FlowKV, RoPE, and DynamicRoPE Retrieval Accuracy on MLNeedle Beyond 200K Tokens

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 8 peer-reviewed papers addressing the following research question: How does the retrieval accuracy of FlowKV compare to RoPE and DynamicRoPE on the MLNeedle benchmark for Llama-3-8B and Llama-3-70B at context lengths exceeding 200K tokens. 9 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.6/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Retrieval Models Aren't Tool-Savvy: Benchmarking Tool Retrieval for Large Language Models. Research question: How does the retrieval accuracy of FlowKV compare to RoPE and DynamicRoPE on the MLNeedle benchmark for Llama-3-8B and Llama-3-70B at context lengths exceeding 200K tokens?.

2 Methodology

Systematic literature search across multiple databases yielded 8 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.6/10.

3 Results

8 papers retrieved. 9 claims extracted; 1 independently verified. Quality review score: 4.6/10.

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
The TOOLRET benchmark comprises 7.6k diverse retrieval tasks and a corpus of 43k tools.	✓	0.24
The best model (NV-embedd-v1) achieves an nDCG@10 of only 33.83 in the TOOLRET benchmark.	×	0.03
The TOOLRET-train dataset contains more than 200k retrieval tasks.	×	0.11
The pass rate with pre-annotated toolset (oracle) decreases by 10.1 for e5-large-v2, bge-base-v1.5, and bge-large-v1.5.	×	0.05
The Rouge-L score between generated instruction and seed instruction follows a normal distribution with a kernel density	×	0.05
BM25 achieves a score of 18.98, 4.64, 24.62, 15.20, 21.20, 3.37, 28.23, 26.96, 26.76, 5.86, 32.39, and 24.40 in various	×	0.04
NV-Embed-v1 achieves a score of 60 in the benchmark.	×	0.02
The Pearson coefficient is 0.790 and the Spearman coefficient is 0.441 in the evaluation.	×	0.01
COLT achieves scores of 28.91, 4.61, 40.64, 38.83, 20.06, 4.71, 27.78, 18.84, 31.29, 6.05, 42.19, and 34.01 in various m	×	0.01

References

- http://arxiv.org/abs/2302.11947v2
- http://arxiv.org/abs/2503.01763v2
- http://arxiv.org/abs/2408.11848v2