Multi-Stage Retrieval and Contrastive Learning in Dense Retrievers for Multi-Hop QA

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does the integration of multi-stage retrieval with contrastive learning in dense retrievers impact answer accuracy and latency in multi-hop QA benchmarks compared to single-stage retrieval. 16 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Retriever Portfolios: A Principled Approach to Adaptive RAG. Research question: How does the integration of multi-stage retrieval with contrastive learning in dense retrievers impact answer accuracy and latency in multi-hop QA benchmarks compared to single-stage retrieval?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

12 papers retrieved. 16 claims extracted; 1 independently verified. Quality review score: 4.5/10.

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
Retriever portfolios were evaluated on four QA benchmarks: HotpotQA, 2WikiMultiHopQA, TriviaQA, and MusiQue.	×	0.11
Two answer models were used for evaluation: Gemma-3-27B-It and Llama-3.1-70B-Instruct.	×	0.02
The evaluation addressed three questions: (1) do learned portfolios provide better retrieval coverage as the portfolio s	×	0.11
The best-of-k retrieval score was used to evaluate a size-k portfolio by taking the maximum support-document score achie	×	0.05
The portfolio was trained once on the pooled training queries from all four benchmarks and then evaluated on the corresp	×	0.03
The portfolio selection is not equivalent to picking the best retrievers on average.	×	0.04
At k = 5, the greedy portfolio selection reached 0.594 support recall and 0.500 support F1, while the top-k average base	×	0.05
The learned portfolio includes lower-average but complementary Vendi and GraphDense variants that cover queries missed b	×	0.05
The top-k average list is dominated by closely related GraphDense/E5 configurations, so additional members add little ne	×	0.02
The evaluation was conducted on diverse open-domain and multi-hop QA benchmarks: HotpotQA, 2WikiMultihopQA, TriviaQA, an	×	0.10
The method consistently yields better retrieval recall and answer accuracy compared to single-retriever baselines and in	✓	0.18
The method significantly reduces latency and token usage.	×	0.06
Retrieval-augmented generation (RAG) has become a standard approach for grounding large language models (LLMs) in extern	×	0.12
RAG improves factual accuracy and knowledge coverage on open-domain and knowledge-intensive tasks.	×	0.04
Early work combined neural retrievers with sequence-to-sequence generators for open-domain QA.	×	0.04
Subsequent work has extended 4the RAG paradigm to more complex settings, including multi-hop reasoning and conversational	×	0.09

References

- http://arxiv.org/abs/2605.31176v1
- http://arxiv.org/abs/2403.10939v1
- http://arxiv.org/abs/2205.02303v1