Dynamic Retriever Portfolios Outperform Static Methods in Real-Time Question Answering

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How does the dynamic retriever portfolio in REALTIME QA compare to traditional static retrieval methods in terms of answer accuracy and latency when evaluated on recent news events. 14 claims were extracted from source literature; 2 were 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 dynamic retriever portfolio in REALTIME QA compare to traditional static retrieval methods in terms of answer accuracy and latency when evaluated on recent news events?.

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 4.5/10.

3 Results

13 papers retrieved. 14 claims extracted; 2 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
The study evaluates retriever portfolios on four QA benchmarks: HotpotQA, 2WikiMultiHopQA, TriviaQA, and MusiQue.	×	0.10
The evaluation uses two answer models: Gemma-3-27B-It and Llama-3.1-70B-Instruct.	×	0.03
Retrieval performance is measured using a best-of-k retrieval score, defined as the maximum support-document score achie	×	0.05
The candidate pool for portfolio selection consists of 360 candidates, including DS and Vendi retrievers with MPNet and	×	0.04
The portfolio is trained once on pooled training queries from all four benchmarks and evaluated on their corresponding t	×	0.03
At portfolio size k=5, the top-k average baseline achieves 0.492 support recall and 0.432 support F1.	×	0.04
At portfolio size k=5, the learned portfolio achieves 0.594 support recall and 0.500 support F1.	×	0.03
The learned portfolio includes lower-average but complementary Vendi and GraphDense variants that cover queries missed b	×	0.05
The top-k average baseline list is dominated by closely related GraphDense/E5 configurations.	×	0.01
The proposed method yields better retrieval recall and answer accuracy compared to single-retriever baselines.	✓	0.15
The proposed method yields better retrieval recall and answer accuracy compared to inference-time tuning methods like Ve	×	0.11
The proposed method significantly reduces latency and token usage compared to baselines.	×	0.07
Retrieval-augmented generation (RAG) grounds large language models in external knowledge by conditioning generation on b	✓	0.17
RAG improves factual accuracy and knowledge coverage on open-domain and knowledge-intensive tasks.	×	0.05

References

- http://arxiv.org/abs/2502.17521v2
- http://arxiv.org/abs/2605.31176v1
- http://arxiv.org/abs/2207.13332v2