Semantic Search and Hybrid Retrieval Enhancements in RAG for Multi-Hop QA Reasoning

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: How does the combination of semantic search and hybrid query-based retrievers in RAG systems affect the reasoning capabilities of LLMs on complex multi-hop QA tasks, as measured by answer accuracy. 13 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: MedBioRAG: Semantic Search and Retrieval-Augmented Generation with Large Language Models for Medical and Biological QA. Research question: How does the combination of semantic search and hybrid query-based retrievers in RAG systems affect the reasoning capabilities of LLMs on complex multi-hop QA tasks, as measured by answer accuracy and logical consistency scores?.

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 4.2/10.

3 Results

16 papers retrieved. 13 claims extracted; 2 independently verified. Quality review score: 4.2/10.

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.

5 Extracted Claims

Claim	Verified	Confidence
BioRAG outperforms previous state-of-the-art models and the GPT-4o base model in all evaluated tasks.	✓	0.24
MedBioRAG improves NDCG and MRR scores for document retrieval, while achieving higher accuracy in close-ended QA and ROU	✓	0.43
LLMs have demonstrated strong capabilities in retrieving medical knowledge, structured reasoning, and evidence-based res	×	0.10
Fine-tuning has played a critical role in enhancing LLM performance for biomedical QA.	×	0.13
Domain-specific models have leveraged task-specific fine-tuning to improve accuracy and contextual understanding.	×	0.08
Some models integrate uncertainty-guided search strategies, allowing them to refine responses using external retrieval m	×	0.04
Others employ preference-based optimization frameworks, iteratively refining generated responses through synthetic prefe	×	0.05
These approaches have achieved state-of-the-art performance by leveraging retrieval-based adaptation and human-aligned e	×	0.04
RAG has emerged as a pivotal approach in biomedical QA.	×	0.09
MedBioRAG improves upon lexical search methods by leveraging semantic search for precise retrieval and fine-tuned LLMs f	×	0.14
MedBioRAG achieves 85.0 NDCG and 88.2 MRR on NFCorpus.	×	0.07
MedBioRAG achieves 88.2 NDCG and 64.3 MRR on TREC-COVID.	×	0.09
MedBioRAG achieves 61.9 NDCG and 88.2 MRR on PubMedQA.	×	0.07

References

- http://arxiv.org/abs/2512.10996v1
- http://arxiv.org/abs/2502.11228v2

- http://arxiv.org/abs/2508.05197v2