Context Window Expansion and Multi-Hop Reasoning Degradation in Retrieval-Augmented Llama-3 Models

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: What is the correlation between context window expansion from 100K to 500K tokens and the degradation of multi-hop reasoning scores in retrieval-augmented Llama-3 models compared to standard. 12 claims were extracted from source literature; 1 was 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: Vendi-RAG: Adaptively Trading-Off Diversity And Quality Significantly Improves Retrieval Augmented Generation With LLMs. Research question: What is the correlation between context window expansion from 100K to 500K tokens and the degradation of multi-hop reasoning scores in retrieval-augmented Llama-3 models compared to standard pre-training baselines?.

2 Methodology

Systematic literature search across multiple databases yielded 16 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

16 papers retrieved. 12 claims extracted; 1 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
Experiments were conducted on three multi-hop QA benchmark datasets: MuSiQue, HotpotQA, and 2WikiMultiHopQA.	✓	0.16
The sensitivity analysis of the VSR process used 100 randomly sampled queries from the dataset.	×	0.03
Setting the parameter s = 0.0 represents a baseline pure similarity search scenario relying exclusively on cosine simila	×	0.03
In the sensitivity analysis, increasing the parameter s from 0.0 to 1.0 causes both Kendall's τ and Spearman's ρ to decr	×	0.02
At s = 0.2, the Kendall's τ value is 0.797 and the Spearman's ρ value is 0.828.	×	0.02
At s = 1.0, the Kendall's τ value is 0.074 and the Spearman's ρ value is 0.078.	×	0.02
Higher s values in the Vendi-RAG retrieval process promote retrieval diversity by prioritizing documents that may be les	×	0.14
The Vendi Score (VS) explicitly quantifies semantic diversity in a set of documents.	×	0.13
The Vendi Score attains its maximum value n when all documents in the set are orthogonal (fully diverse).	×	0.05
Standard similarity search (SS) often results in redundant documents with high similarity.	×	0.04
Maximal Marginal Relevance (MMR) struggles to capture global semantic diversity compared to the Vendi Score approach.	×	0.09
Figure 3 presents a performance comparison of Vendi-RAG and Adaptive-RAG variants across the MuSiQue, HotpotQA, and 2Wik	×	0.15

References

- http://arxiv.org/abs/2510.25621v1
- http://arxiv.org/abs/2502.11228v2
- http://arxiv.org/abs/2404.14464v1