# Retriever Portfolios Enhance Robustness in Open-Domain QA Against Adversarial Queries

Assignee Research

June 9, 2026

**Abstract**

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: Can retriever portfolios improve robustness against adversarial queries in open-domain QA by reducing answer faithfulness gaps (measured by QA-SA scores) compared to single-retriever systems on the. 14 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.0/10. 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: Can retriever portfolios improve robustness against adversarial queries in open-domain QA by reducing answer faithfulness gaps (measured by QA-SA scores) compared to single-retriever systems on the AmbiEval 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 5.0/10.

## 3 Results

14 papers retrieved. 14 claims extracted; 2 independently verified. Quality review score: 5.0/10.

## 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. | × | 0.11 |
| The evaluation uses two answer models: Gemma-3-27B-It and Llama-3.1-70B-Instruct. | × | 0.03 |
| A size-k portfolio is evaluated by its best-of-k retrieval score, defined as the maximum support-document score achieved | × | 0.05 |
| The full heterogeneous candidate pool consists of 360 candidates, including DS and Vendi retrievers with MPNet and E5 ba | × | 0.05 |
| The portfolio is trained once on pooled training queries from all four benchmarks and evaluated on their corresponding t | × | 0.03 |
| At k=5, the top-k average baseline achieves 0.492 support recall and 0.432 support F1. | × | 0.03 |
| At k=5, the learned portfolio achieves 0.594 support recall and 0.500 support F1. | × | 0.03 |
| The top-k average baseline list is dominated by closely related GraphDense/E5 configurations. | × | 0.02 |
| The learned portfolio adds lower-average but complementary Vendi and GraphDense variants to cover queries missed by earl | × | 0.05 |
| The method yields better retrieval recall and answer accuracy compared to single-retriever baselines. | ✓ | 0.16 |
| The method yields better retrieval recall and answer accuracy compared to inference-time tuning methods like Vendi-RAG. | × | 0.11 |
| The method significantly reduces latency and token usage compared to baselines. | × | 0.08 |
| Retrieval-augmented generation (RAG) grounds large language models in external knowledge by conditioning generation on b | ✓ | 0.16 |
| RAG improves factual accuracy and knowledge coverage on open-domain and knowledge-intensive tasks. | × | 0.07 |

## References


- http://arxiv.org/abs/2605.31176v1
- http://arxiv.org/abs/2601.11722v1
- http://arxiv.org/abs/2207.13332v2