Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ30058

The Development Status of Retrieval-Augmented Generation (RAG) Technology

Chenhui Zhu

Institute of School of Software, College of Computer Science and Technology, Southwest University, Chongqing, China
Z.3095820544@outlook.com

Abstract. Currently, Retrieval-Augmented Generation (RAG) technology is the core method for breaking through the knowledge limitations of large models. This article provides an in-depth analysis of the development history and mainstream techniques of RAG. Early advanced RAG systems mostly followed a passive "retrieve-then-generate" process. The milestone SELF-RAG technology introduced self-reflection and decision-making capabilities. It enabled a RAG system to actively determine whether retrieval is necessary and if the retrieved content is useful for the first time. This has pushed RAG from static process optimization to a new stage of development guided by intelligent decision-making. Building on this, the paper further dissects the limitations of SELF-RAG in terms of cost and decision logic and outlines the emerging trends of the post-SELF-RAG era, characterized by adaptivity, modularity, and agentic behavior. By integrating subsequent works such as Adaptive RAG and Corrective Retrieval Augmented Generation (CRAG), this paper aims to delineate the technological evolutionary path, explore how to leverage new advancements to address its shortcomings, and envision future paradigms for more efficient and intelligent RAG applications.

Keywords: Retrieval-Augmented Generation, Self-Reflective Retrieval-Augmented Generation, Self-Reflection, Adaptive Retrieval-Augmented Generation, Intelligent Agent

1. Introduction

Retrieval-Augmented Generation (RAG) has emerged as a pivotal technology in tackling the difficulties presented by Large Language Models (LLMs). These difficulties encompass issues like factual inaccuracies, outdated world knowledge, and queries that are knowledge-intensive or specific to certain domains. The RAG architecture manages to bridge the gap between information retrieval and text generation by dynamically introducing external knowledge into the language model. At present, the research outcomes in the RAG field seem to be somewhat disorganized and lack a cohesive framework, which underscores the need for a thorough and methodical comprehensive review. Even though research into RAG technology is thriving, most of the surveys concentrate on enumerating and classifying various advanced RAG techniques without offering a clear, unbroken evolutionary account that elucidates the underlying forces propelling its intrinsic paradigm shifts. The motivation and originality of this study reside in its characterization of Self-

DOI: 10.54254/2755-2721/2026.TJ30058

Reflective Retrieval-Augmented Generation (SELF-RAG) not just as one of many enhanced methods, but as a crucial node that separates two distinct eras of RAG technology. Prior to the introduction of SELF-RAG (roughly from around 2022 to mid-2023), RAG development was mainly centered on refining a fixed retrieval→generation pipeline. Researchers endeavored to boost the accuracy of the retriever, for instance, through Dense Passage Retrieval (DPR) [1], and to optimize the integration of the generator with the retrieved content, as evidenced in Fusion-in-Decoder (FiD) and Retrieval-Enhanced Transformer (RETRO) [2,3]. Nevertheless, these initial improvements were inherently passive and lacked the capability to modify their behavior dynamically based on specific queries. This resulted in unnecessary computational burden and potential performance deterioration when confronted with simple questions that did not necessitate retrieval or when irrelevant information was retrieved.SELF-RAG (Self-Reflective Retrieval-Augmented Generation), first introduced in October 2023 [4], represents a significant milestone. For the first time, it integrates self-reflection and active decision-making capabilities into the RAG framework. It performs self-evaluation at each step of the process, dynamically deciding whether retrieval is necessary, if the retrieved content is relevant, and whether the generated output is reliable. This moves beyond the previous paradigm of indiscriminately executing retrieval before generation. This shift in paradigm signals the potential for RAG to evolve from a fixed data-processing logic into an intelligent system with rudimentary reasoning capabilities, laying the foundation for subsequent research on adaptive and agentic RAG.

This paper is intended for researchers in Natural Language Processing (NLP) and Retrieval-Augmented Generation (RAG), as well as other interested parties. It aims to help readers identify research gaps and promising future directions, and to offer practical guidance for those seeking to apply RAG technology.

2. Research methods and evaluation benchmarks

2.1. Overview of mainstream methods

2.1.1. The pre-SELF-RAG era (approx. 2020 – mid-2023): internal optimization of a fixed pipeline

The research focus during this stage was on optimizing the fixed paradigm of retrieving information and then directly generating a response. The goal was to enhance the efficiency and effectiveness of this passive process, without yet incorporating decision-making. This period saw many key advancements. In the area of end-to-end retrieval learning, REALM pioneered the end-to-end training of retrievers, while DPR [1,5], centered on contrastive learning, proposed dense passage retrieval, which became a mainstream retriever baseline. Simultaneously, researchers also focused on optimizing the mechanisms for both generation and retrieval. For instance, FiD improved information utilization by fusing multiple documents at the decoder level, and RETRO deeply embedded retrieval into the model [2,3], demonstrating that the combination of a smaller model with retrieval could rival the performance of much larger models. Subsequent research explorations shifted towards more lightweight and flexible architectures. RePlug designed the retriever as an external plugin, and HyDE utilized an LLM to generate a hypothetical document as a query, respectively advancing the concept of modularity and cleverly addressing the semantic gap problem [6].

DOI: 10.54254/2755-2721/2026.TJ30058

2.1.2. SELF-RAG (October 2023): introducing self-reflection: a paradigm shift this section must be in one column

SELF-RAG introduces capabilities for self-evaluation and reflection, breaking from the fixed pipeline of previous retrieve-and-generate models. This marks a fundamental shift for RAG from passive optimization to active decision-making.

Asai et al. proposed a paradigm-shifting breakthrough: by having the model generate special "reflection tokens," they constructed a closed-loop decision process of retrieval, generation [4], and critique. This enables the model to dynamically judge when retrieval is needed, whether the retrieved content is relevant, and whether the generated content is reliable. This active decision-making capability is pivotal, as it demonstrates that RAG systems can evaluate and self-correct their own behavior. This not only highlights the immense potential of RAG but also lays a crucial foundation for the subsequent development of agentic paradigms.

This trend toward active decision-making was not an isolated event but a reflection of a broader technological current at the time. Proposed almost concurrently with SELF-RAG, FLARE (Forward-Looking Active Retrieval Augmented Generation) adopted a forward-looking active retrieval strategy, proactively pausing generation to retrieve information when it predicted low-confidence content might appear [7]. Another representative work, Corrective Retrieval Augmented Generation (CRAG) [8], focused on correcting retrieval results. It used a lightweight evaluator to judge the relevance of retrieved documents and triggered supplementary search or rewriting strategies when quality was poor. Although the specific implementations of these methods differ—SELF-RAG focuses on comprehensive self-critique, FLARE on proactive retrieval, and CRAG on retrieval correction—they collectively pushed RAG from a passive process to an intelligent framework with dynamic feedback and correction capabilities.

2.1.3. The post-SELF-RAG era (late 2023 to present): towards adaptivity, modularity, and agentic systems

Inspired by SELF-RAG, subsequent research has evolved towards greater efficiency, robustness, and intelligence, exhibiting three major trends. The first is adaptive and controllable retrieval. Works like Adaptive RAG and FLARE aim to let the model dynamically decide the timing of retrieval to balance cost and effectiveness [7,9], reducing unnecessary computational overhead. The second is modularity and multi-source fusion. For example, CRAG introduced a corrective module that automatically triggers supplementary searches when retrieval quality is poor [8], enhancing robustness. Meanwhile, works like Graph-Toolformer address the need for structured knowledge and multi-hop reasoning by training models to call graph computation tools [10]. The most cutting-edge trend is the evolution of RAG towards an agentic paradigm. In frameworks like LlamaIndex and LangChain, RAG is treated as a tool or sub-module that an agent can call upon to perform multi-step reasoning, enabling it to handle tasks far more complex than simple question-answering.

2.1.4. Performance and adoption comparison across stages

To visually illustrate this evolutionary process, Table 1 summarizes the representative methods from the three development stages across two key dimensions: core performance and practical application. The table clearly reveals the progressive enhancement of RAG technology in both performance metrics and practical value.

DOI: 10.54254/2755-2721/2026.TJ30058

Table 1. Performance and application comparison of RAG technology development stages

Development Stage/Representative Method	Core Performance	Application & Adoption
Advanced RAG (approx. 2020 – Mid-2023)	• 20-40% improvement in open-domain QA. • RETRO achieved GPT-3-level performance with a smaller scale.	Good academic results, but limited industrial application due to tight coupling and high costs.
SELF-RAG (October 2023)	• ~10pp accuracy increase on PopQA vs. traditional RAG. • 20-30pp increase in citation precision for long-form text, with significantly reduced hallucination.	Strong academic baseline, but mostly used in research due to high training and inference overhead.
Adaptive/Agentic RAG (Late 2023 to Present)	• Efficiency: Adaptive/FLARE reduce costs by 30-50%. • Robustness: CRAG improves accuracy on noisy data by 10-15%. • Complex Reasoning: Methods like Graph-Toolformer significantly enhance multi-hop reasoning on graph structures.	Entering production use (e.g., Copilot, enterprise knowledge bases) due to modularity, low cost, and high robustness.

2.2. Core technical architecture

The core idea of SELF-RAG is to transform the LLM itself into a decision-maker and evaluator. Its process is no longer linear but a closed loop involving retrieval, generation, and reflection. Specifically, during generation, the LLM proactively generates special reflection tokens to answer a series of key questions. First, when faced with a user query, it decides whether it is necessary to Retrieve information from an external knowledge base. After retrieving documents, it evaluates whether each document is relevant to the query (Relevance) and whether it can provide support for generating the answer (Support). Finally, after generating each segment of the answer, it assesses whether the segment is supported by the retrieved information, and is relevant and useful (IsUse). In this way, SELF-RAG transforms a passive process into a dynamic, active reasoning process capable of on-demand retrieval and real-time self-correction.

2.3. Evaluation benchmarks

To systematically evaluate the effectiveness of RAG technology, the research community relies on a series of standardized benchmark datasets that challenge a model's knowledge acquisition, factuality, and reasoning capabilities from different angles. Among these, Open-Domain Question Answering (Open-Domain QA) is a classic testing ground for RAG. In addition to Natural Questions (NQ) and TriviaQA, it includes MS MARCO, derived from real search engine queries, and HotpotQA, which tests multi-hop reasoning. These datasets require the model to retrieve and generate precise factual answers from vast amounts of knowledge. To comprehensively test a model's generalization ability, Knowledge-Intensive Tasks (KILT) benchmarks were developed, integrating various scenarios such as fact-checking (e.g., FEVER) and long-form explanation (e.g., ELI5). Furthermore, to evaluate RAG's core value in suppressing hallucinations, researchers use factuality and robustness test sets. For example, PopQA is used to test its grasp of long-tail knowledge, TruthfulQA tests its ability to resist common misinformation, and datasets like FreshQA evaluate its timeliness in acquiring recent knowledge.

DOI: 10.54254/2755-2721/2026.TJ30058

2.4. Evaluation metrics

Evaluating the performance of RAG systems usually entails a set of multi-dimensional metrics to guarantee a thorough assessment. A key metric is Accuracy, which gauges the model's performance in question-answering tasks and is measured using Exact Match (EM) or F1 score. Citation Precision and Recall are employed to assess the accuracy and comprehensiveness of the cited sources, an important element in evaluating the system's reliability. On the other hand, the Hallucination Rate determines the percentage of factually incorrect information in the generated content, directly showing the degree to which the model fabricates information. To fully highlight a system's technical advantages, researchers often carry out comparative analyses with traditional RAG models and language models of different scales, like ChatGPT.

3. Experimental results and performance evaluation

3.1. Benchmark comparison

SELF-RAG was compared against several baseline methods, including standard LLMs (no retrieval), traditional RAG (fixed retrieval), and large conversational models like ChatGPT.

3.2. Performance and evolutionary analysis

In terms of performance, SELF-RAG demonstrates significant advantages over traditional RAG methods and has established new evaluation dimensions for subsequent technological evolution. Specifically, on the PopQA dataset, which tests long-tail knowledge, its accuracy increased by approximately 10 percentage points. Furthermore, in long-form generation tasks with stricter factuality requirements, its citation precision improved substantially by 20 to 30 percentage points, while significantly reducing the occurrence of hallucinations. This outstanding performance quickly established it as a powerful baseline model in academia and led to its integration into mainstream RAG frameworks like Llama Index.

This performance improvement is even more striking when compared to its predecessors. Although early advanced RAG methods like DPR and RETRO had already achieved 20-40% performance gains on open-domain QA tasks, their core contribution was in improving the accuracy of a fixed process. In contrast, the value of SELF-RAG lies not only in accuracy but also in achieving a qualitative leap in credibility and autonomy through its self-reflection mechanism. However, due to its high inference overhead and training complexity, direct application of SELF-RAG in industrial production environments that demand low latency and high efficiency remains relatively limited.

Its limitations become evident upon comparison with methods that emerged after the SELF-RAG era, given that later research managed to attain targeted performance improvements in numerous aspects. In terms of efficiency, approaches such as Adaptive RAG and FLARE, by means of more meticulously designed on-demand retrieval strategies, succeeded in reducing the computational cost by 30 to 50% without compromising performance. This effectively dealt with the issue of high evaluation overhead associated with SELF-RAG. As for robustness, CRAG’s corrective mechanism enhanced accuracy by 10 to 15% when confronted with low-quality or noisy retrieval outcomes, thereby making up for SELF-RAG’s heavy reliance on the quality of initial retrieval. With regard to complex reasoning, Graph-Tool Former secured notable performance advancements on multi-hop

DOI: 10.54254/2755-2721/2026.TJ30058

reasoning tasks by empowering the model to utilize graph computation tools. This broke through the constraints of traditional RAG in managing structured, interconnected knowledge.

In summary, the experimental results clearly outline an evolutionary path: from the pursuit of higher accuracy in early RAG, to the more reliable and autonomous SELF-RAG, and finally to the multi-dimensional optimization for greater efficiency, robustness, and intelligence that characterizes current cutting-edge methods.

4. Model limitations and technical challenges

Even though decision-making RAG, with SELF-RAG as a typical example, has achieved notable advancements and brought about a paradigmatic change, it is yet to confront a string of quite daunting challenges along the way towards greater intelligence. Such challenges can be seen not only in the application and design of the particular SELF-RAG model but also in the general bottlenecks that the whole RAG technology paradigm is up against at present. This chapter will carry out a thorough analysis from these two aspects.

4.1. Specific limitations of SELF-RAG

As a key node in the paradigm shift, SELF-RAG, while demonstrating great potential, also reveals specific limitations in its practical application and popularization through its own design and validation process. At the application and deployment level, its generalizability and cost are major constraints. The effectiveness of SELF-RAG has been primarily validated on English general-domain datasets; its performance in low-resource languages, specific vertical domains (e.g., medical, legal), or few-shot scenarios has not been fully examined, limiting the breadth of its direct application. Moreover, experiments were mainly conducted on medium-sized open-source models. For very large-scale closed-source models like GPT-4, their powerful internal knowledge and reasoning capabilities may diminish the marginal benefits of on-demand retrieval. Additionally, its high training and inference costs pose a severe deployment challenge. The model requires extra fine-tuning to learn and understand reflection tokens, and its inference process involves multiple model calls for evaluation, significantly increasing computational overhead and response latency. This is a significant constraint for production environments that prioritize low cost and high efficiency.

Beyond external application challenges, the innovative mechanism of SELF-RAG also introduces a series of internal design trade-offs, introducing new risks while solving old problems. Table 2 provides a detailed analysis of the built-in strategies designed by SELF-RAG to address specific challenges and their associated limitations. The table intuitively reveals the new trade-offs introduced by its innovative mechanisms.

DOI: 10.54254/2755-2721/2026.TJ30058

Table 2. Analysis of SELF-RAG's built-in solutions and their limitations

Problem Type	SELF-RAG's Solution	Potential Shortcomings
Handling Complex Queries	Dynamically decides whether to retrieve and use information through self-reflection.	The decision logic is relatively rigid (yes/no), lacking more fine-grained strategies, such as how much to retrieve or what type of information to retrieve.
Ensuring Factuality	Evaluates whether generated content is supported using IsSup and IsUse tokens.	Relies on the model's own evaluation capability. If the critic model itself hallucinates, the evaluation fails.
Improving Efficiency	Skips retrieval for simple questions to avoid unnecessary overhead.	The evaluation step itself adds overhead. For the majority of questions that require retrieval, the overall latency actually increases.

These issues specifically manifested in SELF-RAG—such as the rigidity of decision logic and the reliability risks of self-evaluation—also reflect the deeper, common challenges faced by the entire RAG paradigm at its current stage.

4.2. Common challenges of the RAG paradigm

Building on the limitations of SELF-RAG, we can identify broader and more fundamental challenges facing the entire RAG technology paradigm. At the methodological level, current RAG frameworks commonly face several core bottlenecks. First, the performance of most frameworks is constrained by a "retrieval bottleneck"; if the initial retrieval results have low recall or poor quality, subsequent stages are rendered ineffective, and systems generally lack built-in mechanisms for "corrective retrieval" or "query rewriting." Second, even when relevant information is retrieved, how to deeply and non-conflictingly fuse it with the model's internal parametric knowledge, rather than simple concatenation, remains a key challenge that directly affects the coherence and quality of the answer. Finally, the reliability and risk of error accumulation in the entire system cannot be ignored. While the risk in traditional RAG stems from retrieval quality, the introduction of self-evaluation mechanisms adds a new risk point: the critic's judgment may be flawed. Errors can accumulate and be amplified in subsequent steps, reducing the reliability of the final output.

Aside from methodological constraints, the advancement of RAG technology is hindered by the external evaluation ecosystem as well. For one thing, numerous existing benchmark datasets, despite being standardized, have rather simplistic question formats which might fail to completely mirror the intricate and ambiguous query situations in the real world, resulting in a gap between the model's performance in testing environments and its practical applications. For another thing, present evaluation standards heavily depend on objective metrics like accuracy, but insufficiently take into account higher-dimensional quality aspects such as comprehensiveness, logical consistency, timeliness, and explainability. This one-dimensional evaluation approach might cause research to excessively optimize for specific metrics while overlooking the overall user experience in real-world scenarios.

DOI: 10.54254/2755-2721/2026.TJ30058

5. Feasibility analysis of improvement solutions

5.1. Post-SELF-RAG technological evolution and improvement schemes

The ideas of SELF-RAG have guided the subsequent development of RAG technology. By analyzing cutting-edge research from late 2023 to the present, it becomes clear that its limitations are being addressed by subsequent modular, adaptive, and agentic approaches.

To address the issues of high cost and rigid logic, adaptive and controllable retrieval methods have emerged. Works like Adaptive RAG and FLARE train models to dynamically decide the timing and quantity of retrieval [7,9], focusing on reducing cost and latency. These can be seen as lightweight and engineering-focused implementations of SELF-RAG's on-demand retrieval concept.

In order to make up for its reliance on retrieval quality, researchers have come up with solutions that include correction as well as the fusion of multi-source knowledge. For instance, CRAG presents a lightweight corrective module [8]. If the model determines that the retrieval results are of low quality, it will automatically initiate an extra web search or query rewriting, which effectively alleviates the shortcomings of SELF-RAG when dealing with low-quality retrievals. At the same time, works such as Graph-Toolformer address the limitation of traditional RAG's dependence on unstructured text by giving the model the ability to use graph tools for structured reasoning, allowing it to handle complex problems that require multi-hop reasoning [10].

To transcend its mere single-step decision logic, the cutting-edge trend lies in the progression of RAG toward an agentic paradigm. Within an Agentic RAG framework, RAG is regarded as a tool that an Agent has the ability to utilize. A central planning LLM determines when and in what manner to employ RAG, and can even synchronize it with other tools such as code interpreters and search engines. This significantly enhances SELF-RAG's decision-making capabilities, allowing it to carry out intricate multi-step tasks.

These technologies have the potential to be combined with the SELF-RAG paradigm. As an illustration, the reflection capability of SELF-RAG might function as a crucial evaluation module inside an agent's decision loop, thereby giving rise to an even more potent system.

5.2. Future research directions and outlook

To tackle the challenges present in RAG's methodology as well as its evaluation systems, future research is anticipated to delve deeper in these crucial directions so as to construct RAG systems that are smarter, more dependable, and more pragmatic.

Toward a Complete Agent Framework: In the future, RAG won't merely be a static process or a simple callable tool; instead, it will be thoroughly integrated into an agent framework that's propelled by LLMs. This implies that RAG systems will have the abilities of iterative planning as well as dynamic policy adjustment. They'll be capable of independently determining when to retrieve information, how to rephrase queries, if multi-source cross-validation should be carried out, and how to merge different tools such as retrieval, code execution, and web browsing. This will facilitate profound coordination between the retrieval and generation processes, thereby enabling the solving of complex tasks that extend far beyond the present scope. Constructing More Dynamic and Realistic Evaluation Benchmarks: In order to deal with the limitations of present evaluation systems, the research community has to create dynamic evaluation benchmarks which are capable of being continually updated. These benchmarks ought to encompass a greater number of adversarial questions, like ones featuring factual conflicts, lacking clear-cut answers, or necessitating long-chain reasoning, so as to more thoroughly put the model's robustness and reasoning capabilities to the test.

DOI: 10.54254/2755-2721/2026.TJ30058

Moreover, crafting end-to-end evaluation procedures that are customized for particular application scenarios instead of standalone academic tasks will be vital for gauging the practical worth of RAG systems. Developing Multi-dimensional, Automated Evaluation Paradigms: Future evaluations will move beyond traditional accuracy metrics towards a more comprehensive quality assessment system. This includes promoting automated evaluation methods like LLM-as-a-Judge, which score generated content across multiple dimensions such as fluency, relevance, and factual consistency. At the same time, strengthening the analysis of Attribution and explainability will be vital to ensure that every conclusion from the system is verifiable, thereby comprehensively enhancing its trustworthiness for applications in critical domains like healthcare and finance.

6. Conclusion

This study, with SELF-RAG serving as a crucial anchor point, methodically maps out the development process of Retrieval-Augmented Generation (RAG), from its initial stage of merely being a passive process optimization tool to becoming an actively engaged, intelligent decision-making platform. The research clearly shows that the self-reflection mechanism and the active decision-making mechanism brought about by SELF-RAG mark a highly important turning point in the historical course of the technology. The improvements it has made in boosting factual accuracy as well as the capability it has endowed for on-demand retrieval have paved the way for new research paths in the following stages.

However, SELF-RAG is not without its limitations. In comparison with subsequent adaptive, modular, and agentic RAG methods, it exhibits clear constraints in terms of training and inference overhead, the flexibility of its decision-making logic, and its robustness against varying retrieval quality. Current technological trends indicate that RAG is evolving from a monolithic, integrated model into an intelligent system composed of multiple pluggable and interoperable tool modules.

Tracing the whole evolutionary process of RAG - starting from its initial stage of optimizing fixed pipelines, going through the introduction of self-reflection in SELF-RAG, up to the emergence of today's agentic and modular paradigms - it is quite clear that these systems are getting more and more autonomous, flexible, and powerful day by day. The future research will concentrate on enabling RAG agents to better plan, coordinate, and integrate various knowledge sources and tools so as to deal with the ever-increasingly complicated real-world problems.

References

[1] Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D. and Yih, W. (2020) Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).

[2] Izacard, G. and Grave, E. (2021) 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 (EACL).

[3] Borgeaud, S., et al. (2022) Improving language models by retrieving from trillions of tokens. In Proceedings of the 39th International Conference on Machine Learning (ICML).

[4] Asai, A., Wu, Z., Wang, Y., Sil, A. and Hajishirzi, H. (2023) Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In Advances in Neural Information Processing Systems 36 (NeurIPS).

[5] Guu, K., Lee, K., Tung, Z., Pasupat, P. and Chang, M.W. (2020) REALM: Retrieval-augmented language model pre-training. In Proceedings of the 8th International Conference on Learning Representations (ICLR).

[6] Gao, L., Ma, X., Lin, J. and Guo, J. (2023) Precise zero-shot dense retrieval without relevance labels. In Proceedings of the 11th International Conference on Learning Representations (ICLR).

[7] Jiang, Z., et al. (2023) Active retrieval augmented generation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP).

DOI: 10.54254/2755-2721/2026.TJ30058

[8] Yan, S.Q., Gu, J.C., Zhu, Y. and Ling, Z.H. (2024) Corrective retrieval augmented generation. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL).
[9] Jeong, S., Baek, J., Cho, S., Hwang, S.J. and Park, J.C. (2024) Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).
[10] Zhang, Y., et al. (2024) Graph-Toolformer: A large language model that can use graph tools. In Proceedings of the 12th International Conference on Learning Representations (ICLR).