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 η and batch size B. While techniques like μP (Yang et al., 2022) provide scaling rules for optimal η transfer in the infinite model size limit, the optimal scaling behavior in the infinite data size limit ($T \to \infty$) remains unknown. We fill in this gap by observing for the first time an intricate dependence of optimal η scaling on B and its relation to the critical batch size B_{crit}, which we measure to evolve in $T \to \infty$ as $B_{\mathrm{crit}} \propto T$. Furthermore, we show that the optimal batch size is positively correlated with B_{crit}: keeping it fixed becomes suboptimal over time even if learning rate is scaled optimally. Surprisingly, our results demonstrate that the observed optimal η and B dynamics are preserved with μP model scaling, challenging the conventional view of B_{crit} dependence solely on loss value. Complementing optimality, we examine the sensitivity of loss to changes in learning rate, where we find the sensitivity to decrease with $T \to \infty$ and to remain constant with μP 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, μTransfer (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 (μP). 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 μTransfer 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 μTransfer 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 μTransfer'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 μTransfer.

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 η^* 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 η^* evolution with T on the batch size B and its relation to the critical batch size B_{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_{crit} and T, which we find compatible with experimental observations. Furthermore, we find these dynamics to be largely preserved within μP (Appendix A.11).
- **Optimal batch size scaling:** assuming η 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_{crit} (Fig. 2, left), with $B^*(T) < B_{\mathrm{crit}}(T)$ in our measured range of T. Importantly, we show that naïve application of optimal η scaling rules in the $T \to \infty$ limit with B being indefinitely fixed becomes suboptimal over time: a joint (η, B) scaling is required.
- **Critical batch size:** we experimentally find B_{crit} (see definition in Sec. 2.1) to evolve in time with $B_{\mathrm{crit}} \propto T^{\alpha_B}$ and $\alpha_B 1.0 \pm 0.2$ (Fig. 2, left). This dynamic affects optimal η scaling via Eq. 3.1 and drives the transition between various scaling behaviors (Sec. 3.2). Surprisingly, we show evidence that B_{crit} is not exclusively defined by the value of the loss function (Eq. 8) as suggested by McCandlish et al. (2018): models within μP share the same B_{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 μP base model and within the μP 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* $T \to \infty$, as opposed to a fixed budget regime with $T = \mathrm{const}$, refers to an (infinite) increase of the number of tokens seen by the model during pretraining.

μP: we refer to a model with width $d_{\mathrm{model}}^{\mathrm{base}}$ as a *base model* if μP 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 μP where the base model "pinpoints" the training dynamics for all the other models obtained either by scaling up or down the base $d_{\mathrm{model}}^{\mathrm{base}}$ width. Together with the base model, we refer to this set of models as a *μP model family* or as a *μP 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 μP model family, while the latter is a model used to tune hyperparameters to be transferred with μTransfer to a target model.

Critical batch size (B_{crit}): following Li et al. (2024), we define B_{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_{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 η/η^*. 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.

μP is implemented according to Table 8 of Yang et al. (2022), so that when d_{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 μP. 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_{head} being always fixed and the number of heads being scaled proportionally to the width scaling.

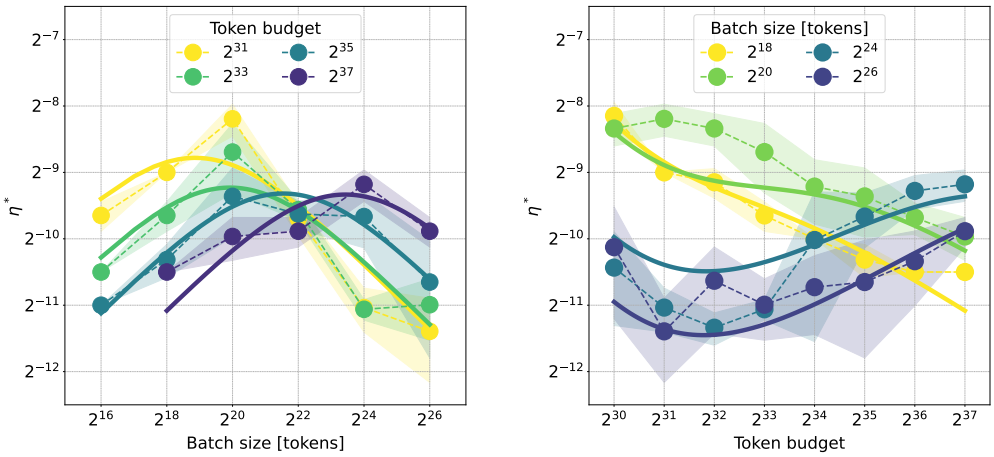
The models are trained with the causal language modeling task on the train split of the Colossal Clean Crawled Corpus (C4) dataset (Raffel et al., 2020), tokenized with the GPT2 tokenizer (Radford et al., 2019) with a vocabulary size of 50257 and a context length of 1024 tokens. As a metric to evaluate 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 $T \to \infty$, we define a 5D grid spanned by the following axes: η, B, T, d_{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 η^* evolve in time T depending on the choice of batch size B. As this measurement is moreover conditioned on the μP 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 μP trajectory with different widths d_{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 μP 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 $\eta_{\max}$ and a constant phase with learning rate fixed at $\eta_{\max}$, hereafter notated as η. Our version omits the decay phase to simplify experimentation as we observe that it does not affect the optimal η 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.



(a) (b)

Figure 1: **(a)**: Optimal learning rate η^* 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 μP 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_{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 μP

First, we begin with setting $d_{\text{model}}^{\text{base}} = 1024$ and scanning learning rate across different batch sizes and d_{model}. We present results for the η^* 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 μP models[1] 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 (η^*, 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.

[1]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 μTransfer to share the same optimal learning rate. From the experimental side, we also observe no significant differences across the three models (Appendix A.11).

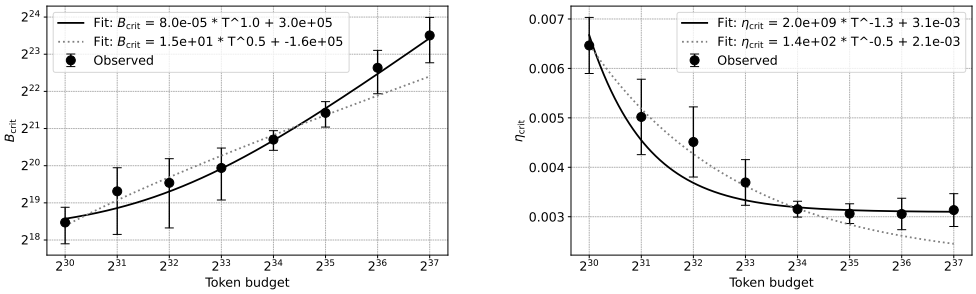


Figure 2: Critical batch size B_{crit} (left) and critical learning rate η_{crit} (right), as extracted from the fit with the power law $p_{\text{crit}} = a_p T^{\alpha_p} + b_p$, where $p \in \{\eta, B\}$, following the procedure from Appendix A.5, as a function of token budget. Solid line represents 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.

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:

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

where B_{crit} and η_{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. $B(T) = B_{\text{crit}}(T)$.

After performing the fit for each token budget, we analyze the η_{crit} and B_{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:

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

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

Lastly, there is a difference of B_{crit} evolution between the T and μP infinite width limits. Specifically, for a fixed token budget, we observe no significant change of the curves' shapes and peak positions across d_{model} values within the same μP trajectory, and also with the change of the base model (Appendix A.11). At the same time, there is a noticeable drift of B_{crit} in the $T \to \infty$ limit with the model being fixed. As both limits are accompanied with a comparable change of the model performance[2], 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.

[2]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 $\mathcal{L}_{\text{val}} = 3.4 \to 2.8$ with a token budget increase $2^{31} \to 2^{37}$, resulting in B_{crit} drifting by 2^4. For the same $(d_{\text{model}}^{\text{base}}, B)$ configuration, there is no significant B_{crit} drift with a change of width by 2^2 within μP, but the corresponding loss change is $\mathcal{L}_{\text{val}} = 3.5 \to 2.9$.

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 η^* in time T for a representative set of batch size values, with the full set of batch size values in Appendix A.9. Overlayed, 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 η^* 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[3] 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 η and B would better constrain the model fit and therefore constitute an important next step in future work.

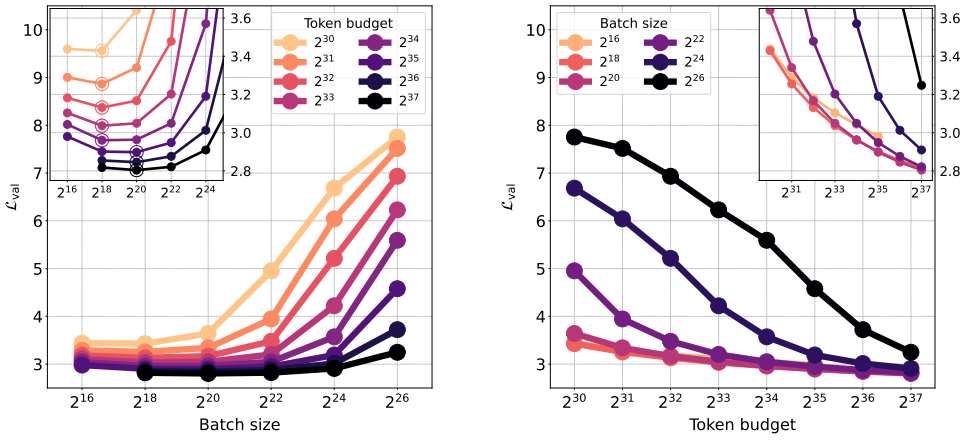
It is instructive to consider several limiting scaling scenarios of Eq. 3.1. First, when $B \ll B_{\text{crit}}$, one obtains $\eta^*(T, B) \propto \eta_{\text{crit}}(T)/\sqrt{B_{\text{crit(T)}}}$, which we do not observe, since $\min(B_{\text{crit}}) = 2^{18}$ in our experiments. Second, when $B \gg B_{\text{crit}}$, 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 $T \to \infty$ 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.



(a) (b)

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 μP

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 μP 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 μP 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 $\mathcal{L}_{\text{val}} - \mathcal{L}_{\text{val}}^{\min}$ value, where $\mathcal{L}_{\text{val}}^{\min} = \mathcal{L}_{\text{val}}(\eta^*)$ 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 μP 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 μP (Fig. 5 with and Fig. 21 without $\mathcal{L}_{\text{val}}$ normalization). Evaluated for the data horizon of $T = 2^{35} \approx 34$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

[3]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 (η^*, B) behavior per token budget (illustrated in Fig. 1a), followed by fitting the time dynamics (illustrated in Fig. 2).

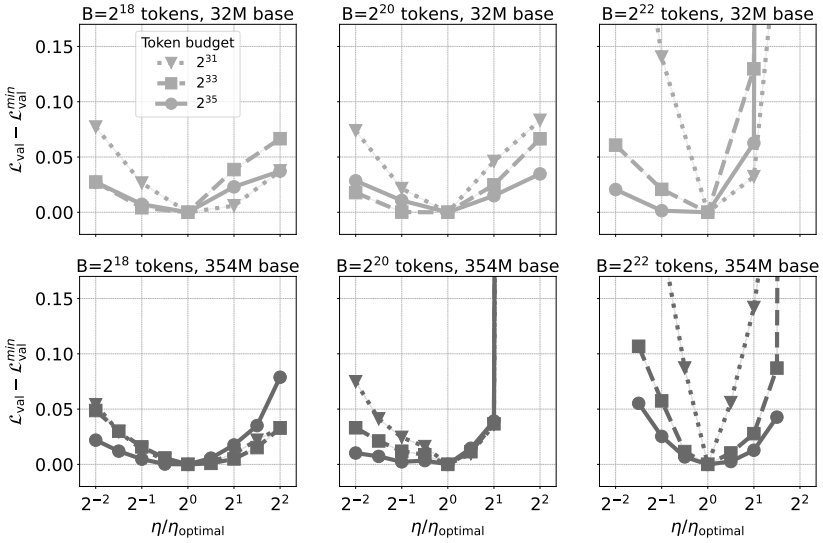


Figure 4: Learning rate sensitivity $\mathcal{L}_{\text{val}} - \mathcal{L}_{\text{val}}^{\min}$ as a function of the learning rate deviation from the optimal value $\eta/\eta_{\text{optimal}}$, measured for batch sizes of $B = 2^{18}$ (left column), 2^{20} (middle column), and 2^{22} (right column) tokens, separately for the μP 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.

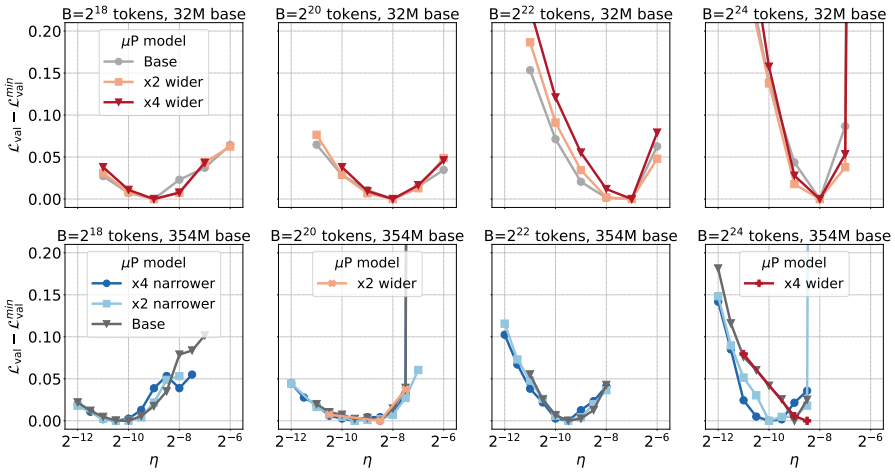


Figure 5: Learning rate sensitivity $\mathcal{L}_{\text{val}} - \mathcal{L}_{\text{val}}^{\min}$ as a function of learning rate η, measured for batch sizes of $B = 2^{18}$ (leftmost column), 2^{20} (middle left column), 2^{22} (middle right column) and 2^{24} (rightmost column) tokens, separately for the μP 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 μP 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 hyper-parameter optimization in the infinite data and model size limit. We provide our recommendations in Appendix A.1.

As future work, it would important to improve the resolution in data points with a finer grid of (η, 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 μP 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

(η, 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 (Granziol 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 η), 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 (η, B) scaling laws.

μP Originally developed within the Tensor Program series studying feature learning in the infinite width limit (Yang & Hu, 2022; Yang et al., 2022), μP 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 μTransfer 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 μP and experimentally measured (η, B) scaling rules to allow for the hyperparameter transfer in the $T \to \infty$ 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 μP 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 and batch size dynamics. We observed an intricate interplay of three optimal learning rate scaling regimes in the infinite data limit, controlled by the batch size in its relation to the critical batch size as it evolves in time. This dynamic is preserved during model scaling with μP, as well as the loss sensitivity to the learning rate variation, highlighting the intriguing difference in how μP 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-μp: 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.

DeepSeek-AI, :, Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, Guangbo Hao, Zhewen Hao, Ying He, Wenjie Hu, Panpan Huang, Erhang Li, Guowei Li, Jiashi Li, Yao Li, Y. K. Li, Wenfeng Liang, Fangyun Lin, A. X. Liu, Bo Liu, Wen Liu, Xiaodong Liu, Xin Liu, Yiyuan Liu, Haoyu Lu, Shanghao Lu, Fuli Luo, Shirong Ma, Xiaotao Nie, Tian Pei, Yishi Piao, Junjie Qiu, Hui Qu, Tongzheng Ren, Zehui Ren, Chong Ruan, Zhangli Sha, Zhihong Shao, Junxiao Song, Xuecheng Su, Jingxiang Sun, Yaofeng Sun, Minghui Tang, Bingxuan Wang, Peiyi Wang, Shiyu Wang, Yaohui Wang, Yongji Wang, Tong Wu, Y. Wu, Xin Xie, Zhenda Xie, Ziwei Xie, Yiliang Xiong, Hanwei Xu, R. X. Xu, Yanhong Xu, Dejian Yang, Yuxiang You, Shuiping Yu, Xingkai Yu, B. Zhang, Haowei Zhang, Lecong Zhang, Liyue Zhang, Mingchuan Zhang, Minghua Zhang, Wentao Zhang, Yichao Zhang, Chenggang Zhao, Yao Zhao, Shangyan Zhou, Shunfeng Zhou, Qihao Zhu, and Yuheng Zou. Deepseek llm: Scaling open-source language models with longtermism, 2024. URL https://arxiv.org/abs/2401.02954.

Nolan Dey, Quentin Anthony, and Joel Hestness. The practitioner's guide to the maximal update parameterization, 2024. URL https://www.cerebras.ai/blog/the-practitioners-guide-to-the-maximal-update-parameterization.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, 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

Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Katie Everett, Lechao Xiao, Mitchell Wortsman, Alexander A. Alemi, Roman Novak, Peter J. Liu, Izzeddin Gur, Jascha Sohl-Dickstein, Leslie Pack Kaelbling, Jaehoon Lee, and Jeffrey Pennington. Scaling exponents across parameterizations and optimizers, 2024. URL https://arxiv.org/abs/2407.05872.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour, 2018. URL https://arxiv.org/abs/1706.02677.

Diego Granziol, Stefan Zohren, and Stephen Roberts. Learning rates as a function of batch size: A random matrix theory approach to neural network training, 2021. URL https://arxiv.org/abs/2006.09092.

Tom Gunter, Zirui Wang, Chong Wang, Ruoming Pang, Andy Narayanan, Aonan Zhang, Bowen Zhang, Chen Chen, Chung-Cheng Chiu, David Qiu, Deepak Gopinath, Dian Ang Yap, Dong Yin, Feng Nan, Floris Weers, Guoli Yin, Haoshuo Huang, Jianyu Wang, Jiarui Lu, John Peebles, Ke Ye, Mark Lee, Nan Du, Qibin Chen, Quentin Keunebroek, Sam Wiseman, Syd Evans, Tao Lei, Vivek Rathod, Xiang Kong, Xianzhi Du, Yanghao Li, Yongqiang Wang, Yuan Gao, Zaid Ahmed, Zhaoyang Xu, Zhiyun Lu, Al Rashid, Albin Madappally Jose, Alec Doane, Alfredo Bencomo, Allison Vanderby, Andrew Hansen, Ankur Jain, Anupama Mann Anupama, Areeba Kamal, Bugu Wu, Carolina Brum, Charlie Maalouf, Chinguun Erdenebileg, Chris Dulhanty, Dominik Moritz, Doug Kang, Eduardo Jimenez, Evan Ladd, Fangping Shi, Felix Bai, Frank Chu, Fred Hohman, Hadas Kotek, Hannah Gillis Coleman, Jane Li, Jeffrey Bigham, Jeffery Cao, Jeff Lai, Jessica Cheung, Jiulong Shan, Joe Zhou, John Li, Jun Qin, Karanjeet Singh, Karla Vega, Kelvin Zou, Laura Heckman, Lauren Gardiner, Margit Bowler, Maria Cordell, Meng Cao, Nicole Hay, Nilesh Shahdadpuri, Otto Godwin, Pranay Dighe, Pushyami Rachapudi, Ramsey Tantawi, Roman Frigg, Sam Davarnia, Sanskruti Shah, Saptarshi Guha, Sasha Sirovica, Shen Ma, Shuang Ma, Simon Wang, Sulgi Kim, Suma Jayaram, Vaishaal Shankar, Varsha Paidi, Vivek Kumar, Xin Wang, Xin Zheng, Walker Cheng, Yael Shrager, Yang Ye, Yasu Tanaka, Yihao Guo, Yunsong Meng, Zhao Tang Luo, Zhi Ouyang, Alp Aygar, Alvin Wan, Andrew Walkingshaw, Andy Narayanan, Antonie Lin, Arsalan Farooq, Brent Ramerth, Colorado Reed, Chris Bartels, Chris Chaney, David Riazati, Eric Liang Yang, Erin Feldman, Gabriel Hochstrasser, Guillaume Seguin, Irina Belousova, Joris Pelemans, Karen Yang, Keivan Alizadeh Vahid, Liangliang Cao, Mahyar Najibi, Marco Zuliani, Max Horton, Minsik Cho, Nikhil Bhendawade, Patrick Dong, Piotr Maj, Pulkit Agrawal, Qi Shan, Qichen Fu, Regan Poston, Sam Xu, Shuangning Liu, Sushma Rao, Tashweena Heeramun, Thomas Merth, Uday Rayala, Victor Cui, Vivek Rangarajan Sridhar, Wencong Zhang, Wenqi Zhang, Wentao Wu, Xingyu Zhou, Xinwen Liu, Yang Zhao, Yin Xia, Zhile Ren, and Zhongzheng Ren. Apple intelligence foundation language models, 2024. URL https://arxiv.org/abs/2407.21075.

Jacob Hilton, Karl Cobbe, and John Schulman. Batch size-invariance for policy optimization, 2022. URL https://arxiv.org/abs/2110.00641.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022. URL https://arxiv.org/abs/2203.15556.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm: Unveiling the potential of small language models with scalable training strategies, 2024. URL https://arxiv.org/abs/2404.06395.

Alexander Hägele, Elie Bakouch, Atli Kosson, Loubna Ben Allal, Leandro Von Werra, and Martin Jaggi. Scaling laws and compute-optimal training beyond fixed training durations, 2024. URL https://arxiv.org/abs/2405.18392.

Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats L. Richter, Quentin Anthony, Timothée Lesort, Eugene Belilovsky, and Irina Rish. Simple and scalable strategies to continually pre-train large language models, 2024. URL https://arxiv.org/abs/2403.08763.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks, 2020. URL https://arxiv.org/abs/1806.07572.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima, 2017. URL https://arxiv.org/abs/1609.04836.

Atli Kosson, Bettina Messmer, and Martin Jaggi. Analyzing & eliminating learning rate warmup in GPT pre-training. In *High-dimensional Learning Dynamics 2024: The Emergence of Structure and Reasoning*, 2024. URL https://openreview.net/forum?id=RveSp5oESA.

Tim Large, Yang Liu, Minyoung Huh, Hyojin Bahng, Phillip Isola, and Jeremy Bernstein. Scalable optimization in the modular norm, 2024. URL https://arxiv.org/abs/2405.14813.

Shuaipeng Li, Penghao Zhao, Hailin Zhang, Xingwu Sun, Hao Wu, Dian Jiao, Weiyan Wang, Chengjun Liu, Zheng Fang, Jinbao Xue, Yangyu Tao, Bin Cui, and Di Wang. Surge phenomenon in optimal learning rate and batch size scaling, 2024. URL https://arxiv.org/abs/2405.14578.

Lucas Lingle. A large-scale exploration of μ-transfer, 2024. URL https://arxiv.org/abs/2404.05728.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Sadhika Malladi, Kaifeng Lyu, Abhishek Panigrahi, and Sanjeev Arora. On the sdes and scaling rules for adaptive gradient algorithms, 2023. URL https://arxiv.org/abs/2205.10287.

Sam McCandlish, Jared Kaplan, Dario Amodei, and OpenAI Dota Team. An empirical model of large-batch training, 2018. URL https://arxiv.org/abs/1812.06162.

MosaicML. Introducing mpt-7b: A new standard for open-source, commercially usable llms, 2023. URL www.mosaicml.com/blog/mpt-7b. Accessed: 2023-05-05.

MosaicML. Llm foundry, 2024. URL https://github.com/mosaicml/llm-foundry.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, 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

Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Daniel S. Park, Jascha Sohl-Dickstein, Quoc V. Le, and Samuel L. Smith. The effect of network width on stochastic gradient descent and generalization: an empirical study, 2019. URL https://arxiv.org/abs/1905.03776.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Christopher J. Shallue, Jaehoon Lee, Joseph Antognini, Jascha Sohl-Dickstein, Roy Frostig, and George E. Dahl. Measuring the effects of data parallelism on neural network training, 2019. URL https://arxiv.org/abs/1811.03600.

Yikang Shen, Matthew Stallone, Mayank Mishra, Gaoyuan Zhang, Shawn Tan, Aditya Prasad, Adriana Meza Soria, David D. Cox, and Rameswar Panda. Power scheduler: A batch size and token number agnostic learning rate scheduler, 2024. URL https://arxiv.org/abs/2408.13359.

Samuel L. Smith and Quoc V. Le. A bayesian perspective on generalization and stochastic gradient descent, 2018. URL https://arxiv.org/abs/1710.06451.

Samuel L. Smith, Pieter-Jan Kindermans, Chris Ying, and Quoc V. Le. Don't decay the learning rate, increase the batch size, 2018. URL https://arxiv.org/abs/1711.00489.

Samuel L. Smith, Erich Elsen, and Soham De. On the generalization benefit of noise in stochastic gradient descent, 2020. URL https://arxiv.org/abs/2006.15081.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023. URL https://arxiv.org/abs/2104.09864.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal

Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, 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 Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan,

Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kepa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Woj-

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

Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. Gemini: A family of highly capable multimodal models, 2024. URL https://arxiv.org/abs/2312.11805.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Gal Kaplun, Sham Kakade, and Boaz Barak. Beyond implicit bias: The insignificance of sgd noise in online learning, 2024. URL https://arxiv.org/abs/2306.08590.

Mitchell Wortsman, Peter J. Liu, Lechao Xiao, Katie Everett, Alex Alemi, Ben Adlam, John D. Co-Reyes, Izzeddin Gur, Abhishek Kumar, Roman Novak, Jeffrey Pennington, Jascha Sohl-dickstein, Kelvin Xu, Jaehoon Lee, Justin Gilmer, and Simon Kornblith. Small-scale proxies for large-scale transformer training instabilities, 2023. URL https://arxiv.org/abs/2309.14322.

Greg Yang and Edward J. Hu. Feature learning in infinite-width neural networks, 2022. URL https://arxiv.org/abs/2011.14522.

Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer, 2022. URL https://arxiv.org/abs/2203.03466.

Greg Yang, Dingli Yu, Chen Zhu, and Soufiane Hayou. Tensor programs vi: Feature learning in infinite-depth neural networks, 2023. URL https://arxiv.org/abs/2310.02244.

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

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

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

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

A Appendix

A.1 Hyperparameter optimization in the infinite data and model size limit

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

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

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

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

$$\eta^*_{\text{target}} = \begin{cases} \eta^{\text{crit}}_{\text{target}} \cdot \sqrt{B^*_{\text{target}}/B^{\text{crit}}_{\text{target}}} & \text{if } B^*_{\text{target}} \leq B^{\text{crit}}_{\text{target}} \\ \eta^{\text{crit}}_{\text{target}} \cdot \sqrt{B^{\text{crit}}_{\text{target}}/B^*_{\text{target}}} & \text{if } B^*_{\text{target}} > B^{\text{crit}}_{\text{target}} \end{cases}, \tag{3}$$

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

4. Apply optimal values of learning rate η^*_{target} and batch size B^*_{target} to the target model, scaled up with μP, and to the target training horizon. As we show in this work, μP 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 η^{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_{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:

$$B_{\text{crit}} \coloneqq \frac{E_{\min}}{S_{\min}}, \tag{4}$$

where $E_{\min}$ ($S_{\min}$) 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):

$$B^{\text{curv}}_{\text{noise}} \coloneqq \frac{tr(H\Sigma)}{G^T H G}, \tag{5}$$

where G is the noiseless true gradient, H is the true hessian of the loss function and Σ is the minibatch covariance. For $B \ll B^{\text{curv}}_{\text{noise}}$ one obtains the linear learning rate scaling rule, while for $B \gg B^{\text{curv}}_{\text{noise}}$ 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:

$$B^{\text{curv}}_{\text{simple}} \coloneqq \frac{tr(\Sigma)}{|G^2|}, \tag{6}$$

and McCandlish et al. (2018) argue that

$$B_{\text{crit}} \approx B^{\text{curv}}_{\text{noise}} \propto B^{\text{curv}}_{\text{simple}}, \tag{7}$$

thus bridging together mathematical loss curvature and pragmatical compute resource utilization views. Approximation with $B^{\text{curv}}_{\text{simple}}$, 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 α_B to be determined empirically (Kaplan et al., 2020):

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

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

$$B_{\text{noise}}^{\text{SDE}} \coloneqq \eta(\frac{T}{B} - 1) \approx \eta\frac{T}{B}, \tag{9}$$

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):

$$B_{\text{noise}}^{\text{norm}} \coloneqq \frac{B_{\text{noise}}^{\text{SDE}}}{|w|^2}, \tag{10}$$

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

The second perspective on B_{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_{peak} and transforms into a $\eta \propto 1/\sqrt{B}$ rule via:

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

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

- Learning rate η:
  - $\{2^{-12},\ 2^{-11.5}, \ldots,\ 2^{-7}\}$ for $d_{\text{model}}^{\text{base}} = 1024$
  - $\{2^{-11},\ 2^{-10}, \ldots,\ 2^{-6}\}$ for $d_{\text{model}}^{\text{base}} = 256$
- Batch size $B = \{2^{16},\ 2^{18}, \ldots,\ 2^{26}\}$ tokens
- Data horizon $T = \{2^{30},\ 2^{31} \ldots,\ 2^{35}\}$ 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 $2^{35} \approx 34$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 μP (Sec. 3.4). The models share the same μP 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 η_{crit} and B_{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 μP 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 + ϵ:
  - $(a_\eta, \alpha_\eta, b_\eta) = (2.3 \cdot 10^9, -1.31, 2.7 \cdot 10^{-3})$
  - $(a_B, \alpha_B, b_B) = (2.3 \cdot 10^{-2}, 0.75, 1.9 \cdot 10^5)$
  - $(\sigma_{a_\eta}, \sigma_{\alpha_\eta}, \sigma_{b_\eta}) = (1.0 \cdot 10^{10}, 0.21, 1.1 \cdot 10^{-4})$
  - $(\sigma_{a_B}, \sigma_{\alpha_B}, \sigma_{b_B}) = (5.1 \cdot 10^{-2}, 0.09, 8.8 \cdot 10^4)$
- With errors + mean uncertainty attribution:
  - $(a_\eta, \alpha_\eta, b_\eta) = (4.5 \cdot 10^{12}, -1.68, 2.9 \cdot 10^{-3})$

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

Since we are primarily interested in the power exponents α_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 α_p fixed to the ones obtained above. This gives us the final model parameters, which we visualize oi Fig. 2 and 1b and discuss throughout the main text:

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

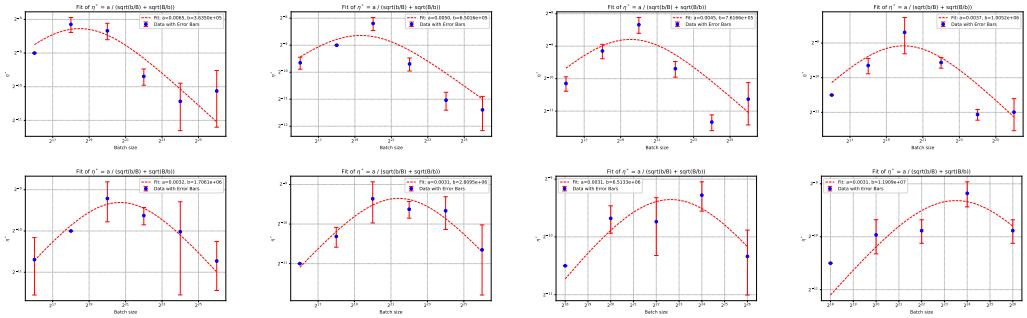


Figure 6: Fits to token budgets $T = 2^{30}, 2^{31}, \ldots, 2^{37}$ (from upper left to bottom right) with Eq. 3.1 to the data points in Fig. A.9.

A.6 Random seed variation

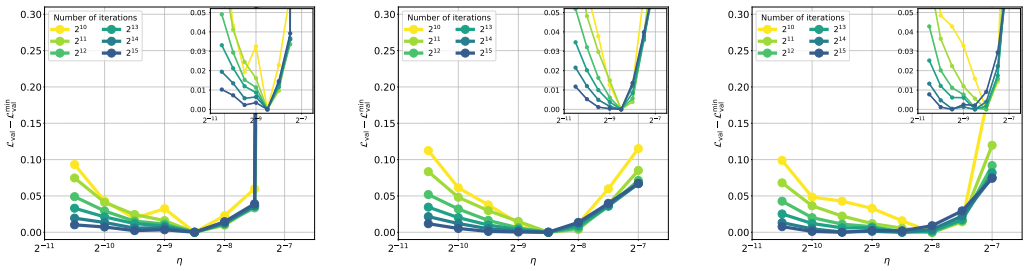


Figure 7: Loss profile $\mathcal{L}_{\text{val}} - \mathcal{L}_{\text{val}}^{\min}$ as a function of maximum learning rate η for three different random seeds for the model configuration ($d_{\text{model}} = d_{\text{model}}^{\text{base}} = 1024$).

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):

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

where T is the total length of the training horizon, T_{warmup} (T_{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 η^* 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).

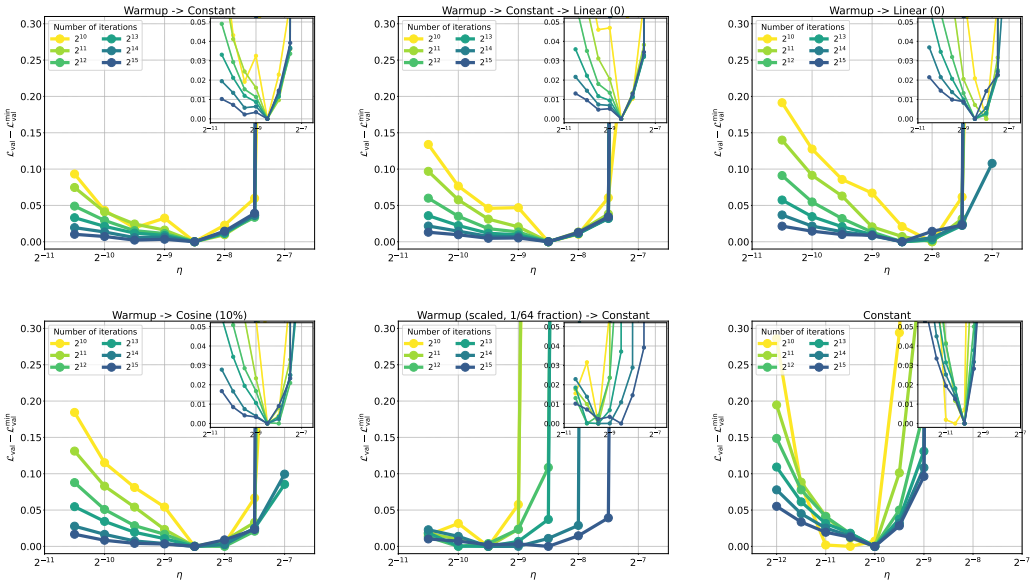


Figure 8: Loss profile $\mathcal{L}_{\text{val}} - \mathcal{L}_{\text{val}}^{\min}$ as a function of maximum learning rate η for schedules with the following phases: warmup and constant (top left); warmup, constant and linear decay to 0 (top middle); warmup and linear decay to 0 (top right); warmup and cosine decay to 10% of the maximum η (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 $(d_{\text{model}}^{\text{base}}, d_{\text{model}})$ CONFIGURATION

A.8.1 $d_{\text{model}}^{\text{base}} = 1024,\ d_{\text{model}} = 1024$

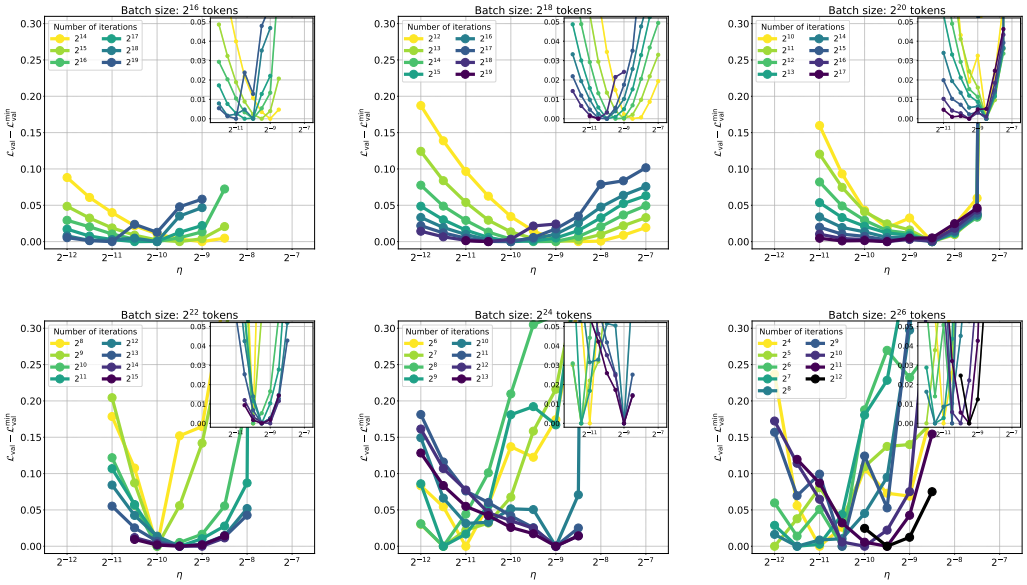


Figure 9: Loss profile $\mathcal{L}_{\text{val}} - \mathcal{L}_{\text{val}}^{\min}$ as a function of maximum learning rate η for $(d_{\text{model}}^{\text{base}} = 1024,\ d_{\text{model}} = 1024)$ for batch size $B = 2^{16}$ (top left), $B = 2^{18}$ (top middle), $B = 2^{20}$ (top right), $B = 2^{22}$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

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

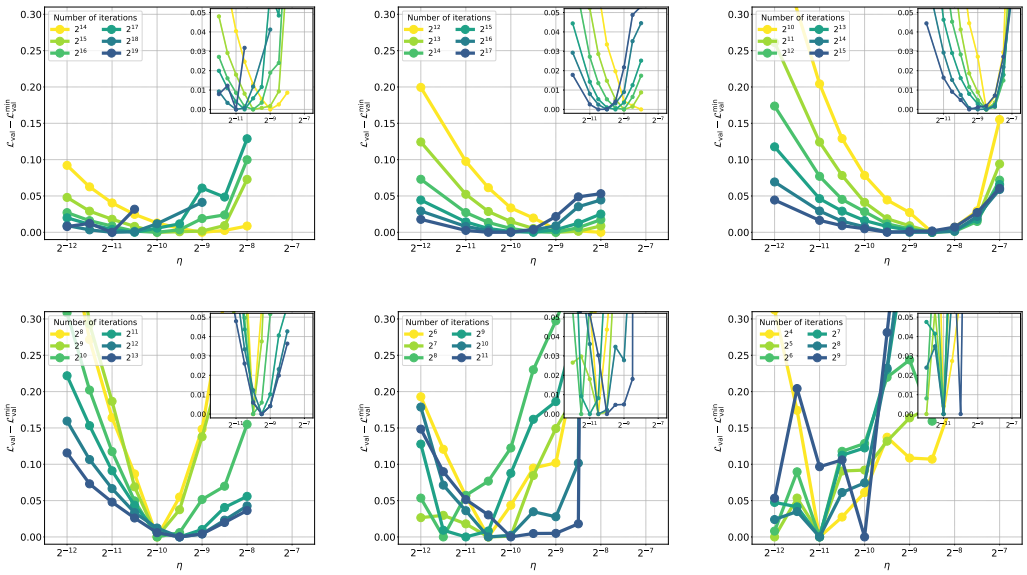


Figure 10: Loss profile $\mathcal{L}_{\text{val}} - \mathcal{L}_{\text{val}}^{\min}$ as a function of maximum learning rate η for $(d_{\text{model}}^{\text{base}} = 1024,\ d_{\text{model}} = 512)$ for batch size $B = 2^{16}$ (top left), $B = 2^{18}$ (top middle), $B = 2^{20}$ (top right), $B = 2^{22}$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

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

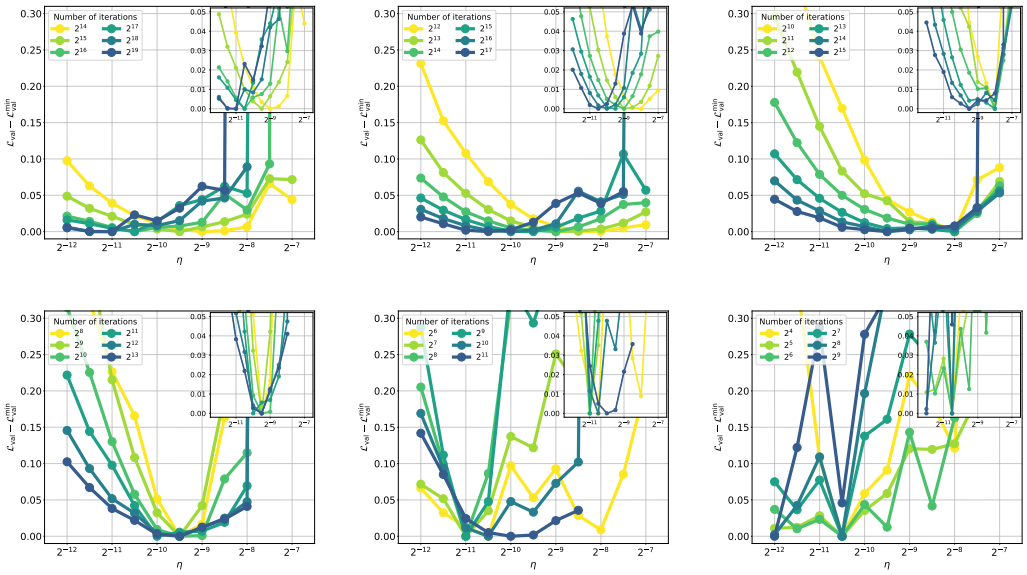


Figure 11: Loss profile $\mathcal{L}_{\text{val}} - \mathcal{L}_{\text{val}}^{\min}$ as a function of maximum learning rate η for $(d_{\text{model}}^{\text{base}} = 1024,\ d_{\text{model}} = 256)$ for batch size $B = 2^{16}$ (top left), $B = 2^{18}$ (top middle), $B = 2^{20}$ (top right), $B = 2^{22}$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

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

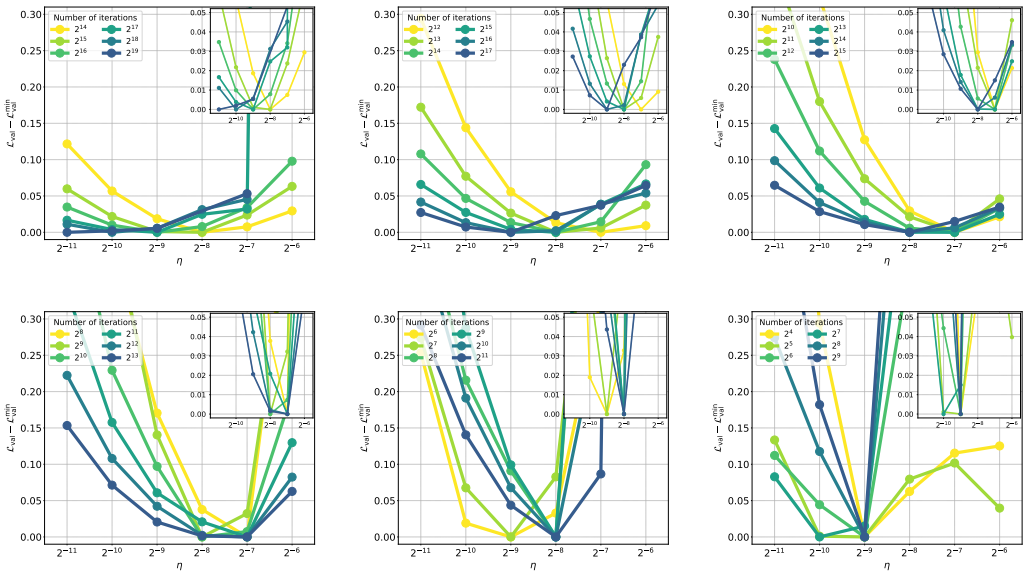


Figure 12: Loss profile $\mathcal{L}_{\text{val}} - \mathcal{L}_{\text{val}}^{\min}$ as a function of maximum learning rate η for ($d_{\text{model}}^{\text{base}} = 256,\ d_{\text{model}} = 256$) for batch size $B = 2^{16}$ (top left), $B = 2^{18}$ (top middle), $B = 2^{20}$ (top right), $B = 2^{22}$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

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

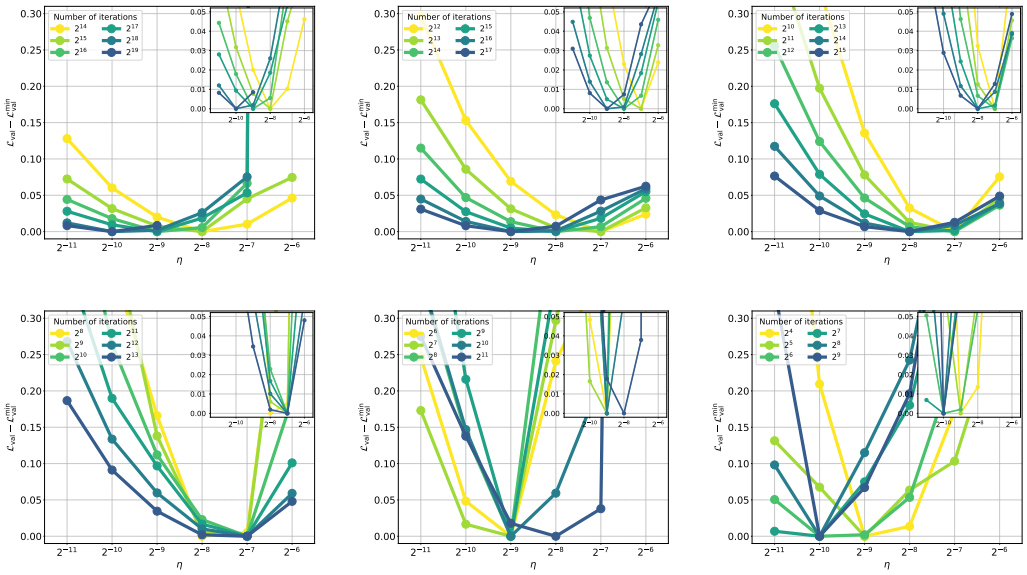


Figure 13: Loss profile $\mathcal{L}_{\text{val}} - \mathcal{L}_{\text{val}}^{\min}$ as a function of maximum learning rate η for ($d_{\text{model}}^{\text{base}} = 256,\ d_{\text{model}} = 512$) for batch size $B = 2^{16}$ (top left), $B = 2^{18}$ (top middle), $B = 2^{20}$ (top right), $B = 2^{22}$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

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

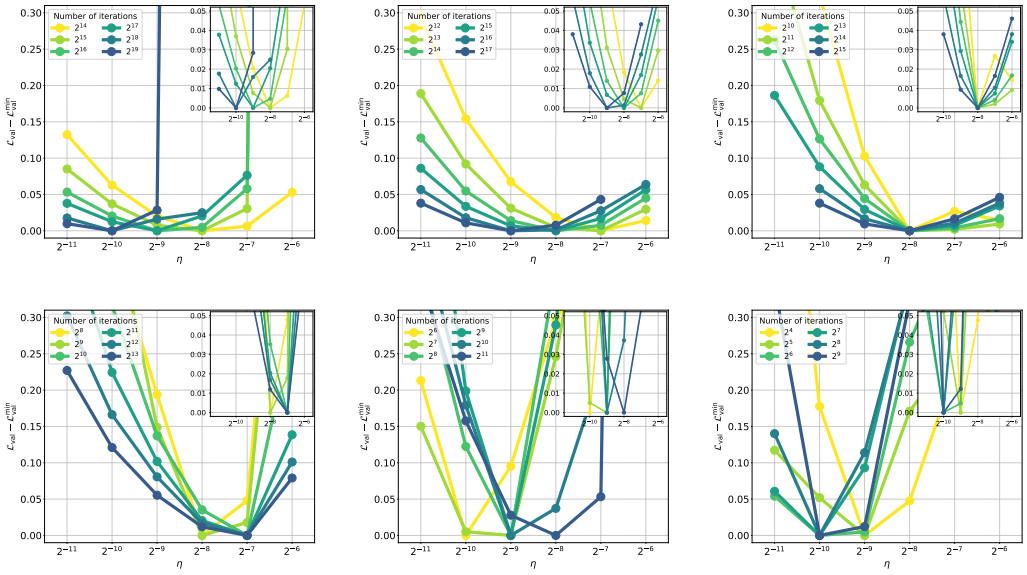


Figure 14: Loss profile $\mathcal{L}_{\text{val}} - \mathcal{L}_{\text{val}}^{\min}$ as a function of maximum learning rate η for ($d_{\text{model}}^{\text{base}} = 256,\ d_{\text{model}} = 1024$) for batch size $B = 2^{16}$ (top left), $B = 2^{18}$ (top middle), $B = 2^{20}$ (top right), $B = 2^{22}$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

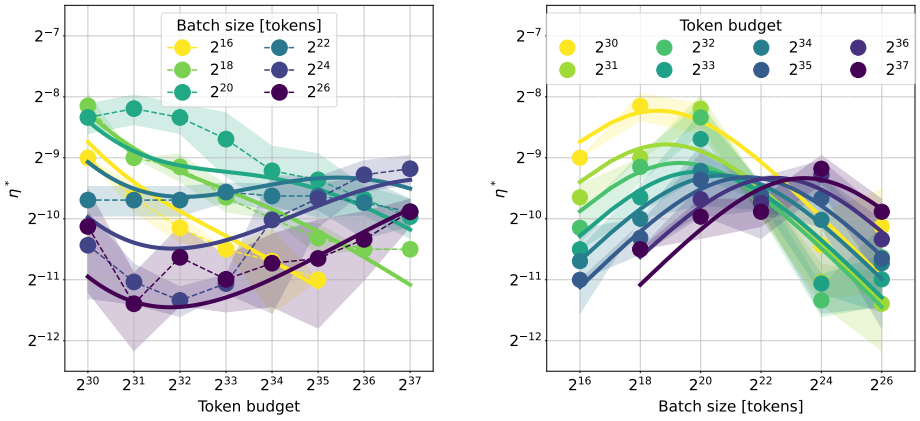


Figure 15: Same as Fig. 1 with the full set of batch size (left) and token budget (right) values.

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

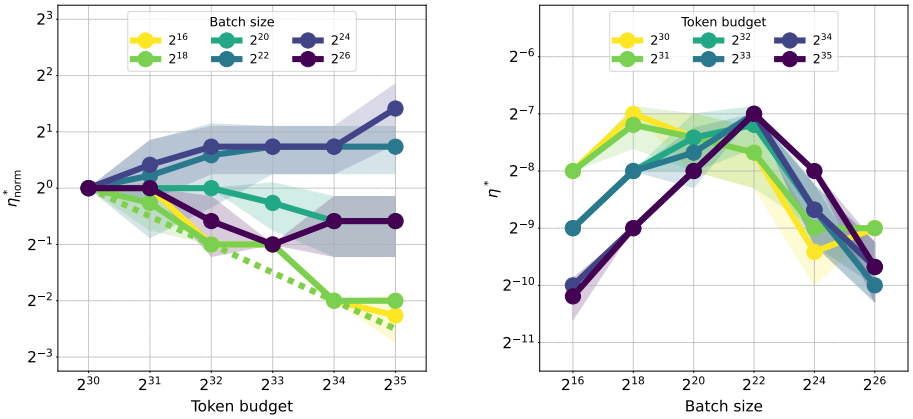


Figure 16: (left) Evolution of the optimal learning rate with an increase of the pretraining token budget $\eta^*_{\text{norm}}(T)$, normalized to $\eta^*|_{T=2^{30}}$, for a set of batch sizes (in tokens). Each point is obtained by averaging optimal learning rate values across μP model family, as described 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 η^* per batch size, against a range of pretraining token budgets. Each point is μP-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 $d_{\text{model}}^{\text{base}} = 256$

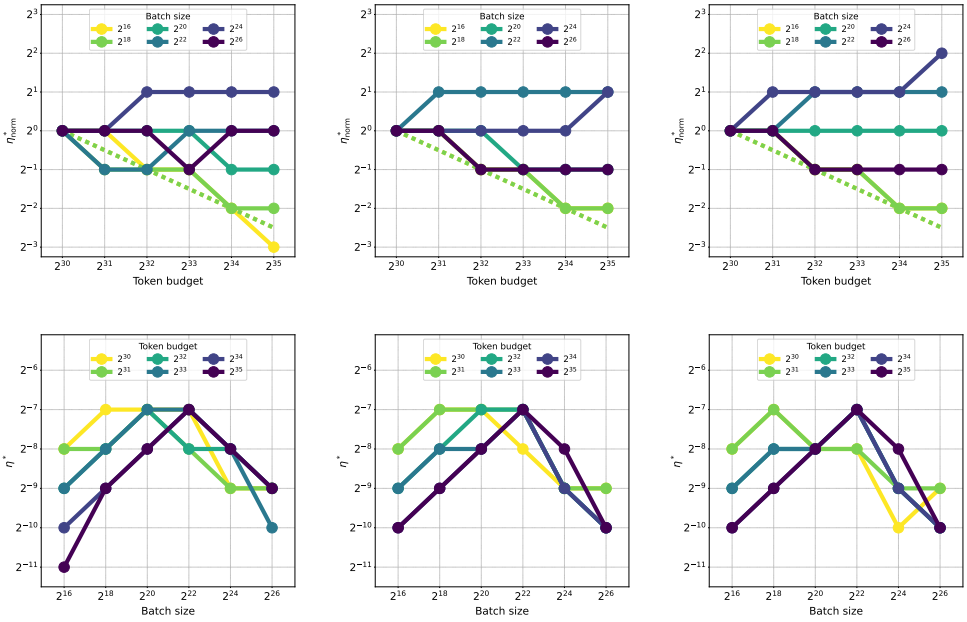


Figure 17: Individual curves contributing to Fig. A.10 for models with $d_{\text{model}} = 256$ (left column), 512 (middle column), 1024 (right column) showing evolution of the normalized to $T = 2^{30}$ tokens optimal learning rate η^*_{norm} in time per batch size (top row), and joint optimal (η, B)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

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

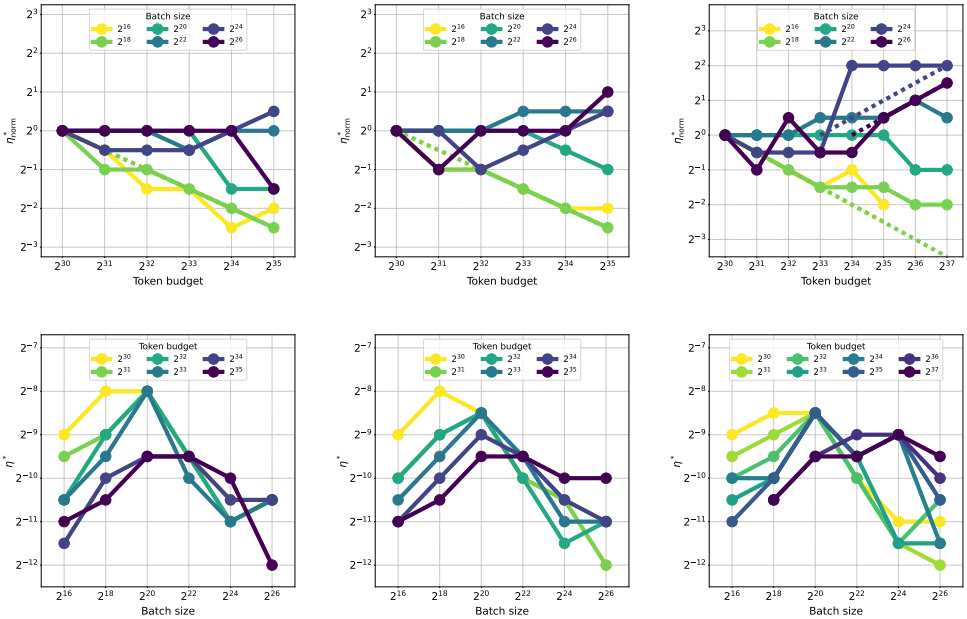


Figure 18: Individual curves contributing to Fig. 1 for models with $d_{\text{model}} = 256$ (left column), 512 (middle column), 1024 (right column) showing evolution of the normalized to $T = 2^{30}$ tokens optimal learning rate η^*_{norm} in time per batch size (top row), and joint optimal (η, B) curves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

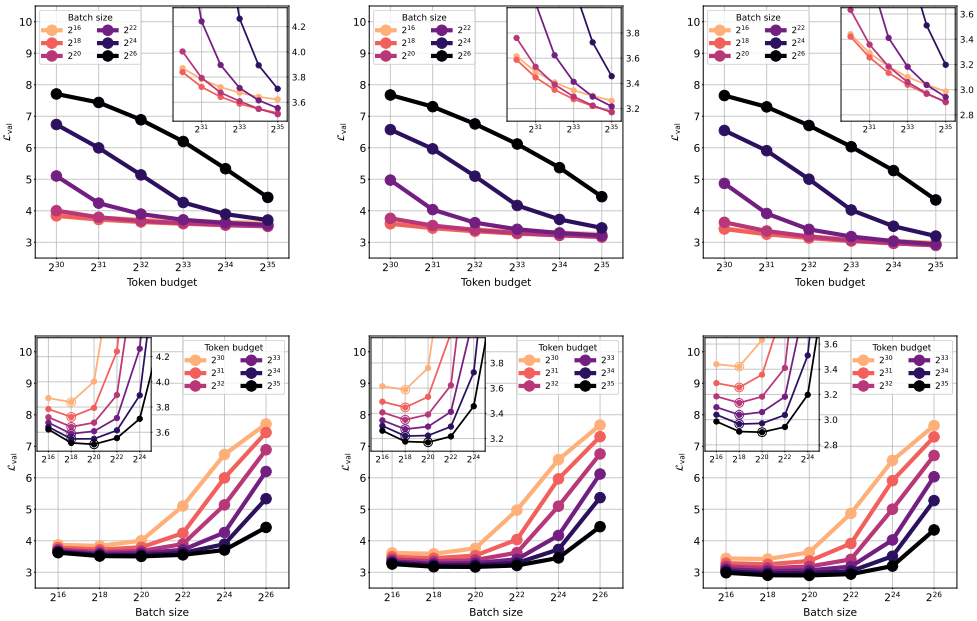


Figure 19: Analogue of Fig. 3b (top row) and Fig. 3a (bottom row) for models with widths $d_{\text{model}} = 256$ (left column), 512 (middle column), 1024 (right column) and the base model width $d_{\text{model}}^{\text{base}} = 256$.

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

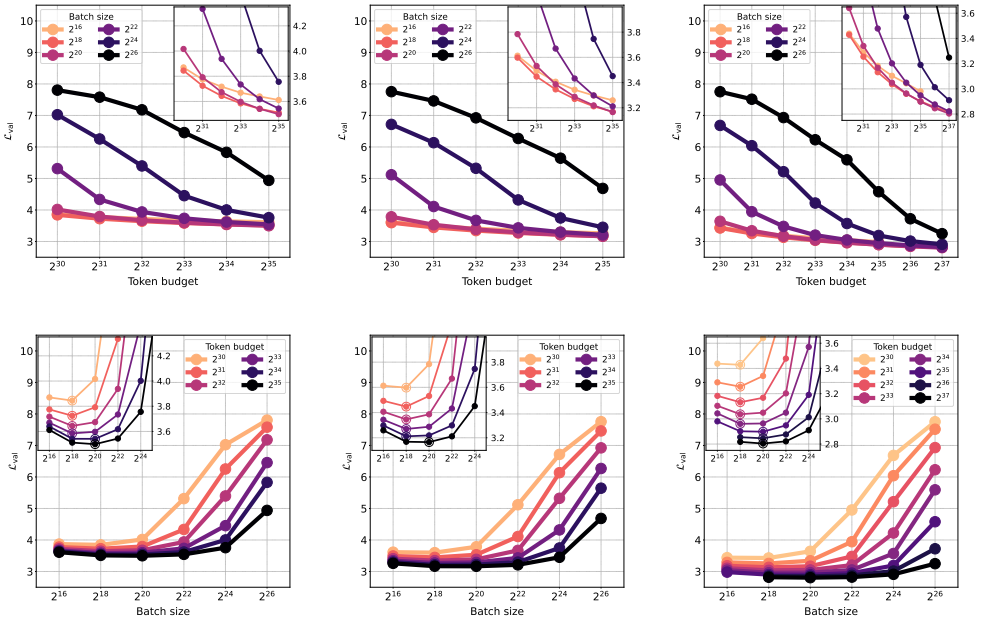


Figure 20: Analogue of Fig. 3b (top row) and Fig. 3a (bottom row) for models with widths $d_{\text{model}} = 256$ (left column), 512 (middle column), 1024 (right column) and the base model width $d_{\text{model}}^{\text{base}} = 1024$.

A.13 Learning rate sensitivity in the μP width limit

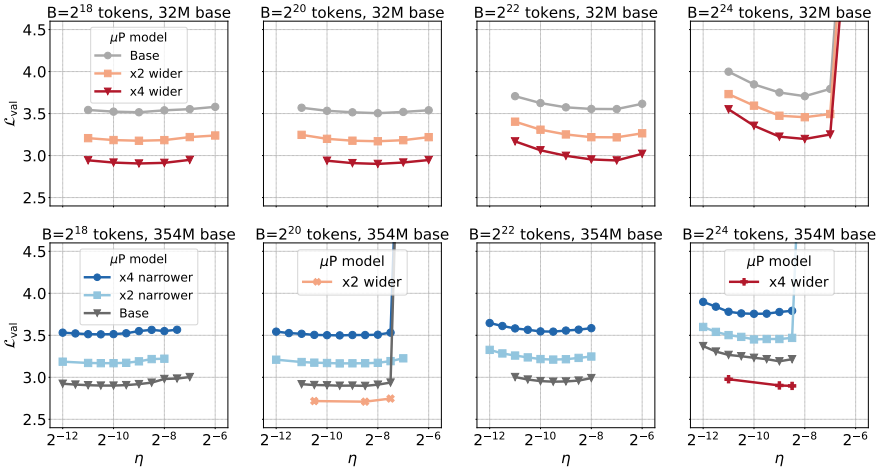


Figure 21: Same as Fig. 5 but without y-axis normalization with $\mathcal{L}_{\text{val}}^{\min}$.