SOVEREIGN: How does the retrieval efficiency of Llama-3-8B-128K vary across context lengths 32K, 64K, and 128K on the MuS

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Processing long contexts presents a significant challenge for large language models (LLMs). While recent advancements allow LLMs to handle much longer contexts than before (e.g., 32K or 128K tokens), it is computationally expensive and can still be insufficient for many applications. Retrieval-Augmented Generation (RAG) is considered a promising strategy to address this problem. However, conventional RAG methods face inherent limitations because of two underlying requirements: 1) explicitly stated queries, and 2) well-structured knowledge. These conditions, however, do not hold in general long

1 Introduction

Analysis of: MemoRAG: Boosting Long Context Processing with Global Memory-Enhanced Retrieval Augmentation. Research goal: How does the retrieval efficiency of Llama-3-8B-128K vary across context lengths 32K, 64K, and 128K on the MuSiQue benchmark when using the Tree of Reviews framework compared to chain-based retrieval?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

6 papers retrieved. 12 claims extracted, 12 verified. Tribunal: 8.5/10 $\rightarrow$ APPROVE (revision_round=0). Policy: AUTO_APPROVE.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
Processing long contexts presents a significant challenge for large language models (LLMs).	✓	0.40
Recent advancements allow LLMs to handle much longer contexts than before (e.g., 32K or 128K tokens).	✓	0.39
Retrieval-Augmented Generation (RAG) is considered a promising strategy to address the problem of processing long contex	✓	0.43
Conventional RAG methods face inherent limitations because of two underlying requirements: 1) explicitly stated queries,	✓	0.45
MemoRAG is a novel RAG framework empowered by global memory-augmented retrieval.	✓	0.32
MemoRAG features a dual-system architecture.	✓	0.17
MemoRAG employs a light but long-range system to create a global memory of the long context.	✓	0.29
MemoRAG generates draft answers, providing useful clues for the retrieval tools to locate relevant information within th	✓	0.34
MemoRAG leverages an expensive but expressive system, which generates the final answer based on the retrieved informatio	✓	0.27
MemoRAG realizes the memory module in the form of KV compression.	✓	0.17
MemoRAG reinforces its memorization and cluing capacity from the Generation quality's Feedback (a.k.a. RLGF).	✓	0.20
MemoRAG achieves superior performances across a variety of long-context evaluation tasks.	✓	0.27

References

- https://doi.org/10.48550/arxiv.2409.05591
- https://doi.org/10.1145/3696410.3714805
- https://doi.org/10.18653/v1/2025.findings-acl.903