Query-Focused Retrieval Heads Improve Long-Context Reasoning and Re-ranking

Wuwei Zhang[♠] **Fangcong Yin**[♢] **Howard Yen**[♠] **Danqi Chen**[♠] **Xi Ye**[♠]
[♠] Princeton Language and Intelligence, Princton University
[♢] The University of Texas at Austin
[♠] {wuwei.zhang,hyen,danqic}@cs.princeton.edu xi.ye@princeton.edu
[♢] fangcongyin@utexas.edu

Abstract

Recent work has identified retrieval heads (Wu et al., 2025), a subset of attention heads responsible for retrieving salient information in long-context language models (LMs), as measured by their copy-paste behavior in Needle-in-a-Haystack tasks. In this paper, we introduce QRHEAD (Query-Focused Retrieval Head), an improved set of attention heads that enhance retrieval from long context. We identify QRHEAD by aggregating attention scores *with respect to the input query*, using a handful of examples from real-world tasks (e.g., long-context QA). We further introduce QRRETRIEVER, an efficient and effective retriever that uses the accumulated attention mass of QRHEAD as retrieval scores. We use QRRETRIEVER for long-context reasoning by selecting the most relevant parts with the highest retrieval scores. On multi-hop reasoning tasks LongMemEval and CLIPPER, this yields over 10% performance gains over full context and outperforms strong dense retrievers. We also evaluate QRRETRIEVER as a re-ranker on the BEIR benchmark and find that it achieves strong zero-shot performance, outperforming other LLM-based re-rankers such as RankGPT. Further analysis shows that both the query-context attention scoring and task selection are crucial for identifying QRHEAD with strong downstream utility. Overall, our work contributes a general-purpose retriever and offers interpretability insights into the long-context capabilities of LMs.