Token-Level Precision in Long-Context Code Completion: Mistral 7B with Sliding Window vs. Standard Attention

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: How does the token-level precision of code completion differ between Mistral 7B with sliding window attention and standard attention mechanisms when processing inputs longer than 32k tokens on. 15 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Dilated Neighborhood Attention Transformer. Research question: How does the token-level precision of code completion differ between Mistral 7B with sliding window attention and standard attention mechanisms when processing inputs longer than 32k tokens on LongCodeEval?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

4 papers retrieved. 15 claims extracted; 1 independently verified. Quality review score: 4.5/10.

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
DiNAT achieves state-of-the-art image segmentation performance when integrated with advanced segmentation frameworks.	×	0.06
The authors extended NATTEN (NA's CUDA extension for PyTorch) to include dilation support and bfloat16 utilization.	×	0.02
The combination of Neighborhood Attention (NA) and Dilated Neighborhood Attention (DiNA) maintains linear complexity.	✓	0.24
The combination of NA and DiNA expands the receptive field exponentially.	×	0.10
The modified NATTEN implementation reduces runtime by orders of magnitude compared to naive implementations.	×	0.00
Dot product self-attention has a time complexity of O(n^2d) and a space complexity of O(n^2) for attention weights.	×	0.09
DiNAT-L with 200M parameters achieves 87.4% Top-1 accuracy on ImageNet.	×	0.01
DiNAT-L with 200M parameters achieves 87.5% Top-1 accuracy on ImageNet in a configuration with 112 dilation levels.	×	0.01
DiNAT-M (40M parameters, 225G FLOPs) achieves 69.1% APb on MSCOCO, outperforming NAT-M which achieves 68.1%.	×	0.01
DiNAT-T (48M parameters) achieves 70.2% APb on MSCOCO, outperforming Swin-T which achieves 68.1%.	×	0.03
DiNAT-S (70M parameters) achieves 70.8% APb on MSCOCO, outperforming NAT-S which achieves 69.8%.	×	0.02
DiNAT-Tiny with a maximum dilation configuration achieves 83.2% Top-1 accuracy on ImageNet.	×	0.01
DiNAT-Tiny with a gradual dilation configuration achieves 48.8% mIoU on ADE20K.	×	0.04
DiNAT-M (77M parameters) achieves 69.8% APb on MSCOCO, outperforming NAT-M which achieves 68.9%.	×	0.02
DiNAT-S (108M parameters) achieves 71.8% APb on MSCOCO, outperforming Swin-S which achieves 70.4%.	×	0.03

References

- http://arxiv.org/abs/2605.26355v1
- http://arxiv.org/abs/2209.15001v3
- http://arxiv.org/abs/2303.15105v1