INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 10, Issue 5 - V10I5-1242)
Available online at: https://www.ijariit.com

Self-Corrective Retrieval-Augmented Generation

Priya Jadam
priya.jadam@cmr.edu.in
CMR University, Bengaluru, Karnataka

Syeeda Mujeebunnisa
syeeda.m@cmr.edu.in
CMR University, Bengaluru, Karnataka

ABSTRACT

Though they are quite good at producing text, large language models (LLMs) frequently make mistakes or give incorrect information. This occurs as a result of LLMs heavy reliance on training material, which may eventually become outmoded or lacking. Retrieval-Augmented Generation (RAG) was developed as a solution to this problem. In RAG, pertinent data is retrieved and integrated from outside sources by the model. RAG does have several drawbacks, though, like the ability to retrieve superfluous or irrelevant data, which might confuse the model and produce inaccurate or ineffective results. Self-Corrective Retrieval-Augmented Generation (SCRAG), a novel method, attempts to address these issues by merging the internal knowledge of the model with the world data systems. In SCRAG, the model uses a technique called reflection tokens to assess the value of the information it retrieves in addition to retrieving it. This enables the model to modify its behavior according on the task and the caliber of the data it has acquired. In order to address this, SCRAG includes a simple method for evaluating the accuracy of the data that is retrieved. The model conducts a more thorough search—it even retrieves information from the internet to identify more reliable sources if the data is erroneous or insufficient. SCRAG also employs a decompose-then-recompose procedure that aids in the model's ability to dissect the recovered data, concentrate on the most pertinent portions, and eliminate unimportant information. This guarantees that the model produces accurate and trustworthy replies by using only high-quality data.

KEYWORDS: *Self - Corrective RAG, Retrieval Augmented Generation, LLMs Based RAG, Knowledge Based RAG*

1. Introduction

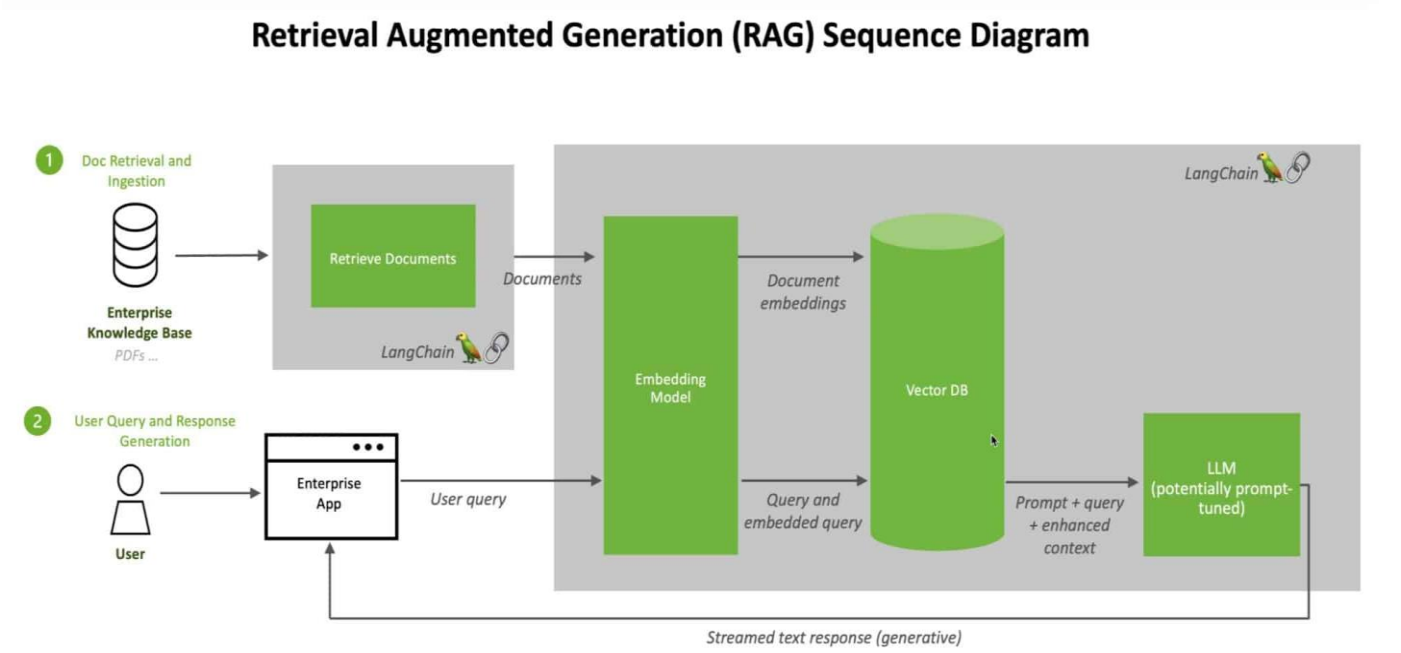
Even with greater model sizes and more data, large language models (LLMs) still suffer with factual errors and hallucinations, despite their impressive abilities to interpret and produce fluent language. The reason for this issue is that LLMs mostly rely on their internal, sometimes out-of-date, parametric information. In order to improve factual accuracy in knowledge-based activities, Retrieval-Augmented Generation (RAG) has been introduced to enable models to retrieve pertinent information from external sources. Unfortunately, current RAG approaches frequently retrieve a predetermined number of documents arbitrarily, which may include information that is superfluous or irrelevant. This results in poor generation quality and limits the applicability of the model. Our suggestion is to use Self-Corrective Retrieval-Augmented Generation (Self Corrective-RAG) to address these issues. Two fundamental concepts are combined in this new method: corrective retrieval and self-reflection. Using unique tokens known as reflection tokens, self-reflection allows the model to assess the value of retrieved passages and its own generated content in real-time. With the help of these tokens, the model may modify its retrieval behavior in response to task requirements, figuring out when and how to use retrieval. This guarantees that retrieval is applied only when it will significantly enhance the generation.

Apart from introspection, Self-Corrective-RAG presents a corrective retrieval procedure that assesses the caliber of documents that are recovered. When necessary, a lightweight retrieval evaluator can initiate further retrieval operations, such as extensive online searches, in addition to verifying the accuracy and relevancy of the retrieved sections. This guarantees that the model can access more varied and extensive data, rather than being restricted to static, pre-existing corpora. Moreover, the retrieved content is refined using a decompose-then-recompose method, which eliminates superfluous information and concentrates only on crucial insights that are essential to the work.

Because of its plug and play and flexible nature, Self-Corrective-RAG can be integrated with a variety of RAG based methodologies. This method improves the factual correctness and general quality of created content by fusing the benefits of corrective retrieval and self-reflection, which makes LLMs more dependable and adaptable for a variety of applications.

2. Problem statement

Despite significant advancements in large language models (LLMs), challenges regarding factual accuracy and coherent content generation remain. LLMs often suffer from "hallucination," producing outputs that, while plausible, may be factually incorrect. This is particularly concerning as LLMs are increasingly used in critical fields like education and healthcare, where accuracy is vital. Below is the RAG Diagram which shows how it works.



Retrieval-Augmented Generation (RAG) has emerged as a solution, enhancing LLMs by integrating external knowledge sources to ground their outputs in factual information. However, current RAG methods face limitations, such as indiscriminately retrieving fixed numbers of documents without assessing their relevance, leading to low-quality outputs. Furthermore, these models often fail to effectively utilize retrieved passages, lacking mechanisms to evaluate the quality of generated content against the retrieved information.

To address these challenges, we propose Self Corrective Retrieval Augmented Generation (SCRAG), a novel framework that combines adaptive retrieval with self-reflection. SCRAG allows the model to critically assess its outputs and the information retrieved using reflection tokens, enhancing the relevance and reliability of generated content. By implementing a lightweight retrieval evaluator, SCRAG aims to ensure that LLM outputs are both accurate and contextually appropriate, bridging the gap in current generative models.

3. Content

3.1 METHODOLOGY

3.1.1 Research Design:

The design of this study adopts a mixed-method approach, combining quantitative evaluations with qualitative analysis to understand the effectiveness of the Self Corrective Retrieval-Augmented Generation (SCRAG) framework. This approach explores how SCRAG improves the factual accuracy and coherence of language models by integrating adaptive retrieval and self-reflection mechanisms. The study is both descriptive and exploratory, aiming to assess SCRAG potential for improving generation quality in various application domains.

3.1.2 Data Collection Methods:

- Literature Review: A thorough review of existing literature on retrieval-augmented generation (RAG) systems, corrective retrieval strategies, and large language models (LLMs) will be conducted. This will identify critical insights, current challenges, and gaps in conventional RAG methods.
- Simulation Studies: Multiple small simulations will be conducted on datasets to assess the performance of SCRAG in generating accurate and contextually appropriate content. These datasets will include both short-form (e.g., question-answering) and long-form (e.g., document generation) tasks.
- Performance Evaluation: SCRAG will be implemented and tested against other RAG based models, using a variety of evaluation metrics such as factual accuracy, retrieval relevance, and content coherence.

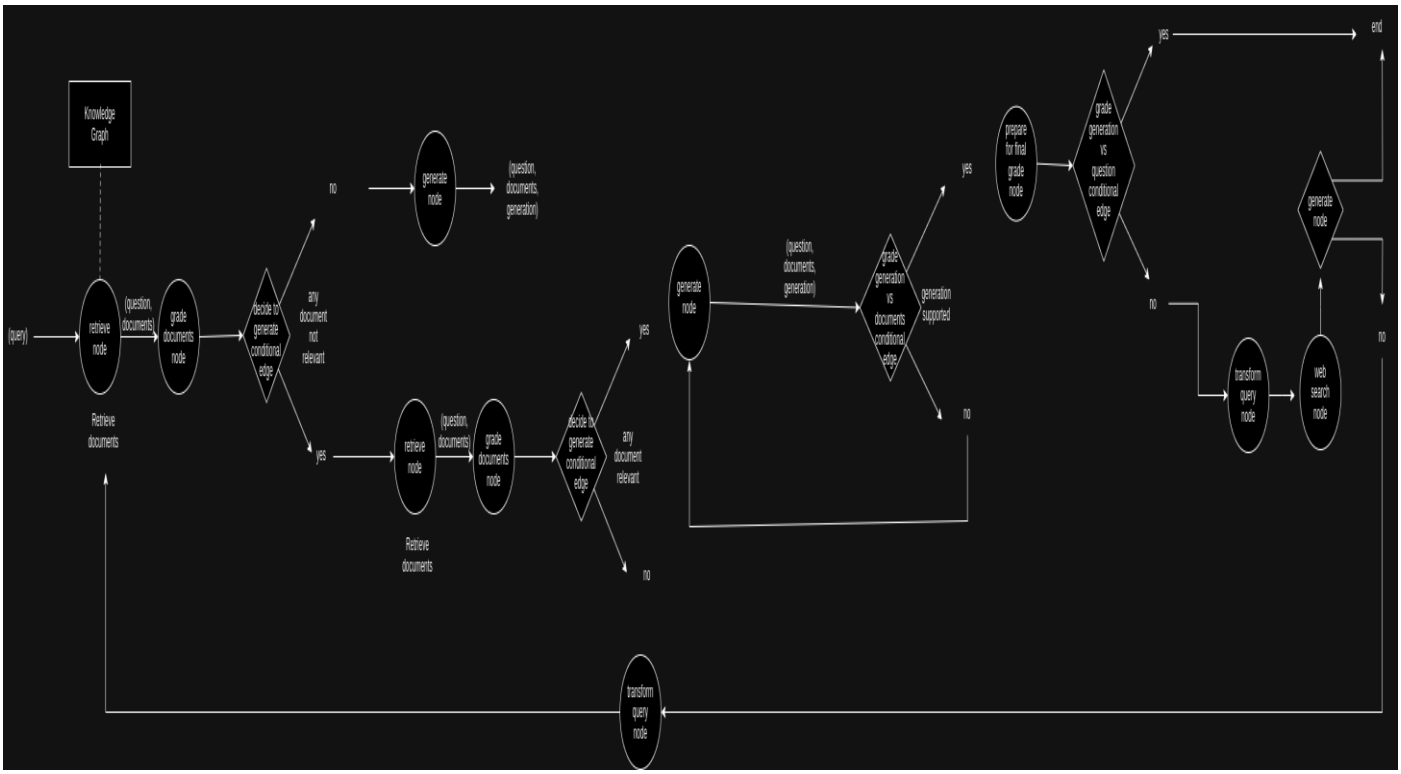
3.1.3 Data Analysis Techniques:

- Comparative Analysis: The performance of SCRAG will be compared against other RAG approaches like traditional RAG and Self-Reflective RAG (SRAG) to quantify improvements in factual accuracy, retrieval precision, and generation fluency.

- Error Analysis: A detailed error analysis will be conducted to assess cases where retrieval or content generation fails. This will help refine the self-correction mechanisms in SCRAG.

3.2 PROPOSED SYSTEM

- ***Adaptive Retrieval Mechanism:*** SC-RAG employs an adaptive retrieval system that dynamically evaluates when and how external knowledge should be retrieved. The model generates reflection tokens, allowing it to decide whether retrieved information is required, optimizing retrieval frequency based on task complexity and context.
- ***Self-Correction Component:*** A core feature of SCRAG is its self-correction capability. After generating content, the model generates critique tokens that assess the quality of its output and the relevance of retrieved passages. This self-assessment mechanism helps refine the generated content, reducing factual errors and improving coherence.
- ***Reflection Tokens Integration:*** SCRAG introduces special reflection tokens that allow the model to reflect on its output throughout the generation process. These tokens guide the model in deciding whether additional retrieval is necessary or whether the current generation aligns with the retrieved facts.
- ***Retrieval Evaluator:*** SCRAG includes a lightweight retrieval evaluator to assess the relevance and quality of the retrieved documents. This evaluator enables the model to categorize retrieved content as Correct, Incorrect, or Ambiguous, triggering appropriate retrieval actions when needed.
- ***Performance Optimizations:*** The system employs a customizable decoding algorithm, allowing users to adjust the retrieval frequency and tailor the model's behavior to specific tasks or preferences. This flexibility ensures that SCRAG can adapt to both short- and long-form generation tasks efficiently.



The diagram above illustrates how the Self Corrective Retrieval-Augmented Generation (SCRAG) system operates. SCRAG can work with datasets in various forms, such as knowledge graphs or large collections of text. If the model is unable to generate a suitable response based on the initial query, it initiates a web search to retrieve additional information for better accuracy. Throughout the process, SCRAG generates multiple responses, assigning rewards to the best-performing outputs. If all relevant possibilities fail to produce a satisfactory answer, the system rewrites the query and retries the generation process.

4. Conclusion

In this work, I present Self-Corrective Retrieval-Augmented Generation (SCRAG), a comprehensive framework designed to enhance the quality and factual accuracy of large language models (LLM) through on-demand retrieval and self-reflection. SCRAG trains the model to not only retrieve and generate text but also to critique its own outputs and the retrieved passages using a unified approach that incorporates both its original vocabulary and newly introduced special tokens, known as reflection tokens. This capability allows SCRAG to adaptively tailor the model's behavior during inference, optimizing its performance based on specific task requirements.

Additionally, SCRAG addresses the challenges faced by traditional RAG methods when retrieval fails, which can lead to the incorporation of inaccurate or misleading information.

By implementing a lightweight retrieval evaluator, SCRAG effectively estimates the relevance and quality of retrieved documents, triggering appropriate knowledge retrieval actions such as own Dataset or additional web searches. This dual mechanism of self-reflection and corrective retrieval significantly improves the model's robustness and efficiency in utilizing retrieved documents.

References

[1] (2020) Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In International Conference on Machine Learning.

[2] (2022) Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. Transactions on Machine Learning Research.

[3] (2020) Tom B Brown, Benjamin Mann, Nick Ryder. Language models are few-shot learners.

[4] (2022) Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. A survey on retrieval-augmented text generation.

[5] (2023) Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil,Hannaneb Hajishirzi. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection

[6] (2024) Shi-Qi Yan, Jia-Chen Gu, Yun Zhu,Zhen-Hua Ling. Corrective Retrieval Augmented Generation