Hybrid Retrieval in RAG Systems: Latency and Accuracy Trade-offs on Multi-Track Music QA

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: How does the integration of hybrid retrieval methods (combining dense and sparse) in RAG systems impact inference latency and accuracy trade-offs on multi-track music QA benchmarks compared to. Retrieval-Augmented Generation (RAG) is a prevalent approach to infuse a private knowledge base of documents with Large Language Models (LLM) to build Generative Q\&A (Question-Answering) systems. However, RAG accuracy becomes increasingly challenging as the corpus of documents. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. 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 integration of hybrid retrieval methods (combining dense and sparse) in RAG systems impact inference latency and accuracy trade-offs on multi-track music QA benchmarks compared to single-track datasets?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

3 Results

11 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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/2402.12317v2
- http://arxiv.org/abs/2507.23334v2
- http://arxiv.org/abs/2404.07220v2