MA-DPR and BM25 Inference Latency in RAG-Based Code Generation at Scale

Assignee Research

June 2, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the inference latency of MA-DPR compare to BM25 in RAG-based code generation pipelines when processing the HumanEval dataset at scale. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Blended RAG: Improving RAG (Retriever-Augmented Generation) Accuracy with Semantic Search and Hybrid Query-Based Retrievers. Research question: How does the inference latency of MA-DPR compare to BM25 in RAG-based code generation pipelines when processing the HumanEval dataset at scale?.

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 3.2/10.

3 Results

14 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 3.2/10.

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/2605.18561v1
- http://arxiv.org/abs/2404.07220v2
- http://arxiv.org/abs/2402.12317v2