SELF-RAG: LEARNING TO RETRIEVE, GENERATE, AND CRITIQUE THROUGH SELF-REFLECTION

Akari Asai[†], **Zeqiu Wu**[†], **Yizhong Wang**[†§], **Avirup Sil**[‡], **Hannaneh Hajishirzi**[†§]
[†]University of Washington [§]Allen Institute for AI [‡]IBM Research AI
{akari,zeqiuwu1,yizhongw,hannaneh}@cs.washington.edu, avi@us.ibm.com

ABSTRACT

Despite their remarkable capabilities, large language models (LLMs) often produce responses containing factual inaccuracies due to their sole reliance on the parametric knowledge they encapsulate. Retrieval-Augmented Generation (RAG), an ad hoc approach that augments LMs with retrieval of relevant knowledge, decreases such issues. However, indiscriminately retrieving and incorporating a fixed number of retrieved passages, regardless of whether retrieval is necessary, or passages are relevant, diminishes LM versatility or can lead to unhelpful response generation. We introduce a new framework called **Self-Reflective Retrieval-Augmented Generation (SELF-RAG)** that enhances an LM's quality and factuality through retrieval and self-reflection. Our framework trains a single arbitrary LM that adaptively retrieves passages on-demand, and generates and reflects on retrieved passages and its own generations using special tokens, called *reflection* tokens. Generating reflection tokens makes the LM controllable during the inference phase, enabling it to tailor its behavior to diverse task requirements. Experiments show that SELF-RAG (7B and 13B parameters) significantly outperforms state-of-the-art LLMs and retrieval-augmented models on a diverse set of tasks. Specifically, SELF-RAG outperforms ChatGPT and retrieval-augmented Llama2-chat on Open-domain QA, reasoning and fact verification tasks, and it shows significant gains in improving factuality and citation accuracy for long-form generations relative to these models.[1]

1 INTRODUCTION

State-of-the-art LLMs continue to struggle with factual errors (Mallen et al., 2023; Min et al., 2023) despite their increased model and data scale (Ouyang et al., 2022). Retrieval-Augmented Generation (RAG) methods (Figure 1 left; Lewis et al. 2020; Guu et al. 2020) augment the input of LLMs with relevant retrieved passages, reducing factual errors in knowledge-intensive tasks (Ram et al., 2023; Asai et al., 2023a). However, these methods may hinder the versatility of LLMs or introduce unnecessary or off-topic passages that lead to low-quality generations (Shi et al., 2023) since they retrieve passages indiscriminately regardless of whether the factual grounding is helpful. Moreover, the output is not guaranteed to be consistent with retrieved relevant passages (Gao et al., 2023) since the models are not explicitly trained to leverage and follow facts from provided passages.

This work introduces **Self-Reflective Retrieval-augmented Generation (SELF-RAG)** to improve an LLM's generation quality, including its factual accuracy without hurting its versatility, via on-demand retrieval and self-reflection. We train an arbitrary LM in an end-to-end manner to learn to reflect on its own generation process given a task input by generating both task output and intermittent special tokens (i.e., *reflection tokens*). Reflection tokens are categorized into *retrieval* and *critique* tokens to indicate the need for retrieval and its generation quality respectively (Figure 1 right). In particular, given an input prompt and preceding generations, SELF-RAG first determines if augmenting the continued generation with retrieved passages would be helpful. If so, it outputs a **retrieval** token that calls a retriever model on demand (Step 1). Subsequently, SELF-RAG concurrently processes multiple retrieved passages, evaluating their relevance and then **generating** corresponding task outputs (Step 2). It then generates critique tokens to **criticize** its own output and choose best one (Step 3) in terms of factuality and overall quality. This process differs from conventional RAG (Figure 1 left), which

[1]Our code and trained models are available at https://selfrag.github.io/.

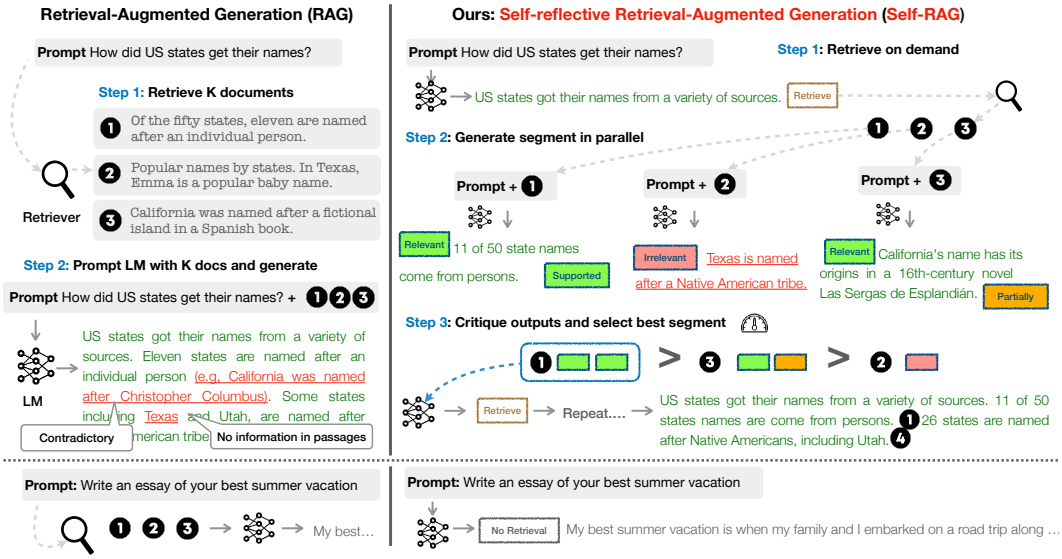

Figure 1: Overview of SELF-RAG. SELF-RAG learns to retrieve, critique, and generate text passages to enhance overall generation quality, factuality, and verifiability.

consistently retrieves a fixed number of documents for generation regardless of the retrieval necessity (e.g., the bottom figure example does not require factual knowledge) and never second visits the generation quality. Moreover, SELF-RAG provides citations for each segment with its self-assessment of whether the output is supported by the passage, leading to easier fact verification.

SELF-RAG trains an arbitrary LM to generate text with reflection tokens by unifying them as the next token prediction from the expanded model vocabulary. We train our generator LM on a diverse collection of text interleaved with reflection tokens and retrieved passages. Reflection tokens, inspired by reward models used in reinforcement learning (Ziegler et al., 2019; Ouyang et al., 2022), are inserted offline into the original corpus by a trained *critic* model. This eliminates the need to host a critic model during training, reducing overhead. The critic model, in part, is supervised on a dataset of input, output, and corresponding reflection tokens collected by prompting a propriety LM (i.e., GPT-4; OpenAI 2023). While we draw inspiration from studies that use control tokens to start and guide text generation (Lu et al., 2022; Keskar et al., 2019), our trained LM uses critique tokens to assess its own predictions after each generated segment as an integral part of the generation output.

SELF-RAG further enables a customizable decoding algorithm to satisfy hard or soft constraints, which are defined by reflection token predictions. In particular, our inference-time algorithm enables us to (1) flexibly adjust retrieval frequency for different downstream applications and (2) customize models' behaviors to user preferences by leveraging reflection tokens through segment-level beam search using the weighted linear sum of the reflection token probabilities as segment score.

Empirical results on six tasks, including reasoning and long-form generation, demonstrate that SELF-RAG significantly outperforms pre-trained and instruction-tuned LLMs that have more parameters and widely adopted RAG approaches with higher citation accuracy. In particular, SELF-RAG outperforms retrieval-augmented ChatGPT on four tasks, Llama2-chat (Touvron et al., 2023) and Alpaca (Dubois et al., 2023) on all tasks. Our analysis demonstrates the effectiveness of training and inference with reflection tokens for overall performance improvements as well as test-time model customizations (e.g., balancing the trade-off between citation previsions and completeness).

2 RELATED WORK

Retrieval-Augmented Generation. Retrieval-Augmented Generation (RAG) augments the input space of LMs with retrieved text passages (Guu et al., 2020; Lewis et al., 2020), leading to large improvements in knowledge-intensive tasks after fine-tuning or used with off-the-shelf LMs (Ram et al., 2023). A more recent work (Luo et al., 2023) instruction-tunes an LM with a fixed number