SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14824
e-ISSN : 2541-2019
p-ISSN : 2541-044X

A Systematic Review of Retrieval-Augmented Generation for Enhancing Domain-Specific Knowledge in Large Language Models

Murtiyoso[1)*], Imam Tahyudin[2)], Berlilana[3)]
1)2)3) Universitas Amikom Purwokerto, Indonesia
[1)]moertiyoso@gmail.com, [2)] imam.tahyudin.amikompurwokerto.ac.id, [3)]berli@amikompurwokerto.ac.id

Submitted :May 26, 2025 | **Accepted** : Jun 19, 2025 | **Published** : Jun 20, 2025

Abstract: This literature review examines the use of Retrieval-Augmented Generation (RAG) in enhancing Large Language Models (LLM) for domain-specific knowledge. RAG integrates retrieval techniques with generative models to access external knowledge sources, addressing the limitations of LLMs in handling specialized information. By leveraging external data, RAG improves the accuracy and relevance of generated content, making it particularly useful in fields that require detailed and up-to-date knowledge. This review highlights the effectiveness of RAG in overcoming challenges such as data sparsity and the dynamic nature of specialized knowledge. Furthermore, it discusses the potential of RAG to enhance LLM performance, scalability, and the ability to generate contextually accurate responses in knowledge-intensive applications. Key challenges and future research directions in the implementation of RAG for domain-specific knowledge are also identified.

Keywords: information retrieval; retrieval augmented generation; large language model; text generation; contextualized language models

INTRODUCTION

In recent years, the development of Large Language Models (LLMs), such as GPT and BERT, has led to significant breakthroughs in Natural Language Processing (NLP). Large language models (LLMs) are artificial intelligence (AI) tools specifically trained to process and generate text. LLMs attracted substantial public attention after OpenAI's ChatGPT was made publicly available in November 2022. LLMs can often answer questions, summarize, paraphrase, and translate text on a level that is nearly indistinguishable from human capabilities. (Clusmann et al., 2023). These models demonstrate exceptional capabilities across various language tasks, such as translation, text generation, and context understanding. However, despite the vast capacity of LLMs to learn patterns from available data, they face major challenges in terms of specific knowledge that may not be present or adequately represented in their training data. This results in limitations in terms of accuracy and completeness of answers, particularly in highly technical or specialized domains.

As an effort to overcome these limitations, the concept of Retrieval-Augmented Generation (RAG) has emerged as an innovative solution. RAG combines two main components in language processing: retrieval and generation. Essentially, RAG enhances the capabilities of generative models by integrating external information through the retrieval of relevant documents from a larger knowledge base. This allows LLMs to access and utilize specific information that is not part of the model's parameters but is relevant to the context or query at hand.

The main advantage of using RAG in LLMs lies in its ability to improve the accuracy and relevance of answers in contexts that require specialized knowledge. By accessing up-to-date, evidence-based information through retrieval, LLMs can provide more detailed and data-driven responses, especially in domains with rapidly evolving knowledge. (Li et al., 2023) Additionally, RAG can help reduce overfitting to training data by expanding the model's knowledge base, making it more flexible and responsive to questions that require new knowledge or are beyond the scope of the model's prior training.

The application of RAG in LLM is particularly useful in scenarios where specific knowledge is critical, such as in areas such as energy, construction, and health to traditional medicine (Zijuan et al., 2023). Through intelligent information retrieval, the model can extract evidence-based, up-to-date information that cannot be learned directly from training data alone. Therefore, the integration of information-based techniques in the text generation process gives LLM the ability to provide more informative, detailed, contextually relevant answers and the ability to perform source code summarization (Choi et al., 2023).

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14824
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Although research on Retrieval-Augmented Generation (RAG) in Large Language Models (LLM) has been growing, there has been no systematic study specifically addressing the application of RAG in a specific knowledge domain. Each domain, such as energy, construction, and health, has unique challenges related to the accuracy and relevance of the information generated. Therefore, the application of RAG in these domains requires further research to optimize the selection of data sources, reduce bias, and improve the quality of text generation. This research will provide important insights into how RAG can be tailored to the specific needs of each domain. This article aims to present a literature review on the implementation of RAG in LLMs, with a specific focus on its use in specialized knowledge domains. We will discuss various approaches that have been developed to integrate retrieval into generative frameworks, the challenges encountered, and the applications and potential for further development in highly technical and detailed knowledge areas.

METHOD

This study used a systematic literature review method according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) protocol. This protocol ensures that only relevant and high-quality articles are analyzed, so that the research results can effectively capture the latest trends in the application of RAG in the process of providing knowledge for Large Language Models. The search was conducted across Scopus and IEEE parameters manually to ensure accuracy and consistency. The process analysis was carried out with a thematic approach and sectoral categorization to identify trends, challenges, and solutions in the application of RAG. Of the 317 articles identified, the articles were entered into Microsoft Excel, then the articles were filtered based on the title and abstract. After a thorough quality assessment, 21 articles with a score above six were included in the analysis. Non-scientific articles, articles with inadequate evaluation results, and discussions limited to conventional methods were excluded from the observation. Each selected article was evaluated based on its relevance to the topic, clarity of methodology, and the results of using RAG in improving the quality of answers. This study uses descriptive analysis to reveal patterns, trends, challenges, and solutions in the application of RAG to improve the performance of Large Language Models.

Research Questions (RQ)

This literature review is focused on answering the following research questions:

RQ1: What specific areas most commonly use RAG for enhancing Large Language Models?
RQ2: How is the quality of responses when RAG is used for enhancing Large Language Models?
RQ3: What challenges and solutions have been identified in the application of Retrieval Augmented Reality for enhancing knowledge in Large Language Models?

The analysis of 21 selected articles reveals several key trends in the application of Retrieval Augmented Generation (RAG) for enhancing knowledge in Large Language Models. These trends include the dominant use of RAG for specific applications. It suggests the use of RAG models to improve the ability of Large Language Models to access and update knowledge accurately. This method is a relatively new approach designed to enhance the capabilities of trained language models by integrating them with external data. With the addition of this external data, the knowledge possessed by the Large Language Model will increase, and naturally, the quality of the responses provided should become more aligned with user expectations.

The importance of combining RAG is increasingly evident in various studies, where both approaches deliver maximum accuracy for the model. The use of RAG to obtain relevant information for domain-specific adjustments has been shown to significantly enhance the model's performance. These studies demonstrate that although RAG sometimes requires additional data, its initial use will provide an update of additional knowledge to the Large Language Model being used.

Search Process

The literature search process was conducted systematically through several academic databases, namely Scopus, IEEE Xplore, and ScienceDirect. These sources were selected based on their comprehensive coverage of high-quality scientific publications and their relevance to the field of information technology. A combination of keywords was used to ensure comprehensive search results. The search focus was limited to the time range of 2020–2025 to ensure the relevance and currency of the research.

The search process followed these stages:

1. Keywords were identified based on the main topic and related terms (e.g., large language model, retrieval augmented generation).
2. Initial search results were filtered based on relevant titles and abstracts.
3. Keywords were combined with Boolean operators such as AND and OR to expand or narrow the scope of journal searches.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14824
e-ISSN : 2541-2019
p-ISSN : 2541-044X

4. Further selection was conducted to ensure the literature aligned with the research focus.

The following table shows the sources and keywords used in the search:

Table 1. Sources and Keywords

Sumber	Keywords
Scopus	("large language model" OR "retrieval augmented generation") AND ("large language model")
IEEE Xplore	("large language model" OR "retrieval augmented generation") AND ("large language model")
ScienceDirect	("large language model" OR "retrieval augmented generation") AND ("large language model")

Table 1 shows the combination of keywords used for literature searches in Scopus, IEEE Xplore, and ScienceDirect. These combinations were designed to cover research on Retrieval-Augmented Generation (RAG) in Large Language Models. The results in this table ensure that the search scope encompasses the methods of using RAG.

Inclusion and Exclusion Criteria

The selected literature for this review was filtered based on six strict inclusion and exclusion criteria. With this filtering, these criteria ensure that only relevant and up-to-date research is considered.

Table 2. Inclusion and Exclusion Criteria

Inclusion	Exclusion
Publications in the 2020–2025 timeframe	Publications before 2020
English articles or valid translations	Articles in languages other than English without translation
Focuses on the use of Retrieval Augmented Generation for updating large language model information	An article that only discusses large language model methods.
Studies related to large language model in various sectors	Research outside the large language model
Peer-reviewed journal or conference articles	Books, theses, or articles that have not been peer- reviewed
Contains experimental and evaluation results with clear metrics	Conceptual studies without empirical evaluation

Table 2 summarizes the inclusion and exclusion criteria used to filter the literature in this study. The included articles focus on the application of RAG to enhance knowledge in Large Language Models, published between 2020 and 2025, and provide empirical evaluation with clear metrics. Articles in languages other than English or those without the application of Retrieval-Augmented Generation and Large Language Models were excluded. The application of these criteria ensures that only relevant and high-quality articles are analyzed, thereby supporting the focus of the research on current trends and the use of RAG in enhancing knowledge in Large Language Models.

Quality Assessment Process

Quality assessment (QA) ensures that only relevant and high-quality literature is used in this research. Each piece of literature selected through the initial search stage is evaluated based on quality criteria to ensure reliable sources, reliability, and contribution to the research topic. This assessment includes content relevance, research methods used, data completeness, and transparency in reporting results and conclusions. Here is the checklist or rubric used to conduct the quality assessment:

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14824
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 3. Quality Assessment Checklist or Rubric

Criteria	Assessment Questions	Score (0-2)
Relevance	Is the article relevant to the topic of augmented reality generation on large language models?	0-2
Methodology	Is the research methodology described clearly and replicable?	0-2
Evaluation Matrix	Does the study use appropriate evaluation metrics for the use of augmented reality generation on large language models?	0-2
Data Completeness	Is the data used complete and valid enough for analysis?	0-2
Conclusions and Implications	Are the conclusions supported by the results obtained and are their implications clear?	0-2
Contains experimental and evaluation results with clear metrics.	Conceptual study without empirical evaluation	0-2

From the Quality Assessment Process, a score will be obtained. The maximum score is 10. Literature with a score below five will be excluded from the review, ensuring the quality of the findings is maintained.

Data Collection and Data Analysis

Data collection process in this study consists of several steps:

1. Articles that passed the initial screening and QA stages were analyzed. Each article was summarized based on the methods used, namely retrieval augmented generation on large language models, reading evaluation metrics, and challenges found.
2. Data from each relevant study was stored in a spreadsheet with specific attributes (e.g., publication year, source, method, and results).
3. Descriptive analysis techniques were used to identify patterns and trends in the use of retrieval augmented generation on large language models. Key indicators analyzed included the use of retrieval augmented generation, model performance, and the quality of generated answers.

PRISMA Diagram for Literature Selection Process

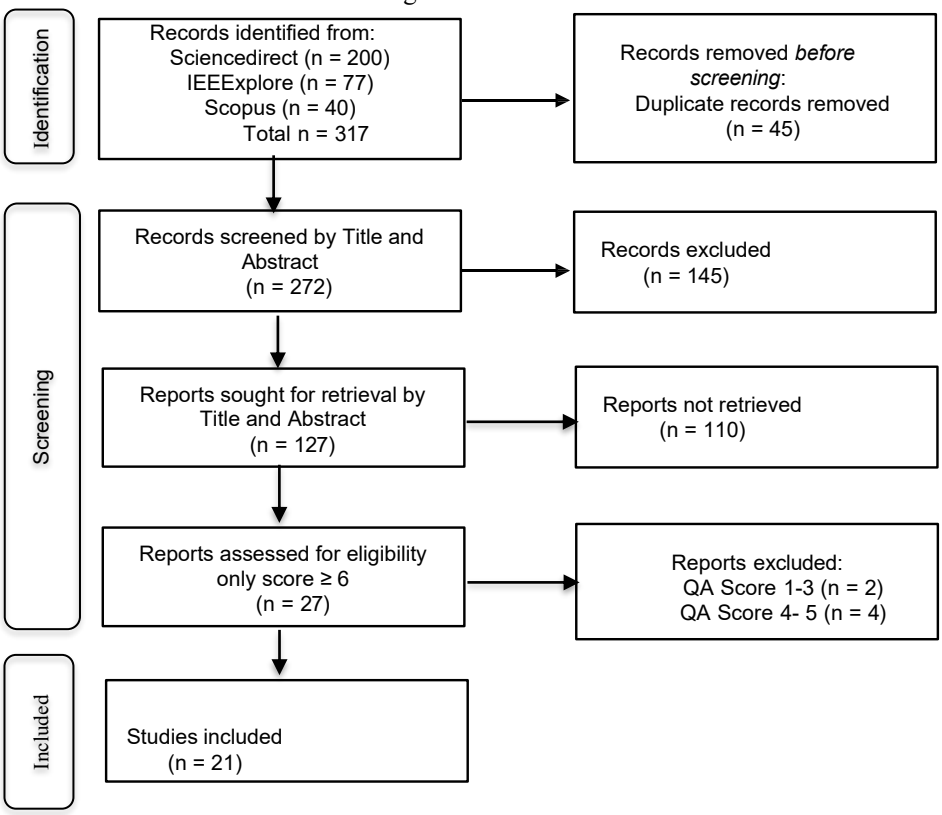

Fig. 1 Systematic process used in this literature review

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14824
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Analysis of 21 selected articles reveals several key trends in the application of Retrieval-Augmented Generation (RAG) to augment knowledge in Large Language Models (LLMs). These trends include the dominance of RAG for specific knowledge domains. This article suggests the use of RAG models to enhance the ability of LLMs to access and manipulate knowledge accurately. This method is a relatively new approach designed to enhance the capabilities of previously mined language models by integrating them through external information knowledge retrieval mechanisms, which help in acquiring through training on large language models, or updating knowledge using available external data. The use of RAG can improve the accuracy of answers up to 99% (Alkhalaf et al., 2024).

Results and Discussion of Each Research Question

RQ1: What fields of science most commonly use RAG for enhancing Large Language Models?

Literature analysis reveals that Retrieval Augmented Generation (RAG) methods dominate the use in enhancing the knowledge of Large Language Models. Retrieval Augmented Generation (RAG) has become the dominant approach in updating knowledge in large language models, mainly due to its ability to dynamically integrate external information into the text generation process. Using retrieval techniques, RAG can extract relevant information from external sources such as large databases or documents, which are then processed and combined with the model's generation capabilities to produce more informative, relevant, and up-to-date knowledge-based answers or texts. (Saha et al., 2024). The most common fields of science using RAG include healthcare, manufacturing, energy, and agriculture.

RQ2: How is the quality of answers when RAG is used to enhance Large Language Models?

Meanwhile, RAG brings additional benefits by combining retrieval techniques to obtain information from a broader range of external sources, and then using generation techniques to compose more informative and detailed text. This method is highly effective in overcoming the limitations of static knowledge possessed by the model, allowing it to access more up-to-date and relevant information without requiring intensive retraining. This makes development easier.

Compared to conventional methods that rely entirely on knowledge within the model's parameters, the quality of answers provided after using Retrieval Augmented Generation increases, meeting expectations.

Table 4.
Specific fields of science that use RAG in Large Language Models

No	Study	Implementation Sector	Result
1.	Rongfan Liu et al., 2025	Battery Recycling	The accuracy of this system reaches around 95%
2.	Jingchi Jiang et al., 2025	Agriculture	The average increase in BertScore for models using RAG (TSRAG) is around +26.66%. The average increase in BLEU (Mean BLEU) is around +20.92%.
3.	Muhammad Arslan et al., 2024	Sustainable Energy	Precision increased to 94.2%. Recall increased to 95%. Accuracy reached 94%.
4.	K.B. Mustapha et al., 2024	Manufaktur	Accuracy of answers increased by 30-40%
5.	Libo Qin et al., 2024	Language	The accuracy of the answers increased by 70%-80%.
6.	Lun-Chi Chen et al., 2024	Industry	Answer accuracy increased by 30%. Answer search is faster.
7.	Jiawei Shao et al., 2024	Wireless Communications	Accuracy of answers increased by 40%
8.	Akhila Abdulnazar et al., 2024	Medical	F1 score increased by 6.87%
9.	Vallidevi Krishnamurthy et al., 2024	Fact Checking	Accuracy of answers increased by 30%
10.	Luis Bernardo Hernandez Salinas et al., 2024	Intelligent Driver Assistance Systems	F1 score increased by 94 %

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14824
e-ISSN : 2541-2019
p-ISSN : 2541-044X

11.	Menna Fateen et al., 2024	Education	Accuracy of the answers increased by 9%.
12.	Biplov Paneru et al., 2024	Agriculture	Accuracy of the answers increased by 95%.
13.	Xuan Liu et al., 2024	Manufactur	Accuracy of the answers increased by 91%.
14.	Seongwoo Cho Xuan Liu et al., 2024	Manufactur	Accuracy of the answers increased by 79%.
15.	Jean Pierre Nayinzira et al., 2024	Health	Accuracy of the answers increased by 10%.
16.	Mohammad Alkhalaf Jean Pierre Nayinzira et al., 2024	Health	With RAG, the accuracy increased to 99% which was only 93% without RAG.
17.	Jos´e Antonio Heredia Alvaro et al., 2024	Manufactur	Accuracy of the answers increased by 60%.
18.	Miyoung Uhm et al., 2024	Construction	Accuracy of the answers increased by 80%.
19.	Jungwon Lee et al., 2024	Construction	Accuracy of the answers increased by 21.5%.
20.	Saha, B et al., 2024	Language	Accuracy of the answers increased by 63%
21.	Zhao, Z et al., 2023	Health	Precision value reaches 63%

Table 4 Performance of the retrieval augmented generation method on large language models. Data shows that the most widely used retrieval augmented generation is in the construction and manufacturing sectors. However, the highest accuracy value achieved can reach 99% in the health sector (Alkhalaf et al., 2024). This proves that the use of Retrieval Augmented Generation on large language models is a solution to improve the quality of answers.

RQ3: **What challenges and solutions are found in applying Retrieval Augmented Generation to enhancing Large Language Model knowledge?**

The application of retrieval augmented generation in optimizing large language model knowledge faces several significant challenges. One of the main challenges is the availability of data to be used for enhancing the large language model, which often occurs when using retrieval augmented generation. However, this challenge can be resolved by collecting more data. Additionally, the larger the data used, the larger the database required, thus necessitating optimization techniques such as the use of knowledge graphs.

Table 5. Challenges and Solutions for Implementation and RAG

No	Study	Challenge	Solution
1.	Rongfan Liu et al., 2025	Manually Trained Database Update	Using API for automatic database updates
2.	Jingchi Jiang et al., 2025	Lack of High-Quality Data in Agriculture	Development of High Quality Agricultural Dialogue Dataset Development using Fine Tuning method
3.	Muhammad Arslan et al., 2024	Limited Data Available	Using data from various sources to using knowledge graphs.
4.	K.B. Mustapha et al., 2024	Difficulty in handling unstructured data (sketches, images and others)	Development using Fine Tuning and In Context Learning methods
5.	Libo Qin et al., 2024	Inefficient Tokenization for Languages with Complex Morphology	Use of Knowledge graphs and cross-lingual retrieval.
6.	Lun-Chi Chen et al., 2024	There is a lot of data that has not been srecorded in the source data.	Added data used as a RAG data source.
7.	Jiawei Shao et al., 2024	Limited knowledge and high cost of model training.	Use of Knowledge graphs and training using fine tuning.

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

8.	Akhila Abdulnazar et al., 2024	Limited knowledge such as medical terms and jargon.	Manual error analysis by medical experts
9.	Vallidevi Krishnamurthy et al., 2024	Data limitations, including informal words.	Use of manual validation and contextual learning.
10.	Luis Bernardo Hernandez Salinas et al., 2024	Limitations of data used for training.	Adding training data from external sources. Using Knowledge Graphs.
11.	Menna Fateen et al., 2024	Limitations of data used for training.	Adding training data from external sources.
12.	Biplov Paneru et al., 2024	Data limitations and training image quality.	Addition of training data and further image processing.
13.	Xuan Liu Set al., 2024	Limitations of data used for training.	Use of External Knowledge and Special Corpus, as well as the use of Fine Tuning.
14.	Seongwoo Cho Xuan Liu et al., 2024	Difficulty in handling complex and unstructured data, such as diagrams and terms.	Use of External Knowledge and use of Fine-Tuning.
15	Jean Pierre Nayinzira et al., 2024	Limitations of the data used for training.	Addition of training data.
16	Mohammad Alkhalaf Jean Pierre Nayinzira et al., 2024	Use of tokens during data processing.	Use of chunking.
17	Jos´e Antonio Heredia Alvaro et al., 2024	Limitations of the data used for training.	Use of External Knowledge.
18	Miyoung Uhm et al., 2024	Difficulty in Using Diverse Data as Terms in Construction.	Restructuring data to make it easier to process by the LLM model.
19	Jungwon Lee et al., 2024	Limitations of data used for training.	Addition of data to be trained and use of real-time data.
20	Saha, B et al., 2024	The data volume is too large.	Use of special datasets and fine tuning methods.
21	Zhao, Z et al., 2023	Limitations of structured medical data and the complexity of herbal compatibility rules	Addition of training data and use of fine-tuning.

In Table 5, the most significant challenge when using augmented generation retrieval is the limited data to be drilled, which is the main obstacle. However, this obstacle can be overcome by preparing a larger data set and combining it with Fine Tuning. (Jingchi Jiang et al., 2025). Although some may still provide less accurate responses, overall, the accuracy value remains relatively high, which validates the effectiveness of the augmented generation retrieval method.

DISCUSSIONS

The results of the analysis of 21 selected articles show that the application of Retrieval-Augmented Generation (RAG) in improving knowledge in Large Language Models (LLM) provides significant results. RAG has proven effective in improving the accuracy of answers, especially in areas that require specific knowledge, such as health, manufacturing, and energy. In the health sector, the accuracy reached 99%, while other sectors such as manufacturing and construction also showed a significant increase in accuracy, at 91% and 80% respectively. This study shows that the use of RAG is able to dynamically update the knowledge model, allowing the model to produce more relevant, accurate, and up-to-date answers without the need for intensive retraining.

The Retrieval-Augmented Generation (RAG) method for enhancing knowledge in Large Language Models has proven to be highly effective in improving answer quality by combining the retrieval process to access relevant external information with the generation capability to produce more informative and contextual text. This method offers significant advantages, particularly in tasks that require deep understanding and access to up-to-date knowledge, such as providing more accurate and relevant answers.

Overall, Retrieval-Augmented Generation (RAG) method for augmenting knowledge in Large Language Models (LLMs) has proven to be very effective in improving answer quality by combining external information retrieval and more informative and contextual text generation capabilities. This method offers significant advantages, especially in tasks that require deep understanding and access to current knowledge. Overall, this method shows great potential in updating and improving knowledge on large language models, with RAG

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

excelling in applying external information that provides higher accuracy for certain tasks. Several articles also mention that RAG is more effective when combined with other methods, such as fine-tuning. With this combination, it is expected that knowledge on Large Language Models can be updated, allowing the model to provide more precise answers. Although promising results have been obtained, this trend is still relatively new and has not been widely discussed in the literature, indicating that a deeper understanding of the specific challenges of its application in various domains is still needed. Several similar studies have also highlighted the importance of data quality and relevance in the retrieval process, which can affect the generation results, and although RAG has great potential, its application in specific domains requires further research to optimize its effectiveness and reduce bias.

CONCLUSION

Research shows that Retrieval-Augmented Generation (RAG) is widely and frequently used to improve the performance of Large Language Models (LLM) in certain knowledge domains, especially in healthcare, manufacturing, energy, and construction. The answers generated by the model integrated with RAG also show higher accuracy, with an accuracy rate of up to 99%. However, to achieve such high accuracy, the model still requires additional training, such as prompting and fine-tuning. With the great potential of RAG, its application in various sectors shows significant results in updating and improving model knowledge. Further research needs to be focused on developing methods to reduce bias, improve the quality of retrieval data, and optimize the application of RAG in more specific domains, to support broader and more effective applications. In addition, the development of Agentic RAG as an approach that allows models to act independently by integrating more context and control in the retrieval and generation process also needs to be a major focus of future research.

REFERENCES

Abdulnazar, A., Roller, R., Schulz, S., & Kreuzthaler, M. (2024). Large Language Models for Clinical Text Cleansing Enhance Medical Concept Normalization. *IEEE Access*, *12*(September), 147981–147990. https://doi.org/10.1109/ACCESS.2024.3472500

Alkhalaf, M., Yu, P., Yin, M., & Deng, C. (2024). Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records. *Journal of Biomedical Informatics*, *156*(May), 104662. https://doi.org/10.1016/j.jbi.2024.104662

Arslan, M., Mahdjoubi, L., & Munawar, S. (2024). Driving sustainable energy transitions with a multi-source RAG-LLM system. *Energy and Buildings*, *324*(July), 114827. https://doi.org/10.1016/j.enbuild.2024.114827

Chen, G., Alsharef, A., Ovid, A., Albert, A., & Jaselskis, E. (2025). Meet2Mitigate: An LLM-powered framework for real-time issue identification and mitigation from construction meeting discourse. *Advanced Engineering Informatics*, *64*(December 2024), 103068. https://doi.org/10.1016/j.aei.2024.103068

Chen, L. C., Pardeshi, M. S., Liao, Y. X., & Pai, K. C. (2025). Application of retrieval-augmented generation for interactive industrial knowledge management via a large language model. *Computer Standards and Interfaces*, *94*(December 2024), 103995. https://doi.org/10.1016/j.csi.2025.103995

Cho, S., Park, J., & Um, J. (2024). Development of Fine-Tuned Retrieval Augmented Language Model specialized to manual books on machine tools. *IFAC-PapersOnLine*, *58*(19), 187–192. https://doi.org/10.1016/j.ifacol.2024.09.157

Fateen, M., Wang, B., & Mine, T. (2024). Beyond Scores: A Modular RAG-Based System for Automatic Short Answer Scoring with Feedback. *IEEE Access*, *12*(November), 185371–185385. https://doi.org/10.1109/ACCESS.2024.3508747

Heredia Álvaro, J. A., & Barreda, J. G. (2025). An advanced retrieval-augmented generation system for manufacturing quality control. *Advanced Engineering Informatics*, *64*(May 2024). https://doi.org/10.1016/j.aei.2024.103007

Hernandez-Salinas, B., Terven, J., ChaveZ-Urbiola, E. A., Cordova-Esparza, D. M., Romero-Gonzalez, J. A., Arguelles, A., & Cervantes, I. (2024). IDAS: Intelligent Driving Assistance System using RAG. *IEEE Open Journal of Vehicular Technology*, *5*(August), 1139–1165. https://doi.org/10.1109/OJVT.2024.3447449

Jiang, J., Yan, L., Liu, H., Xia, Z., Wang, H., Yang, Y., & Guan, Y. (2025). Knowledge assimilation: Implementing knowledge-guided agricultural large language model. *Knowledge-Based Systems*, *314*(January). https://doi.org/10.1016/j.knosys.2025.113197

Krishnamurthy, V., & Balaji, V. (2024). Yours Truly: A Credibility Framework for Effortless LLM-powered Fact Checking. *IEEE Access*, *12*(November). https://doi.org/10.1109/ACCESS.2024.3520187

Lee, J., Ahn, S., Kim, D., & Kim, D. (2024). Performance comparison of retrieval-augmented generation and fine-tuned large language models for construction safety management knowledge retrieval. *Automation in Construction*, *168*(PB), 105846. https://doi.org/10.1016/j.autcon.2024.105846

Liu, R., Zou, Z., Chen, S., Liu, Y., & Wan, J. (2025). Harnessing AI for understanding scientific literature:

*name of corresponding author

Innovations and applications of chat-agent system in battery recycling research. *Materials Today Energy*, *49*(January), 101818. https://doi.org/10.1016/j.mtener.2025.101818

Liu, X., Erkoyuncu, J. A., Fuh, J. Y. H., Lu, W. F., & Li, B. (2025). Knowledge extraction for additive manufacturing process via named entity recognition with LLMs. *Robotics and Computer-Integrated Manufacturing*, *93*(November 2024), 102900. https://doi.org/10.1016/j.rcim.2024.102900

Mustapha, K. B. (2025). A survey of emerging applications of large language models for problems in mechanics, product design, and manufacturing. *Advanced Engineering Informatics*, *64*(July 2024), 103066. https://doi.org/10.1016/j.aei.2024.103066

Nayinzira, J. P., & Adda, M. (2024). SentimentCareBot: Retrieval-Augmented Generation Chatbot for Mental Health Support with Sentiment Analysis. *Procedia Computer Science*, *251*, 334–341. https://doi.org/10.1016/j.procs.2024.11.118

Paneru, B., Thapa, B., & Paneru, B. (2024). Leveraging AI in ayurvedic agriculture: A RAG chatbot for comprehensive medicinal plant insights using hybrid deep learning approaches. *Telematics and Informatics Reports*, *16*(August), 100181. https://doi.org/10.1016/j.teler.2024.100181

Qin, L., Chen, Q., Zhou, Y., Chen, Z., Li, Y., Liao, L., Li, M., Che, W., & Yu, P. S. (2025). A survey of multilingual large language models. *Patterns*, *6*(1), 101118. https://doi.org/10.1016/j.patter.2024.101118

Shao, J., Tong, J., Wu, Q., Guo, W., Li, Z., Lin, Z., & Zhang, J. (2024). WirelessLLM: Empowering Large Language Models Towards Wireless Intelligence. *Journal of Communications and Information Networks*, *9*(2), 99–112. https://doi.org/10.23919/JCIN.2024.10582827

Uhm, M., Kim, J., Ahn, S., Jeong, H., & Kim, H. (2025). Effectiveness of retrieval augmented generation-based large language models for generating construction safety information. *Automation in Construction*, *170*(November 2024), 105926. https://doi.org/10.1016/j.autcon.2024.105926

*name of corresponding author