PROJECT REPORT

A HYBRID ADAPTIVE AND SEMANTIC CACHING MECHANISM FOR EFFICIENT RETRIEVAL-AUGMENTED GENERATION IN LLMs

VALENTINO RYO KOESDARTO

22.K3.0002

Faculty of Computer Science
Soegijapranata Catholic University
2026

ABSTRACT

Large Language Models (LLMs) have the ability to understand and generate natural language, but they still face limitations in providing accurate answers for knowledge beyond the training data, which triggers hallucination issues. Retrieval-Augmented Generation (RAG) presents itself as a solution by adding an external context retrieval stage to enhance answer relevance, but it causes significant overhead due to the embedding process, document search, and context integration, which slow down response time and increase computational costs. This research develops a hybrid caching system that combines an Adaptive LRFU caching mechanism that dynamically adjusts cache priorities based on access recency, frequency patterns, and cache hit/miss behavior, with semantic caching based on embedding to detect meaning similarity between queries. The system is designed using a dataset of 10,000 question-answer pairs in ChromaDB with the BAAI/bge-large-en-v1.5 embedding model, as well as FAISS and Chroma-based semantic matching techniques. Evaluation was conducted using five main metrics: Cache Hit Ratio (CHR), Average Response Time (ART), Answer Accuracy (AA), Memory Utilization Efficiency (MUE), and Cache Update Overhead (CUO). The result is expected that the hybrid approach can outperform static baselines and single caching methods by delivering faster responses, reducing redundant computations, and maintaining answer quality.

Keyword: Large Language Models, Retrieval-Augmented Generation, Caching Mechanisms, LRFU, Semantic Similarity