Sliding Window Attention Degradation in Mistral 7B on LongCodeEval Beyond 32k Tokens

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 7 peer-reviewed papers addressing the following research question: How does the accuracy of Mistral 7B with sliding window attention degrade on the LongCodeEval benchmark compared to full attention baselines when context length exceeds 32k tokens. 8 claims were extracted from source literature; 6 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: Large Language Models: A Survey. Research question: How does the accuracy of Mistral 7B with sliding window attention degrade on the LongCodeEval benchmark compared to full attention baselines when context length exceeds 32k tokens?.

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 7.5/10.

3 Results

7 papers retrieved. 8 claims extracted; 6 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.

5 Extracted Claims

Claim	Verified	Confidence
ChatGPT was released in November 2022.	×	0.09
Large Language Models (LLMs) have demonstrated strong performance on a wide range of natural language tasks.	✓	0.26
LLMs acquire general-purpose language understanding and generation abilities by training billions of parameters on massi	✓	0.25
The relationship between model performance, parameter count, and data volume is predicted by scaling laws.	×	0.10
GPT, LLaMA, and PaLM are three popular LLM families.	✓	0.22
The paper reviews techniques developed to build and augment LLMs.	✓	0.18
The paper surveys datasets prepared for LLM training, fine-tuning, and evaluation.	✓	0.23
The paper compares the performance of several popular LLMs on a set of representative benchmarks.	✓	0.22

References

- https://doi.org/10.48550/arxiv.2402.06196
- https://doi.org/10.18653/v1/2024.acl-long.642
- https://doi.org/10.1007/s10916-024-02045-3