

# TIME TRANSFER: ON OPTIMAL LEARNING RATE AND BATCH SIZE IN THE INFINITE DATA LIMIT

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

One of the main challenges in optimal scaling of large language models (LLMs) is the prohibitive cost of hyperparameter tuning, particularly learning rate  $\eta$  and batch size  $B$ . While techniques like  $\mu$ P (Yang et al., 2022) provide scaling rules for optimal  $\eta$  transfer in the infinite model size limit, the optimal scaling behavior in the infinite data size limit remains unknown. We fill in this gap by observing for the first time an intricate dependence of optimal  $\eta$  scaling on the pretraining token budget  $T$ ,  $B$  and its relation to the critical batch size  $B_{\text{crit}}$ , which we measure to evolve as  $B_{\text{crit}} \propto T$ . Furthermore, we show that the optimal batch size is positively correlated with  $B_{\text{crit}}$ : keeping it fixed becomes suboptimal over time even if learning rate is scaled optimally. Surprisingly, our results demonstrate that the observed optimal  $\eta$  and  $B$  dynamics are preserved with  $\mu$ P model scaling, challenging the conventional view of  $B_{\text{crit}}$  dependence solely on loss value. Complementary to optimality, we examine the sensitivity of loss to changes in learning rate, where we find the sensitivity to decrease with increase of  $T$  and to remain constant with  $\mu$ P model scaling. We hope our results make the first step towards a unified picture of the joint optimal data and model scaling.

## 1 INTRODUCTION

Large Language Models (LLMs) have increasingly become a prominent area of study in the field of Natural Language Processing (NLP) and beyond. They have demonstrated significant improvement in performance across a wide range of tasks, such as language understanding, text generation, translation, and summarization, showing results comparable or outperforming those of an average domain expert (Dubey et al., 2024; OpenAI et al., 2024; Team et al., 2024). The primary advantage of LLMs is their ability to scale well with increased computational resources, which results in predictive improved performance (Kaplan et al., 2020; Hoffmann et al., 2022).

One of the main challenges in LLM scaling lies in the proportional scaling of computational resources required for hyperparameter tuning. To remedy this,  $\mu$ Transfer (Yang et al., 2022) technique was proposed as a way to transfer hyperparameters from a small (proxy) model to a large (target) one by introducing scaling rules for learning rate, weight multipliers and initialization scale, altogether referred to as Maximal Update Parametrization ( $\mu$ P). While significantly reducing the hyperparameter tuning cost coming with model scaling, its applicability is limited by requiring both target and proxy models to share the same batch size and number of training iterations. With current pretraining budgets surpassing trillions of tokens, it makes  $\mu$ Transfer computationally expensive to apply even with tuning a small proxy model.

One solution would be hyperparameter tuning performed both for the small proxy model *and* on the small dataset, followed by  $\mu$ Transfer to the larger model and larger dataset, under assumption of both datasets being sampled from the same underlying data distribution. This raises the question of  $\mu$ Transfer’s applicability in the *infinite data limit*, which can be formalized as an increase in the size of the training dataset, which in the LLM case is measured by the number of tokens. Understanding training dynamics in this limit would unlock hyperparameter transfer not only across model scales, but also across data horizons, thus removing the largest limitation of  $\mu$ Transfer.

The study of optimal hyperparameter evolution throughout the model training should also be complemented with a study of hyperparameter *sensitivity*, i.e. the measure of how the model performance is affected when the training is performed outside the optimal hyperparameter range. In practice, it

is rarely possible to remain within the optimum due to statistical uncertainties in its estimation. It would be of large interest to find training regimes which have small hyperparameter sensitivity and penalize model performance the least if the optimal hyperparameters are missed by a small degree.

Expanding on this line of research, we consider a commonly used LLM pretraining setup and aim towards building a yet missing holistic picture of optimal learning rate and batch size dynamics as one scales up the model training – both in the data and model sizes. Our main contributions are summarized as follows:

- **Optimal learning rate scaling:** by incorporating a dependence on the pretraining token budget into the theoretical model of optimal learning rate  $\eta^*$  scaling Li et al. (2024) via Eq. 3.1 and performing a fit to experimentally observed data (Fig. 1), we establish a dependence of the  $\eta^*$  evolution with  $T$  on the batch size  $B$  and its relation to the critical batch size  $B_{\text{crit}}$  (see definition in Sec. 2.1). From interpreting the model fit results, we obtain scaling behaviors ranging from  $\eta^* \propto \sqrt{T}$  to  $\eta^* \propto 1/\sqrt{T}$  depending on  $B$ ,  $B_{\text{crit}}$  and  $T$ , which we find compatible with experimental observations. Furthermore, we find these dynamics to be largely preserved within  $\mu\text{P}$  (Appendix A.11).
- **Optimal batch size scaling:** assuming  $\eta$  is optimal for a given data horizon  $T$ , we observe a gradual increase of the optimal batch size  $B^*$  with an increase of the token budget (Fig. 3a). The drift is correlated with the evolution of the critical batch size  $B_{\text{crit}}$  (Fig. 2, left), with  $B^*(T) < B_{\text{crit}}(T)$  in our measured range of  $T$ . Importantly, we show that naïve application of optimal  $\eta$  scaling rules in the  $T \rightarrow \infty$  limit with  $B$  being indefinitely fixed becomes suboptimal over time: a joint  $(\eta, B)$  scaling is required.
- **Critical batch size:** we experimentally find  $B_{\text{crit}}$  (see definition in Sec. 2.1) to evolve in time with  $B_{\text{crit}} \propto T^{\alpha_B}$  and  $\alpha_B = 1.0 \pm 0.2$  (Fig. 2, left). This dynamic affects optimal  $\eta$  scaling via Eq. 3.1 and drives the transition between various scaling behaviors (Sec. 3.2). Surprisingly, we show evidence that  $B_{\text{crit}}$  is not exclusively defined by the value of the loss function (Eq. 8) as suggested by McCandlish et al. (2018): models within  $\mu\text{P}$  share the same  $B_{\text{crit}}$  region while having different performance in terms of loss.
- **Learning rate sensitivity:** the sensitivity is generally decreasing with an increase of the training token budget, which is interestingly more pronounced for the batch sizes in the critical batch size region (Fig. 4). We observe no significant change in the learning rate sensitivity with the change of the  $\mu\text{P}$  base model and within the  $\mu\text{P}$  width limit (Fig. 5).

## 2 METHODOLOGY

### 2.1 TERMINOLOGY

**Time ( $T$ ):** we often use the terms *time*, *token budget*, and *data horizon* interchangeably, both to specify the measure of the training data size in tokens, and to pinpoint the specific moment throughout the model training. From this perspective, an *infinite data limit*  $T \rightarrow \infty$ , as opposed to a fixed budget regime with  $T = \text{const}$ , refers to an (infinite) increase of the number of tokens seen by the model during pretraining.

**$\mu\text{P}$ :** we refer to a model with width  $d_{\text{model}}^{\text{base}}$  as a *base model* if  $\mu\text{P}$  scaling multipliers for learning rates, weight multipliers and initialization scale (Sec. 2.2) are computed relative to this width. This brings us to a broader view on  $\mu\text{P}$  where the base model “pinpoints” the training dynamics for all the other models obtained either by scaling up or down the base  $d_{\text{model}}^{\text{base}}$  width. Together with the base model, we refer to this set of models as a  *$\mu\text{P}$  model family* or as a  *$\mu\text{P}$  trajectory* if the direction of scaling is implied. We also slightly distinguish between the base and proxy models, where the former is used to define a  $\mu\text{P}$  model family, while the latter is a model used to tune hyperparameters to be transferred with  $\mu\text{Transfer}$  to a target model.

**Critical batch size ( $B_{\text{crit}}$ ):** following Li et al. (2024), we define  $B_{\text{crit}}$  as the corresponding parameter in Eq. 3.1, also describing the peak position of the bell-shaped curve (Fig. 1a), which was shown by the authors to equal the  $B_{\text{crit}}$  definition of McCandlish et al. (2018). To better clarify the nomenclature appearing in the literature, we provide an extended discussion in Appendix A.2.

**Sensitivity:** as acknowledged by Wortsman et al. (2023), it is difficult to formalize this notion, also in the absence of a theory to be verified. We therefore define it in the most minimal way, namely as the variation of validation loss  $\mathcal{L}_{\text{val}}(\eta) - \mathcal{L}_{\text{val}}(\eta^*)$  for a given learning rate variation from its optimal value  $\eta/\eta^*$ . We refer to the corresponding loss vs. learning rate curve (both with and without  $\mathcal{L}_{\text{val}}(\eta^*)$  normalization) as a *loss profile*.

## 2.2 MODEL CONFIGURATION AND DATASETS

For all our experiments we use a default MPT model architecture (MosaicML, 2023) as implemented in the `llm-foundry` codebase (MosaicML, 2024), with all the models sharing the same training configuration (Appendix A.3). We use the Decoupled AdamW optimizer (Loshchilov & Hutter, 2019) with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.95$ ,  $\epsilon = 10^{-8}$ , weight decay  $\lambda = 0$  and gradient clipping by the  $L_2$  norm value of 1.

$\mu\text{P}$  is implemented according to Table 8 of Yang et al. (2022), so that when  $d_{\text{model}}$  is set to the base model width  $d_{\text{model}}^{\text{base}}$ , it replicates Standard Parametrization (SP). That makes our observations for the base models also applicable to setups that use SP rather than  $\mu\text{P}$ . Model weights are initialized from the normal distribution with the base model standard deviation  $\sigma^{\text{base}} = 1/\sqrt{d_{\text{model}}^{\text{base}}}$ . The models are scaled up/down only in width, with the head dimension  $d_{\text{head}}$  being always fixed and the number of heads being scaled proportionally to the width scaling.

The models are trained with the causal language modeling task on the train split of the Colossal Clean Crawled Corpus (C4) dataset (Raffel et al., 2020), tokenized with the GPT2 tokenizer (Radford et al., 2019) with a vocabulary size of 50257 and a context length of 1024 tokens. As a metric to evaluate model performance, we report the loss on the C4 validation split as  $\mathcal{L}_{\text{val}}$ .

## 2.3 HYPERPARAMETER GRID

To investigate the interplay of learning rate and batch size in the infinite data limit  $T \rightarrow \infty$ , we define a 5D grid spanned by the following axes:  $\eta$ ,  $B$ ,  $T$ ,  $d_{\text{model}}$ ,  $d_{\text{model}}^{\text{base}}$  (see Appendix A.4 for exact definition). Fundamentally, we are interested in measuring how the loss profile  $\mathcal{L}_{\text{val}}(\eta)$  and its optimum value  $\eta^*$  evolve in time  $T$  depending on the choice of batch size  $B$ . As this measurement is moreover conditioned on the  $\mu\text{P}$  trajectory and a specific point therein, we firstly study this evolution for a trajectory pinpointed by one specific base model with  $d_{\text{model}}^{\text{base}}$ . We train a set of models within the defined  $\mu\text{P}$  trajectory with different widths  $d_{\text{model}}$ , ranging in size from 32M up to 354M parameters, and measure for each of them the  $\mathcal{L}_{\text{val}}(\eta)$  profile at specific points in time  $T$ , ranging from 1B up to 275B tokens. Then, we repeat the same measurement for a new  $\mu\text{P}$  trajectory, pinpointed by a different value of  $d_{\text{model}}^{\text{base}}$ . This grid approach allows us to interpret results from multiple perspectives, as we detail in Sec. 3.

## 2.4 LEARNING RATE SCHEDULE SCALING

Since we study the training dynamics in the infinite data limit, it necessarily implies training models across different data horizons. This raises the question of how one should adjust the learning rate schedule in this limit. Motivated by recent work of Hu et al. (2024); Hägele et al. (2024), in all our experiments we use a warmup-stable (WS) version of the warmup-stable-decay (WSD) schedule consisting of a warmup phase with a linear increase of learning rate from 0 to  $\eta_{\text{max}}$  and a constant phase with learning rate fixed at  $\eta_{\text{max}}$ , hereafter notated as  $\eta$ . Our version omits the decay phase to simplify experimentation as we observe that it does not affect the optimal  $\eta$  position (Appendix A.7). The warmup duration is fixed across all horizons and across all experiments at an absolute value of  $T_{\text{warmup}} = 2^{19} = 524288$  tokens. Whenever batch size is varied, we adjust the number of gradient steps in the warmup phase accordingly so that the total amount of tokens seen by the model during warmup equals  $2^{19}$ . We also present additional experiments with different ways to scale the learning rate warmup and an added decay phase in Appendix A.7, with results largely confirming those of Hägele et al. (2024). The WS schedule allows us to reduce computational requirements by approximately a factor of two: contrary to retraining for each of the data horizons in the  $T$  grid, we run indefinitely continued trainings and take evaluation snapshots on the way.

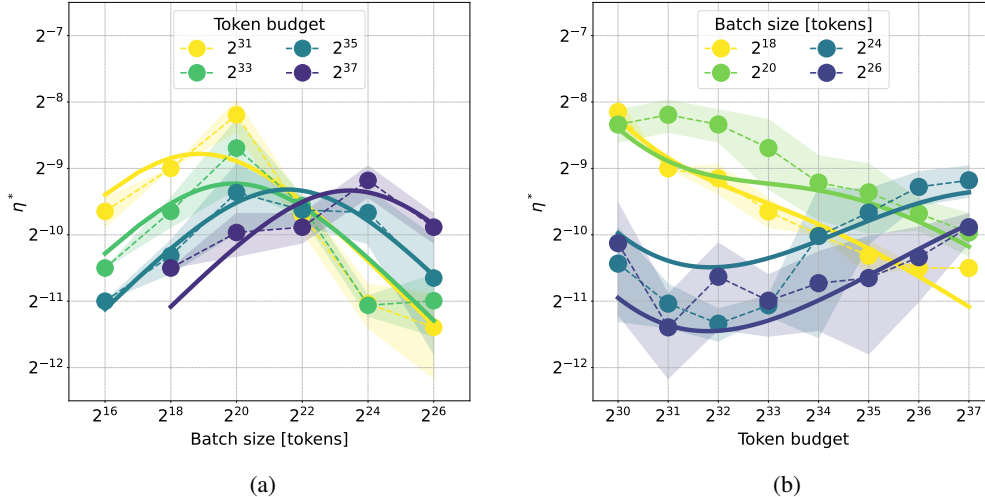


Figure 1: **(a)**: Optimal learning rate  $\eta^*$  per batch size against a set of pretraining token budgets (see Appendix A.9 for a full set). Each point is obtained by averaging experimental observations of optimal learning rate values across  $\mu\text{P}$  model family and random seeds, as described in Sec. 3.1, with color bands visualizing the corresponding standard deviation. Solid lines represent the fitted theoretical model of Li et al. (2024) (Eq. 3.1) as described in Sec. 3.1, dashed lines only connect the data points for visualization purposes. We observe an approximately linear growth of  $B_{\text{crit}}$  (see also a dedicated Fig. 2), defined as the peak position of the fitted curve, in the limit of increased token budget.

**(b)** Transposition of Fig. 1a: evolution of the optimal learning rate with an increase of the pretraining token budget  $\eta^*(T)$  for a representative set of batch sizes, in tokens. We observe the fitted model to describe the scaling behavior of low ( $B = 2^{18}$ ) and high ( $B = 2^{26}$ ) batch sizes, as well as intermediate batch sizes in the high token budget regime. For  $B = 2^{18}$ , the model reduces to  $\eta^* \propto 1/\sqrt{T}$  as discussed in Sec. 3.2, matching the observations.

### 3 RESULTS

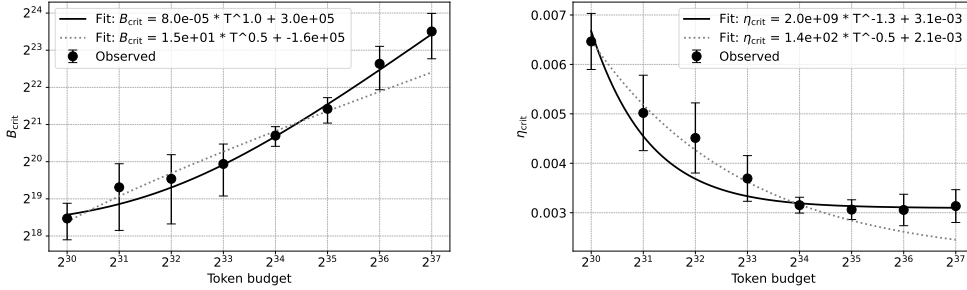
#### 3.1 CRITICAL BATCH SIZE EVOLVES IN TIME, BUT IS UNCHANGED WITHIN $\mu\text{P}$

First, we begin with setting  $d_{\text{model}}^{\text{base}} = 1024$  and scanning learning rate across different batch sizes and  $d_{\text{model}}$ . We present results for the  $\eta^*$  optimum dependence on the batch size  $B$  per data horizon  $T$  in Fig. 1a, with individual  $\mathcal{L}_{\text{val}}(\eta)$  profile scans in Appendix A.8 and the full set of horizons in Appendix A.9. In order to reduce statistical uncertainties, we average results across three  $\mu\text{P}$  models<sup>1</sup> with  $d_{\text{model}} \in \{256, 512, 1024\}$  for tokens budgets  $T \leq 2^{35}$  and additionally across four more random seeds for large batch size values  $B \in \{2^{20}, 2^{22}, 2^{24}, 2^{26}\}$  for the model with  $d_{\text{model}} = 256$  to reduce statistical fluctuations in the low token budget region. We include a similar plot for the other base model with  $d_{\text{model}}^{\text{base}} = 256$  in Appendix A.10 and individual plots for each of the  $(d_{\text{model}}, d_{\text{model}}^{\text{base}})$  configurations in Appendix A.11.

We observe that for a given time horizon, the  $(\eta^*, B)$  curve has a bell-like shape, as predicted by Li et al. (2024). The left-hand side of the peak represents a known  $\eta \propto \sqrt{B}$  scaling rule (Malladi et al., 2023; Shen et al., 2024). However, with our experiments, we uncover a previously unseen right-hand side of the curve, also referred to as ‘‘surge’’ by Li et al. (2024), where the optimal learning rate for a fixed token budget scales inversely proportionally to the batch size scaling via the  $\eta^* \propto 1/\sqrt{B}$  rule.

<sup>1</sup>We believe this averaging approach is justified since all the three models share the same optimization trajectory in terms of the number of steps, batch size and data horizon length, therefore are theoretically guaranteed by  $\mu\text{Transfer}$  to share the same optimal learning rate. From the experimental side, we also observe no significant differences across the three models (Appendix A.11).

216  
217  
218  
219  
220  
221  
222  
223  
224  
225  
226



227  
228  
229  
230  
231  
232  
233

Figure 2: Critical batch size  $B_{\text{crit}}$  (left) and critical learning rate  $\eta_{\text{crit}}$  (right), as extracted from the fit with the power law  $p_{\text{crit}} = a_p T^{\alpha_p} + b_p$ , where  $p \in \{\eta, B\}$ , following the procedure from Appendix A.5, as a function of token budget. Solid line represents the fit result. Dashed line corresponds to the fit with the power exponent fixed to  $\alpha_B = 0.5$  (left) and  $\alpha_\eta = -0.5$  (right). This model fit is visualized only to illustrate the model variation with the exponent change and its parameters are not used in the main analysis.

234  
235  
236

We analyze the observed data points within the theoretical framework of Li et al. (2024). For each of the token budgets we fit the data with the following functional form:

237  
238  
239  
240

$$\eta^*(T, B) = \frac{\eta_{\text{crit}}(T)}{\sqrt{\frac{B}{B_{\text{crit}}(T)} + \sqrt{\frac{B_{\text{crit}}(T)}{B}}}}, \tag{1}$$

241  
242  
243  
244  
245  
246

where  $B_{\text{crit}}$  and  $\eta_{\text{crit}}$  are parameters of the fit. The former corresponds to the peak position of the bell-like curve and was shown in McCandlish et al. (2018); Li et al. (2024) to approximate the critical batch size defined as the balance point between optimal number of training steps and data efficiency. The latter can be interpreted as the optimal learning rate when training in the regime with the batch size tracking the critical one, i.e.  $B(T) = B_{\text{crit}}(T)$ .

247  
248  
249  
250

After performing the fit for each token budget, we analyze the  $\eta_{\text{crit}}$  and  $B_{\text{crit}}$  evolution in time. We fit a power law of the form  $p_{\text{crit}} = a_p T^{\alpha_p} + b_p$ ,  $p \in \{\eta, B\}$  to each of the data sets following the procedure described in Appendix A.5 and present the results in Fig. 2 (solid lines). For the time dependence of critical parameters, we obtain the following power exponents:

251  
252  
253  
254

$$\begin{aligned} B_{\text{crit}} &\propto T^{\alpha_B}, & \alpha_B &= 1.0 \pm 0.2, \\ \eta_{\text{crit}} &\propto T^{\alpha_\eta}, & \alpha_\eta &= -1.3 \pm 0.4, \end{aligned} \tag{2}$$

255  
256  
257

where we find that in both cases the hypothesis of the power exponent  $+1$  ( $-1$ ) for  $\alpha_B$  ( $\alpha_\eta$ ) is compatible with experimental observations within uncertainties. Detailed results for the  $a, b$  coefficients are provided in Appendix A.5.

258  
259  
260  
261  
262  
263  
264  
265  
266

Lastly, there is a difference of  $B_{\text{crit}}$  evolution between the  $T$  and  $\mu P$  infinite width limits. Specifically, for a fixed token budget, we observe no significant change of the curves' shapes and peak positions across  $d_{\text{model}}$  values within the same  $\mu P$  trajectory, and also with the change of the base model (Appendix A.11). At the same time, there is a noticeable drift of  $B_{\text{crit}}$  in the  $T \rightarrow \infty$  limit with the model being fixed. As both limits are accompanied with a comparable change of the model performance<sup>2</sup>, this observation brings evidence that dependence of the critical batch size exclusively on the loss value suggested by Kaplan et al. (2020) (Eq. 8) is not entirely complete. Or, contrary to results in Li et al. (2024), the two definitions of the critical batch size region (Appendix A.2) are not the same and should be disentangled.

267  
268  
269

<sup>2</sup>Back-of-the-envelope calculation from Fig. 3a and Appendix A.12: for  $d_{\text{model}}^{\text{base}} = 1024$ ,  $B = 2^{20}$ , there is a loss change  $\mathcal{L}_{\text{val}} = 3.4 \rightarrow 2.8$  with a token budget increase  $2^{31} \rightarrow 2^{37}$ , resulting in  $B_{\text{crit}}$  drifting by  $2^4$ . For the same  $(d_{\text{model}}^{\text{base}}, B)$  configuration, there is no significant  $B_{\text{crit}}$  drift with a change of width by  $2^2$  within  $\mu P$ , but the corresponding loss change is  $\mathcal{L}_{\text{val}} = 3.5 \rightarrow 2.9$ .

### 3.2 LEARNING RATE OPTIMUM DRIFTS IN TIME, WITH BATCH SIZE INTERPOLATING BETWEEN DIFFERENT SCALING RULES

In Fig. 1b, we reinterpret Fig. 1a by transposing the batch size and token budget axes and by plotting the evolution of the optimal learning rate  $\eta^*$  in time  $T$  for a representative set of batch size values, with the full set of batch size values in Appendix A.9. Overlaid, we also plot the model fitted with Eq. 3.1 (solid lines).

From the data points alone we observe an intricate drift of the optimal learning rate in time as governed by the batch size value. In a simplified way, for small  $B$  values ( $2^{18}$  and  $2^{20}$  in Fig. 1b), we observe a decrease of  $\eta^*$  by  $2^2$  with an increase of the token budget by  $2^7$ , while for the larger  $B$  values ( $2^{24}$  and  $2^{26}$ ), it is oppositely an increase by up to  $2^2$ .

We also find that the fitted model<sup>3</sup> describes the data points for the smallest ( $2^{16}$  and  $2^{18}$ ) and largest ( $2^{26}$ ) probed batch sizes well. For the intermediate batch size values, it captures the behavior in the large token budget regime and the general curvature patterns ( $B = 2^{24}$ ), but sometimes lacks the correct amplitude. We note, however, large uncertainties on the fitted model parameters (Appendix A.5): additional data points with improved resolution in both  $\eta$  and  $B$  would better constrain the model fit and therefore constitute an important next step in future work.

It is instructive to consider several limiting scaling scenarios of Eq. 3.1. First, when  $B \ll B_{\text{crit}}$ , one obtains  $\eta^*(T, B) \propto \eta_{\text{crit}}(T)/\sqrt{B_{\text{crit}}(T)}$ , which we do not observe, since  $\min(B_{\text{crit}}) = 2^{18}$  in our experiments. Second, when  $B \gg B_{\text{crit}}$ , one obtains  $\eta^*(T, B) \propto \eta_{\text{crit}}(T) \cdot \sqrt{B_{\text{crit}}(T)}$ . We observe this regime for high batch size values ( $B = 2^{24}$  and  $B = 2^{26}$ ) in the high token budget regime ( $T > 2^{34}$  tokens). Since in this region, we see  $\eta_{\text{crit}}(T) \sim 1$  (Fig. 2, right) and  $B_{\text{crit}} \propto T^{\alpha_B} \approx T$ , we obtain  $\eta^* \propto \sqrt{T}$  (as can be seen in Fig. 1b). Lastly, a special case is when  $B(T) = B_{\text{crit}}(T)$ , i.e. batch size is tracking the critical one. We observe this regime for the smallest batch sizes ( $B = 2^{16}$  and  $B = 2^{18}$ ) in the low token budget regime ( $T < 2^{33}$  tokens). In that case, the optimal learning rate  $\eta^* \propto \eta_{\text{crit}} \propto T^{\alpha_\eta} = T^{-0.5}$  (see Fig. 2, where the power exponent  $\alpha_\eta = -0.5$  describes data in the low token budget region better).

### 3.3 OPTIMALLY-TUNED BATCH SIZE INCREASES IN TIME

Second, we study how optimal hyperparameter values evolve in time to yield optimal loss values. For each batch size and horizon length, we select the best-performing run across the learning rate grid and plot model loss  $\mathcal{L}_{\text{val}}$  against batch size across time horizons for the configuration with ( $d_{\text{model}} = 1024$ ,  $d_{\text{model}}^{\text{base}} = 1024$ ). Results are presented in Fig. 3, with a full set of plots across various combinations of ( $d_{\text{model}}$ ,  $d_{\text{model}}^{\text{base}}$ ) in Appendix A.12.

We observe an increase of the optimal batch size with increase of the pretraining token budget from  $B^*|_{T=2^{30}} = 2^{18}$  to  $B^*|_{T=2^{35}} = 2^{20}$  (Fig. 3a). Emergence of suboptimality is more pronounced when transposing the token budget and batch size axes (Fig. 3b), where the smallest  $B = 2^{16}$  batch size curve, with each point having learning rate scaled approximately with the inverse scaling rule  $\eta^* \propto 1/\sqrt{T}$ , is being taken over in the  $T \rightarrow \infty$  limit by the curves corresponding to larger batch sizes. Furthermore, comparing the optimal batch size values with the critical batch size evolution (Sec. 3.1), we obtain  $B^*(T) < B_{\text{crit}}$  for the range of our measurements  $T = [2^{30}, 2^{37}]$ .

This result illustrates that, while naïve “pairwise” scaling rules for optimal learning rate, e.g.  $\eta^* \propto 1/\sqrt{T}$ , are convenient for predicting optimal values at scale, they do not necessarily result in the best model performance: taking batch size dynamics into account is required. In other words, the invariant induced solely by, for example, the  $\eta^* \propto 1/\sqrt{T}$  scaling rule is not sufficient for the model performance to be optimal. We believe, similarly to Smith & Le (2018), that some broader notion of noise scale should serve as a more fundamental invariant to optimize for in the joint data and model size limit. We discuss this idea in more detail in Sec. 4.

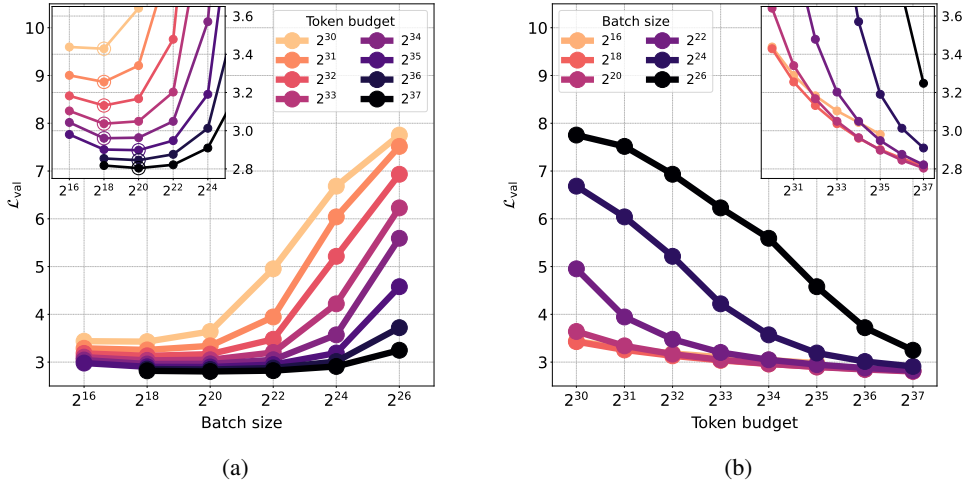


Figure 3: Validation loss  $\mathcal{L}_{\text{val}}$  for a ( $d_{\text{model}} = d_{\text{model}}^{\text{base}} = 1024$ ) model training (354M parameters) with an optimally-tuned learning rate as a function of (a) batch size split in pretraining token budgets (b) pretraining token budget split in batch size, both measured in tokens. Inset plots zoom into the optimum region. We observe that (a) optimal batch size (circled markers in the inset plot) evolves in time, by a  $\times 2^2$  ( $B = 2^{18} \rightarrow 2^{20}$  tokens) increase with an increase of the budget by  $\times 2^5$  ( $T = 2^{30} \rightarrow 2^{35}$  tokens) (b) smaller batch sizes are gradually phased out to become suboptimal as the token budget increases.

### 3.4 LEARNING RATE SENSITIVITY IS REDUCED IN TIME, AND IS UNCHANGED WITHIN $\mu\text{P}$

After having studied the learning rate optimum dynamics, we turn our attention to a broader structure around the optimum from the sensitivity perspective. Specifically, we are interested in how the *shape* of the  $\mathcal{L}_{\text{val}}(\eta)$  curve changes in the time  $T \rightarrow \infty$  and  $\mu\text{P}$  width limits. In Fig. 4, we present our observations for the two base models with  $d_{\text{model}}^{\text{base}} = d_{\text{model}} \in \{256, 1024\}$ , for token budgets  $T \in \{2^{31}, 2^{33}, 2^{35}\}$ . We note that since we implement  $\mu\text{P}$  in a way that the base model is also SP-parametrized, the results should be applicable to this parametrization as well.

We observe that there is a general decrease in the learning rate sensitivity by up to  $2^1$  per each token budget increase by  $2^2$  as measured by  $\mathcal{L}_{\text{val}} - \mathcal{L}_{\text{val}}^{\text{min}}$  value, where  $\mathcal{L}_{\text{val}}^{\text{min}} = \mathcal{L}_{\text{val}}(\eta^*)$  is the validation loss value in the learning rate optimum. This indicates that the model profits from longer training by having lower penalty for the misspecification of the optimal learning rate. Notably, the decrease is more pronounced for batch sizes in the critical region ( $B = 2^{20}$  and  $2^{22}$ ), while for the region with the  $\eta^* \propto 1/\sqrt{T}$  scaling rule ( $B = 2^{18}$ ), the effect is either reduced (base model  $d_{\text{model}}^{\text{base}} = 1024$ ) or shows asymmetric trends w.r.t. the learning rate optimum (base model  $d_{\text{model}}^{\text{base}} = 256$ ). However, within our measurement precision, the sensitivity evens out across batch sizes for the longest  $2^{35}$  token horizon. Overall, our results motivate the choice of the training regime within the critical batch size region in order to minimize the risks of under- or overshooting the learning rate optimum. As we show in Appendix A.6, the learning rate optimum position can vary by a factor of two just depending on the random seed choice.

With respect to the  $\mu\text{P}$  width limit, we observe no significant deviation of the loss profile from the one of the base model, both for up- and down-scaled models within  $\mu\text{P}$  (Fig. 5 with and Fig. 21 without  $\mathcal{L}_{\text{val}}$  normalization). Evaluated for the data horizon of  $T = 2^{35} \approx 34\text{B}$  tokens, this holds across the models with the number of trainable parameters ranging from 32M up to 5B. Likewise, changing the base model does not affect the profile shape, except for the optimum learning rate shift

<sup>3</sup>We note that the model is not fitted to the data points in the representation of Fig. 1b. Instead, as described in Sec. 3.1, the fit is performed in two consecutive steps: by first capturing the  $(\eta^*, B)$  behavior per token budget (illustrated in Fig. 1a), followed by fitting the time dynamics (illustrated in Fig. 2).

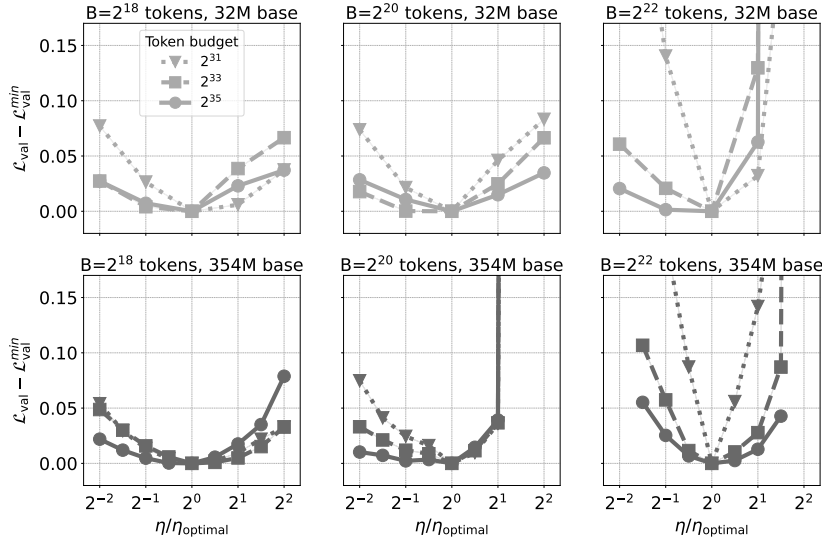


Figure 4: Learning rate sensitivity  $\mathcal{L}_{\text{val}} - \mathcal{L}_{\text{val}}^{\min}$  as a function of the learning rate deviation from the optimal value  $\eta/\eta_{\text{optimal}}$ , measured for batch sizes of  $B = 2^{18}$  (left column),  $2^{20}$  (middle column), and  $2^{22}$  (right column) tokens, separately for the  $\mu\text{P}$  base models with width  $d_{\text{model}}^{\text{base}} = 256$  (top row) and 1024 (bottom row). The former model amounts to 32M and the latter to 354M trainable parameters. With an increase of the pretraining token budget (different marker styles) we observe a general decrease in the learning rate sensitivity, which is more pronounced for batch sizes  $B \in \{2^{20}, 2^{22}\}$  in the critical region (Sec. 2.1) and for the 354M model. At the largest probed token budget  $T = 2^{35}$  tokens, the sensitivity equalizes across the models and batch sizes.

by  $\times 2$ , which is expected for the base models compared here due to our  $d_{\text{model}}^{\text{base}}$ -dependent weight initialization scheme (Sec. 2.2).

## 4 DISCUSSION

While originally, we were aiming to find a golden recipe for hyperparameter transfer in the infinite data limit, we show that there is no simple and straight-forward answer. In Sec. 3.2, we show that the model fit based on Eq. 3.1 describes the observed data points well. However, as illustrated in Sec. 3.3, following the optimal learning rate trajectory in time is not sufficient to obtain optimal performance. That leads us to believe that there exists a deeper underlying perspective on the problem, as opposed to the one of simply tuning learning rate and batch size.

Fundamentally, the field of model parametrization research has originated from and is further converging towards preserving some notion of norm in some infinite (model width and/or depth) limit (Everett et al., 2024; Yang et al., 2024; Large et al., 2024). In fact, any parametrization itself is simply a set of scaling rules to be applied to hyperparameters in order to preserve these norms (e.g. of model weight matrices or weight updates). Expanding on this, one can argue that scaling rules follow from the requirement of keeping some underlying quantity invariant within the infinite limit. From this perspective, hyperparameter transfer is nothing but a consequence of such “conservation laws”.

With this perspective in mind, we draw a parallel between infinite model and data limits, and speculate that a similar notion of “norm” should exist and should be aimed to be preserved in the infinite data limit. In fact, there is already a good candidate for this, namely the *noise scale* (Eq. 5 and 9), which intriguingly also induces scaling rules for hyperparameters (see Appendix A.2 for in-depth discussion). However, the existing definition neither takes into account the adaptive nature of the optimizer, nor the scenario of jointly following the infinite data and model limits.



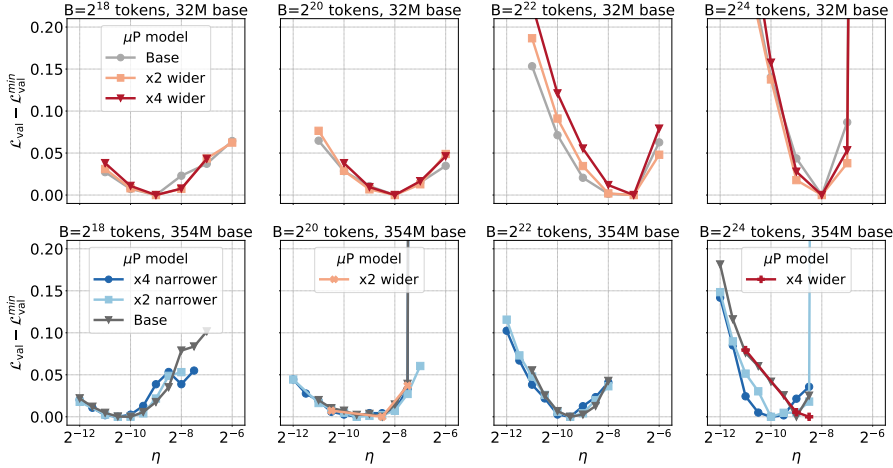


Figure 5: Learning rate sensitivity  $\mathcal{L}_{\text{val}} - \mathcal{L}_{\text{val}}^{\min}$  as a function of learning rate  $\eta$ , measured for batch sizes of  $B = 2^{18}$  (leftmost column),  $2^{20}$  (middle left column),  $2^{22}$  (middle right column) and  $2^{24}$  (rightmost column) tokens, separately for the  $\mu P$  base models with the width  $d_{\text{model}}^{\text{base}} = 256$  (top row) and 1024 (bottom row). Different marker styles correspond to different models within the  $\mu P$  family, with all the models being evaluated at the data horizon of  $T = 2^{35}$  tokens. For the base model with  $d_{\text{model}}^{\text{base}} = 256$ , we scale the width only downwards, while for the base model with  $d_{\text{model}}^{\text{base}} = 1024$ , we scale it both upwards and downwards. We observe no significant difference in the sensitivity across all the  $(d_{\text{model}}^{\text{base}}, d_{\text{model}})$  configurations. Note that for the configuration  $(B = 2^{24}, d_{\text{model}}^{\text{base}} = 1024)$ , the base and  $d_{\text{model}} = 4 \times d_{\text{model}}^{\text{base}}$  models share a different random seed compared to all the other models, to illustrate the loss penalty arising from the learning rate optimum variation.

We hope that our experimental observations, similarly to the discovery of the  $\eta^* \propto T$  scaling rule for SGD, will make the first step towards the theoretical unification of infinite data and model size limits via deriving such a joint scaling invariant. Still, our insights into optimal scaling rules for learning rate and batch size might be valuable for practitioners who approach the problem of hyperparameter optimization in the infinite data and model size limit. We provide our recommendations in Appendix A.1.

As future work, it would be important to improve the resolution in data points with a finer grid of  $(\eta, B)$  values. This is a necessary step to establish the generalization power of Eq. 3.1 and the power law fits of its critical parameters as a function of time (Fig. 2), also across various data sets, model architectures and modalities. Additionally, while we used  $\mu P$  as the main way to incorporate model scaling due to its ability to transfer optimal learning rate across model sizes, recent work of Everett et al. (2024) suggests that this is not the only way to do so. A similar study to ours, but for other model parametrizations, is an exciting direction of future research.

## 5 RELATED WORK

**$(\eta, B)$  scaling rules** In efforts to accelerate model training, the  $\eta \propto B$  rule for the SGD optimizer was found necessary to avoid performance loss due to increased batch size (Goyal et al., 2018), known as generalization gap (Keskar et al., 2017). Afterwards, additional usage of momentum (Smith et al., 2018) and model scaling (Park et al., 2019) was incorporated, and a  $\eta \propto \sqrt{B}$  rule for Adam was observed (Hilton et al., 2022). From the theoretical side, experimentally observed rules were verified with the framework of stochastic differential equations (SDEs) (Smith & Le, 2018; Malladi et al., 2023), loss curvature analysis (Zhang et al., 2019; McCandlish et al., 2018; Li et al., 2024) and random matrix theory (Granzio et al., 2021). While most of the studies were performed in the fixed epoch budget, Shallue et al. (2019) broadened the perspective to other target budget

measures and studied the scope of the  $\eta \propto B$  rule applicability across various datasets and model architectures. Looking beyond fixed budgets, Smith & Le (2018) showed a linear relation between the optimal batch size and the dataset size (for fixed  $\eta$ ), and Smith et al. (2020) similarly presented hints for a linear relation between the optimal learning rate and the dataset size (for fixed  $B$ ), with both works considering the SGD optimizer. In the modern LLM pretraining context, Hu et al. (2024); DeepSeek-AI et al. (2024) approached this problem by deriving the joint  $(\eta, B)$  scaling laws.

**$\mu\text{P}$**  Originally developed within the Tensor Program series studying feature learning in the infinite width limit (Yang & Hu, 2022; Yang et al., 2022),  $\mu\text{P}$  has been gaining traction recently within the LLM community. It has been extensively tested and applied experimentally (Lingle, 2024; Blake et al., 2024; Gunter et al., 2024; Dey et al., 2024), as well as theoretically, with Yang et al. (2023); Bordelon et al. (2024) extending it to the infinite depth limit, and Yang et al. (2024); Bernstein et al. (2023) revisiting it from the spectral normalization perspective. Recently, Everett et al. (2024) showed that other model parametrizations also induce hyperparameter transfer if taking weight alignment into account. Furthermore, they revealed that  $\mu\text{Transfer}$  does not work in the regime of Chinchilla-optimal scaling (Hoffmann et al., 2022). The most closely related work to ours, Shen et al. (2024) expanded on this observation and proposed a learning rate scheduler combining  $\mu\text{P}$  and experimentally measured  $(\eta, B)$  scaling rules to allow for the hyperparameter transfer in the  $T \rightarrow \infty$  limit, however only limited to the  $\eta^* \propto 1/\sqrt{T}$  scaling regime.

**Sensitivity** The topic of loss sensitivity to suboptimal hyperparameter choice is less thoroughly studied, focusing exclusively on learning rate as the most affecting hyperparameter. Wortsman et al. (2023) studied how various optimizer and model interventions, such as weight decay or  $\mu\text{P}$  usage, influence the learning rate sensitivity with the model size scaling. Hägele et al. (2024) investigated the impact of various learning rate schedule choices, such as length and functional form of the decay phase.

## 6 CONCLUSION

In this work, we studied joint model and data scaling in the LLM context from the perspective of optimal learning rate  $\eta$  and batch size  $B$  dynamics. We observed an intricate dependence of optimal  $\eta$  scaling on  $B$  and its relation to the critical batch size  $B_{\text{crit}}$ , as a function of the pretraining token budget  $T$ . This dynamic is preserved during model scaling with  $\mu\text{P}$ , as well as the loss sensitivity to the learning rate variation, highlighting the intriguing difference in how  $\mu\text{P}$  infinite width and time limits evolve the critical batch size. Overall, we hope our observations pave the way towards deeper understanding of the optimal scaling in the unified infinite data and model size limit.

## REFERENCES

- Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. URL <https://arxiv.org/abs/1607.06450>.
- Jeremy Bernstein, Chris Mingard, Kevin Huang, Navid Azizan, and Yisong Yue. Automatic gradient descent: Deep learning without hyperparameters, 2023. URL <https://arxiv.org/abs/2304.05187>.
- Charlie Blake, Constantin Eichenberg, Josef Dean, Lukas Balles, Luke Yuri Prince, Björn Deiseroth, Andres Felipe Cruz-Salinas, Carlo Luschi, Samuel Weinbach, and Douglas Orr.  $u\text{-}\mu\text{P}$ : The unit-scaled maximal update parametrization. In *2nd Workshop on Advancing Neural Network Training: Computational Efficiency, Scalability, and Resource Optimization (WANT@ICML 2024)*, 2024. URL <https://openreview.net/forum?id=44NKKzz1n5>.
- 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>.
- Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning, 2023. URL <https://arxiv.org/abs/2307.08691>.

- 540 DeepSeek-AI, :, Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng,  
541 Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi  
542 Ge, Kang Guan, Daya Guo, Jianzhong Guo, Guangbo Hao, Zhewen Hao, Ying He, Wenjie Hu,  
543 Panpan Huang, Erhang Li, Guowei Li, Jiashi Li, Yao Li, Y. K. Li, Wenfeng Liang, Fangyun Lin,  
544 A. X. Liu, Bo Liu, Wen Liu, Xiaodong Liu, Xin Liu, Yiyuan Liu, Haoyu Lu, Shanghao Lu, Fuli  
545 Luo, Shirong Ma, Xiaotao Nie, Tian Pei, Yishi Piao, Junjie Qiu, Hui Qu, Tongzheng Ren, Zehui  
546 Ren, Chong Ruan, Zhangli Sha, Zhihong Shao, Junxiao Song, Xuecheng Su, Jingxiang Sun,  
547 Yaofeng Sun, Minghui Tang, Bingxuan Wang, Peiyi Wang, Shiyu Wang, Yaohui Wang, Yongji  
548 Wang, Tong Wu, Y. Wu, Xin Xie, Zhenda Xie, Ziwei Xie, Yiliang Xiong, Hanwei Xu, R. X. Xu,  
549 Yanhong Xu, Dejian Yang, Yuxiang You, Shuiping Yu, Xingkai Yu, B. Zhang, Haowei Zhang,  
550 Lecong Zhang, Liyue Zhang, Mingchuan Zhang, Minghua Zhang, Wentao Zhang, Yichao Zhang,  
551 Chenggang Zhao, Yao Zhao, Shangyan Zhou, Shunfeng Zhou, Qihao Zhu, and Yuheng Zou.  
552 Deepseek llm: Scaling open-source language models with longtermism, 2024. URL <https://arxiv.org/abs/2401.02954>.
- 553 Nolan Dey, Quentin Anthony, and Joel Hestness. The practitioner’s guide to the  
554 maximal update parameterization, 2024. URL [https://www.cerebras.ai/blog/  
555 the-practitioners-guide-to-the-maximal-update-parameterization](https://www.cerebras.ai/blog/the-practitioners-guide-to-the-maximal-update-parameterization).
- 556  
557 Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha  
558 Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony  
559 Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark,  
560 Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere,  
561 Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris  
562 Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong,  
563 Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny  
564 Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino,  
565 Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael  
566 Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Ander-  
567 son, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah  
568 Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan  
569 Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Ma-  
570 hadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy  
571 Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak,  
572 Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Al-  
573 wala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini,  
574 Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der  
575 Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo,  
576 Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Man-  
577 nat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova,  
578 Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal,  
579 Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur  
580 Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhar-  
581 gava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong,  
582 Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic,  
583 Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sum-  
584 baly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa,  
585 Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang,  
586 Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende,  
587 Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney  
588 Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom,  
589 Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta,  
590 Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petro-  
591 vic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang,  
592 Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuwei Wang, Yaelle Goldschlag, Yashesh Gaur,  
593 Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre  
Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha  
Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay  
Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda  
Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew

- 594 Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita  
595 Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh  
596 Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De  
597 Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Bran-  
598 don Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina  
599 Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai,  
600 Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymier, Daniel Li,  
601 Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana  
602 Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil,  
603 Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Ar-  
604 caute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco  
605 Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella  
606 Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory  
607 Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang,  
608 Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Gold-  
609 man, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman,  
610 James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer  
611 Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe  
612 Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie  
613 Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun  
614 Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal  
615 Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva,  
616 Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian  
617 Khabza, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson,  
618 Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Ke-  
619 neally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel  
620 Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mo-  
621 hammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navy-  
622 ata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong,  
623 Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli,  
624 Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux,  
625 Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao,  
626 Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li,  
627 Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott,  
628 Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Sa-  
629 tadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lind-  
630 say, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang  
631 Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen  
632 Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho,  
633 Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser,  
634 Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Tim-  
635 othy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan,  
636 Vinay Satish Kumar, Vishal Mangla, Vitor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu  
637 Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Con-  
638 stable, Xiaocheng Tang, Xiaofang Wang, Xiaojuan Wu, Xiaolan Wang, Xide Xia, Xilun Wu,  
639 Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi,  
640 Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef  
641 Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models, 2024.  
642 URL <https://arxiv.org/abs/2407.21783>.
- 643 Katie Everett, Lechao Xiao, Mitchell Wortsman, Alexander A. Alemi, Roman Novak, Peter J.  
644 Liu, Izzeddin Gur, Jascha Sohl-Dickstein, Leslie Pack Kaelbling, Jaehoon Lee, and Jeffrey  
645 Pennington. Scaling exponents across parameterizations and optimizers, 2024. URL <https://arxiv.org/abs/2407.05872>.
- 646 Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, An-  
647 drew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet  
in 1 hour, 2018. URL <https://arxiv.org/abs/1706.02677>.

- 648 Diego Granzio, Stefan Zohren, and Stephen Roberts. Learning rates as a function of batch size:  
649 A random matrix theory approach to neural network training, 2021. URL <https://arxiv.org/abs/2006.09092>.  
650  
651
- 652 Tom Gunter, Zirui Wang, Chong Wang, Ruoming Pang, Andy Narayanan, Aonan Zhang, Bowen  
653 Zhang, Chen Chen, Chung-Cheng Chiu, David Qiu, Deepak Gopinath, Dian Ang Yap, Dong  
654 Yin, Feng Nan, Floris Weers, Guoli Yin, Haoshuo Huang, Jianyu Wang, Jiarui Lu, John Pee-  
655 bles, Ke Ye, Mark Lee, Nan Du, Qibin Chen, Quentin Keunebroek, Sam Wiseman, Syd Evans,  
656 Tao Lei, Vivek Rathod, Xiang Kong, Xianzhi Du, Yanghao Li, Yongqiang Wang, Yuan Gao,  
657 Zaid Ahmed, Zhaoyang Xu, Zhiyun Lu, Al Rashid, Albin Madappally Jose, Alec Doane, Alfredo  
658 Bencomo, Allison Vanderby, Andrew Hansen, Ankur Jain, Anupama Mann Anupama, Areeba  
659 Kamal, Bugu Wu, Carolina Brum, Charlie Maalouf, Chinguun Erdenebileg, Chris Dulhanty, Do-  
660 minik Moritz, Doug Kang, Eduardo Jimenez, Evan Ladd, Fangping Shi, Felix Bai, Frank Chu,  
661 Fred Hohman, Hadas Kotek, Hannah Gillis Coleman, Jane Li, Jeffrey Biggam, Jeffery Cao, Jeff  
662 Lai, Jessica Cheung, Jiulong Shan, Joe Zhou, John Li, Jun Qin, Karanjeet Singh, Karla Vega,  
663 Kelvin Zou, Laura Heckman, Lauren Gardiner, Margit Bowler, Maria Cordell, Meng Cao, Nicole  
664 Hay, Nilesh Shahdadpuri, Otto Godwin, Pranay Dighe, Pushyami Rachapudi, Ramsey Tantawi,  
665 Roman Frigg, Sam Davarnia, Sanskruti Shah, Saptarshi Guha, Sasha Sirovica, Shen Ma, Shuang  
666 Ma, Simon Wang, Sulgi Kim, Suma Jayaram, Vaishaal Shankar, Varsha Paidi, Vivek Kumar,  
667 Xin Wang, Xin Zheng, Walker Cheng, Yael Shrager, Yang Ye, Yasu Tanaka, Yihao Guo, Yun-  
668 song Meng, Zhao Tang Luo, Zhi Ouyang, Alp Ayyar, Alvin Wan, Andrew Walkingshaw, Andy  
669 Narayanan, Antonie Lin, Arsalan Farooq, Brent Ramerth, Colorado Reed, Chris Bartels, Chris  
670 Chaney, David Riazati, Eric Liang Yang, Erin Feldman, Gabriel Hochstrasser, Guillaume Seguin,  
671 Irina Belousova, Joris Pelemans, Karen Yang, Keivan Alizadeh Vahid, Liangliang Cao, Mah-  
672 yar Najibi, Marco Zuliani, Max Horton, Minsik Cho, Nikhil Bhendawade, Patrick Dong, Piotr  
673 Maj, Pulkit Agrawal, Qi Shan, Qichen Fu, Regan Poston, Sam Xu, Shuangning Liu, Sushma  
674 Rao, Tashweena Heeramun, Thomas Merth, Uday Rayala, Victor Cui, Vivek Rangarajan Sridhar,  
675 Wencong Zhang, Wenqi Zhang, Wentao Wu, Xingyu Zhou, Xinwen Liu, Yang Zhao, Yin Xia,  
676 Zhile Ren, and Zhongzheng Ren. Apple intelligence foundation language models, 2024. URL  
677 <https://arxiv.org/abs/2407.21075>.
- 678 Jacob Hilton, Karl Cobbe, and John Schulman. Batch size-invariance for policy optimization, 2022.  
679 URL <https://arxiv.org/abs/2110.00641>.  
680
- 681 Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza  
682 Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hen-  
683 nigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy,  
684 Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre.  
685 Training compute-optimal large language models, 2022. URL <https://arxiv.org/abs/2203.15556>.  
686
- 687 Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang,  
688 Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang,  
689 Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang  
690 Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm: Unveiling the potential of small  
691 language models with scalable training strategies, 2024. URL <https://arxiv.org/abs/2404.06395>.  
692
- 693 Alexander Hägele, Elie Bakouch, Atli Kosson, Loubna Ben Allal, Leandro Von Werra, and Martin  
694 Jaggi. Scaling laws and compute-optimal training beyond fixed training durations, 2024. URL  
695 <https://arxiv.org/abs/2405.18392>.  
696
- 697 Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats L. Richter, Quentin Anthony, Timothée  
698 Lesort, Eugene Belilovsky, and Irina Rish. Simple and scalable strategies to continually pre-train  
699 large language models, 2024. URL <https://arxiv.org/abs/2403.08763>.  
700
- 701 Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and gen-  
702 eralization in neural networks, 2020. URL <https://arxiv.org/abs/1806.07572>.  
703
- 704 Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child,  
705 Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language  
706 models, 2020. URL <https://arxiv.org/abs/2001.08361>.

- 702 Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Pe-  
703 ter Tang. On large-batch training for deep learning: Generalization gap and sharp minima, 2017.  
704 URL <https://arxiv.org/abs/1609.04836>.  
705
- 706 Atli Kosson, Bettina Messmer, and Martin Jaggi. Analyzing & eliminating learning rate warmup in  
707 GPT pre-training. In *High-dimensional Learning Dynamics 2024: The Emergence of Structure*  
708 *and Reasoning*, 2024. URL <https://openreview.net/forum?id=RveSp5oESA>.
- 709 Tim Large, Yang Liu, Minyoung Huh, Hyojin Bahng, Phillip Isola, and Jeremy Bernstein. Scalable  
710 optimization in the modular norm, 2024. URL <https://arxiv.org/abs/2405.14813>.  
711
- 712 Shuaipeng Li, Penghao Zhao, Hailin Zhang, Xingwu Sun, Hao Wu, Dian Jiao, Weiyang Wang,  
713 Chengjun Liu, Zheng Fang, Jinbao Xue, Yangyu Tao, Bin Cui, and Di Wang. Surge phenomenon  
714 in optimal learning rate and batch size scaling, 2024. URL <https://arxiv.org/abs/2405.14578>.  
715
- 716 Lucas Lingle. A large-scale exploration of  $\mu$ -transfer, 2024. URL <https://arxiv.org/abs/2404.05728>.  
717  
718
- 719 Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL <https://arxiv.org/abs/1711.05101>.  
720
- 721 Sadhika Malladi, Kaifeng Lyu, Abhishek Panigrahi, and Sanjeev Arora. On the sdes and scal-  
722 ing rules for adaptive gradient algorithms, 2023. URL <https://arxiv.org/abs/2205.10287>.  
723  
724
- 725 Sam McCandlish, Jared Kaplan, Dario Amodei, and OpenAI Dota Team. An empirical model of  
726 large-batch training, 2018. URL <https://arxiv.org/abs/1812.06162>.  
727
- 728 MosaicML. Introducing mpt-7b: A new standard for open-source, commercially usable llms, 2023.  
729 URL [www.mosaicml.com/blog/mpt-7b](http://www.mosaicml.com/blog/mpt-7b). Accessed: 2023-05-05.
- 730 MosaicML. Llm foundry, 2024. URL <https://github.com/mosaicml/llm-foundry>.  
731
- 732 OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Floren-  
733 cia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red  
734 Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Moham-  
735 mad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher  
736 Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brock-  
737 man, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann,  
738 Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis,  
739 Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey  
740 Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux,  
741 Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila  
742 Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix,  
743 Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gib-  
744 son, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan  
745 Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hal-  
746 lacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan  
747 Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu,  
748 Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun  
749 Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Ka-  
750 mali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook  
751 Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel  
752 Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen  
753 Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel  
754 Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez,  
755 Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv  
Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney,  
Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick,  
Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel

- 756 Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Ra-  
757 jeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe,  
758 Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel  
759 Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe  
760 de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny,  
761 Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl,  
762 Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra  
763 Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders,  
764 Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Sel-  
765 sam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor,  
766 Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky,  
767 Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang,  
768 Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Pre-  
769 ston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vi-  
770 jayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan  
771 Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng,  
772 Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Work-  
773 man, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming  
774 Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao  
775 Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL  
<https://arxiv.org/abs/2303.08774>.
- 776 Daniel S. Park, Jascha Sohl-Dickstein, Quoc V. Le, and Samuel L. Smith. The effect of network  
777 width on stochastic gradient descent and generalization: an empirical study, 2019. URL <https://arxiv.org/abs/1905.03776>.
- 778 Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language  
779 models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- 780 Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi  
781 Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text  
782 transformer. *Journal of machine learning research*, 21(140):1–67, 2020.
- 783 Christopher J. Shallue, Jaehoon Lee, Joseph Antognini, Jascha Sohl-Dickstein, Roy Frostig, and  
784 George E. Dahl. Measuring the effects of data parallelism on neural network training, 2019. URL  
785 <https://arxiv.org/abs/1811.03600>.
- 786 Yikang Shen, Matthew Stallone, Mayank Mishra, Gaoyuan Zhang, Shawn Tan, Aditya Prasad, Adria-  
787 na Meza Soria, David D. Cox, and Rameswar Panda. Power scheduler: A batch size and token  
788 number agnostic learning rate scheduler, 2024. URL <https://arxiv.org/abs/2408.13359>.
- 789 Samuel L. Smith and Quoc V. Le. A bayesian perspective on generalization and stochastic gradient  
790 descent, 2018. URL <https://arxiv.org/abs/1710.06451>.
- 791 Samuel L. Smith, Pieter-Jan Kindermans, Chris Ying, and Quoc V. Le. Don’t decay the learning  
792 rate, increase the batch size, 2018. URL <https://arxiv.org/abs/1711.00489>.
- 793 Samuel L. Smith, Erich Elsen, and Soham De. On the generalization benefit of noise in stochastic  
794 gradient descent, 2020. URL <https://arxiv.org/abs/2006.15081>.
- 795 Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: En-  
796 hanced transformer with rotary position embedding, 2023. URL <https://arxiv.org/abs/2104.09864>.
- 797 Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut,  
798 Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson,  
799 Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lilli-  
800 crap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hen-  
801 nigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins,  
802 Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk,  
803 Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal

810 Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis  
811 Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah  
812 Liu, Andras Orban, Fabian Gra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapath-  
813 thy, Steven Zheng, HyunJeong Choe, goston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal,  
814 Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Mar-  
815 tin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker,  
816 Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs,  
817 Anas White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lu-  
818 cas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura  
819 Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban  
820 Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi  
821 Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober,  
822 Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William  
823 Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan  
824 Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iaki Iturrate, Ruibo Liu, Yunxuan Li,  
825 Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hart-  
826 man, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego  
827 de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adri Puigdomnech Badia, David Re-  
828 itter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane  
829 Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi  
830 Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Bala-  
831 guer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Gana-  
832 pathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting  
833 Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy  
834 Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault  
835 Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli,  
836 Sbastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin,  
837 Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodgkinson, Pranav Shyam, Johan  
838 Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander  
839 Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipan-  
840 jan Das, Dominika Rogoziska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka,  
841 Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei  
842 Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim,  
843 Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan  
844 Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo,  
845 Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Lan-  
846 don, Miteyan Patel, Ruizhe Zhao, Kevin Vilella, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai  
847 Gimnez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal,  
848 Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fer-  
849 nando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex  
850 Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek,  
851 Ross McIlroy, Mario Lui, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul  
852 Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin  
853 Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc,  
854 Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua  
855 Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash  
856 Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose  
857 Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth,  
858 Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay  
859 Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina  
860 Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si,  
861 Jeremy Greer, Guolong Su, Martin Polacek, Raphal Lopez Kaufman, Simon Tokumine, Hexi-  
862 ang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Toma-  
863 sev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada  
Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimentko, Chih-Kuan Yeh, Soravit Chang-  
pinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan,  
Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu  
Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe  
Sjsund, Sbastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan,



864 Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate  
865 Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio  
866 Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kass-  
867 ner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El  
868 Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao  
869 Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec,  
870 Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson,  
871 Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlias, Arpi Vezer, Marco  
872 Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan  
873 Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pel-  
874 lat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi,  
875 Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang,  
876 Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Has-  
877 san, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal,  
878 Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević,  
879 Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot,  
880 Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks,  
881 Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang,  
882 Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert,  
883 Nate Hurlley, Motoki Sano, Anhad Mohanane, Jonah Joughin, Egor Filonov, Tomasz Kepa,  
884 Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil  
885 Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Nor-  
886 bert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou,  
887 Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej  
888 Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan  
889 Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan  
890 Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lom-  
891 briser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas  
892 Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii  
893 Duzhyi, Anton Aligmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh  
894 Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Sub-  
895 habrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Mau-  
896 rya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M,  
897 Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu,  
898 Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet,  
899 Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-  
900 Eleonora Marinescu, Martin Bølle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang,  
901 Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi,  
902 Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa  
903 Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrz, Pengcheng Yin, Jon Simon, Mal-  
904 colm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka,  
905 Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn,  
906 Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie  
907 Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik  
908 Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam  
909 Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee,  
910 Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang,  
911 Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Do-  
912 oley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao,  
913 Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, An-  
914 drew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim,  
915 Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali  
916 Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mu-  
917 jika, Igor Petrovski, Pierre-Louis Cedo, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Sid-  
918 dharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury,  
919 Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting  
920 Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor  
921 Åhdel, Sujeewan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent  
922 Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Woj-

918 ciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou,  
 919 Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen,  
 920 Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur  
 921 Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will  
 922 Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi  
 923 Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric  
 924 Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas  
 925 Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey  
 926 Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan  
 927 Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros,  
 928 Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan  
 929 Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David  
 930 Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu,  
 931 Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieil-  
 932 lard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong  
 933 Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper,  
 934 Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, John-  
 935 son Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng  
 936 Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Sukanthan, Alfonso Castaño, Irene  
 937 Giannoumis, Wooyeol Kim, Mikolaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David So-  
 938 ergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Di-  
 939 ana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li,  
 940 Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Mar-  
 941 cus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan,  
 942 Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom  
 943 van der Weide, Priya Ponnappalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse,  
 944 Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel  
 945 Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan  
 946 Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili  
 947 Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon,  
 948 Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi  
 949 Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bu-  
 950 lanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer,  
 951 Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Han-  
 952 nah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan  
 953 Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J.  
 954 Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Den-  
 955 nis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li,  
 956 Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila  
 957 Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nico-  
 958 las Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian  
 959 Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu  
 960 Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko,  
 961 Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver  
 962 Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina,  
 963 Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton,  
 964 Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Be-  
 965 nigno Urias, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing  
 966 Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim,  
 967 Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann,  
 968 Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim  
 969 Pöder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu  
 970 Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun  
 971 Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christo-  
 pher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan,  
 Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona  
 Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson,  
 Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei  
 Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas

- 972 Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson,  
973 Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew  
974 Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi  
975 Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yad-  
976 lowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan  
977 Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Ram-  
978 moham Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan  
979 Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy  
980 Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshv, Nina Martin,  
981 Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier,  
982 Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie,  
983 Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason  
984 Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng,  
985 Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia  
986 Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Gold-  
987 enson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki,  
988 Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria  
989 Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal  
990 Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimír Miku-  
991 lik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua  
992 Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang,  
993 Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku,  
994 Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini,  
995 Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan  
996 Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush,  
997 Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum,  
998 Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel  
999 Elkind, Aviël Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikuś,  
1000 Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirn-  
1001 schall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng,  
1002 Keshav Dhandhanja, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao,  
1003 Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu,  
1004 Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna  
1005 Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar,  
1006 Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mu-  
1007 dit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha  
1008 Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger,  
1009 Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao,  
1010 Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu,  
1011 Slav Petrov, Jeffrey Dean, and Oriol Vinyals. Gemini: A family of highly capable multimodal  
1012 models, 2024. URL <https://arxiv.org/abs/2312.11805>.
- 1013 Nikhil Vyas, Depen Morwani, Rosie Zhao, Gal Kaplun, Sham Kakade, and Boaz Barak. Beyond  
1014 implicit bias: The insignificance of sgd noise in online learning, 2024. URL <https://arxiv.org/abs/2306.08590>.
- 1015 Mitchell Wortsman, Peter J. Liu, Lechao Xiao, Katie Everett, Alex Alemi, Ben Adlam, John D. Co-  
1016 Reyes, Izzeddin Gur, Abhishek Kumar, Roman Novak, Jeffrey Pennington, Jascha Sohl-dickstein,  
1017 Kelvin Xu, Jaehoon Lee, Justin Gilmer, and Simon Kornblith. Small-scale proxies for large-scale  
1018 transformer training instabilities, 2023. URL <https://arxiv.org/abs/2309.14322>.
- 1019 Greg Yang and Edward J. Hu. Feature learning in infinite-width neural networks, 2022. URL  
1020 <https://arxiv.org/abs/2011.14522>.
- 1021 Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ry-  
1022 der, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs v: Tuning large neural  
1023 networks via zero-shot hyperparameter transfer, 2022. URL <https://arxiv.org/abs/2203.03466>.
- 1024 Greg Yang, Dingli Yu, Chen Zhu, and Soufiane Hayou. Tensor programs vi: Feature learning in  
1025 infinite-depth neural networks, 2023. URL <https://arxiv.org/abs/2310.02244>.

1026 Greg Yang, James B. Simon, and Jeremy Bernstein. A spectral condition for feature learning, 2024.  
 1027 URL <https://arxiv.org/abs/2310.17813>.

1028 Guodong Zhang, Lala Li, Zachary Nado, James Martens, Sushant Sachdeva, George E. Dahl,  
 1029 Christopher J. Shallue, and Roger Grosse. Which algorithmic choices matter at which batch  
 1030 sizes? insights from a noisy quadratic model, 2019. URL [https://arxiv.org/abs/](https://arxiv.org/abs/1907.04164)  
 1031 [1907.04164](https://arxiv.org/abs/1907.04164).

1032 Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright,  
 1033 Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania,  
 1034 Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. Pytorch fsdp: Experi-  
 1035 ences on scaling fully sharded data parallel, 2023. URL [https://arxiv.org/abs/2304.](https://arxiv.org/abs/2304.11277)  
 1036 [11277](https://arxiv.org/abs/2304.11277).

1037 Çağatay Yıldız, Nishaanth Kanna Ravichandran, Prishruit Punia, Matthias Bethge, and Beyza Ermis.  
 1038 Investigating continual pretraining in large language models: Insights and implications, 2024.  
 1039 URL <https://arxiv.org/abs/2402.17400>.

## 1042 A APPENDIX

### 1043 A.1 HYPERPARAMETER OPTIMIZATION IN THE INFINITE DATA AND MODEL SIZE LIMIT

1044 We believe our observations provide useful hints on how to scale learning rate and batch size jointly  
 1045 in the infinite data and model size limits. We take the general  $\mu$ Transfer approach of tuning hyper-  
 1046 parameters for a small proxy model and then transferring them either zero-shot or according to some  
 1047 scaling rules via extrapolation, across model sizes and data horizons.

- 1048 1. If one can afford tuning a  $\mu$ P proxy model on the data horizon of the target model, then it  
 1049 is sufficient to simply perform a grid search over learning rate and batch size values to find  
 1050 the best combination, following  $\mu$ Transfer (Yang et al., 2022). As we describe in Sec. 5,  
 1051  $\mu$ Transfer has been established to successfully transfer hyperparameters to O(10B) model  
 1052 sizes, albeit with potential limitations arising from very long range extrapolation in the  
 1053 infinite width limit (Blake et al., 2024; Gunter et al., 2024).
- 1054 2. Otherwise, a proxy model has to be tuned on a shorter data horizon than the target one. In  
 1055 that case, we suggest running a 2D grid search across learning rate and batch size values  
 1056 roughly around the optimal ones, where each training follows a WSD schedule (Sec. 2.4),  
 1057 for as long as compute budget allows. We suggest both the warmup and decay of the  
 1058 schedule to be fixed to the one of the target model in *absolute number of tokens*, which in  
 1059 turn should be about 10–20% fraction of the target model horizon to be optimal (Kosson  
 1060 et al., 2024; Hägele et al., 2024). This is due to the observed drift of the learning rate  
 1061 optimum with the change of the number of steps (Appendix A.7). It is still not yet clear  
 1062 how scaling of warmup/decay length and Adam’s  $\beta_{1,2}$  parameters (which we keep constant  
 1063 in our experiments) can be incorporated into the total horizon scaling. We leave this as an  
 1064 interesting direction for future work.
- 1065 3. After the grid search, one should be able to obtain a plot similar to Fig. 1a and Fig. 3a.  
 1066 Provided long enough WSD horizon, a drift in time of the critical batch size region, asso-  
 1067 ciated to the peak of the fixed token budget curve in Fig. 1a, should be visible. Likewise,  
 1068 there should be a drift of the optimally tuned (i.e. assuming optimal learning rate is used)  
 1069 batch size in time as in Fig. 3a. Since we observe a strong correlation but still a mismatch  
 1070 between the optimally-tuned batch size and the critical batch size, we suggest the following  
 1071 approach for selecting optimal hyperparameter values:  
 1072 (a) Derive scaling rule by extrapolating the batch size optimum drift in time  $T$  based on  
 1073 Fig. 3a (in our case, approximately  $B^* \propto \sqrt{T}$ ). Estimate the expected optimal batch  
 1074 size value  $B_{\text{target}}^*$  for the target data horizon  $T_{\text{target}}$  under assumption of the optimally  
 1075 tuned learning rate.  
 1076 (b) Perform a fit to fixed-budget curves per token budget step based on data similar to  
 1077 Fig. 1a with Eq. 3.1, following the procedure of Sec. 3.1. Fit a set of extracted  $B_{\text{crit}}$  per  
 1078  
 1079

token budget with a power law function  $B_{\text{crit}} = aT^{\alpha_B} + b$  to extract the corresponding exponent  $\alpha_B$  (in our case,  $\alpha_B \approx 1$ ) and derive the expected critical batch size for the target horizon  $B_{\text{target}}^{\text{crit}}$ .

- (c) Perform the same power law fit to the  $\eta^{\text{crit}}(T)$  data and extrapolate its value to the target horizon, obtaining  $\eta_{\text{target}}^{\text{crit}}$ . Set optimal learning rate for the target horizon as:

$$\eta_{\text{target}}^* = \begin{cases} \eta_{\text{target}}^{\text{crit}} \cdot \sqrt{B_{\text{target}}^*/B_{\text{target}}^{\text{crit}}} & \text{if } B_{\text{target}}^* \leq B_{\text{target}}^{\text{crit}} \\ \eta_{\text{target}}^{\text{crit}} \cdot \sqrt{B_{\text{target}}^{\text{crit}}/B_{\text{target}}^*} & \text{if } B_{\text{target}}^* > B_{\text{target}}^{\text{crit}} \end{cases}, \quad (3)$$

where we correct the learning rate value for the corresponding  $\eta^*(B)$  scaling regime.

4. Apply optimal values of learning rate  $\eta_{\text{target}}^*$  and batch size  $B_{\text{target}}^*$  to the target model, scaled up with  $\mu\text{P}$ , and to the target training horizon. As we show in this work,  $\mu\text{P}$  does not impact the dynamics of the critical batch size evolution in the infinite data limit, therefore we expect no interference between the two limits.

We suppose it is also possible to adjust the recipe above to the continual learning setting (Çağatay Yıldız et al., 2024; Ibrahim et al., 2024): under assumption of  $\eta^{\text{crit}}$  being constant in time and of the golden path hypothesis (Vyas et al., 2024), one could indefinitely run the model training with the same learning rate but dynamically adjust the batch size to follow the critical one (peak of the fixed budget curve in Fig. 1a), or, alternatively viewed, to remain on the pareto curve of Fig. 3b (inset plot).

## A.2 ON CRITICAL BATCH SIZE AND NOISE SCALE

There are two perspectives on the critical batch size  $B_{\text{crit}}$ . Firstly, McCandlish et al. (2018) define it as a batch size which results in an optimal trade-off between data sample efficiency and gradient update step efficiency:

$$B_{\text{crit}} := \frac{E_{\text{min}}}{S_{\text{min}}}, \quad (4)$$

where  $E_{\text{min}}$  ( $S_{\text{min}}$ ) are the minimum possible number of training examples (steps) to reach a specified level of performance. Additionally, they introduce a notion of a *noise scale* (for SGD-like optimizers):

$$B_{\text{noise}}^{\text{curv}} := \frac{\text{tr}(H\Sigma)}{G^T H G}, \quad (5)$$

where  $G$  is the noiseless true gradient,  $H$  is the true hessian of the loss function and  $\Sigma$  is the minibatch covariance. For  $B \ll B_{\text{noise}}^{\text{curv}}$  one obtains the linear learning rate scaling rule, while for  $B \gg B_{\text{noise}}^{\text{curv}}$  increasing  $B$  does not yield any loss improvement.

Under assumption of the Hessian being a multiple of the identity matrix, one obtains a simplified version:

$$B_{\text{simple}}^{\text{curv}} := \frac{\text{tr}(\Sigma)}{|G|^2}, \quad (6)$$

and McCandlish et al. (2018) argue that

$$B_{\text{crit}} \approx B_{\text{noise}}^{\text{curv}} \propto B_{\text{simple}}^{\text{curv}}, \quad (7)$$

thus bridging together mathematical loss curvature and pragmatical compute resource utilization views. Approximation with  $B_{\text{simple}}^{\text{curv}}$ , being computationally less expensive to estimate, is shown to be to a good degree applicable across multiple tasks, datasets and model architectures. Both the

critical batch size and the noise scale are shown to grow in time as one progresses in the training, with the only dependence on the loss value via a power law, with parameters  $B_0$  and  $\alpha_B$  to be determined empirically (Kaplan et al., 2020):

$$B_{\text{crit}} = \frac{B_0}{L^{1/\alpha_B}}. \quad (8)$$

Notably, Smith & Le (2018) introduce from a different SDE perspective another definition of the noise scale:

$$B_{\text{noise}}^{\text{SDE}} := \eta \left( \frac{T}{B} - 1 \right) \approx \eta \frac{T}{B}, \quad (9)$$

where  $T$  is the training set size. It is suggested that one should aim at finding the optimal noise scale in the first place, rather than optimal batch size and learning rate. Within the suggested Bayesian framework, Smith & Le (2018) argue that the optimality arises from the trade-off between depth and breadth in the Bayesian evidence. In a follow-up work, Park et al. (2019) take one step further and extend the noise scale to a model width limit and introduce a modified noise scale accounting for the change of the model width in the standard (SP) and Neural Tangent Kernel (NTK) parametrizations (Jacot et al., 2020):

$$B_{\text{noise}}^{\text{norm}} := \frac{B_{\text{noise}}^{\text{SDE}}}{|w|^2}, \quad (10)$$

where  $|w|^2$  is model weight norm, normalizing  $B_{\text{noise}}^{\text{SDE}}$  to have the unit 1/loss.

The second perspective on  $B_{\text{crit}}$  is as a region where *batch invariance* breaks. Introduced by Hilton et al. (2022), batch invariance refers to a regime where the model performance remains invariant with the change of either learning rate or batch size within the corresponding scaling rule. As shown by Shallue et al. (2019), the breaking of batch invariance appears with an increase of batch size to sufficiently large values and looks like plateauing of the optimal learning rate. Zhang et al. (2019) further investigated how the critical batch size is affected by using momentum, optimizer pre-conditioning and exponential moving average (EMA).

Intriguingly, Li et al. (2024) expanded the approach of McCandlish et al. (2018) and showed that in the case of Adam, the batch invariance does not break conventionally as in the SGD case. In fact, it is always preserved, with the only difference that the  $\eta \propto \sqrt{B}$  scaling rule breaks at the peak value  $B_{\text{peak}}$  and transforms into a  $\eta \propto 1/\sqrt{B}$  rule via:

$$\eta^* = \frac{\eta^{\text{crit}}}{\frac{1}{2} \left( \sqrt{\frac{B_{\text{peak}}}{B}} + \sqrt{\frac{B}{B_{\text{peak}}}} \right)}. \quad (11)$$

They also show that  $B_{\text{peak}} \approx B_{\text{crit}}$  in the definition of McCandlish et al. (2018), therefore bridging together the two  $B_{\text{crit}}$  perspectives outlined above.

### A.3 MODEL TRAINING CONFIGURATION (CONT.)

- 24 layers, FFN expansion factor  $f_{\text{ffn}} = d_{\text{ffn}}/d_{\text{model}} = 4$ , multihead attention with the head dimension  $d_{\text{head}} = 128$ .
- GeLU activation function, Layer Normalization initialized with 1 (Ba et al., 2016), RoPE with  $\theta = 10000$  (Su et al., 2023).
- Dropout is disabled and biases are included in all layers (initialized with 0), weights are shared between the input and output embedding layers.
- FSDP parallelization scheme (Zhao et al., 2023), bfloat16 precision, FlashAttention-2 (Dao, 2023).

#### A.4 HYPERPARAMETER GRID (CONT.)

The  $(\eta, B, T, d_{\text{model}}^{\text{base}}, d_{\text{model}})$  grid is defined with the following values:

- Learning rate  $\eta$ :
  - $\{2^{-12}, 2^{-11.5}, \dots, 2^{-7}\}$  for  $d_{\text{model}}^{\text{base}} = 1024$
  - $\{2^{-11}, 2^{-10}, \dots, 2^{-6}\}$  for  $d_{\text{model}}^{\text{base}} = 256$
- Batch size  $B = \{2^{16}, 2^{18}, \dots, 2^{26}\}$  tokens
- Data horizon  $T = \{2^{30}, 2^{31}, \dots, 2^{35}\}$  tokens
- Base model width  $d_{\text{model}}^{\text{base}} = \{256, 1024\}$
- Model width  $d_{\text{model}} = \{256, 512, 1024\}$

For a configuration with  $(B = 2^{20}, d_{\text{model}}^{\text{base}} = 1024)$ , we perform longer runs with an extended set of horizons with  $\{2^{36}, 2^{37}\}$  token budgets, except for the smallest  $B = 2^{16}$  due to limited computational resources and low GPU utilization of this batch size on our hardware. A configuration for the largest batch size  $(B = 2^{26}, d_{\text{model}}^{\text{base}} = 1024)$ , we train until  $T = 2^{38}$  tokens to further establish the learning rate optimum drift (Sec. 3.2).

The total number of trainable parameters is 32M, 101M, 354M for the models with widths  $d_{\text{model}} = \{256, 512, 1024\}$ , respectively. We also train 1.3B and 5B models up until  $2^{35} \approx 34\text{B}$  tokens with three selected learning rate values for a fixed batch size of  $2^{20}$  and  $2^{24}$  tokens, respectively, in order to study learning rate sensitivity change within  $\mu\text{P}$  (Sec. 3.4). The models share the same  $\mu\text{P}$  base model with  $d_{\text{model}}^{\text{base}} = 1024$  and have the corresponding width  $d_{\text{model}} = 2048$  (1.3B) and  $d_{\text{model}} = 4096$  (5B).

#### A.5 FITTING PROCEDURE

In Sec. 3.1, we introduce the fitting procedure to the data points of Fig. 1a with a functional form of Eq. 3.1 with two parameters  $\eta_{\text{crit}}$  and  $B_{\text{crit}}$ . We perform the fit separately per token budget, as illustrated in Fig. 6, using `scipy.optimize.curve_fit()` and firstly without including error bars (which stem from the variation of the model with  $\mu\text{P}$  and random seeds). Since we observe some data points as having no uncertainties, which makes the fit computationally unstable, we repeat the fit two more times: one with adding small  $\epsilon = 10^{-15}$  as uncertainty for such data points. Then, we attribute the mean uncertainty across the other points to the points without uncertainties and perform the same fit. This procedure results in three data sets for each  $\eta_{\text{crit}}(T)$  and  $B_{\text{crit}}(T)$ , corresponding to three variations of the fitting procedure, which we treat as “systematic” uncertainty. For each data point, we assign an uncertainty as a square root of the corresponding covariance matrix element, as obtained from the fit.

We then perform a full power law  $p_{\text{crit}} = a_p T^{\alpha_p} + b_p$ ,  $p \in \{\eta, B\}$  fit with uncertainties to each of the three data sets, for each of the critical parameters  $p$ . We obtain the following values with corresponding fit uncertainties:

- No error:
  - $(a_\eta, \alpha_\eta, b_\eta) = (1.9 \cdot 10^5, -0.85, 2.9 \cdot 10^{-3})$
  - $(a_B, \alpha_B, b_B) = [8.2 \cdot 10^{-5}, 1.00, 3.0 \cdot 10^5]$
  - $(\sigma_{a_\eta}, \sigma_{\alpha_\eta}, \sigma_{b_\eta}) = (1.2 \cdot 10^6, 0.32, 2.8 \cdot 10^{-4})$
  - $(\sigma_{a_B}, \sigma_{\alpha_B}, \sigma_{b_B}) = (4.5 \cdot 10^{-4}, 0.23, 1.5 \cdot 10^5)$
- With errors +  $\epsilon$ :
  - $(a_\eta, \alpha_\eta, b_\eta) = (2.3 \cdot 10^9, -1.31, 2.7 \cdot 10^{-3})$
  - $(a_B, \alpha_B, b_B) = (2.3 \cdot 10^{-2}, 0.75, 1.9 \cdot 10^5)$
  - $(\sigma_{a_\eta}, \sigma_{\alpha_\eta}, \sigma_{b_\eta}) = (1.0 \cdot 10^{10}, 0.21, 1.1 \cdot 10^{-4})$
  - $(\sigma_{a_B}, \sigma_{\alpha_B}, \sigma_{b_B}) = (5.1 \cdot 10^{-2}, 0.09, 8.8 \cdot 10^4)$
- With errors + mean uncertainty attribution:
  - $(a_\eta, \alpha_\eta, b_\eta) = (4.5 \cdot 10^{12}, -1.68, 2.9 \cdot 10^{-3})$

$$\begin{aligned}
& - (a_B, \alpha_B, b_B) = (4.9 \cdot 10^{-7}, 1.20, 2.8 \cdot 10^5) \\
& - (\sigma_{a_\eta}, \sigma_{\alpha_\eta}, \sigma_{b_\eta}) = (4.5 \cdot 10^{13}, 0.47, 1.4 \cdot 10^{-4}) \\
& - (\sigma_{a_B}, \sigma_{\alpha_B}, \sigma_{b_B}) = (2.1 \cdot 10^{-6}, 0.17, 8.0 \cdot 10^4)
\end{aligned}$$

Since we are primarily interested in the power exponents  $\alpha_p$ , we average their values across the three variations to produce a central value for each  $p$ . For the uncertainties, we add in quadratures the variance across the three fit variations and the mean uncertainty obtained from each of the individual fits. This produces the results we supply in the main text (Eq. 3.1). To further constrain large uncertainties on  $a_p$ , we refit  $\eta_{\text{crit}}(T)$  and  $B_{\text{crit}}(T)$  time dependence data with the power exponents  $\alpha_p$  fixed to the ones obtained above. This gives us the final model parameters, which we visualize in Fig. 2 and 1b and discuss throughout the main text:

- $a_\eta = (2.0 \pm 0.3) \cdot 10^9$
- $\alpha_\eta = -1.3 \pm 0.4$
- $b_\eta = (3.1 \pm 0.1) \cdot 10^{-3}$
- $a_B = (8.0 \pm 1.3) \cdot 10^{-5}$
- $\alpha_B = 1.0 \pm 0.2$
- $b_B = (3.0 \pm 1.1) \cdot 10^5$

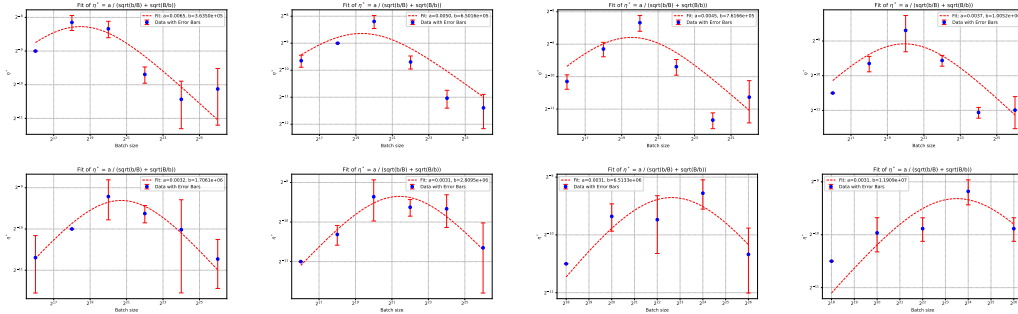


Figure 6: Fits to token budgets  $T = 2^{30}, 2^{31}, \dots, 2^{37}$  (from upper left to bottom right) with Eq. 3.1 to the data points in Fig. A.9.

## A.6 RANDOM SEED VARIATION

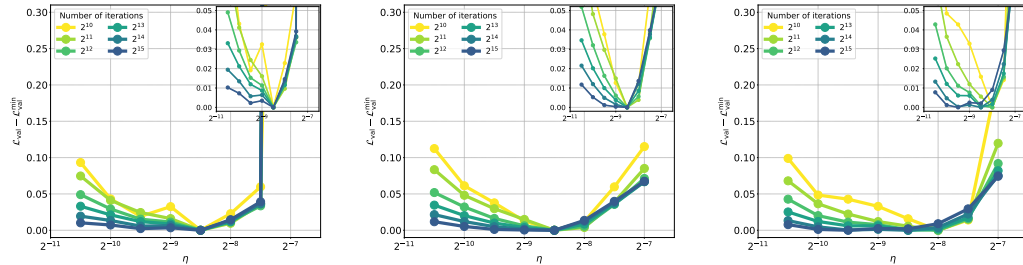


Figure 7: Loss profile  $\mathcal{L}_{\text{val}} - \mathcal{L}_{\text{val}}^{\min}$  as a function of maximum learning rate  $\eta$  for three different random seeds for the model configuration ( $d_{\text{model}} = d_{\text{model}}^{\text{base}} = 1024$ ).

## A.7 LEARNING RATE SCHEDULE SCALING (CONT.)

Conventionally, the learning rate schedule consists of a warmup phase, followed by either a constant phase or a decay phase. When all of the three phases are enabled, one obtains a warmup-stable-



decay (WSD) schedule (Hu et al., 2024):

$$\eta(t) = \begin{cases} \frac{t}{T_{\text{warmup}}} \cdot \eta_{\text{max}} & \text{if } t < T_{\text{warmup}} \\ \eta_{\text{max}} & \text{if } T_{\text{warmup}} \leq t < T - T_{\text{decay}} \\ \left(1 - \frac{t - (T - T_{\text{decay}})}{T_{\text{decay}}}\right) \cdot \eta_{\text{max}} & \text{if } T - T_{\text{decay}} \leq t < T \end{cases}, \quad (12)$$

where  $T$  is the total length of the training horizon,  $T_{\text{warmup}}$  ( $T_{\text{decay}}$ ) is the length of the warmup (decay) phases, all measured in tokens.

As Hägele et al. (2024) showed, there is no significant difference in terms of the final loss value and learning rate sensitivity between using cosine decay and WSD schedules. We run additional ablations in our setup and also arrive at the same conclusions: the structure of the learning rate optimum is marginally affected by the decay phase of the schedule and its type. Even though there appears to be a small increase in learning rate sensitivity if learning rate is decayed comparing to the schedule without decay, it does not affect the optimal  $\eta^*$  location (Fig. 8).

Furthermore, we vary the warmup scaling strategy with an increase of the data horizon, specifically where all the horizons either share the same warmup length, or warmup is scaled together with the horizon length (with the fixed  $f = T_{\text{warmup}}/T = 1/64$  fraction of the total horizon), or warmup is disabled. We observe that the addition of warmup decreases learning rate sensitivity and, interestingly, that scaling of the warmup proportionally with the horizon length leads to a drift of the learning rate optimum, as also indirectly observed earlier by Kosson et al. (2024).

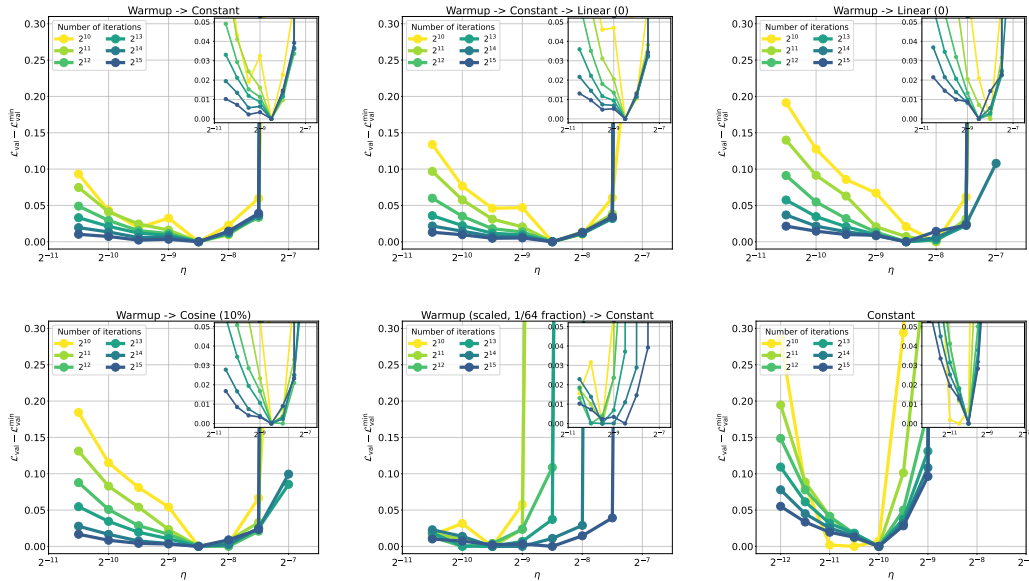


Figure 8: Loss profile  $\mathcal{L}_{\text{val}} - \mathcal{L}_{\text{val}}^{\min}$  as a function of maximum learning rate  $\eta$  for schedules with the following phases: warmup and constant (top left); warmup, constant and linear decay to 0 (top middle); warmup and linear decay to 0 (top right); warmup and cosine decay to 10% of the maximum  $\eta$  (bottom left); warmup scaled as 1/64 fraction of the total horizon and constant (bottom middle); constant (bottom right). Warmup duration is always set to  $T_{\text{warmup}} = 2^{19} = 524288$  tokens, except for the case with warmup phase scaling. The model configuration is ( $d_{\text{model}}^{\text{base}} = 1024$ ,  $d_{\text{model}} = 1024$ ,  $B = 2^{20}$ ).

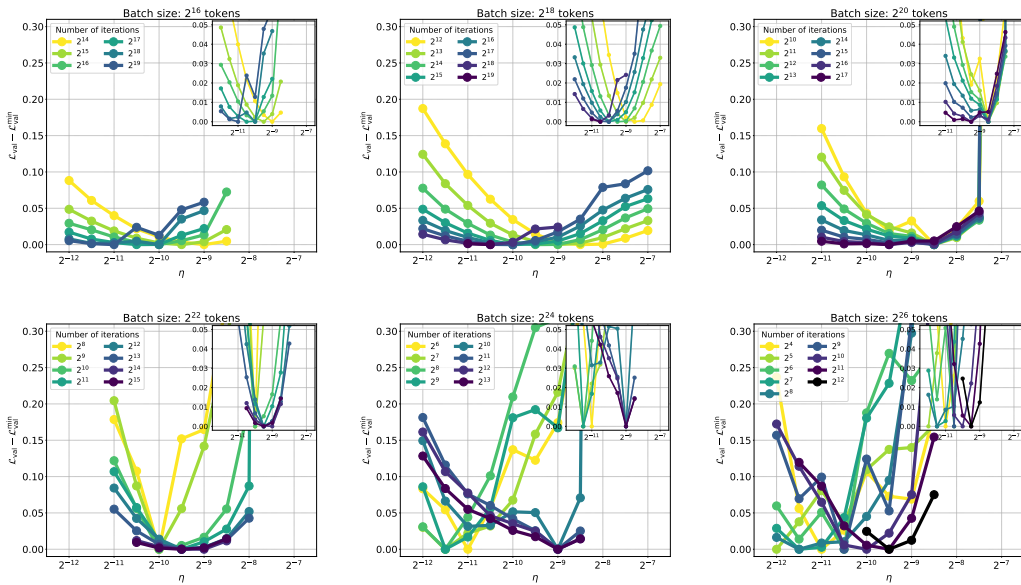
A.8 LOSS PROFILES PER  $(d_{\text{model}}^{\text{base}}, d_{\text{model}})$  CONFIGURATIONA.8.1  $d_{\text{model}}^{\text{base}} = 1024, d_{\text{model}} = 1024$ 

Figure 9: Loss profile  $\mathcal{L}_{\text{val}} - \mathcal{L}_{\text{val}}^{\min}$  as a function of maximum learning rate  $\eta$  for  $(d_{\text{model}}^{\text{base}} = 1024, d_{\text{model}} = 1024)$  for batch size  $B = 2^{16}$  (top left),  $B = 2^{18}$  (top middle),  $B = 2^{20}$  (top right),  $B = 2^{22}$  (bottom left),  $B = 2^{24}$  (bottom middle),  $B = 2^{26}$  (bottom right) across various token budgets.

1404 A.8.2  $d_{\text{model}}^{\text{base}} = 1024, d_{\text{model}} = 512$ 

1405

1406

1407

1408

1409

1410

1411

1412

1413

1414

1415

1416

1417

1418

1419

1420

1421

1422

1423

1424

1425

1426

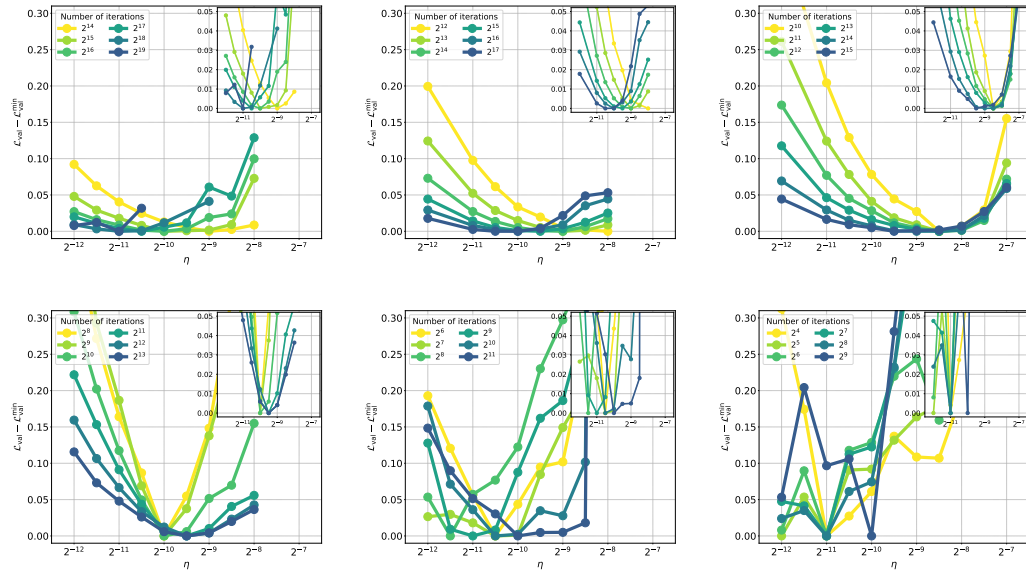


Figure 10: Loss profile  $\mathcal{L}_{\text{val}} - \mathcal{L}_{\text{val}}^{\min}$  as a function of maximum learning rate  $\eta$  for ( $d_{\text{model}}^{\text{base}} = 1024, d_{\text{model}} = 512$ ) for batch size  $B = 2^{16}$  (top left),  $B = 2^{18}$  (top middle),  $B = 2^{20}$  (top right),  $B = 2^{22}$  (bottom left),  $B = 2^{24}$  (bottom middle),  $B = 2^{26}$  (bottom right) across various token budgets.

1430

1431

1432

1433

1434

1435

1436

1437

1438

1439

1440

1441

1442

1443

1444

1445

1446

1447

1448

1449

1450

1451

1452

1453

1454

1455

1456

1457

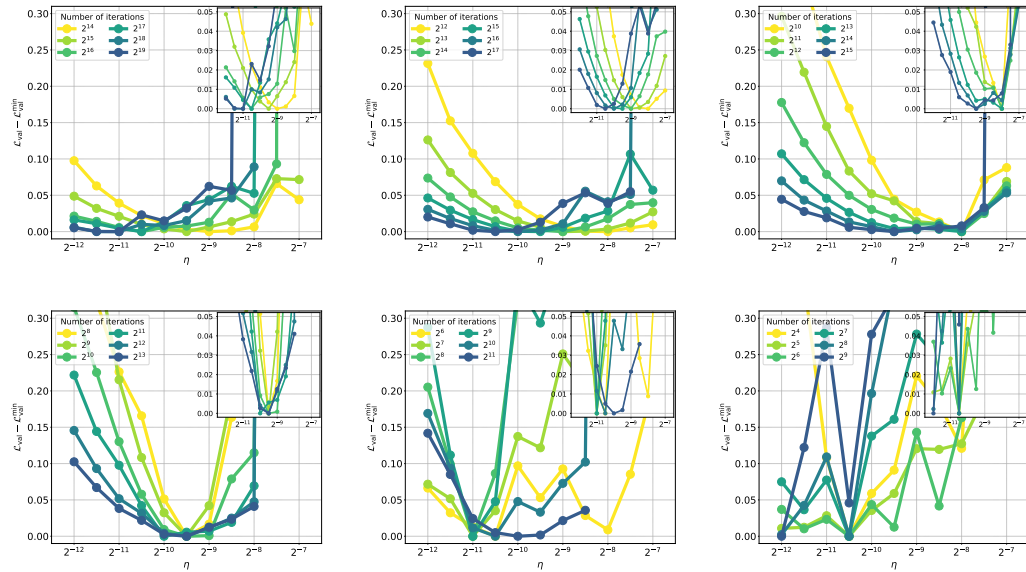
A.8.3  $d_{\text{model}}^{\text{base}} = 1024, d_{\text{model}} = 256$ 

Figure 11: Loss profile  $\mathcal{L}_{\text{val}} - \mathcal{L}_{\text{val}}^{\min}$  as a function of maximum learning rate  $\eta$  for ( $d_{\text{model}}^{\text{base}} = 1024, d_{\text{model}} = 256$ ) for batch size  $B = 2^{16}$  (top left),  $B = 2^{18}$  (top middle),  $B = 2^{20}$  (top right),  $B = 2^{22}$  (bottom left),  $B = 2^{24}$  (bottom middle),  $B = 2^{26}$  (bottom right) across various token budgets.

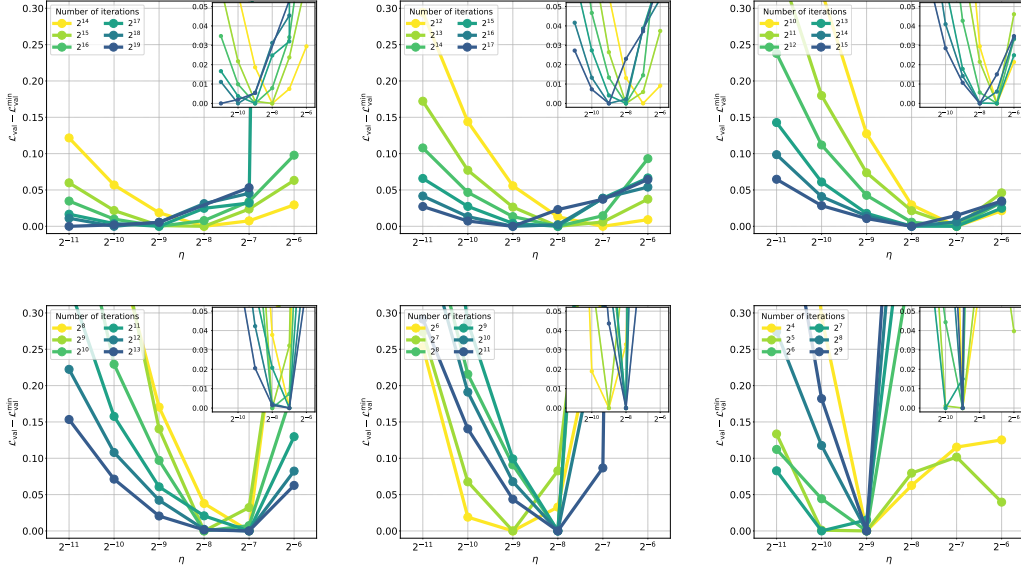
A.8.4  $d_{\text{model}}^{\text{base}} = 256, d_{\text{model}} = 256$ 

Figure 12: Loss profile  $\mathcal{L}_{\text{val}} - \mathcal{L}_{\text{val}}^{\min}$  as a function of maximum learning rate  $\eta$  for ( $d_{\text{model}}^{\text{base}} = 256, d_{\text{model}} = 256$ ) for batch size  $B = 2^{16}$  (top left),  $B = 2^{18}$  (top middle),  $B = 2^{20}$  (top right),  $B = 2^{22}$  (bottom left),  $B = 2^{24}$  (bottom middle),  $B = 2^{26}$  (bottom right) across various token budgets.

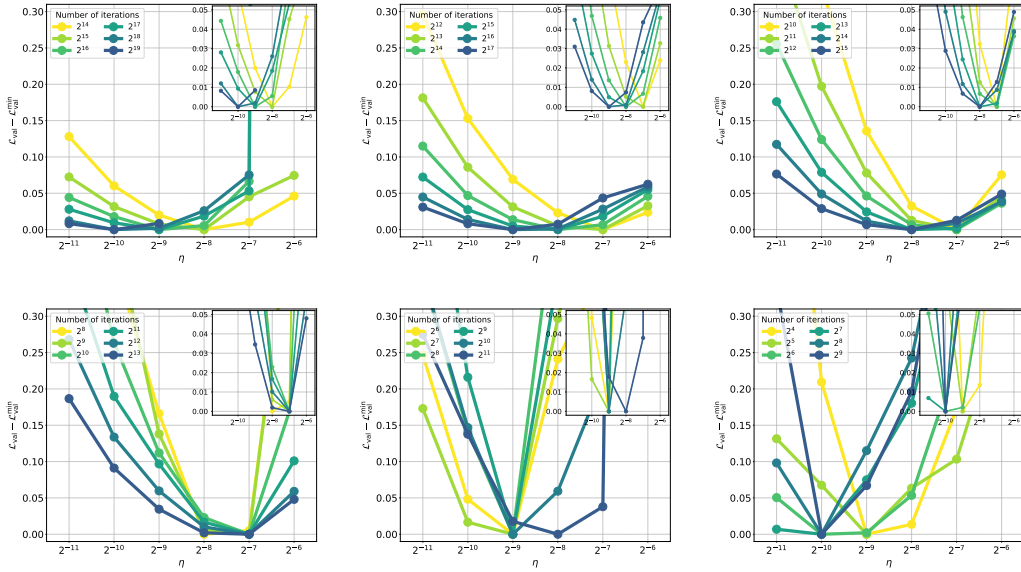
A.8.5  $d_{\text{model}}^{\text{base}} = 256, d_{\text{model}} = 512$ 

Figure 13: Loss profile  $\mathcal{L}_{\text{val}} - \mathcal{L}_{\text{val}}^{\min}$  as a function of maximum learning rate  $\eta$  for ( $d_{\text{model}}^{\text{base}} = 256, d_{\text{model}} = 512$ ) for batch size  $B = 2^{16}$  (top left),  $B = 2^{18}$  (top middle),  $B = 2^{20}$  (top right),  $B = 2^{22}$  (bottom left),  $B = 2^{24}$  (bottom middle),  $B = 2^{26}$  (bottom right) across various token budgets.

A.8.6  $d_{\text{model}}^{\text{base}} = 256, d_{\text{model}} = 1024$

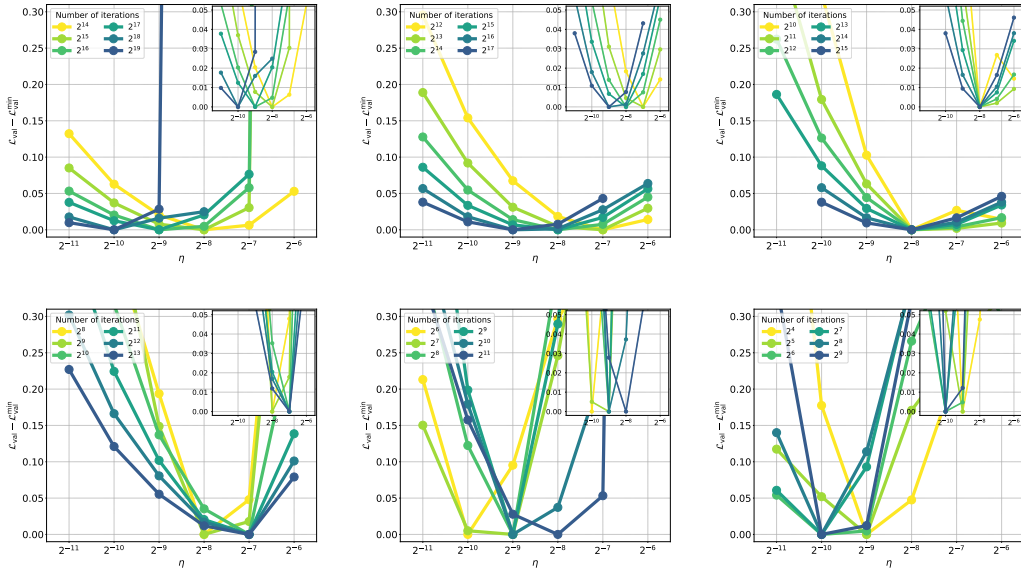


Figure 14: Loss profile  $\mathcal{L}_{\text{val}} - \mathcal{L}_{\text{val}}^{\min}$  as a function of maximum learning rate  $\eta$  for ( $d_{\text{model}}^{\text{base}} = 256, d_{\text{model}} = 1024$ ) for batch size  $B = 2^{16}$  (top left),  $B = 2^{18}$  (top middle),  $B = 2^{20}$  (top right),  $B = 2^{22}$  (bottom left),  $B = 2^{24}$  (bottom middle),  $B = 2^{26}$  (bottom right) across various token budgets.

A.9 FIG. 1 WITH THE FULL SET OF BATCH SIZE AND TOKEN BUDGET VALUES

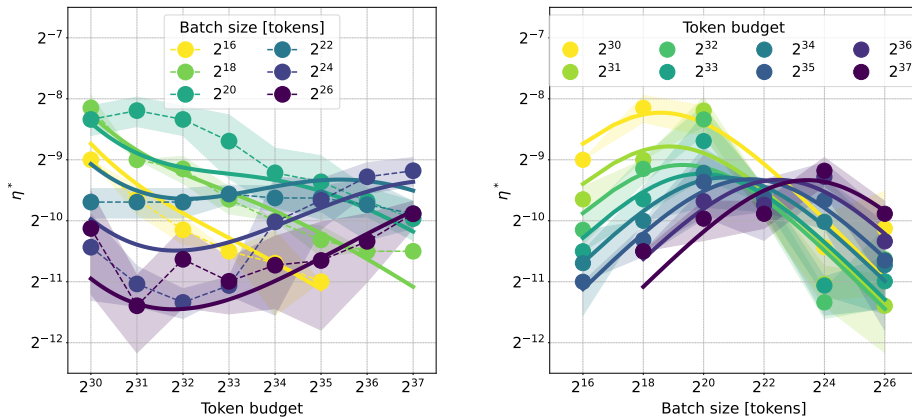


Figure 15: Same as Fig. 1 with the full set of batch size (left) and token budget (right) values.

A.10  $\mu$ P-AVERAGED OPTIMAL LEARNING RATE AND BATCH SIZE JOINT SCALING FOR  
 $d_{\text{model}}^{\text{base}} = 256$

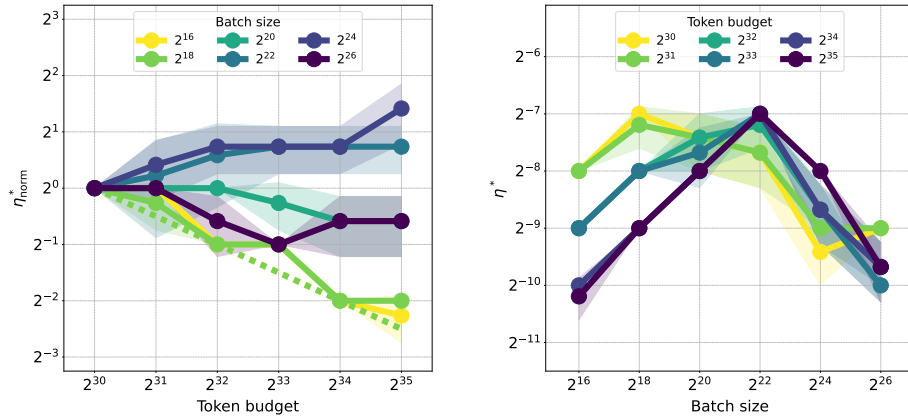


Figure 16: (left) Evolution of the optimal learning rate with an increase of the pretraining token budget  $\eta_{\text{norm}}^*(T)$ , normalized to  $\eta^*|_{T=2^{30}}$ , for a set of batch sizes (in tokens). Each point is obtained by averaging optimal learning rate values across  $\mu$ P model family, as described in Sec. 3.2. Dashed lines correspond to a square-root  $\eta^* \propto \sqrt{T^{-1}}$  scaling rule. (right) Transposition of (left): optimal learning rate  $\eta^*$  per batch size, against a range of pretraining token budgets. Each point is  $\mu$ P-averaged as in (left), with color bands visualizing the corresponding standard deviation. We note that experiments were performed with a coarser learning rate resolution of  $2^1$  compared to a  $2^{0.5}$  step in experiments with  $d_{\text{model}}^{\text{base}} = 1024$ .

## A.11 PER-MODEL OPTIMAL LEARNING RATE AND BATCH SIZE JOINT SCALING

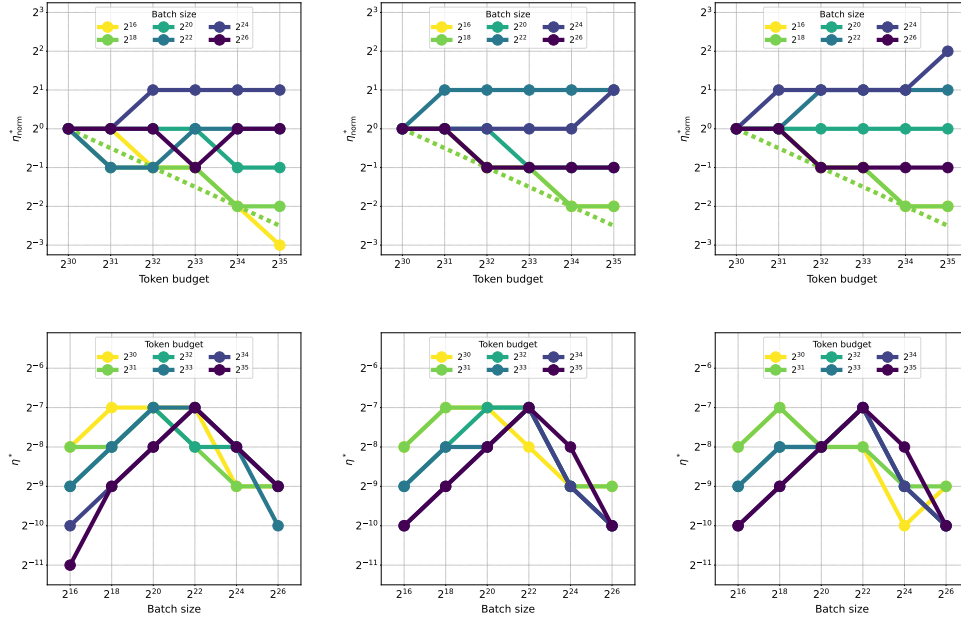
A.11.1  $d_{\text{model}}^{\text{base}} = 256$ 

Figure 17: Individual curves contributing to Fig. A.10 for models with  $d_{\text{model}} = 256$  (left column), 512 (middle column), 1024 (right column) showing evolution of the normalized to  $T = 2^{30}$  tokens optimal learning rate  $\eta_{\text{norm}}^*$  in time per batch size (top row), and joint optimal  $(\eta, B)$  curves per token budget (bottom row), for  $d_{\text{model}}^{\text{base}} = 256$ .

1674 A.11.2  $d_{\text{model}}^{\text{base}} = 1024$

1675  
1676  
1677  
1678  
1679  
1680  
1681  
1682  
1683  
1684  
1685  
1686  
1687  
1688  
1689  
1690  
1691  
1692  
1693  
1694  
1695  
1696  
1697  
1698  
1699  
1700  
1701  
1702  
1703  
1704  
1705  
1706  
1707  
1708  
1709  
1710  
1711  
1712  
1713  
1714  
1715  
1716  
1717  
1718  
1719  
1720  
1721  
1722  
1723  
1724  
1725  
1726  
1727

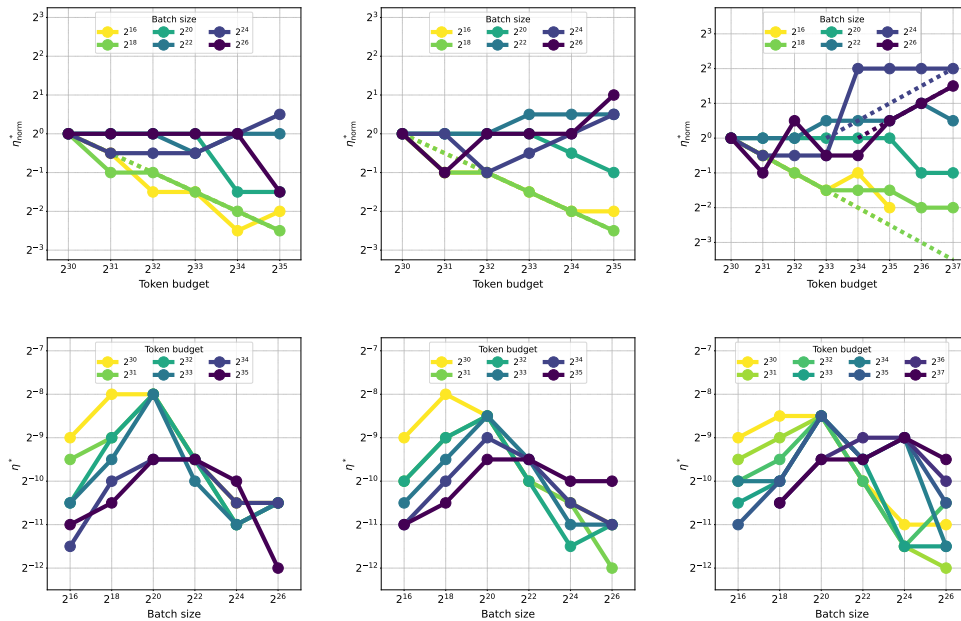


Figure 18: Individual curves contributing to Fig. 1 for models with  $d_{\text{model}} = 256$  (left column), 512 (middle column), 1024 (right column) showing evolution of the normalized to  $T = 2^{30}$  tokens optimal learning rate  $\eta_{\text{norm}}^*$  in time per batch size (top row), and joint optimal  $(\eta, B)$  curves per token budget (bottom row), for  $d_{\text{model}}^{\text{base}} = 1024$ .



A.12 PER-MODEL VALIDATION LOSS EVOLUTION IN TIME DEPENDING ON BATCH SIZE WITH OPTIMALLY-TUNED LEARNING RATE

A.12.1  $d_{\text{model}}^{\text{base}} = 256$

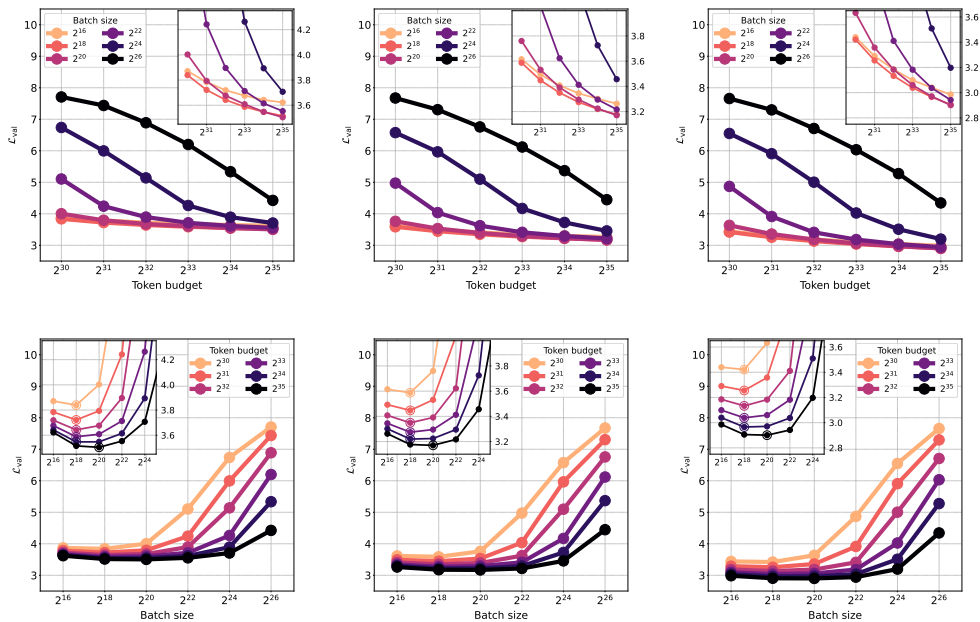


Figure 19: Analogue of Fig. 3b (top row) and Fig. 3a (bottom row) for models with widths  $d_{\text{model}} = 256$  (left column), 512 (middle column), 1024 (right column) and the base model width  $d_{\text{model}}^{\text{base}} = 256$ .

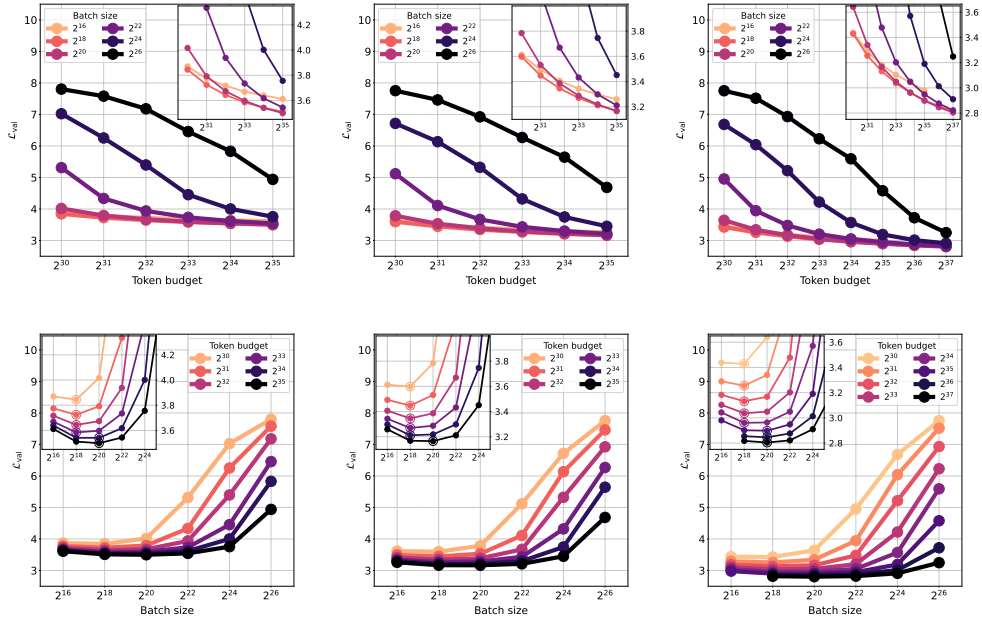
A.12.2  $d_{\text{model}}^{\text{base}} = 1024$ 

Figure 20: Analogue of Fig. 3b (top row) and Fig. 3a (bottom row) for models with widths  $d_{\text{model}} = 256$  (left column), 512 (middle column), 1024 (right column) and the base model width  $d_{\text{model}}^{\text{base}} = 1024$ .

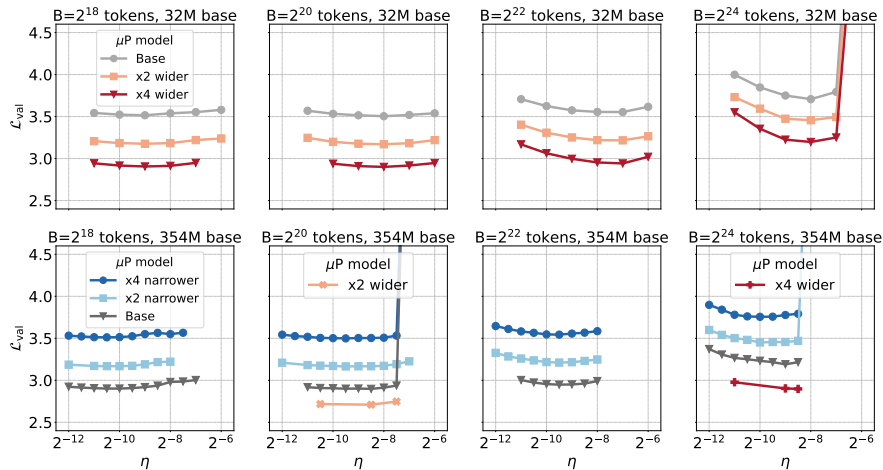
A.13 LEARNING RATE SENSITIVITY IN THE  $\mu\text{P}$  WIDTH LIMIT

Figure 21: Same as Fig. 5 but without y-axis normalization with  $\mathcal{L}_{\text{val}}^{\text{min}}$ .