Token Prioritization in Vcc: Perplexity Analysis on PG-19 for Long Sequences

Assignee Research

June 12, 2026

Abstract

The computational burden of attention in long-context language models has motivated two largely independent lines of work: sparse attention mechanisms that reduce complexity by attending to selected tokens, and gated attention variants that improve training sta-bility while mitigating the attention sink phenomenon. We observe that these approaches address complementary weaknesses and propose Gated Sparse Attention (GSA), an architecture that realizes the benefits of both. GSA incorporates a gated lightning indexer with sigmoid activations that produce bounded, interpretable selection scores, a

1 Introduction

This paper examines: Gated Sparse Attention: Combining Computational Efficiency with Training Stability for Long-Context Language Models. Research question: How does the token prioritization strategy in Vcc affect perplexity scores on the PG-19 benchmark compared to sparse attention patterns like those in LongNet for sequences exceeding 64K tokens?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.0/10.

3 Results

14 papers retrieved. 13 claims extracted; 11 independently verified. Quality review score: 8.0/10.

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
Training uses a 4K context window; evaluation extends to 128K via YaRN positional interpolation.	✓	0.23
All runs use 8$\times$ H100 GPUs.	×	0.14
Gating alone closes most of the gap to GSA, but sparsity contributes an additional reduction, and the combination outper	✓	0.20
GSA leads across the board; the largest gains appear on MMLU (+2.6 points over standard) and GSM8K (+3.1 points), sugges	✓	0.32
All methods perform comparably up to 32K, but standard attention collapses beyond this point. GSA maintains strong perfo	✓	0.29
Standard attention allocates nearly half of its probability mass to the first token; GSA reduces this to under 4%.	✓	0.20
Maximum activation magnitudes drop by an order of magnitude, which we attribute to the regularizing effect of sigmoid ga	✓	0.21
Gating dramatically reduces spike frequency, permitting a 2$\times$ higher learning rate without instability.	✓	0.22
Prefill cost drops by roughly 11$\times$; decode improves similarly. Memory overhead from gating parameters is negligible.	✓	0.26
Output gating (G1) accounts for most of the quality gain; value gating (G2) adds a smaller but consistent improvement. C	✓	0.30
GSA augments a standard transformer layer by applying a value gate (G2), using a gated lightning indexer to score all po	✓	0.18
The gated lightning indexer replaces DSA's ReLU activations with sigmoids, yielding bounded importance scores.	✓	0.22
A fixed budget k may be suboptimal; GSA modulates the selection budget based on score variance.	×	0.15

References

- http://arxiv.org/abs/2601.15305v1
- http://arxiv.org/abs/2602.03216v3
- http://arxiv.org/abs/2307.02486v2