Enhancing Retrieval-Augmented Generation: A Study of Best Practices

Siran Li **Linus Stenzel** **Carsten Eickhoff** **Seyed Ali Bahrainian**
University of Tübingen
siran.li@uni-tuebingen.de, stenzel@student.uni-tuebingen.de,
{carsten.eickhoff, seyed.ali.bahreinian}@uni-tuebingen.de

Abstract

Retrieval-Augmented Generation (RAG) systems have recently shown remarkable advancements by integrating retrieval mechanisms into language models, enhancing their ability to produce more accurate and contextually relevant responses. However, the influence of various components and configurations within RAG systems remains underexplored. A comprehensive understanding of these elements is essential for tailoring RAG systems to complex retrieval tasks and ensuring optimal performance across diverse applications. In this paper, we develop several advanced RAG system designs that incorporate query expansion, various novel retrieval strategies, and a novel Contrastive In-Context Learning RAG. Our study systematically investigates key factors, including language model size, prompt design, document chunk size, knowledge base size, retrieval stride, query expansion techniques, Contrastive In-Context Learning knowledge bases, multilingual knowledge bases, and Focus Mode retrieving relevant context at sentence-level. Through extensive experimentation, we provide a detailed analysis of how these factors influence response quality. Our findings offer actionable insights for developing RAG systems, striking a balance between contextual richness and retrieval-generation efficiency, thereby paving the way for more adaptable and high-performing RAG frameworks in diverse real-world scenarios. Our code and implementation details are publicly available [1].

1 Introduction

Language Models (LMs) such as GPT, BERT, and T5 have demonstrated remarkable versatility, excelling in a wide range of NLP tasks, including summarization (Bahrainian et al., 2022), extracting relevant information from lengthy documents, question-answering, and storytelling (Brown et al., 2020b; Devlin et al., 2019; Raffel et al., 2020). However, their static knowledge and opaque reasoning raise concerns about maintaining factual accuracy and reliability as language and knowledge evolve (Huang et al., 2024; Jin et al., 2024). As new events emerge, and scientific advancements are made, it becomes crucial to keep models aligned with current information (Shi et al., 2024a). However, continuously updating models is both costly and inefficient. To address this, RAG models have been proposed as a more efficient alternative, integrating external knowledge sources during inference to provide up-to-date and accurate information (Lewis et al., 2020; Borgeaud et al., 2022; Lee et al., 2024). RAG models augment language models by incorporating verifiable information, improving factual accuracy in their responses (Gao et al., 2023; Kim et al., 2023). This approach not only mitigates some conceptual limitations of traditional LMs but also unlocks practical, real-world applications. By integrating a domain-specific knowledge base, RAG models transform LMs into specialized experts, enabling the development of highly targeted applications and shifting them from generalists to informed specialists (Siriwardhana et al., 2023). In recent years, this advancement has led to many proposed architectures and settings for an optimal RAG model (Li et al., 2024; Dong et al., 2024). However, the best practices for designing RAG models are still not well understood.

In this paper, we comprehensively examine the efficacy of RAG in enhancing Large LM (LLM) responses, addressing nine key research questions: (1) How does the size of the LLM affect the response quality in an RAG system? (2) Can subtle differences in prompt significantly affect the alignment of retrieval and generation? (3) How does the retrieved document chunk size impact the response quality? (4) How does the size of the knowledge base impact the overall performance? (5) In the retrieval strides (Ram et al., 2023), how

[1] https://github.com/ali-bahrainian/RAG_best_practices

often should context documents be updated to optimize accuracy? (6) Does expanding the query improve the model's precision? (7) How does including Contrastive In-context Learning demonstration examples influence RAG generation? (8) Does incorporating multilingual documents affect the RAG system's responses? (9) Does focusing on a few retrieved sentences sharpen RAG's responses? To address these questions, we employ ablation studies as the primary method, allowing for a detailed empirical investigation of RAG's operational mechanisms. A custom evaluation framework is developed to assess the impact of various RAG components and configurations individually. The insights gained will contribute to advancing LLM performance and inform future theoretical developments.

The Main Contributions of this paper are: (1) We conduct an extensive benchmark to help explain the best practices in RAG setups. **(2)** While the first five research questions above are based on previous literature, the methods that address the last four research questions, namely, Query Expansion, Contrastive In-context Learning demonstration, multilingual knowledge base, and Focus Mode RAG are novel contributions of this study which we believe will advance the field.

The remainder of this paper is organized as follows: Section 2 provides an overview of important related work. Section 3 presents novel methods that improve RAG responses and outlines the methodology. Section 4 presents two evaluation datasets, knowledge base, and evaluation metrics and explains the implementation details. Section 5 discusses the extensive results of our carefully designed benchmark comparison and Section 6 highlights the key findings of this study. Section 7 concludes this paper and suggests avenues for future research. Finally, Section 8 discusses the limitations of our study.

2 Related Works

RAG systems have emerged as a promising solution to the inherent limitations of LLMs, particularly their tendency to hallucinate or generate inaccurate information (Semnani et al., 2023; Chang et al., 2024). By integrating retrieval mechanisms, RAG systems fetch relevant external knowledge during the generation process, ensuring that the model's output is informed by up-to-date and contextually relevant information (Gao et al., 2023; Tran and Litman, 2024). Guu et al. (2020) show that language models could retrieve relevant documents in real time and use them to inform text generation, significantly enhancing factual accuracy without increasing model size. Shi et al. (2024b) demonstrate how retrieval modules can be applied even to black-box models without direct access to their internals. In-Context Retrieval-Augmented Language Models further dynamically incorporate retrievals into the generation process, allowing for more flexible and adaptive responses (Ram et al., 2023). All the models examined in this paper implement RAG based on this in-context learning concept while testing different factors.

Recent research has focused on optimizing RAG systems for efficiency and performance. Several strategies for improving the system's retrieval components are outlined, such as optimizing document indexing and retrieval algorithms to minimize latency without compromising accuracy (Wang et al., 2024). Additionally, Hsia et al. (2024) examine the architectural decisions that can enhance the efficacy of RAG systems, including corpus selection, retrieval depth, and response time optimization. Furthermore, Wu et al. (2024) illustrate how optimization strategies can be designed to balance the model's internal knowledge with the retrieved external data, addressing the potential conflict between these two sources of information. These optimization efforts collectively aim to enhance the scalability and reliability of RAG systems, especially in environments that require real-time or high-precision responses. Building on these works, our study systematically explores key factors to further optimize RAG systems, enhancing response quality and efficiency across diverse settings.

3 Methods

Augmenting LLMs with real-time, up-to-date external knowledge bases, allows the resulting RAG system to generate more accurate, relevant, and timely responses without the need for constant retraining (Fan et al., 2024). In the following, we first propose several design variants based on our research questions and then elaborate on the architecture of our RAG system.

3.1 RAG Design Variations

To explore the strategy that influences the efficacy of RAG, we propose the following research questions to guide our investigation: