KDA vs Full Attention Accuracy-Throughput Trade-offs on Long-Sequence GEMM Benchmarks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: What is the accuracy-throughput trade-off of Kimi Delta Attention (KDA) versus full attention on the GEMM benchmark when processing sequences longer than 8k tokens. 9 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Kimi Linear: An Expressive, Efficient Attention Architecture. Research question: What is the accuracy-throughput trade-off of Kimi Delta Attention (KDA) versus full attention on the GEMM benchmark when processing sequences longer than 8k tokens?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

3 Results

12 papers retrieved. 9 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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
Kimi Linear achieves a score of 84.3 on the RULER benchmark with a 128k context length.	×	0.06
Kimi Linear achieves a score of 51.0 on the MMLU-Pro benchmark with a 4k context length.	×	0.04
Kimi Linear achieves a 3.98$\times$ acceleration on the RULER benchmark with a 128k context length.	×	0.07
Kimi Linear maintains a low time per output token (TPOT) and matches GDN-H while outperforming MLA at long sequences.	×	0.07
Kimi Linear achieves a 6.3$\times$ faster TPOT (1.84ms vs. 11.48ms) than MLA at 1M tokens.	×	0.05
Kimi Linear uses a consistent model configuration of 2 layers with 2 attention heads, each having a head dimension of 12	×	0.06
Kimi Linear is trained for at most 20,000 steps with a grid search over learning rates in {5 $\times$ 10-5, 1 $\times$ 10-4, 5 $\times$ 10-4,	×	0.03
Kimi Linear is evaluated on three synthetic tasks: Palindrome, Multi Query Associative Recall (MQAR), and another unspec	×	0.05
Kimi Linear achieves the best performance in the ablation study on the hybrid ratio of KDA to MLA attention and other ke	×	0.09

References

- http://arxiv.org/abs/2510.26692v2
- http://arxiv.org/abs/2508.06447v2
- http://arxiv.org/abs/2603.20586v2