Sliding Window Attention Mismatch and Accuracy Degradation in Long-Context Code Retrieval

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: Does the training-inference mismatch in sliding window attention cause significant accuracy degradation on the Needle In A Haystack test for code repositories larger than 32k tokens. 12 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Pay Attention to What You Need. Research question: Does the training-inference mismatch in sliding window attention cause significant accuracy degradation on the Needle In A Haystack test for code repositories larger than 32k tokens?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.0/10.

3 Results

16 papers retrieved. 12 claims extracted; 0 independently verified. Quality review score: 3.0/10.

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
LLaMA-2-7B achieves an average improvement of 4.7% over the original model in the LandMark PASS KEY retrieval task.	×	0.04
Compared to the LandMark variant of LLaMA-7B, SRA delivers an average improvement of 8.5%.	×	0.01
SRA achieves a notable performance boost, with an average retrieval accuracy improvement exceeding 10% over original mod	×	0.05
SRA substantially improves retrieval capabilities across various models and tasks, highlighting its versatility and effe	×	0.06
SRA delivers consistent performance improvements without requiring any additional fine-tuning or retraining.	×	0.09
The benefits of SRA become increasingly evident as text length grows, with a declining number of ties and a steadily ris	×	0.03
Beyond a context length of 3000 for LLaMA-2-13B-Chat and 3500 for LLaMA-3-8B-Instruct, over half of the total samples sh	×	0.03
SRA has demonstrated significant advantages in long-text comprehension and summarization, with these benefits becoming i	×	0.04
SRA successfully improved the performance of LongChat-7B-16K and LLaMA-3-8B on the LongChat retrieval task by over 10% c	×	0.04
SRA significantly outperformed the original models with LLaMA-3-8B-Instruct and LLaMA-2-13B-Chat on the XSUM summarizati	×	0.04
On the public datasets such as LongBench v1 (v2), SRA improved the performance of a series of LLMs by above 1.5%.	×	0.04
SRA is a plug-and-play method that enhances the comprehension and retrieval capabilities of LLMs without the need for fi	×	0.12

References

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

- http://arxiv.org/abs/2505.09561v2
- http://arxiv.org/abs/2307.13365v3