Llama-3-70B Accuracy-Robustness Trade-offs under FlowKV and SmoothEvict on LongBench

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 9 peer-reviewed papers addressing the following research question: How does the accuracy-robustness trade-off curve of Llama-3-70B vary across LongBench subsets when applying FlowKV versus SmoothEvict at 500K token contexts. 12 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.7/10. 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 the accuracy-robustness trade-off curve of Llama-3-70B vary across LongBench subsets when applying FlowKV versus SmoothEvict at 500K token contexts?.

2 Methodology

Systematic literature search across multiple databases yielded 9 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.7/10.

3 Results

9 papers retrieved. 12 claims extracted; 0 independently verified. Quality review score: 2.7/10.

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
LongBench Pro organizes tasks into a two-level taxonomy with 11 primary categories and 25 secondary categories.	×	0.11
LongBench Pro covers all core capability dimensions evaluated by existing benchmarks.	×	0.06
LongBench Pro introduces a 'context requirement' dimension distinguishing between full and partial context requirements.	×	0.10
Documents in LongBench Pro are collected from diverse domains and formats including news, medicine, science, literature,	×	0.03
LongBench Pro includes documents in both English and Chinese, and varies in length from 8k to 256k tokens.	×	0.13
Raw documents are assigned to length buckets if their token count falls within $\pm$20% of the target length.	×	0.02
All documents in LongBench Pro undergo a compliance review to exclude privacy-sensitive or copyrighted content.	×	0.03
LongBench Pro uses the Qwen tokenizer to measure token length.	×	0.05
Each prompt in LongBench Pro is reviewed using a non-thinking and a thinking prompt template.	×	0.04
The thinking prompt requires step-by-step reasoning before producing a final answer, while the non-thinking prompt does	×	0.01
Model predictions are collected from five advanced models for quality assurance in answer review.	×	0.04
Two annotators independently verify each sample in the answer review process.	×	0.02

References

- http://arxiv.org/abs/2308.14508v2
- http://arxiv.org/abs/2601.02872v1
- http://arxiv.org/abs/2505.19293v2