Retrieval-Augmented Generation: Advances, Applications, and Future Directions in Knowledge-Grounded Language Modeling

Hancheng Yu

Zhejiang University Edinburgh United College, Zhejiang University, Haining, China
Hancheng.22@intl.zju.edu.cn

Abstract. Retrieval-Augmented Generation (RAG) has emerged as a pivotal innovation in natural language processing (NLP), integrating information retrieval with generative modeling to overcome the limitations of static, parameter-based large language models. By dynamically retrieving relevant external knowledge during text generation, RAG enhances factual accuracy, contextual relevance, and adaptability across knowledge-intensive tasks. This paper presents a detailed survey of RAG' s development, methods, and applications. The paper first present the evolution of mainstream RAG architectures including Retrieval-Enhanced Language Model (REALM), Dense Passage Retrieval (DPR), Fusion-in-Decoder (FiD), Hypothesis-Driven Enhancement (HyDE), Self-Retrieval-Augmented-Generation (Self-RAG), Corrective Retrieval- Augmented Generation (CRAG), Graph-based Retrieval-Augmented Generation (GraphRAG) - and then discuss their retrieval, reasoning, and error correction mechanisms respectively. Afterwards, we review the gradually widening applications of RAG including open-domain question answering, summarization, and domain-specific applications such as medicine and multimodal learning. We also address several challenges faced by RAG, such as retrieval level, computational efficiency, hallucination suppression, and interpretability. Finally, we explore several research directions for RAG, including adaptive retriever, self-reflective generation, and structured and multimodal knowledge sources. We believe that RAG is an important step toward more accurate, explainable, and knowledge-based AI, with significant potential for both theoretical research and practical applications in intelligent language understanding.

Keywords: Retrieval-Augmented Generation, Natural language processing, Large language models, Information retrieval, Knowledge-intensive tasks

1 Introduction

Natural Language Processing is a cross-field between linguistics, computer science, and artificial intelligence. Its goal is to allow computers to understand, generate human language, and even interact with it. It has tasks such as text classification, machine translation, speech recognition, and question-answering. In recent years, NLP has

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_48

greatly improved and greatly propelled the development of artificial intelligence and machine learning [1]. Despite this, NLP still has limitations when applied to tasks that require a large amount of external knowledge. That is, tasks where models need to understand and utilize domain-specific knowledge or time-dependent information [2]. In such a context, a new model called Retrieval-Augmented Generation (RAG) comes into play. By integrating information retrieval with generative models, RAG enables models to dynamically retrieve and incorporate relevant information from external knowledge bases during text generation. This retrieval process compensates for gaps in the model's training data, yielding more specialized and accurate responses [3]. This retrieval mechanism grants RAG substantial advantages when tackling knowledge-intensive tasks, allowing it to rely not only on training data but also to access external information in real time.

With the rapid development of RAG technology, many improved methods are proposed constantly. REALM is proposed by Guu et al. [4], in which retrieval process is added into pre-training procedure. Finally, the language model can reuse knowledge from external memory in the process of response generation. The ability of completing open-ended QA task is greatly improved. Subsequently, DPR is proposed by Karpukhin et al. [5]. In which, dual-tower architecture is used to vectorize query and document. The retrieval accuracy and efficiency are greatly improved by this method, and better document input is provided for RAG model. Subsequently, Izacard and Grave proposed a new method named Fusion-in-Decoder (FiD). In this method, multiple retrieved documents are used during the decoding process. More information is synthesized externally in the generation process, and excellent performance is achieved in QA and knowledge-intensive tasks. Based on these researches, we will propose new methods to further extend the application scenario of RAG [6]. Gao et al. proposed a method called Hypothesis-Driven Enhancement (HyDE), which generates hypotheses to guide retrieval, and thus improve the retrieved results [7]. Self-Retrieval-Augmented Generation (Self-RAG) proposed by Asai et al. allows models to reflect on their outputs and retrieve information to self-retrieve and self-correct the results more reliably [8]. corrective retrieval-Augmented Generation (CRAG) proposed by Yan et al. corrects the possible noise and deviation in the retrieved results, and further improves the accuracy of the model output [9]. Yu et al. proposed Graph-based Retrieval-Augmented Generation (GraphRAG), which uses structured knowledge graphs to help models deal with complex and multimodal tasks. These technologies show that RAG is applicable in complex application scenarios [10].

Although RAG models have many advantages over traditional models, there are also some challenges.

RAG models depend on external knowledge bases to improve the generation process. The accuracy and relevance of the retrieved results directly affect the accuracy and relevance of the generated results. The wrong or unmatched external information will cause the generated answer to be inaccurate or even false. In addition, as the knowledge base continues to expand, the problem of fast generation also exists in RAG for practical applications [11].

This paper aims to provide a systematic review of RAG-related technologies. The fundamental concepts and technical frameworks of RAG models and their variants will

be thoroughly introduced. Their applications across various domains, along with their respective strengths and weaknesses, will be analyzed. Furthermore, the primary challenges facing this technology will be discussed, and potential directions for future research will be explored.

2 Mainstream Techniques Overview

The emergence of RAG has sparked a wave of research integrating retrieval mechanisms into generative models. Current mainstream approaches primarily focus on optimizing various aspects of this paradigm, including retrieval efficiency, knowledge utilization, and error correction and reasoning for structured data. This section reviews the most representative models in the field, examining their architectures, contributions, and limitations.

2.1 Retrieval-Enhanced Language Model (REALM)

Proposed by Guu et al., REALM was the first model to integrate retrieval into the pre-training phase of language models. Unlike standard pre-trained models that rely solely on static corpora, REALM incorporates a differentiable retriever, enabling dynamic access to external knowledge during both pre-training and fine-tuning. Specifically, the retriever maps queries and documents into a dense vector space, while the generator uses retrieved paragraphs to produce output. REALM jointly optimizes both components, enabling language models to acquire factual knowledge without memorizing all information during training. In open-domain question answering tasks, REALM demonstrates higher accuracy than models without retrieval systems, validating the effectiveness of retrieval-augmented approaches. However, REALM's reliance on dense retrieval introduces new challenges: retrieval quality heavily depends on learned embeddings, meaning retrieval errors directly impact final generation. Additionally, its high pre-training cost limits scalability across extensive knowledge bases [4].

2.2 Dense Passage Retrieval (DPR)

Karpukhin et al. proposed DPR, representing a key advancement in the retrieval component of RAG systems. DPR employs a dual-encoder architecture, encoding queries and paragraphs as dense vectors respectively, and measures relevance through dot-product similarity. This approach replaces traditional sparse retrieval methods like BM25, achieving higher recall and precision across large-scale document collections. DPR's primary contribution lies in its effectiveness and efficiency for open-domain retrieval, providing high-quality document candidates that downstream generative models—including RAG—can leverage to produce more accurate outputs. Empirical studies demonstrate that DPR significantly outperforms sparse methods on popular QA benchmarks like Natural Questions and TriviaQA. However, when applied to multi-domain scenarios, DPR faces limitations due to its reliance on learned embeddings, which may not generalize well to unseen topics. Additionally, the computational cost of building and querying dense indexes increases as knowledge bases grow larger [5].

2.3 Fusion-in-Decoder (FiD)

FiD, proposed by Izacard and Grave, adopts a novel approach to integrating retrieved documents into the generation process. Unlike previous approaches, FiD does not concatenate retrieved paragraphs before encoding. Instead, it encodes each paragraph independently and fuses them within the decoder during text generation. This design enables the model to process large volumes of retrieved documents, providing the generator with richer contextual text without overburdening the encoder. Compared to original RAG models, FiD achieves higher accuracy on question-answering benchmarks. By enabling the decoder to process multiple documents simultaneously, FiD integrates information from diverse sources, yielding more precise and comprehensive responses. However, this resource-intensive task of separately encoding multiple channels and processing them during decoding demands higher computational power. Whether FiD can deliver fast, real-time responses in practical applications remains an unresolved challenge [6].

2.4 Hypothesis-Driven Enhancement (HyDE)

Gao et al. further optimized the retrieval process by introducing an intermediate reasoning step in HyDE. This reasoning step means that HyDE does not retrieve documents directly based on the user's query. Instead, it first generates hypotheses, reasonably rewrites or expands the query, and uses this content to guide retrieval. This mechanism broadens the scope of retrieval and increases the probability of discovering relevant documents, especially when the initial query is ambiguous. The introduction of hypothesis generation bridges the gap between query understanding and document retrieval. Experiments demonstrate that HyDE improves retrieval quality and downstream generation accuracy, particularly in tasks requiring complex reasoning or multi-hop retrieval. Similar to the aforementioned models, HyDE introduces additional complexity, and errors generated during the hypothesis phase can significantly impact subsequent retrieval and generation. Maintaining hypotheses that remain faithful to the original query meaning remains an open challenge [7].

2.5 Self-Retrieval-Augmented-Generation (Self-RAG)

Self-RAG, developed by Asai et al., introduces a self-reflection mechanism into RAG systems. This mechanism enables the model not only to utilize retrieved documents but also to evaluate its own outputs, identify potential errors and inconsistencies, and trigger additional retrieval when necessary. This iterative process enhances the robustness and reliability of the final generated results. Self-RAG addresses a key weakness of traditional RAG: the lack of self-monitoring during generation. By employing self-reflection, the model reduces hallucinations and improves factual accuracy. Experimental results demonstrate that Self-RAG outperforms standard RAG in knowledge-intensive QA tasks. However, its iterative mechanism increases inference time, revealing limitations in highly time-sensitive task [8].

2.6 Corrective Retrieval-Augmented Generation (CRAG)

Unlike Self-RAG, CRAG proposed by Yan et al. focuses on correcting errors generated during retrieval before they influence the generation process. Unlike traditional RAG

models that directly feed retrieved documents to the generator, CRAG introduces a correction module to filter and refine retrieved results, ensuring only high-quality, highly relevant documents are used for generation. This approach reduces noise in the input and lowers the risk of misleading outputs caused by erroneous inputs. CRAG demonstrates improvements in both accuracy and consistency of generated content. Its correction mechanism holds particular value in domains where generated outcomes carry significant implications, such as legal or medical fields. However, the added correction module also increases the overall pipeline complexity, and balancing correction with computational efficiency remains a current challenge [9].

2.7 Graph-based Retrieval-Augmented Generation (GraphRAG)

Yu et al. proposed GraphRAG, which extends RAG from another perspective—by integrating structured knowledge sources such as knowledge graphs. Unlike unstructured document retrieval, GraphRAG leverages graph-based relationships to retrieve and reason about entities and their linked content. This enables the model to capture more complex dependencies, providing more coherent answers in tasks requiring relational reasoning. GraphRAG represents a significant advancement in RAG systems by combining symbolic and neural approaches. Its ability to integrate structured and unstructured data enables it to tackle multimodal and complex reasoning tasks. However, constructing and maintaining such large-scale knowledge graphs is resource-intensive, and ensuring consistency between graph-based retrieval and unstructured text generation remains a challenge [10].

Overall, mainstream RAG technologies have evolved across multiple dimensions. REALM introduced retrieval into pre-training, DPR enhanced retrieval efficiency, and FiD optimized information fusion during decoding. Subsequently, HyDE, Self-RAG, CRAG, and GraphRAG addressed early model limitations such as retrieval relevance, self-reflection, noise correction, and structured data reasoning. These approaches demonstrate the rapid advancement of RAG research, showcasing the immense potential of retrieval-augmented methods for tackling knowledge-intensive NLP tasks. Despite this promising landscape, each new RAG paradigm involves trade-offs in efficiency, scalability, and reliability, underscoring the necessity for continued innovation in this field.

3 Applications, Challenges, and Future Directions of RAG

From a conceptual perspective, RAG has matured from being a framework into a mainstream paradigm in multiple fields, especially in AI. Due to its capability of leveraging retrieved knowledge in a dynamic way during generation, RAG is able to transcend one of the biggest challenges in NLP: the boundedness of parameter-dependent knowledge [1]. With the maturity of RAG technologies, their applications have been extensively extended to knowledge-intensive scenarios, domain-specific fields, and even multimodal settings. However, these extensions also bring some critical challenges regarding efficiency, reliability, and interpretability into the open. In this section, we present a detailed review of the current RAG applications, challenges, and future works.

3.1 Expanding Applications of RAG

RAG has shown great promise in knowledge-intensive NLP tasks such as open-domain question answering (QA), information retrieval, and summarization. By recalling from external knowledge sources, RAG models maintain both factual correctness and up-to-datedness, which is the features beyond the static large language models (LLMs). For example, systems built with RAG architectures obtain state-of-the-art performance on several open-domain QA benchmarks such as Natural Questions and TriviaQA [1, 6]. The retrieval process allows the model to back up its answers with reliable sources, which results in a great drop of hallucinations and increases user's confidence in the system. In addition to answering QA, RAG also improves the capability of generating abstract summaries and reports. Maintaining the coherence between original data and generated descriptions is essential in this task. RAG introduces a retrieval part of original data and is less likely to make mistakes such as factual drift. In customer service and education scenarios, a knowledge-informed assistant powered by RAG is able to provide up-to-date and contextually relevant answers for users, making it more reliable in a rapidly changing knowledge space. In all the above applications, RAG connects the static parameters of models with running knowledge in the real world.

Due to its flexibility, RAG can also be employed for domain-specific tasks, especially in domains where factual accuracy is critical. In medicine, we see retrieval-Augmented models being used for clinical applications such clinical support, medical QA , and literature-based reports. For example, models retrieving from biomedical text repositories such as PubMed show enhanced diagnostic reasoning and interpretability [12]. At the same time, RAG has gradually been extended to new types of scenarios. GraphRAG is a Retrieval-Augmented framework that incorporates relational information alongside textual data, allowing models to reason about entities and their relations [10]. Multimodal RAG frameworks have also since been developed that include visual retrieval capabilities for downstream visual question answering and cross-modal summarization tasks.

3.2 Core Challenges in RAG Systems

Although RAG has made great strides forward, there are still many fundamental problems that need to be solved before it can be widely used. RAG success rates depend strongly on the effectiveness of the retrieved documents. Irrelevant or outdated documents may be retrieved and thus may affect the generation module to produce incorrect or contradicting answers. This is a particularly serious concern in domains such as medicine and law, where the correctness of facts is a matter of life and death, and simple retrieval of relevant, up-to-date, and trustworthy information sources is a top priority. The amount of external knowledge bases is increasing dramatically. How to efficiently retrieve relevant information has become a new problem. Large vector databases in dense retrieval models such as DPR are very accurate but expensive to maintain. These models also have low document encoding workload; however, the models inference cost is high due to strong performance. Finding a balance between retrieval accuracy and computational efficiency is an important goal for next generation RAG architectures. Hallucination also remains a challenge for current RAG systems. While RAG mitigates hallucinations to some extent by relying on retrieved evidence,

it cannot eliminate them entirely. The generation module may still produce outputs inconsistent with retrieved documents or misinterpret retrieved facts. Recent approaches like Self-RAG and CRAG attempt to address this by incorporating reflection and correction mechanisms, but the problem of achieving a fully reliable factual foundation remains unresolved. Finally, RAG also faces interpretability challenges.

RAG models introduce additional complexity by requiring tracking of which retrieved sources influenced the final output. In high-stakes applications, users and regulators demand transparent reasoning paths. However, most existing RAG implementations treat retrieval and generation as black-box processes. Building interpretable RAG frameworks that explicitly justify their outputs represents an important yet under-explored direction.

3.3 Future Research Directions

Overcoming these limitations is beyond the capabilities of current methods and systems. Ongoing research on RAG may pursue the following directions: First, RAG should be endowed with adaptive and self-improving retrieval. Designing retrievers that can adapt to changing knowledge bases will be an important research direction. Methods that combine RL with online updates can help RAG systems to keep up to date without retraining from scratch. Second, although Self-RAG can already engage in a limited form of reflection, this may generalize into more general metacognitive RAG architectures that can evaluate the reliability of generated outputs, detect factual errors, and even automatically trigger corrective retrieval upon the detection of an error. Finally, future RAG systems may enable more effective integration of graph-based multimodal representations. RAG systems that combine symbolic reasoning with neural generation will support even more expressive reasoning capabilities.

4 Conclusion

Retrieval-Augmented Generation (RAG) represents a significant milestone in the evolution of natural language processing. By combining the strengths of information retrieval and text generation, RAG effectively addresses a fundamental limitation of large language models—their reliance on static, parameter-based knowledge. Through dynamic access to external information sources, RAG enhances factual accuracy, expands situational awareness, and improves the reliability of model outputs in knowledge-intensive tasks.

The rapid evolution of RAG-related architectures—such as REALM, DPR, FiD, HyDE, Self-RAG, CRAG, and GraphRAG—demonstrates how integrating retrieval and generation can boost performance across multiple dimensions. Each model uniquely contributes to this paradigm: REALM introduces retrieval during pretraining, DPR improves retrieval precision, FiD optimizes document fusion, HyDE refines retrieval relevance through hypothesis generation, and Self-RAG and CRAG enhance factual consistency via reflection and correction mechanisms. GraphRAG further extends the paradigm by integrating structured and multimodal data, demonstrating

RAG's potential beyond textual domains. Collectively, these advancements underscore the versatility and maturing sophistication of the RAG framework.

Despite these achievements, RAG still faces several unresolved challenges. Reliance on high-quality retrieval results remains a critical constraint, as irrelevant or outdated information can propagate errors during generation. Computational efficiency, scalability, and interpretability also present ongoing hurdles, particularly as knowledge bases expand and user expectations rise. Addressing hallucinations, improving retrieval-generation alignment, and establishing transparent reasoning paths are crucial for RAG's transition from research innovation to widespread practical deployment.

Looking ahead, future research should focus on developing adaptive, self-correcting, and explainable RAG systems. Integrating reinforcement learning, dynamic retrievers, and multimodal reasoning frameworks can enable models to continually refine their understanding and maintain factual consistency in real time. As RAG continues to evolve, it has the potential not only to redefine the boundaries of knowledge-based language modeling but also to form the foundation for more trustworthy, explainable, and knowledge-aware AI systems.

References

1. Jamuie, A., Blessing, F.: Natural Language Processing: A Comprehensive Survey. ResearchGate. (2025)
2. Lewis, P., Perez, E., Piktus, A., et al.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401. (2020)
3. Huang, Y., Huang, J.: A Survey on Retrieval-Augmented Text Generation for Large Language Models. https://doi.org/10.48550/arXiv.2404.10981. (2024)
4. Guu, K., Lee, K., Tung, Z., et al.: REALM: Retrieval-Augmented Language Model Pre-Training. https://doi.org/10.48550/arXiv.2002.08909. (2020)
5. Karpukhin, V., Oğuz, B., Min, S., et al.: Dense Passage Retrieval for Open-Domain Question Answering. https://doi.org/10.48550/arXiv.2004.04906. (2020)
6. Izacard, G., Grave, E.: Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Presented at the Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, pp. 874–880. https://doi.org/10.18653/v1/2021.eacl-main.74. (2021)
7. Gao, L., Ma, X., Lin, J., et al.: Precise Zero-Shot Dense Retrieval without Relevance Labels. https://doi.org/10.48550/arXiv.2212.10496. (2022)
8. Asai, A., Wu, Z., Wang, Y., et al.: Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. https://doi.org/10.48550/arXiv.2310.1151. (2023)
9. Yan, Z., Bi, M., Zhang, Q., et al.: LncRNA TUG1 promotes the progression of colorectal cancer via the miR-138-5p/ZEB2 axis. Biosci. Rep., 40, BSR20201025. (2020). https://doi.org/10.1042/BSR20201025.
10. Yu, H., Gan, A., Zhang, K., et al.: Evaluation of Retrieval-Augmented Generation: A Survey. pp. 102–120. https://doi.org/10.1007/978-981-96-1024-2_8. (2025)
11. Edge, D., Trinh, H., Cheng, N., et al.: From Local to Global: A Graph RAG Approach to Query-Focused Summarization. arXiv:2404.16130. (2025). https://doi.org/10.48550/arXiv.2404.16130

12. Singhal, K., Azizi, S., Tu, T., et al.: Large language models encode clinical knowledge. Nature, 620(7972), 172–180. (2023). https://doi.org/10.1038/s41586-023-06291-2