ReST-KV, H2O, and SnapKV Retrieval Accuracy on LongEval's Multilingual Needle-in-a-Haystack Task

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: What is the comparative retrieval accuracy of ReST-KV versus H2O and SnapKV on the Needle-in-a-Haystack task within LongEval for sequences exceeding 128k tokens. 12 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Multilingual Needle in a Haystack: Investigating Long-Context Behavior of Multilingual Large Language Models. Research question: What is the comparative retrieval accuracy of ReST-KV versus H2O and SnapKV on the Needle-in-a-Haystack task within LongEval for sequences exceeding 128k tokens?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.8/10.

3 Results

10 papers retrieved. 12 claims extracted; 3 independently verified. Quality review score: 4.8/10.

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
This is the first study to investigate the multilingual long-context behavior of LLMs.	✓	0.31
MLNeedle assesses model performance across seven languages in both monolingual and cross-lingual settings.	✓	0.18
The performance of LLMs is affected by the position of the needle in the haystack.	×	0.10
LLMs show relative robustness to variations in the language of distractor passages.	×	0.05
The source code and datasets are available at https://github.com/AmeyHengle/multilingual-needle-in-a-haystack.	×	0.08
The MLNeedle benchmark evaluates the performance of LLMs on a multilingual question-answering task.	✓	0.17
The MLNeedle benchmark systematically changes the position of the needle, the language of the needle, and the language o	×	0.12
The performance of LLMs should remain relatively stable regardless of changes in language or needle position if they can	×	0.14
The MLNeedle dataset includes examples in English, Arabic, German, Hindi, Spanish, Vietnamese, and Chinese.	×	0.03
The performance of Llama3-8b-instruct, Cohere-aya-23-8b, and Mistral-7b-instruct-v0.2 is evaluated in the MLNeedle bench	×	0.01
The performance of Llama2-7B-Ch, Llama3-8B-Ins, Cohere-Aya-23, and Mistral-7B-Ins is evaluated at different context leng	×	0.04
The performance of Mistral-7B-Instruct-v0.2 and Llama3-8B-Instruct is evaluated across different languages (English, Ger	×	0.02

References

- http://arxiv.org/abs/2408.10151v1
- http://arxiv.org/abs/2605.08840v1
- http://arxiv.org/abs/2406.11230v2