Context-Aware Retriever Ensembles Outperform Single Retrievers in Multi-Hop Reasoning

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: How does the performance of retriever portfolios compare to single-retriever systems in multi-hop reasoning tasks on the AmbiEval benchmark when measured by answer accuracy and retrieval coverage. 14 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Evaluating Multi-Hop Reasoning in RAG Systems: A Comparison of LLM-Based Retriever Evaluation Strategies. Research question: How does the performance of retriever portfolios compare to single-retriever systems in multi-hop reasoning tasks on the AmbiEval benchmark when measured by answer accuracy and retrieval coverage?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.8/10.

3 Results

10 papers retrieved. 14 claims extracted; 2 independently verified. Quality review score: 3.8/10.

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 uses the HotPotQA, MuSiQue, and SQuAD datasets to simulate a RAG system.	×	0.15
The study compares three LLM-as-judge strategies: an indirect approach derived from eRAG, a direct approach based on ARE	×	0.07
On the HotPotQA dataset, the CARE method achieved an Accuracy of 0.827$\pm$0.02.	×	0.03
On the HotPotQA dataset, the CARE method achieved an F1-Score of 0.814$\pm$0.02.	×	0.03
On the HotPotQA dataset, the Indirect method achieved an Accuracy of 0.642$\pm$0.03.	×	0.02
On the HotPotQA dataset, the Direct method achieved an Accuracy of 0.720$\pm$0.03.	×	0.01
On the MuSiQue dataset, the CARE method achieved an Accuracy of 0.755$\pm$0.02.	×	0.03
On the MuSiQue dataset, the Indirect method achieved a Precision of 0.994$\pm$0.01.	×	0.03
The difference in performance between CARE and the baseline methods on HotPotQA and MuSiQue is statistically significant	×	0.05
For the LLaMa 3.1-8b model, the CARE method experienced a significant decline in overall performance, with accuracy fall	×	0.03
CARE consistently outperformed other approaches across all tested models except for the LLaMa 3.1-8b model.	×	0.04
Performance gains for CARE are most pronounced in models with larger parameter counts and longer context windows.	✓	0.23
Single-hop queries show minimal sensitivity to context-aware evaluation.	✓	0.30
The experimental data is available at https://github.com/lorenzbrehme/CARE.	×	0.13

References

- http://arxiv.org/abs/2605.31176v1
- http://arxiv.org/abs/2604.26649v1
- http://arxiv.org/abs/2604.18234v1