FlowKV and DynamicRoPE Performance Scaling on MLNeedle with Extended Context Lengths in Llama-3-8B

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the performance degradation of FlowKV scale against DynamicRoPE on the MLNeedle benchmark as context length increases from 128K to 512K tokens for Llama-3-8B. 11 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Data Engineering for Scaling Language Models to 128K Context. Research question: How does the performance degradation of FlowKV scale against DynamicRoPE on the MLNeedle benchmark as context length increases from 128K to 512K tokens for Llama-3-8B?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.8/10.

3 Results

14 papers retrieved. 11 claims extracted; 0 independently verified. Quality review score: 3.8/10.

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
The base LLaMA 2 7B/13B models were continued pretraining on data produced by the 'Per-source Upsampling' strategy.	×	0.09
Long context continual pretraining is feasible under academic-level resources.	×	0.13
The configuration on 8$\times$80G A100s takes 5 days, which is about 1% budget than existing works such as Xiong et al. (2023),	×	0.02
Using only the validation loss is insufficient because two data recipes that result in similar loss may have substantial	×	0.03
The Needle-in-a-Haystack and a real-world book-long question answering benchmark (Zhang et al., 2023) were used for eval	×	0.05
The model achieves better long context task performance (Needle performance) without compromising short context performa	×	0.14
The model was trained on 5B tokens, which translates to 2000 optimization steps.	×	0.04
The batch size was set to be 4M tokens.	×	0.03
The hardware used for training was 8 $\times$ 80G A100.	×	0.04
The training took 10 days for 10B tokens with a context length of 80K for LLaMA 2 7B.	×	0.10
The training took 13 days for 10B tokens with a context length of 64K for LLaMA 2 13B.	×	0.07

References

- http://arxiv.org/abs/2410.02660v4
- http://arxiv.org/abs/2404.19553v1

- http://arxiv.org/abs/2402.10171v1