Retrieval Method Impact on RAG Accuracy-Throughput Trade-offs in Domain-Specific Benchmarks

Assignee Research

June 12, 2026

Abstract

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 scales up, with Retrievers playing an outsized role in the overall RAG accuracy by extracting the most relevant document from the corpus to provide context to the LLM. In this paper, we propose the 'Blended RAG' method of leveraging semantic search techniques, such as Dense Vector indexes and Sparse Encoder indexes, ble

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 accuracy-throughput trade-off in RAG systems vary when using different retrieval methods (e.g., dense vs. sparse) across model sizes (1B-7B) on domain-specific benchmarks like MedQA?.

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 7.1/10.

3 Results

11 papers retrieved. 17 claims extracted; 12 independently verified. Quality review score: 7.1/10.

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
On the NQDataset, the Blended RAG approach achieved a P@20 score of 0.633.	×	0.09
On the NQDataset, the Blended RAG approach achieved an NDCG@10 score of 0.67.	✓	0.19
On the NQDataset, the monoT5-3B model achieved an NDCG@10 score of 0.633.	✓	0.23
On the Trec Covid dataset, the Blended RAG approach achieved an NDCG@10 score of 0.87.	✓	0.25
On the Trec Covid dataset, the COCO-DR Large model achieved an NDCG@10 score of 0.804.	✓	0.29
On the HotpotQA dataset, the system achieved an F1 and EM score of 0.85.	×	0.07
On the SqUAD dataset, the KNN+BF configuration achieved a Top-5 retrieval accuracy of 94.89.	✓	0.16
On the SqUAD dataset, the KNN+BF configuration achieved a Top-10 retrieval accuracy of 97.43.	✓	0.16
On the SqUAD dataset, the KNN+BF configuration achieved a Top-20 retrieval accuracy of 98.58.	✓	0.15
On the COQA dataset, the Sparse Encoder + Best Fields (SE+BF) configuration achieved a Top-5 retrieval accuracy of 49.94	×	0.10
In RAG evaluation on SqUAD, the Blended RAG pipeline achieved an Exact Match (EM) score of 57.63.	✓	0.18
In RAG evaluation on SqUAD, the Blended RAG pipeline achieved an F1 score of 68.4.	×	0.14
In RAG evaluation on SqUAD, the RAG-original pipeline achieved an Exact Match (EM) score of 28.12.	✓	0.16
In RAG evaluation on SqUAD, the RAG-end2end pipeline achieved an Exact Match (EM) score of 40.02.	×	0.14
All RAG pipeline experiments in this study were performed using the FLAN-T5-XXL model.	✓	0.15
The study evaluated three primary indices: BM25 for keyword-based search, KNN for vector-based search, and Elastic Learn	✓	0.25
The Sparse Encoder-based semantic search combined with the 'Best Fields' hybrid query provides superior results compared	✓	0.20

References

- http://arxiv.org/abs/2507.23334v2
- http://arxiv.org/abs/2604.00715v1
- http://arxiv.org/abs/2404.07220v2