Diverse Retriever Portfolios Enhance Multi-Hop Reasoning on HotpotQA

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How does the diversity of retriever portfolios impact multi-hop reasoning performance on the HotpotQA benchmark compared to single-retriever systems when measured by EM (Exact Match) and F1 scores. 12 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.2/10. 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 diversity of retriever portfolios impact multi-hop reasoning performance on the HotpotQA benchmark compared to single-retriever systems when measured by EM (Exact Match) and F1 scores?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.2/10.

3 Results

13 papers retrieved. 12 claims extracted; 2 independently verified. Quality review score: 5.2/10.

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
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.	×	0.03
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 reached 0.594 support recall and 0.500 support F1, while the top-k average baseline reach	×	0.04
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 method was evaluated on diverse open-domain and multi-hop QA benchmarks: HotpotQA, 2WikiMultihopQA, TriviaQA, and Mu	✓	0.20
The method consistently yields better retrieval recall and answer accuracy compared to single-retriever baselines and in	✓	0.17
The method significantly reduces latency and token usage.	×	0.06

References

- http://arxiv.org/abs/2605.31176v1
- http://arxiv.org/abs/2502.11228v2
- http://arxiv.org/abs/2404.14464v1