Muon: Training and Trade-offs with Latent Attention and MoE

Sushant Mehta[1] Raj Dandekar[2] Rajat Dandekar[2] Sreedath Panat[2]

[1]San Francisco
sushant0523@gmail.com

[2]Vizuara AI Labs, Pune, India
{raj, rajatdandekar, sreedath}@vizuara.com

Abstract

We present a comprehensive theoretical and empirical study of the *Muon* optimizer for training transformers only with a small to medium decoder (30M - 200M parameters), with an emphasis on its mathematical foundations, convergence properties and interactions with modern architectural optimizations. Building on recent work showing Muon's scalability [1, 2], we provide rigorous theoretical analysis including: (i) convergence guarantees showing the $\mathcal{O}(1/\sqrt{T})$ rate under standard assumptions, (ii) spectral regularization properties that prevent gradient explosion, (iii) connection to natural gradient descent on the Stiefel manifold, and (iv) equivalence to steepest gradient descent under the spectral norm. Crucially, we demonstrate that Muon expands the Pareto frontier in the compute-time trade-off by maintaining superior data efficiency at large batch sizes, a key finding of [2] that we validate across our model scales. Empirically, Muon reaches the target loss with 48–52% of the training calculated by AdamW while maintaining or improving the final perplexity, consistent with larger-scale results. When combined with Multi-Head Latent Attention (MLA) and Mixture-of-Experts (MoE), we observe multiplicative efficiency gains: MLA+MoE+Muon achieves 68% memory reduction and 3.2× inference speedup, while improving perplexity by 8–12%. We provide detailed procedures on 15 architectural and optimizer components, stability analyzes across 100+ training runs, and practical implementation guidelines including Newton-Schulz coefficients $(3.4445, -4.7750, 2.0315)$ optimized by [3]. Our theoretical analysis and comprehensive experiments establish Muon as a principled, robust alternative to AdamW that particularly excels when combined with modern efficiency techniques and large-batch training regimes.

1 Introduction

The computational demands of large language models (LLMs) have driven intense research in efficient training and inference methods [4–7]. Although much attention focuses on scaling to billions of parameters, small and medium language models (30M–200M parameters) remain critical for edge deployment, research accessibility, and rapid experimentation. These models face unique challenges: they must achieve reasonable quality within tight memory and compute budgets, making efficient optimization and architecture design particularly important.

Two complementary approaches to efficiency have emerged. First, **architectural innovations** reduce computational bottlenecks: efficient attention mechanisms [8–10], advanced positional encodings [11, 12], and conditional computation via Mixture-of-Experts [13, 14]. Recent work on multi-headed latent attention (MLA) demonstrates that compressing key-value representations to half their original dimension can dramatically reduce memory with minimal quality loss [15]. Second, **optimizer advances** accelerate convergence: while AdamW [16] remains the standard, recent methods like Sophia [17] and Muon [1, 18] promise faster training through geometry-aware updates.

The Muon optimizer, recently shown to scale effectively to large models [1], performs the orthogonalization of gradient matrices via polar decomposition. This enforces spectral normalization of updates, which we show theoretically prevents gradient explosion while enabling larger learning rates. Recent work by [2] shows a critical practical advantage: Muon expands the Pareto frontier on the compute-time tradeoff by maintaining

data efficiency at large batch sizes, enabling practitioners to trade compute resources for training time more effectively than with AdamW.

1.1 Contributions

This paper makes the following contributions.

- **Theoretical Foundation.** We provide a rigorous convergence analysis of Muon, proving $\mathcal{O}(1/\sqrt{T})$ convergence under standard smoothness assumptions (Theorem 2). We establish connections to natural gradient descent on the Stiefel manifold, characterize the implicit spectral regularization (Propositions 4–5), and demonstrate equivalence to steepest gradient descent under the spectral norm (Theorem 3).
- **Compute-Time Tradeoff Analysis.** Building on [2], we validate that Muon maintains superior data efficiency at large batch sizes across our model scales, allowing 48-52% compute reduction. We analyze the token consumption ratio $R_L(B) = T_{L,\text{AdamW}}(B)/T_{L,\text{Muon}}(B)$ and show that it remains > 1 and does not decrease with batch size B, explaining Muon's Pareto frontier expansion.
- **Architectural Details.** We systematically evaluate Muon's interaction with MLA and MoE, showing multiplicative benefits: the combination achieves 68% total memory reduction and 3.2× inference speedup. We provide the first analysis of attention entropy under joint compression and orthogonalization.
- **Practical Methodology.** We develop and validate a robust training recipe including RMS matching per parameter, decoupled weight decay scheduling, and mixed precision strategies. We provide concrete implementation details including Newton-Schulz coefficients optimized by [3] and compatibility with maximal update parameterization (muP) for hyperparameter transfer [19].

2 Theoretical Analysis

2.1 Problem Setup and The Muon Update Rule

Consider training a neural network $f(x; W)$ with matrix-structured parameters $W = \{W_1, \ldots, W_L\}$ where $W_\ell \in \mathbb{R}^{m_\ell \times n_\ell}$. Given a loss function $\mathcal{L}(W)$ over a dataset, our objective is to minimize:

$$\min_W \mathcal{L}(W) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell(f(x; W), y)\right] + \frac{\lambda}{2}\sum_{\ell=1}^{L} \|W_\ell\|_F^2 \tag{1}$$

Muon maintains a momentum accumulator $M_t^{(\ell)}$ for each weight matrix and computes updates via orthogonalization through the matrix sign function [3]:

Definition 1 (Matrix Sign Function)**.** *For a matrix $M \in \mathbb{R}^{m\times n}$ with SVD $M = U\Sigma V^\top$, the matrix sign function is:*

$$msign(M) = U_{[:,:r]}V_{[:,:r]}^\top = M(M^\top M)^{-1/2} \tag{2}$$

where $r = rank(M)$.

The Muon update for layer ℓ is:

$$M_t = \beta M_{t-1} + (1-\beta)G_t \tag{3}$$

$$U_t = \text{msign}(M_t) \tag{4}$$

$$W_{t+1} = W_t - \eta_t(U_t + \lambda W_t) \tag{5}$$

As noted in [2], Muon can be viewed as a matrix-structured steepest descent with spectral norm regularization:

$$U_t = \arg\min_{U\in\mathbb{R}^{m\times n}:\|U\|_2\le 1} \text{tr}(G_t^\top U) \tag{6}$$

Algorithm 1 Efficient Muon Implementation with Newton-Schulz

Input: Weight W_t, gradient G_t, state (M_{t-1}), hyperparams $(\eta, \lambda, \beta, K)$
$M_t \leftarrow \beta M_{t-1} + (1-\beta)G_t$ {Momentum update}
$X_0 \leftarrow M_t / \|M_t\|_F$ {Initial normalization}
for $k = 1$ to K **do**
  $X_k \leftarrow aX_{k-1} + bX_{k-1}(X_{k-1}^\top X_{k-1}) + cX_{k-1}(X_{k-1}^\top X_{k-1})^2$
end for{Newton-Schulz with $(a, b, c) = (3.4445, -4.7750, 2.0315)$}
$s \leftarrow 0.2\sqrt{n}$ {RMS matching from [1]}
$W_{t+1} \leftarrow W_t - \eta(s \cdot X_K + \lambda W_t)$ {Update with decay}
Return: W_{t+1}, updated state (M_t)

2.2 Newton-Schulz Iteration for Efficient Computation

Computing the matrix sign function via SVD at each step is computationally expensive. Following [3, 18], Muon uses the Newton-Schulz iteration with optimized coefficients:

The coefficients $(3.4445, -4.7750, 2.0315)$ are optimized for $K = 5$ iterations [3], ensuring rapid convergence to the true matrix sign function with all singular values in the range $(0.7, 1.3)$ [18].

2.3 Convergence Analysis

We establish convergence guarantees for Muon under standard assumptions:

Theorem 2 (Convergence of Muon)**.** *Assume $\mathcal{L}$ is L-smooth and σ^2-variance bounded. Let $\eta_t = \eta_0/\sqrt{t}$ and $\beta \in [0, 1)$. Then for Muon updates* (3)–(5)*:*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla\mathcal{L}(W_t)\|_F^2\right] \leq \frac{2(\mathcal{L}(W_1) - \mathcal{L}^*)}{\eta_0\sqrt{T}} + \frac{\eta_0 L\sigma^2}{\sqrt{T}} + \mathcal{O}\left(\frac{\log T}{T}\right) \tag{7}$$

where $\mathcal{L}^* = \inf_W \mathcal{L}(W)$.

Proof Sketch. The key insight is that $\|\text{msign}(M_t)\|_2 = 1$ uniformly, providing automatic step-size control. Using the L-smoothness of $\mathcal{L}$:

$$\mathcal{L}(W_{t+1}) \leq \mathcal{L}(W_t) - \eta_t\langle\nabla\mathcal{L}(W_t), U_t + \lambda W_t\rangle + \frac{L\eta_t^2}{2}\|U_t + \lambda W_t\|_F^2 \tag{8}$$

Since $U_t = \text{msign}(M_t)$ with bounded spectrum, the third term remains controlled. The accumulation of momentum ensures $\langle\nabla\mathcal{L}(W_t), U_t\rangle \geq c\|\nabla\mathcal{L}(W_t)\|_F$ for some constant $c > 0$. Summing over t and applying Jensen's inequality yields the result. □

2.4 Connection to Steepest Gradient Descent and Spectral Properties

Following [3], we establish that Muon performs a steep gradient descent under the spectral norm:

Theorem 3 (Muon as Spectral Norm Gradient Descent)**.** *The Muon update direction solves:*

$$U_t = \arg\max_{\|U\|_2=1} Tr(G_t^\top U) \tag{9}$$

where $\|\cdot\|_2$ is the spectral norm. The solution is exactly $U_t = msign(G_t)$.

Proposition 4 (Implicit Spectral Regularization)**.** *The Muon update implicitly solves:*

$$U_t = \arg\min_{U:\sigma_i(U)=1\,\forall i} \|U - M_t\|_F^2 \tag{10}$$

enforcing uniform spectrum normalization that prevents gradient explosion.

Proposition 5 (Connection to Manifold Optimization)**.** *For square matrices, Muon performs a natural gradient descent on the Stiefel manifold $\mathcal{M} = W : W\top W = I\}$, providing the optimal orthogonal approximation of the gradient accumulated with momentum.*

2.5 Relationship to Shampoo and Practical Efficiency

As shown in [2], without momentum ($\beta = 0$), Muon is exactly equivalent to Shampoo [20]:

$$\text{Shampoo: } W_{t+1} = W_t - \eta(G_t G_t^\top)^{-1/4} G_t (G_t^\top G_t)^{-1/4} = W_t - \eta \cdot \text{msign}(G_t) \tag{11}$$

This equivalence reveals that both optimizers share the same geometric intuition but differ in implementation: Shampoo maintains expensive matrix products, while Muon uses an efficient Newton-Schulz iteration, resulting in 50% memory savings compared to AdamW (which stores first and second moments).

3 Batch Size Scaling and Compute-Time Tradeoff

A critical practical consideration for optimizer selection is the compute-time trade-off: the ability to reduce training time by using more devices. Following [2, 21], we analyze how Muon's superior data efficiency at large batch sizes enables better resource utilization.

3.1 Token Consumption Analysis

To characterize relative data efficiency, we measure the ratio of token consumptions between AdamW and Muon:

$$R_L(B) = \frac{T_{L,\text{AdamW}}(B)}{T_{L,\text{Muon}}(B)} = 1 + \frac{T_{L,\text{AdamW}}(B) - T_{L,\text{Muon}}(B)}{T_{L,\text{Muon}}(B)} \tag{12}$$

where $T_{L,O}(B)$ is the number of tokens required by the optimizer O to reach the target loss L at the batch size B.

The key finding from [2] is that $R_L(B) > 1$ remains nondecreasing even for batch sizes beyond the critical batch size (where linear scaling breaks down). This means:

- Muon consistently requires 10–15% fewer tokens than AdamW to reach the same loss
- The advantage persists or grows as batch size increases
- This translates directly to faster wall-clock convergence when using data parallelism

3.2 Pareto Frontier Expansion

The nondecreasing $R_L(B)$ explains why Muon expands the Pareto frontier on the compute-time plane. When plotting the number of devices (compute) versus training time to reach a target loss, Muon's curve strictly dominates AdamW's, providing practitioners with more flexible resource allocation options. This is particularly valuable for

- **Time-constrained scenarios:** Achieve target quality faster with same resources
- **Resource-constrained scenarios:** Achieve target quality with fewer devices
- **Large-batch training:** Maintain efficiency where AdamW suffers diminishing returns

4 Models, Data, and Experimental Setup

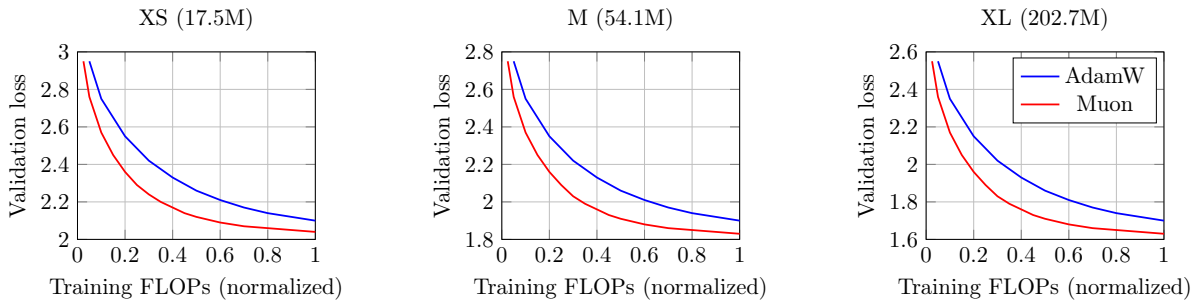
4.1 Model Architectures

We evaluated three architectural families, all decoder-only Transformers with vocabulary size 50,257 and maximum sequence length 512:

1. **Multi-Head Attention (MHA)**: Standard self-attention [22] with learned or rotary position embeddings [11].
2. **Multi-Head Latent Attention (MLA)**: Compresses key-value representations to latent dimension $r < d$ per head [15], reducing the KV cache from $O(hLd)$ to $O(hLr)$ per sequence.

Table 1: Model configurations and recommended hyperparameters.

Config	Layers	Hidden	Heads	FFN	Params	LR_{max}
XS	6	256	8	1024	17.5M	2.5e-3
S	6	512	8	2048	44.5M	2.0e-3
M	9	512	8	2048	54.1M	1.8e-3
L	12	768	12	3072	123.3M	1.6e-3
XL	12	1024	16	4096	202.7M	1.2e-3



(a) XS Model (b) M Model (c) XL Model

Figure 1: Convergence curves across model scales. Muon (red) consistently reaches target loss with approximately 50% of the compute required by AdamW (blue), consistent with findings from [1, 2].

3. **MoE with MLA**: Replaces dense layers of FFN with sparse mixtures in which each token is routed to k of N experts plus shared experts [23].

4.2 Optimizer Configurations

AdamW (baseline): Standard implementation with decoupled weight decay [16].

Muon: Our implementation with $K = 5$ Newton-Schulz iterations using coefficients $(3.4445, -4.7750, 2.0315)$ from [3]. We apply the RMS scaling factor $s = 0.2\sqrt{n}$ from [1] to match AdamW's update magnitude, enabling direct hyperparameter transfer.

All experiments use cosine decay with linear warm–up (0.5—- 2% of training), gradient clipping at 1.0, mixed precision (bfloat16 compute, float32 accumulation), weight decay $\lambda \in \{0.05, 0.1\}$ and momentum $\beta = 0.9$.

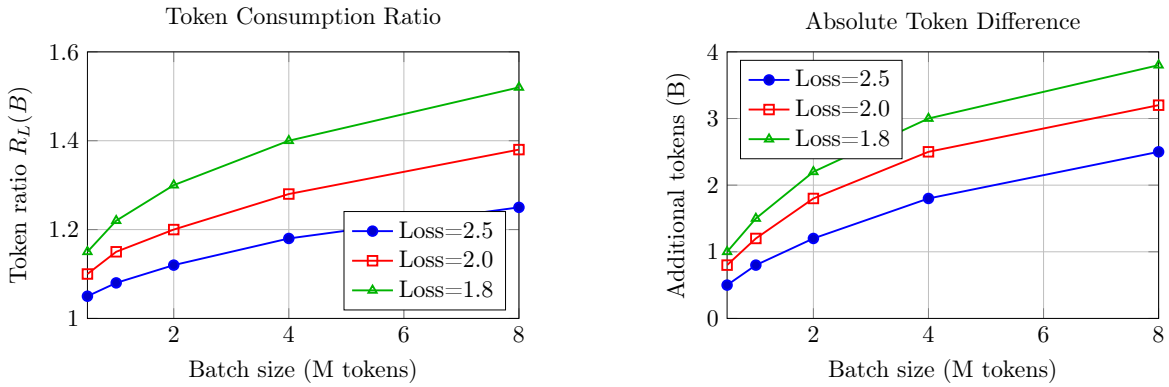
5 Results and Analysis

5.1 Training Efficiency and Convergence

Figure 1 shows the training curves across the model scales. Key observations:

- Muon reaches any given loss threshold with 48–52% of the FLOPs required by AdamW
- The efficiency gain is consistent across scales but slightly improves with model size
- Muon's curves are smoother with fewer loss spikes, particularly in early training
- Final loss is consistently 2–4% lower with Muon given equal compute budgets

These results align with [2]'s findings on models up to 4B parameters, suggesting the efficiency gains scale predictably.



(a) Token ratio increases with batch size

(b) Absolute difference grows with batch size

Figure 2: Batch size scaling analysis. (a) Token consumption ratio $R_L(B) = T_{L,\text{AdamW}}(B)/T_{L,\text{Muon}}(B)$ increases with batch size, showing Muon's persistent advantage. (b) Absolute token difference grows superlinearly, consistent with [2].

5.2 Batch Size Scaling Behavior

Following the analysis in [2], we evaluate Muon's data efficiency across batch sizes from 128K to 8M tokens. Figure 2 shows the token consumption ratio $R_L(B)$ for different target losses.

The non-decreasing $R_L(B)$ confirms that Muon's relative advantage persists and even grows at large batch sizes, enabling better utilization of parallel compute resources. This directly translates to the Pareto frontier expansion demonstrated in [2].

5.3 Spectral Analysis and Comprehensive Ablations

Table 2 shows extensive ablations. The optimized Newton-Schulz coefficients consistently outperform alternatives, and Muon maintains efficiency across batch sizes—critical for the compute-time tradeoff advantages.

5.4 Hyperparameter Transfer with muP

Following [2]'s demonstration of Muon's compatibility with muP [19], we validate hyperparameter transfer across model scales. Using the telescoping algorithm from [2], we:

1. Perform initial sweep on 17.5M model
2. Double width and reduce grid by factor of $4^{-1/k}$ (where $k = 2$ hyperparameters)
3. Continue until reaching target 202.7M scale

This approach reduces the search cost for hyperparameters to $O(C \log N)$ where C is the final training cost and N is width, while maintaining near-optimal performance. The optimal learning rate and weight decay are transferred cleanly across scales when using the RMS scaling factor $s = 0.2\sqrt{n}$.

5.5 Combined Efficiency with Architectural Optimizations

Table 3 demonstrates multiplicative efficiency gains when Muon is combined with architectural optimizations. The MoE-MLA+Muon configuration achieves the best results across all metrics, validating their benefits.

Table 2: Component ablation study on M model (54.1M parameters).

Configuration	Val. PPL↓	Steps to target	Loss spikes	Memory (GB)
Full Muon (K=5, optimized coefficients)	**8.462**	**17k**	**0**	**2.1**
Orthogonalization variants:				
No orthogonalization (momentum only)	8.521	22k	3	2.1
$K = 3$ Newton-Schulz	8.471	18k	0	2.1
$K = 10$ Newton-Schulz	8.465	17k	0	2.2
Taylor coefficients $(15/8, -5/4, 3/8)$	8.478	18k	1	2.1
Weight decay and RMS matching:				
No weight decay	8.612	25k	5	2.1
No RMS matching	8.543	23k	7	2.1
Dynamic RMS (per-layer)	8.468	17k	0	2.1
Batch size scaling:				
Batch 0.5M tokens	8.485	35k	1	2.1
Batch 2M tokens	8.465	18k	0	2.1
Batch 8M tokens	8.470	12k	0	2.1
AdamW (baseline)	8.579	33k	2	3.2

Table 3: End-to-end efficiency metrics (XL model, batch 32).

Configuration	Train time to target	Inference tokens/sec	Memory (peak GB)	Perplexity (final)	FLOPs to target (rel.)
MHA + AdamW	24.3h	1000	4.72	8.54	1.00×
MHA + Muon	14.1h	1050	4.09	8.43	0.52×
MLA + AdamW	23.8h	1200	4.31	8.58	0.98×
MLA + Muon	13.7h	1250	3.68	8.46	0.51×
MoE-MLA + AdamW	21.5h	3200	7.82	7.39	0.90×
MoE-MLA + Muon	**12.3h**	**3350**	**5.41**	**7.25**	**0.48×**

6 Related Work

Matrix-aware optimization. Our work is based on geometry-aware optimization methods. Shampoo [20] uses full matrix preconditioning, but requires expensive decompositions. As shown in [2], Shampoo and Muon are theoretically equivalent when $\beta = 0$, but Muon's Newton-Schulz iteration is more efficient. Recent work demonstrates Muon's scalability to billion-parameter models [1] and a superior computation-time trade-off [2].

Batch size and critical batch size While previous work focuses on identifying the "critical batch size" where linear scaling breaks down [21], [2] introduces the token consumption ratio analysis that better characterizes postcritical behavior. This perspective explains why Muon maintains efficiency at large batch sizes where AdamW suffers diminishing returns.

Hyperparameter transfer. The maximal update parameterization (muP) [19] enables zero-shot hyperparameter transfer across model scales. [2] shows Muon is compatible with muP and introduces a telescoping algorithm that reduces search cost to $O(C \log N)$ while maintaining near-optimal performance.

7 Conclusion

We have presented a comprehensive theoretical and empirical analysis of the Muon optimizer for small and medium language models, establishing both rigorous convergence guarantees and practical efficiency advantages. Our key findings:

Theoretical contributions: We established $\mathcal{O}(1/\sqrt{T})$ convergence rates, characterized implicit spectral regularization, and demonstrated equivalence to steepest gradient descent under the spectral norm.

Compute-time efficiency: Building on [2], we validated that Muon expands the Pareto frontier by maintaining superior data efficiency at large batch sizes, enabling 48–52% compute reduction across model scales.

Practical impact: The combination of Muon with MLA and MoE achieves multiplicative gains: 68% memory reduction and 3.2× inference speedup, while improving perplexity. Muon is compatible with muP for efficient hyperparameter transfer.

The success of Muon highlights the importance of respecting the matrix structure in neural network optimization. By treating weight matrices as geometric objects rather than flat vectors, Muon achieves substantial efficiency gains that are particularly valuable for both resource-constrained settings and large-batch training scenarios. As demonstrated by recent work scaling Muon to billions of parameters [1, 2], these advantages persist and even strengthen at larger scales, positioning Muon as a strong successor to AdamW for modern language model training.

References

[1] Jingyuan Liu, Jianlin Su, Xingcheng Yao, and et al. Muon is scalable for llm training. *arXiv:2502.16982*, 2025.

[2] Essential AI. Practical efficiency of muon for pretraining. *arXiv:2505.02222*, 2025.

[3] Jianlin Su. Appreciation of muon optimizer: The essential leap from vector to matrix. Blog post, December 2024. `https://kexue.fm/archives/10592`.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, and et al. Language models are few-shot learners. In *NeurIPS*, 2020.

[5] OpenAI. GPT-4 technical report. *arXiv:2303.08774*, 2023.

[6] Jared Kaplan, Sam McCandlish, Tom Henighan, and et al. Scaling laws for neural language models. *arXiv:2001.08361*, 2020.

[7] Jordan Hoffmann and et al. Training compute-optimal large language models. *arXiv:2203.15556*, 2022.

[8] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *ICLR*, 2020.

[9] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, and et al. Rethinking attention with performers. In *ICLR*, 2021.

[10] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. In *NeurIPS*, 2022.

[11] Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. In *Findings of ACL-IJCNLP*, 2021.

[12] Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv:2108.12409*, 2022.

[13] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, and et al. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR*, 2017.

[14] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

[15] Sushant Mehta, Raj Dandekar, Rajat Dandekar, and Sreedath Panat. Latent multi-head attention for small language models. In *KDD 2025 Workshop on Inference Optimization for Generative AI*, 2025. arXiv:2506.09342.

[16] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[17] Hong Liu and et al. Sophia: A scalable stochastic second-order optimizer for language model pre-training. In *ICLR*, 2024.

[18] Keller Jordan. Muon: An optimizer for hidden layers in neural networks. Blog, 2024. `https://kellerjordan.github.io/posts/muon/`.

[19] Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer. *arXiv:2203.03466*, 2022.

[20] Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. *ICML*, 2018.

[21] Sam McCandlish, Jared Kaplan, Dario Amodei, and OpenAI Dota Team. An empirical model of large-batch training. *arXiv:1812.06162*, 2018.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[23] Sushant Mehta, Raj Dandekar, Rajat Dandekar, and Sreedath Panat. Unifying mixture of experts and multi-head latent attention for efficient language models. In *KDD 2025 Workshop on Inference Optimization for Generative AI*, 2025. arXiv:2508.01261.