A Competitive LLM-based Hybrid RAG System with Optimized Integration of Dense, Sparse, and Cached Retrieval on Edge Devices

Jhing-Fa Wang *
Department of Electrical Engineering, National Cheng-Kung University, Tainan, Taiwan

Din-Yuen Chan *
Department of Computer Science Information Engineering, National Chiayi University, Chiayi, Taiwan

Kuo-Sheng Hu
Department of Electrical Engineering, National Cheng-Kung University, Tainan, Taiwan

Abstract

In this paper, an innovative hybrid RAG system is constructed by optimally integrating dense retrieval, sparse keyword-based retrieval, and a retrieval cache mechanism. The proposed system can harness the main challenge for deploying LLMs on edge devices. The challenge includes the limited memory, constrained compute resources, and high latency. The edge-device implementation of common RAG systems often cause lower 70% retrieval accuracy and more 5 seconds end-to-end response latency for standard question-answering benchmarks. The proposed system can effectively reduce the redundant computation and the inference latency as well as improve the retrieval precision. Experiments demonstrate that our system termed CLH-RAG can achieve the retrieval accuracy of over 80% and the average response latency of less than 2 seconds. Consequently, CLH-RAG have the high competition against the existing RAG systems for the edge device deployment supplying the high-quality real-time LLM inference.

Keywords

Edge AI, Hybrid Retrieval, Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Retrieval Cache.