Scaling Kimi Delta Attention vs Full Attention: Accuracy-Throughput Trade-offs on Long-Sequence Pile Benchmarks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: What is the trade-off between accuracy and throughput when scaling Kimi Delta Attention (KDA) versus full attention on the Pile benchmark with sequences of 16k tokens or longer. 9 claims were extracted from source literature; 9 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: A Survey of Large Language Models. Research question: What is the trade-off between accuracy and throughput when scaling Kimi Delta Attention (KDA) versus full attention on the Pile benchmark with sequences of 16k tokens or longer?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.7/10.

3 Results

13 papers retrieved. 9 claims extracted; 9 independently verified. Quality review score: 8.7/10.

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
The survey reviews LLM advancements across four key dimensions: pre-training methodologies, post-training techniques, ut	✓	0.23
Pre-training methodologies establish core model capabilities through large-scale self-supervised training, architectural	✓	0.36
Post-training techniques include supervised fine-tuning and reinforcement learning.	✓	0.19
Post-training techniques adapt foundational models to downstream tasks and enhance their alignment and safety.	✓	0.26
Utilization strategies include in-context learning, prompt engineering, and agentic reasoning.	✓	0.20
Utilization strategies optimize real-world deployment and enable effective interaction with external environments.	✓	0.25
Evaluation methods encompass benchmarks for core language capabilities, reasoning, and safety.	✓	0.20
The survey identifies critical research issues concerning theoretical foundations, efficient scaling, alignment, and age	✓	0.25
Large language models are distinguished from their predecessors by unprecedented scale and advanced capabilities.	✓	0.25

References

- https://doi.org/10.48550/arxiv.2107.06419
- https://doi.org/10.1007/s11704-026-60308-3
- https://doi.org/10.48550/arxiv.2311.16867