BM25 and Dense Retriever Hybridization in RAG Pipelines: Latency and Throughput Trade-offs

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: What is the effect of combining BM25 and dense retrievers on the inference latency and throughput of RAG pipelines in production environments. Retrieval-Augmented Generation (RAG) enhances Large Language Models (LLMs) by integrating external knowledge bases, achieving state-of-the-art results in various coding tasks. The core of RAG is retrieving demonstration examples, which is essential to balance effectiveness. 14 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.1/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Evaluating the Effectiveness and Efficiency of Demonstration Retrievers in RAG for Coding Tasks. Research question: What is the effect of combining BM25 and dense retrievers on the inference latency and throughput of RAG pipelines in production environments?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.1/10.

3 Results

16 papers retrieved. 14 claims extracted; 0 independently verified. Quality review score: 3.1/10.

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.

5 Extracted Claims

Claim	Verified	Confidence
In a preliminary evaluation of the Assertion Generation Task with 18,027 test samples and a knowledge base of 150,523 en	×	0.05
In a preliminary evaluation of the Assertion Generation Task with 18,027 test samples and a knowledge base of 150,523 en	×	0.06
In the preliminary evaluation of the Assertion Generation Task, BM25 achieved only a 1% increase in Exact Match rate com	×	0.08
The Atlas dataset contains 150,523 entries in its knowledge base and 18,027 test samples for the Assertion Generation ta	×	0.05
The NNGen dataset contains 22,112 entries in its knowledge base and 2,521 test samples for the Commit Generation task.	×	0.07
The CoNaLa dataset contains 2,300 entries in its knowledge base and 477 test samples for the Program Synthesis task.	×	0.05
For the Program Synthesis task, the ANNOY retriever parameter 'Search K' was evaluated using the values [1, 2, 5, 15, 30	×	0.05
For the Assertion Generation task, the HNSW retriever parameter 'M' was evaluated using the values [3, 4, 8, 16, 64].	×	0.06
In the Program Synthesis task, the BM25 retriever achieved a Code Bleu score of 0.249.	×	0.05
In the Program Synthesis task, the Exhaustive SBERT retriever achieved a Code Bleu score of 0.261.	×	0.06
In the Program Synthesis task, the HNSW retriever achieved a speedup of 0.43x relative to the baseline.	×	0.06
In the Program Synthesis task, the BM25L retriever resulted in a 5.75% reduction in Code Bleu score compared to the base	×	0.05
Sparse retrievers like BM25 operate at the token level, while dense retrievers operate at the level of latent semantics.	×	0.09
The Llamaindex RAG system supports both BM25 and custom embedding encoders.	×	0.03

References

- http://arxiv.org/abs/2410.09662v2
- http://arxiv.org/abs/2510.20609v1
- http://arxiv.org/abs/2512.24268v1