Adaptive Retriever Portfolios Outperform Static Multimodal RAG in Latency-Accuracy Trade-offs

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: How does the adaptive selection mechanism in retriever portfolios impact latency and accuracy trade-offs on the AmbiEval benchmark compared to static single-retriever multimodal RAG systems. 13 claims were extracted from source literature; 1 was 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: Retriever Portfolios: A Principled Approach to Adaptive RAG. Research question: How does the adaptive selection mechanism in retriever portfolios impact latency and accuracy trade-offs on the AmbiEval benchmark compared to static single-retriever multimodal RAG systems?.

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 4.2/10.

3 Results

11 papers retrieved. 13 claims extracted; 1 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
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.	×	0.05
The evaluation addressed three questions: (1) do learned portfolios provide better retrieval coverage as the portfolio s	×	0.11
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	×	0.03
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	×	0.05
The baseline of taking the best-of-k score over the top-k candidates by average training score is nearly flat: at k = 5,	×	0.04
The reason for the improved recall scores is that the greedy objective adds lower-average but complementary Vendi and Gr	×	0.03
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.	×	0.03
The method was evaluated on diverse open-domain and multi-hop QA benchmarks (HotpotQA, 2WikiMultihopQA, TriviaQA, and Mu	×	0.11
The method consistently yields better retrieval recall and answer accuracy compared to single-retriever baselines and in	✓	0.17

References

- http://arxiv.org/abs/2402.12317v2

- http://arxiv.org/abs/2605.31176v1
- http://arxiv.org/abs/2502.08826v3