Scaling Behavior of Retriever Portfolio Diversity in Large-Scale Multimodal Multi-Hop Retrieval

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: What is the scaling behavior of retriever portfolio diversity when applied to large-scale multimodal datasets, and how does it correlate with performance gains on multi-hop retrieval tasks. 13 claims were extracted from source literature; 4 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.2/10. 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: What is the scaling behavior of retriever portfolio diversity when applied to large-scale multimodal datasets, and how does it correlate with performance gains on multi-hop retrieval tasks?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.2/10.

3 Results

12 papers retrieved. 13 claims extracted; 4 independently verified. Quality review score: 5.2/10.

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
CARE consistently outperforms existing methods for evaluating multi-hop reasoning in RAG systems.	✓	0.38
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.	✓	0.31
The indirect evaluation approach led to a significant improvement in F1-Score for the small LLaMa model.	×	0.02
The direct approach resulted in a decline in F1-Score for the reasoning model o4-mini.	×	0.02
For CARE, the reasoning model o4-mini exhibited a decrease in accuracy, F1-Score, and recall compared to GPT-4.1, while	×	0.03
The LLaMa 3.1-8b model experienced a significant decline in overall performance, with substantial drops in both F1-Score	×	0.03
CARE consistently outperformed other approaches across all models except for the LLaMa 3.1-8b model.	×	0.05
The indirect method labels a context as non-relevant if the LLM cannot answer the query using only that context.	×	0.04
The direct method evaluates whether a context is crucial to answering the query with the ground-truth answer.	×	0.03
CARE evaluates whether a context is crucial to answering the query with the ground-truth answer, given a list of context	×	0.05
The experiments were conducted using the HotPotQA, MuSiQue, and SQuAD datasets.	×	0.12
The complete data of the experiments is available at https://github.com/lorenzbrehme/CARE.	✓	0.19

References

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

- http://arxiv.org/abs/2507.23334v2
- http://arxiv.org/abs/2604.18234v1