FlowKV Selective Eviction and Retrieval Accuracy on LongBench Pro Needle-in-Haystack Tasks

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 7 peer-reviewed papers addressing the following research question: How does FlowKV's selective eviction impact retrieval accuracy on the LongBench Pro needle-in-a-haystack tasks compared to LongNet and SmoothFormer for Llama-3-70b at 500K+ context lengths. 14 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: LongBench Pro: A More Realistic and Comprehensive Bilingual Long-Context Evaluation Benchmark. Research question: How does FlowKV's selective eviction impact retrieval accuracy on the LongBench Pro needle-in-a-haystack tasks compared to LongNet and SmoothFormer for Llama-3-70b at 500K+ context lengths?.

2 Methodology

Systematic literature search across multiple databases yielded 7 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.2/10.

3 Results

7 papers retrieved. 14 claims extracted; 0 independently verified. Quality review score: 3.2/10.

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
LongBench Pro is a fully natural benchmark with 11 primary categories, diverse context requirements, and supports both E	×	0.11
LongBench Pro covers all the core capability dimensions evaluated by existing benchmarks.	×	0.04
LongBench Pro organizes tasks into a two-level taxonomy with 11 primary categories and 25 secondary categories based on	×	0.11
LongBench Pro includes documents from diverse domains and formats, balanced across single-document and multi-document se	×	0.06
Documents in LongBench Pro are assigned to a length bucket if their length falls within $\pm$20% of the target.	×	0.04
All documents in LongBench Pro undergo a compliance review by human annotators to exclude privacy-sensitive, copyrighted	×	0.02
LongBench Pro uses two types of prompts: non-thinking and thinking prompts, with the latter requiring explicit step-by-s	×	0.03
LongBench Pro collects predictions from five advanced models for each sample and uses human annotators to review and imp	×	0.05
Samples in LongBench Pro are reviewed by two annotators independently, and samples with potential issues are evaluated b	×	0.09
Gemini-2.5-Pro achieved an overall score of 55.92 in the LongBench Pro benchmark.	×	0.05
Gemini-2.5-Flash achieved an overall score of 36.14 in the LongBench Pro benchmark.	×	0.03
Gemma-3-27B-It achieved an overall score of 32.16 in the LongBench Pro benchmark.	×	0.04
Gemma-3-12B-It achieved an overall score of 21.76 in the LongBench Pro benchmark.	×	0.03
Gemma-3-4B-It achieved an overall score of 20.89 in the LongBench Pro benchmark.	×	0.04

References

- http://arxiv.org/abs/2605.09649v1

- http://arxiv.org/abs/2601.02872v1
- http://arxiv.org/abs/2308.14508v2