SpikingBrain and Mistral 7B Throughput in Long-Context Code Completion

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: How does the throughput of SpikingBrain compare to Mistral 7B with sliding window attention on long-context code completion tasks exceeding 10k tokens. 13 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: SWAA: Sliding Window Attention Adaptation for Efficient and Quality Preserving Long Context Processing. Research question: How does the throughput of SpikingBrain compare to Mistral 7B with sliding window attention on long-context code completion tasks exceeding 10k tokens?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.5/10.

3 Results

10 papers retrieved. 13 claims extracted; 0 independently verified. Quality review score: 3.5/10.

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
SWA is identical to full attention when the context length is within the window size.	×	0.08
LongMemEval is a benchmark consisting of various types of long context QA tasks with moderate difficulty.	×	0.07
LongMemEval 24k is constructed by sampling 10 sessions, resulting in 500 samples ranging from 16k to 32k with an average	×	0.02
LongBench-V2 retains only the samples whose context length is under 128k due to GPU memory limitations; thus, 311 of 500	×	0.04
Ruler's Multi-Query task contains 500 samples, and the context length is controlled to 128k (counted by Qwen3 tokenizer)	×	0.03
LLM as judge with GPT-5-Mini is used for LongMemEval and exact match for LongBench-V2 and Ruler.	×	0.02
LongAlign has a sample count of $\sim$10,000.	×	0.00
Fusang-v1-long is a more comprehensive corpus of over 40,000 long context samples that includes LongAlign as a subset.	×	0.07
SWA aware fine tuning is performed using LoRA with rank r = 16 and α = 128.	×	0.04
LoRA is applied only to the query, key, and value projection modules.	×	0.03
A learning rate of 1e-4 with a cosine decay schedule is used.	×	0.02
Models are fine tuned for a single epoch on the sampled long context dataset.	×	0.08
Full Attention (FA) Decode applies SWA only to the prefilling stage and reverts to full attention during the decoding st	×	0.11

References

- http://arxiv.org/abs/2410.13187v3
- http://arxiv.org/abs/2512.10411v5
- http://arxiv.org/abs/2505.09561v2