SOVEREIGN: How does the accuracy of Tree of Reviews on MuSiQue at 128K context degrade when the number of distractor pass

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Long-context capability is considered one of the most important abilities of LLMs, as a truly long context-capable LLM shall enable its users to effortlessly process many originally exhausting tasks -e.g., digesting a long-form document to find answers v.s., directly asking an LLM about it.However, existing realtask-based long-context evaluation benchmarks have a few major shortcomings.For instance, some Needle-in-a-Haystack-like benchmarks are too synthetic, and therefore do not represent the real world usage of LLMs.While some real-task-based benchmarks like Long-Bench avoid this problem, su

1 Introduction

Analysis of: 100-LongBench: Are de facto Long-Context Benchmarks Literally Evaluating Long-Context Ability?. Research goal: How does the accuracy of Tree of Reviews on MuSiQue at 128K context degrade when the number of distractor passages is increased from 5 to 20, relative to chain-based retrieval, using Llama-3-128K?.

2 Methodology

Multi-query arXiv search (1 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

1 papers retrieved. 6 claims extracted, 6 verified. Tribunal: 8.0/10 → REVISE (revision_round=1). Policy: SOFT_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
Long-context capability is considered one of the most important abilities of LLMs	✓	0.24
Some Needle-in-a-Haystack-like benchmarks are too synthetic and do not represent real world usage of LLMs	✓	0.26
Real-task-based benchmarks like Long-Bench have data samples with fixed sequence length	✓	0.24
Most benchmarks do not provide proper metrics to separate long-context performance from the model's baseline ability	✓	0.31
The paper introduces a length-controllable, real-life reflective benchmark with a novel metric	✓	0.18
The introduced benchmark disentangles baseline knowledge from long-context capabilities	✓	0.19

References

- https://doi.org/10.18653/v1/2025.findings-acl.903