# Macro Roundup Artcile

## Headline: Scalable MatMul-free Language Modeling

**Article Link:** https://arxiv.org/abs/2406.02528

| | |
|---|---|
| Author(s) | Rui-Jie Zhu, Yu Zhang, Ethan Sifferman, et al. |
| Publication | Arxiv |
| Publication Date | June 18, 2024 |

**Tweet:** A new approach to LLM development eliminates the need for resource-intensive matrix multiplication (MatMul), reducing memory requirements by 61% for training and 90% for generating responses to queries, potentially expanding AI use in consumer devices.

**Summary:** Matrix multiplication (MatMul) typically dominates the overall computational cost of large language models (LLMs). In this work, we show that MatMul operations can be completely eliminated from LLMs while maintaining strong performance at billion-parameter scales. Our experiments show that our proposed MatMul-free models achieve performance on par with state-of-the-art Transformers that require far more memory during inference at a scale up to at least 2.7B parameters. We provide a GPU-efficient implementation of this model which reduces memory usage by up to 61% over an unoptimized baseline during training. Our model's memory consumption can be reduced by more than 10x compared to unoptimized models. This work not only shows how far LLMs can be stripped back while still performing effectively but also points at the types of operations future accelerators should be optimized for in processing the next generation of lightweight LLMs.

**Related Articles:** Are We on the Brink of an AI Investment Arms Race? and Artificial Intelligence’s ‘Insatiable’ Energy Needs Not Sustainable, Arm CEO Says and The AI Transition One Year Later: On Track, but Macro Impact Still Several Years Off

**Primary Topic:** Science

**Topics:** Academic paper, Science

**Permalink:** https://www.edwardconard.com/macro-roundup/a-new-approach-to-llm-development-eliminates-the-need-for-resource-intensive-matrix-multiplication-matmul-reducing-memory-requirements-by-61-for-training-and-90-for-generating-responses-to-queries?view=detail

**Featured Image Link:** https://www.edwardconard.com/wp-content/uploads/2024/06/21335-scalable-matmul-free-language-modeling-featured-thumbnail-image.png