GMR
The Original

RETRIEVAL-AUGMENTED GENERATION FOR SCIENTIFIC KNOWLEDGE SYNTHESIS: A REVIEW OF RETRIEVAL, QUESTION ANSWERING, SUMMARIZATION, AND EVALUATION

Ahmed Hassan[1], Kiran Macwan[2], AnilKumar Patel[3]

1M.Tech (Artificial Intelligence & Data Science), Parul University, Vadodara, India. ORCID ID: 0009-0005-0910-8186, 2503032340047@paruluniversity.ac.in
[2]Department of Artificial Intelligence and Data Science, Parul University, Vadodara, India (Supervisors) ORCID ID: 0000-0003-2610-2593
[3]Department of Artificial Intelligence and Machine Learning Parul University, Vadodara, India (Supervisors) ORCID ID: 0000-0002-4039-9672
***Corresponding author:** Ahmed Hassan, 2503032340047@paruluniversity.ac.in

ABSTRACT

The growth of literature is happening faster, making it hard for researchers to find the information they need. They have to look through a lot of things to get the evidence and put together what they find from different studies and make sure their knowledge is accurate. Large Language Models are good for natural language understanding and generation, but they depend on static knowledge, which leads to hallucinations, outdated information and limited transparency in the research domain. One way to solve this problem is the use of Retrieval Augmented Generation. It integrates external knowledge with LLMs to produce evidence for Scientific Questions and Document Summarization. It explores the current improvements in retrieval architectures, including adaptive, hierarchical, graph-based and self-improving strategies, with their applications in literature retrieval, multi-document reasoning and knowledge synthesis. This review discusses evaluation frameworks and benchmark datasets used to assess retrieval quality, citation grounding, answer reliability and faithfulness to the retrieved evidence. In addition, it also identifies current challenges related to retrieval scalability, long-context reasoning, multi-modal research, standardized data evaluation and computational efficiency. Unlike existing surveys that explore individual aspects of RAG, this review uses a unified framework that integrates retrieval architectures, research applications, and evaluation methodologies. It also provides a direction toward trustworthy, scalable and evidence-grounded knowledge systems.

KEYWORDS: Retrieval-Augmented Generation (RAG), Large Language Models (LLMs), Scientific Question-Answering, Document Summarization, Literature Retrieval, Information Retrieval and Scientific Knowledge Systems

1. INTRODUCTION

The growth of scientific literature has fundamentally transformed the landscape of knowledge discovery and dissemination. Millions of scholarly publications are produced each year across diverse scientific disciplines, making it increasingly difficult for researchers to efficiently locate relevant evidence, synthesize findings from multiple studies and remain current with rapidly evolving research. As the volume and complexity of scientific information continue to increase, conventional keyword-based search engines and information retrieval systems often struggle to support evidence synthesis, multi-document reasoning and complex scientific question answering. These challenges have created an urgent demand for intelligent knowledge systems capable of retrieving trustworthy evidence while generating accurate and context-aware scientific insights [1-4],[28]. Large Language Models have shown abilities in natural language understanding, reasoning and generating text. This has led to improvements in question answering, making summaries and finding new scientific knowledge. However, their reliance on static parametric knowledge limits their ability to provide current, domain-specific and verifiable information. Consequently, LLMs remain susceptible to hallucinations, outdated knowledge and insufficient factual grounding, particularly in scientific domains where precision, transparency and citation-supported evidence are essential. These limitations have motivated the development of Retrieval Augmented Generation (RAG), which combines external knowledge retrieval with language generation to ground responses in relevant documents and improve factual accuracy,reliability and interpretability [1-4], [16]. In Recent years, we have witnessed faster progress in RAG research. Beyond conventional vector retrieval, researchers have proposed adaptive retrieval, graph-based retrieval, hierarchical retrieval, retrieval optimization, self-reflective generation and domain-specific RAG frameworks that substantially improve scientific question answering, literature retrieval, document reasoning and multi-document summarization. Concurrently, dedicated evaluation frameworks and benchmarks have been introduced to assess retrieval effectiveness, answer faithfulness, citation accuracy and overall system reliability, enabling more systematic evaluation of RAG-based scientific knowledge systems [17-33]. Despite these important advances, the existing literature remains fragmented. Current studies primarily focus on individual aspects of RAG, including retrieval architectures, scientific applications or evaluation methodologies, while existing surveys largely emphasize general-purpose RAG techniques. Despite these advances, no review has systematically brought together retrieval architectures, scientific question answering, document summarization and evaluation

frameworks within the broader context of scientific knowledge systems. As a result, it remains difficult to understand how developments in retrieval mechanisms, generation strategies and evaluation methodologies work together to support trustworthy scientific AI systems. This review is going to look at the research on Retrieval-Augmented Generation for scientific knowledge systems. We want to understand what is missing and how we can make it better. The main focus is on RAG research because it is very important for knowledge systems. By studying RAG, we can find out what works and what does not work. This helps us to make scientific knowledge systems more useful. This review examines the evolution of retrieval architecture, advanced retrieval paradigms, scientific question answering, document summarization and evaluation frameworks within a unified perspective. This work is different from reviews that looked at these topics on their own. It brings them together and shows how they are connected. It finds patterns in the research and points out problems that are still unsolved. In the future, we will work on making systems that can deal with a lot of information. These systems have to be ones that researchers can trust. The information in these systems has to be based on evidence

2. LITERATURE REVIEW

2.1. Introduction to Retrieval Augmented Generation

Large Language Models have really changed the way we do Natural Language Processing. They are good at many things like answering questions, translating languages, summarizing documents and generating code. It also used in different areas and they are very helpful. It can do things that were not possible before. That is why they are important for Natural Language Processing. Despite all these achievements, existing LLMs remain constrained by their reliance on static parametric knowledge acquired during pre-training. They are susceptible to hallucinations, difficulties in incorporating newly emerging knowledge without expensive retraining and often lack specialized expertise for domain-specific applications. These limitations affect their reliability in knowledge-intensive environments where factual accuracy, transparency and continuously updated information are important [1]-[4]. Retrieval-Augmented Generation (RAG) has emerged to address these shortcomings by augmenting LLMs with external knowledge retrieval during inference. Instead of relying on exclusive knowledge stored within model parameters, RAG retrieves relevant information from external knowledge sources and integrates the retrieved evidence into the response generation process. This retrieval mechanism enables LLMs to generate responses that are factual, transparent and adaptable to evolving knowledge, reducing hallucinations and improving performance in knowledge-intensive tasks. Thus, RAG has become a foundational framework for applications such as open-domain question answering, scientific literature analysis, enterprise knowledge management and document summarization [1]-[4] Although the primary objective of RAG is defined as language generation through external knowledge retrieval, the four authors analyze the paradigm from different perspectives. Gao et al. described the evolutionary development of RAG through the Naive, Advanced and Modular paradigms, illustrating how retrieval-augmented systems have progressed from simple retrieve-then-generate pipelines toward flexible architectures composed of independently optimized functional modules [1]. In contrast, Wu et al. focused on the interaction between retrieval and generation by introducing a taxonomy of retrieval fusion strategies, including query-based, logits-based, latent and parametric fusion, thereby highlighting different knowledge integration mechanisms in LLMs [2]. There is also an architecture-oriented survey categorized existing RAG systems into retriever-oriented, generation-oriented, hybrid and robustness-oriented architectures, providing a systematic understanding of optimization strategies across the entire RAG framework [3]. Meanwhile, Haung and Huang examined RAG from an Information Retrieval (IR) perspective by decomposing the workflow into pre-retrieval, retrieval, post-retrieval and generation stages, enabling researchers to identify optimization opportunities throughout the complete retrieval-generation pipeline [4]. Rather than presenting competing viewpoints, these taxonomies collectively illustrate the multidimensional evolution of RAG research. The evolutionary taxonomy explains how RAG architectures have matured from basic retrieval-generation pipelines to modular systems [1], the retrieval fusion taxonomy investigates how retrieved information can be incorporated into language models more effectively [2], and the IR-oriented framework identifies where optimization can be performed within the retrieval-generation pipeline [4]. Together, these perspectives demonstrate that modern RAG research has evolved beyond traditional document retrieval to encompass retrieval optimization, knowledge integration, modular architectures and reliable text generation [1]-[4]. Building on these foundations, recent RAG research has increasingly focused on improving retrieval effectiveness through query optimization, retrieval fusion, document re-ranking, context compression, adaptive retrieval and iterative retrieval strategies [1]-[4]. These techniques aim to improve the relevance, completeness and quality of retrieved evidence before generation, thereby enhancing factual consistency while reducing the propagation of retrieval errors. Recent studies have shown some progress. They also point out problems with retrieval inaccuracies, redundant contextual information and computational overhead limitations in long - context studies. The main issues are retrieval inaccuracies and redundant contextual information. Also, there are limitations in long-context processing and computational overhead. There is also a trade-off between retrieval efficiency and response quality, as shown in the studies [1]-[4]. The development of RAG shows a change from ways of using language models that rely on retrieval information to systems that focus on knowledge. It helps get information, adjust thinking and create answers that are based evidence. As RAG models have grown more advanced, researchers have explored areas such as literature retrieval, autonomous retrieval planning, graph-based retrieval and trustworthy generation. These improvements lead to the topics that are discussed next, which look at the areas of research and new trends that are helping shape how RAG systems are built [1]-[4].

Figure 1 illustrates the evolution and taxonomy of RAG architectures. It shows the progression of RAG from conventional retrieval- generation pipelines toward advanced and modular approaches.

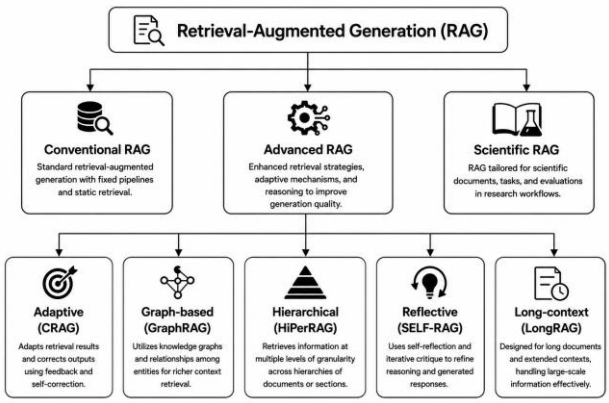


Figure 1: Evolution and taxonomy of Retrieval-Augmented Generation (RAG).

The evolution includes improvements in retrieval, knowledge integration and generation strategies. These developments support more reliable and evidence-grounded scientific knowledge systems.

2.2. Retrieval Strategies in RAG

The effectiveness of a RAG system depends heavily on the retrieval component. High-quality retrieved knowledge enables the model to produce responses that are more accurate, contextually relevant and reliable. Although early RAG systems demonstrated the effectiveness of augmenting LLMs with external knowledge, following research identified retrieval as the primary performance bottleneck. Inaccurate document selection, irrelevant contexts, incomplete evidence and inefficient retrieval workflows often degrade downstream generation quality. Recent research has shifted from simply retrieving semantically similar documents to designing intelligent retrieval strategies that optimize the entire retrieval process [2]-[4]. Early studies concentrated on strengthening the retrieval pipeline through improvements in query processing, retrieval fusion, document indexing, re-ranking and post-retrieval refinement. Wu et al. Further organized retrieval integration into query-based, latent, logits-based and parametric fusion strategies, showing multiple mechanisms for incorporating external knowledge into LLMs [2]. Similarly, the architecture-oriented survey emphasized retrieval-oriented optimization to improve document selection and evidence quality [3], while the information retrieval (IR)-oriented survey showed that retrieval effectiveness depends on coordinated optimization across the pre-retrieval, retrieval and post-retrieval stages rather than on retrieval alone [4]. These studies show that retrieval is a process that happens in stages. They also made it possible for other retrieval systems to be created that can adapt to situations. The retrieval process is really about finding the way to get what you need. These studies helped create a base for systems to be built on, so we can have better retrieval frameworks. The RAG system got better over time. Researchers saw that they could not just use retrieval to deal with queries that were complicated in different ways or to handle the fact that the quality of retrieval was not always the same. They knew the RAG system had to do more to help with this. The system was not good enough when it came to dealing with queries of varying complexity or with retrieval quality that was not consistent. This led to the development of adaptive retrieval approaches that dynamically adjust retrieval behavior during inference. SELF-RAG introduced a self-reflection mechanism that enables the model to determine when additional external knowledge is required, thereby reducing unnecessary retrieval while improving factual grounding [16]. In contrast, CRAG focuses on retrieval quality assurance by evaluating retrieval documents before generation and performing corrective retrieval when the initial evidence is insufficient or unreliable [17]. Although both methods improve retrieval dynamically, SELF-RAG optimizes when retrieval occurs, whereas CRAG optimizes the quality of retrieval evidence, representing two different directions in adaptive retrieval research.

Beyond adaptive retrieval, recent studies have expanded retrieval optimization to improve overall pipeline efficiency, robustness and semantic matching. Auto-RAG automatically identifies effective combinations of retrieval, re-ranking, chunking and generation modules, reducing the need for manual pipeline design [18]. Trust-RAG extends retrieval optimization by addressing corpus poisoning attacks and improving the reliability of retrieved evidence under adversarial conditions [20]. Meanwhile, QulM-RAG introduces a question-to-question retrieval paradigm that aligns user queries with previously indexed questions instead of relying on document similarity, thereby improving semantic matching for complex information needs [25]. These approaches demonstrate that retrieval research has evolved beyond improving retrieval accuracy towards building retrieval systems that are adaptive, optimized, trustworthy and task-aware. The effectiveness of these retrieval strategies is reflected in downstream applications. Retrieval-enhanced systems such as TSQA and document-oriented RAG frameworks demonstrate that improved retrieval quality directly enhances evidence relevance, answer accuracy and scientific document understanding
[7],[10]. These findings consistently indicate that retrieval quality remains the most influential factor affecting overall RAG performance across diverse knowledge-intensive tasks. Collectively, the evolution of retrieval strategies reflects a transition from static semantic retrieval to intelligent retrieval optimization. Current research increasingly emphasizes adaptive decision-making, retrieval quality evaluation, automated pipeline optimization, trustworthy retrieval and semantic-aware retrieval paradigms to improve the reliability and effectiveness of RAG systems [2]-[4], [16]-[18],[20],[25]. Other than these advances, most existing approaches continue to rely primarily on document-level retrieval and provide limited support for modeling complex relationships among documents and entities. This limitation has motivated the development of graph-based retrieval methods, which are discussed in the following section.

2.3. Advanced Retrieval Paradigms

As Retrieval Augmented Generation (RAG) has changed over time, researchers have seen that regular dense retrieval is often not enough for tasks that need a lot of knowledge, especially in scientific areas. Even though semantic retrieval can find documents that have similar meanings, it often misses the way things are connected, how ideas link over long distances and how documents interact with each other. Because of this, recent studies have started to focus on ways to retrieve information. These new ways use graph structures, systems that can grow smart ways to pick evidence and learning methods that change as needed. These improvements help RAG systems to think clearly when dealing with extensive and complicated knowledge [19],[23]-[32]. One big area of research is graph-enhanced retrieval. This is where relationships between things and documents are used to give context than the usual way of retrieving information. The light-RAG system is a way to do this. It uses a graph to get information from global contexts. This makes retrieving information efficient and it also keeps the meaning of the relationships between things. The Graph-RAG system takes this idea further. It builds graphs that show how things are related in a hierarchy. It summarizes what groups of things have in common. This allows systems to answer questions that need a picture understanding of a collection of documents rather than just looking at one document at a time. The CG-RAG system is similar. It uses information about which publications cite each other to improve how it answers questions about research. It looks at the words used, what they mean and how they are connected to publications. By looking at the connections between publications, graph-enhanced retrieval can find more complete evidence, in scientific literature than the old way of just looking for similar things.

Beyond graph-based methods, recent studies have focused on improving retrieval for large-scale scientific repositories and long-context reasoning. Hiper-RAG solves the problem of finding information in millions of articles by using fast computing along with smart ways to process documents, break them into meaningful parts and find the right information based on the question. This shows that being able to retrieve information quickly is getting more and more important for discovering knowledge. On the hand, Long-RAG focuses on answering questions that need a lot of context. It uses a way of looking at information from two angles. This helps to get both the big picture and the small details at the same time. These approaches show that modern retrieval systems need to deal with a large amount of information and also understand the context. This is important for RAG applications that are getting more and more complicated. RAG applications need retrieval systems that can handle a lot of things at the same time and also understand what is going on.

Another emerging area of research focuses on intelligent retrieval, where retrieved evidence is refined before response generation. INSTRUCT-RAG strengthens retrieval robustness by guiding language models to identify and justify relevant evidence using self-synthesized rationales. As a result, the model becomes less susceptible to noisy or misleading retrieved documents [29]. Rank-RAG takes a different approach by integrating context ranking with answer generation in a single instruction-tuned language model. These models generate additional training data and they are getting better retrieval performance while reducing manual labeled datasets. Compared with conventional retrieval pipelines, self-improving retrieval helps RAG systems keep up with new information. As a result, these models still maintain retrieval quality over time and achieve better performance in specific areas of knowledge.

RAG retrieval has changed over time. It used to find things that are similar in meaning. Now it is more about understanding relationships between things. This is because of approaches that use graphs to figure out how things are connected and they also make it possible how RAG systems to get better on their own. They can do this because they can handle a lot of information and adapt to things they learn. It's also good at working with areas of knowledge and can keep getting better as it learns more about these areas. RAG retrieval systems can keep improving because they can use information to make themselves better. This is really helpful because it has to work with a lot of kinds of information.

With the great progress from retrieval methods, there still remains the challenge of balancing the retrieval effectiveness, computational efficiency and generalization across different scientific corpora. Next section we focus on solving these limitations that are critical for building reliable RAG systems that support scientific question answering and document summarization effectively.

2.4. Scientific Question Answering and Summarization

In this part, scientific question answering and literature summarization are two important uses of RAG. When we are talking about Scientific question answering, which is not like general-purpose question answering, it requires scientific application responses that are accurate, supported by verifiable citations and capable of integrating evidence for multiple research articles. There are many scientific papers being published and it's getting hard for researchers to manually look through all the literature. This means need an intelligent system that can retrieve and give answers that are based on real references.

Because of these problems, the researchers studying RAG are now working on more than just making the system better at finding information. They are also building systems that support scientific thinking, summarize the literature, and find evidence-based knowledge, as seen in the work of [5]-[15]. Question answering and scientific literature summarization are the main focus of RAG research PaperQA created an agent-based RAG architecture. This system finds text in scientific articles. It checks how relevant the parts it finds are. Then it creates answers that have citations. This approach works better than language models on scientific question-answering tests [5]. PaperQA2 takes what we already know. Adds to it with a new way of working that brings together finding evidence, summing things up in context, looking at citations and thinking things through step by step. When we tested PaperQA2 on LitQA2 and other science tasks, we found that it does as well as, or even better than expects in the field at finding the right information, summing up science and spotting mistakes, which shows that using many steps to look at evidence really works for thinking about science [6].

Several studies have investigated RAG-based document question answering for specific document collections. Retrieval-Augmented Large Language Models have been evaluated for scientific document reasoning, revealing that

although retrieval improves explainability, Models may still generate fabricated evidence when reasoning over scientific documents, emphasizing the need for trustworthy evidence verification [8]. Similarly, PDF Retrieval Augmented Question Answering improves RAG by adding text and non-text parts from PDF files. These parts include tables, pictures, charts and images. This allows answers to questions about complicated scientific papers [9]. The document-oriented framework uses GPT-3.5 with RAG. This helps with document question answering. It does this by combining vector-based retrieval with language generation. Document question answering with GPT-3.5 and RAG shows improvements. These improvements are seen in ROUGE, BLEU, BERTScore and Jaccard similarity. These improvements are compared to other document question answering systems [10]. Beyond question answering, recent research is focusing more on summarizing literature. TSQA adds text summarization into the retrieval process by creating versions of documents before answering questions. By using SBERT-based summarization along with RAG, the system improves retrieval. Answers are more accurate. It shows improvements in text similarity and BERTScore compared to RAG systems that do not use summarization [7]. Open-Scholar takes scientific literature synthesis to the next level with a special retrieval- augmented language model. This model looks through 45 million open-access scientific papers. It finds the evidence, then rearranges the results. It keeps improving the answers by using what it learns from itself. The model is backed by the ScholarQABench benchmark. OpenScholar shows results in citation accuracy. It also gives accurate facts and helpful answers. People often like the answers more than those written by experts according to human testing [11]. More recent work has shifted toward multi-document scientific summarization and knowledge synthesis. Ask Retrieve, Summarize (XSum) introduces a modular RAG pipeline in which dynamically generated questions guide evidence retrieval before an editor module integrates retrieved information into coherent scientific summaries.

This modular design improves summaries. Keeps the right citations from many different doc-uments [12]. In addition, it helps with summarizing documents and RAG has also been used to help with integrating information. This helps cut down on manual effort when reviewing information [13].

Research on summarization is also looking at what is important in a study. This means finding points like why the research was done. What was discovered? What are researchers saying? And what proves it. So, these methods make summaries that really get to the heart of what the scientific articles are saying, rather than conventional abstracts [14].

Improving the robustness of scientific summarization has become an important focus of recent research. Summ-RAG introduces Logic-Summ, an evaluation framework for assessing RAG-based summarization under realistic retrieval scenarios, together with a fine-tuning strategy that improves the logical consistency and robustness of generated summaries. Experimental results demonstrate that summ-RAG maintains high summarization quality even when irrelevant retrieved documents are included, reducing the sensitivity of RAG systems to irrelevant evidence while improving logical coherence during multi-document summarization [15]. The research shows that there has been a change in how people answer questions. In the past, studies focused mostly on finding information to answer one question at a time. Now new systems are combining steps. They bring in information, improve the evidence, use citations to think through problems, summarize documents and put everything together into one system. This is all done using a

RAG structure [5]-[15]. There are still some problems with this stuff. For example, it is hard for computers to figure out what is true when they have conflicting information. They also have to make sure they give credit to the people who wrote the things they are using. Sometimes computers make things up when they are trying to summarize something. That is not good. We need to come up with ways to test how well computers can answer questions and summarize things. These problems are why we need to talk about evaluation frameworks and benchmarks in the following section.

2.5. Evaluation Frameworks and Benchmarks

The quick progress of RAG systems has made it more important to find ways to check how well they perform. These models are different from language models because they have parts that work together. The evaluation of RAG systems brings challenges because performance depends on multiple connected parts, including document retrieval, evidence selection and answer generation. Traditional metrics like BLEU, ROUGE and Exact Match mainly look at text similarity, which makes them not good enough for checking retrieval quality, factual grounding, citation accuracy and reasoning performance. These problems have led to the creation of evaluation frameworks and benchmark datasets that are meant to test RAG systems on many different retrieval and question answering tasks [21]-[24].

The evaluation of RAG systems is important. Researchers are working on making this process automatic. ARES is a system that helps with this. It creates training data and uses it to teach language models. These language models then check if the answers are good. They look at how the answer fits the context, whether the answer is true to the original information and if the answer is relevant to what was asked

A different approach is provided by RAGAS, which focuses on reference-free evaluation of RAG pipelines. Instead of requiring manually annotated reference answers, RAGAS evaluates RAG systems using automated metrics that measure answer faithfulness, answer relevance and context relevance through large language models. This helps developers check how good the search part is and how well the answers are made, even when there are no sets of answers to use. RAGAS works inside LangChain and LlamaIndex. This makes it easy to check things quickly while building the system. RAGAS has become very popular for building retrieval-augmented generation systems because it provides checks without needing answers made by people and works with the most common tools for building these systems [22]. Alongside automated evaluation frameworks, benchmark datasets have become very important for testing retrieval and reasoning abilities in real-life situations. SPIQA solves a problem with earlier scientific question-answering benchmarks by creating the first large-scale dataset that combines different types of information for scientific paper questions. Unlike datasets that only look at text, SPIQA asks models to understand pictures, tables, captions and text all at the same time. The dataset has 270,000 questions taken from 25,859 scientific papers. It adds three ways to test

models: answering questions using pictures and tables, answering questions about the whole paper and using retrieval to think step by step. These capabilities help us look closely at how RAG models deal with kinds of information and long documents. This is important for RAG systems that will be used in science in the future [23]. Retrieval-level evaluation is also supported by BEIR. BEIR is one of the complete zero-shot information retrieval benchmarks. It is a benchmark that helps test and evaluate information retrieval systems. It's known for being one of the comprehensive zero-shot information retrieval benchmarks [24].

It has 18 datasets that are open for everyone to use. These datasets cover nine tasks for finding information. The tasks include answering questions, checking facts, predicting citations, finding arguments and getting documents. The test checks ways of finding information, like using simple words, sparse methods, more complex methods combining ideas at a later stage and improving results again. It shows that no one way of finding information is the best in all situations. Even though some methods are faster, BEIR finds that BM25 works well without needing any special training.

When we look at systems that use late interaction and re-ranking usually do better at finding the right information. However, this comes at a computational cost. One thing to remember here is that we need to test retrieval systems in different areas before we use them in the real world [24]. This is because retrieval systems need to be tested across different domains before we actually use them in real-world applications.

When you look at all these evaluation resources, you can see that RAG research is getting more mature because of these resources. ARES and RAGAS are tools that help us check how good our retrieval quality is, how well our facts are grounded and how well we can generate responses. On the hand, SPIQA and REIR give us datasets to test how well we can reason and retrieve information in a scientific way and how well we can apply what we learn to new situations. RAG research is really improving with these resources. These resources help us look at it in a way they do not just check if two pieces of text are similar.

The resources also check how well we can find the information we need, whether the facts are correct, if we can understand things that are shown to us in a way and if we can apply what we learn to many different areas. These resources are helpful for evaluating the information we have. The resources are very good at helping us evaluate information across different domains. There are still some problems to solve when it comes to figuring out how good RAG systems are. We need to find ways to test how well they can find the information, reason correctly, get citations right and not fall for false information. We also need to see if they can generate pieces of scientific knowledge. RAG systems have to be able to do all these things if we want to make sure we are testing them in a consistent way. This is important for RAG systems to keep getting better.

3. Comparative Discussion and Research Gaps

The development of RAG shows a move from ways of using retrieval with language generation to systems that focus on knowledge and can help with complicated scientific research tasks. As explained in the previous parts. New research has made progress in different parts of the RAG process, like how to find information, better methods for looking things up and answering science questions, making summaries of papers and checking how good the results are. All these improvements have made it easier for language models to find the right information, create answers that have references and bring together data from big collections of scientific corpora. With all this improvement, there are still big problems that have not been solved yet, showing where more research is needed [1]-[32].

A main thing that researchers have found when they look at the research is that information is getting better. Old RAG systems relied on a method that looked at semantic retrieval; new systems can adopt and fix mistakes and they can even look at a lot of information and get better at getting the right information, especially when it comes to scientific applications, and it's important for scientific as it get information from documents. The quality of retrieval still depends on factors like how the documents are organized (document indexing), how words are represented (embedding representations), what the documents are about (corpus characteristics) and how systems decide what information is most important (re-ranking strategies). It's also hard to make these systems work well and not use much computational power, especially when there are millions of publications [16]-[20], [25]-[32].

Table 1: Comparative Evolution of Retrieval-Augmented Generation

Research Direction	Major Contribution	Remaining Limitation
Evaluation	Established retrieval augmented generation paradigm	Limited scientific specialization
Retrieval Strategies	Adaptive, corrective and trustworthy retrieval	Retrieval quality remain corpus dependent
Advanced Retrieval	Graph-based, long-context and intelligent retrieval	Higher computational complexity
Scientific QA & Summarization	Citation-grounded scientific reasoning and knowledge synthesis	Difficulty handling conflicting evidence
Evaluation Frameworks	Automated evaluation and standardized benchmarks	Lack of unified evaluation protocols

Table 1 presents a comparative overview of the evolution of RAG across major research directions. It summarizes the major contributions associated with retrieval strategies, advanced retrieval, scientific question answering and

summarization and evaluation frameworks. The table also highlights the remaining limitations of these approaches. This comparison demonstrates the progression and challenges of modern RAG systems.

A second big trend is the increasing focus on scientific knowledge synthesis. Earlier RAG applications were about answering single questions, but now systems are more helpful with large scientific tasks. These tasks include looking through research papers, using citations properly, checking for contradictions and summarizing documents. Projects like paperQA2 and openScholar show this change by bringing together repeated searching, improving the evidence and creating content that pays attention to citations in one scientific help system. Even though these systems help a lot with putting together research, making sure the facts match across documents and resolving conflicting scientific information are still big challenges that need better thinking skills [5]-[15].

Table 2 summarizes the major research gaps identified from the reviewed literature. These gaps include retrieval robustness, citation reliability, multi-document reasoning, multi-modal reasoning and evaluation standardization. The identified limitations continue to affect the reliability and effectiveness of scientific RAG systems. Addressing these gaps is important for developing trustworthy scientific knowledge systems.

Table 2: Major Research Gaps Identified

Research Gap	Evidence from Literature
Retrieval robustness	Retrieval remains sensitive to corpus quality and indexing strategies
Citation reliability	Citation errors and unsupported claims still occur
Multi-document reasoning	Difficulty integrating conflicting evidence
Multi-modal reasoning	Limited support for figures, tables and diagrams
Evaluation standardization	Existing benchmarks evaluate different aspects independently

The studies that were looked at also show that a lot of progress has been made in how we evaluate things. The old ways of measuring how similar texts are are being replaced by frameworks that check how well we can retrieve information, how accurate our answers are and how well we can reason. We have tools like ARES and RAGAS that make it easier to evaluate RAG pipelines. We also have datasets like SPIQA and BEIR that give us a way to test how well we can understand and retrieve information in a scientific way. However, the way we evaluate things is still not very organized. Each benchmark only looks at one part of how RAG systems work, rather than giving us a complete picture. This makes it hard to compare RAG systems, especially when we are looking at complicated things like scientific reasoning, citation quality and how we can use information from many documents [21]-[24]. Another big issue is figuring out how to stop RAG systems from making things up and how to make sure the citation reliable. Although retrieval substantially reduces hallucinations compared with standalone language models, several reviewed studies demonstrate that retrieval alone cannot guarantee factual correctness. Generated responses may still contain unsupported claims, inaccurate citations or reasoning inconsistencies when retrieved evidence is incomplete, conflicting or only partially relevant. Developing mechanisms that improve evidence verification, citation fidelity and uncertainty estimation therefore remains an essential direction for future research, particularly in domains such as scientific publishing, medicine and legal analysis [6],[11],[15],[21],[22].

Table 3 presents the major future research opportunities identified for RAG-based scientific knowledge systems. These opportunities include adaptive retrieval, graph-based RAG, Multi-modal RAG, genetic RAG and unified evaluation. These research directions aim to improve reasoning, retrieval, knowledge integration and system reliability. These also provide potential pathways for developing more advanced scientific AI systems.

Table 3: Future Research Opportunities

Research Direction	Expected Development
Adaptive Retrieval	Dynamic retrieval strategies based on query complexity
Graph-based RAG	Improved multi-hop reasoning over scientific knowledge
Multimodal RAG	Unified reasoning across text, figures, table and diagrams
Agentic RAG	Autonomous literature review and scientific discovery assistants
Evaluation	Unified benchmarks integrating retrieval, reasoning, citation quality and hallucination assessment

Finally, the literature indicates an increasing movement toward adaptive and autonomous RAG systems capable of continuously improving retrieval quality, integrating heterogeneous knowledge sources, and supporting increasingly complex reasoning tasks. Future RAG research is expected to combine graph-based knowledge representation, multi-modal retrieval, self-improving retrieval strategies, and agentic reasoning within unified retrieval-generation architectures. At the same time, greater emphasis will be required on standardized evaluation, computational efficiency, explainability, and reproducibility to ensure that increasingly sophisticated RAG systems remain reliable for real-world scientific applications [26]–[32].

Overall, the reviewed literature demonstrates that RAG has evolved into a mature and rapidly expanding research area with significant potential for scientific question answering and document summarization. Substantial progress has been

achieved in retrieval optimization, knowledge synthesis and evaluation. Important challenges remain in retrieval robustness, reasoning over conflicting evidence, citation reliability, multi-modal knowledge integration and standardized evaluation. Addressing these challenges will play a central role in advancing the next generation discussed in the following chapter.

Figure 2 illustrates the overall scientific RAG ecosystem and major components. It shows the relationship between scientific literature, retrieval techniques, RAG, question answering, summarization, knowledge synthesis and evaluation frameworks. The figure also connects these components with future research directions. This provides an overall view of how RAG can support scientific knowledge exploration and synthesis.

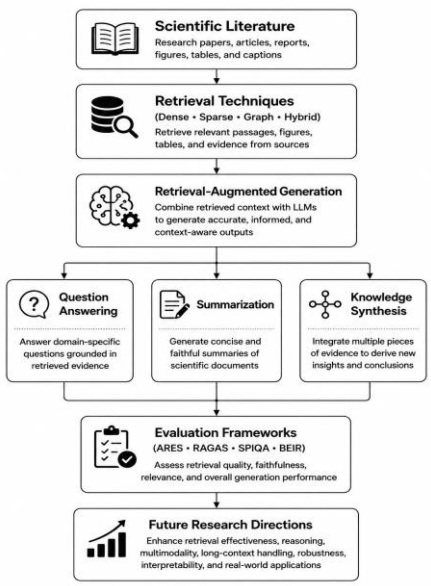


Figure 2: Scientific RAG ecosystem showing scientific literature, retrieval techniques, retrieval-augmented generation, question answering, summarization, knowledge synthesis, evaluation frameworks, and future research directions.

4. Open Challenges and Future Directions

Despite the significant progress achieved by RAG, several open challenges continue to limit its effectiveness in scientific question answering and document summarization. As we discussed throughout this review, recent studies have substantially improved retrieval accuracy, knowledge synthesis, citation-grounded response generation and automated evaluation. However, increasing publication volumes, heterogeneous scientific knowledge and the demand for trustworthy evidence-based responses require further advances in retrieval architectures, reasoning mechanisms, multi-modal understanding, evaluation methodologies and system scalability [1] – [32]. One of the most important research directions is the development of adaptive retrieval mechanisms capable of dynamically selecting retrieval strategies according to query complexity, information requirements, and corpus characteristics. Recent studies have demonstrated considerable improvements through adaptive, corrective, graph-enhanced, and long-context retrieval approaches, enabling more effective evidence selection and contextual understanding [16]–[20], [25]–[32]. However, retrieval quality remains highly dependent on document indexing strategies, embedding representations, corpus quality, and computational resources. Future research should therefore investigate retrieval frameworks that automatically optimize retrieval depth, evidence diversity, re-ranking strategies, and retrieval confidence while maintaining computational efficiency for large-scale scientific repositories.

4.1. Adaptive Retrieval Mechanism

One of the most important research directions is the development of adaptive retrieval mechanisms capable of dynamically selecting retrieval strategies according to query complexity, information requirements and corpus characteristics. Recent studies have demonstrated considerable improvement through adaptive, corrective, graph-enhanced and long-context retrieval approaches, enabling more effective evidence selection and contextual understanding [16]– [20], [25] -[32]. However, retrieval quality remains highly dependent on document indexing strategies and embedding retrieval frameworks that automatically optimize retrieval depth, evidence diversity, re-ranking strategies and retrieval confidence while maintaining computational efficiency for large-scale scientific repositories.

4.2. Reasoning Across Multiple Scientific Documents

Another important challenge involves reasoning across multiple scientific documents. Although recent systems have substantially improved citation-grounded scientific question answering and literature summarization, generating logically consistent responses from complementary, contradictory or evolving scientific evidence remains difficult. Scientific assistants increasingly require the ability to compare experimental findings, identify disagreements, estimate evidence confidence and synthesize comprehensive conclusions from multiple publications. Future RAG systems should therefore incorporate more sophisticated evidence aggregation, contradiction detection, uncertainty estimation and citation-aware reasoning mechanisms to improve the reliability and transparency of generated scientific knowledge [5] – [15].

4.3. Multi-modal Scientific Information

The growing availability of multi-modal scientific information presents another promising re-search direction. Scientific publications communicate knowledge not only through text but also through figures, tables, mathematical expressions, experimental workflows and diagrams. Recent benchmark datasets have highlighted the importance of multi-modal reasoning for scientific question answering, yet most existing RAG systems remain predominantly text-oriented [9], [23]. Future research should investigate unified retrieval and reasoning architecture capable of jointly processing heterogeneous information sources, enabling comprehensive understanding of complex scientific documents and improving performance on multimodal research tasks

4.4. Agentic RAG Systems

Another emerging direction is the development of agentic RAG systems. Instead of relying on a single retrieval-generation cycle, future scientific assistants are expected to perform iter-ative retrieval, evidence refinement, planning, self-verification and citation validation through autonomous reasoning processes. Recent scientific retrieval and advanced retrieval paradigms demonstrate the benefits of iterative evidence refinement and intelligent retrieval strategies, providing a foundation for more autonomous scientific assistants [5], [6], [11],[13],[26]-[32]. Future agentic architectures may coordinate multiple specialized agents responsible for literature retrieval, evidence verification, contradiction analysis, citation checking and summary generation, thereby improving robustness, explainability and reproducibility.

4.5. Standardized Evaluation Protocols

Although considerable progress has been achieved in automated evaluation, standardized evalu-ation protocols remain an important research challenge. Existing evaluation frameworks and benchmark datasets assess different aspects of RAG, including retrieval relevance, factual ground-ing, answer faithfulness, multi-modal reasoning and retrieval generalization. However, these evaluation methodologies remain fragmented, making direct comparison between RAG systems difficult [21]-[24]. Future research should establish unified evaluation frameworks capable of jointly measuring retrieval effectiveness, reasoning quality, citation accuracy, hallucination resistance, computational efficiency and user trust across diverse scientific applications. Such standardized benchmarks would facilitate reproduction evaluation and accelerate the development of more reliable RAG systems.

4.6. Scalability and Real-World Deployment

Finally, scalability and real-world deployment remain critical challenges as scientific repositories continue to expand rapidly. Modern scientific assistants must efficiently retrieve, index and reason over millions of continuously evolving publications while maintaining low latency and acceptable computational cost. Recent studies have demonstrated the importance of scalable retrieval architectures, efficient indexing strategies and robust evaluation across heterogeneous corpora, yet practical deployment at large scale remains computationally demanding [10], [11], [20], [24], [31], [32]. Future work should therefore focus on scalable vector databases, incremental embedding updates, distributed retrieval architectures and efficient retrieval-generator collaboration to support real-time scientific knowledge management. Overall, the literature we reviewed in this paper demonstrates that RAG has evolved from conventional retrieval-assisted language generation into an increasingly sophisticated paradigm for scientific knowledge synthesis. Continued progress will require advances in adaptive retrieval, multi-document reasoning, multi-modal understanding, autonomous agentic workflows, standardized evaluation and scalable deployment. Addressing these challenges will improve the reliability, transparency and practical applicability of future RAG systems, enabling more effective support for scientific literature exploration, evidence-based question answering and automated knowledge synthesis [1] – [32].

Table 4 summarizes the major open challenges and corresponding future research directions for scientific RAG systems. The challenges include adaptive retrieval, multi-document reasoning, multi-modal scientific understanding, agentic RAG evaluation standardization and scalability. The table links each limitation with a potential research direction. These directions can help improve the reliability, efficiency and practical deployment of future RAG systems.

Table 4: Open Challenges and Future Research Directions

Research Challenge	Current Limitation	Future Research Direction
Adaptive Retrieval	Static retrieval strategies and corpus dependency	Query-aware adaptive retrieval and dynamic re-ranking
Multi-document Reasoning	Difficulty synthesizing conflicting evidence	Citation-aware reasoning and evidence verification
Multi-modal Scientific Understanding	Predominantly text-based retrieval	Joint reasoning over text, figures, tables and diagrams
Agentic RAG	Limited autonomous planning and verification	Multi-agent retrieval, planning and self-verification
Evaluation Standardization	Fragmented evaluation methodologies	Unified benchmarks integrating retrieval, reasoning, citation quality and hallucination assessment
Scalability and Deployment	High computational cost for large scientific repositories	Efficient indexing, distributed retrieval and incremental knowledge updating

5. CONCLUSION

Retrieval Augmented Generated has become a way to make large language models better by adding external knowledge retrieval to language generation. This review explores how Retrieval Augmented Generation has changed from retrieval-based language generation to more complex systems that use adaptive retrieval, graph-based methods, and intelligent evidence selection. Also, scientific question answering, document summarization, and specialized evaluation frameworks are other advanced applications. These changes have made Retrieval Augmented Generation systems better at generating responses that are based on facts and supporting applications that need a lot of knowledge. This reviewed literature shows that recent developments have taken RAG beyond retrieving documents to integrating scientific knowledge. Modern Retrieval Augmented Generation systems can support answering questions based on citations, reasoning with documents, exploring literature and summarizing evidence and new evaluation frameworks can assess how well the system retrieves and generates information. With all this progress, there are still some challenges, like making retrieval more robust, reasoning with conflicting evidence, understanding multiple types of information, standardizing evaluation and making the system more efficient and scalable for large scientific databases. In general RAG is an area of research that has a lot of potential to improve how we access and discover scientific information. We need research on how to retrieve information, how to reason with that information, how to evaluate the system and how to make the system scalable to improve the reliability, transparency and effectiveness of future Retrieval Augmented Generation systems. These developments could lead to scientific assistants that help researchers find the right evidence, combine knowledge and make good decisions. This will be very helpful for researchers who use RAG systems. This will be helpful for researchers who use RAG systems. The future of Retrieval Augmented Generation looks very promising. It will continue to play a role in improving research because researchers will keep using RAG to make discoveries.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to Parul University, Vadodara, India, for providing the academic environment and support necessary for this research. I also acknowledge the valuable guidance and supervision provided by Kiran Macwan and AnilKumar Patel throughout the development of this review.

Funding

Funding not applicable. there was no financial support or external funding for this research

Authors' Contributions

Ahmed Hassan: Contributed to Conceptualization, literature review and analysis, synthesis of the reviewed research, identification of research gaps, manuscript preparation and revision

Kiran Macwan: contributed to supervision, academic guidance, critical and revision of the manuscript.

AnilKumar Patel: is a supervisor and contributed to academic guidance, critical review and revision of the manuscript.

All authors: Reviewed and approved the final version of the manuscript

Conflict of Interest

The authors declare that they have no conflict of interest

Data Availability

Data sharing does not apply to this article as no new datasets were generated or analyzed during the study. This manuscript is a review of previously published literature, and all sources used in the review are appropriately cited in the reference list.

REFERENCES

1. Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang, "Retrieval-Augmented Generation for Large Language Models: A Survey," arXiv preprint arXiv:2312.10997, 2024.
2. S. Wu, Y. Xiong, Y. Cui, H. Wu, C. Chen, Y. Yuan, L. Huang, X. Liu, T.-W. Kuo, N. Guan, and C. J. Xue, "Retrieval-Augmented Generation for Natural Language Processing: A Survey," arXiv preprint arXiv:2407.13193, 2026.
3. C. Sharma, "Retrieval-Augmented Generation: A Comprehensive Survey of Architectures, Enhancements, and Robustness Frontiers," arXiv preprint arXiv:2506.00054, 2025.
4. Y. Huang and J. X. Huang, "A Survey on Retrieval-Augmented Text Generation for Large Language Models," ACM Computing Surveys, vol. 58, no. 12, Art. no. 300, May 2026, doi: 10.1145/3805774.
5. J. Lala, O. O'Donoghue, A. Shtedritski, S. Cox, S. G. Rodriques, and A. D. White, "Pa-perQA: Retrieval-Augmented Generative Agent for Scientific Research," arXiv preprint arXiv:2312.07559, 2023.
6. M. D. Skarlinski, S. Cox, J. M. Laurent, J. D. Braza, M. Hinks, M. J. Hammerling, M. Ponnapati, S. G. Rodriques, and A. D. White, "Language Agents Achieve Superhuman Synthesis of Scientific Knowledge," arXiv preprint arXiv:2409.13740, 2024
7. A. S. Jaddoa, J. Karimpour, and P. Salehpour, "TSQA: Integrating Text Summariza-tion and Question Answering to Improve Information Retrieval from Documents Using Retrieval-Augmented Generation," Information, vol. 17, no. 4, Art. no. 372, 2026, doi: 10.3390/info17040372.
8. S. Munikoti, A. Acharya, S. Wagle, and S. Horawalavithana, "Evaluating the Effectiveness of Retrieval-Augmented Large Language Models in Scientific Document Reasoning," in Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024), 2024,pp. 84–89.
9. T. T. U. Hoang, M. Rajendran, K. Zhang, Y. Wu, and V. A. Nguyen, "PDF Retrieval Augmented Question Answering," arXiv preprint, 2024.

10. K. Muludi, K. M. Fitria, J. Triloka, and Sutedi, "Retrieval-Augmented Generation Approach: Document Question Answering Using Large Language Model," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 15, no. 3, pp. 776–785, 2024.
11. A. Asai, J. He, R. Shao, W. Shi, A. Singh, J. C. Chang, K. Lo, L. Soldaini, S. Feldman,
M. D'Arcy, et al., "Synthesizing Scientific Literature with Retrieval-Augmented Language Models," Nature, vol. 650, pp. 857–867, Feb. 2026.
12. P. Achkar, T. Gollub, and M. Potthast, "Ask, Retrieve, Summarize: A Modular Pipeline for Scientific Literature Summarization," in Proceedings of the 2025 Conference on Scholarly Document Processing, 2025.
13. B. Han, T. Susnjak, and A. Mathrani, "Automating Systematic Literature Reviews with Retrieval-Augmented Generation: A Comprehensive Overview," Applied Sciences, vol. 14, no. 19, Art. no. 9103, Oct. 2024, doi: 10.3390/app14199103.
14. M. Majadly and M. Last, "Leveraging Peer-Review Aspects for Extractive and Abstractive Summarization of Scientific Articles," International Journal of Data Science and Analytics, vol. 19, pp. 537–550, 2025, doi: 10.1007/s41060-024-00665-z.
15. S. Liu, J. Wu, J. Bao, W. Wang, N. Hovakimyan, and C. G. Healey, "Towards a Robust Retrieval-Based Summarization System," arXiv preprint arXiv:2403.19889, 2024.
16. A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "SELF-RAG: Learning to Retrieve, Gen-erate, and Critique Through Self-Reflection," in Proceedings of the International Conference on Learning Representations (ICLR), 2024.
17. Anonymous Authors, "Corrective Retrieval Augmented Generation (CRAG)," International Conference on Learning Representations (ICLR) 2025, under review, 2025.
18. H. Zhou, K.-H. Lee, Z. Zhan, Z. Li, Y. Chen, Z. Wang, H. Haddadi, and E. Yilmaz, "TrustRAG: Enhancing Robustness and Trustworthiness in Retrieval-Augmented Genera-tion," arXiv preprint arXiv:2501.00879, 2025.
19. Z. Guo, L. Xia, Y. Yu, T. Ao, and C. Huang, "LightRAG: Simple and Fast Retrieval-Augmented Generation," arXiv preprint, 2024.
20. D. Kim, B. Kim, D. Han, and M. Eibich, "AutoRAG: Automated Framework for Optimization of Retrieval-Augmented Generation Pipeline," arXiv preprint arXiv:2410.20878, 2024.
21. J. Saad-Falcon, O. Khattab, C. Potts, and M. Zaharia, "ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems," in Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2024), pp. 338–354, 2024.
22. S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated Evaluation of Retrieval Augmented Generation," in Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL 2024), pp. 150–158, 2024.
23. S. Pramanick, R. Chellappa, and S. Venugopalan, "SPIQA: A Dataset for Multimodal Question Answering on Scientific Papers," in Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024), Datasets and Benchmarks Track, 2024.
24. N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, "BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models," arXiv preprint arXiv:2104.08663, 2021.
25. B. Saha, U. Saha, and M. Z. Malik, "QuIM-RAG: Advancing Retrieval-Augmented Genera-tion with Inverted Question Matching for Enhanced QA Performance," IEEE Access, vol. 12, 2025, doi: 10.1109/ACCESS.2024.3513155.
26. D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitansky, R.Ness, and J. Larson, "From Local to Global: A GraphRAG Approach to Query-Focused Summarization," arXiv preprint arXiv:2404.16130, 2025.
27. Y. Hu, Z. Lei, Z. Dai, A. Zhang, A. Angirekula, Z. Zhang, and L. Zhao, "CG-RAG: Re-search Question Answering by Citation Graph Retrieval-Augmented LLMs," arXiv preprint arXiv:2501.15067, 2025.
28. Gokdemir, C. Siebenschuh, A. Brace, A. Wells, B. Hsu, K. Hippe, P. V. Setty, A. Ajith,J. G. Pauloski, V. Sastry, et al., "HiPerRAG: High-Performance Retrieval Augmented Generation for Scientific Insights," in Proceedings of the Platform for Advanced Scien-tific Computing Conference (PASC '25), Brugg-Windisch, Switzerland, Jun. 2025, doi: 10.1145/3732775.3733586.
29. Z. Wei, W.-L. Chen, and Y. Meng, "INSTRUCTRAG: Instructing Retrieval-Augmented Generation via Self-Synthesized Rationales," in Proceedings of the International Conference on Learning Representations (ICLR 2025), 2025.
30. J. Chen, H. Lin, X. Han, and L. Sun, "Benchmarking Large Language Models in Retrieval-Augmented Generation," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, 2024.
31. Q. Zhao, R. Wang, Y. Cen, D. Zha, S. Tan, Y. Dong, and J. Tang, "LongRAG: A Dual-Perspective Retrieval-Augmented Generation Paradigm for Long-Context Question Answer-ing," arXiv preprint arXiv:2410.18050, 2024.
32. Y. Yu, W. Ping, Z. Liu, B. Wang, J. You, C. Zhang, M. Shoeybi, and B. Catanzaro, "RankRAG: Unifying Context Ranking with Retrieval-Augmented Generation in LLMs," arXiv preprint arXiv:2407.02485, 2024.
33. R. Xu, H. Liu, S. Nag, Z. Dai, Y. Xie, X. Tang, C. Luo, Y. Li, J. C. Ho, C. Yang, and Q. He, "SimRAG: Self-Improving Retrieval-Augmented Generation for Adapting Large Language Models to Specialized Domains," in Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 11534–11550, 2025.