

# Don't be lazy: CompleteP enables compute-efficient deep transformers

Nolan Dey\*  
Cerebras Systems

Bin Claire Zhang\*  
Cerebras Systems

Lorenzo Noci  
ETH Zurich  
Princeton University

Mufan Li  
Princeton University

Blake Bordelon  
Harvard University

Shane Bergsma  
Cerebras Systems

Cengiz Pehlevan  
Harvard University  
Kempner Institute

Boris Hanin  
Princeton University

Joel Hestness  
Cerebras Systems

## Abstract

We study compute efficiency of LLM training when using different *parameterizations*, i.e., rules for adjusting model and optimizer hyperparameters (HPs) as model size changes. Some parameterizations fail to *transfer* optimal base HPs (such as learning rate) across changes in model depth, requiring practitioners to either re-tune these HPs as they scale up (expensive), or accept sub-optimal training when re-tuning is prohibitive. Even when they achieve HP transfer, we develop theory to show parameterizations may still exist in the *lazy learning* regime where layers learn only features close to their linearization, preventing effective use of depth and nonlinearity. Finally, we identify and adopt the parameterization we call *CompleteP* that achieves both depth-wise HP transfer and non-lazy learning in all layers. CompleteP enables a wider range of model width/depth ratios to remain compute-efficient, unlocking shapes better suited for different hardware settings and operational contexts. Moreover, CompleteP enables 12-34% compute efficiency improvements over the prior state-of-the-art. All experiments were run on Cerebras CS-3 systems. A minimal implementation is available at <https://github.com/EleutherAI/nanoGPT-mup/tree/completep>.

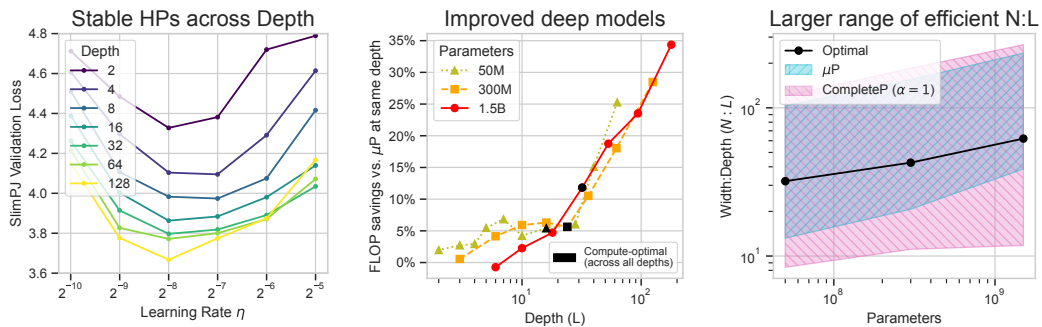


Figure 1: We introduce **CompleteP**, which offers depth-wise HP transfer (**Left**), FLOP savings when training deep models (**Middle**), and a larger range of compute-efficient width/depth ratios (**Right**).

\*Equal contribution. Email: {nolan,claire.zhang,joel}@cerebras.net

# 1 Introduction

A hallmark of modern deep learning is that training larger models leads to better performance [1]. In Large Language Models (LLMs), for instance, the paradigm of pre-training larger models on larger datasets has led to remarkable results in a wide range of downstream evaluations. However, these gains come at a substantial increase in computational cost. As compute budgets grow, practitioners must navigate a complex design space to choose model width and depth, dataset size, batch size, number of training steps, and a variety of other hyperparameters (HPs) in order to find an optimal allocation of resources to minimize a pretraining objective, given a fixed compute budget [2, 3].

The prohibitive cost of naively conducting such a search in large models can result in suboptimal HPs and hence an inefficient use of computational resources. This motivated techniques such as the maximal update parameterization ( $\mu\text{P}$ ), which ensures that optimal HPs remain approximately constant when scaling model *width* [4, 5] and enables a “tune small and train large” strategy.

We refine and thoroughly compare extensions of  $\mu\text{P}$  for simultaneously scaling *depth and width* [6–8]. These scaling strategies define *parameterizations* — sets of rules for how to scale model and optimizer hyperparameters with model size. The core difference between  $\mu\text{P}$  and these depth-aware refinements is how they re-scale the outputs of the transformer’s [9] residual block (as a function of depth  $L$ ) before the output gets added to the residual stream variables  $h^\ell$ :

$$\mathbf{h}^{\ell+1} = \mathbf{h}^\ell + L^{-\alpha} \mathcal{F}_\ell(\mathbf{h}^\ell), \ell \in \{1, \dots, L\}, \quad (1)$$

where  $\mathcal{F}_\ell$  is the  $\ell$ ’th residual block (for transformers, these are MLP and attention blocks). The depth-dependent re-scaling factor is governed by a single parameter  $\alpha \in [0.5, 1]$ .

Yang et al. [7] argue  $\alpha = 0.5$  works best in practice and that HP transfer is not possible at any  $\alpha$ , while Bordelon et al. [8] find  $\alpha = 1$  allows better learning as  $L \rightarrow \infty$ . Since  $\alpha \in \{0.5, 1\}$  are the two most promising potential candidates for depth scaling [6–8], we compare these two values and find  $\alpha = 1$  is consistently more compute-efficient through better HP transfer and faster pre-training (Figure 1). Setting  $\alpha = 1$  gives a parameterization we call *CompleteP* because it is the unique  $\alpha$  ensuring *complete feature learning* as width and depth are scaled (Section 6).

To realize these gains, however, we must extend the parameterization to include principled re-scalings of LayerNorm (LN) and bias learning rates, AdamW’s weight decay  $\lambda$ , and AdamW’s  $\epsilon$  (Table 1). CompleteP is simple to implement, yet yields superior upstream and downstream performance, with gains over alternative approaches increasing with model depth. To summarize, our contributions are:

- **HP transfer across depth.** We compare transfer of learning rate and weight initialization standard deviation across depth (2-128 layers) for the standard parameterization (SP),  $\mu\text{P}$ , and  $\alpha = \{0.5, 1\}$ , when training Pre-LN transformers (Figure 2, 3). We find only  $\alpha = 1$  enables depthwise HP transfer, a feat thought impossible by Yang et al. [7].
- **Optimal shape and compute efficiency.** For SP,  $\mu\text{P}$ , and  $\alpha = \{0.5, 1\}$ , we construct scaling laws over models ranging from 75M to 1.9B total parameters, trained at a compute-optimal frontier of 20 tokens per parameter (TPP) [3]. We use this to revisit the question of compute-efficient transformer *shapes* (i.e. width-to-depth ratio  $N : L$ ). The 1.9B parameter models (i.e. with 1.5B non-embedding parameters) trained with  $\alpha = 1$  achieve 11.8% and 34.4% FLOP savings over  $\mu\text{P}$  for optimally-shaped and 179-layer models, respectively (Figure 4).
- **Refined desiderata for HP transfer.** In Section 6, we describe the differences between SP,  $\mu\text{P}$ , and  $\alpha \in \{0.5, 1\}$ . Given the strong empirical evidence for HP transfer for  $\alpha = 1$ , we devise a refined set of desiderata for HP transfer that includes the notion of *complete feature learning*: we require that the learned representation in every layer of the model remains non-lazy [10] (i.e. non-linear) with respect to the parameter in both that layer and all earlier ones. Only  $\alpha = 1$  ensures complete feature learning, thus we name it CompleteP.
- **Extended parameterizations for modern training.** Modern LLMs train with pre-LN and AdamW. In Table 1 and Appendix D, we extend existing parameterizations [8] to include prescriptions for LayerNorm learning rates  $\eta$ , weight decay, bias, and AdamW  $\epsilon$  as functions of depth and width. These changes are **essential** for stable training with  $\alpha = 0.5$  (Figure 7).

## 2 Related work

**Theoretical approach to HP tuning.** Early methods for selecting HPs analyzed networks at initialization to ensure numerical stability in the initial forward and backward passes [11–28]. Subsequent work devised two parameterizations with consistent training dynamics at infinite width: the *Neural Tangent Kernel (NTK)* parameterization in which the model converges to its linearization around initialization [29–31] and the *mean-field/ $\mu$ P* parameterization [4, 6, 8, 10, 32–39]. Training at the simultaneous limits of width and depth was explored in [40, 41] for fully connected networks, in [6, 7] for vanilla ResNets and in [8] for transformers. Infinite-limit descriptions of training in the mean-field/ $\mu$ P parameterization are difficult to study analytically but give a clear proposal for HP transfer by requiring consistent dynamics of hidden layer representations across model scale [41–44, 32].

This approach to HP transfer was taken up in [6–8], which derive a family of mean-field parameterizations for ResNets indexed by  $\alpha \in [0.5, 1]$  (see Table 1). Based on a heuristic related to feature diversity, [7] argued HP transfer was not truly possible at any  $\alpha$  but that  $\alpha = 0.5$  was the best in practice. In contrast, we show CompleteP ( $\alpha = 1$ ) yields both HP transfer and FLOP savings during pre-training and provide theoretical justification in Section 6. Other interesting mean-field approaches to HP transfer include adaptations to sharpness-aware optimization [45], to low-precision training [46], to sparse training [47], and finally to state space models [48].

**Empirical approaches to depth scaling.** Empirical approaches to HP transfer in [49, 50] normalize layer outputs and learning rates to yield HP transfer across width, consistent with the  $\mu$ P prescription. HP transfer across depth is often more ad hoc: [49] proposes something resembling  $\alpha = 0.5$ , while Large et al. [50] proposes  $\mathbf{h}^{\ell+1} = \frac{L-1}{L} \cdot \mathbf{I} + L^{-1} \cdot \mathcal{F}_\ell(\mathbf{h}^\ell)$  per layer, reminiscent of setting  $\alpha = 1$ . Several other works, though not directly focused on HP transfer, aim to stabilize training in deep ResNets through LN modifications. Sun et al. [51] advocates multiplying pre-LN output in layer  $\ell$  by  $\ell^{-0.5}$ , similar to taking  $\alpha = 0.5$ . Similarly, [52] applies LN to both the input and output of each residual block, avoiding exponential-in-depth activation and gradient variance. Finally, [53] adjusts the scale of weights in each residual block, up-weights residual branch contributions, and inserts LN after the residual add. While this allows one to train very deep networks, it does not achieve depth HP transfer.

**Compute-optimal transformer  $N : L$  ratios.** Theoretical and empirical work have studied the optimal transformer shape. Notably Kaplan et al. [2], using SP and large-scale tests, showed a wide range of  $N : L$  were close to compute-optimality, with  $N : L \approx 100$  being the optimal ratio. This finding guides modern LLM shapes, which often fix  $N : L \approx 100$  when parameters and tokens are scaled [54–72]. Yet SP does not fairly admit stable width and depth scaling, which undermines the prior conclusion. McLeish et al. [73] adopt a  $\eta = \eta_{\text{base}}/\sqrt{L}$  parameterization resembling incomplete  $\alpha = 0.5$  and study compute-optimal  $N : L$ . However, their parameterization does not admit training stability (Figure 7) or HP transfer so their deeper models were disadvantaged, causing them to conclude shallower models are optimal. In Section 5 we revisit the transformer  $N : L$  study with proper width and depth scaling control with CompleteP and show that even  $N : L \approx 10$  remains close to compute optimality. Rather than fixing  $N : L$ , other works perform empirical searches to estimate scaling exponents for  $N$  and  $L$  [74, 75]. Levine et al. [76] propose a theory for transformer capacity based on separation rank, predict that optimal  $N : L$  increases with parameter count, and provide empirical validation. Mixture of experts (MoE) [77] enable transformer capacity to scale without increasing depth. Similarly, [78] propose parallel sub-networks and argue that large depth isn’t necessary for competitive performance. We focus on dense transformers and consider MoE and parallel sub-networks as out of scope.

## 3 Methodology

For all experiments in this work, we train decoder-only Transformer language models [79] with pre-normalization, untied embeddings, ALiBi position embeddings [80] and ReLU<sup>2</sup> nonlinearity [81, 82], using the AdamW optimizer with decoupled weight decay [83] with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.95$ ,  $\epsilon = 1e-16$ . The learning rate  $\eta$  schedule follows a linear warmup of min(10% of steps, 375M tokens), then a linear decay to zero [84]. We pretrain our models using an autoregressive loss (i.e. the next token prediction objective) on the SlimPajama dataset [85] with a maximal sequence length of 2048 tokens using the GPT-2 tokenizer [79]. All models use  $d_{\text{head}} = 64$  for each attention head and

feedforward dimension  $4N$ . For all parameterizations, we use  $\mathbf{Q}^\top \mathbf{K}/N$  as proposed in [5] to account for correlation between  $Q$  and  $K$ . We use  $N_{\text{base}} = 256, L_{\text{base}} = 2$  and scale parameters according to Table 1. All experiments were performed using Cerebras CS-3 systems. We refer the reader to Appendix G for full methodology of all experiments.

Table 1 provides an overview of how to implement the parameterizations we test in this paper.<sup>2</sup> We consider models with depth  $L$  (or  $2L$  residual blocks to account for both MLP and attention blocks), and width  $N$  (e.g., residual activations  $h^\ell \in \mathbb{R}^N$ ). We define adjustments in terms of width multiplier  $m_N = N/N_{\text{base}}$  and depth multiplier  $m_L = L/L_{\text{base}}$ , where  $N_{\text{base}}, L_{\text{base}}$  are the width and the depth of a base model. For the base model, i.e., when  $m_N = 1, m_L = 1$ , all parameterizations are equivalent. We use  $N_{\text{base}} = 256, L_{\text{base}} = 2$ . In experiments where only  $m_N = 1$ , SP is equivalent to  $\mu\text{P}$ ; such results are labelled “SP/ $\mu\text{P}$ ”. We extend the infinite depth parameterizations in the literature [6–8] with corrections for bias learning rate  $\eta$ , LayerNorm  $\eta$ , AdamW  $\epsilon$ , and weight decay  $\lambda$ . See Appendix D for derivations of these extensions and practical advice for applying CompleteP to different architectures. Appendix E for empirical verification of  $\epsilon$  scaling<sup>3</sup>. We provide a minimal implementation of Table 1 which reproduces Figure 7 at: <https://github.com/EleutherAI/nanoGPT-mup/tree/complestep>.

Table 1: Summary of SP,  $\mu\text{P}$ , and  $\alpha \in \{0.5, 1\}$  for a pre-LN transformer language model. Terms related to **width** and **depth** control are highlighted in **orange** and **green** respectively. **Additional tunable parameters** are highlighted in **blue**. *Hidden* refers to all linear layers in the transformer backbone.

Parameterization	SP	$\mu\text{P}$	$\alpha \in \{0.5, 1\}$
Emb. Init. Var.	$\sigma_{\text{base}}^2$	$\sigma_{\text{base}}^2$	$\sigma_{\text{base}}^2$
Emb. LR (AdamW)	$\eta_{\text{base}}$	$\eta_{\text{base}}$	$\eta_{\text{base}}$
Pre-LN Init. Var.	$\sigma_{\text{base}}^2$	$\sigma_{\text{base}}^2$	$\sigma_{\text{base}}^2$
Pre-LN LR (AdamW)	$\eta_{\text{base}}$	$\eta_{\text{base}}$	$\eta_{\text{base}} m_L^{\alpha-1}$
Hidden Init. Var.	$\sigma_{\text{base}}^2$	$\sigma_{\text{base}}^2 \cdot m_N^{-1}$	$\sigma_{\text{base}}^2 \cdot m_N^{-1}$
Hidden LR (AdamW)	$\eta_{\text{base}}$	$\eta_{\text{base}} \cdot m_N^{-1}$	$\eta_{\text{base}} \cdot m_N^{-1} \cdot m_L^{\alpha-1}$
Hidden Bias LR (AdamW)	$\eta_{\text{base}}$	$\eta_{\text{base}}$	$\eta_{\text{base}} m_L^{\alpha-1}$
Hidden WD (AdamW)	$\lambda_{\text{base}}$	$\lambda_{\text{base}} \cdot m_N$	$\lambda_{\text{base}} \cdot m_N$
MHA Residual	$\mathbf{X}^l + \text{MHA}(\text{LN}(\mathbf{X}^l))$	$\mathbf{X}^l + \text{MHA}(\text{LN}(\mathbf{X}^l))$	$\mathbf{X}^l + m_L^{-\alpha} \cdot \text{MHA}(\text{LN}(\mathbf{X}^l))$
MLP Residual	$\mathbf{Z}^l + \text{MLP}(\text{LN}(\mathbf{Z}^l))$	$\mathbf{Z}^l + \text{MLP}(\text{LN}(\mathbf{Z}^l))$	$\mathbf{Z}^l + m_L^{-\alpha} \cdot \text{MLP}(\text{LN}(\mathbf{Z}^l))$
Final-LN Init. Var.	$\sigma_{\text{base}}^2$	$\sigma_{\text{base}}^2$	$\sigma_{\text{base}}^2$
Final-LN LR (AdamW)	$\eta_{\text{base}}$	$\eta_{\text{base}}$	$\eta_{\text{base}}$
Unemb. Init. Var.	$\sigma_{\text{base}}^2$	$\sigma_{\text{base}}^2$	$\sigma_{\text{base}}^2$
Unemb. LR (AdamW)	$\eta_{\text{base}}$	$\eta_{\text{base}}$	$\eta_{\text{base}}$
Unemb. Fwd.	$\mathbf{X}^L \mathbf{W}_{\text{unemb}}^\top$	$\mathbf{X}^L \mathbf{W}_{\text{unemb}}^\top \cdot m_N^{-1}$	$\mathbf{X}^L \mathbf{W}_{\text{unemb}}^\top \cdot m_N^{-1}$
AdamW $\epsilon$ (Residual blocks)	$\epsilon_{\text{base}}$	$\epsilon_{\text{base}} \cdot m_N^{-1}$	$\epsilon_{\text{base}} \cdot m_N^{-1} \cdot m_L^{-\alpha}$
AdamW $\epsilon$ (Emb. & Unemb.)	$\epsilon_{\text{base}}$	$\epsilon_{\text{base}}$	$\epsilon_{\text{base}}$

## 4 Depth-wise HP transfer and $\alpha$

Here, we investigate the HP transfer abilities of  $\mu\text{P}$  and  $\alpha \in \{0.5, 1\}$  as model depth  $L$  is varied.

**Traditional HP transfer** First, we investigate the depth-wise transfer of  $\eta_{\text{base}}$  and  $\sigma_{\text{base}}$  while training all models for 300M tokens with batch size  $B = 128$  and  $\lambda_{\text{base}} = 0$ . We remark that this is the traditional setting where the phenomenon of HP transfer was originally observed [5]. Figure 2 shows that SP,  $\mu\text{P}$ , and  $\alpha = 0.5$  do not have stable optimal  $\eta_{\text{base}}$  or  $\sigma_{\text{base}}$  as depth  $L$  is varied. When using the optimal HPs for  $L = L_{\text{base}} = 2$ , the right column shows these parameterizations do not achieve consistent loss improvements with depth.

**Finding 1:** With SP,  $\mu\text{P}$ , and  $\alpha = 0.5$ , as  $L$  is varied, models do not share the same optimal HPs.

<sup>2</sup>For the embedding layer, we choose an initialization variance scale independent of input dimension due to the input data being one-hot encoded. If the input data is dense, then we would require a pre-factor of  $N_{\text{in}}^{-1/2}$ .

<sup>3</sup>In practice, we observe that a “small enough”  $\epsilon$  is sufficient to achieve consistent dynamics and transfer across width depth. See Appendix G for details.

For CompleteP ( $\alpha = 1$ ), the optimal  $\eta_{\text{base}}$  and  $\sigma_{\text{base}}$  remain stable across depths as evidenced by concentric curves with stable minima. It is the only parameterization to consistently improve loss for deeper models without HP tuning, as shown in Figure 2 right column. Such stable HP transferability dramatically reduces HP tuning budgets for deep models. We demonstrate HP transfer from 2 to 128 layers, exceeding the depth of LLaMA-70B (80 layers) and LLaMA-405B (126 layers) [58].

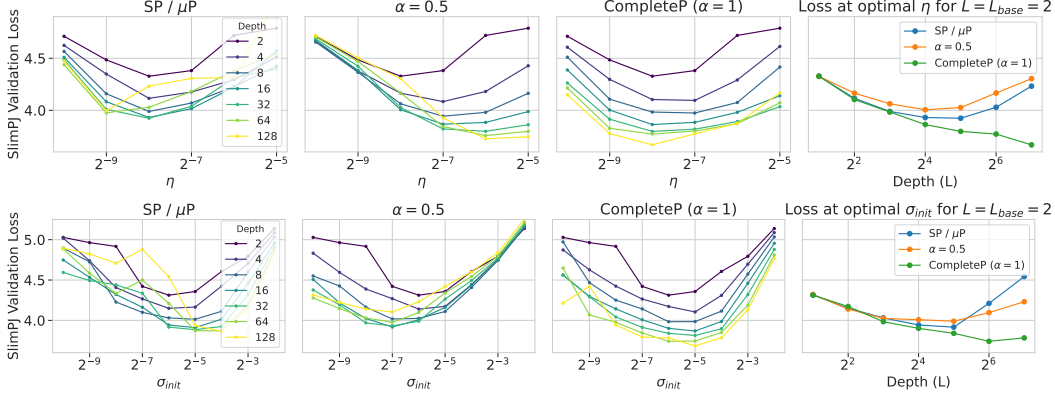


Figure 2: Depth-wise HP transfer, with 300M training tokens. **Top:** Learning rate ( $\eta$ ) transfer. **Bottom:** Initialization standard deviation ( $\sigma_{\text{init}}$ ) transfer. See Table 4 for experimental details.

**Compute-efficient HP transfer** The parameterizations we test are grounded in theories derived under a fixed token count [8]. We now investigate whether the depth-wise transfer of  $\eta_{\text{base}}$  extends to the compute-optimal setup prescribed in Hoffmann et al. [3], where all the models are trained for 20 TPP. We also select compute-efficient batch sizes based on total training FLOPs [2, 86–88], and ensure a well-tuned  $\lambda_{\text{base}}$  [89] (see Appendix G for further details). Figure 3 shows this compute-optimal setup is much less sensitive to the choice of  $\eta_{\text{base}}$  compared to training for 300M tokens with  $B = 128$  and  $\lambda_{\text{base}} = 0$  in Figure 2. Despite this reduced sensitivity, the right column shows  $\alpha = 1$  achieves superior losses to SP,  $\mu\text{P}$ , and  $\alpha = 0.5$  without additional  $\eta_{\text{base}}$  tuning from  $L_{\text{base}}$ . These results lead us to conclude:

**Finding 2:** Only CompleteP ( $\alpha = 1$ ) achieves reliable depth-wise HP transfer.

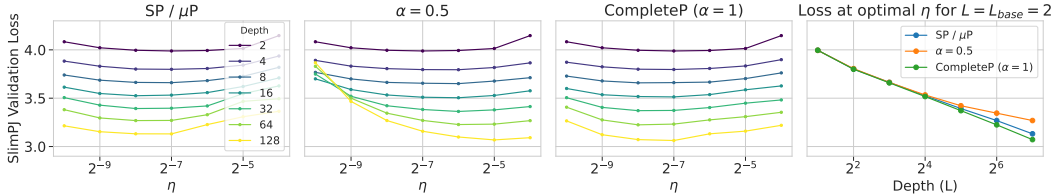


Figure 3: Learning rate transfer test under compute-optimal setting (20 TPP, batch size based on BS-FLOP power law, optimal weight decay  $\lambda_{\text{base}}$ ). See Table 4 for experimental details.

**Finding 3:** Empirically, we observe larger TPP reduces HP sensitivity (Figure 2 vs. 3).

## 5 Re-examining compute-optimal N:L ratio

We now turn to the study of the optimal  $N : L$  ratio in transformers, as its tuning can significantly impact compute efficiency. Kaplan et al. [2] popularized the practice of fixing  $N : L \approx 100$  and scaling both model parameters and tokens [54–72], but without adopting a parameterization that allows for infinite width and depth control. The consequences of this are exemplified in Section 4, where we have shown that deeper models suffer under SP/ $\mu\text{P}$  due to hyperparameter detuning and instability, leading  $\alpha = 1$  to outperform. This depth-disadvantage represents an important confounding variable in these studies on the optimal  $N : L$ . In other words, without an adequate



depth parameterization, deeper models are not given their best chance to succeed. In this section, we conduct the first study of compute-optimal  $N:L$  with simultaneous control of infinite depth and width and show that  $\alpha = 1$  enables a wider range of close-to-optimal aspect ratios, where even  $N:L \approx 10$  remain close to compute-optimal.

**Experimental setup** We train models in the compute-optimal setting of 20 TPP, select compute-efficient batch sizes based on FLOPs [2, 86–88], and ensure well-tuned  $\eta, \lambda, \sigma$  for  $L_{\text{base}} = 2$  [89] as in the previous section. See Appendix G for extensive experimental details and plotting methodology. This time, we vary  $N, L$  to approximately maintain the total number of non embedding parameters  $P_{\text{non-emb}} = 12N^2L$ , for values of  $P_{\text{non-emb}} \in \{50M, 300M, 1.5B\}$ . This setup matches the model sizes from Kaplan et al. [2], but while they use the same number of tokens for all runs (compute-inefficient), we follow the compute-optimal prescription of Hoffmann et al. [3]. Instead of pure web text [2], we train on SlimPajama: a mix of web text, academic prose, and code. We use the maximal update parameterization meaning our HPs have superior tuning with respect to changing width compared to Kaplan et al. [2] who use the standard parameterization. These differences modernize our setup and make our results more applicable to contemporary LLM training.

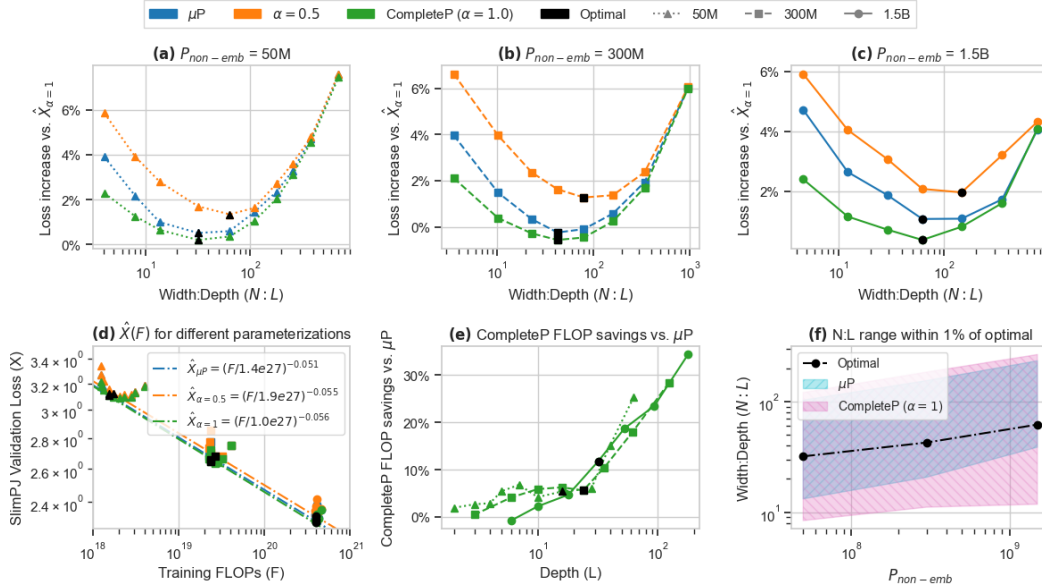


Figure 4: Optimal  $N:L$  across model sizes. (a)-(c) For models of size  $\{50M, 300M, 1.5B\}$  we see an increase in optimal  $N:L$  as size increases but optimal  $N:L$  is less than  $\approx 60$  in this range of model sizes. (d) Scaling laws with FLOPs for the optimal aspect ratios in each parameterization. (e) FLOP savings of CompleteP compared to  $\mu P$  baseline, as a function of depth  $L$ . (f) Shaded regions represent the  $N:L$  range with  $\leq 1\%$  loss increase relative to compute-optimal  $N:L$ .

**Results** We first fit the power law of the form  $\hat{X}(F) = (F/a)^{-b}$  describing how validation cross entropy loss ( $X$ ) scales with a power law in FLOPs ( $F$ ) [1, 2, 62]. The scaling law  $\hat{X}(F)$ , reported in Figure 4d is fitted with the most compute-efficient  $N:L$  from each parameter count and it defines the compute-efficient frontier in our setup. To fairly compare model loss obtained with different FLOP budgets, we then evaluate each  $N:L$  configuration in terms of the distance to the fitted  $\hat{X}_{\alpha=1}(F)$ . This approach is similar to Figure 5 from Kaplan et al. [2]; in our case it provides a measure of the loss difference with respect to the reference scaling law of models trained with CompleteP. We report  $N:L$  as a function of this measure in Figure 4a-c. Notice that the optimal aspect ratio is the same in  $\mu P$  and CompleteP ( $\alpha = 1$ ), while  $\alpha = 0.5$  prefers slightly wider models. In all cases, CompleteP ( $\alpha = 1$ ) gives lower loss values. In particular, we observe the following trend, (see also Figure 4f) :

**Finding 4:** Compute-optimal  $N:L$  trends larger with increasing scale, for all parameterizations.

**Advantages of CompleteP at large depth.** The gap between  $\mu\text{P}$  and CompleteP increases for deeper models. We quantify this gap in terms of FLOPs savings in Figure 4e, where we can see:

**Finding 5:** *The deeper the model, the more FLOP savings CompleteP ( $\alpha = 1$ ) has over  $\mu\text{P}$ .*

We attribute this gap to the fact that at higher depth  $\mu\text{P}$  is more detuned due to lack of HP transfer. In the 1.5B parameter models, CompleteP saves 11.8% of FLOPs for optimal  $N:L$  and 34.4% FLOPs in the deepest models (lowest  $N:L$ ). The shaded regions in Figure 1-right and 4f represent the  $N:L$  range with  $\leq 1\%$  loss increase relative to compute-optimal  $N:L$  (see Appendix G.2 for more detail).

**Finding 6:** *As model scale increases, CompleteP ( $\alpha = 1$ ) enables deep-narrow models (small  $N:L$ ) to remain close to compute-optimal.*

For  $P_{\text{non-emb}}=1.5\text{B}$ ,  $\alpha = 1$  enables  $N:L = 11.8$  to remain within 1% of compute-optimal, compared to  $N:L = 38.7$  from  $\mu\text{P}$ . Hardware latency worsens with increasing depth, making shallow-wide models preferable for latency-sensitive applications. However, low-memory hardware can benefit from narrow-deep models by streaming one layer at a time into memory. Prominent examples of weight streaming exist in both training [90] and inference [91, 92] settings.

**Downstream Performance.** In Table 2 we evaluate the 20 TPP  $P_{\text{non-emb}}=1.5\text{B}$  models at the minimum and optimal  $N:L$  settings, and confirm the upstream gains also translate to gains across five downstream tasks [93–98] that collectively test for common sense reasoning, world knowledge, and reading comprehension. The details of downstream evaluation setup are in Appendix G.3.

Table 2: Zero-shot downstream evaluation accuracy for 20 TPP  $P_{\text{non-emb}} \approx 1.5\text{B}$  models at the optimal and minimum  $N:L$  ratios. We report average task accuracy  $\pm$  standard error. If the average task accuracy is within the standard error, both columns are **bolded**.

Task	Random	Optimal ( $N = 1984, L = 32$ )			Deepest ( $N = 832, L = 179$ )		
		$\mu\text{P}$	$\alpha = 0.5$	CompleteP ( $\alpha = 1$ )	$\mu\text{P}$	$\alpha = 0.5$	CompleteP ( $\alpha = 1$ )
SlimPJ Val. xent. ( $\downarrow$ )	4.701	2.302	2.326	<b>2.286</b>	2.386	2.417	<b>2.330</b>
FLOP savings vs. $\mu\text{P}$ ( $\uparrow$ )	-	0%	-19.6%	<b>11.8%</b>	0%	-23.7%	<b>34.4%</b>
HellaSwag ( $\uparrow$ )	25.0	53.3 $\pm$ 0.5	51.7 $\pm$ 0.5	<b>54.2 <math>\pm</math> 0.5</b>	49.1 $\pm$ 0.5	46.7 $\pm$ 0.5	<b>52.7 <math>\pm</math> 0.5</b>
ARC-Easy ( $\uparrow$ )	25.0	54.4 $\pm$ 1.0	54.5 $\pm$ 1.0	<b>55.6 <math>\pm</math> 1.0</b>	50.0 $\pm$ 1.0	49.2 $\pm$ 1.0	<b>54.6 <math>\pm</math> 1.0</b>
LAMBADA ( $\uparrow$ )	0.0	<b>54.3 <math>\pm</math> 0.7</b>	51.7 $\pm$ 0.7	<b>54.9 <math>\pm</math> 0.7</b>	51.8 $\pm$ 0.7	43.8 $\pm$ 0.7	<b>53.3 <math>\pm</math> 0.7</b>
RACE ( $\uparrow$ )	25.0	<b>34.9 <math>\pm</math> 1.5</b>	<b>34.7 <math>\pm</math> 1.5</b>	<b>34.1 <math>\pm</math> 1.5</b>	33.5 $\pm$ 1.5	32.7 $\pm$ 1.5	<b>35.6 <math>\pm</math> 1.5</b>
PIQA ( $\uparrow$ )	50.0	<b>70.7 <math>\pm</math> 1.1</b>	<b>71.6 <math>\pm</math> 1.1</b>	<b>71.5 <math>\pm</math> 1.1</b>	<b>69.6 <math>\pm</math> 1.1</b>	70.5 $\pm$ 1.1	<b>70.6 <math>\pm</math> 1.1</b>
BoolQ ( $\uparrow$ )	50.0	58.4 $\pm$ 0.9	<b>61.3 <math>\pm</math> 0.9</b>	<b>60.7 <math>\pm</math> 0.9</b>	57.8 $\pm$ 0.9	57.9 $\pm$ 0.9	<b>59.0 <math>\pm</math> 0.9</b>
Downstream Avg. ( $\uparrow$ )	29.2	54.3 $\pm$ 0.3	54.3 $\pm$ 0.3	<b>55.2 <math>\pm</math> 0.3</b>	52.0 $\pm$ 0.3	50.1 $\pm$ 0.3	<b>54.3 <math>\pm</math> 0.3</b>

Given the popularity of TPP>20 training for enhanced parameter efficiency [56], we also compare parameterizations at 200 TPP for  $P_{\text{non-emb}} \in \{50M, 300M\}$  in Table 3. CompleteP ( $\alpha = 1$ ) consistently achieves the best loss across all parameterizations, even in the 200 TPP regime.

Table 3: SlimPJ validation loss for  $P_{\text{non-emb}} \in \{50M, 300M\}$  models trained for 200 TPP.

50M optimal ( $N = 512, L = 16$ )			50M deepest ( $N = 256, L = 63$ )			300M optimal ( $N = 1024, L = 24$ )			300M deepest ( $N = 448, L = 125$ )		
$\mu\text{P}$	$\alpha=0.5$	CompleteP ( $\alpha=1$ )	$\mu\text{P}$	$\alpha=0.5$	CompleteP ( $\alpha=1$ )	$\mu\text{P}$	$\alpha=0.5$	CompleteP ( $\alpha=1$ )	$\mu\text{P}$	$\alpha=0.5$	CompleteP ( $\alpha=1$ )
2.867	2.874	<b>2.859</b>	2.973	2.998	<b>2.950</b>	2.451	2.455	<b>2.443</b>	2.537	2.534	<b>2.497</b>

## 6 Desiderata for Hyperparameter Transfer

Our empirical results suggest that, among all the parameterizations we tested,  $\alpha = 1$  yields the best performance in terms of both HP transfer and pre-training efficiency. In this section, we provide a theoretical heuristic for why this might be and how one might have arrived *a priori* at the CompleteP ( $\alpha = 1$ ) parameterization. Our heuristic consists of proposing three desiderata for constructing a good parameterization. While variants of our first two desiderata were already proposed in [4, 7, 8], Desideratum 3 is novel and distinguishes  $\alpha = 1$ . To frame our discussion, recall that parameterizations are typically constructed to ensure consistent, numerically stable, and meaningful updates for both the hidden layer representations and the network outputs, during training at any model size.

**Setup** Since we are interested in finding such a parameterization for transformers, let us consider a residual network with  $L$  residual layers of width  $N$  in which the representation  $h^\ell \in \mathbb{R}^N$  of a fixed input after  $\ell$  residual layers satisfies the following recursion

$$\mathbf{h}^{\ell+1} = \mathbf{h}^\ell + L^{-\alpha} \mathcal{F}_\ell(\mathbf{h}^\ell; \boldsymbol{\theta}^\ell), \quad \ell = \{1, \dots, L\}. \quad (2)$$

The residual block  $\mathcal{F}_\ell$  is a fixed depth neural network such as an attention or MLP block. The value of  $\alpha$  determines at what scale residual blocks contribute to the residual stream. The desiderata below constrain  $\alpha \in [0.5, 1]$ , consistent with prior work on depth transfer [6–8]. We will denote by  $\theta_\ell$  the trainable parameters in layer  $\ell \in [L] = \{1, \dots, L\}$  and consider updates of the form

$$\boldsymbol{\theta}^\ell \leftarrow \boldsymbol{\theta}^\ell + \Delta\boldsymbol{\theta}^\ell, \quad \Delta\boldsymbol{\theta}^\ell = -\eta^\ell \cdot \mathbf{g}_{\text{AdamW}}^\ell. \quad (3)$$

Here  $g_{\text{AdamW}}$  is the AdamW update and, as we shall see from Desideratum 2, we shall have to take

$$\eta^\ell = \Theta(L^{\alpha-1})$$

to ensure that the change in  $\mathbf{h}^\ell$  entries is  $\Theta(1)$  for any depth or width (see Appendix C.2 or [8]). We focus our presentation on the AdamW optimizer because it is widely used for LLM training, but the desiderata below can be equally well applied to any optimizer (see [6–8]).

**Stable Initialization.** Our first desideratum is a numerical stability requirement for hidden layer representations  $h^\ell$  and network outputs  $f$  at initialization.

**Desideratum 1** (Stable Initialization). *Hidden layers and output remain stable at initialization. More precisely, for all layers  $\ell \in [L]$ ,  $\frac{1}{N}\|\mathbf{h}^\ell\|_2^2 \in \Theta(1)$  and  $f \in O(1)$ , as  $N \rightarrow \infty, L \rightarrow \infty$ .*

This desideratum prescribes how to scale weight variances (or pre-factors) as in Table 1. It also constrains the value of  $\alpha$  to be at least 0.5 (see Appendix C.1). It can be viewed as a numerical stability condition and has been well-studied in the signal-propagation literature [11, 16, 17, 19–21, 40]. We review why this desideratum imposes  $\alpha \geq 1/2$  in Appendix C.1.

**Maximal Residual Stream Update.** Many parameterization schemes, including those from works on signal propagation [16, 17, 27, 12], the NTK parameterization, and the Mean Field /  $\mu\text{P}$  approaches [29, 33, 4], satisfy Desideratum 1 for width scaling  $N \rightarrow \infty$  at fixed  $L$ . The core distinction between them is the presence or absence of feature learning. While there are several ways to make this precise, we follow the original  $\mu\text{P}$  work [4] and formalize feature learning by considering the change in hidden layer representations after each step of training.

Consider the simple case of fixed-depth fully connected networks  $\mathbf{h}^{\ell+1} = \mathbf{W}^\ell \phi(\mathbf{h}^\ell)$  in which the change  $\Delta\mathbf{h}^{\ell+1}$  to first order in the learning rate can be naturally written as a sum of two terms:

$$\Delta\mathbf{h}^{\ell+1} = \Delta\mathbf{W}^\ell \phi(\mathbf{h}^\ell) + \mathbf{W}^\ell \Delta\phi(\mathbf{h}^\ell).$$

The first term captures the change in  $\Delta\mathbf{h}^{\ell+1}$  from the immediately preceding weights  $\mathbf{W}^\ell$  and the second term reflects the change in  $\mathbf{h}^{\ell+1}$  from updates to representations in previous layers.

To derive a parameterization for HP transfer across width, [4] required not only that the relative change  $\|\Delta\mathbf{h}^{\ell+1}\|_2 / \|\mathbf{h}^{\ell+1}\|_2$  to pre-activations in layer  $\ell$  be  $\Theta(1)$  but also that the proportion of this change  $\|\Delta\mathbf{W}^\ell \phi(\mathbf{h}^\ell)\|_2 / \|\mathbf{h}^{\ell+1}\|_2$  attributable directly to the parameters  $W_\ell$  must be  $\Theta(1)$  as well. This excludes degenerate situations such as when one trains only the first few layers of a network.

While the maximal update requirement leads to a unique parameterization for HP transfer across width in a fixed-depth network, it is not directly applicable to the setting where one also scales the network depth. For example, in the case when  $\alpha = 0.5$  and  $\mathcal{F}_\ell(\mathbf{h}^\ell; \boldsymbol{\theta}^\ell) = \mathbf{W}^\ell \phi(\mathbf{h}^\ell)$ , one must actually require  $\|\Delta\mathbf{W}^\ell \phi(\mathbf{h}^\ell)\|_2^2 / \|\mathbf{h}^{\ell+1}\|_2^2 = \Theta(L^{-0.5})$  in order for activations to remain stable in the sense that  $\|\Delta\mathbf{h}^{\ell+1}\|_2 / \|\mathbf{h}^{\ell+1}\|_2 = \Theta(1)$  (see Appendix C.2 & [6, 7]). To cover residual networks of growing depth, we therefore require that the maximal update prescription hold *per residual block*.

**Desideratum 2** (Maximal Residual Stream Update). *Each residual block’s weights should contribute order  $1/L$  to feature movements, and each non-residual block should contribute constant order. More precisely, for all  $\ell \in [L-1]$ , each block’s parameter update  $\boldsymbol{\theta}^\ell \mapsto \boldsymbol{\theta}^\ell + \Delta\boldsymbol{\theta}^\ell$  should contribute the change  $\frac{1}{N}\|\Delta_{\boldsymbol{\theta}^\ell} \mathbf{h}^{\ell+1}\|_2^2 \in \Theta(1/L)$ . Moreover, for the embedding and unembedding layers we require  $\frac{1}{N}\|\Delta\mathbf{W}^0 x\|_2^2 \in \Theta(1)$  and  $\frac{1}{N}\|\Delta\mathbf{W}^L \mathbf{h}^L\|_2^2 \in \Theta(1)$ .*



Desideratum 2 matches the maximal update prescription of [4] when depth  $L$  is finite, as  $\Theta(1/L) = \Theta(1)$ . Furthermore, Desideratum 2 uniquely determines the depth-dependence of the learning rate in AdamW to be  $\eta = \Theta(L^{1-\alpha})$  for the update to satisfy the  $\Theta(1/L)$  scale [6]. Note also that requiring  $\mathcal{F}(h^\ell; \theta_\ell) = \Theta(1)$  with respect to  $L$  constrains  $\alpha \leq 1$ . When combined with Desideratum 1 this explains why we consider  $0.5 \leq \alpha \leq 1$ . We also emphasize that additional care must be taken to correctly determine learning rates for biases and LayerNorm (LN) parameters, otherwise  $\alpha = 0.5$  fails Desideratum 2 for a pre-LN transformer (see Figure 7 and Appendix D).

**Complete Feature Learning.** Variants of Desiderata 1 and 2 have been proposed in prior work and are satisfied by any  $\alpha \in [0.5, 1]$ . In this section we provide one possible intuition for what distinguishes  $\alpha = 1$  and why it might work so well in our experiments. We do so by showing that only  $\alpha = 1$  gives parameterization in which every layer remains uniformly non-linear in its parameters, regardless of depth and width (see Desideratum 3). More precisely, we define the **linearization**  $h^{\text{lin}, \theta}$  of a function  $\mathbf{h}(\theta)$  with respect to  $\theta$  about  $\theta_0$  is

$$\mathbf{h}^{\text{lin}, \theta}(\theta, \theta_0) = \mathbf{h}(\theta_0) + \langle \nabla_{\theta} \mathbf{h}(\theta) |_{\theta_0}, \theta - \theta_0 \rangle.$$

We say  $\mathbf{h}$  is **linear** in  $\theta$  if,  $\mathbf{h} = \mathbf{h}^{\text{lin}, \theta}$ .

**Definition.** We say a layer  $\mathbf{h}^\ell$  is **lazy** with respect to a subset of parameters  $\theta \subset \{\theta_j\}_{j < \ell}$  if at finite depth and width,  $\mathbf{h}^\ell$  is not linear in  $\theta$  and the change  $\Delta_{\theta} \mathbf{h}^\ell$  at initialization from updating only  $\theta$  (i.e. replacing  $\theta \mapsto \theta + \Delta\theta$ ) is asymptotically the same as the change to the linearization of  $\mathbf{h}$ :

$$\frac{|\Delta_{\theta} \mathbf{h}^\ell - \Delta_{\theta} \mathbf{h}_\ell^{\text{lin}, \theta}|}{|\Delta_{\theta} \mathbf{h}_\ell^{\text{lin}, \theta}|} = o(1), \quad \text{as } N, L \rightarrow \infty. \quad (4)$$

**Desideratum 3** (Complete Feature Learning). *The network parameterization satisfies complete feature learning, i.e. neither the hidden layers  $\{\mathbf{h}^\ell\}_{\ell \in [L]}$  nor the model output  $f$  are lazy with respect to any subset of model parameters.*

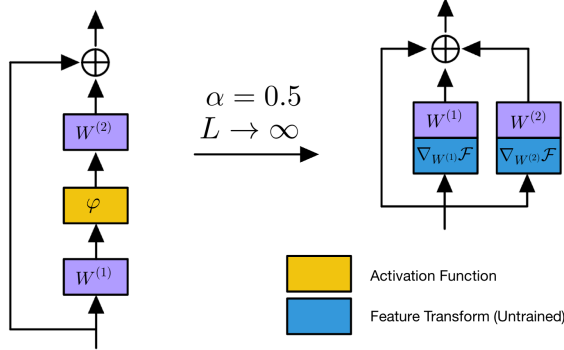


Figure 5: As  $L \rightarrow \infty$ ,  $\alpha = 0.5$  asymptotically linearizes residual blocks (Eqn. 2). **Left:** Depth 2 MLP residual block. For  $\alpha = 0.5$ , this block is only nonlinear with respect to parameters when  $L$  is finite. **Right:** Linearized MLP residual block, with  $\nabla_{W^{(k)}} \mathcal{F}$  a non-trainable transform of  $h^\ell$ .

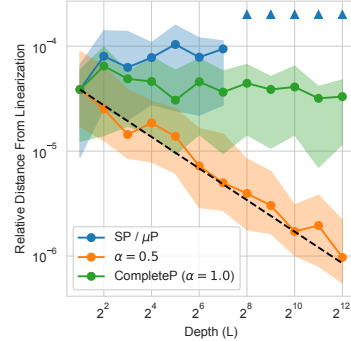


Figure 6: Only  $\alpha = 1$  achieves complete feature learning with stability. SP/ $\mu$ P diverges at large depth (triangles are NaNs), while  $\alpha = 0.5$  converges to the linearization at a rate of  $1/\sqrt{L}$ . The y axis is the left hand side of Eq. 4.

To put this definition into context, note that in the NTK regime, where we do not scale depth, the model output are lazy with respect to all non-output parameters  $\{\theta_\ell\}_{\ell=0}^{L-1}$  [99]. As we will illustrate through a simple example, only  $\alpha = 1$  gives complete feature learning for the  $L \rightarrow \infty$  limit.

**Simple Block Depth 2 Example** To illustrate why Desideratum 3 distinguishes between  $\alpha = 1$  and other values of  $\alpha$ , consider a width  $N = 1$  toy model where we only scale depth  $L \rightarrow \infty$

$$h^{\ell+1} = h^\ell + L^{-\alpha} W_{(2)}^\ell W_{(1)}^\ell h^\ell, \quad \ell = \{1, \dots, L\} \quad (5)$$

To satisfy Desiderata 1 and 2, we need the weight update to satisfy  $\Delta w_\ell^i = \Theta(L^{\alpha-1})$  for both  $i = 1, 2$ . Next, consider the update to  $h^\ell$  from the parameters of preceding residual block parameters:

$(W_{(1)}^\ell, W_{(2)}^\ell) = \theta^\ell \mapsto \theta^\ell + \Delta\theta^\ell$ . Taylor expanding yields:

$$\begin{aligned}\Delta_{\theta^\ell} h^{\ell+1} &= \langle \nabla_{\theta^\ell} h^{\ell+1}, \Delta\theta^\ell \rangle + \frac{1}{2} \nabla_{\theta^\ell}^2 h^{\ell+1} [\Delta\theta^\ell, \Delta\theta^\ell] \\ &= L^{-\alpha} (W_{(2)}^\ell h^\ell \underbrace{\Delta W_{(1)}^\ell}_{L^{\alpha-1}} + W_{(1)}^\ell h^\ell \underbrace{\Delta W_{(2)}^\ell}_{L^{\alpha-1}}) + L^{-\alpha} h^\ell \underbrace{\Delta W_{(1)}^\ell \Delta W_{(2)}^\ell}_{L^{2(\alpha-1)}}.\end{aligned}\quad (6)$$

Note the first term is exactly the change  $\Delta h_{\ell+1}^{\text{lin}, \theta^\ell}$  to the linearization of  $h^{\ell+1}$ , which has the correct order  $\Theta(L^{-1})$ . The second term therefore captures the difference between updating  $\mathbf{h}^{\ell+1}$  and  $\mathbf{h}_{\ell+1}^{\text{lin}, \theta^\ell}$  and has order  $\Theta(L^{\alpha-2})$ . The two terms have the same order only when  $\alpha = 1$ . For all  $\alpha < 1$ , as we increase the depth  $L$  of the network, the contribution of  $\mathbf{h}^{\ell+1}$  to the non-linear term diminishes with scale. We empirically verify this, for the same toy model, in Figure 6 (see Appendix G.4 for details). Since we expect the optimal HPs in shallow models to depend on linear and non-linear dynamics of  $\mathbf{h}_{\ell+1}$ , we do not expect these HPs to be optimal when scaling depth with  $\alpha < 1$ .

While we carried out our computations above only in the simple toy model (Equation (5)), the form of the calculation extends to any arbitrary  $\mathcal{F}_\ell(\mathbf{h}^\ell; \theta^\ell)$  with bounded depth, including large wide MLPs and self-attention blocks. The core observation is that higher order terms in the Taylor expansion ( $k > 1$ ) of  $\Delta_{\theta^\ell} \mathbf{h}^{\ell+1}$  will always have the form

$$\nabla^k \mathbf{h}^{\ell+1} \cdot [[\Delta\theta^\ell]^{\otimes k}] = \Theta(L^{-\alpha} \cdot L^{k(\alpha-1)}), \quad [\Delta\theta^\ell]^{\otimes k} = \Theta(L^{k(\alpha-1)}) \quad (7)$$

We summarize the above discussion on lazy learning into the following statement.

**Finding 7:** Only CompleteP ( $\alpha = 1$ ) ensures stable training, maximal updates, and complete feature learning as  $N \rightarrow \infty, L \rightarrow \infty$ .

Given that  $\alpha = 1$  is the unique  $\alpha$  that ensures complete feature learning while scaling both width and depth, we dub it *CompleteP*.

## 7 Limitations

Our theoretical analysis considers scaling width and depth at fixed token count, which limits the direct applicability to the fixed TPP compute-optimal regime. However, the scaling predictions derived in the fixed token setting (Figure 2) have predicted success in the fixed TPP compute-optimal regime (Figure 3). The  $\hat{X}$  fits in Section 5 were only fit to 3 points due to budget constraints with training models larger than 1.5B parameters. However the three points span a 256x FLOP range and are each the most compute-efficient point from each group of 7-10 N:L values for each parameter count (Table 5), so they are likely to be an accurate picture of the compute-efficient frontier. We use the scaling law fits for interpolation rather than extrapolation, where fitting to limited data can make predictions unreliable.

## 8 Conclusion

We studied in large pre-LN transformers a variety of parameterizations, i.e., prescriptions for adjusting hyperparameters such as learning rates as functions of network architecture. Among them, we found that CompleteP ( $\alpha = 1$ ) gives HP transfer when varying depth and width. We showed that CompleteP achieves significant FLOP savings during pre-training compared to other parameterizations, even in compute-optimal settings with jointly scaled batch, dataset, and model sizes. To achieve this, we extended the theoretical analysis of CompleteP to include prescriptions for how to scale LN parameters and AdamW’s  $\epsilon$  across depth and width.

To understand the salutary effects of CompleteP, we formalized a set of desiderata for designing parameterizations. They include variants of desiderata from prior work as well as a novel property, which we called *complete feature learning*, that distinguishes CompleteP. This parameterization is simple to implement, and we released a public code base to facilitate reproduction and further study. Our empirical tests were conducted on models with up to 1.5B non-embedding parameters. In future work, we plan to train large, state-of-the-art LLMs that leverage CompleteP’s utility in HP transfer and compute-efficient training.

## Acknowledgments and Disclosure of Funding

BB and LN thank the Google PhD fellowship for support. BH is supported by a Sloan Fellowship, NSF DMS-2143754, and NSF DMS-2133806. BB and C.P. acknowledge support by NSF grant DMS-2134157, NSF CAREER Award IIS-2239780, DARPA grant DIAL-FP-038, a Sloan Research Fellowship, and The William F. Milton Fund from Harvard University. C.P. and B.B.’s work has been made possible in part by a gift from the Chan Zuckerberg Initiative Foundation to establish the Kempner Institute for the Study of Natural and Artificial Intelligence.

## References

- [1] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.
- [2] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361*, 2020.
- [3] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack William Rae, and Laurent Sifre. An Empirical Analysis of Compute-optimal Large Language Model Training. In *The Conference on Neural Information Processing Systems (NeurIPS)*, 2022.
- [4] Greg Yang and Edward J Hu. Tensor programs IV: Feature learning in infinite-width neural networks. In *International Conference on Machine Learning*, pages 11727–11737. PMLR, 2021.
- [5] Greg Yang, Edward Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tuning large neural networks via zero-shot hyperparameter transfer. *Advances in Neural Information Processing Systems*, 34: 17084–17097, 2021.
- [6] Blake Bordelon, Lorenzo Noci, Mufan Bill Li, Boris Hanin, and Cengiz Pehlevan. Depthwise hyperparameter transfer in residual networks: Dynamics and scaling limit. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=KZJehvRKGD>.
- [7] Greg Yang, Dingli Yu, Chen Zhu, and Soufiane Hayou. Feature learning in infinite-depth neural networks. In *NeurIPS 2023 Workshop on Mathematics of Modern Machine Learning*, 2023.
- [8] Blake Bordelon, Hamza Tahir Chaudhry, and Cengiz Pehlevan. Infinite limits of multi-head transformer dynamics. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=p0BBKhD5aI>.
- [9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf).
- [10] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32, 2019.
- [11] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994.

- [12] Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurelien Lucchi. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. *Advances in Neural Information Processing Systems*, 35:27198–27211, 2022.
- [13] Emily Dinan, Sho Yaida, and Susan Zhang. Effective theory of transformers at initialization. *arXiv preprint arXiv:2304.02034*, 2023.
- [14] Yann LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–50. Springer, 2002.
- [15] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.
- [16] Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. *Advances in Neural Information Processing Systems*, 29, 2016.
- [17] Samuel S Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. *arXiv preprint arXiv:1611.01232*, 2016.
- [18] Lechao Xiao, Yasaman Bahri, Jascha Sohl-Dickstein, Samuel Schoenholz, and Jeffrey Pennington. Dynamical isometry and a mean field theory of CNNs: How to train 10,000-layer vanilla convolutional neural networks. In *International Conference on Machine Learning*, pages 5393–5402. PMLR, 2018.
- [19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.
- [20] Boris Hanin and David Rolnick. How to start training: The effect of initialization and architecture. *Advances in Neural Information Processing Systems*, 31, 2018.
- [21] Boris Hanin. Which neural net architectures give rise to exploding and vanishing gradients? *Advances in Neural Information Processing Systems*, 31, 2018.
- [22] Boris Hanin and Mihai Nica. Products of many large random matrices and gradients in deep neural networks. *Communications in Mathematical Physics*, 376(1):287–322, 2020.
- [23] Lorenzo Noci, Gregor Bachmann, Kevin Roth, Sebastian Nowozin, and Thomas Hofmann. Precise characterization of the prior predictive distribution of deep ReLU networks. *Advances in Neural Information Processing Systems*, 34:20851–20862, 2021.
- [24] Mufan Li, Mihai Nica, and Dan Roy. The neural covariance SDE: Shaped infinite depth-and-width networks at initialization. *Advances in Neural Information Processing Systems*, 35: 10795–10808, 2022.
- [25] Hongyi Zhang, Yann N Dauphin, and Tengyu Ma. Fixup initialization: Residual learning without normalization. *arXiv preprint arXiv:1901.09321*, 2019.
- [26] Lorenzo Noci, Chuning Li, Mufan Li, Bobby He, Thomas Hofmann, Chris J Maddison, and Dan Roy. The shaped transformer: Attention models in the infinite depth-and-width limit. *Advances in Neural Information Processing Systems*, 36, 2024.
- [27] Soufiane Hayou. Commutative scaling of width and depth in deep neural networks. *Journal of Machine Learning Research*, 25(299):1–41, 2024.
- [28] Akhil Kedia, Mohd Abbas Zaidi, Sushil Khyalia, Jungho Jung, Harshith Goka, and Haejun Lee. Transformers get stable: An end-to-end signal propagation theory for language models. *arXiv preprint arXiv:2403.09635*, 2024.
- [29] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

- [30] Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022.
- [31] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. *Advances in Neural Information Processing Systems*, 32, 2019.
- [32] Blake Bordelon and Cengiz Pehlevan. Self-consistent dynamical field theory of kernel evolution in wide neural networks. *Advances in Neural Information Processing Systems*, 35: 32240–32256, 2022.
- [33] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33): E7665–E7671, 2018.
- [34] Grant Rotskoff and Eric Vanden-Eijnden. Trainability and accuracy of artificial neural networks: An interacting particle system approach. *Communications on Pure and Applied Mathematics*, 75(9):1889–1935, 2022.
- [35] Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in Neural Information Processing Systems*, 31, 2018.
- [36] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020.
- [37] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852, 2020.
- [38] Katie Everett, Lechao Xiao, Mitchell Wortsman, Alexander A Alemi, Roman Novak, Peter J Liu, Izzeddin Gur, Jascha Sohl-Dickstein, Leslie Pack Kaelbling, Jaehoon Lee, and Jeffrey Pennington. Scaling exponents across parameterizations and optimizers. *arXiv preprint arXiv:2407.05872*, 2024.
- [39] Lucas Lingle. A large-scale exploration of  $\mu$ -transfer. *arXiv preprint arXiv:2404.05728*, 2024.
- [40] Daniel A Roberts, Sho Yaida, and Boris Hanin. *The principles of deep learning theory*, volume 46. Cambridge University Press Cambridge, MA, USA, 2022.
- [41] Sho Yaida. Meta-principled family of hyperparameter scaling strategies. *arXiv preprint arXiv:2210.04909*, 2022.
- [42] Nikhil Vyas, Alexander Atanasov, Blake Bordelon, Depen Morwani, Sabarish Sainathan, and Cengiz Pehlevan. Feature-learning networks are consistent across widths at realistic scales. *Advances in Neural Information Processing Systems*, 36:1036–1060, 2023.
- [43] Lorenzo Noci, Alexandru Meterez, Thomas Hofmann, and Antonio Orvieto. Super consistency of neural network landscapes and learning rate transfer. In *Advances in Neural Information Processing Systems*, 2024.
- [44] Blake Bordelon and Cengiz Pehlevan. Dynamics of finite width kernel and prediction fluctuations in mean field neural networks. *Advances in Neural Information Processing Systems*, 36, 2024.
- [45] Moritz Haas, Jin Xu, Volkan Cevher, and Leena Chennuru Vankadara.  $\mu P^2$ : Effective sharpness aware minimization requires layerwise perturbation scaling. *arXiv preprint arXiv:2411.00075*, 2024.
- [46] Charlie Blake, Constantin Eichenberg, Josef Dean, Lukas Balles, Luke Y Prince, Björn Deiseroth, Andres Felipe Cruz-Salinas, Carlo Luschi, Samuel Weinbach, and Douglas Orr.  $u\text{-}\mu P$ : The unit-scaled maximal update parametrization. *arXiv preprint arXiv:2407.17465*, 2024.



- [47] Nolan Simran Dey, Shane Bergsma, and Joel Hestness. Sparse maximal update parameterization: A holistic approach to sparse training dynamics. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [48] Leena Chennuru Vankadara, Jin Xu, Moritz Haas, and Volkan Cevher. On feature learning in structured state space models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [49] Edward Milsom, Ben Anson, and Laurence Aitchison. Function-space learning rates. *arXiv preprint arXiv:2502.17405*, 2025.
- [50] Tim Large, Yang Liu, Minyoung Huh, Hyojin Bahng, Phillip Isola, and Jeremy Bernstein. Scalable optimization in the modular norm. *arXiv preprint arXiv:2405.14813*, 2024.
- [51] Wenfang Sun, Xinyuan Song, Pengxiang Li, Lu Yin, Yefeng Zheng, and Shiwei Liu. The curse of depth in large language models. *arXiv preprint arXiv:2502.05795*, 2025.
- [52] Jeonghoon Kim, Byeongchan Lee, Cheonbok Park, Yeontaek Oh, Beomjun Kim, Taehwan Yoo, Seongjin Shin, Dongyoon Han, Jinwoo Shin, and Kang Min Yoo. Peri-LN: Revisiting layer normalization in the transformer architecture. *arXiv preprint arXiv:2502.02732*, 2025.
- [53] Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, and Furu Wei. DeepNet: Scaling transformers to 1,000 layers. *arXiv preprint arXiv:2203.00555*, 2022.
- [54] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.
- [55] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.
- [56] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [57] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. LLaMA 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [58] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [59] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, et al. Gemma: Open models based on Gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.
- [60] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.
- [61] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

- [62] Nolan Dey, Gurpreet Gosal, Hemant Khachane, William Marshall, Ribhu Pathria, Marvin Tom, and Joel Hestness. Cerebras-GPT: Open compute-optimal language models trained on the Cerebras wafer-scale cluster. *arXiv preprint arXiv:2304.03208*, 2023.
- [63] Nolan Dey, Daria Soboleva, Faisal Al-Khateeb, Bowen Yang, Ribhu Pathria, Hemant Khachane, Shaheer Muhammad, Zhiming, Chen, Robert Myers, Jacob Robert Steeves, Natalia Vassilieva, Marvin Tom, and Joel Hestness. BTLM-3B-8K: 7B parameter performance in a 3B parameter model. *arXiv preprint arXiv:2309.11568*, 2023.
- [64] Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, et al. OLMo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*, 2024.
- [65] Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, et al. 2 OLMo 2 Furious. *arXiv preprint arXiv:2501.00656*, 2025.
- [66] Paolo Glorioso, Quentin Anthony, Yury Tokpanov, James Whittington, Jonathan Pilault, Adam Ibrahim, and Beren Millidge. Zamba: A compact 7B SSM hybrid model. *arXiv preprint arXiv:2405.16712*, 2024.
- [67] Paolo Glorioso, Quentin Anthony, Yury Tokpanov, Anna Golubeva, Vasudev Shyam, James Whittington, Jonathan Pilault, and Beren Millidge. The Zamba2 suite: Technical report. *arXiv preprint arXiv:2411.15242*, 2024.
- [68] Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. *arXiv preprint arXiv:2304.01373*, 2023.
- [69] Mosaic Research Team. Introducing MPT-7B: A new standard for open-source, commercially usable LLMs. <https://www.databricks.com/blog/mpt-7b>, May 2023.
- [70] Mosaic Research Team. MPT-30B: Raising the bar for open-source foundation models. <https://www.databricks.com/blog/mpt-30b>, June 2023.
- [71] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.
- [72] Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, et al. DataComp-LM: In search of the next generation of training sets for language models. *arXiv preprint arXiv:2406.11794*, 2025.
- [73] Sean McLeish, John Kirchenbauer, David Yu Miller, Siddharth Singh, Abhinav Bhatele, Micah Goldblum, Ashwinee Panda, and Tom Goldstein. Gemstones: A model suite for multi-faceted scaling laws. *arXiv preprint arXiv:2502.06857*, 2025.
- [74] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2020.
- [75] Ibrahim Alabdulmohsin, Xiaohua Zhai, Alexander Kolesnikov, and Lucas Beyer. Getting ViT in shape: Scaling laws for compute-optimal model design. *arXiv preprint arXiv:2305.13035*, 2024.
- [76] Yoav Levine, Noam Wies, Or Sharir, Hofit Bata, and Amnon Shashua. The depth-to-width interplay in self-attention. *arXiv preprint arXiv:2006.12467*, 2021.

- [77] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarczyk, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.
- [78] Ankit Goyal, Alexey Bochkovskiy, Jia Deng, and Vladlen Koltun. Non-deep networks. *arXiv preprint arXiv:2110.07641*, 2021.
- [79] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners, 2019. URL <https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf>.
- [80] Ofir Press, Noah Smith, and Mike Lewis. Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation. In *International Conference on Learning Representations*, 2022.
- [81] David R. So, Wojciech Mańke, Hanxiao Liu, Zihang Dai, Noam Shazeer, and Quoc V. Le. Primer: Searching for efficient transformers for language modeling. *arXiv preprint arXiv:2109.08668*, 2022.
- [82] Zhengyan Zhang, Yixin Song, Guanghui Yu, Xu Han, Yankai Lin, Chaojun Xiao, Chenyang Song, Zhiyuan Liu, Zeyu Mi, and Maosong Sun. ReLU<sup>2</sup> wins: Discovering efficient activation functions for sparse LLMs. *arXiv preprint arXiv:2402.03804*, 2024.
- [83] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.
- [84] Shane Bergsma, Nolan Dey, Gurpreet Gosal, Gavia Gray, Daria Soboleva, and Joel Hestness. Straight to zero: Why linearly decaying the learning rate to zero works best for LLMs. In *International Conference on Learning Representations*, 2025.
- [85] Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R Steeves, Joel Hestness, and Nolan Dey. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama. <https://www.cerebras.net/blog/slimpajama-a-627b-token-cleaned-and-deduplicated-version-of-redpajama>, 2023. URL <https://huggingface.co/datasets/cerebras/SlimPajama-627B>.
- [86] Sam McCandlish, Jared Kaplan, Dario Amodei, and OpenAI Dota Team. An empirical model of large-batch training. *arXiv preprint arXiv:1812.06162*, 2018.
- [87] Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. DeepSeek LLM: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.
- [88] Tomer Porian, Mitchell Wortsman, Jenia Jitsev, Ludwig Schmidt, and Yair Carmon. Resolving discrepancies in compute-optimal scaling of language models. *arXiv preprint arXiv:2406.19146*, 2024.
- [89] Xi Wang and Laurence Aitchison. How to set AdamW’s weight decay as you scale model and dataset size. *arXiv preprint arXiv:2405.13698*, 2024.
- [90] Cerebras Systems. Linear Scaling Made Possible with Weight Streaming - Cerebras. <https://www.cerebras.ai/blog/linear-scaling-made-possible-with-weight-streaming>, September 2022.
- [91] Keivan Alizadeh, Iman Mirzadeh, Dmitry Belenko, Karen Khatamifard, Minsik Cho, Carlo C Del Mundo, Mohammad Rastegari, and Mehrdad Farajtabar. LLM in a flash: Efficient large language model inference with limited memory. *arXiv preprint arXiv:2312.11514*, 2024.
- [92] Ying Sheng, Lianmin Zheng, Binhang Yuan, Zhuohan Li, Max Ryabinin, Beidi Chen, Percy Liang, Christopher Ré, Ion Stoica, and Ce Zhang. FlexGen: High-throughput generative inference of large language models with a single GPU. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

- [93] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.
- [94] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.
- [95] Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016.
- [96] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. RACE: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*, 2017.
- [97] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. *arXiv preprint arXiv:1911.11641*, 2019.
- [98] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*, 2019.
- [99] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in Neural Information Processing Systems*, 32, 2019.
- [100] David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluís-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*, 2021.
- [101] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.
- [102] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. *arXiv preprint arXiv:1906.02243*, 2019.
- [103] Mitchell Wortsman, Peter J Liu, Lechao Xiao, Katie Everett, Alex Alemi, Ben Adlam, John D Co-Reyes, Izzeddin Gur, Abhishek Kumar, Roman Novak, Jeffrey Pennington, Jascha Sohl-dickstein, Kelvin Xu, Jaehoon Lee, Justin Gilmer, and Simon Kornblith. Small-scale proxies for large-scale transformer training instabilities. *arXiv preprint arXiv:2309.14322*, 2023.
- [104] Saaketh Narayan, Abhay Gupta, Mansheej Paul, and Davis Blalock.  $\mu$ nit scaling: Simple and scalable fp8 llm training. *arXiv preprint arXiv:2502.05967*, 2025.
- [105] Greg Yang and Etai Littwin. Tensor programs IVb: Adaptive optimization in the infinite-width limit. *arXiv preprint arXiv:2308.01814*, 2023.
- [106] Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac’h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 07 2024. URL <https://zenodo.org/records/12608602>.
- [107] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: All of the claims made in the abstract and introduction are supported by both theoretical and empirical results we present in Section 4, 5, and 6.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: In Section 7 we discuss the limitations of our work. In Section 6 and Appendix D we state our theoretical assumptions.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?



Answer: [Yes]

Justification: In Section 6 and Appendix D we state our theoretical assumptions.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: In Appendix G as well as Section 4 and Section 5, we disclose all the information needed to reproduce the main experimental results of the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

#### 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide a minimal open-source implementation of the parameterizations used in this work in our supplementary material. We also train all our models on SlimPajama [85], an open-source dataset.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: In Table 4 as well as Section 4 and Section 5, specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: In Table 2, along with task accuracy we report the standard error for each measurement. Other tests involve reporting upstream validation cross entropy loss on a single seed so we have no method of quantifying the variability.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: In Table 4 as well as Section 4 and Section 5, we provide extensive experimental details which can be used to easily derive compute requirements based on a reader's specific hardware setting.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: This work conforms to the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss the potential broader impacts of our work in Appendix A.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

#### 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We do not release any models or datasets.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

#### 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite and use all training and evaluation datasets in accordance with their licenses.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

### 13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The only new asset introduced in this work is a minimal code implementation of our parameterizations included in our supplementary material. This is accompanied by a "README.md" file containing clear documentation.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

### 14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: We do not conduct any research involving crowdsourcing or human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

### 15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: We do not conduct any research involving human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.



- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

#### 16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

## A Broader impacts

As LLM compute budgets grow, compute efficient training methods are becoming increasingly important for reducing carbon emissions [100] and offsetting environmental and financial costs of large model training [101]. This work presents methods which improve the compute efficiency of LLM training, especially for large deep models. There is also growing recognition that HP tuning is a key contributor to these costs. HP tuning is costly, possibly undermining equity in AI research due to financial resources [102]. During model retraining, sensitivity to HPs also leads to downstream costs [102]. This work presents methods which can reduce these costs and sensitivities and thus improve equity.

## B Coordinate check test for verification of stable training

To verify our theoretical expectations for training stability match empirical results, we perform “coordinate check” tests where we scale  $L$  and measure the change in activation size at the final residual addition  $h_L$  in Figure 7. Interestingly,  $\alpha = 0.5$  as described in the literature [6–8] did not achieve transformer training stability as we initially expected (2nd column). We reasoned this was due to a lack of consideration for how the bias and LayerNorm parameters can affect training stability. After adopting the bias and LayerNorm  $\eta$  adjustments proposed in Table 1 and Appendix D, we achieved training stability with  $\alpha = 0.5$  (3rd column). Figure 7 shows both  $\alpha \in \{0.5, 1\}$  achieve stability for any depth. This test is a necessary but not sufficient condition for satisfying Desiderata 2.

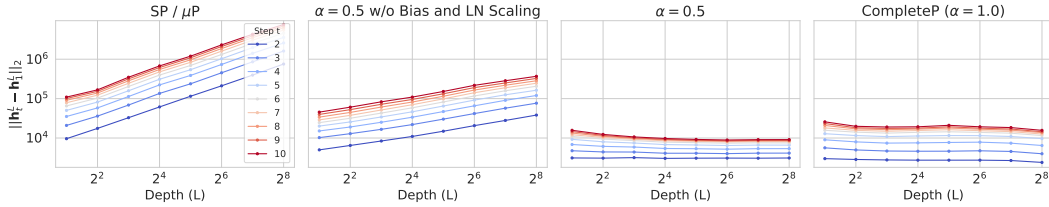


Figure 7: Coordinate check. Frobenius norm of activations after merged residual streams from MHSA and MLP blocks across models of increasing depths after training step  $t$ . 1st column: SP/ $\mu$ P without depth control. 2nd column:  $\alpha = 0.5$  as described in literature [6–8]. 3rd column:  $\alpha = 0.5$  (Table 1). 4th column:  $\alpha = 1.0$ .  $\alpha \in \{0.5, 1\}$  achieve stability for any depth.

## C Scalings that Guarantee Our Desiderata

In the following analysis, we argue why  $\alpha \geq \frac{1}{2}$  is necessary for Desideratum 1 (Stable Initialization) and we argue for AdamW LR to be  $\eta = L^{\alpha-1}$  for each residual block to achieve Desideratum 2 (Maximal Residual Stream Update).

### C.1 Desideratum 1: Stable Initialization

In this section, we demonstrate that  $\alpha \geq \frac{1}{2}$  is necessary for stable signal propagation at initialization. For concreteness in the simplest model that captures this effect, we consider a residual MLP network with block depth 1, the model analyzed in [6],

$$\mathbf{h}^{\ell+1} = \mathbf{h}^\ell + \frac{1}{L^\alpha} \mathbf{W}^\ell \phi(\mathbf{h}^\ell). \quad (8)$$

The weights are initialized with  $W_{ij}^\ell \sim \mathcal{N}(0, \frac{\sigma_W^2}{N})$ . Defining  $H^\ell \equiv \frac{1}{N} |\mathbf{h}^\ell|^2$  as the norm of the residual stream variables, we have the following recursion for  $H^\ell$  at large  $N$

$$H^{\ell+1} = H^\ell + L^{-2\alpha} \sigma_W^2 \mathbb{E}_{h \sim \mathcal{N}(0, H^\ell)} \phi(h)^2, \quad (9)$$

where the average  $\mathbb{E}_h[\cdot]$  represents an average over the hidden neurons at residual layer  $\ell$  [6]. Taking for concreteness  $\phi(h) = \text{ReLU}(h)$ , the above equation gives

$$H^{\ell+1} = H^\ell + \frac{\sigma_W^2}{2} L^{-2\alpha} H^\ell = \left[ 1 + \frac{\sigma_W^2}{2} L^{-2\alpha} \right]^\ell H^0 \quad (10)$$

where  $H^0$  is the scale of the first residual variable. The last layer variance therefore has the following large depth limit

$$\lim_{L \rightarrow \infty} H^L = \begin{cases} H^0 & \alpha > \frac{1}{2} \\ \exp\left(\frac{\sigma_W^2}{2}\right) H^0 & \alpha = \frac{1}{2} \\ \infty & \alpha < \frac{1}{2} \end{cases}. \quad (11)$$

We see that  $\alpha \geq \frac{1}{2}$  is necessary for the residual stream variance to converge to a finite value as  $L \rightarrow \infty$ . For  $\alpha = \frac{1}{2}$  the residual stream covariance depends on the activation function and the initialization variance  $\sigma_W^2$  while for the  $\alpha > \frac{1}{2}$ , the signal propagation becomes trivial in the infinite depth limit as the contribution from the nonlinearity and the initial weight variance disappears at large  $L$ .

## C.2 Desideratum 2: Maximal Update

In this section, we consider the Adam learning rate necessary to achieve  $\Theta(1)$  changes to the residual stream variables. First, we note that the backward pass variables  $\mathbf{g}^\ell \equiv \frac{\partial \mathcal{L}}{\partial \mathbf{h}^\ell}$  are strongly correlated across layers since, from the chain rule,

$$\mathbf{g}^\ell = \mathbf{g}^{\ell+1} + \frac{1}{L^\alpha} \dot{\phi}(\mathbf{h}^\ell) \odot \left[ (\mathbf{W}^\ell)^\top \mathbf{g}^{\ell+1} \right]. \quad (12)$$

We therefore note that any weighted average of over layers is  $\Theta_L(1)$ , such as

$$\frac{1}{L} \sum_{\ell=1}^L C^\ell h_i^\ell = \Theta_L(1), \quad \frac{1}{L} \sum_{\ell=1}^L C^\ell g_i^\ell = \Theta_L(1), \quad (13)$$

where  $C^\ell$  is an arbitrary sequence with increments of scale  $C^{\ell+1} - C^\ell = \Theta_L(L^{-1})$ . Consider now the update under a sign-GD update

$$W_{ij}^\ell(t+1) = W_{ij}^\ell(t) + \eta^\ell \frac{1}{Z_{ij}^\ell} g_i^{\ell+1} \phi(h_j^\ell), \quad Z_{ij}^\ell = \sqrt{(g_i^{\ell+1})^2 \phi(h_j^\ell)^2} \quad (14)$$

where  $\eta^\ell$  is the learning rate for this hidden layer. Note that AdamW will contain the same scaling (with  $N, L$ ) for the numerator and denominator but will also contain momentum (which does not change the scaling with  $N, L$ ) and an additional  $\epsilon$  parameter which we discuss how to scale with  $N, L$  in Appendix D.5.

We now discuss how to choose  $\eta^\ell$  to obtain  $\Theta_{N,L}(1)$  updates to the residual stream (Desideratum 2). First, we can express the residual variables at layer  $\ell + 1$  at step  $t$  of training as

$$\begin{aligned} \mathbf{h}^{\ell+1}(t) &= \mathbf{h}^\ell(t) + L^{-\alpha} \mathbf{W}^\ell(0) \phi(\mathbf{h}^\ell(t)) \\ &+ L^{-\alpha} \sum_{t' < t} \left[ \frac{\eta^\ell}{Z_{ij}^\ell(t')} \odot \mathbf{g}^{\ell+1}(t') \phi(\mathbf{h}^\ell(t'))^\top \right] \phi(\mathbf{h}^\ell(t)) \end{aligned} \quad (15)$$

The entries of the above vector from the update to  $\mathbf{W}^\ell(t)$  have the following scaling with  $N$  and  $L$

$$v_i^\ell(t, t') \equiv \frac{1}{N} \sum_{j=1}^N \frac{g_i^{\ell+1}(t') \phi(h_j^\ell(t'))}{Z_{ij}^\ell(t')} \phi(h_j^\ell(t)) = \Theta_{N,L}(1), \quad (16)$$

as the variables  $\phi(h_j^\ell(t'))$  and  $\phi(h_j^\ell(t))$  are random variables with  $\Theta(1)$  correlation. We desire the final layer of the residual stream to achieve a  $\Theta(1)$  updates

$$\mathbf{h}^L(t) = \mathbf{h}^0(t) + \underbrace{L^{-\alpha} \sum_{\ell=0}^{L-1} \mathbf{W}^\ell(0) \phi(\mathbf{h}^\ell(t))}_{\text{Stable by D1 for } \alpha \geq 1/2} + \underbrace{L^{-\alpha} N \sum_{\ell=0}^{L-1} \eta^\ell \sum_{t' < t} \mathbf{v}^\ell(t, t')}_{\text{Desired scale} = \Theta(1)} \quad (17)$$

Since  $\alpha \geq \frac{1}{2}$ , the stability of the middle term is guaranteed from the D1 analysis. The  $\mathbf{v}^\ell$  variables are strongly correlated across layers, so a sum over all  $L$  layers will give a result of order  $\Theta(L)$ . To make this final term  $\Theta_{N,L}(1)$  we must choose a learning rate of scale

$$\eta^\ell = \eta_{\text{base}} N^{-1} L^{\alpha-1}. \quad (18)$$

This choice causes the final term above in the  $\mathbf{h}^L(t)$  formula to be  $\Theta(1)$ , resulting in a  $\Theta(1)$  change to the residual variables due to the weight updates as desired.

## D Derivation for bias, LN, WD, and AdamW $\epsilon$ parameterization adjustments

### D.1 On Generalizability of the Parameterization

In the later parts of this subsection, we will discuss derivations of parameterizations for several different components of the architecture via a heuristic method, which we find very accurately represents the scaling important quantities. However, this remains limited as there are other architecture components that are not contained by our manuscript, which may require further adjustments and derivations.

With this in mind, we can divide up the future generalizations into three categories:

1. No modifications required. This includes LayerNorm at any position (pre, post, QK-norm etc.), adding biases at different locations, LR schedule, and other well-normalized Adam-like algorithms (e.g. SignGD, Adagrad, Lion), or position embeddings (learned, RoPE, ALiBi, NoPE, etc.). The reason is that these do not change the scale of both the forward pass and hidden layer updates due to training, with respect to width and depth. LayerNorm, for example, maintains the  $\Theta(1)$  scale of each neuron, and as long as the update maintains the  $\Theta(1)$  scale, then no changes are required.
2. Slight modifications are required and known. A good example of this is SGD, where the learning rate scaling is derived in Table 1 of Bordelon et al. [8] (up to an equivalent ABC-reparameterization). Notice here that compared to Adam, the learning rate of SGD needs to be scaled up due to the prefactors in front of both the weight matrices and the residual branch, where as these extra factors are canceled out in Adam due to normalization.
3. More theoretical derivations required. This includes MoE, long context, batch size (and schedule), gradient clipping, LAMB, and momentum. Some of these are a relatively straight forward exercise to derive (e.g. LAMB and other optimization algorithms), others are less straight forward. In particular, any scaling that requires increasing the number of data points and training steps along with width and depth remains theoretically challenging. Our approach in this work computes the theoretical scaling for finitely many data points and training steps, and testing whether or not hyperparameter transfer empirically, which has been yielding very good results (see e.g. Figure 3).

For the third category, we would effectively need to derive the scale of the updates and the downstream effects for each new architectural modification. A good example is consider the effect of bias weights in the next subsection. Here biases  $\mathbf{b}^\ell$  are updated such that it contributes the same amount of changes to  $\mathbf{h}^{\ell+1}$  as the usual weights  $\mathbf{W}^\ell$ . In order to do that, we need to choose the appropriate initialization and learning scale to satisfy this requirement. All other modifications follow the same mechanism, and the rest of this section serve as an example derivation for all future generalizations.

### D.2 Bias Parameters

In this section, we consider the effect of a bias parameter. To start, we will illustrate the effect of bias parameters in a depth 1 MLP residual block

$$f = \frac{1}{N} \mathbf{W}^L \phi(\mathbf{h}^L), \mathbf{h}^{\ell+1} = \mathbf{h}^\ell + \frac{1}{L^\alpha} [\mathbf{W}^\ell \phi(\mathbf{h}^\ell) + \mathbf{b}^\ell], \mathbf{h}^1 = \mathbf{W}^0 \mathbf{x} \quad (19)$$

The entries of the bias vectors  $\mathbf{b}^\ell$  are initialized with unit variance  $\mathbf{b}^\ell \sim \mathcal{N}(0, \mathbf{I})$  for all layers while  $\mathbf{W}^\ell$  entries are  $\mathcal{N}(0, \frac{1}{N})$  for intermediate layers  $\ell \in \{1, \dots, L-1\}$  and  $\mathcal{N}(0, 1)$  for encoder and decoder layers  $\ell \in \{0, L\}$ . With this choice, stable signal propagation is guaranteed for  $\alpha \geq \frac{1}{2}$ . To obtain the desired  $\frac{1}{L}$  increment in the residual stream, the weight and bias parameters must be updated with scale

$$\Delta W_{ij}^\ell = \Theta(N^{-1} L^{\alpha-1}), \Delta b_i^\ell = \Theta(L^{\alpha-1}), \ell \in \{1, \dots, L-1\} \quad (20)$$

For the Adam optimizer this directly sets the desired scaling of learning rate for each of these parameters

$$\eta_{W^\ell} = \Theta(N^{-1} L^{\alpha-1}), \eta_{b^\ell} = \Theta(L^{\alpha-1}). \quad (21)$$

This argument extends to more complicated blocks such as self attention layers and deeper MLPs. However, for the first and last layer of the transformer which we want to achieve  $\Theta(1)$  changes to their block outputs we need to set

$$\Delta W_{ij}^\ell = \Theta(1), \Delta b_i^\ell = \Theta(1), \ell \in \{0, L\}. \quad (22)$$

### D.3 LayerNorm

While a static LayerNorm operation was examined in [8], most LayerNorm operations include trainable gain and bias parameters for each neuron.

$$\text{LN}^\ell(\mathbf{h}^\ell) = \frac{1}{\sqrt{\sigma_h^2 + \epsilon}} [\mathbf{h}^\ell - \mu_h \mathbf{1}] \odot \mathbf{g}^\ell + \mathbf{b}^\ell \quad (23)$$

where  $\mu_h$  and  $\sigma_h^2$  are the mean and variance of the entries of  $\mathbf{h}^\ell$ . For any hidden block in the residual stream, the entries of  $\mathbf{g}$  and  $\mathbf{b}$  must be updated with

$$\forall \ell \in \{1, \dots, L-1\}, \Delta g_i = \Theta(L^{\alpha-1}), \Delta b_i^\ell = \Theta(L^{\alpha-1}). \quad (24)$$

while for the first and last layer (before and after the residual stream),

$$\Delta g_i = \Theta(L^{\alpha-1}), \Delta b_i^\ell = \Theta(L^{\alpha-1}), \ell \in \{0, L\}. \quad (25)$$

### D.4 Weight Decay

In this section, we will perform a heuristic calculation the weight decay scaling required in AdamW, which we will approximate with a SignGD-like update of the type

$$\mathbf{W}^\ell(t+1) = \mathbf{W}^\ell(t) - \eta \frac{1}{\mathbf{Z}^\ell} \odot \nabla_{\mathbf{W}^\ell} L(\theta(t)) - \eta \lambda \mathbf{W}^\ell(t), \quad (26)$$

where  $\mathbf{Z}^\ell$  is a collection of constant such that  $\frac{1}{\mathbf{Z}^\ell} \nabla_{\mathbf{W}^\ell} L(\theta(t))$  have  $\Theta(1)$  entries with respect to width  $N$  and depth  $L$ ,  $\lambda > 0$  is the weight decay parameter.

Suppose in either an MLP or inside a residual block, we have a fully connected layer (including the cases for  $Q, K, V$  blocks)

$$\mathbf{h}^{\ell+1} = \mathbf{W}^\ell \phi(\mathbf{h}^\ell), \quad (27)$$

where  $\mathbf{W}^\ell \in \mathbb{R}^{N \times N}$  is initialized with  $\mathcal{N}(0, \sigma_{\text{base}}^2/N)$ .

Suppose we want  $\mathbf{h}^{\ell+1}$  to change by order  $\Theta(1)$  in an SignGD update with one data point, then we get that

$$\mathbf{W}^\ell(t+1) = \mathbf{W}^\ell(t) - \frac{\eta}{\mathbf{Z}^\ell} \odot \mathbf{g}^{\ell+1}(t) \phi(\mathbf{h}^\ell(t))^\top - \eta \lambda \mathbf{W}^\ell(t), \quad (28)$$

where  $\mathbf{Z} \in \mathbb{R}^{N \times N}$  is once again the normalization constants, and  $\mathbf{g}^{\ell+1}$  is the backward pass variable [32, Equation 16]. However, in this case we have that  $g$  and  $\phi(h)$  are already  $\Theta(1)$  in the previous iterate, so  $\mathbf{Z}$  is itself  $\Theta(1)$ .

Next we will calculate  $\Delta \mathbf{h}^{\ell+1}(t+1) = \mathbf{h}^{\ell+1}(t+1) - \mathbf{h}^{\ell+1}(t)$  to get

$$\Delta \mathbf{h}^{\ell+1} = \Delta \mathbf{W}^\ell \phi(\mathbf{h}^\ell) + \mathbf{W}^\ell \Delta \phi(\mathbf{h}^\ell), \quad (29)$$

where we abuse notations slightly and drop the  $t$  indices when clear. Here we will also assume as induction hypothesis that  $\Delta \phi(\mathbf{h}^\ell)$  is well behaved so we will focus on the first term only, which is

$$\begin{aligned} \Delta \mathbf{W}^\ell \phi(\mathbf{h}^\ell) &= -\eta \left[ \frac{1}{\mathbf{Z}^\ell} \odot \mathbf{g}^{\ell+1} \phi(\mathbf{h}^\ell)^\top \right] \phi(\mathbf{h}^\ell) - \eta \lambda \mathbf{W}^\ell \phi(\mathbf{h}^\ell) \\ &= -\eta \left[ \frac{1}{\mathbf{Z}^\ell} \odot \mathbf{g}^{\ell+1} \phi(\mathbf{h}^\ell)^\top \right] \phi(\mathbf{h}^\ell) - \eta \lambda \underbrace{\mathbf{h}^{\ell+1}}_{\Theta(1)}. \end{aligned} \quad (30)$$

where the entries of  $\mathbf{h}^{\ell+1}$  are  $\Theta_{N,L}(1)$  by the arguments of Appendix C.1. Following the arguments of Appendix C.2, the learning rate  $\eta$  must be set as

$$\eta = \eta_{\text{base}} \frac{L^{\alpha-1}}{N}, \quad (31)$$

to ensure the correct change to the residual stream variables. Since we desire the final term to be on the same scale this means that  $\lambda = \Theta_{N,L}(1)$ . Our proposed in weight decay scaling rule (Table 1) is also used in a few recent works [89, 103, 104]. The contribution of this section is to show no additional depth correction to the weight decay parameter  $\lambda$  is required.



## D.5 Adam $\epsilon$ Parameter

In principle, the  $\epsilon$  parameter for Adam may need to be scaled with width or depth to obtain stable behavior. The key problem can be seen from the update for a parameter  $\theta$

$$\Delta\theta^\ell = \frac{\eta^\ell}{\sqrt{v_t^\ell + \epsilon^\ell}} m_t^\ell \quad (32)$$

where  $v_t$  is a moving average of squared gradients and  $m_t$  is a moving average of gradients. Prior work outlined various strategies to achieve this stability as width is scaled [105, 38]. We will now describe strategies to achieve of the  $\epsilon$  parameter across depths.

*Layer-wise  $\epsilon, \eta$ :* One strategy is to utilize the current parameterization provided in the main text and to modify the  $\epsilon$  parameter and learning rate for each type of layer. The parameterization is

$$f = \frac{1}{N} \mathbf{W}^L \phi(\mathbf{h}^L), \mathbf{h}^{\ell+1} = \mathbf{h}^\ell + \frac{1}{L^\alpha} \mathbf{W}^\ell \phi(\mathbf{h}^\ell), \mathbf{h}^1 = \mathbf{W}^0 \mathbf{x}, \quad (33)$$

where the hidden weights  $W_{ij}^\ell$  are initialized with random entries from

$$W_{ij}^\ell(0) \sim \mathcal{N}(0, N^{-1}), \ell \in \{1, \dots, L-1\} \quad (34)$$

and the first and last layer are initialized with unit variance

$$W_{ij}^0(0) \sim \mathcal{N}(0, 1), W_{ij}^L(0) \sim \mathcal{N}(0, 1) \quad (35)$$

We note that the  $m_t$  and  $v_t$  variables have the following scalings in the residual stream

$$m_t^\ell = \Theta(L^{-\alpha} N^{-1}), \sqrt{v_t^\ell} = \Theta(L^{-\alpha} N^{-1}), \ell \in \{1, \dots, L-1\} \quad (36)$$

$$m_t^\ell = \Theta(N^{-1}), \sqrt{v_t^\ell} = \Theta(N^{-1}), \ell \in \{0, L\} \quad (37)$$

This parameterization demands the following learning rate for hidden layers for an  $\epsilon \rightarrow 0$  limit to be stable

$$\eta^\ell = \eta_0 N^{-1} L^{\alpha-1}, \ell \in \{1, \dots, L-1\} \quad (38)$$

$$\eta^\ell = \eta_0, \ell \in \{0, L\} \quad (39)$$

Now, to consider the effect of  $\epsilon$ , we desire it to match the scale of the raw gradients  $m_t$  so we take

$$\begin{aligned} \epsilon^\ell &= \epsilon_0 L^{-\alpha} N^{-1}. \\ \epsilon^\ell &= \epsilon_0 N^{-1}, \ell \in \{0, L\}. \end{aligned} \quad (40)$$

This will ensure that  $\epsilon$  is of the same order as  $\sqrt{v_t}$  for all hidden layers. For the read-in and readout we can pre-multiply gradients by a constant.

*Reparameterize the Definition of the Model:* Another strategy that one could adopt is to reparameterize the model so that a single value of  $\epsilon$  can be used for every layer. In this case,

$$f = \frac{1}{NL^\alpha} \mathbf{W}^L \phi(\mathbf{h}^L), \mathbf{h}^{\ell+1} = \mathbf{h}^\ell + \frac{1}{L^\alpha} \mathbf{W}^\ell \phi(\mathbf{h}^\ell), \mathbf{h}^1 = \frac{1}{L^\alpha} \mathbf{W}^0 \mathbf{x}, \quad (41)$$

where the readin and readout weights are initialized as

$$W_{ij}^0(0) \sim \mathcal{N}(0, L^{2\alpha}), W_{ij}^L(0) \sim \mathcal{N}(0, L^{2\alpha}). \quad (42)$$

The global  $\epsilon$  parameter can now be set as

$$\epsilon = \epsilon_0 L^{-\alpha} N^{-1}. \quad (43)$$

The learning rates must be set as

$$\eta^\ell = \eta_0 L^\alpha, \ell \in \{0, L\} \quad (44)$$

$$\eta^\ell = \eta_0 L^{\alpha-1} N^{-1}, \ell \in \{1, \dots, L-1\}. \quad (45)$$

These rules will ensure the correct scale updates in each layer.

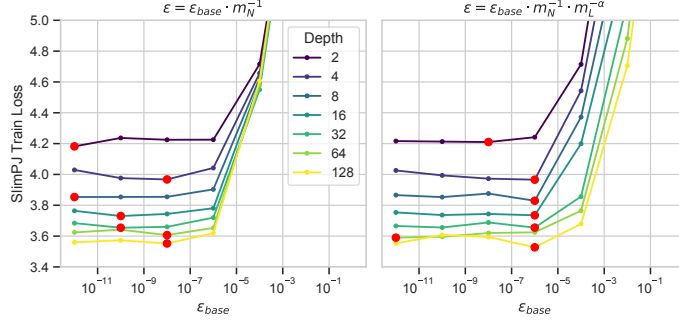


Figure 8: Using ComplteP ( $\alpha = 1$ ), we ablate the effect of the depth-wise Adam  $\epsilon$  scaling rule. The scaling rule enables a wider range of  $\epsilon_{\text{base}}$  to achieve competitive loss.

## E Depth-wise transfer of AdamW $\epsilon$

Everett et al. [38] show appropriate AdamW  $\epsilon$  scaling is important for compute-efficient large models. Building on the width adjustment for AdamW  $\epsilon$  introduced by Yang and Littwin [105], we add a depth correction for AdamW  $\epsilon$ . Figure 8 shows this correction enables stable training for a wider range of  $\epsilon_{\text{base}}$ .

## F Depth Scaling Training Dynamics

We illustrate the training dynamics in the form of loss curves in Figure 9 and 10, which slice the  $\eta$  transfer test in Figure 2 at CompleteP ( $\alpha = 1$ )’s and  $\alpha = 0.5$ ’s optimal  $\eta$  respectively. CompleteP ( $\alpha = 1$ ) consistently achieves lower training loss when model depth scales up. In  $\alpha = 0.5$ , we point out the loss "crossing" behavior where deeper models continually take longer steps to scale better than shallower models. This aligns with our observation in Figure 2 where  $\alpha = 0.5$  desires larger  $\eta$  with increasing model depth and laziness (defined in Section 6).

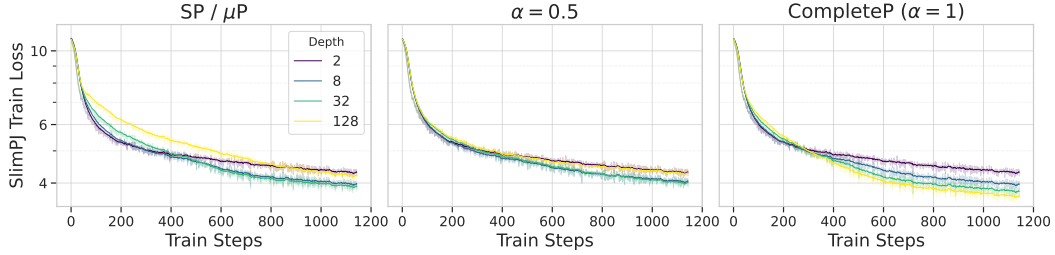


Figure 9: Training dynamics of Figure 2 for all parameterizations at CompleteP ( $\alpha = 1$ )’s optimal  $\eta = 2^{-8}$ .

## G Additional experimental details

Table 4 contains extensive details for all experiments. The rest of this section describes the remaining experimental details.

### G.1 Compute-efficient setup

In Figure 4, to study the compute-optimal setting, first our training setup must represent compute-optimal. Following [3], we train models for 20 tokens per parameter (TPP) to ensure a compute-optimal tradeoff of parameters and tokens. We also scale our batch size according to a power law in

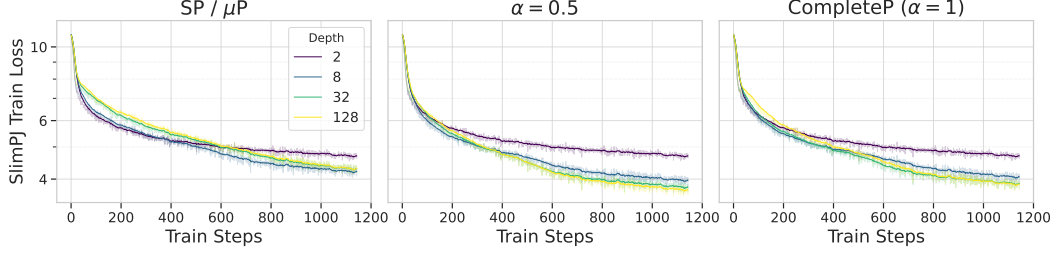


Figure 10: Training dynamics of Figure 2 for all parameterizations at  $\eta = 2^{-6}$ , where  $\alpha = 0.5$  first demonstrates deeper-model-lower-loss ordering.

Table 4: Additional experimental details. Cells marked as “vary” mean the quantity is varied.

Figure	$N$	$L$	$\sigma_{\text{init}}$	$\eta_{\text{base}}$	$\lambda_{\text{base}}$	$T_{\text{EMA}}$	$B$	Steps	Tokens	TPP
1 left, 2 top, 9, 10	256	Vary	0.02	Vary	0	N/A	128	1144	300M	Vary
2 bottom	256	Vary	Vary	2E-08	0	N/A	128	1144	300M	Vary
3	256	Vary	0.02	Vary	Vary	0.1407	Eqn. 46	Vary	Vary	20
4, 13, 1 middle & right	Vary	Vary	0.02	2E-08	Vary	0.1407	Eqn. 46	Vary	Vary	20
7	256	Vary	0.06	2E-03	0	N/A	4	10	82K	Vary
8	256	Vary	0.02	2E-08	0	N/A	128	1144	300M	Vary
11	Vary	6	0.09	Vary	Vary	Vary	Eqn. 46	Vary	Vary	20
12	Vary	Vary	0.02	2E-08	Vary	Vary	Eqn. 46	Vary	Vary	20
Table 3	Vary	Vary	0.02	2E-08	Vary	0.1407	Eqn. 46	Vary	Vary	200

FLOPS. Since, McCandlish et al. [86], Kaplan et al. [2] showed the critical batch size scales with a power law in loss, and loss scales with a power law in FLOPS, the critical batch size scales with a power law in FLOPS [87, 88]. We follow Equation 46 based on empirical fits and rounded to the nearest multiple of 8, where  $B, F$  are batch size and training FLOPS respectively.

$$B = \max(32, 0.7857 \cdot C^{0.1527} - 306.8) \quad (46)$$

We adopt the  $\sigma_{\text{base}} = 0.02$  and the optimal  $\eta_{\text{base}}$  for  $L = L_{\text{base}} = 2$  from Figure 2:  $\eta_{\text{base}} = 0.0039$ .

To ensure optimal weight decay  $\lambda$  as we scale model and dataset size, we follow Wang and Aitchison [89] by setting  $\lambda_{\text{base}}$  to maintain the optimal  $\tau_{\text{EMA}} = (\eta_{\text{base}} \lambda_{\text{base}} n_{\text{steps}})^{-1}$ . A major shortcoming of  $\tau_{\text{EMA}}$  is that it does not take the learning rate schedule into account, even though learning rates are decayed in practice. Despite this we observe reliable transfer of  $\tau_{\text{EMA}}$  across model sizes, provided we use the same learning rate schedule. For SP and  $\mu\text{P}$ , Figure 11 validates that the claims of Wang and Aitchison [89] that the optimal  $\tau_{\text{EMA}}$  remains stable across different learning rates and weight decay values. We then tune  $\tau_{\text{EMA}}$  for SP,  $\mu\text{P}$ ,  $\alpha = 0.5$ , and CompleteP ( $\alpha = 1$ ). Figure 12 shows  $\tau_{\text{EMA}} = 0.1407$  close to optimal for all parameterizations and this adopted throughout this paper. When a non-zero weight decay is adopted, we use  $\tau_{\text{ema}}, \eta_{\text{base}}, n_{\text{steps}}$  to decide the  $\lambda_{\text{base}}$  used in each run.

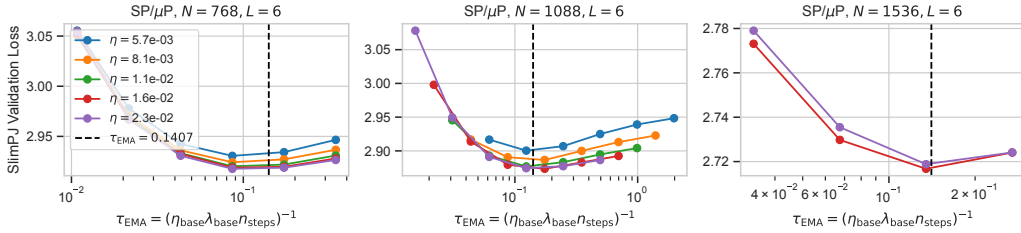


Figure 11: When grid searching learning rate  $\eta_{\text{base}}$  and weight decay  $\lambda_{\text{base}}$ , the optimal  $\tau_{\text{EMA}}$  remains stable.

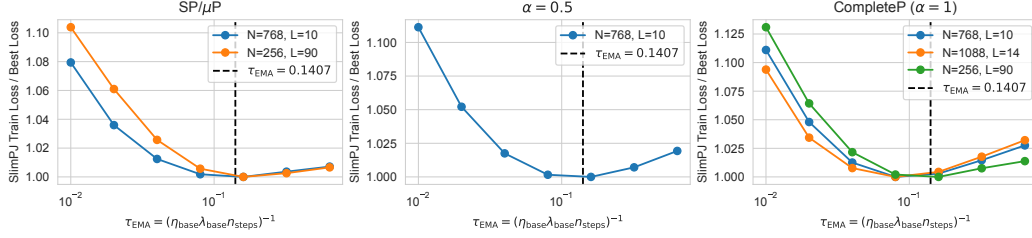


Figure 12:  $\tau_{EMA}$  sweep for  $\mu P$  and  $\alpha \in \{0.5, 1\}$ .  $\tau_{EMA}$  changes minimally across parameterizations. The  $\tau_{EMA}$  values we empirically chose sit in the bowl of optimal loss values.  $\tau_{EMA}$  for  $\alpha = 0.5$  seems good but for  $\mu P$  and  $\alpha = 1$  it seems slightly off. Looks like aspect ratio affects the optimal  $\tau_{EMA}$ .

## G.2 N:L study details

Here we include additional experimental details for Figure 4. We vary  $N, L$  to approximately maintain  $P_{non-emb} = 12N^2L$ . Table 5 contains all model shapes, token counts, and FLOPs used in Figure 4. Approximations for training FLOPs like  $F = 6ND$  do not account for embedding, attention, LN, nonlinearity or bias FLOPs. We use a very granular function to measure FLOPs. Although we maintain  $P_{non-emb}$ , the FLOPs are not equal mainly because of embedding and attention costs. We train  $P_{non-emb} \in \{50M, 300M, 1.5B\}$  models, matching sizes from [2], but all at 20 TPP. [2] used a similar study design, but with the same number of tokens for all runs, meaning it didn’t follow the compute-optimal prescription of [3].

Table 5: Details of model sizes used in the aspect ratio study. The compute-optimal configurations for  $\mu P$  and  $\alpha = 1$  are **bolded**. The compute-optimal configurations for  $\alpha = 0.5$  are *italicized*

Label	$P_{non-emb}$ (M)	$P_{total}$ (M)	Tokens (B)	Train FLOPS	Steps	B	N	L	N:L
50M	49.8	75.5	1.5	1.25E+18	4849	152	256	63	4.1
50M	49.3	81.5	1.6	1.26E+18	5235	152	320	40	8.0
50M	49.7	88.3	1.8	1.34E+18	5671	152	384	28	13.7
<b>50M</b>	<b>50.4</b>	<b>101.9</b>	<b>2.0</b>	<b>1.56E+18</b>	<b>5923</b>	<b>168</b>	<b>512</b>	<b>16</b>	<b>32.0</b>
<i>50M</i>	<i>49.2</i>	<i>113.6</i>	<i>2.3</i>	<i>1.76E+18</i>	<i>6301</i>	<i>176</i>	<i>640</i>	<i>10</i>	<i>64.0</i>
50M	49.6	126.8	2.5	2.07E+18	6730	184	768	7	109.7
50M	48.2	138.3	2.8	2.35E+18	7033	192	896	5	179.2
50M	50.4	153.3	3.1	2.82E+18	7198	208	1024	4	256.0
50M	47.8	163.6	3.3	3.11E+18	7397	216	1152	3	384.0
50M	47.6	189.1	3.8	4.02E+18	7696	240	1408	2	704.0
<hr/>									
300M	301.8	346.8	6.9	2.38E+19	8301	408	448	125	3.6
300M	305.3	369.6	7.4	2.32E+19	8846	408	640	62	10.3
300M	299.4	383.1	7.7	2.27E+19	9352	400	832	36	23.1
<b>300M</b>	<b>302.3</b>	<b>405.2</b>	<b>8.1</b>	<b>2.38E+19</b>	<b>9699</b>	<b>408</b>	<b>1024</b>	<b>24</b>	<b>42.7</b>
<i>300M</i>	<i>314.8</i>	<i>443.5</i>	<i>8.9</i>	<i>2.70E+19</i>	<i>10214</i>	<i>424</i>	<i>1280</i>	<i>16</i>	<i>80.0</i>
300M	307.4	468.2	9.4	2.85E+19	10784	424	1600	10	160.0
300M	302.1	508.0	10.2	3.20E+19	11275	440	2048	6	341.3
300M	298.7	588.2	11.8	4.06E+19	12169	472	2880	3	960.0
<hr/>									
1.5B	1488.8	1572.5	31.4	4.10E+20	19195	800	832	179	4.6
1.5B	1498.4	1614.2	32.3	3.96E+20	19903	792	1152	94	12.3
1.5B	1501.6	1656.0	33.1	3.91E+20	20418	792	1536	53	29.0
<b>1.5B</b>	<b>1512.3</b>	<b>1711.8</b>	<b>34.2</b>	<b>3.99E+20</b>	<b>21106</b>	<b>792</b>	<b>1984</b>	<b>32</b>	<b>62.0</b>
<i>1.5B</i>	<i>1487.9</i>	<i>1751.6</i>	<i>35.0</i>	<i>4.00E+20</i>	<i>21597</i>	<i>792</i>	<i>2624</i>	<i>18</i>	<i>145.8</i>
1.5B	1487.3	1841.1	36.8	4.26E+20	22474	800	3520	10	352.0
1.5B	1487.0	1943.8	38.9	4.62E+20	23262	816	4544	6	757.3

For each  $(F, X)$  point in Figure 4d, we calculate the “Loss increase vs.  $\hat{X}_{\alpha=1}$ ” as  $d = \frac{X - \hat{X}_{\alpha=1}}{\hat{X}}$  and plot it in Figure 4a-c. To understand how the FLOP savings in Figure 4e are calculated, first recall the functional form of  $\hat{X}(F) = (F/a)^{-b}$ . Inverting yields  $F(X) = aX^{-\frac{1}{b}}$ . Due to the scale invariance

of power laws, we can calculate:

$$\text{FLOP savings vs. } \mu P = 1 - (1 - (d_{\mu P} - d_{\alpha=1}))^{-1/b} \quad (47)$$

Figure 13 shows how Figure 4f was created by fitting cubic functions to data  $\{x = \log(N : L), y = d - d_{\alpha=1}^{\text{optimal}}\}$  and finding intersections with  $y = 1$ . Cubic functions were used instead of quadratics to enable better fits to functions without an axis of symmetry.

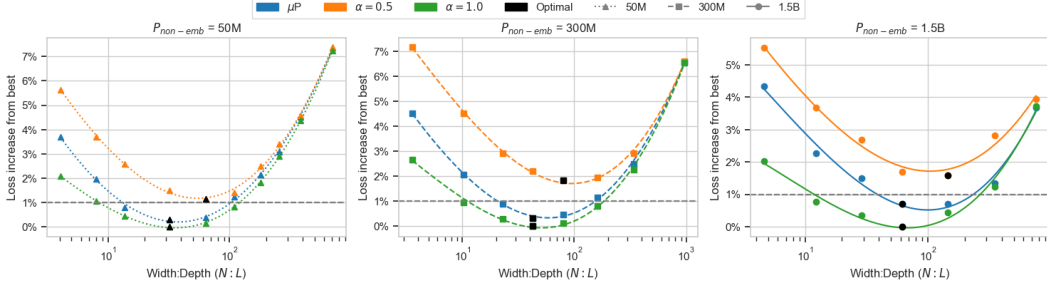


Figure 13: Plotted loss increase from best point. Fit cubic functions to log-transformed data. Found intersections with  $y = 1$  for Figure 4f.

### G.3 Downstream evaluation details

We evaluated our models on downstream tasks using Eleuther Harness Evaluation (EEH) [106]. The evaluation details are reported in Table 6.

Table 6: Details on downstream tasks evaluations. **acc** stands for accuracy. **acc\_norm** stands for accuracy normalized by target string’s byte-length.

Evaluation	EEH Task Name	Metric
Hellaswag	hellaswag	acc_norm
ARC-easy	arc_easy	acc_norm
LAMBADA	lambada_openai	acc
RACE	race	acc
PIQA	piqa	acc_norm
BoolQ	boolq	acc

### G.4 Figure 6 details

To visualize the notion of laziness in (4):

$$\frac{\Delta_{\theta} \mathbf{h}^{\ell} - \Delta_{\theta} [\mathbf{h}^{\ell}]^{\text{lin}, \theta}}{\Delta_{\theta} [\mathbf{h}^{\ell}]^{\text{lin}, \theta}} = o(1), \quad \text{as } N, L \rightarrow \infty, \quad (48)$$

we consider the toy architecture of a residual network with two layers per residual branch:

$$\mathbf{h}^{\ell+1} = \mathbf{h}^{\ell} + L^{-\alpha} \mathbf{W}_{(2)}^{\ell} \mathbf{W}_{(1)}^{\ell} \mathbf{h}^{\ell}, \quad \ell = \{1, \dots, L\}, \quad (49)$$

with width  $N = 256$ , and perform the learning rate updates according to our parameterizations, i.e.:

$$\eta = \begin{cases} \eta_0 L^{\alpha-1} & \alpha \in [1/2, 1] \\ \mu P. & \end{cases} \quad (50)$$

We then perform a single gradient step only with respect to  $\mathbf{W}_{(2)}^{\ell}$  for a fixed layer  $\ell$ , and calculate the left hand side of the laziness equation above for varying depths across 50 seeds per run. We report the median with lines and shade the region between the first and third quartiles. We compute  $\Delta_{\theta} \mathbf{h}_{\ell}^{\text{lin}, \theta}$  explicitly by differentiating the block with respect to  $\mathbf{W}_{(2)}^{\ell}$  and evaluate at the initial parameter values, i.e. we compute:

$$\mathbf{h}^{\text{lin}, \theta}(\theta, \theta_0) = \mathbf{h}(\theta_0) + \langle \nabla_{\theta} \mathbf{h}(\theta) |_{\theta_0}, \theta - \theta_0 \rangle,$$

where  $\theta = \mathbf{W}_{(2)}^\ell$ . We use  $\eta_0 = 0.0001$  and train on MNIST [107]. We provide the code to reproduce this plot in supplementary material.