FlowKV Scaling in Multilingual Long-Context Needle-in-a-Haystack Benchmarks

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does FlowKV's performance on the Needle-in-a-Haystack benchmark scale with model size when applied to LLaMA-3 variants (8B, 70B) at context lengths beyond 100K tokens, and how does this compare. 9 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.7/10. 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: How does FlowKV's performance on the Needle-in-a-Haystack benchmark scale with model size when applied to LLaMA-3 variants (8B, 70B) at context lengths beyond 100K tokens, and how does this compare to similar evaluations on Vicuna-13B?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.7/10.

3 Results

12 papers retrieved. 9 claims extracted; 2 independently verified. Quality review score: 4.7/10.

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 MLNeedle benchmark is available at https://github.com/AmeyHengle/multilingual-needle-in-a-haystack.	×	0.11
The performance of LLMs remains relatively stable regardless of changes in language or needle position if they can use i	×	0.15
The model is provided with a question Q to answer and K documents, where exactly one contains the correct answer.	×	0.03
The content of N and H is sampled from Wikipedia articles.	×	0.00
The performance of Llama3-8b-instruct, cohere-aya-23-8b, and mistral-7b-instruct-v0.2 is evaluated across seven language	×	0.02
The performance of Llama2-7B-Ch, Llama3-8B-Ins, Cohere-Aya-23, and Mistral-7B-Ins is evaluated at different context leng	×	0.02
The performance of Mistral-7B-Instruct-v0.2 and Llama3-8B-Instruct is evaluated with different languages for the needle	×	0.06

References

- http://arxiv.org/abs/2402.13718v3
- http://arxiv.org/abs/2408.10151v1
- http://arxiv.org/abs/2407.20114v3