Enhancing Retrieval-Augmented Large Language Models with Iterative Retrieval-Generation Synergy

Zhihong Shao[1], **Yeyun Gong**[2], **yelong shen**[3], **Minlie Huang**[1]*, **Nan Duan**[2], **Weizhu Chen**[3]
[1] The CoAI Group, DCST, Institute for Artificial Intelligence,
[1] State Key Lab of Intelligent Technology and Systems,
[1] Beijing National Research Center for Information Science and Technology,
[1] Tsinghua University, Beijing 100084, China
[2] Microsoft Research Asia [3] Microsoft Azure AI
szh19@mails.tsinghua.edu.cn aihuang@tsinghua.edu.cn

Abstract

Retrieval-augmented generation has raise extensive attention as it is promising to address the limitations of large language models including outdated knowledge and hallucinations. However, retrievers struggle to capture relevance, especially for queries with complex information needs. Recent work has proposed to improve relevance modeling by having large language models actively involved in retrieval, i.e., to guide retrieval with generation. In this paper, we show that strong performance can be achieved by a method we call ITER-RETGEN, which synergizes retrieval and generation in an iterative manner: a model's response to a task input shows what might be needed to finish the task, and thus can serve as an informative context for retrieving more relevant knowledge which in turn helps generate a better response in another iteration. Compared with recent work which interleaves retrieval with generation when completing a single output, ITER-RETGEN processes all retrieved knowledge as a whole and largely preserves the flexibility in generation without structural constraints. We evaluate ITER-RETGEN on multi-hop question answering, fact verification, and commonsense reasoning, and show that it can flexibly leverage parametric knowledge and non-parametric knowledge, and is superior to or competitive with state-of-the-art retrieval-augmented baselines while causing fewer overheads of retrieval and generation. We can further improve performance via generation-augmented retrieval adaptation.

1 Introduction

Generative Large Language Models (LLMs) have powered numerous applications, with well-perceived utility. Despite being powerful, LLMs lack knowledge that is under-represented in their training data, and are prone to hallucinations, especially in open-domain settings (OpenAI, 2023). Retrieval-augmented LLMs, therefore, have raised widespread attention as LLM outputs can be potentially grounded on external knowledge.

Previous retrieval-augmented LMs (Izacard et al., 2022b; Shi et al., 2023) typically adopted one-time retrieval, i.e., to retrieve knowledge using only the task input (e.g., a user question for open-domain question answering). One-time retrieval should suffice to fulfill the information needs if they are clearly stated in the original input, which is applicable to factoid question answering (Kwiatkowski et al., 2019) and single-hop fact verification (Thorne et al., 2018), but not to tasks with complex information needs, e.g., multi-hop reasoning (Yang et al., 2018) and long-form question answering (Fan et al., 2019).

To fulfill complex information needs, recent work proposes to gather required knowledge multiple times throughout the generation process, using partial generation (Trivedi et al., 2022a; Press et al., 2022)) or forward-looking sentence(s) (Jiang et al., 2023) as search queries. However, such structured workflows of interleaving retrieval with generation have the following limitations: (1) as intermediate generation is conditioned on knowledge retrieved before, with no awareness of knowledge retrieved afterwards, they fail to process all retrieved knowledge as a whole during the generation process; (2) they require multi-round retrieval to gather a comprehensive set of knowledge, and may frequently change the prompts by updating newly retrieved knowledge, thus increasing the overheads of both retrieval and generation.

In this paper, we find it simple but effective to enhance retrieval-augmented LLMs through iterative retrieval-generation synergy (ITER-RETGEN, Fig 1). ITER-RETGEN iterates *retrieval-augmented generation* and *generation-augmented retrieval*: Retrieval-augmented generation outputs a response to a task input based on all retrieved knowledge (initially using the task input as the query). This

*Corresponding author: Minlie Huang.