

LongBench Pro Multi-Hop QA: FlowKV and DynamicRoPE Reasoning Divergence at Extended Context Lengths

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: Does scaling the context length beyond 32k tokens reveal a divergence in reasoning accuracy between FlowKV and DynamicRoPE implementations on the multi-hop QA subset of LongBench Pro. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 7.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: NovelHopQA: Diagnosing Multi-Hop Reasoning Failures in Long Narrative Contexts. Research question: Does scaling the context length beyond 32k tokens reveal a divergence in reasoning accuracy between FlowKV and DynamicRoPE implementations on the multi-hop QA subset of LongBench Pro?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.5/10.

3 Results

13 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 7.5/10.

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.

References

- <http://arxiv.org/abs/2601.02872v1>
- <http://arxiv.org/abs/2506.02000v2>
- <http://arxiv.org/abs/2605.02173v1>