Evaluation of Retrieval-Augmented Generation: A Survey

Hao Yu[1,2], Aoran Gan[3], Kai Zhang[3], Shiwei Tong[1†], Qi Liu[3], and Zhaofeng Liu[1]

[1] Tencent Company
[2] McGill University
[3] State Key Laboratory of Cognitive Intelligence,
University of Science and Technology of China
hao.yu2@mail.mcgill.ca
gar@mail.ustc.edu.cn
{shiweitong†,zhaofengliu}@tencent.com
{kkzhang08,qiliuql}@ustc.edu.cn

Abstract. Retrieval-Augmented Generation (RAG) has recently gained traction in natural language processing. Numerous studies and real-world applications are leveraging its ability to enhance generative models through external information retrieval. Evaluating these RAG systems, however, poses unique challenges due to their hybrid structure and reliance on dynamic knowledge sources. To better understand these challenges, we conduct *A Unified Evaluation Process of RAG* (*Auepora*) and aim to provide a comprehensive overview of the evaluation and benchmarks of RAG systems. Specifically, we examine and compare several quantifiable metrics of the Retrieval and Generation components, such as relevance, accuracy, and faithfulness, within the current RAG benchmarks, encompassing the possible output and ground truth pairs. We then analyze the various datasets and metrics, discuss the limitations of current benchmarks, and suggest potential directions to advance the field of RAG benchmarks.

1 Introduction

Retrieval-Augmented Generation (RAG) [34] efficiently enhances the performance of generative language models through integrating information retrieval techniques. It addresses a critical challenge faced by standalone generative language models: the tendency to produce responses that, while plausible, may not be grounded in facts. By retrieving relevant information from external sources, RAG significantly reduces the incidence of hallucinations [23] or factually incorrect outputs, thereby improving the content's reliability and richness. [73] This fusion of retrieval and generation capabilities enables the creation of responses that are not only contextually appropriate but also informed by the most current and accurate information available, making RAG a development in the pursuit of more intelligent and versatile language models [73,64].

† Corresponding Author
Paper Homepage: https://github.com/YHPeter/Awesome-RAG-Evaluation

Numerous studies of RAG systems have emerged from various perspectives since the advent of Large Language Models (LLMs) [55,45,59,42,41,69,16]. The RAG system comprises two primary components: ***Retrieval*** and ***Generation***. The retrieval component aims to extract relevant information from various external knowledge sources. It involves two main phases, *indexing* and *searching*. Indexing organizes documents to facilitate efficient retrieval, using either inverted indexes for sparse retrieval or dense vector encoding for dense retrieval [16,12,28]. The searching component utilizes these indexes to fetch relevant documents on the user's query, often incorporating the optional rerankers [4,39,6,52] to refine the ranking of the retrieved documents. The generation component utilizes the retrieved content and question query to formulate coherent and contextually relevant responses with the prompting and inferencing phases. As the "Emerging" ability [59] of LLMs and the breakthrough in aligning human commands [42], LLMs are the best performance choices model for the generation stage. Prompting methods like Chain of Thought (CoT) [60], Tree of Thgouht [65], Rephrase and Respond (RaR) [8] guide better generation results. In the inferencing step, LLMs interpret the prompted input to generate accurate and in-depth responses that align with the query's intent and integrate the extracted information [35,9] without further finetuning, such as fully finetuning [16,1,67,68] or LoRA [21]. Appendix A details the complete RAG structure. Figure 1 illustrates the structure of the RAG systems as mentioned.

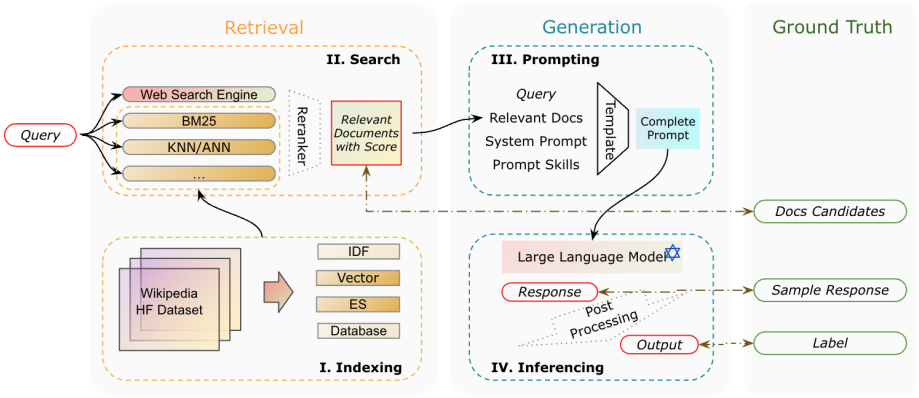

Fig. 1: The structure of the RAG system with retrieval and generation components and corresponding four phrases: indexing, search, prompting and inferencing. The pairs of "Evaluable Outputs" (EOs) and "Ground Truths" (GTs) are highlighted in read frame and green frame, with brown dashed arrows.

The importance of evaluating RAG is increasing in parallel with the advancement of RAG-specific methodologies. On the one hand, RAG is a complex system intricately tied to specific requirements and language models, resulting in various evaluation methods, indicators, and tools, particularly given the black-box LLM generation. Evaluating RAG systems involves specific components and the complexity of the overall system assessment. On the other hand, the complexity of RAG systems is further compounded