ESP-JETA
ESP Journal of Engineering & Technology Advancements
ISSN: 2583-2646 / Volume 6 Issue 2 April to June 2026
Paper Id: JETA-V6I2P108 / Page No: 67-75

Original Article

Advances and Challenges in Retrieval-Augmented Generation Models for Knowledge-Driven NLP Tasks

Sumeet Mathur
University of Waikato NZ Joint Institute at Zhejiang University, Hangzhou, China.

Received Date: 14 March 2026 Revised Date: 23 March 2026 Accepted Date: 12 April 2026

Abstract *- The advancement in artificial intelligence continues to transform knowledge-based applications, and Retrieval-Augmented Generation (RAG) has become a prominent framework for this activity. By adding wide-scale information retrieval, RAG enhances huge language models, allowing them to produce responses based on applicable and current knowledge instead of simply using the pre-trained memory. Originally intended as open-domain question answering, RAG has since been applied in general domains such as healthcare, legal reasoning, education, enterprise analytics and conversational AI. RAG enhances the accuracy of facts, relevant to the context and intelligibility by retrieving domain-specific evidence and combining it with generative reasoning. This review describes the basic structure of RAG with the focus on retrieval pipelines, embedding-based indexing, reranking strategies, and knowledge fusion processes. It also shows how techniques like hybrid dense-sparse retrieval, graph-based knowledge modeling and adaptive query reformulation reinforce retrieval accuracy and reasoning power. An overall overview of the latest literature shows the increase in the sophistication and variety of RAG implementations, including modular architectures and more refined domain models, as well as scalable enterprise-ready systems. The empirical evidence continuously demonstrates that RAG performs better than standalone language models in tasks that involve grounded reasoning, contextual fidelity, and evidence-based response, providing it with a key paradigm of the next generation of intelligent systems.*

Keywords *- Retrieval-Augmented Generation (RAG), Large Language Models (LLMs), Semantic Search, Hybrid Retrieval, Text Generation, Knowledge-Driven NLP, Question Answering, RAFT, Dense and Sparse Indexing, Natural Language Processing.*

I. INTRODUCTION

The recent studies have aimed at implementing the retrieval augmented generation (RAG) technology in the field especially in the case of traffic law documents to effectively find complex laws and regulations relating to traffic [1]. RAG is a system of higher order that retrieves the information and generates the text simultaneously, which allows processing the rules pertaining to road traffic in the form of specific language and various regulations in a timely and accurate manner [2][3]. Respecting a jurisdiction's traffic-related rules and regulations is crucial for unmanned systems, particularly autonomous driving.

Retrieval-Augmented Generation (RAG) models use a retrieval technique during generation to get around this flaw. Information pertinent to the external sources can be accessed by the model thanks to the retrieval method, such as a database or the internet, when it is generating a response. In this way, RAG models can improve the text generated by them by adding additional accuracy, relevance, and up-to-date information, which greatly increases the usefulness and applicability of the LLMs in practice [4][5][6]. This survey discusses the combination of LLMs and the retrieval systems, main methodologies, applications, advantages, and difficulties of RAG models.

Noteworthy techniques for domain-specific LM adaptation are Retrieval-Augmented Generation (RAG) and Fine-Tuning (FT). RAG uses in-context learning (ICL) to improve LM's response production by extracting pertinent information from a corpus of documents. The FT technique, on the other hand, modifies model weights to improve its memorizing capacities during inference and becomes skilled at recalling certain facts [7][8][9]. As a first step towards FT, data augmentation techniques are applied to create synthetic training data in the context of less well-known information, when there is a dearth of data. Although there has been research on using RAG to improve LM's memorization, RAG and knowledge gained via FT have not been compared in any research, particularly for less well-known material.

As a component of the application or implementation of artificial intelligence (AI), natural language processing (NLP) converts human language into output that can be understood by computers or by humans [10]. NLP is the capacity of a machine to respond in a human-understandable language. Algorithms that translate text into words can be created. According

to their definitions, the terms can be categorized [11]. The majority of NLP methods rely on machine learning to extract meaning from human languages. NLP is divided into two fields: computer science and linguistics [12]. The study of language's structure, syntax, meaning, and different phrase forms is known as linguistics. The study of linguistics or natural language processing, which encompasses related fields like machine learning, deep learning, and artificial intelligence, is one of the quickest and most expansive new technologies in computer science [13].

In Retrieval-Augmented Generation (RAG) models have become a ground-breaking breakthrough in natural language processing, resolving the basic problems with conventional Large Language Models (LLMs) [14]. RAG systems are a major breakthrough in AI and NLP because of the ability to use the strengths of the LLM to access external knowledge dynamically. The new method can be used to create more accurate, relevant, and up-to-date responses in a large variety of applications. The importance of RAG research is that it is capable of altering AI applications on a fundamental basis. These systems make text more accurate and relevant, increase the ability of AI to work on complex and knowledge-intensive tasks, and allow updating the knowledge continuously without necessarily retraining the model[15].

A. Structure of the paper

This paper is structured as follows: Section II describes the RAG framework and its basic components. Section III discusses new developments in RAG models. Section IV discusses some of the important NLP applications. Section V underlines major issues in RAG models. In section VI, the literature review is given. Future directions are put at the end of section VII.

II. OVERVIEW OF RETRIEVAL-AUGMENTED GENERATION

A classic usage of RAG is described in Figure 1, below, where a user asks a question regarding recent, popular news. Since is based on pretraining data, it does not have the ability to give updates on the latest events at first. RAG accesses and adds knowledge from external databases to bridge this information gap. Here, it gathers news data relevant to the user's inquiry. Together with the starting query, these articles create one large prompt that allows LLMs to produce an informed response [16]. The RAG research paradigm, which can be divided into three tiers, is likewise constantly evolving. Naive RAG, Modular RAG, and Advanced RAG. Although the RAG method is affordable and performs better than the original LLM, it has a number of limitations as well. These specific shortcomings in Naive RAG prompted the development of Modular RAG and Advanced RAG.

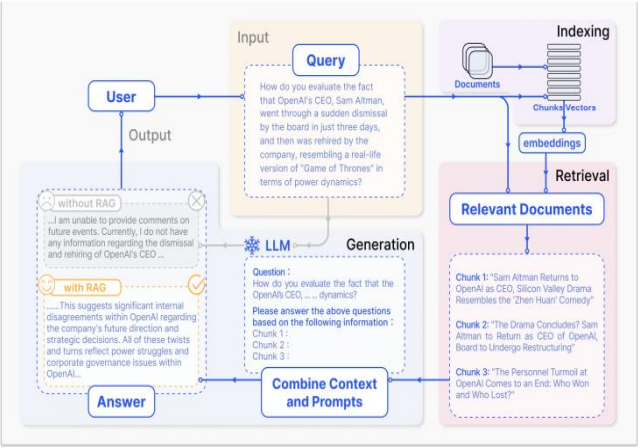

Figure 1: Workflow of Retrieval-Augmented Generation (RAG)

A. Naive RAG

Naive RAG is one of the earliest methods for using pertinent information obtained from outside knowledge sources to root LLMs and produce text [17]. It is the simplest form of RAG and does not have the sophistication needed to handle sophisticated queries. Indexing, retrieval, and generating are the three processes that make up the Naive RAG process. The following section discusses Naive RAG implementations in the medical field. Customized liver disease and condition guidelines and advisory materials using text embeddings. The text-embedding-ada-002 model was then used to instantly translate user queries into embeddings. They employed GPT-3.5-turbo or GPT-4-32k to reply after locating matches for the embeddings in a vector database. Their findings demonstrate that, while utilizing GPT-3.5, they were able to produce more detailed responses than regular.

A. Definition and Core Components

An AI system called Retrieval-Augmented Generation (RAG) enhances content creation with outside information sources by fusing generative AI models with retrieval-based search. RAG-based AI systems query external knowledge repositories to get pertinent information prior to producing replies, rather than depending exclusively on parametric memory (model weights) [18].

The core components of RAG include

a) Retriever

Searches for relevant external knowledge based on the user query.

- Sparse Retrieval (BM25, TF-IDF): Matches queries using keyword-based search
- Dense Retrieval (DPR, Colbert, ANCE): Uses neural embeddings for semantic similarity retrieval
- Hybrid Retrieval: combines sparse and dense approaches to provide the best outcomes.

b) Generator

It uses retrieved knowledge to produce a logical response.

- Uses transformer-based architectures.
- Ensures responses are grounded in the retrieved evidence.

c) Indexing Mechanisms

- Dense embeddings are stored in vector databases for effective searching.
- Knowledge Graphs (KGs) organize retrieval enhancements according to domains[19][20].

B. RAG Architecture

The RAG is a sophisticated AI system that overcomes the shortcomings mentioned above by integrating the merits of information retrieval and text generation. RAG as a hybrid strategy is developed in two main components [21]:

a) Retrieval Component

The retrieval system has the duty of searching external knowledge bases, such as indexed documents, structured databases, or live web queries. The most relevant content is identified using techniques such as dense passage retrieval (DPR), BM25, or transformer-based embeddings.

b) Generative Component

The relevant context is then given as input into a generative language model which generates responses by taking into account both the query and the retrieved information. This makes sure that the text generated is supported with factual information and is user-friendly.

III. ADVANCES IN RAG MODELS

The last developments of Retrieval-Augmented Generation (RAG) models were mostly devoted to the improvement of the retrieval component, which is the most important factor in improving the overall accuracy of knowledge-intensive tasks [22]. Among the contributions in this respect, there is the introduction of the Blended RAG, which suggests a more advanced methodology in search terms, a method of combining hybrid query-based retrievers with semantic search. In the research, the authors stress that the use of the traditional RAG systems usually presupposes the application of the keyword-based or single methods of retrieval, which restricts the capability of the generator to retrieve the corresponding context, especially in the case of large-scale or domain-specific corpora. In response to this, Blended RAG is a mixture of several retrieval indices BM25, dense vector search (KNN), and Sparse Encoder-based search and hybrid query methods, including Cross Fields and Best Fields[23]. This combined approach yields higher accuracy in retrieval when compared to a variety of benchmark systems, such as TREC-COVID, Natural Questions (NQ), and Squad, and is also higher when compared to individually fine-tuned models in zero-shot scenarios. Particularly, the framework shows that the use of the index of Sparse Encoder queries using Best Fields provides the best performance in terms of top-k retrieval measurements and the quality of RAG output and does not need data-specific fine-tuning. This discovery demonstrates that RAG development now focuses on retraining large language models not only to make them more relevant and useful across scenarios but also to make the retriever architecture more efficient and effective.

A. Blended Retrieval Strategies in RAG

In the case of RAG systems, three different search strategies were considered: similarity search by using keywords, dense vector-based search and semantic-based sparse encoders, and combined them to create hybrid queries [24]. As opposed to traditional keyword matching, semantic search looks deeper into the depths of a query typed by a user and reads between the lines. This paper is a systematic comparison of a collection of search methods using three main indices: BM25 to use key-

based, KNN to use the vector-based, and Elastic Learned Sparse Encoder (ELSER) to use the sparse encoder-based semantic search.

- **BM25 Index:** The BM25 index is adept at using fuzzy matching technologies to use full-text search capabilities, which paves the way for more complex query operations.
- **Dense Vector Index:** It builds a sparse index of vectors with the strength of sentence transformers. It finds the closeness of the vectors' representations of the document and query contents.
- **Sparse Encoder Index:** In order to extract the nuanced links between words, the Sparse Encode Retriever Model index combines similarity-based retrieval with semantic understanding. This results in a more accurate representation of user intent and document relevance.

A. Retrieval Augmented Generation

RAG is the process of improving the model's performance by supplementing the input prompt with data from a collection of publications (the Knowledge Base). The process is simple: first, use any embedding transformer to embed a query (the question that has to be answered) as a semantic vector [25]. The knowledge base is embedded, pre-segmented, and kept in a vector database for quick similarity searches. The top-n documents that were provided as context in the LLM's prompt had the query embedding's highest semantic similarity (lowest cosine distance).

However, there are a few caveats

- In order for semantic vectors to be a particularly useful tool for information retrieval, the topic or subject must remain constant throughout the data. This is due to the fact that a fixed-size vector's representation capacity is finite, and adding more than one subject to it would lead to representation conflicts.
- Furthermore, determining semantic vectors is a language-dependent problem, and the majority of unconventional models performed poorly on Ukrainian in early tests.
- A significant factor in determining how much material to obtain from the KB is the LLM's context window. Compared to commercial-grade LLMs, open-source models often have narrower context windows, necessitating a decrease in the quantity of input data collected from the RAG phase.

B. Retrieval-Augmented Fine Tuning (RAFT)

To enhance LLM performance in domain-specific Retrieval-Augmented Generation (RAG) scenarios, a unique training method known as Retrieval-Augmented Fine-Tuning (RAFT) was developed. By successfully recreating an open-book exam environment where models must use retrieved materials to create valid answers, the RAFT technique tackles major obstacles in training LLMs to operate with retrieval mechanisms [26]. Training the model to distinguish between significant and irrelevant documents—known as "oracle" and "distractor" materials, respectively—is the goal of the RAFT technique. Through these variations, RAFT improves the LLM's ability to provide proper answers to questions regardless of whether the learner's context is incorrect or unrelated. This is similar to studying for an open-book test by identifying and applying useful information while ignoring unimportant material.

IV. APPLICATIONS IN KNOWLEDGE-DRIVEN NLP

NLP knowledge-based inference systems provide partial solutions to several effective applications in text processing. Database queries, information retrieval and question-answering, and conceptually-based document retrieval, information extraction, knowledge acquisition of natural-language texts, automatic summarization, knowledge-based machine translation, understanding of literal expressions, such as metonymy, metaphor and idioms, document classification, intelligent authoring systems, and intelligent agents that communicate using natural languages are some of the major uses [27][28]. The KB-NLP systems of most of these applications are yet to be developed to some extent.

A. Open-Domain Question Answering

The reader and the retriever are the two components that make up the ODQA system. The objective of the passage retriever is to search a vast collection of passages for information based on a query. The retrieved information is then as the input of the reader module which produces answers to the questions. The retriever, in contrast to the reader, is not as concerned about accuracy since it utilizes a much larger amount of data because its goal is efficiency. Hence, the retriever's architecture design is often simpler. Previous literature used non-parametric sparse techniques, including BM25 and TF-IDF, as the retriever. These methods are inadequate for interpreting real language and instead focus on matching important words. Due to the development of DL, the recent literature has embraced techniques based on deep neural networks to improve retrieval performance, where PLMs like BERT and T5 were utilized to initialize the retriever-initiated question and passage encoders with BERT and the similarity of question passages is determined by taking the dot product of the output from two encoders [29].

B. Text Classification

Text classification, which involves classifying a text based on its content into predefined groups, is one of the fundamental tasks in natural language processing (NLP) [30]. Topics, feelings, and other characteristics might be included in the categories. To address this problem, a number of machine learning techniques have been created, such as Naive Bayes, Support Vector Machines, and Neural Networks. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are two DL approaches that have significantly improved text categorization problems in recent years. Based on its content, this algorithm divides a document into one or more pre-established groups or subjects. Text classification has several uses in a variety of fields, such as news categorization, spam detection, and sentiment analysis. Among the ML techniques that have been suggested to solve this issue are neural networks, support vector machines, and naive Bayes.

C. Dialogue Systems and Conversational Agents

The dialogue systems that emerged throughout the 1960s and 1970s relied on text. Some prominent examples include BASE-BALL, SHRDLU, and GUS. BASEBALL was a method for answering questions about baseball games. The system simply rejected queries that it could not answer and was capable of answering queries with a limited syntactic framework. SHRDLU was more advanced linguistically, with a vast English grammar, semantic understanding of the objects in its domain (a world of blocks), and a pragmatic component that processed input from non-linguistic domains.

A computer-generated animated figure known as an Embodied Conversational Agent (ECA) uses voice, hand gestures, body posture, and facial expression to create a more engaging and human-like interaction. Screen-based characters and virtual agents are examples of ECAs. Examples include [31]:

- Smarta Kus, an animated figure that presents information as part of the Smart Kom project.
- REA, a multi-modal, life-sized ECA that functions as a real estate agent in real time.
- GRET, a three-dimensional ECA that can converse and show movement of the head, gaze, and facial emotions in real time.

V. KEY CHALLENGES IN RAG MODELS

Dynamic knowledge management presents complex challenges in keeping the knowledge base up-to-date while maintaining system performance. This is particularly critical in domains with rapidly evolving information, such as news, scientific research, or social media trends.

A. Scalability

The fundamental challenge of scalability is the curse of dimensionality. The search space of nearest neighbor search increases exponentially as the amount of data increases[32], and thus it is computationally intractable to carry out exact nearest neighbor search in high-dimensional space [33][34]. It is especially an issue with dense representations of vectors in modern retrieval systems. High-dimensional spaces are characterized by the fact that the notion of the nearest neighbor is less significant because of the so-called distance concentration. As the dimensionality grows, the distances between the closest and the farthest neighbors of a point become equalized to 1, hence, it becomes hard to tell the difference between points that are close and those that are far away. The phenomenon has a lot of influence on the efficiency of the classic similarity search algorithms.

B. Query Reformulation

In RAG systems, query reformulation is an extremely difficult issue due to the semantic discrepancy between user inquiries and knowledge base material. The process is a complex task comprising of natural language understanding and generation and thus highly advanced models are needed to capture subtle semantic relationships[35][36]. One of the issues is the management of prejudiced or biased queries and being objective. As an example, a search query such as Why are vaccines harmful? has a partial assumption that the system has to acknowledge and counter-check to bring out balanced information retrieval. A question that is yet to be answered is how to find a way to identify and control such biases and still consider user intent. The other issue that is very important is the adaptation of queries to particular areas or times. The queries made by the user usually include jargon within the domain, or they refer to the events happening at the time that only specific knowledge would help interpret them appropriately. Adopting both specific domain knowledge and generic language insight is a complicated problem that contemporary systems cannot always cope with successfully.

C. Latency

The core challenge of latency in RAG systems is the inherent trade-off between reaction time and the quality of results. The system should also exercise caution in the interactive applications, in terms of depth of retrieval versus the patience limit of the user. This balancing manipulation is especially important since the depth of retrieval has a great influence on the latency and the quality of the results; the deeper the retrieval, the more detailed it may be, but at the cost of more time to respond [37]. Multi-step retrieval techniques are sometimes necessary for processing complicated queries or multi-hop reasoning, although their intricacy is one of the main sources of delay. The more retrieval stages there are, the more difficult it

is to control cumulative delay. Every extra step not only lengthens the response time overall but also creates possible inconsistencies or areas of failure in the retrieval process.

VI. LITERATURE REVIEW

The literature overview of Advances and Difficulties in the next section presents Models of Knowledge-Driven Natural Language Processing with Retrieval-Augmented Generation, and Table I offers a succinct summary.

Al-Qatf et al. (2025) If the RAG framework is seen to be an efficient method of improving LLMs by offering a suitable retrieval method to obtain relevant external information, although it still has drawbacks when it comes to obtaining high-quality knowledge from a variety of data sources. To take advantage of RAG's capabilities inside data spaces, a complementary integration of RAG and data spaces is suggested. Through direct data exchange arrangements and safe data-sharing procedures, RAG may obtain a range of high-quality data sources from several data producers through data spaces. Additionally, RAG improves data space support services. A high-level architecture is used to deliver RAG data space models (RAG-DSMs) with a single lifetime for both RAG and data spaces. This study highlights potential prospects as well as potential obstacles of the suggested connection. Additionally, examples of using RAG-DSMs in the health and mobility domains [38]

Meng et al. (2025) researched the fine-tuning technique, and RAG, the general framework design for the text generation system of fusion strategy, is put forward, and several typical systems in specific fields are analyzed. The realization and development direction of technology provides a reference for studying fusion strategy, which can improve the quality of text generation system design [39].

Gu (2025) constitutes a significant development in the area of NLP, cleverly combining dynamic external information retrieval with LLMs. The factual accuracy issues with conventional generative models are resolved by this combination and knowledge update velocity, while also improving the models' reactivity to real-world information. The problems with RAG systems and their fixes are covered in this review. It examines the fundamental architecture of RAG systems, which consists of knowledge bases, generating components, and retrieval components, paying particular focus to the most recent advancements that expanded the range of usefulness and performance [14].

Joseph et al. (2024) propose RAG models by combining an advanced generative transformer architecture with a complex information retrieval mechanism, thus improving the reliability of the information obtained and the coherence of the responses produced. The main aspect of the HRATN model is that there is a dense-based retrieval system that can skilfully tap the rich knowledge base into the collection of relevant information, and an optimization fusion response generation is performed by a transformer network with sophisticated attention mechanisms. The striking innovation within the HRATN is its hybrid fusion layer, which is effective in the combination of both retrieved information and the outputs of the generative model to provide more accurate responses that are also more pertinent to the situation. To assess HRATN's efficacy, comprehensive experiments were conducted, comparing its performance with the best RAG models on the ratings of correctness of response, relevance and computational efficiency [40].

Ali et al. (2024) RAG with a Large Language Model (LLM) and many Natural Language Processing (NLP) models. The growing quantity of research publications is a significant obstacle for manual literature reviews. As a result, there is now more need for automation. As a result, there is now more need for automation. A number of NLP techniques, including the transformer model (Simple T5), the frequency-based method (spaCy), and the ability of RAG with Large Language Model (GPT-3.5-turbo) to achieve the main goal is assessed. The ROUGE ratings are used to assess each of the three approaches. The Large Language Model GPT-3.5-turbo has the greatest ROUGE-1 score, 0.364, according to the examination [41].

Destan et al. (2024) provide the capability to integrate LLMs with techniques for information retrieval to provide more comprehensive and contextually relevant responses to user inquiries. The language model may immediately access external information sources thanks to its design; thus, it generates more accurate and contextual responses armed with existing information. These features of RAG provide appropriate solutions to users' information-based demands by better understanding the complexity of natural language. In this study, it is emphasized that the integration of RAG architecture with information retrieval systems and LLMs provides more sensitive and accurate solutions in information-intensive tasks. This work highlights how applications in the field of NLP are strengthened by the RAG architecture's capacity to extract information by dynamically utilizing the lessons learned from vast datasets of LLMs [42].

The research on knowledge-driven natural language processing tasks using retrieval-augmented generation models: Advances and Challenges is summarized in Table I, along with the methodology, important discoveries, difficulties, and future prospects

Table 1. The Studies no Retrieval-Augmented Generation (Rag) Systems in the Recent Past

Author	Study On	Approach	Key Findings	Challenges	Future Directions
Al-Qatf et al., (2025)	Integration of RAG with data spaces	RAG-Data Space Model (RAG-DSM)	RAG can leverage secure, high-quality, diverse data sources through data spaces	Acquiring reliable knowledge from varied sources; integration complexity	Unified RAG-data space lifecycle; applications in mobility and health
Meng et al., (2025)	Fusion strategy in RAG-based generation systems	Fine-tuning and fusion design framework	Framework improves design quality of text generation through effective strategy fusion	Limited applicability in diverse NLP fields	Guide for designing fusion-based RAG systems
Gu, (2025)	Core architecture and innovation in RAG systems	Component-level analysis of RAG	RAG enhances factual accuracy and responsiveness by integrating LLMs with dynamic retrieval	Knowledge update velocity, retrieval quality	Expand capabilities of retrieval and generation components
Joseph et al., (2024)	High-Performance RAG model (HRATN)	Dense vector retrieval + Transformer + Hybrid fusion layer	HRATN boosts precision and contextual relevance by combining retrieved data with generative outputs	Computational cost; optimizing hybrid fusion	Improved attention mechanisms and retrieval techniques
Ali et al., (2024)	Comparison of NLP techniques for literature review automation	spaCy, Simple T5, RAG with GPT-3.5-turbo	RAG (GPT-3.5) outperformed others with highest ROUGE-1 score of 0.364	Volume and complexity of scientific literature	Enhance automation tools using RAG-LLMs
Destan et al., (2024)	Enhancing contextual accuracy through RAG in NLP	RAG integration with LLMs and IR systems	RAG improves contextual understanding and accuracy in complex natural language tasks	Real-time access and integration with diverse data	Wider NLP task adaptation and use in information-intensive applications

VII. CONCLUSION AND FUTURE WORK

RA generation has quickly evolved out of a research concept into a substantive paradigm on par with the way intelligent systems gain access to, synthesize and reason about knowledge. It has the merits of an intermediate between retrieval accuracy and generative flexibility, which allows results that are contextually appropriate and factual. Based on this review, it is established that RAG makes an immensely larger difference in terms of question answering and value extending into realms like healthcare, education, dialogue systems and enterprise analytics. But the quest is not over yet. Prolonged difficulties, such as noisy retrieval, weak integration of reasoning, poor domain generalization and computational complexity, have continued to filter high-stakes performance. Going forward, future studies are supposed to emphasize short-term enhancements, i.e., stronger retrieval pipelines, dynamic reranking, adaptive prompting, and scalable data management. Mid-term research has the ability to pursue more reasoning concordance by using graph representations, planning-informed architectures and multi-hop knowledge chaining. The long-term directions are towards autonomous systems that can be capable of self-retrieval, ongoing learning and interpretable reasoning in changing knowledge space. Finally, the interdisciplinary approach that connects information retrieval, knowledge graphs, cognitive reasoning and efficient model

design will be needed to advance RAG. When such attempts come into reality, RAG will end up being a fundamental facilitator of reliable, domain-sensitive and decision-support-ready AI systems and run with reliability in complicated real-world situations.

VIII. REFERENCES

[1] S. B. Karri, S. Gawali, S. Rayankula, and P. Vankadara, "AI Chatbots in Banking: Transforming Customer Service and Operational Efficiency," 2025. doi: 10.3233/FAIA251498.

[2] Y. Choi, S. Kim, Y. C. F. Bassole, and Y. Sung, "Enhanced Retrieval-Augmented Generation Using Low-Rank Adaptation," *Appl. Sci.*, vol. 15, no. 8, pp. 1–19, 2025, doi: 10.3390/app15084425.

[3] U. Dodda, H. Volikatla, and J. R. Vummadi, "Exploring the Role of AI-Enhanced Chatbots in Automating Recruitment Processes in Human Capital Management Systems," *Int. J. Emerg. Trends Comput. Sci. Inf. Technol.*, vol. 6, no. 3, July, pp. 28–36, 2025, doi: https://doi.org/10.63282/3050-9246.IJETCSIT-V6I3P104.

[4] P. Nutalapati, J. R. Vummadi, S. Dodda, and N. Kamuni, "Advancing Network Intrusion Detection: A Comparative Study of Clustering and Classification on NSL-KDD Data," in *2025 International Conference on Data Science and Its Applications (ICoDSA)*, Jakarta, Indonesia: IEEE, 2025, pp. 880–885, July. doi: 10.1109/ICoDSA67155.2025.11157595.

[5] N. Shaik, B. Harichandana, and P. Chitralingappa, "RAG Models : Integrating Retrieval for Enhanced Natural Language Generation," *Int. J. Res. Eng. Sci.*, vol. 12, no. 6, pp. 129–138, 2024.

[6] P. R. Marapatla, "NEXT-GEN ENTERPRISE BI: A STRATEGIC GUIDE TO AI-INFUSED REPORTING SOLUTIONS," *TPM – Testing, Psychom. Methodol. Appl. Psychol.*, vol. 32, 2025.

[7] H. Soudani, E. Kanoulas, and F. Hasibi, "Fine Tuning vs. Retrieval Augmented Generation for Less Popular Knowledge," pp. 12–22, 2024, doi: 10.1145/3673791.3698415.

[8] S. Amrale, "A Novel Generative AI-Based Approach for Robust Anomaly Identification in HighDimensional Dataset," *Int. J. Adv. Res. Sci. Commun. Technol.*, vol. 4, no. 2, pp. 709–721, 2024, doi: 10.48175/IJARSCT-19900D.

[9] F. J. C. Faust *et al.*, "Embedding-based retrieval techniques for feeds," 11960550, 2024

[10] K. M. R. Seetharaman and S. Pandya, "IMPORTANCE OF ARTIFICIAL INTELLIGENCE IN TRANSFORMING SALES, PROCUREMENT, AND SUPPLY CHAIN PROCESSES," *Int. J. Recent Technol. Sci. Manag.*, vol. 8, no. 7, July, pp. 140–148, 2023, [Online]. Available: https://ijrtsm.com/wp-content/uploads/2025/05/July-2023-Karthika-140-148.pdf

[11] B. Priya, N. J.M, and G. Thangavel, "An Analysis of the Applications of Natural Language Processing in Various Sectors," in *Advances in Parallel Computing*, vol. 38, 2021, pp. 598–602. doi: 10.3233/APC210109.

[12] D. Patel, "AI-Enhanced Natural Language Processing for Improving Web Page Classification Accuracy," *ESP J. Eng. Technol. Adv.*, vol. 4, no. 1, pp. 133–140, 2024, doi: 10.56472/25832646/JETA-V4I1P119.

[13] S. Garg, "Predictive Analytics and Auto Remediation using Artificial Intelligence and Machine Learning in Cloud Computing Operations," *Int. J. Innov. Res. Eng. Multidiscip. Phys. Sci.*, vol. 7, no. 2, March-April, pp. 01–05, 2019, doi: http://dx.doi.org/10.5281/zenodo.15362327.

[14] J. Gu, "A Research of Challenges and Solutions in Retrieval Augmented Generation (RAG) Systems," *Highlights Sci. Eng. Technol.*, vol. 124, pp. 132–138, 2025, doi: 10.54097/364hex16.

[15] S. K. Chintagunta and S. Amrale, "AI in Code , Testing , and Deployment : A Survey on Productivity Enhancement in Modern Software Engineering," *Int. J. Curr. Eng. Technol.*, vol. 13, no. 6, December, pp. 627–634, 2023, doi: https://doi.org/10.14741/ijcet/v.13.6.16.

[16] Y. Gao *et al.*, "Retrieval-Augmented Generation for Large Language Models: A Survey," pp. 1–21, 2023.

[17] L. M. Amugongo, P. Mascheroni, S. Brooks, S. Doering, and J. Seidel, "Retrieval augmented generation for large language models in healthcare: A systematic review," *PLOS Digit. Heal.*, vol. 4, no. 6, Jun. 2025, doi: 10.1371/journal.pdig.0000877.

[18] A. Ramachandran, "Advancing Retrieval-Augmented Generation RAG Innovations, Challenges and the Future of AI Reasoning," 2025.

[19] A. Dudhipala, R. Karne, and P. K. Pativada, "Prompt2Graph: Leveraging LLMs to Construct Knowledge Graphs from Technical Manuals," in *2025 4th International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, Tirupur, India: IEEE, 2025, pp. 912–919, September. doi: 10.1109/ICIMIA67127.2025.11200177.

[20] A. R. Toorpu, S. K. Vududala, A. Nerella, and B. P. Madupati, "Hybrid AI Models for Privacy-Preserving Big Data Analytics in Distributed Environments," in *2025 Global Conference in Emerging Technology (GINOTECH)*, PUNE, India: IEEE, 2025, pp. 1–8, July. doi: 10.1109/GINOTECH63460.2025.11076666.

[21] D. C. Youvan, "Retrieval-Augmented Generation (RAG): Advancing AI with Dynamic Knowledge Integration," no. January, 2025, doi: 10.13140/RG.2.2.30888.89606.

[22] A. Mishra, "Retrieval Augmented Generation (RAG) Model," pp. 115–120, 2025, doi: 10.63169/GCARED2025.p16.

[23] S. S. Saisuman Singamsetty, "Hy-Search: A Hybrid Retrieval-Augmented Framework for Factual and Context-Aware Enterprise Knowledge Discovery," in *Proceedings of the 1st Engineering Data Analytics and Management Conference (EAMCON 2025)*, Springer Nature, 2025, pp. 431, Dec. doi: https://doi.org/10.2991/978-94-6463-978-0_37.

[24] K. Sawarkar, A. Mangal, and S. R. Solanki, "Blended RAG: Improving RAG (Retriever-Augmented Generation) Accuracy with Semantic Search and Hybrid Query-Based Retrievers," in *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, IEEE, Aug. 2024, pp. 155–161. doi: 10.1109/MIPR62202.2024.00031.

[25] T. Boros, R. Chivereanu, S. D. Dumitrescu, and O. Purcaru, "Fine-Tuning and Retrieval Augmented Generation for Question Answering Using Affordable Large Language Models," *3rd Ukr. Nat. Lang. Process. Work. UNLP 2024 Lr. 2024 - Work. Proc.*, no. May, pp. 75–82, 2024.

[26] S. A. Akheel, "Fine-Tuning Pre-Trained Language Models for Improved Retrieval in RAG Systems for Domain-Specific Use," *Int. J. Multidiscip. Res.*, vol. 6, no. 5, pp. 1–10, 2024.

[27] K. Mahesh and S. Nirenburg, "Knowledge-based systems for natural language processing Knowledge-Based Systems for Natural Language Processing Kavi Mahesh MCCS-96-296," no. February, 1997.

[28] R. Patel, "Artificial Intelligence-Powered Optimization of Industrial IoT Networks Using Python-Based Machine Learning," *ESP J. Eng. Technol. Adv.*, vol. 3, no. 4, pp. 138–148, 2023, doi: 10.56472/25832646/JETA-V3I8P116.

[29] Q. Zhang, M. Zheng, S. Chen, H. Liu, and M. Fang, "Self Data Augmentation for Open Domain Question Answering," *ACM Trans. Inf. Syst.*, vol. 43, no. 2, pp. 1–35, Mar. 2025, doi: 10.1145/3707449.

[30] A. A. Dande and M. A. Pund, "A Review Study on Applications of Natural Language Processing," *Int. J. Sci. Res. Sci. Eng. Technol.*, vol. 4099, pp. 122–126, Mar. 2023, doi: 10.32628/IJSRSET2310214.

[31] M. Mctear, "Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots," *Synth. Lect. Hum. Lang. Technol.*, vol. 13, pp. 1–251, 2020, doi: 10.2200/S01060ED1V01Y202010HLT048.

[32] D. Bhattacharjee, "Design and Evaluation of Deep Generative AI Model for Intrusion Detection in Cyber Threat Monitoring," in *2025 7th International Symposium on Advanced Electrical and Communication Technologies (ISAECT)*, Mohali, Punjab, India: IEEE, 2025, pp. 1–6, December. doi: https://doi.org/10.1109/ISAECT68904.2025.11318752.

[33] D. Peng, Z. Gui, and H. Wu, "Interpreting the Curse of Dimensionality from Distance Concentration and Manifold Effect," 2024. doi: 10.48550/arXiv.2401.00422.

[34] Y. Macha and S. K. Pulichikkunnu, "A Survey of DevOps Practices for Machine Learning and Artificial Intelligence Workflows in Modern Software Development," *ESP J. Eng. Technol. Adv.*, vol. 4, no. 3, pp. 200–208, 2024, doi: 10.56472/25832646/JETA-V4I3P121.

[35] S. Singamsetty, "AI-Enabled Data Stewardship Real-Time Alignment of Privacy and Storage Policies Across Global Systems using Deep CNN-RNN Techniques," in *2025 5th Asian Conference on Innovation in Technology (ASIANCON)*, PIMPRI, India: IEEE, 2025, pp. 1–6, August. doi: 10.1109/ASIANCON66527.2025.11281010.

[36] S. Hassantabar, Z. Wang, and N. K. Jha, "SCANN: Synthesis of Compact and Accurate Neural Networks," *IEEE Trans. Comput. Des. Integr. Circuits Syst.*, vol. 41, no. 9, pp. 3012–3025, Sep. 2022, doi: 10.1109/TCAD.2021.3116470.

[37] G. Maddali, "Enhancing Database Architectures with Artificial Intelligence (AI)," *SSRN Electron. J.*, 2025, doi: 10.2139/ssrn.5276667.

[38] M. Al-Qatf *et al.*, "RAG4DS: Retrieval-Augmented Generation for Data Spaces—A Unified Lifecycle, Challenges, and Opportunities," *IEEE Access*, vol. 13, pp. 39510–39522, 2025, doi: 10.1109/ACCESS.2025.3545387.

[39] Q. Meng, Z. Wu, Z. Zhao, and X. Lian, "Analysis of Text Generation System Design Combining Retrieval Augmented Generation and Fine-Tuning Strategy," in *2025 2nd International Conference on Smart Grid and Artificial Intelligence (SGAI)*, 2025, pp. 204–208. doi: 10.1109/SGAI64825.2025.11009349.

[40] V. S, Seshasai, and L. Joseph, "A Hybrid Deep Learning Algorithm for Improved ChatBot Accuracy and Relevance Through Advanced Retrieval-Augmented Generation," in *2024 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, 2024, pp. 1–7. doi: 10.1109/ICSES63760.2024.10910846.

[41] N. F. Ali, M. M. Mohtasim, S. Mosharrof, and T. G. Krishna, "Automated Literature Review Using NLP Techniques and LLM-Based Retrieval-Augmented Generation," in *2024 International Conference on Innovations in Science, Engineering and Technology (ICISET)*, 2024, pp. 1–6. doi: 10.1109/ICISET62123.2024.10939517.

[42] B. Tural, Z. Örpek, and Z. Destan, "Retrieval-Augmented Generation (RAG) and LLM Integration," in *2024 8th International Symposium on Innovative Approaches in Smart Technologies (ISAS)*, 2024, pp. 1–5. doi: 10.1109/ISAS64331.2024.10845308.