrase frequencies to

determine relevance.

- **Hybrid Retrieval:** By combining dense and sparse methods, systems achieve a balance between semantic understanding and term-based matching.

B. Types of RAG

RAG classify retrieval-augmented systems according to how external knowledge is retrieved, processed, and integrated with the language model. There are different types that aim to enhance retrieval precision, contextual relevance, reasoning ability, and the ability to generate responses in varying knowledge-intensive tasks.

- **Naive RAG:** The Naive RAG research paradigm is the first paradigm that came to the fore right after ChatGPT became popular. The Naive RAG is a traditional approach with three steps: indexing, retrieval, and generation that is also referred to as a "Retrieve-Read" approach [8].

- **Advanced RAG:** Advanced RAG are some improvements built on top of Naive RAG. Focusing on enhancing retrieval quality, it employs pre-retrieval and post-retrieval strategies.

- **Modular RAG:** The modular RAG architecture goes beyond these two RAG approaches, providing greater flexibility and adaptability. It also contains various approaches to enhance its parts, including fine tuning the retriever and adding a search module for similarity searches.

- **Graph-Based RAG:** Graph RAG, also begins with a query based on the user's input same as Vector RAG. The main difference between Vector RAG and Graph RAG lies in the retrieval part. The question here is now used to query the KG and extract relevant nodes (entities) and edges (relationships) from the KG that are relevant to the query [9].

C. RAG Architecture

It explains the operation of the process by which external knowledge is pulled into a language model. Generally, consists of document collection, document pre-processing, context expansion, and response generation.

- A. Document Collection: Documents are
- B. collected from multiple sources such as
- C. PDF, DOCX, TXT, and HTML files.

- **Document Collection:** Support for multiple document sources like PDF, DOCX, TXT, HTML files.
- **Preprocessing:** The extracted text is cleaned, normalized, tokenized, and divided into smaller chunks.
- **Embedding Generation:** Using a pre-trained

embedding model, every document chunk is transformed into a dense vector embedding.

- **Vector Storage:** A vector database is used to store and index the created embeddings in order to facilitate efficient retrieval.

- **User Query Processing:** Vector embeddings are generated from the user's natural language query.

- **Semantic Retrieval:** The vector database retrieves the most semantically relevant document chunks based on similarity search.

- **Context Augmentation:** Users' queries are integrated with the obtained document fragments to offer context [10].

- **Response Generation:** The Large Language Model generates a context-aware and evidence-based response.

- **Result Presentation:** The generated answer is displayed to the user along with references or supporting document excerpts.

D. Retrieval Techniques in RAG

It determines how relevant information is identified and retrieved from external knowledge sources. Common approaches include sparse, dense, and knowledge graph-based retrieval.

- **Sparse Retrieval:** Early systems relied on Boolean matching, which was based on exact logical matching of

query terms; documents were either retrieved or not[11]. This binary approach neglected a crucial feature, which was term importance, and produced either extremely large or empty results. By the late 1960s, heuristic weighting schemes such as term frequency (TF) were being used in systems like SMART.

- **Dense Retrieval:** A modern information retrieval technique based on DL, utilizing dense vector representations generated by NNs to match queries with documents[12]. Unlike traditional sparse retrieval methods, dense retrieval does not rely on exact word matching but instead captures semantic similarity between the query and the documents.

- **Knowledge Graph Retrieval:** The knowledge graph retrieval section retrieves the structured biomedical facts in the form of triples, where e1 and e2 refer to entities and r represents the relation between them [13]. This module allows for explicit reasoning over mechanistic associations, e.g., pathogen–host interactions or drug–disease relationships, unlike text-based retrieval, which is based on lexical or semantic similarity.

E. Knowledge Sources for Retrieval-Augmented Models

These methods assist in determining key information from external knowledge sources and offer helpful context to the language model. Table I presents the major retrieval techniques used in RAG systems. It contrasts various methods according to their retrieval mechanism, characteristics, and typical uses.

TABLE I. MAJOR KNOWLEDGE SOURCES USED IN RETRIEVAL-AUGMENTED MODELS

Knowledge Source Category	Source Type	Content Format	Examples	Primary Knowledge	Retrieval Approach	Main Application
Textual Sources	Web pages, books, articles	Unstructured text	Websites, reports, news	General and domain knowledge	Keyword / Dense Retrieval	Question Answering
Scientific Sources	Research papers journals	Text + metadata	IEEE papers, PubMed articles	Scientific and technical knowledge	Semantic Retrieval	Research Assistance
Structured Sources	Databases, knowledge bases	Tables, records, triples	SQL databases, Wiki data	Factual and structured knowledge	Structured Retrieval	Fact-Based QA
Knowledge Graphs	Entity-relation graphs	Graph / Triples	Biomedical and enterprise KGs	Entity relationships	Graph Retrieval	Multi-Hop Reasoning
Enterprise Sources	Internal documents	Text, tables, PDFs	Policies, manuals, reports	Organization-specific knowledge	Hybrid Retrieval	Enterprise QA

Multimodal Sources	Images, audio, video	Text visual/audio data +	Diagrams, medical images, videos	Multimodal knowledge	Multimodal Retrieval	Multimodal QA
User-Provided Sources	Uploaded documents	PDFs, notes, files	Personal documents, datasets	Task-specific knowledge	Document Retrieval	Personalized QA

3. Applications of Retrieval-Augmented Language Models in Knowledge-Intensive Tasks

The applications of RALMs in Knowledge-Intensive Tasks span a broad spectrum of tasks that demand the ability to access external knowledge, domain-specific knowledge, or current knowledge. RALMs are applied to improve the accuracy and relevance of the generated response in applications like question answering, knowledge-intensive reasoning, document analysis, summarization, and information extraction.

A. Question Answering System

Question Analysis, Passage Retrieval, and Answer Extraction are the three main modules that make up a question answering system's pipeline structure.

- **Question Analysis Module:** Usually, in order to get solutions, queries sent to QA systems must be processed and comprehended. This means that the Question Analysis module is where all the processing of questions should begin. Useful for analysis in subsequent modules, this step takes the user query as input and returns representations of the query as output. Extracting useful information from other sources and determining the user's purpose are both aided by the processed text.

- **Passage Retrieval Module:** Naturally, a traditional search engine is the backbone of passage retrieval, which uses a database to find a group of relevant candidate passages or phrases [14]. In this step, leverage the queries created in the question analysis module to get relevant responses by searching various information sources. The collected text gives the RAG model the necessary context to provide correct results.

- **Answer Extraction Module:** Any Question Answering system worth its salt have an answer extraction component. It derives the precise solution from the produced passages. For this purpose, it employs a scoring system to construct a set of possible responses from the produced passages, which are subsequently ranked. A meaningful and contextually grounded solution is then generated by the RAG model using the selected text.

B. Knowledge-Intensive Reasoning

Large-scale models are necessary for knowledge-intensive tasks in order to efficiently use external information during inference. A common solution is RAG which retrieves relevant knowledge from external corpora based on the input and supplements the model's internal reasoning process[15]. Building on this idea, ICL provides input-output exemplars directly within the prompt, allowing the model to generalize task-specific behaviors from observed patterns. The retrieved text provides additional evidence that helps the model perform accurate and context-aware reasoning.

C. Knowledge-Grounded Dialogue Systems

Knowledge-grounded dialog systems are methods that may produce varied and instructive replies by drawing on relevant external knowledge [16]. Knowledge selection and answer creation are the two main components of knowledge-grounded conversation systems. Based on the content of the retrieved knowledge and the context of the dialog, the response generation module provides a response. The knowledge selection module picks relevant information from the candidate set and uses it in the next response. In order to provide insightful and correct replies, the recovered text supplies necessary background information.

D. Document Analysis and Information Extraction

Gathering documents from a variety of sources, including PDFs, text files, research papers, legal documents, and domain-specific reports, is the first step in the technique [17]. Because no two papers can possibly have the same structure, format, or quality, they are all converted to a common text format. In doing so, ensure that the downstream processing modules are compatible and that all document types may be processed in accordance with standard protocols. The standardized text is then indexed and used as an external knowledge source for retrieval-based document analysis and information extraction.

E. RAG in Healthcare

RAG in Healthcare enables language models to retrieve relevant medical information from trusted knowledge sources such as clinical guidelines, research literature, and electronic health records [18]. It supports applications including clinical question answering,

medical decision support, patient information retrieval, and healthcare document analysis.

- **Diagnostic Assistance:** RAG can be used to identify similar patient cases or clinical guidelines from structured databases and literature to aid in diagnostic decision making [19]. The image caption or structured report is used in radiology and pathology to query the repositories, which helps to formulate a differential diagnosis or recommendations based on evidence [20].

- **Model Training on Labeled Threat Data:** Supervised learning models learn from historical data by

- **Summarization of EHRs and Discharge Notes:** RAG provides clinical summarization with longitudinal patient information. The retriever accesses relevant medical information from EHRs, including medications, lab reports, and previous hospitalizations, which the generator then uses to generate a brief discharge summary or a progress note [21]. The contextualization of extractive and abstractive outputs is evidence-based, which minimizes leaving out and inaccuracies in the clinical aspects.

- **Medical Question Answering:** Medical Question Answering systems, which retrieve information from well-curated biomedical corpora, are used to solve complex medical queries. These systems can limit hallucinations in standalone LLMs by grounding answers in peer-reviewed literature or organized medical knowledge. These systems can minimize the hallucinations typically observed in standalone LLMs when answering questions by referencing literature or structured medical knowledge [22][23]. Advanced indexing techniques such as BM25 provide relevance in long-form or open-ended queries, making them more reliable and transparent.

F. RAG Applications in Education

RAG Applications in Education use language models to search for relevant information in textbooks, academic materials, course content, or institutional documents. They offer a source of contextually appropriate information that is reliable for personalized learning, educational question answering, content generation, and student support.

- **RAG-powered interactive learning systems:** RAG-Powered Interactive Learning Systems leverage RAG to facilitate dynamic, personalized and context-aware interactions in education. These are systems which enable conversational interaction, recall of domain knowledge, and adaptability to a wide range of educational contexts.

- **Educational content development and assessment:** RAG has also been applied in the field of

educational content development and assessment [24]. The emphasis in this area is on improving the development and testing of learning resources. Different from interactive learning systems, this approach automates content generation without requiring multiple interactions between users and the system.

- **Scaling RAG in educational ecosystems:** The integration of RAG into educational ecosystems is changing the way educational institutions provide accurate, personalized, and user-friendly solutions. This work identifies three major trends in how RAG is scaled and integrated within educational ecosystems for AI in Educational Policy and Student Support, Domain-Specific Applications in Various Disciplines, and RAG Enhancement for Language Models and Re-search Directions.

G. RAG In Financial Document Analysis

The difficulty of RAG makes it an attractive choice for financial document automation. In this context, by way of illustration, financial documents such as annual reports, tax filings, and regulatory submissions are filled with specialized terminology; hence the appropriate facilities must be accorded to these documents, particularly in the face of an ever-changing environment with new regulations. Real-life examples abound of traditional models having difficulties with such situations [25]. A notable strength of RAG is the ability to pull the most relevant data about any financial issue from internal systems and external sources that are so much regulatory updates and market- reports.

4. Literature Review

This section features related research in the areas of RAG and RALMs, and their impact on enhancing factual accuracy, contextual relevance, knowledge retrieval, reasoning, document understanding, and domain adaptability in a variety of knowledge-intensive applications.

I. Ahmad et al. (2026) presents an overview of the key aspects that make up LLMs and introduces the main methodologies for specializing and grounding these models in factual information, focusing on methods like Retrieval Augmented Generation and fine-tuning. The contribution builds on this by examining the legal implications from their use, including issues like liability, algorithmic transparency, personal data protection, bias mitigation, and the reliability of outputs generated by these models, while also conducting a survey of the main applications of LLMs in legal and administrative settings, offering a timely and critically informed overview of the opportunities and challenges arising from their adoption [26].

Yuepeng Zhang et al. (2025) involves an optimization-

oriented RAG system for large-scale document understanding, incorporating advanced FAISS-based retrieval and multi-model generation architecture to enhance retrieval precision, consistency of responses, and crossdomain adaptability. Evaluated system using two large-scale data sets, one based on the ArXiv scientific corpus and another on a domain-specific EV dataset [27].

Samuel Kiruba et al. (2025) gives a comprehensive analysis of RAG architectures, including its elements—retrieval, generation, and augmentation—and how to integrate them to improve the accuracy of facts and the relevance of context. Lastly, talk about how technology has recently progressed, how many knowledge-intensive work areas have used it, and how sustained knowledge sources may be created without retraining being expensive. By supplanting the static nature of AI knowledge with real-time retrieval of information; RAG enables a new approach in which durable, scalable, and domain-adaptive content may be created by AI [28].

Garima Agrawal et al. (2024) Identify eight major areas of concern, such as inaccurate relation mapping, inefficient ambiguity resolution, and misreading of question context, by critically analyzing the failure points in current KG-based RAG approaches. The goal is to make KG-RAG systems more reliable and effective by making reasoning skills better, constraint identification better, and knowledge graph structural limits addressed. I created a proof-of-concept by combining an existing KG-RAG system with the ideas of Mindful-RAG in order to verify this method. With possible uses in fields such as healthcare, law, research, and scientific discovery—where trustworthiness and accuracy are of the utmost importance—Mindful-RAG

aspires to provide AI-driven knowledge retrieval systems that are more stable, precise, and contextually relevant [29].

Zecheng Wang et al. (2023) provide SUFFIX RETRIEVAL-Augmented LM (SUREALM), a new language model that uses autoregressive methods to mimic a bidirectional contextual impact. In order to generate sequences with comparable word histories, SUREALM uses an embedding retriever to look for training sentences in a database. The recovered phrases' suffix parts, in particular, are quite similar to the "future" context. Relative to competing baselines, the suggested model demonstrated encouraging results in reducing word perplexity on both the validation and test sets when evaluated on the DSTC9 spoken dialog corpus. The code may be found on GitHub [30].

Alexander Long et al. (2022) implement RAC on the long-tail classification issue and show that it outperforms the state-of-the-art on Places365-LT by 14.5% and iNaturalist-2018 by 6.7%, respectively, even if the external information source is only the training datasets. prove that the retrieval module of RAC can learn to accurately identify tail classes on its own. In turn, this allows the basic encoder to enhance its performance on common classes. In addition to paving the way for more efficient use of external memory in popular computer vision architectures, RAC offers an alternate method of using big, pretrained models without fine-tuning [31].

Table II provides a summary of representative studies on RAG and retrieval-based language models, along with a summary of their methodologies, key results, limitations, and future directions. It shows the increasing importance of retrieval skills to enhance factual accuracy, context, scalability, and domain adaptability.

TABLE II. COMPARATIVE ANALYSIS OF RETRIEVAL-AUGMENTED LANGUAGE MODELS FOR KNOWLEDGE-INTENSIVE TASK

Reference	Study Focus	Approach / Methodology	Key Findings	Challenges / Limitations	Future Directions
I. Ahmad et al. (2026)	LLM specialization and factual grounding	Review of LLM methodologies, including RAG and fine-tuning, with analysis of legal and administrative applications	RAG improves factual grounding and supports reliable LLM applications	Liability, transparency, data protection, bias, and reliability of generated outputs	Develop responsible and reliable LLM grounding methods for legal and administrative domains
Yuepeng Zhang et al. (2025)	Large-scale document understanding	FAISS-based retrieval with metadata-rich embeddings, adaptive text	Improved retrieval efficiency, response quality,	Large-scale processing and computational requirements;	Develop scalable and efficient RAG systems for broader

		tiling, multi-model generation, dynamic scheduling, and weighted answer aggregation	consistency, and cross-domain adaptability compared with traditional RAG	system complexity	knowledge-intensive applications
Samuel Kiruba et al. (2025)	RAG architectures and applications	Comprehensive review of retrieval, augmentation, and generation components, applications, and continuous knowledge sources	RAG improves factual accuracy, contextual relevance, scalability, and domain adaptability without retraining	Efficient retrieval and evaluation remain challenging	Improve retrieval, evaluation, sustainability, and continuous knowledge integration
Garima Agrawal et al. (2024)	Knowledge Graph-based RAG	Analysis of KG-RAG failure points and development of intent-driven and context-aware Mindful-RAG	Improves reasoning, constraint identification, contextual alignment, and retrieval reliability	Question-context misinterpretation, incorrect relation mapping, ambiguity resolution, and KG structural limitations	Develop robust KG-RAG for healthcare, legal, research, and scientific applications
Zecheng Wang et al. (2023)	Retrieval-Augmented Language Modeling	SUREALM uses an embedding retriever to retrieve similar training sentences and suffix information during generation	Reduced word perplexity on the DSTC9 spoken dialogue corpus compared with competitive baselines	Evaluation focused on spoken dialogue; dependence on retrieved training sentences	Extend retrieval-based language modeling to broader dialogue and generation tasks
Alexander Long et al. (2022)	Retrieval-Augmented Classification	Retrieval-Augmented Classification (RAC) uses training data as an external information source for long-tail classification	Improved performance on Places365-LT and iNaturalist-2018, particularly for tail classes	Focused on computer vision and long-tail classification rather than text-based RAG	Extend external-memory retrieval approaches to broader domains and applications

5. Conclusion and Future Work

RALMs have proven to be a successful technique in the realm of language generation for knowledge-intensive applications of text, which combine retrieval of external information with language generation. This survey has covered the key elements of RAG: architecture, types, knowledge sources, and retrieval methods. It pointed out the use of textual documents, scientific literature, databases, knowledge bases, knowledge graphs, enterprise documents and multimodal sources as sources for providing relevant contextual information. In addition, survey applications in the areas of question answering, knowledge-intensive reasoning, document analysis, summarization, information extraction, healthcare, and education were explored. RAG can enhance the contextuality, factuality, and contextual awareness of the generated responses and decrease reliance on static model knowledge. But there are difficulties like retrieval quality, knowledge freshness, hallucination, contextual relevance and scalability. In summary, RAG offers a versatile approach to creating robust and domain-specific language applications in different knowledge-intensive contexts.

Future research should explore the following aspects further: retrieval accuracy, knowledge freshness, multimodal retrieval, efficient indexing, and multi-hop reasoning. To address these challenges, additional research is required to address the problem of hallucination, increase domain adaptability, and build trustworthy RAG systems for specialized domains like healthcare and education.

REFERENCES

1. S. Devarakonda, R. Lingam, and V. Challa, "Confidence-Gated RAG for Adaptive Retrieval in Sequential Agents," in ICLR 2026 Workshop on Logical Reasoning of Large Language Models, ICLR, 2026, pp. 01–10, Apr.
2. Y. Gao et al., "Retrieval-Augmented Generation for Large Language Models : A Survey," pp. 1–21, 2024.
3. W. Fan et al., "A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models," in Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024, pp. 6491–6501. doi: 10.1145/3637528.3671470.
4. C. P. Singh, "Enterprise-Grade AI-Driven Text Analytics and Insight Extraction Using Transformer-Based NLP Models," in 2026 14th International Symposium on Digital Forensics and Security (ISDFS), IEEE, Mar. 2026, pp. 1–6. doi: 10.1109/ISDFS69419.2026.11459059.
5. S. Gupta, R. Ranjan, and S. N. Singh, "A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions," no. ML, 2024, doi: 10.48550/arXiv.2410.12837.
6. M. N. Paralath, "Bridging Astronomical Imagery and Text with Advanced Multi-Modal AI," in 2025 IEEE 7th International Conference on Computing, Communication and Automation (ICCCA), IEEE, Nov. 2025, pp. 1–8. doi: 10.1109/ICCCA66364.2025.11325649.
7. S. K. Kota, "Retrieval-Augmented Generation (RAG) Systems: Architectures, Strategies, and Evaluation," pp. 350–358, 2024, doi: 10.32996/jcsts.
8. Y. Gong, P. He, H. Zhao, N. Duan, and C. Engineering, "Query Rewriting for Retrieval-Augmented Large Language Models," 2023. doi: 10.48550/arXiv.2305.14283.
9. B. Sarmah, B. Hall, and S. Pasquali, "HybridRAG : Integrating Knowledge Graphs and Vector Retrieval Augmented Generation for Efficient Information Extraction," 2024.
10. T. Manohar and M. G.priyanka, "Document Retrieval System Using Rag," Int. J. Data Sci. IoT Manag. Syst., vol. 5, pp. 1–12, 2026, doi: 10.64751/ijdim.2026.v5.n3.1123.
11. S. Mohammed, "From Sparse to Dense and Beyond: A Literature Review of Retrieval-Augmented Generation," 2025.
12. Y. Wang, Evaluating Sparse and Dense Retrieval in Retrieval-Augmented Generation Systems : A Study, vol. 1, no. 1. Association for Computing Machinery, 2024. doi: 10.1145/3708657.3708747.
13. S. Xu, Z. Yan, and C. Dai, "MEGA-RAG : a retrieval-augmented generation framework with multi-evidence guided answer refinement for mitigating hallucinations of LLMs in public health".
14. B. Ojokoh and E. Adebisi, "A Review of Question Answering Systems," J. Web Eng., vol. 17, pp. 717–758, 2019, doi: 10.13052/jwe1540-9589.1785.
15. B. Zhao, Unbiased Reasoning for Knowledge-Intensive Tasks in Large Language Models via Conditional Front-Door Adjustment, vol. 1, no. 1. Association for Computing Machinery. doi: 10.1145/3746252.3761103.
16. H. Onozeki, "Knowledge-Grounded Dialogue Systems for Generating Interesting and Engaging Responses," 2024. doi: 10.18653/v1/2024.yrrsds-1.9.

17. S. K. Singh, A. Kumar, and D. M. Sharma, "Content Summarizer And Document Gpt Using Rag," vol. 13, no. 6, pp. 99–104, 2026.
18. M. Mittal and V. A. J. Raja, "Ai-Driven Clinical Decision Support Systems: Evaluating Impact On Diagnosis And Treatment Accuracy," *Int. J. Eng. Appl. Artif. Intell.*, vol. 3, no. 1, pp. 10–29, Jun. 2025, doi: 10.34218/IJEAAI_03_01_002.
19. F. Neha, D. Bhati, and D. K. Shukla, "Retrieval-Augmented Generation (RAG) in Healthcare: A Comprehensive Review," vol. 6, no. 9, 2025, doi: 10.3390/ai6090226.
20. P. S. Papageorgiou et al., "The Role of Large Language Models in Improving Diagnostic-Related Groups Assignment and Clinical Decision Support in Healthcare Systems: An Example from Radiology and Nuclear Medicine," *Appl. Sci.*, vol. 15, no. 16, pp. 1–15, 2025, doi: 10.3390/app15169005.
21. M. Alkhalaf, P. Yu, M. Yin, and C. Deng, "Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records," *J. Biomed. Inform.*, vol. 156, p. 104662, 2024, doi: 10.1016/j.jbi.2024.104662.
22. E. Asgari et al., "A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation," *npj Digit. Med.*, vol. 8, no. 1, p. 274, 2025, doi: 10.1038/s41746-025-01670-7.
23. P. Sundaramoorthy and J. Lingam, "Enhancing Supply Chain Resilience and Efficiency with Large Language Models," in 2026 IEEE 18th International Conference on Computational Intelligence and Communication Networks (CICN), IEEE, Jun. 2026, pp. 367–373. doi: 10.1109/CICN70047.2026.11594388.
24. Z. Li, Z. Wang, W. Wang, K. Hung, H. Xie, and F. Lee, "Computers and Education : Artificial Intelligence Retrieval-augmented generation for educational application : A systematic survey," *Comput. Educ. Artif. Intell.*, vol. 8, no. May, p. 100417, 2025, doi: 10.1016/j.caeai.2025.100417.
25. N. Malali, "The Role of Retrieval-Augmented Generation (RAG) in Financial Document Processing: Automating Compliance and Reporting," *Int. J. Manag. Technol.*, vol. 12, no. 3, pp. 26–46, 2025, doi: 10.37745/ijmt.2013/vol12n32646.
26. I. Ahmad, A. Amelio, M. Aracne, L. Caroprese, E. Z. Gill, and C. Morbidoni, "Technologies, Applications, and Challenges of Large Language Models," in 2026 49th MIPRO ICT and Electronics Convention (MIPRO), 2026, pp. 2116–2121. doi: 10.1109/MIPRO70003.2026.11591567.
27. Y. Zhang, J. Zeng, J. Wang, Y. Gao, and W. Wu, "Optimization-Oriented Retrieval-Augmented Generation for Large-Scale Document Understanding," in 2025 International Conference on Graphics and Signal Processing (ICGSP), 2025, pp. 86–93. doi: 10.1109/ICGSP66091.2025.11379732.
28. S. Kiruba, C. Sam Anish, R. S. B. Krishna, and S. J. Rayen, "A Comprehensive Study-Retrieval-Augmented Generation(RAG)," in 2025 10th International Conference on Smart Structures and Systems (ICSSS), 2025, pp. 1–8. doi: 10.1109/ICSSS66939.2025.11346240.
29. G. Agrawal, T. Kumarage, Z. Alghamdi, and H. Liu, "Mindful-RAG: A Study of Points of Failure in Retrieval Augmented Generation," in 2024 2nd International Conference on Foundation and Large Language Models (FLLM), 2024, pp. 607–611. doi: 10.1109/FLLM63129.2024.10852457.
30. Z. Wang and Y.-C. Tam, "Suffix Retrieval-Augmented Language Modeling," in ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10096450.
31. A. Long et al., "Retrieval Augmented Classification for Long-Tail Visual Recognition," in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6949–6959. doi: 10.1109/CVPR52688.2022.00683.