Dynamic RAG Knowledge Base Evolution Latency and Inference Throughput in Code Generation

Assignee Research

May 29, 2026

Abstract

Large Language Models (LLMs) have showcased impressive reasoning abilities, but often suffer from hallucinations or outdated knowledge. Knowledge Graph (KG)-based Retrieval-Augmented Generation (RAG) remedies these shortcomings by grounding LLM responses in structured external information from a knowledge base. However, many KG-based RAG approaches struggle with (i) aligning KG and textual representations, (ii) balancing retrieval accuracy and efficiency, and (iii) adapting to dynamically updated KGs. In this work, we introduce Walk&Retrieve, a simple yet effective KG-based framework that

1 Introduction

This paper examines: Walk&Retrieve: Simple Yet Effective Zero-shot Retrieval-Augmented Generation via Knowledge Graph Walks. Research question: How does the retrieval latency overhead of dynamic RAG knowledge base evolution affect inference throughput (tokens per second) for code generation tasks on the CodeXGLUE benchmark, relative to static retrieval baselines under controlled GPU memory constraints?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.5/10.

3 Results

14 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 6.5/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

References

- http://arxiv.org/abs/2411.19443v1
- http://arxiv.org/abs/2601.11863v1
- http://arxiv.org/abs/2505.16849v2