

Context-Aware Retrieval Evaluation for Multi-Hop Reasoning Beyond AmbiEval

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How does the diversity of retriever portfolios impact downstream performance on multi-hop reasoning tasks beyond the AmbiEval benchmark, as measured by accuracy on HotPotQA and TriviaQA. 12 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.8/10. 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 diversity of retriever portfolios impact downstream performance on multi-hop reasoning tasks beyond the AmbiEval benchmark, as measured by accuracy on HotPotQA and TriviaQA?.

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 4.8/10.

3 Results

13 papers retrieved. 12 claims extracted; 3 independently verified. Quality review score: 4.8/10.

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
CARE consistently outperforms existing methods for evaluating multi-hop reasoning in RAG systems.	✓	0.37
The performance gains of 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 indirect evaluation approach derived from the eRAG method [24] involves comparing generated answers with ground truth	×	0.02
The direct evaluation method based on the ARES framework [23] involves determining if a context is crucial to answering	×	0.05
The CARE method involves determining if a context is crucial to answering a question with a ground truth answer, given a	×	0.05
In the indirect method, an LLM attempts to answer the query using only a single context document, and if the answer is e	×	0.06
CARE achieved an accuracy of 0.827 ± 0.02 , F1-Score of 0.814 ± 0.02 , recall of 0.757 ± 0.04 , and precision of 0.880 ± 0.03 on th	×	0.03
CARE achieved an accuracy of 0.755 ± 0.02 , F1-Score of 0.678 ± 0.03 , recall of 0.517 ± 0.04 , and precision of 0.987 ± 0.01 on th	×	0.03
The indirect approach led to a significant improvement in F1-Score for the small LLaMa model on the HotPotQA dataset.	×	0.03
The direct approach resulted in a decline in F1-Score for the reasoning model o4-mini on the HotPotQA dataset.	×	0.04
CARE consistently outperformed other approaches across all models except for the LLaMa 3.1-8b model on the HotPotQA data	×	0.05

References

- <http://arxiv.org/abs/2502.11228v2>
- <http://arxiv.org/abs/2404.14464v1>
- <http://arxiv.org/abs/2604.18234v1>