Scalability and Performance of Retriever Portfolios in Large-Scale RAG Systems

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How scalable is the retriever portfolio approach in terms of latency and throughput when deployed in large-scale production RAG systems, and how does it compare to single-retriever baselines on the. 17 claims were extracted from source literature; 1 was 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 scalable is the retriever portfolio approach in terms of latency and throughput when deployed in large-scale production RAG systems, and how does it compare to single-retriever baselines on the BEIR benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 14 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

14 papers retrieved. 17 claims extracted; 1 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.	×	0.10
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	×	0.12
Retrieval performance was measured in isolation, independent of the answer model, using a size-k portfolio evaluated by	×	0.04
The portfolio was trained once on the pooled training queries from all four benchmarks and then evaluated on the corresp	×	0.04
The portfolio selection is not equivalent to picking the best retrievers on average.	×	0.04
A natural alternative to portfolio optimization is to perform a grid search over retriever configurations, rank candidat	×	0.04
At k = 5, the baseline reaches only 0.492 support recall and 0.432 support F1, while the learned portfolio reaches 0.594	×	0.02
The top-k average list is dominated by closely related GraphDense/E5 configurations, so additional members add little ne	×	0.01
Gains are not explained by retrieving more documents.	×	0.04
The method was evaluated on diverse open-domain and multi-hop QA benchmarks (HotpotQA, 2WikiMultihopQA, TriviaQA, and Mu	×	0.08
The method consistently yields better retrieval recall and answer accuracy compared to single-retriever baselines and in	✓	0.16
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	×	0.15
RAG improves factual accuracy and knowledge coverage on open-domain and knowledge-intensive tasks.	×	0.09
Early work combined neural retrievers with sequence-to-sequence generators 4for open-domain QA.	×	0.07
Subsequent work has extended the RAG paradigm to more complex settings, including multi-hop reasoning and conversational	×	0.10

References

- http://arxiv.org/abs/2605.31176v1
- http://arxiv.org/abs/2605.02623v1
- http://arxiv.org/abs/2104.08663v4