Hierarchical Memory Indexing Effects on Retrieval Precision in Multi-Hop RAG Systems

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: What is the effect of hierarchical memory indexing on retrieval precision in RAG systems using multi-hop QA benchmarks such as HotPotQA when applied to tabular data versus unstructured text. 11 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.4/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: The Reasoning Bottleneck in Graph-RAG: Structured Prompting and Context Compression for Multi-Hop QA. Research question: What is the effect of hierarchical memory indexing on retrieval precision in RAG systems using multi-hop QA benchmarks such as HotPotQA when applied to tabular data versus unstructured text?.

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 4.4/10.

3 Results

13 papers retrieved. 11 claims extracted; 1 independently verified. Quality review score: 4.4/10.

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 accuracy is determined by a heuristic-first pipeline: normalized string matching (lowercasing, stripping articles/pu	×	0.04
A 102-judgment audit with Claude Opus 4.6 found 96.1% agreement with the pipeline.	×	0.01
With the budget components (free 384-dimensional embeddings, $0.05/M-token indexing LLM), KET-RAG matches or exceeds its	×	0.09
A budget 8B model, augmented with SPARQL CoT and graph-walk compression, surpasses the unaugmented 70B baseline on 2Wiki	✓	0.17
SPARQL CoT alone improves every configuration, with gains up to +14.2 pp on 2WikiMHQA for 8B and +12.2 pp for 70B.	×	0.12
When both augmentations are applied, 70B + SPARQL reaches the overall best accuracy (61.0% on 2WikiMHQA, 80.2% on Hotpot	×	0.06
Difficulty tracks reasoning depth as expected: HotpotQA (2-hop, 67-80%), 2WikiMHQA (mixed, 30-61%), MuSiQue (2-4 hop, 19	×	0.03
On MuSiQue, 8B abstentions drop from 52.0% to 20.6% (-31.4 pp), suggesting that expressing relationships as triple patte	×	0.08
Both CoT methods improve significantly over the baseline, confirming decomposition as the primary driver.	×	0.03
Reasoning failures account for 73% to 84% of all errors across datasets.	×	0.13
Graph-walk context compression reduces input tokens by $\sim$60% with no LLM calls, directly reducing the needle-in-a-haystac	×	0.13

References

- http://arxiv.org/abs/2603.14045v2
- http://arxiv.org/abs/2510.14278v1
- http://arxiv.org/abs/2506.08074v1