Parallel Key-Value Cache Fusion for Position Invariant RAG

Philhoon Oh **Jinwoo Shin** **James Thorne**

KAIST AI

{philhoonoh,jinwoos,thorne}@kaist.ac.kr

Abstract

Recent advancements in Large Language Models (LLMs) underscore the necessity of Retrieval Augmented Generation (RAG) to leverage external information. However, LLMs are sensitive to the position of relevant information within contexts and tend to generate incorrect responses when such information is placed in the middle, known as 'Lost in the Middle' phenomenon. In this paper, we introduce a framework that generates consistent outputs for decoder-only models, irrespective of the input context order. Experimental results for three open domain question answering tasks demonstrate position invariance, where the model is not sensitive to input context order, and superior robustness to irrelevent passages compared to prevailing approaches for RAG pipelines.

1 Introduction

In Retrieval Augmented Generation (RAG) (Guu et al., 2020; Lewis et al., 2021; Izacard et al., 2022), models first extract relevant information from a knowledge base and then incorporate this extracted information with its parameteric knowledge to generate the response. This two-step approach is the de-facto approach for knowledge-intensive tasks (Lewis et al., 2021; Petroni et al., 2021).

However, decoder-only models exhibit an intrinsic positional bias, assigning more attention to tokens at the beginning or end of the input sequence while often overlooking relevant context located in the middle, a problem known as the 'Lost in the Middle' (Liu et al., 2023). Previous works to address this issue involves training with specific prompt (He et al., 2024) or data-intensive training (An et al., 2024). Other works aimed at modifying positional embeddings (Hsieh et al., 2024b) or reducing positional attention bias in LLMs (Yu et al., 2024a). Yet, none of these methods fully guarantee a solution to this intrinsic bias in LLMs for RAG.

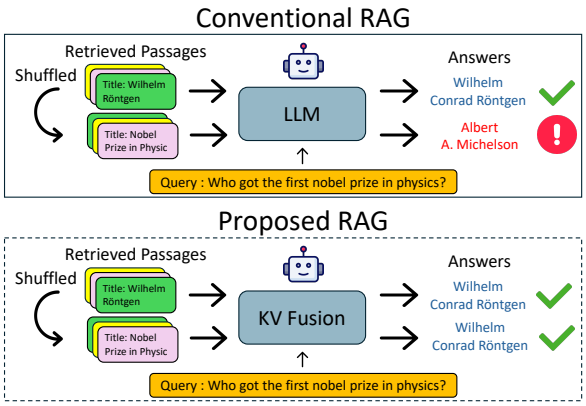

Figure 1: Illustration of the KV-Fusion model: Generated tokens remain consistent even when the retrieved passages are shuffled.

In this paper, we introduce a framework for decoder-only models, called **K**ey **V**alue **Fusion** (**KV Fusion**), to generate consistent outcomes regardless of input order as illustrated in Figure 1. **KV Fusion** consists of two components: a *prefill* decoder that extract key-values caches in parallel and a *trainable* decoder that utilizes extracted key value caches to produce consistent outcome. This architecture injects uniform positional information to each input contexts, ensuring consistent output generation even when the input order varies. Experiments on open domain question answering datasets, including NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and POPQA (Mallen et al., 2023), demonstrate KV-Fusion's position-invariant nature, achieving accuracy improvements of 21.4%, 6.4%, and 6.6% over baseline models in shuffled settings. Furthermore, KV-Fusion models exhibit robust and stable accuracies even with additional contexts compared to other approaches.

2 Method

Notation Our KV-Fusion architecture is illustrated in Figure 2. For clarity, we refer to this prefill decoder as $\mathcal{D}_{\text{p}}$, which is characterized by the