BALANCING EARLY PERFORMANCE SACRIFICES WITH LONG-TERM GAINS: SCALING LEARNING-RATE WARMUP DURATION ACROSS TRAINING HORIZONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning-rate warmup is a standard technique in language-model training, yet its duration remains largely heuristic. Common approaches use either a fixed number of updates or a fixed fraction of the training horizon, two choices that imply very different scaling as training gets longer. When should warmup stay fixed, and when should it grow with the horizon? We address this question with a quadratic model whose modes respond differently to the peak learning rate. Warmup slows progress in directions that already contract well at the peak rate, but can remove persistent error in directions near the stability edge, with higher peak rates shifting the balance toward longer warmup durations. This yields a compact horizon scaling law that captures regimes ranging from essentially no warmup, through fixed-duration warmup, to durations that grow with the training horizon, and explains how the preferred regime changes with peak learning rate. Because the law captures the tradeoff between giving up early progress and improving the trajectory that follows, it can be fit using shorter runs and used to predict warmup at substantially longer horizons. Together, our results explain several familiar properties of warmup through a single tradeoff and suggest treating warmup duration as a horizon-dependent hyperparameter rather than a fixed training heuristic.

1 INTRODUCTION

Learning-rate warmup is ubiquitous in language-model training, yet there is little agreement on how long it should last. Some training setups rely on a fixed number of warmup steps (Vaswani et al., 2017; Groeneveld et al., 2024; Kaplan et al., 2020), while others use a fixed fraction of the training horizon (Black et al., 2022). These choices suggest different views of warmup: a fixed duration treats it as an early stabilization phase, whereas a horizon-dependent duration suggests a longer-term tradeoff whose optimal balance changes with the training budget. Why should an intervention applied only at the start of training depend on when training stops?

Warmup is standard in large-batch and language-model training, including frontier-scale systems (Grattafiori et al., 2024; DeepSeek-AI, 2024; Kimi Team, 2025). Its duration must balance two competing effects: too little warmup can destabilize training at larger peak learning rates, while too much sacrifices early optimization progress. This tradeoff also couples warmup duration to the peak learning rate itself: a rate that appears suboptimal under short warmup can become competitive once longer warmup is allowed, so fixing warmup in advance can change which learning rate appears best.

Prior work explains warmup through adaptive-optimizer statistics and update magnitudes (Liu et al., 2020; Ma & Yarats, 2021), loss curvature and access to larger subsequent learning rates (Gilmer et al., 2022; Kalra & Barkeshli, 2024), and early changes in parameters and representations (Gotmare et al., 2019; Kosson et al., 2024). Recent theory also derives warmup-like schedules from suboptimality-dependent smoothness (Liu et al., 2025b; Alimisis et al., 2026; Riabinin et al., 2026). However, explaining why warmup helps does not determine its optimal duration. That requires balancing the persistent benefit of a longer warmup against the progress lost to smaller initial steps.

We begin with a broad, systematic empirical study of warmup duration, varying peak learning rate and warmup across training horizons to understand how the loss-optimal duration behaves in practice. These experiments show that warmup can remain nearly fixed at lower peak rates, grow sub-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

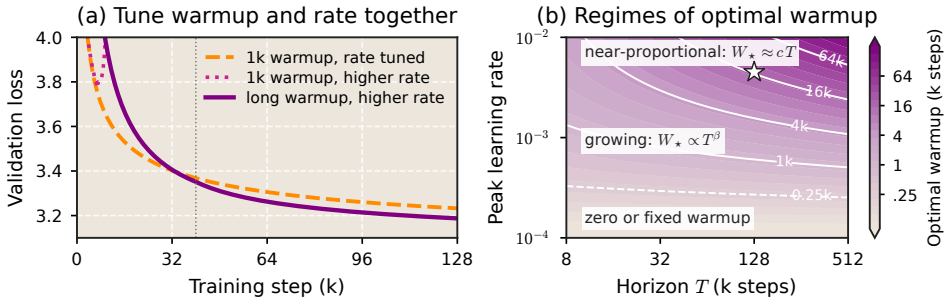


Figure 1: **Overview. (a)** Longer warmup can trade slower early progress for better performance at aggressive peak learning rates. **(b)** Schematic of the loss-optimal warmup regimes, from little or bounded warmup to durations that grow with the training horizon, depending on peak learning rate. The star marks the approximate location of the best measured runs

stantially with the training horizon at larger rates, and change which learning rates are competitive. To understand these patterns, we identify a simple quadratic model that reproduces them and explains why higher peak rates favor longer warmup. The resulting tradeoff between slower early progress and reduced persistent error motivates a compact loss law over warmup duration and training horizon, which we fit to short training trajectories and use to predict warmup choices at longer horizons, including from as few as three short runs.

Our results place warmup duration within the scaling-law view of training. Like other optimization hyperparameters (Li et al., 2025; Zhang et al., 2025; Bergsma et al., 2025), its preferred value changes systematically with the training budget rather than remaining fixed. Prior work has studied how loss (Kaplan et al., 2020; Hoffmann et al., 2022) and learning-rate optima (Bjorck et al., 2025; Filatov et al., 2024) vary with model size, compute, and training horizon. We show that warmup duration has its own horizon dependence, shaped by the peak learning rate. Modeling this dependence turns warmup from a startup heuristic into a hyperparameter with its own scaling law.

2 Warmup regimes across learning rates and training horizons

2.1 Experimental design

Our starting point is an empirical study of warmup duration in Llama-style language models (Grattafiori et al., 2024). We vary peak learning rate and warmup duration under linear warmup followed by a constant rate, and evaluate validation loss across training horizons. Let W denote warmup duration and T the evaluation horizon; for fixed model size and peak rate, varying W defines a *family* of runs. The ranges and training protocol are given in Section A.1.

We use a warmup–stable schedule because one trajectory can be evaluated at many horizons, whereas schedules whose decay depends on the final horizon generally require separate runs. Warmup–stable is also the reusable prefix of warmup–stable–decay (WSD) (Hägele et al., 2024; Wen et al., 2025), a popular schedule in recent large-scale systems (DeepSeek-AI, 2024; Kimi Team, 2025). Existing WSD theoretical analyses focus mainly on the stable and decay phases (Wen et al., 2025; Schaipp et al., 2025). Schedule-aware scaling laws have primarily focused on the post-warmup decay phase rather than on how the initial warmup duration itself should scale with training horizon (Tissue et al., 2025; Luo et al., 2025). By holding the post-warmup learning rate constant, we isolate this question. In Section C.1, we show that appending decay to warmup–stable trajectories largely preserves the ordering of warmup durations.

2.2 From short warmup to horizon-dependent warmup

At conservative peak rates (Figure 2a), little or no warmup can suffice and the preferred duration changes little with training horizon. At larger rates, the lowest-loss region shifts toward longer warmups as training continues, so the duration that is best early need not remain best later; depending on the family, this growth ranges from weak horizon dependence to nearly proportional.

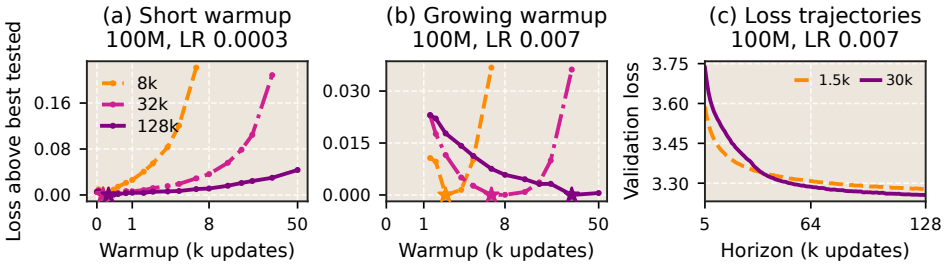


Figure 2: **Peak learning rate changes how optimal warmup evolves with horizon. (a,b)** Lower rates favor short warmup, while higher rates increasingly favor longer durations as training continues. **(c)** At high peak rate, a longer warmup can start worse but become better later, so its benefit persists beyond the warmup duration.

This horizon dependence is not solely due to avoiding unstable runs. In Figure 2c, two successful warmup schedules reverse their ordering after both have ended: the longer warmup is initially worse but becomes better later. Thus, even among stable runs, the loss-optimal warmup can depend on the evaluation horizon. Across the grid, we observe both nearly fixed and horizon-dependent warmup, with longer optimal warmups becoming more common at larger peak learning rates (Figure 3).

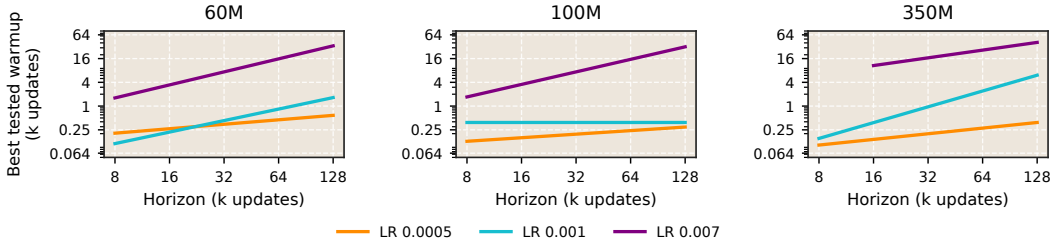


Figure 3: **Fixed and horizon-dependent warmup regimes coexist.** Lines summarize how the measured loss-optimal warmup changes with training horizon within each family.

Warmup needed to avoid training failure. Warmup can make otherwise failing configurations trainable. We measure the *shortest successful warmup*: the smallest tested duration with at least one completed run that converges. This threshold increases with higher peak learning rates, smaller batches, and larger models (Figure 4). The warmup required for a successful start therefore depends strongly on the training regime. However, avoiding failure is only one role of warmup. The shortest successful warmup is often far shorter than the loss-optimal one, so stability and final-loss optimization impose different requirements.

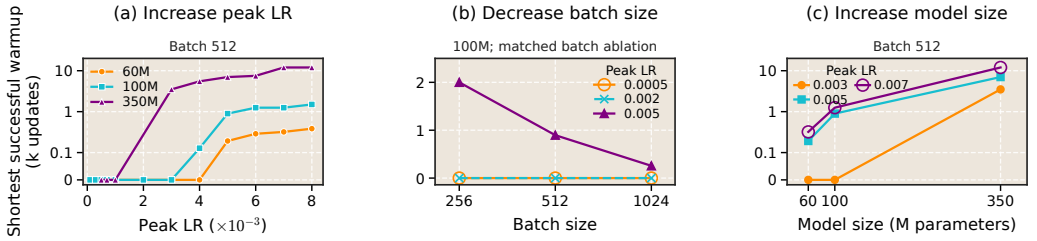


Figure 4: **Shortest successful warmup durations.** Shortest tested successful warmup versus **(a)** peak learning rate, **(b)** batch size, and **(c)** model size.

2.3 Long warmups unlock competitive high learning rates

The strongest horizon dependence appears at larger peak learning rates, but these are not merely pathological settings. In the best-performing region, high peak rates are often paired with long warmups (Table 1), so restricting warmup can make otherwise competitive rates appear suboptimal.

Table 1: **The best tested runs use long warmups.** Learning rate and warmup are jointly tuned.

Model	Peak LR	Warmup (k)	Fraction of horizon (T=128k)	Loss
60M	0.004	30	23.4%	3.404
100M	0.005	50	39.1%	3.246
350M	0.004	40	31.3%	2.983

We also compare unrestricted tuning with a short-warmup constraint, $W \leq 1\text{k}$. Allowing longer warmups improves the loss attainable after tuning warmup at each peak rate and can shift the preferred learning rate (Figure 5).

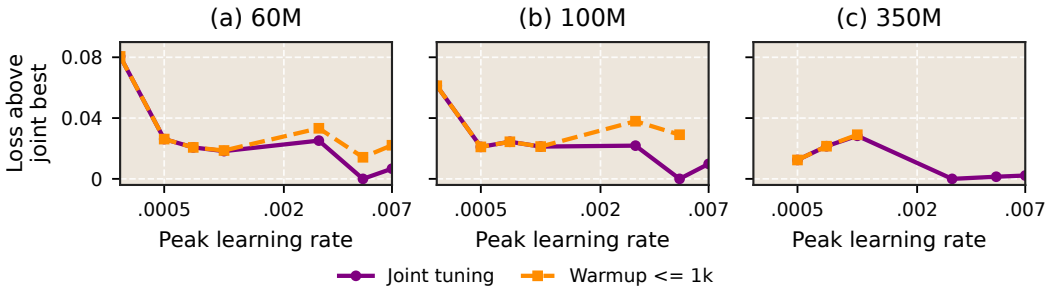


Figure 5: **Long warmups improve the loss attainable at high peak rates.** Validation loss at 128k with warmup tuned freely or capped at 1k. Loss is relative to the best joint configuration at each scale; capped curves end when no successful short-warmup run remains. Additional learning rates used for the stability comparison are omitted because they used a sparser warmup grid.

Prior work finds that the best peak learning rate often decreases as training gets longer, but those studies use different warmup rules (Bjorck et al., 2025; Filatov et al., 2024). In fact, changing warmup can itself change which learning rate works best (Filatov et al., 2024), consistent with Porian et al. (2024), who identify warmup duration as one source of confounding in scaling results. Our results show the same effect: the warmup rule can change which learning rate looks optimal at a given training horizon.

3 A Mechanism for Horizon-Dependent Warmup

To understand the empirical patterns above and ultimately predict how the loss-optimal warmup should change with the training horizon, we seek a simple model that captures the key tradeoffs while remaining analytically tractable. We therefore turn to quadratic dynamics, a standard lens for studying learning-rate, batch-size, and stability effects in optimization (Jain et al., 2018b; Zhang et al., 2019; Bordelon & Pehlevan, 2022; Meterez et al., 2026b). A quadratic is particularly useful here because different curvature directions can respond very differently to the same learning-rate schedule, allowing us to isolate how warmup can trade slower early progress for lower error later in training. Consider gradient descent on the quadratic

$$f(\theta) - f_\star = \tfrac{1}{2}(\theta - \theta_\star)^\top H(\theta - \theta_\star), \qquad H \succeq 0.$$

In a direction of curvature $h > 0$, one update with learning rate η_t multiplies that direction's loss by $|1 - \eta_t h|^2$. We call this multiplicative decrease *contraction*: a smaller multiplier means that error in the direction is removed more quickly. Writing $\alpha = \eta h$, define the contraction rate at the peak learning rate and its average over a linear warmup as

$$\rho(\alpha) = -2\log|1-\alpha|, \qquad g(\alpha) = \int_0^1 \rho(\alpha u)\,\mathrm{d}u. \tag{1}$$

After n peak-rate updates, the loss in this direction is multiplied by $e^{-n\rho(\alpha)}$, while $g(\alpha)$ averages the contraction rates encountered as the learning rate increases from 0 to η. Approximating the linear warmup as continuous, the loss remaining in this direction after warmup and the subsequent peak-rate phase is

$$M_{T,W}(\alpha) \approx e^{-g(\alpha)W - \rho(\alpha)(T-W)}. \tag{2}$$

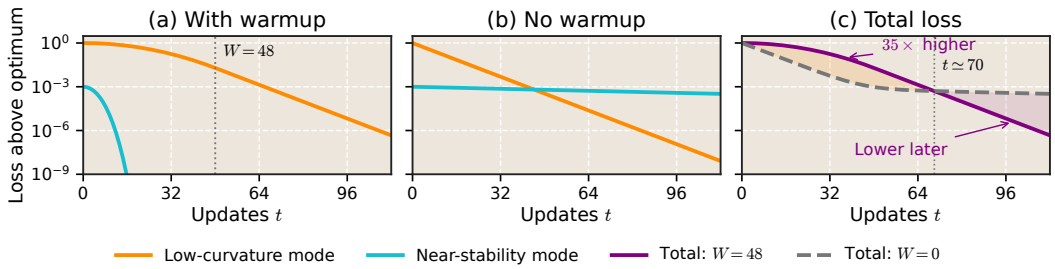


Figure 6: **Warmup trades early progress for lower persistent error. (a,b)** A longer duration slows progress in one direction but reduces error in another that improves only slowly at the peak learning rate. **(c)** As a result, the longer warmup is worse early in training but better later, matching the ranking reversal observed empirically.

For small α, $g(\alpha) \approx \alpha$ and $\rho(\alpha) \approx 2\alpha$: warmup slows progress. Near the upper stable boundary, the comparison reverses. As $\alpha \uparrow 2$, $\rho(\alpha) \to 0$ while $g(\alpha) \to 2$. A longer warmup can remove error that would persist at a stable peak rate.

Higher peak rates favor longer warmup. Consider two modes distinguished by how their contraction compares between warmup and the peak-rate phase. For a *low-curvature mode*, $\rho_s > g_s$, so replacing peak-rate updates with warmup updates slows its decay; define $d_s := \rho_s - g_s > 0$. For a *near-stability mode*, $g_e > \rho_e$, so warmup contracts it more strongly than the subsequent peak-rate phase; define $a_e := g_e - \rho_e > 0$. Their loss is

$$L_{\text{quad}}(W, T) = A_s e^{-\rho_s T + d_s W} + A_e e^{-\rho_e T - a_e W}, \qquad A_s, A_e > 0. \tag{3}$$

Balancing these terms gives

$$W^*_{\text{quad}}(T, \eta) = \Pi_{[0,T]}\left[\frac{(\rho_s - \rho_e)T + \log(A_e a_e/(A_s d_s))}{d_s + a_e}\right]. \tag{4}$$

If $0 < h_s < h_e/2$ and $\eta h_e < 2$, the long-horizon optimum changes at $\eta_c = 2/(h_s + h_e)$. It is zero for $\eta < \eta_c$, bounded at $\eta = \eta_c$, and a fraction of T for $\eta > \eta_c$, a fraction that grows with η. This matches the empirical transition in Figures 2 and 3. Section B.1 gives the proof.

Warmup above the stability limit. The same model also captures the stability trend seen empirically: as peak learning rate increases, completing training without divergence requires progressively longer warmup (Figure 4). Once the near-stability mode crosses the stability boundary, $\alpha_e = \eta h_e > 2$, the constant-rate phase amplifies its error. Writing $u_e = -\rho_e > 0$, its contribution scales as $A_e \exp\{u_e T - (g_e + u_e)W\}$. Preventing net growth requires

$$\frac{W}{T} \geq \kappa(\alpha_e) := \frac{-\rho(\alpha_e)}{g(\alpha_e) - \rho(\alpha_e)} = \frac{\alpha_e \log(\alpha_e - 1)}{\alpha_e + \log(\alpha_e - 1)}. \tag{5}$$

In the continuous-warmup approximation, the minimum warmup fraction required to avoid net growth increases with the peak rate and eventually reaches one. The model therefore reproduces both empirical trends: higher peak rates favor longer loss-optimal warmup before instability, and require longer warmup to avoid net growth by a given horizon after crossing the stability edge.

What the quadratic isolates. The fixed-quadratic analysis deliberately separates the effect of warmup duration from the effect of schedule ordering. Because the update matrices commute, its horizon dependence arises purely from how different curvature directions respond to the rates encountered during warmup and at the peak rate and it does not require an ordering effect. In neural-network training, ordering can matter as well. Section B.5 shows that finite-batch sampling alone makes the order of learning rates matter and can produce the same early–late tradeoff, while evolving curvature provides another mechanism (Gilmer et al., 2022; Kalra & Barkeshli, 2024; Alimisis et al., 2026).

4 A LOSS LAW FOR WARMUP DURATION

Section 3 showed the basic tradeoff in two quadratic modes: longer warmup sacrifices early progress but can reduce error that persists at the peak learning rate. This simple picture is consistent with behavior observed in neural networks, where Hessians can contain isolated large-eigenvalue directions (Sagun et al., 2018; Ghorbani et al., 2019) and training can operate near the gradient-descent stability boundary (Cohen et al., 2021; Damian et al., 2023). We now use the same tradeoff to motivate a compact loss law over warmup duration and training horizon.

The cost of slower initial progress. With linear warmup of duration W, the accumulated learning rate by horizon T is approximately

$$\eta W/2 + \eta(T - W) = \eta\tau, \qquad \tau := T - \frac{W}{2}.$$

Thus τ measures peak-rate-equivalent updates. For small ηh, modal losses decay approximately as $e^{-2\eta h\tau}$. We adopt the common spectral power-law assumption that initial loss below curvature h scales as h^p near zero ($p > 0$) (Bordelon et al., 2024; Velikanov & Yarotsky, 2024). Summing these responses yields $A_\eta \tau^{-p}$ (Section B.3). Longer warmup reduces τ, so this term increases.

From persistent error to a horizon–warmup law. Near the stability edge, warmup can remove error that the peak-rate phase contracts only slowly. For a reference direction, $\rho_\eta = \rho(\eta h_e)$, $g_\eta = g(\eta h_e)$, and its accumulated contraction is

$$C_\eta(W, T) = \rho_\eta \tau + b_\eta W, \qquad b_\eta = g_\eta - \frac{\rho_\eta}{2} > 0. \tag{6}$$

We model the aggregate persistent error by an analogous mixture of slowly relaxing components governed by C_η, with a power-law density of effective decay rates near zero. This gives $E_\eta \propto C_\eta^{-\gamma}$ for large positive C_η. Expanding the logarithm of this power-law surrogate to first order in $\log \tau$ and $\log W$ around a representative training point yields the local product approximation

$$E_\eta(W, T) \approx K_\eta \tau^{-q_\eta} W^{-s_\eta}, \qquad q_\eta + s_\eta = \gamma.$$

In this local approximation, approaching the stability edge shifts more of the persistent term's dependence toward warmup; the derivation is in Section B.3. Combining this term with the cost of delayed progress, and using $w_0 > 0$ to keep the predictor well-defined at zero warmup, gives

$$\boxed{L(W, T) = L_\infty + A\tau^{-p} + K\tau^{-q}(W + w_0)^{-s}} \tag{7}$$

4.1 WHY OPTIMAL WARMUP DEPENDS ON THE HORIZON

With the coefficients fixed, longer training changes the balance between warmup's initial cost and its lasting benefit. When $w_0 \ll W_* \ll T$, the cost of another warmup step scales as $T^{-(p+1)}$, while its benefit scales as $T^{-q} W_*^{-(s+1)}$. Balancing these terms gives

$$W_* \sim \left(\frac{2sK}{Ap}\right)^{1/(s+1)} T^\beta, \qquad \beta = \frac{p + 1 - q}{s + 1}. \tag{8}$$

For $0 < \beta < 1$, the benefit fades more slowly than the cost at a fixed warmup duration, so W_* grows as a sublinear power of T. The full law also covers bounded or zero warmup ($\beta \le 0$) and warmup proportional to T ($\beta \ge 1$); see Section B.4 for the limits.

4.2 FITTING THE MEASURED LOSS CURVES

We fit $(L_\infty, A, K, p, q, s)$ separately for each family with a bounded multistart Huber fit similar to Hoffmann et al. (2022). The same functional form captures short- and long-warmup families (Figure 7), explains most loss variation, and selects low-regret warmups (Table 2). The fitted bounds keep $\beta > 0$ for every family, so nearly fixed warmup at low peak rates appears through a near-zero warmup scale instead (Figure 8b,c). All tables report RMSE and regret in 10^{-3} loss units, and $\pm$ is the sample standard deviation across horizons after any averaging over families.

Table 2: **Absolute-fit accuracy and selection quality.** R^2 and RMSE are medians across families; regret is the mean loss above the best tested warmup. RMSE and regret are in 10^{-3} loss units.

Evaluation	Median R^2	RMSE	Mean regret
All horizons (descriptive)	0.996	4.92	2.00 ± 1.51
Alternating holdout	0.996	4.98	2.05 ± 1.61

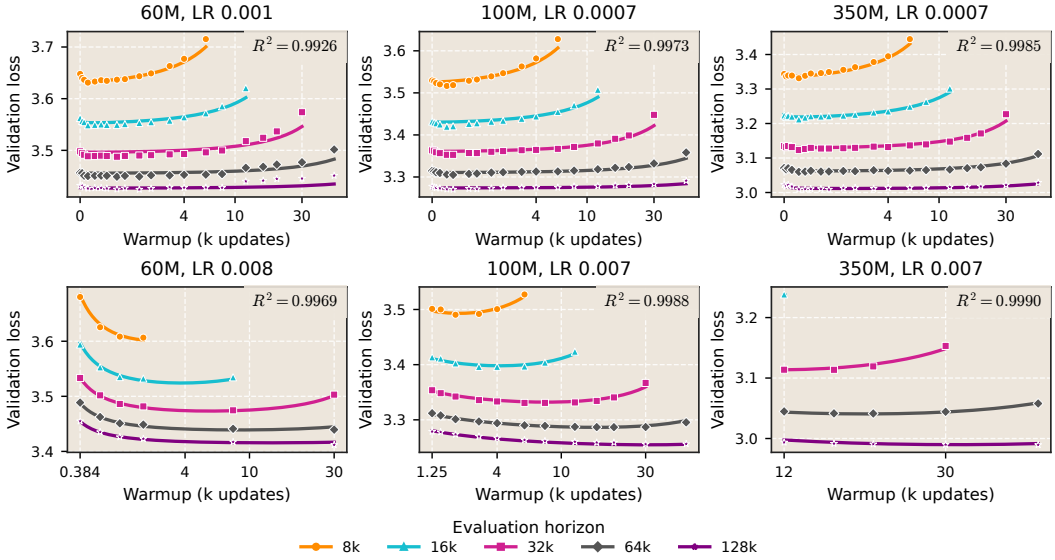


Figure 7: **The loss law captures qualitatively different warmup regimes.** Selected illustrative full-data fits span low and high peak learning rates across model scales.

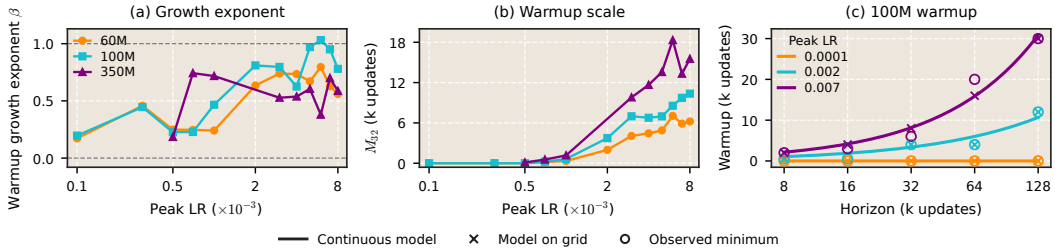


Figure 8: **Growth and scale summarize the fitted warmup regimes.** **(a)** Warmup growth exponent $\beta = (p + 1 - q)/(s + 1)$ from the fitted loss law. **(b)** Corresponding warmup scale $M_{32} = (2sK/(Ap))^{1/(s+1)} 32^{\beta}$ for the interior approximation $W_* \approx M_{32}(T/32)^{\beta}$. **(c)** Representative 100M families comparing the continuous, model-selected and measured optimum.

5 Predicting warmup at later horizons

The horizon dependence in Equation (7) suggests an extrapolation rule. We fit the loss law on shorter trajectories, freeze its coefficients, and use the predicted loss ordering to choose warmup at later horizons. With the full warmup grid, this tests the law. With only three short runs, it becomes a practical protocol.

5.1 From the loss law to a selection rule

Warmup selection depends only on differences between candidate losses. Subtracting a reference warmup therefore preserves the minimizing duration:

$$\Delta L(W, T; W_{\text{ref}}) = L(W, T) - L(W_{\text{ref}}, T). \tag{9}$$

With $\tau_W = T - W/2$ and $\tau_{\text{ref}} = T - W_{\text{ref}}/2$, subtracting Equation (7) gives

$$\boxed{\Delta L(W, T; W_{\text{ref}}) = A[\tau_W^{-p} - \tau_{\text{ref}}^{-p}] + C\left[\tau_W^{-q}\left(\frac{W + w_0}{W_{\text{ref}} + w_0}\right)^{-s} - \tau_{\text{ref}}^{-q}\right].} \tag{10}$$

We fit (A, C, p, q, s) to measured loss differences at the fit horizons and choose the available duration with the lowest predicted loss. The shortest successful warmup serves as the reference. This removes L_∞ while preserving the competing effects of delayed progress and persistent error.

5.2 Horizon extrapolation

We fit the absolute and difference forms independently using checkpoints through 32k updates, freeze their parameters, and select warmup at $\mathcal{H} = \{50\text{k}, 75\text{k}, 100\text{k}, 128\text{k}\}$. We also evaluate interleaved held-out checkpoints within the fit window. At every target horizon, all methods choose from the same retained successful warmup candidates. For a selected duration $\widehat{W}_T$ and candidate set $\mathcal{G}_T$, we measure

$$\mathcal{R}_T = L(\widehat{W}_T, T) - \min_{W \in \mathcal{G}_T} L(W, T) \geq 0, \tag{11}$$

and average regret equally across target horizons and then across families.

Table 3: **Fits through 32k select low-regret warmups at later horizons.** Mean regret in 10^{-3} loss units. Both fits use checkpoints through 32k updates for the four-horizon and 128k evaluations; the alternating column uses interleaved held-out checkpoints.

Selector	Parameters	Alternating	Four horizons	128k
Absolute fit	6	2.05 ± 1.61	2.26 ± 0.66	3.05
Difference fit	5	$\mathbf{1.45} \pm 2.22$	$\mathbf{1.58} \pm 0.77$	**2.61**
1k target	0	15.47 ± 10.80	13.79 ± 0.22	13.56
10% target	0	13.39 ± 4.10	12.58 ± 0.90	11.72

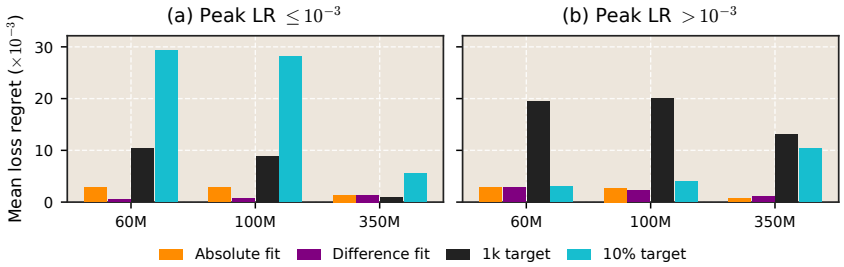


Figure 9: **Warmup selection extrapolates beyond the fit window.** Bars average regret over 50k, 75k, 100k, and 128k after fitting through 32k, then equally over all fitted peak learning rates within each scale and indicated range.

The fixed 1k and 10% targets give a scale for regret. On this scale both fits select low-regret warmups within the fit window and at later horizons, with the difference fit best in every column of Table 3. The full grid thus measures how faithfully the fitted law carries the horizon dependence of the loss forward, and Section 5.3 turns this into a practical protocol.

5.3 Practical extrapolation from three short runs

We next replace the full warmup grid with only three separated, stable warmup trajectories per family. Using their shared post-warmup checkpoints through 20k or 32k, we fit the loss law, freeze its coefficients, and select warmup at the same four target horizons. We retain both fitting methods and use the absolute fit as the primary practical predictor. Both loss-law fits outperform the baselines common in practice and naive extrapolation of the best fitting run (Table 4), so horizon extrapolation does not need a dense grid once a few stable fitting runs cover the relevant warmup range. Figure 10 breaks down the absolute-fit and baseline results at 32k by scale and peak rate. The absolute fit has lower mean regret at both fit horizons. With three fitting runs, the difference fit uses one as a

reference, leaving only two curves of loss differences to constrain its five parameters. The absolute fit instead retains all three loss curves, and its stronger performance in the three-run setting motivates using this method for practical and long-horizon prediction.

Table 4: **Three short runs support longer-horizon warmup prediction.** Mean regret over 50k, 75k, 100k, and 128k, in 10^{-3} loss units. Fixed dur. and fixed frac. keep the best fitting-run duration or its fraction of the fit horizon.

Fit horizon	Absolute	Difference	Fixed dur.	Fixed frac.	1k	10%
20k (29 families)	$\mathbf{2.83} \pm 0.16$	4.22 ± 0.29	9.31 ± 1.21	11.72 ± 0.82	14.28 ± 0.30	13.42 ± 0.96
32k (33 families)	$\mathbf{2.51} \pm 0.27$	2.62 ± 0.36	7.61 ± 1.12	8.47 ± 0.62	13.79 ± 0.22	12.58 ± 0.90

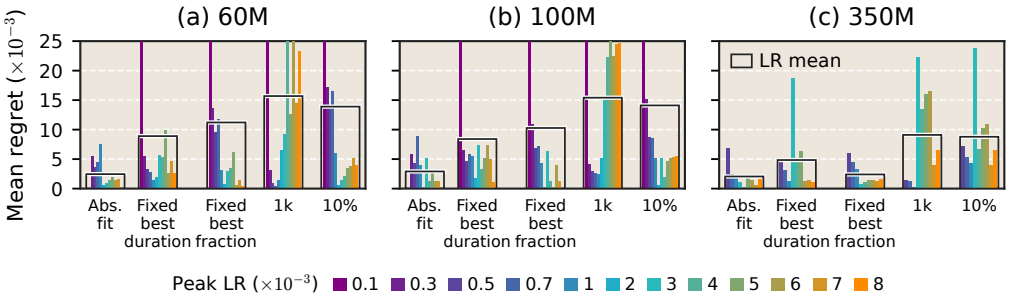


Figure 10: **Three-run selection by scale and peak learning rate.** Absolute-fit selection and the two baselines after fitting through 32k; regret averaged over 50k, 75k, 100k, and 128k.

5.4 Long-horizon extrapolation

We extend the three-run procedure to horizons of up to one million updates using the 100M model at a smaller batch size. Following the three-run extrapolation, we use the absolute fit. For each peak learning rate, we fit the loss law to all three fitting trajectories using prefixes through 50k or 75k updates, freeze the fit, and predict warmup separately at 250k, 500k, and one million updates. We compare the forecasts with a fixed 1k-update warmup, which trains without divergence at every peak rate tested, and a 10% warmup ratio. At the two higher peak rates the forecasts give the lowest loss at every horizon, and at the lowest rate they match the 1k baseline (Table 5).

Table 5: **Long-horizon validation-loss comparison.** Three-run absolute-fit forecasts from 50k or 75k prefixes, compared with 1k-warmup and 10% baselines. Horizons and predicted warmups are in thousands of updates.

	LR $= 0.5 \times 10^{-3}$			LR $= 2 \times 10^{-3}$			LR $= 5 \times 10^{-3}$		
T (k)	250	500	1000	250	500	1000	250	500	1000
1k-warmup loss	**3.370**	3.337	**3.311**	3.433	3.421	3.402	3.371	3.347	3.322
10% loss	3.375	3.347	3.339	3.418	3.367	3.340	3.363	3.335	3.314
$\widehat{W}_{50}$	0.6	0.7	0.8	47.7	77.5	125.9	129.5	241.7	451.1
Loss	3.372	**3.336**	**3.311**	3.412	3.362	3.338	3.359	3.325	**3.303**
$\widehat{W}_{75}$	0.8	1.1	1.2	56.6	95.2	160.0	116.0	213.5	392.9
Loss	3.371	3.336	3.312	**3.411**	**3.359**	**3.336**	**3.358**	**3.324**	**3.303**

6 Conclusion

Warmup duration is neither fixed across training budgets nor tied to a fixed fraction of the run. In our experiments, lower peak learning rates often favor short or bounded warmup, while higher rates favor durations that grow with the training horizon. A simple quadratic model explains this shift as a balance between giving up early progress and reducing error that would otherwise persist later in training. This motivates a compact loss law that captures the observed warmup regimes and can be fit using only three short trajectories to guide warmup choices at longer horizons. Our results therefore suggest treating warmup duration as a horizon-dependent hyperparameter, tuned jointly with peak learning rate and training budget.

AI USE STATEMENT

We used generative AI tools to assist with coding, writing and editing, and checking the manuscript for mistakes and inconsistencies. The authors reviewed and verified all final content.

REPRODUCIBILITY STATEMENT

We provide all experimental details and hyperparameters needed to reproduce our results. Section A.1 gives the data, model configurations, training recipe, and the grid of peak learning rates, warmup durations, and horizons; Section A.2 gives the fitting procedure, parameter bounds, and baselines; and Section A.3 defines candidate selection and all evaluation metrics. The supplementary code includes a small subset of our training curves and Python code to fit the horizon scaling law in Equation (7) to these curves and to evaluate its loss predictions and warmup selections.

REFERENCES

Foivos Alimisis, Rustem Islamov, and Aurelien Lucchi. Why do we need warm-up? A theoretical perspective. In *International Conference on Machine Learning (ICML)*, 2026. URL https://openreview.net/forum?id=a6fo32UnpU. arXiv:2510.03164.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024. doi: 10.1073/pnas.2311878121. URL https://doi.org/10.1073/pnas.2311878121.

Shane Bergsma, Nolan Dey, Gurpreet Gosal, Gavia Gray, Daria Soboleva, and Joel Hestness. Power lines: Scaling laws for weight decay and batch size in LLM pre-training. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025. doi: 10.52202/085713-4171. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/b5f78a17a94da3e34c935515d1b6adae-Abstract-Conference.html.

Johan Bjorck, Alon Benhaim, Vishrav Chaudhary, Furu Wei, and Xia Song. Scaling optimal LR across token horizons. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/cffa22c56c0df3b3edb1df8a9ad67804-Abstract-Conference.html. arXiv:2409.19913.

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pp. 95–136, 2022. doi: 10.18653/v1/2022.bigscience-1.9. URL https://aclanthology.org/2022.bigscience-1.9/.

Blake Bordelon and Cengiz Pehlevan. Learning curves for stochastic gradient descent on structured features. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2106.02713.

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1024–1034, 2020. URL https://proceedings.mlr.press/v119/bordelon20a.html. arXiv:2002.02561.

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. In *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4345–4382, 2024. URL https://proceedings.mlr.press/v235/bordelon24a.html. arXiv:2402.01092.

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature Communications*, 12:2914, 2021. doi: 10.1038/s41467-021-23103-1. URL https://www.nature.com/articles/s41467-021-23103-1.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Jeremy M. Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2103.00065. arXiv:2103.00065.

Alex Damian, Eshaan Nichani, and Jason D. Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2209.15594. arXiv:2209.15594.

DeepSeek-AI. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. URL https://arxiv.org/abs/2412.19437.

Oleg Filatov, Jan Ebert, Jiangtao Wang, and Stefan Kesselheim. Time transfer: On optimal learning rate and batch size in the infinite data limit. *arXiv preprint arXiv:2410.05838*, 2024. URL https://arxiv.org/abs/2410.05838.

Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via Hessian eigenvalue density. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2232–2241, 2019. URL https://proceedings.mlr.press/v97/ghorbani19b.html.

Justin Gilmer, Behrooz Ghorbani, Ankush Garg, Sneha Kudugunta, Behnam Neyshabur, David Cardoze, George Dahl, Zachary Nado, and Orhan Firat. A loss curvature perspective on training instability in deep learning. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=OcKMT-36vUs. arXiv:2110.04369.

Akhilesh Gotmare, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=r14EOsCqKX.

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. *arXiv preprint arXiv:1706.02677*, 2017. URL https://arxiv.org/abs/1706.02677.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15789–15809, 2024. doi: 10.18653/v1/2024.acl-long.841. URL https://aclanthology.org/2024.acl-long.841/.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2405.18392. arXiv:2405.18392.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan

Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c1e2faff6f588870935f114ebe04a3e5-Abstract.html.

Prateek Jain, Sham M. Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Accelerating stochastic gradient descent for least squares regression. In *Conference on Learning Theory (COLT)*, volume 75 of *Proceedings of Machine Learning Research*, pp. 545–604, 2018a.

Prateek Jain, Sham M. Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Parallelizing stochastic gradient descent for least squares regression: Mini-batching, averaging, and model misspecification. *Journal of Machine Learning Research*, 18(223):1–42, 2018b. URL https://jmlr.org/papers/v18/16-595.html.

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

Dayal Singh Kalra and Maissam Barkeshli. Why warmup the learning rate? Underlying mechanisms and improvements. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2406.09405. arXiv:2406.09405.

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

Kimi Team. Kimi K2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025. URL https://arxiv.org/abs/2507.20534.

Atli Kosson, Bettina Messmer, and Martin Jaggi. Analyzing & reducing the need for learning rate warmup in GPT training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2410.23922. arXiv:2410.23922.

Houyi Li, Wenzhen Zheng, Qiufeng Wang, Hanshan Zhang, Zili Wang, Shijie Xuyang, Yuantao Fan, Zhenyu Ding, Haoying Wang, Ning Ding, Shuigeng Zhou, Xiangyu Zhang, and Daxin Jiang. Predictable scale: Part I, Step Law – optimal hyperparameter scaling law in large language model pretraining. *arXiv preprint arXiv:2503.04715*, 2025. URL https://arxiv.org/abs/2503.04715.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for LLM training. *arXiv preprint arXiv:2502.16982*, 2025a. URL https://arxiv.org/abs/2502.16982.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://arxiv.org/abs/1908.03265. arXiv:1908.03265.

Yuxing Liu, Yuze Ge, Rui Pan, Kang An, and Tong Zhang. Theoretical analysis on how learning rate warmup accelerates convergence. *arXiv preprint arXiv:2509.07972*, 2025b. URL https://arxiv.org/abs/2509.07972.

Kairong Luo, Haodong Wen, Shengding Hu, Zhenbo Sun, Zhiyuan Liu, Maosong Sun, Kaifeng Lyu, and Wenguang Chen. A multi-power law for loss curve prediction across learning rate schedules. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2503.12811. arXiv:2503.12811.

Jerry Ma and Denis Yarats. On the adequacy of untuned warmup for adaptive optimization. In *AAAI Conference on Artificial Intelligence*, volume 35, pp. 8828–8836, 2021. doi: 10.1609/aaai.v35i10.17069. URL https://ojs.aaai.org/index.php/AAAI/article/view/17069. arXiv:1910.04209.

Alexandru Meterez, Depen Morwani, Jingfeng Wu, Costin-Andrei Oncescu, Cengiz Pehlevan, and Sham Kakade. Seesaw: Accelerating training by balancing learning rate and batch size scheduling. In *International Conference on Learning Representations (ICLR)*, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/187f9ba4cd5a59477305ac712282a1cd-Paper-Conference.pdf. arXiv:2510.14717.

Alexandru Meterez, Pranav Ajit Nair, Depen Morwani, Cengiz Pehlevan, Sham Kakade, and Alex Damian. A defense of the quadratic model. *arXiv preprint arXiv:2607.21716*, 2026b. URL https://arxiv.org/abs/2607.21716.

Frank W. J. Olver. *Asymptotics and Special Functions*. Academic Press, New York, 1974.

Tomer Porian, Mitchell Wortsman, Jenia Jitsev, Ludwig Schmidt, and Yair Carmon. Resolving discrepancies in compute-optimal scaling of language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2406.19146. arXiv:2406.19146.

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. URL https://www.jmlr.org/papers/v21/20-074.html.

Artem Riabinin, Andrey Veprikov, Arman Bolatov, Martin Takáč, and Aleksandr Beznosikov. Where does warm-up come from? adaptive scheduling for norm-constrained optimizers. *arXiv preprint arXiv:2602.05813*, 2026. URL https://arxiv.org/abs/2602.05813.

Levent Sagun, Utku Evci, V. Ugur Güney, Yann Dauphin, and Léon Bottou. Empirical analysis of the Hessian of over-parametrized neural networks. In *International Conference on Learning Representations (ICLR), Workshop Track*, 2018. URL https://iclr.cc/virtual/2018/workshop/563. arXiv:1706.04454.

Fabian Schaipp, Alexander Hägele, Adrien Taylor, Umut Simsekli, and Francis Bach. The surprising agreement between convex optimization theory and learning-rate scheduling for large model training. In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 53267–53294, 2025. URL https://proceedings.mlr.press/v267/schaipp25a.html. arXiv:2501.18965.

Howe Tissue, Venus Wang, and Lu Wang. Scaling law with learning rate annealing. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025. doi: 10.52202/085713-3044. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/830b1abc6d2da85f23d41169fa44d185-Abstract-Conference.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. URL https://arxiv.org/abs/1706.03762.

Maksim Velikanov and Dmitry Yarotsky. Tight convergence rate bounds for optimization under power law spectral conditions. *Journal of Machine Learning Research*, 25(81):1–78, 2024. URL https://jmlr.org/papers/v25/23-0698.html.

Kaiyue Wen, Zhiyuan Li, Jason Wang, David Hall, Percy Liang, and Tengyu Ma. Understanding warmup-stable-decay learning rates: A river valley loss landscape view. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/6a1fe80a9e2dcda0b3e5fd0fd87eb097-Abstract-Conference.html. arXiv:2410.05192.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/8df7c2e3c3c3be098ef7b382bd2c37ba-Abstract.html.

Guodong Zhang, Lala Li, Zachary Nado, James Martens, Sushant Sachdeva, George Dahl, Christopher Shallue, and Roger B. Grosse. Which algorithmic choices matter at which batch sizes? Insights from a noisy quadratic model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. URL https://arxiv.org/abs/1907.04164. arXiv:1907.04164.

Hanlin Zhang, Depen Morwani, Nikhil Vyas, Jingfeng Wu, Difan Zou, Udaya Ghai, Dean Foster, and Sham Kakade. How does critical batch size scale in pre-training? In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a6f14f95d9c9443927638bcd5d917a7a-Abstract-Conference.html.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. GaLore: Memory-efficient LLM training by gradient low-rank projection. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 61121–61143. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhao24s.html.

Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham M. Kakade. Benign overfitting of constant-stepsize SGD for linear regression. *Journal of Machine Learning Research*, 24 (326):1–58, 2023. URL https://jmlr.org/papers/v24/21-1297.html.

APPENDIX CONTENTS

A Experimental and fitting protocols **16**

A.1 Training setup, observations, and units 16

A.2 Fitting and baselines 17

A.3 Candidate selection and metrics 17

B Theory supporting the main text **17**

B.1 Quadratic warmup tradeoff 18

B.2 Warmup above the stability limit 20

B.3 From many modes to the fitted loss law 21

B.4 Optimal duration under the fitted law 23

B.5 A stochastic mechanism for horizon dependence 24

C Robustness of warmup behavior **26**

C.1 Warmup ordering after decay 26

C.2 Horizon and peak-rate dependence with Muon 27

C.3 Horizon dependence under different warmup shapes 28

D Additional selection results **30**

D.1 Direct warmup-scaling baseline 30

D.2 Near-optimal measured warmups 30

E Extrapolation from short runs **31**

E.1 Three-run extrapolation 31

E.2 Recommendations beyond the observed warmups 31

E.3 Sensitivity to the choice of fitting runs 32

E.4 Sensitivity to the number of fitting runs 33

E.5 Fit horizon and short-run budget 33

E.6 Long-horizon extrapolation protocol 34

E.7 Model-scale extrapolation 34

F Sensitivity of the loss law **35**

F.1 Warmup exponent 35

F.2 Zero-warmup offset 35

F.3 Sharing the persistent exponent sum across peak rates 35

G Complete fit gallery **37**

H Related work **39**

A EXPERIMENTAL AND FITTING PROTOCOLS

A.1 TRAINING SETUP, OBSERVATIONS, AND UNITS

We analyze Llama-style language models trained with linear warmup followed by a constant peak learning rate. Validation checkpoints after warmup are used for fitting; repeated observations at the same model-size, peak-rate, warmup, and horizon coordinate are averaged. The analysis contains the 33 model-size and peak-rate families listed in Table 8. Time in the fitted laws is measured in thousands of optimizer updates, with $w_0 = 0.032$. A run is *eligible* at horizon T if it completed without diverging and $W < T$; fitting, warmup selection, and all figures use only eligible runs.

Data and training. We follow the standard Llama-on-C4 pretraining setup used in GaLore (Zhao et al., 2024).[1] Models are pretrained on C4 English with the `t5-base` SentencePiece tokenizer (Raffel et al., 2020). Training examples are randomly sampled packed sequences, and validation is next-token cross-entropy on a fixed held-out set. The base recipe and model configurations are summarized below; additional experiments inherit these settings unless stated otherwise.

Table 6: **Base training configuration.**

Setting	Value
Training corpus	C4 English; 1,024 shards; 184,565,529,484 tokens
Tokenizer / storage	`t5-base` SentencePiece / `uint16`
Sequence length	256
Batch size	512, 131,072 tokens per update
Training budget	128,000 updates; 16,777,216,000 tokens
AdamW (β_1, β_2) / ϵ	$(0.9, 0.99)$ / 10^{-6}
Weight decay / global gradient clip	0 / 1.0
Learning-rate schedule	Linear warmup, then constant peak rate
Precision	bfloat16

Table 7: **Model and execution settings.** All models use Llama decoder blocks with SwiGLU, rotary position embeddings, RMSNorm with $\epsilon = 10^{-6}$, initialization standard deviation 0.02, and untied input embeddings and output heads. The configured maximum sequence length is 1,024; training uses 256 tokens.

Setting	60M	100M	350M
Layers	8	12	24
Hidden dimension	512	640	1,024
MLP intermediate dimension	1,376	1,708	2,736
Attention heads / head dimension	8 / 64	10 / 64	16 / 64
Vocabulary size	32,000	32,100	32,000
Total parameters	58,073,600	100,117,120	367,969,280
Non-embedding parameters	25,305,600	59,029,120	302,433,280
Embedding and output parameters	32,768,000	41,088,000	65,536,000
Microbatch / accumulation	512 / 1	128 / 4	64 / 8
Validation tokens	10,092,544	10,027,008	10,010,624

Batch-size ablation. The batch-size comparison in Figure 4b trains the 100M model at peak learning rates 0.5, 2, and 5×10^{-3} with batches of 256 and 1,024 sequences, alongside the base batch of 512. All other settings match Tables 6 and 7, including the 128,000-update budget; the peak learning rate is not rescaled with batch size. Tokens per update are therefore 65,536, 131,072, and 262,144, respectively.

[1] Reference implementation: https://github.com/jiaweizzhao/GaLore.

Table 8: **Experimental grid.** Peak learning rates and warmup ranges for the 33 model-size and peak-rate families; warmup grids vary with model size and peak learning rate.

Scale	Peak LRs (10^{-3})	Families	Warmup range (updates)
60M	0.1, 0.3, 0.5, 0.7, 1, 2, 3, 4, 5, 6, 7, 8	12	0–50,000
100M	0.1, 0.3, 0.5, 0.7, 1, 2, 3, 4, 5, 6, 7, 8	12	0–50,000
350M	0.5, 0.7, 1, 3, 4, 5, 6, 7, 8	9	0–64,000

A.2 Fitting and baselines

The absolute and difference fits estimate $(L_\infty, A, K, p, q, s)$ and (A, C, p, q, s) independently. For the difference fit, the shortest successful warmup is the reference and only shared checkpoints are used. Positive parameters are optimized in log space with a bounded Huber objective and deterministic multistart initialization. Held-out fits use only their fitting data.

Table 9: **Fitting settings.** Time is measured in thousands of updates. Widening the exponent bounds to $[0.05, 2]$, which allows $\beta \leq 0$, gave slightly higher selection regret.

Setting	Choice
Fitting optimizer	Trust-region reflective (SciPy `least_squares`)
Huber transition	0.02 loss units
Warmup offset	$w_0 = 0.032$
Exponent bounds	$p, q, s \in [0.05, 1]$
Absolute amplitudes	$A,\ K/w_0^s \in [10^{-9}, 10^5]$
Difference amplitude	$C \in [10^{-9}, 10^4]$
Extrapolation fit window	checkpoints through 32k updates
Free-s multistart	35 deterministic starts

Direct warmup-scaling baseline. The two-parameter baseline in Section D.1 fits a power law directly to the best observed warmup at the fit horizons. It is included to test whether extrapolating the observed duration trend is already sufficient without modeling the loss surface.

Tuned rules. The fixed-duration rule selects the warmup with the lowest loss at the last fit checkpoint, among all eligible warmups on the full grid or among the three fitting runs, and keeps it for every later horizon. The fixed-fraction rule keeps that warmup's fraction of the fit horizon and scales it with T. Both are mapped to the nearest eligible candidate, like the 1k and 10% targets.

A.3 Candidate selection and metrics

All methods choose from the same successful warmup candidates at each target horizon. Loss-law fits select the candidate with lowest predicted loss; the direct scaling, 1k, and 10% baselines are mapped to the nearest available candidate. Regret is measured relative to the best observed candidate and is averaged equally across target horizons and then across families. RMSE and R^2 are computed from unmodified validation loss. In aggregate tables, $\pm$ denotes the sample standard deviation across horizon-wise family means.

B Theory supporting the main text

We follow Sections 3 and 4: first the quadratic warmup tradeoff and warmup above the stability limit, then the passage from many modes to the fitted loss law, and finally the fitted law's optimal duration. Section B.5 gives a complementary discrete result: sampling breaks order invariance and can change the loss-optimal warmup with horizon, even with fixed population curvature. Unless otherwise specified, the peak rate η is fixed when W and T vary.

B.1 QUADRATIC WARMUP TRADEOFF

On a quadratic loss the effect of warmup can be computed one eigendirection (*mode*) at a time. For a mode with curvature h and relative peak rate $\alpha = \eta h$, Figure 11 compares the log-contraction of one step at the peak rate, $\rho(\alpha)$, with that of one warmup step on average, $g(\alpha)$. Warmup helps the mode when $g > \rho$, which splits the peak rates into four ranges:

- $\alpha < \alpha_{\text{cross}} \simeq 1.28$: warmup slows the mode ($g < \rho$);
- $\alpha_{\text{cross}} < \alpha < 2$: warmup speeds it up ($g > \rho > 0$);
- $2 < \alpha < \alpha_{\max} \simeq 4.59$: the peak rate makes the mode grow ($\rho < 0$), but warmup still shrinks it on average ($g > 0$);
- $\alpha > \alpha_{\max}$: warmup makes the mode grow too ($g < 0$).

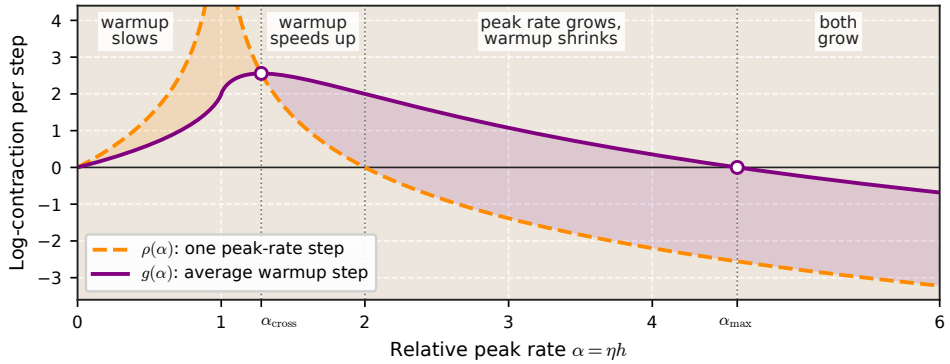


Figure 11: What warmup does to one mode. $\rho(\alpha)$: log-contraction of one step at the peak rate. $g(\alpha)$: its average over a linear warmup from 0 to α (Equation (13)). Dots: α_{cross}, where $g = \rho$, and $\alpha_{\max}$, where $g = 0$.

Exact modal loss. Let $e = \theta - \theta_\star$ and $f(\theta) - f_\star = \frac{1}{2} e^\top H e$ with $H \succeq 0$, let $h_i > 0$ be the curvatures of the modes that carry loss, and let A_i be their initial losses. A gradient step at rate η_t multiplies the loss of mode i by $(1 - \eta_t h_i)^2$. With the linear warmup $\eta_t = \eta \min\{t/W, 1\}$ ($W = 0$ is the constant schedule) and $\alpha_i = \eta h_i$,

$$f(\theta_T) - f_\star = \sum_i A_i \, M_{T,W}(\alpha_i), \qquad M_{T,W}(\alpha) = \prod_{t=1}^{W} \left|1 - \frac{\alpha t}{W}\right|^2 |1 - \alpha|^{2(T-W)}. \tag{12}$$

At the peak rate a mode shrinks exactly when $0 < \alpha < 2$; we call $\alpha = 2$ the stability limit.

Contraction rates. One step at relative rate α multiplies a mode's loss by $e^{-\rho(\alpha)}$. The warmup factor in Equation (12) runs over W equally spaced rates in $(0, \alpha]$; replacing the sum of their logarithms by W times its average gives

$$\rho(\alpha) = -2 \log |1 - \alpha|, \qquad g(\alpha) = \int_0^1 \rho(\alpha u) \, \mathrm{d}u = 2\left[1 + \frac{1 - \alpha}{\alpha} \log |1 - \alpha|\right], \tag{13}$$

where the closed form uses the antiderivative $w \log |w| - w$ of $\log |w|$ (g is finite for every $\alpha > 0$, and $g(1) = 2$), and the continuous-warmup approximation

$$\overline{M}_{T,W}(\alpha) = \exp\{-g(\alpha) W - \rho(\alpha)(T - W)\}. \tag{14}$$

The results below use the continuous-warmup approximation; exact integer warmup schedules can exhibit isolated cancellations when a step lands at $\eta_t h = 1$.

When warmup helps a mode. Turning a plateau step into a warmup step changes a mode's log-loss by $\rho - g$. Two identities locate the sign change:

$$g(\alpha) - \rho(\alpha) = \frac{2}{\alpha}\{\alpha + \log(\alpha - 1)\} \quad (\alpha > 1), \qquad g'(\alpha) = \frac{\rho(\alpha) - g(\alpha)}{\alpha}, \tag{15}$$

the first from Equation (13) and the second by differentiating $g(\alpha) = \alpha^{-1}\int_0^\alpha \rho(v)\,\mathrm{d}v$. For $\alpha < 1$, ρ increases, so its average g is smaller. For $\alpha > 1$, $g - \rho$ changes sign once, at the solution $\alpha_{\mathrm{cross}} \simeq 1.27846$ of $\alpha + \log(\alpha - 1) = 0$. By the second identity, g increases where $g < \rho$ and decreases where $g > \rho$.

Two modes. Take a low-curvature mode s with $\eta h_s < 1$ and a mode e near the stability limit, $\alpha_{\mathrm{cross}} < \eta h_e < 2$. With $\rho_j = \rho(\eta h_j)$ and $g_j = g(\eta h_j)$, the approximate loss is

$$L_{\mathrm{quad}}(W, T) = A_s\, e^{-\rho_s T + d_s W} + A_e\, e^{-\rho_e T - a_e W}, \qquad d_s = \rho_s - g_s > 0, \quad a_e = g_e - \rho_e > 0,$$

with W treated as real. Solving $\partial_W L_{\mathrm{quad}} = 0$ gives a line in T,

$$W = x_\eta T + c_\eta, \qquad x_\eta = \frac{\rho_s - \rho_e}{d_s + a_e}, \qquad c_\eta = \frac{1}{d_s + a_e}\log\frac{A_e a_e}{A_s d_s},$$

whose slope changes sign where both modes contract equally fast at the peak rate, $\eta_c = 2/(h_s + h_e)$.

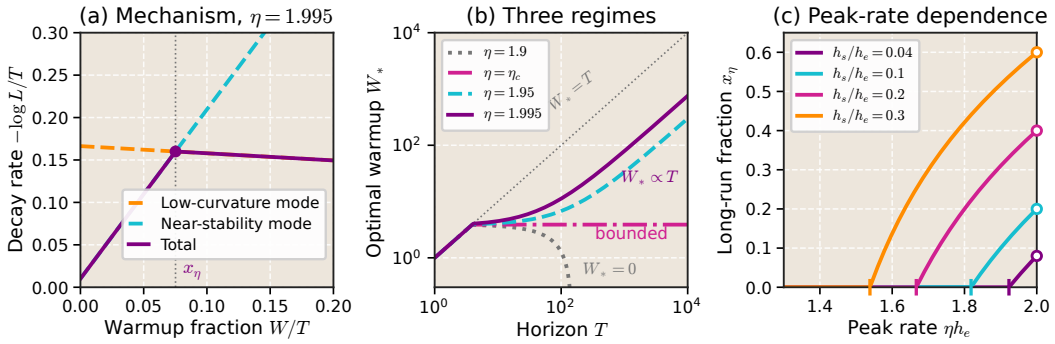


Figure 12: The two-mode optimum (Theorem 1) with the curvatures of Figure 6, $h_e = 1$ and $h_s/h_e = 0.04$, so $\eta_c \simeq 1.923$. **(a)** At $\eta = 1.995$ and $W = xT$, the loss decays at rate $\min(\rho_s - d_s x,\ \rho_e + a_e x)$ (solid), which is largest where the two modal rates (dashed) are equal, at $x = x_\eta$. **(b)** W^*_{quad} against T ($A_e/A_s = 100$); dotted: $W^* = T$. **(c)** x_η against the peak rate. Ticks: η_c. Circles: $2h_s/h_e$.

Theorem 1 (Two-mode optimum and peak-rate transition)**.** *Let $A_s, A_e, d_s, a_e > 0$. For every $T > 0$, the unique minimizer of $L_{\mathrm{quad}}(\cdot, T)$ on $[0, T]$ is*

$$W^*_{\mathrm{quad}} = \Pi_{[0,T]}\left[x_\eta T + c_\eta\right], \tag{16}$$

where $\Pi_{[0,T]}$ clips to $[0, T]$. If moreover $h_s < h_e/2$ and $\alpha_{\mathrm{cross}} < \eta h_e < 2$, then for all large T

$$W^*_{\mathrm{quad}} = \begin{cases} 0, & \eta < \eta_c, \\ \max(c_\eta, 0), & \eta = \eta_c, \\ x_\eta T + c_\eta, & \eta > \eta_c. \end{cases}$$

For $\eta > \eta_c$,

$$\frac{W^*_{\mathrm{quad}}}{T} \longrightarrow x_\eta \in (0, 1), \qquad \frac{\mathrm{d}x_\eta}{\mathrm{d}\eta} > 0, \qquad x_\eta \xrightarrow{\eta \downarrow \eta_c} 0, \qquad x_\eta \xrightarrow{\eta \uparrow 2/h_e} \frac{2h_s}{h_e}. \tag{17}$$

Proof of Theorem 1. $\partial_W^2 L_{\mathrm{quad}} > 0$, so the minimizer is the stationary point clipped to $[0, T]$. On the stated range, $\eta h_s < 2h_s/h_e < 1$ and $\eta h_e > \alpha_{\mathrm{cross}}$, so $d_s, a_e > 0$. Since $\rho(\alpha) = -2\log|1 - \alpha|$,

$$\rho_s > \rho_e \iff 1 - \eta h_s < \eta h_e - 1 \iff \eta > \eta_c,$$

so x_η has the sign of $\eta - \eta_c$. Clipping $x_\eta T + c_\eta$ gives the three cases, using $0 < x_\eta < 1$ for $\eta > \eta_c$.

For the properties of x_η, write $x_\eta = N/(N+G)$ with $N = \rho_s - \rho_e > 0$ and $G = g_e - g_s$. As η grows, ρ_s, g_s increase and ρ_e, g_e decrease (Equation (15)), so $N' > 0$, $G' < 0$, and $G > g(2) - g(2h_s/h_e) > 0$. Hence $0 < x_\eta < 1$ and $x'_\eta = (N'G - NG')/(N+G)^2 > 0$. As $\eta \downarrow \eta_c$, $N \to 0$. As $\eta \uparrow 2/h_e$, $\rho_e \to 0$ and $g_e \to 2$, and Equation (13) gives $x_\eta \to 2h_s/h_e$. □

B.2 Warmup above the stability limit

Above the stability limit, $\alpha = \eta h_e > 2$, a mode grows at every peak-rate step, but warmup still passes through rates at which it shrinks. Write $\rho = \rho(\alpha)$, $g = g(\alpha)$ and $a = g - \rho$. By Equations (13) and (15),

$$\rho = -2\log(\alpha - 1) < 0, \qquad a = \frac{2}{\alpha}\{\alpha + \log(\alpha - 1)\} > 0.$$

In the continuous-warmup approximation of Equation (14), the mode's loss after T steps is $A_e e^{-C}$, with accumulated contraction

$$C = gW + \rho(T - W) = \rho T + aW.$$

Warmup prevents net amplification by horizon T exactly when $C \geq 0$.

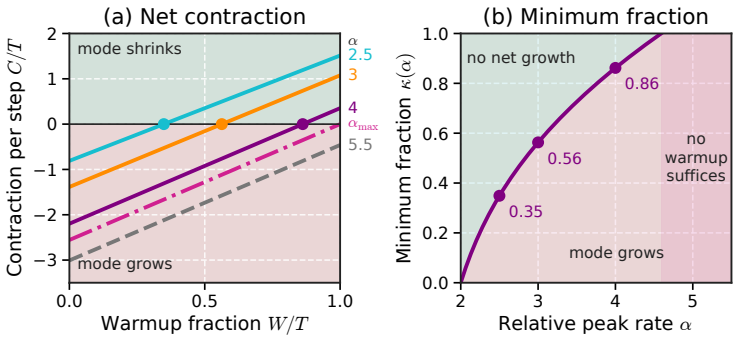


Figure 13: Warmup above the stability limit (Theorem 2). **(a)** $C/T = \rho + a\,W/T$ against the warmup fraction; dots mark $\kappa(\alpha)$. **(b)** The threshold $\kappa(\alpha)$.

Theorem 2 (Warmup required to avoid net growth)**.** *Let $A_e > 0$ and $\alpha > 2$. On $(2, \infty)$, g has a unique zero $\alpha_{\max}$, given by*

$$(\alpha_{\max} - 1)\log(\alpha_{\max} - 1) = \alpha_{\max}, \qquad \alpha_{\max} \simeq 4.59112148. \tag{18}$$

(a) If $2 < \alpha < \alpha_{\max}$, warmup prevents net amplification by horizon T if and only if

$$\frac{W}{T} \geq \kappa(\alpha) = \frac{-\rho}{g - \rho} = \frac{\alpha\log(\alpha - 1)}{\alpha + \log(\alpha - 1)}, \tag{19}$$

and κ increases from $\kappa(2) = 0$ to $\kappa(\alpha_{\max}) = 1$.

(b) If $\alpha \geq \alpha_{\max}$, then $C < 0$ for every $W < T$. At $\alpha = \alpha_{\max}$, $W = T$ gives $C = 0$. For $\alpha > \alpha_{\max}$,

$$\min_{0 \leq W \leq T} A_e e^{-C} = A_e e^{-gT} \longrightarrow \infty \qquad (T \to \infty). \tag{20}$$

Proof of Theorem 2. Since $a > 0$, $C = \rho T + aW$ increases with W, so

$$\max_{0 \leq W \leq T} C = gT.$$

By Equations (13) and (15),

$$g' = \frac{\rho - g}{\alpha} = -\frac{a}{\alpha} < 0, \qquad g = 2 - 2\,\frac{\alpha - 1}{\alpha}\log(\alpha - 1),$$

so g decreases from $g(2) = 2$ to $-\infty$ and has one zero, given by Equation (18).

(a) With $\ell = \log(\alpha - 1)$, so that $-\rho = 2\ell$ and $a = 2 + 2\ell/\alpha$,

$$C \geq 0 \iff \frac{W}{T} \geq \frac{-\rho}{a} = \frac{\alpha\ell}{\alpha + \ell} = \kappa.$$

Moreover $\kappa < 1 \iff g > 0$; $\kappa(2) = 0$ since $\ell = 0$; and $\kappa(\alpha_{\max}) = 1$ since $g = 0$ gives $a = -\rho$. Finally,

$$\kappa'(\alpha) = \frac{\ell^2 + \alpha^2/(\alpha - 1)}{(\alpha + \ell)^2} > 0. \tag{21}$$

(b) If $g = 0$, then $C = \rho(T - W) \leq 0$, with equality only at $W = T$. If $g < 0$, then $C \leq gT < 0$ for every W. □

This is a finite-horizon statement: for any fixed W, an unstable constant-rate phase eventually dominates as $T \to \infty$.

B.3 From many modes to the fitted loss law

A single mode decays exponentially; power laws arise when error is spread over many modes. If components with speed z decay as e^{-zC} and their initial error has density $cz^{\gamma-1}$ near zero, then for large C

$$\int_0^\infty c\, z^{\gamma-1} e^{-zC}\, \mathrm{d}z = c\,\Gamma(\gamma)\, C^{-\gamma}:$$

components faster than $1/C$ are gone, and what remains is the error at speeds below $1/C$. We apply this to low-curvature modes, which give the progress term, and to an effective persistent contribution, and then connect the result to the fitted law by a local product approximation.

Progress term. For a low-curvature mode, $(1 - \eta_t h)^2 \approx e^{-2\eta_t h}$, so its loss is about e^{-2hS_T} with $S_T = \sum_t \eta_t$. For the linear warmup,

$$S_T = \eta\,(T - W/2 + 1/2) \quad (W \geq 1), \qquad S_T = \eta T \quad (W = 0), \tag{22}$$

so $S_T \approx \eta\tau$ with the effective training amount $\tau = T - W/2$: a warmup step counts as half a peak-rate step. We describe the initial loss of these modes by a measure μ_s on curvatures with a power-law density near zero, as is standard in analyses of power-law loss curves (Bordelon et al., 2024; Velikanov & Yarotsky, 2024).

Proposition 1 (Power-law decay from low curvatures)**.** *Fix $\eta > 0$ and $0 < h_{\max} < 1/\eta$. Let μ_s be a finite measure on $(0, h_{\max}]$ with density $\psi(h) \sim c_s h^{p-1}$ near zero, $p, c_s > 0$. Then, for integers $T \to \infty$ and $0 \leq W \leq T$,*

$$P(W, T) = \int M_{T,W}(\eta h)\, \mathrm{d}\mu_s(h) \sim c_s \Gamma(p)\, (2\eta)^{-p}\, \tau^{-p}. \tag{23}$$

Proof. Let $\bar{\alpha} = \eta h_{\max} < 1$. For $0 \leq x \leq \bar{\alpha}$,

$$x \;\leq\; -\log(1 - x) \;\leq\; x + \frac{x^2}{2(1 - \bar{\alpha})}.$$

Applying this to every factor of $M_{T,W}$, with $x = \eta_t h$, and using $\sum_t \eta_t^2 \leq \eta S_T$ places each mode between two exponentials:

$$e^{-2S_T\, h(1+ch)} \;\leq\; M_{T,W}(\eta h) \;\leq\; e^{-2S_T\, h}, \qquad c = \frac{\eta}{2(1 - \bar{\alpha})}.$$

Integrated against μ_s, both bounds are Laplace integrals with large parameter $2S_T \geq \eta T$. By Watson's lemma (Olver, 1974), a density $\psi(h) \sim c_s h^{p-1}$ near zero gives

$$\int e^{-2S_T\, h}\, \mathrm{d}\mu_s(h) \;\sim\; c_s \Gamma(p)\, (2S_T)^{-p},$$

and the same holds for the lower bound after the change of variable $u = h(1 + ch)$, which leaves the density near zero unchanged. Finally, $S_T = \eta(\tau + 1/2) \sim \eta\tau$ by Equation (22), which gives Equation (23). □

The constant is absorbed into the amplitude A_η. The asymptotic power law requires suitable loss-weighted mass at arbitrarily small curvatures; finitely many modes can approximate it over an intermediate range of T.

Persistent term. For a reference direction with $\rho_\eta = \rho(\eta h_e)$, $g_\eta = g(\eta h_e)$, $a_\eta = g_\eta - \rho_\eta > 0$ and $g_\eta > 0$, that is, $\alpha_{\text{cross}} < \eta h_e < \alpha_{\max}$, the accumulated contraction of Section B.2 can be written with $T = \tau + W/2$ as

$$C_\eta = \rho_\eta T + a_\eta W = \rho_\eta \tau + b_\eta W, \qquad b_\eta = g_\eta - \rho_\eta/2 > 0. \tag{24}$$

($b_\eta \geq g_\eta$ if $\rho_\eta \leq 0$, and $b_\eta = a_\eta + \rho_\eta/2$ if $\rho_\eta > 0$.) We model the aggregate persistent error as a mixture of decay rates sharing the accumulated-contraction variable C_η.

Assumption 1 (Effective persistent error)**.** *At each peak rate, the persistent contribution is*

$$E_\eta(W, T) = \int_{(0,\infty)} e^{-zC_\eta(W,T)} \, \mathrm{d}\mu_\eta(z),$$

where μ_η is a finite measure with density $\psi_\eta(z) \sim c_\eta z^{\gamma-1}$ near zero, $c_\eta, \gamma > 0$, and γ is the same for the peak rates being compared.

Here z is an effective decay rate, not a Hessian eigenvalue.

Proposition 2 (Power-law persistent error)**.** *Under Assumption 1, as $C_\eta \to \infty$,*

$$E_\eta \sim B_\eta C_\eta^{-\gamma}, \qquad B_\eta = c_\eta \Gamma(\gamma). \tag{25}$$

Proof. This is Watson's lemma (Olver, 1974) with large parameter C_η, as in the proof of Proposition 1. □

Local product approximation. Proposition 2 motivates the power-law surrogate

$$\bar{E}_\eta(\tau, W) = B_\eta(\rho_\eta \tau + b_\eta W)^{-\gamma},$$

which the fitted law replaces by a product of powers $\tau^{-q} W^{-s}$. The surrogate is the large-C_η form of E_η, not E_η itself: at finite C_η the local exponents of E_η sum to $-C_\eta E_\eta'(C_\eta)/E_\eta(C_\eta)$, which tends to γ only as $C_\eta \to \infty$. We assume $\alpha_{\text{cross}} < \eta h_e < 2$, so that $\rho_\eta, a_\eta, b_\eta > 0$; above the stability limit, $\rho_\eta < 0$ would give a negative time exponent, so the construction does not cover that regime.

Theorem 3 (Product approximation)**.** *Fix a reference point $\tau_0, W_0 > 0$, let $D_{\eta,0} = \rho_\eta \tau_0 + b_\eta W_0$, and let*

$$q_\eta = \frac{\gamma \rho_\eta \tau_0}{D_{\eta,0}}, \qquad s_\eta = \frac{\gamma b_\eta W_0}{D_{\eta,0}}, \qquad q_\eta + s_\eta = \gamma. \tag{26}$$

The product

$$E_{\text{loc},\eta} = \bar{E}_\eta(\tau_0, W_0)\left(\frac{\tau}{\tau_0}\right)^{-q_\eta}\left(\frac{W}{W_0}\right)^{-s_\eta} = K_\eta \, \tau^{-q_\eta} W^{-s_\eta} \tag{27}$$

is the only product of powers that matches $\bar{E}_\eta$ and its derivatives in $\log\tau$ and $\log W$ at (τ_0, W_0). For all $\tau, W > 0$,

$$0 \leq \log \frac{E_{\text{loc},\eta}}{\bar{E}_\eta} \leq \frac{\gamma}{8}\left[\log \frac{\tau/W}{\tau_0/W_0}\right]^2. \tag{28}$$

In particular, it matches the effect of lengthening warmup at a fixed budget $T = \tau + W/2$:

$$\left.\frac{\partial \log \bar{E}_\eta}{\partial W}\right|_{T,\,(\tau_0,W_0)} = -\frac{\gamma a_\eta}{D_{\eta,0}} = \frac{q_\eta}{2\tau_0} - \frac{s_\eta}{W_0}. \tag{29}$$

The exponents split γ according to the share of each term in $\rho_\eta\tau + b_\eta W$ at the reference point, and the error vanishes at a fixed warmup fraction.

Proof. Differentiating $\log \bar{E}_\eta$ in $\log\tau$ and $\log W$ gives $-q_\eta$ and $-s_\eta$ at the reference point, and matching the value fixes K_η. With $\theta = q_\eta/\gamma$ and $v = \log \frac{\tau/W}{\tau_0/W_0}$,

$$\frac{1}{\gamma}\log\frac{E_{\mathrm{loc},\eta}}{\bar{E}_\eta} = \log\big(\theta e^v + 1 - \theta\big) - \theta v,$$

which vanishes with its first derivative at $v = 0$ and has second derivative in $(0, \frac{1}{4}]$; Taylor's theorem gives Equation (28). The warmup derivative follows from $\partial\tau/\partial W = -1/2$ and $b_\eta - \rho_\eta/2 = a_\eta$. □

Corollary 1 (Higher peak rates shift the exponent toward warmup). *Hold h_e, γ and W_0/τ_0 fixed. On $\alpha_{\mathrm{cross}} < \alpha = \eta h_e < 2$, s_η increases and q_η decreases with η, with $s_\eta \to \gamma$ and $q_\eta \to 0$ as $\alpha \uparrow 2$.*

Proof. With $b = g - \rho/2$ and Equation (13),

$$\frac{s_\eta}{q_\eta} = \frac{b(\alpha)}{\rho(\alpha)}\,\frac{W_0}{\tau_0}, \qquad \frac{b(\alpha)}{\rho(\alpha)} = -\frac{1}{\log(\alpha-1)} + \frac{1}{2} - \frac{1}{\alpha}. \tag{30}$$

Its derivative $1/\{(\alpha-1)\log^2(\alpha-1)\} + 1/\alpha^2$ is positive, and it diverges as $\alpha \uparrow 2$. Since $q_\eta + s_\eta = \gamma$, the claims follow. □

Adding the progress term gives the fitted law in Equation (31). We use $W + w_0$ so that zero warmup is finite. The identity $q + s = \gamma$ belongs only to the local approximation, so we neither impose it nor use it to infer the global large-T behavior of the fitted law. Instead, we fit p, q, s freely at each peak learning rate. Each peak rate then has its own law, which can be fitted from a few short runs at that rate and extrapolated to longer horizons without runs at other peak rates. Sharing $q + s$ across peak rates fits the observed losses nearly as well (Section F.3), but it requires fitting runs at several peak rates.

B.4 Optimal duration under the fitted law

With coefficients $A, K, p, q, s, w_0 > 0$ and L_∞ fixed, the fitted law at horizon T is

$$\mathcal{L}_T(W) = L_\infty + A\tau^{-p} + K\tau^{-q}U^{-s}, \qquad U = W + w_0, \quad \tau = T - W/2, \quad 0 \le W \le T. \tag{31}$$

Proposition 3 (Unique finite-horizon optimum). *$\mathcal{L}_T$ is strictly convex on $[0, T]$. Its minimizer W_* is 0 if $\mathcal{L}_T'(0) \ge 0$, is T if $\mathcal{L}_T'(T) \le 0$, and otherwise is the unique solution of*

$$ApU_*^{s+1} = K\tau_*^m\Big(2s - \frac{qU_*}{\tau_*}\Big), \qquad m = p + 1 - q. \tag{32}$$

Proof. With $E = K\tau^{-q}U^{-s}$,

$$\mathcal{L}_T' = \frac{Ap}{2}\tau^{-p-1} + E\Big(\frac{q}{2\tau} - \frac{s}{U}\Big), \qquad \mathcal{L}_T'' = \frac{Ap(p+1)}{4}\tau^{-p-2} + E\Big[\Big(\frac{q}{2\tau} - \frac{s}{U}\Big)^2 + \frac{q}{4\tau^2} + \frac{s}{U^2}\Big] > 0.$$

Multiplying $\mathcal{L}_T' = 0$ by $2\tau^{p+1}U^{s+1}$ gives Equation (32). □

Long-horizon regimes. The balance in Equation (32) weighs the cost of one more warmup step against its lasting benefit. When $w_0 \ll W_* \ll T$, we have $U_* \approx W_*$ and $\tau_* \approx T$, and it reduces to

$$\underbrace{Ap\,W_*^{s+1}}_{\text{cost}} \approx \underbrace{2sK\,T^m}_{\text{benefit}}, \qquad \text{so} \qquad W_* \approx d\,T^\beta, \qquad d = \Big(\frac{2sK}{Ap}\Big)^{1/(s+1)}, \qquad \beta = \frac{p+1-q}{s+1}.$$

The exponent β decides what happens as $T \to \infty$:

Exponent	Optimal warmup	Regime
$\beta < 0$	$W_* = 0$	no warmup
$\beta = 0$	$W_* \to \max\{0,\, d - w_0\}$	bounded warmup
$0 < \beta < 1$	$W_* \approx d\,T^{\beta}$	sublinear growth
$\beta = 1$	$W_*/T \to x_1$	proportional growth
$\beta > 1$	$W_*/T \to x_\infty = \min\{1,\, 2s/(q+s)\}$	proportional growth

For $\beta \le 0$ the balance point does not grow with T, and for $0 < \beta < 1$ it grows but stays small relative to T, so the approximation above is self-consistent. For $\beta \ge 1$ it would give $W_* \gtrsim T$, so instead the optimum settles at a fixed fraction $x = W_*/T$. Writing $W = xT$ and dropping w_0, the loss above L_∞ is, up to the common factor $T^{-(q+s)}$,

$$A\,T^{q+s-p}\,(1 - x/2)^{-p} \;+\; K\,x^{-s}(1 - x/2)^{-q}.$$

For $\beta > 1$, $q + s < p$, so the first term vanishes, and minimizing the second gives x_∞. For $\beta = 1$, both terms remain, and $x_1 = \min\{1,\, y_*/(1 + y_*/2)\}$, where $y_* > 0$ solves

$$A p y_*^{s+1} = K(2s - q y_*). \tag{33}$$

Finite horizons. The regimes above describe $T \to \infty$; the following result holds at every finite horizon.

Corollary 2 (Growth at finite horizons). *If $q \le p+1$, the optimal warmup strictly increases with T wherever it is interior.*

Proof. At an interior optimum, $\mathcal{L}_T'(W_*) = 0$ and $\mathcal{L}_T''(W_*) > 0$, so $W_*'(T) = -\partial_T \mathcal{L}_T'/\mathcal{L}_T''$. Differentiating $\mathcal{L}_T'$ in T at fixed W and using $\mathcal{L}_T' = 0$,

$$\partial_T \mathcal{L}_T' = -\frac{Ap}{2}(p + 1 - q)\,\tau^{-p-2} - \frac{qE}{2\tau^2} < 0,$$

so $W_*'(T) > 0$. □

B.5 A STOCHASTIC MECHANISM FOR HORIZON DEPENDENCE

The quadratic model above is deterministic. Minibatch sampling gives a second route to horizon-dependent warmup, even with fixed curvature and a stable peak rate. We use the exact second-moment dynamics of least squares with Gaussian inputs (Bordelon & Pehlevan, 2022).

Setup. Inputs are Gaussian, labels are noiseless, and each step draws a fresh minibatch of size B:

$$x \sim \mathcal{N}(0, H), \qquad y = x^\top \theta_\star.$$

In the eigenbasis of H, let λ_i be the eigenvalues and $\Lambda = \mathrm{diag}(\lambda)$. We track the expected squared error of each mode and the expected loss,

$$m_{i,t} = \mathbb{E}\big[e_{i,t}^2\big], \qquad R_t = \tfrac{1}{2}\sum_i \lambda_i\, m_{i,t}, \qquad e_t = \theta_t - \theta_\star.$$

A step with learning rate u updates these moments linearly:

$$m_{t+1} = \mathcal{A}(\eta_t)\, m_t, \qquad \mathcal{A}(u) = \underbrace{I - 2u\Lambda + u^2\big(1 + B^{-1}\big)\Lambda^2}_{\text{each mode on its own}} + \underbrace{\frac{u^2}{B}\,\lambda\lambda^\top}_{\text{sampling noise}}. \tag{34}$$

This follows from the Gaussian identity

$$\mathbb{E}\big[\widehat{H}\Sigma\widehat{H}\big] = \big(1 + B^{-1}\big)H\Sigma H + B^{-1}H\,\mathrm{tr}(H\Sigma),$$

where $\widehat{H}$ is the minibatch covariance. The sampling-noise term moves error between modes. As $B \to \infty$ it vanishes, and Equation (34) reduces to gradient descent. Warmup uses $\eta_t = \eta \min\{(t+1)/W,\, 1\}$.

Because the sampling-noise term couples modes with different curvatures, steps at different learning rates no longer commute, so the order of the learning rates now matters.

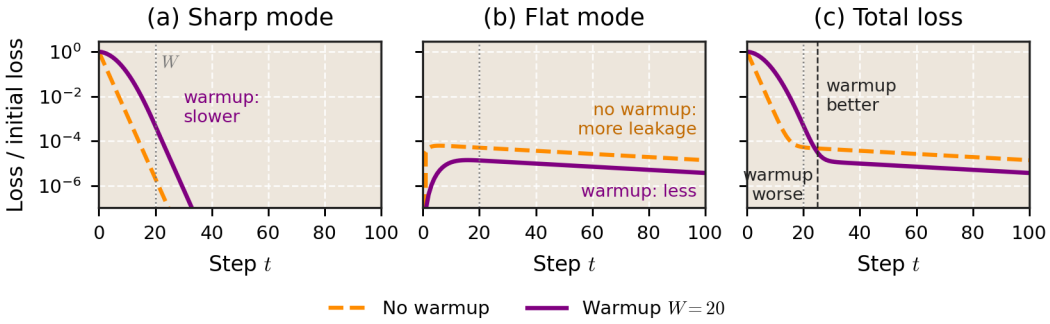


Figure 14: **Sampling flips the warmup ranking.** The two-mode example of Equation (35) with $\epsilon = 0.02$, $\eta h = 0.4$, $B = 2$, and $W = 20$; losses are relative to the initial loss. **(a)** Warmup empties the sharp mode more slowly. **(b)** Warmup leaks less error into the flat mode. **(c)** Warmup is worse until step 25, then settles at about a quarter of the no-warmup loss. The exact recursion in Equation (34) gives the same picture, with the crossing at step 24.

Proposition 4 (Batch size and the stable rate)**.** *At a constant rate $u > 0$, $m_t \to 0$ for every initial state if and only if*

$$\big(1 + B^{-1}\big) u\lambda_i < 2 \text{ for all } i \qquad \text{and} \qquad \frac{u}{B} \sum_i \frac{\lambda_i}{2 - (1 + B^{-1}) u\lambda_i} < 1.$$

Both conditions tighten as B decreases, so the supremum u_B of stable rates increases with B. For a single mode of curvature h,

$$u_B = \frac{2B}{(B+2)\,h},$$

and in general $u_B \to 2/\max_i \lambda_i$ as $B \to \infty$.

Proof. Write $\mathcal{A}(u) = G + (u^2/B)\,\lambda\lambda^\top$, where $G \geq 0$ is diagonal. Such a matrix has spectral radius below one exactly when

$$\max_i G_{ii} < 1 \qquad \text{and} \qquad \frac{u^2}{B}\,\lambda^\top (I - G)^{-1} \lambda < 1.$$

These are the two stated conditions. □

A two-mode example. Consider two modes:

- a *sharp* mode with curvature h, which holds all the initial error;
- a *flat* mode with curvature ϵh, $\epsilon \ll 1$, which starts with no error.

Write $a = uh$ for the step size relative to the sharp curvature, and $c = 1 + 2/B$. Dropping terms of order ϵ^2, one step of Equation (34) reads

$$\begin{aligned} m_{\text{sharp}} &\leftarrow \big(1 - 2a + c\,a^2\big)\, m_{\text{sharp}}, \\ m_{\text{flat}} &\leftarrow \big(1 - 2\epsilon a\big)\, m_{\text{flat}} + \frac{\epsilon a^2}{B}\, m_{\text{sharp}}. \end{aligned} \tag{35}$$

The sharp mode decays quickly. The flat mode decays slowly, and sampling noise leaks error into it from the sharp mode. Per step,

$$\frac{\text{error leaked into the flat mode}}{\text{error removed from the sharp mode}} = \frac{\epsilon a^2/B}{2a - c\,a^2} = \frac{\epsilon}{B} \cdot \frac{a}{2 - c\,a}.$$

This ratio grows with a: large steps are leaky, and small steps are clean.

Why warmup loses early and wins later. Warmup replaces the first large steps with small ones. This has two effects:

- **Cost.** The sharp mode is emptied more slowly (Figure 14a). This extra error decays quickly after warmup.
- **Benefit.** Less error leaks into the flat mode (Figure 14b). This saving decays slowly, because the flat mode shrinks only by a factor $1 - 2\epsilon a$ per step.

After warmup, both schedules take the same steps. The difference between their losses n steps after warmup is therefore a sum of two decaying terms:

$$R(W) - R(0) = \underbrace{C\, r_{\text{fast}}^{n}}_{\text{cost}} - \underbrace{D\, r_{\text{slow}}^{n}}_{\text{benefit}}, \qquad C, D > 0, \qquad r_{\text{fast}} < r_{\text{slow}}.$$

The cost dominates first and the benefit later, so the ranking flips once: warmup is worse early and better late (Figure 14c). Without sampling ($B \to \infty$), nothing leaks, so $D = 0$ and warmup is worse at every horizon. The example shows that sampling alone can make the warmup ranking depend on the horizon.

C Robustness of warmup behavior

C.1 Warmup ordering after decay

We branch the intermediate model's warmup–stable trajectories into linear decay after warmup has ended. For each peak rate and horizon, we compare validation-loss rankings immediately before decay and at its endpoint over the same completed branches. Spearman correlation measures how well the ordering of warmup durations is preserved, regardless of the overall loss reduction from decay. We also report agreement between pairwise orderings and whether the best duration is unchanged. Post-decay regret is the endpoint loss of the warmup that was best before decay, minus the best endpoint loss among those branches.

Table 10: **Decay-branching design.** Model: 100M; AdamW; seed 0. Fork horizons $C \in \{8, 16, 32, 64, 128\}$k updates; linear decay to zero lasts 5k updates, with endpoint $C + 5$k. Only $W \leq C - 2$k is eligible. The comparison uses the 128 branches that completed.

Peak LR (10^{-3})	Source warmups (updates)
0.1	0, 256, 1,000, 4,000, 16,000
0.3	0, 128, 512, 2,000, 8,000
1	0, 512, 2,000, 8,000, 30,000
3	512, 1,500, 8,000, 16,000, 30,000
5	768, 2,000, 6,000, 16,000, 30,000
7	1,000, 3,000, 8,000, 20,000, 50,000

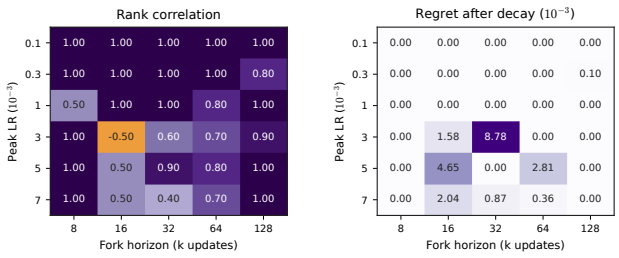


Figure 15: **Warmup ordering before and after decay.** Left: Spearman correlation between warmup rankings at the fork and after 5k decay updates, within each peak-rate family (1: identical ordering; -1: reversed). Right: post-decay regret of the warmup selected before decay, in 10^{-3} loss units. Each cell compares the same warmups before and after decay.

Figure 15 and Table 11 show that decay mostly preserves the ordering of warmup durations. The best duration is unchanged in 22 of 30 fork–family cases, and in 5 of 6 families at the latest fork. Where it changes, at intermediate forks and higher peak rates, decay favors a longer warmup. Warmup–stable comparisons are therefore a reliable guide to warmup choice under decay.

Table 11: **Ordering agreement and selection after decay.** Preserved pairs have the same strict loss ordering before and after decay. "Same best" counts peak-rate families whose best warmup is unchanged by decay. Mean regret weights families equally; maximum regret is the largest family regret. Both are in 10^{-3} loss units.

Fork (k)	Branches	Preserved pairs	Same best	Mean regret	Max regret
8	18	19/20	6/6	0.00	0.00
16	22	27/31	3/6	1.38	4.65
32	29	50/56	4/6	1.61	8.78
64	29	49/56	4/6	0.53	2.81
128	30	57/60	5/6	0.02	0.10

C.2 Horizon and peak-rate dependence with Muon

We repeat the warmup sweep with Muon on the 100M parameter model (Jordan, 2024; Liu et al., 2025a). Muon updates the hidden attention and MLP matrices; auxiliary AdamW updates the embeddings, output head, and normalization weights. We vary only Muon's peak learning rate and warmup duration, keeping the auxiliary AdamW learning rate and warmup fixed across the sweep. Both groups keep their own learning rate constant after warmup. Table 12 gives the optimizer settings and run coverage.

Figure 16 and Table 13 show the same dependence on horizon and peak rate as with AdamW. At the lowest peak rate the best warmup stays short and the loss is nearly flat across short durations. At higher rates it grows with the horizon. The preferred durations and learning-rate scales differ from AdamW, but the warmup tradeoff carries over to Muon.

Table 12: **Muon protocol and controls.** Model: 100M; seed 0; 128k updates. All remaining data, batch, precision, clipping, and evaluation settings match Tables 6 and 7.

Setting	Value
Muon parameters	Hidden matrices: 59,013,120
Auxiliary AdamW parameters	Embeddings, output head, norms: 41,104,000
Muon momentum / Nesterov	0.95 / enabled
Newton–Schulz iterations	5
Muon update scaling	$0.2\sqrt{\max(d_{\text{out}}, d_{\text{in}})}$ after approximate orthogonalization
Muon peak LR (10^{-3})	0.1, 0.2, 0.5, 1, 2
Muon warmups (updates)	0, 512, 2,000, 8,000, 16,000, 50,000
Fixed auxiliary peak LR / warmup	10^{-3} / 2,000 updates
Auxiliary (β_1, β_2) / ϵ	$(0.9, 0.99)$ / 10^{-6}
Weight decay / warmup shape	0 for both groups / linear for each group
Runs	30: 6 warmups at each of 5 Muon peak rates

Table 13: **Best measured Muon warmup (updates).** Columns give the training horizon; only completed runs with $W < T$ are eligible. Ties favor shorter warmup. The loss curves in Figure 16 show how sharply each minimum is determined.

Muon peak LR (10^{-3})	8k	16k	32k	64k	128k
0.1	0	512	512	512	2,000
0.2	512	2,000	2,000	8,000	8,000
0.5	512	2,000	8,000	8,000	8,000
1	512	2,000	8,000	16,000	16,000
2	2,000	2,000	8,000	16,000	50,000

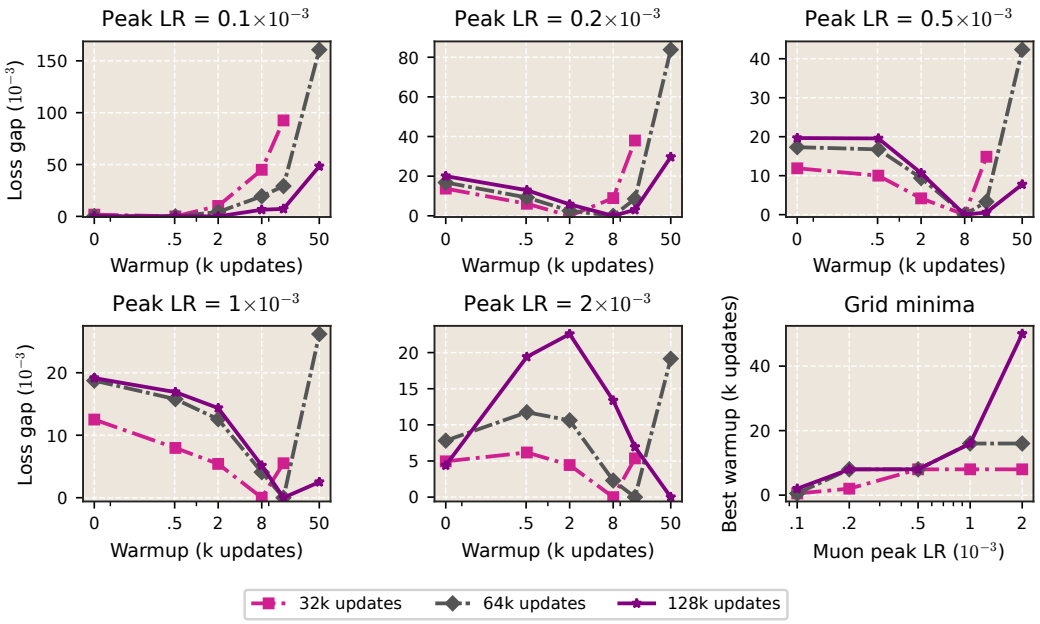


Figure 16: **Warmup dependence with Muon.** Loss gaps are relative to the best tested warmup at each horizon and Muon peak rate, using only $W < T$ from the 30 completed runs.

C.3 Horizon dependence under different warmup shapes

We test whether warmup duration retains its horizon and peak-rate dependence within each warmup shape. On the 60M parameter model, we vary duration under linear, half-cosine, and concave-quadratic warmup, followed by a constant peak rate. The data, initialization seed, optimizer, and batch settings match the base grid. Table 14 defines the schedules and sweep; Figure 17 shows their learning-rate profiles.

For a warmup multiplier m, the accumulated learning rate can be written as

$$\tau_m = T - c_m W, \qquad c_m = 1 - \int_0^1 m(u)\,\mathrm{d}u.$$

Substituting this quantity into Equation (7), the same interior balance as in Section 4.1 gives, for $w_0 \ll W_* \ll T$,

$$W_* \approx \left(\frac{sK}{c_m A p}\right)^{1/(s+1)} T^{\beta}, \qquad \beta = \frac{p+1-q}{s+1}.$$

Holding the other law coefficients fixed, a smaller c_m therefore predicts longer warmup with the same exponent of T. Linear and half-cosine warmup have the same accumulated learning rate. Concave-quadratic warmup has a smaller progress penalty per warmup step, which shifts the balance toward longer durations. The persistent-error coefficients can also depend on warmup shape.

Figure 18 and Table 15 compare durations across horizons within each shape. The shift toward longer warmup is clearest at the highest peak rate, while lower rates favor shorter durations. Linear and half-cosine warmup show similar qualitative trends. At the highest peak rate, the concave-quadratic optimum continues to move toward longer durations at the latest horizon, consistent with its smaller progress penalty.

Table 14: **Warmup shapes and design.** Here $x = t/W$ during warmup, $\eta_t = \eta m(x)$, and $m = 1$ after warmup. The last columns use continuous integrals, with $T \geq W$. Model: 60M; AdamW; peak LRs $\{0.5, 1, 5\} \times 10^{-3}$; nonlinear warmups $\{500, 1000, 5000, 15000, 50000\}$; horizon 128k.

Shape	$m(x)$	$\int_0^1 m(x)\,\mathrm{d}x$	$\tau = \int_0^T \eta_t/\eta\,\mathrm{d}t$
Linear	x	$1/2$	$T - W/2$
Half-cosine	$(1 - \cos(\pi x))/2$	$1/2$	$T - W/2$
Concave quadratic	$2x - x^2$	$2/3$	$T - W/3$

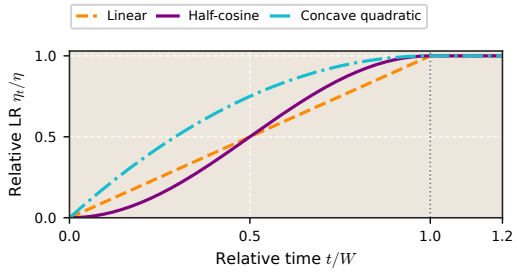


Figure 17: **Warmup learning-rate profiles.** Relative learning rate η_t/η against relative time t/W. All shapes reach the same peak at the end of warmup and remain constant afterward.

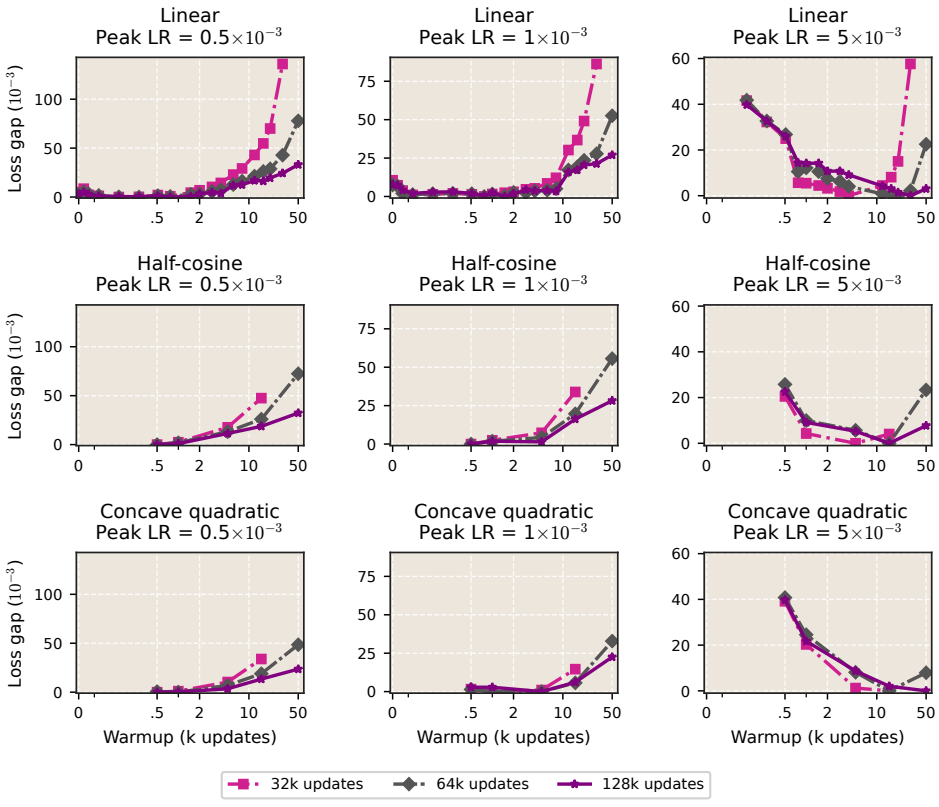


Figure 18: **Horizon-dependent warmup under different shapes.** Each panel compares durations within a fixed shape and peak rate. Rows show linear, half-cosine, and concave-quadratic warmup; columns show peak learning rates. Loss gaps are relative to the best tested duration within each shape, peak rate, and horizon, restricted to $W < T$. Each shape uses its own tested duration grid; linear curves follow the base-grid eligibility rule in Section A.1. Lines connect measured candidates.

Table 15: **Best measured warmup across horizons within each shape (updates).** Each row tracks the loss-minimizing duration for a fixed shape and peak rate. Columns give the training horizon; only $W < T$ is eligible, and exact ties favor shorter warmup.

Shape	Peak LR (10^{-3})	8k	16k	32k	64k	128k
Linear	0.5	256	256	256	384	768
Linear	1	128	128	768	768	1,500
Linear	5	768	2,000	4,000	20,000	30,000
Half-cosine	0.5	500	500	500	500	500
Half-cosine	1	500	500	500	500	500
Half-cosine	5	1,000	1,000	5,000	15,000	15,000
Concave quadratic	0.5	500	500	500	1,000	500
Concave quadratic	1	500	500	1,000	1,000	5,000
Concave quadratic	5	1,000	5,000	15,000	15,000	50,000

D Additional selection results

D.1 Direct warmup-scaling baseline

A simpler alternative is to extrapolate the measured optimal durations directly. We fit

$$\widehat{W}(T) = \min\{T,\ c(T/32\text{k})^{\alpha}\} \tag{36}$$

and map the prediction to the nearest available warmup. This baseline tests whether modeling the loss surface provides information beyond the observed duration trend.

Table 16: **Selection results including the direct warmup-scaling baseline.** All methods are evaluated on the same retained candidate sets. For the tuned rules, the alternating column would measure one checkpoint ahead rather than extrapolation, so it is omitted.

Selector	Parameters	Alternating	Four horizons	128k
Absolute fit	6	2.05 ± 1.61	2.26 ± 0.66	3.05
Difference fit	5	1.45 ± 2.22	$\mathbf{1.58} \pm 0.77$	2.61
Direct warmup scaling	2	$\mathbf{1.34} \pm 4.94$	1.99 ± 0.69	2.86
Fixed best duration	0	–	3.00 ± 1.75	3.87
Fixed best fraction	0	–	1.73 ± 0.28	**2.06**
1k target	0	15.47 ± 10.80	13.79 ± 0.22	13.56
10% target	0	13.39 ± 4.10	12.58 ± 0.90	11.72

With the full warmup grid, extrapolating the measured optimum directly is competitive with the loss-law fits, as are the tuned rules of Section A.2, because the grid supplies accurate pairs of horizon and optimal warmup. That information is unavailable in the three-run setting, where all three rules incur several times the regret of the absolute fit (Table 19). Their anchor is restricted to the three fitted durations, whereas the loss law uses every post-warmup checkpoint and can recommend durations between and beyond them.

D.2 Near-optimal measured warmups

The best tested warmup can make the optimum appear more sharply determined than the measured loss differences imply. For each family and target horizon, we therefore also consider the tested durations whose loss is within ϵ of the best observed loss,

$$\mathcal{S}_\epsilon(T) = \left\{ W \in \mathcal{G}_T : L(W,T) - \min_{u \in \mathcal{G}_T} L(u,T) \le \epsilon \right\}.$$

The set contains only measured warmups; gaps between tested durations are not filled by interpolation. Table 17 shows that several measured durations are often effectively competitive. The sets expand as the loss tolerance increases. This helps explain why two selectors can choose visibly different warmups while incurring similar regret.

Table 17: **Near-optimal measured warmups.** The table summarizes all retained family–horizon comparisons used by the later-horizon benchmark. "Multiple" counts comparisons with more than one tested duration within the stated loss tolerance.

Tolerance (10^{-3})	Multiple	Median count	Median fraction of grid
1	62 / 132	1	16.7%
2	92 / 132	2	20.0%
5	124 / 132	4	40.8%

E Extrapolation from short runs

E.1 Three-run extrapolation

For each model-size and peak-rate family, we fit the loss law using three separated warmup trajectories and checkpoints shared after the longest warmup has ended. The fitted coefficients are frozen before evaluating later horizons.

Table 18: **Three-run extrapolation protocol.**

Item	Choice
Target fitting warmups	2k, 8k, 16k updates
Fit horizons	20k and 32k updates
Target horizons	50k, 75k, 100k, 128k updates
Loss-law fits	absolute and difference forms
Primary practical predictor	absolute fit
Target candidates	same retained warmup grid as the full-grid benchmark

Table 19: **Complete three-run comparison.** Mean regret on the same later-horizon candidate grids as the main-text experiment, in 10^{-3} loss units.

Selector	Through 20k (29 families)	Through 32k (33 families)
Absolute fit	$\mathbf{2.83} \pm 0.16$	$\mathbf{2.51} \pm 0.27$
Difference fit	4.22 ± 0.29	2.62 ± 0.36
Direct warmup scaling	9.09 ± 1.25	7.08 ± 0.82
Fixed best duration	9.31 ± 1.21	7.61 ± 1.12
Fixed best fraction	11.72 ± 0.82	8.47 ± 0.62
1k target	14.28 ± 0.30	13.79 ± 0.22
10% target	13.42 ± 0.96	12.58 ± 0.90

E.2 Recommendations beyond the observed warmups

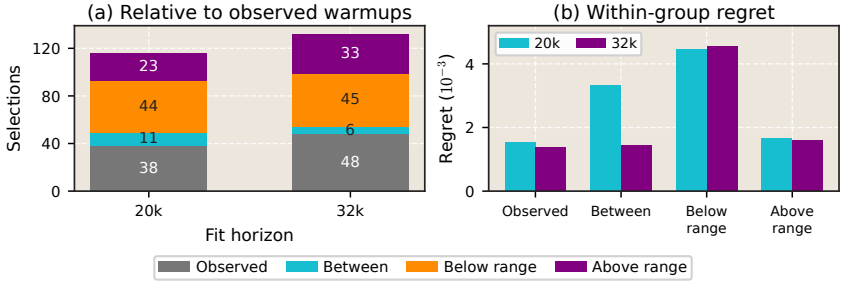


Figure 19: **Where three-run warmup recommendations fall relative to the observed warmups.** Left: number of family–horizon selections that coincide with an observed fitting warmup, fall between observed warmups, or lie below or above their range. Right: mean measured regret within each group. All selections are frozen before target losses are scored.

The three-run predictor is not restricted to selecting one of the warmup durations used for fitting. We group each frozen recommendation according to whether it coincides with an observed fitting warmup, lies between two observed warmups, or falls below or above the observed range. Every recommendation is then scored using the measured target-horizon trajectory at the selected duration.

Figure 19 shows that many selections use durations that were not used for fitting. In particular, recommendations above the observed warmup range remain low-regret, so the fitted loss surface can support useful extrapolation in warmup duration as well as in training horizon.

E.3 Sensitivity to the choice of fitting runs

We next test how the warmup range covered by the three fitting runs affects later-horizon selection. The alternatives either span the available warmup range broadly or concentrate all three runs near its short, middle, or long end. We also evaluate randomly chosen triplets. All alternatives are defined from schedule metadata alone, without using their later-horizon losses.

For the non-default designs, target positions are specified in normalized $\log(W + w_0)$ coordinates over the available warmup range. The wide design targets the two ends and midpoint; the clustered designs place all three targets near the short, middle, or long end. The random baseline samples admissible triplets uniformly without replacement. All designs use the same fit horizon and fitting procedure and require shared post-warmup observations. Changing the longest observed warmup also changes the shared checkpoint window, so this comparison varies both the warmup durations and the available checkpoint window.

Figure 20 and Table 20 show that the practical predictor is not tied to the exact default triplet. Broad log-space coverage performs similarly or better on this retrospective benchmark, and random triplets remain competitive on average. Concentrating all three fitting warmups near the long end is substantially less reliable. These results support choosing fitting runs whose warmups span the relevant range.

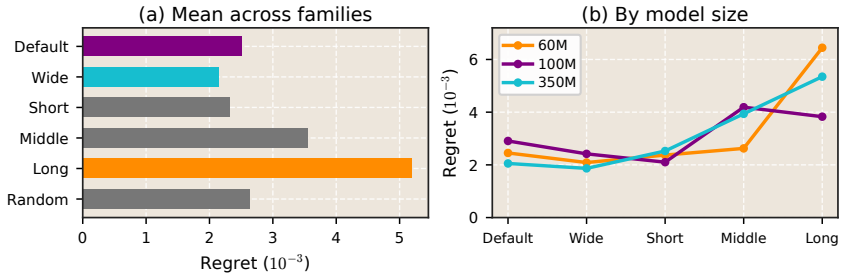


Figure 20: **Sensitivity to the warmup coverage of three fitting runs.** Left: mean four-horizon regret across families. Right: the same comparison split by model size. Random-triplet results are first averaged within family, so families with more admissible triplets receive no additional weight.

Table 20: **Choice of three fitting runs and selection results.** All fits use shared post-warmup checkpoints through 32k, with at least four shared checkpoints. Target positions are normalized coordinates in $\log(W + w_0)$ over each family's available warmup range. The default row uses the exact warmup durations from the main experiment. Family values first average regret over the four target horizons; random triplets are averaged within family before aggregation. Regret is in 10^{-3} loss units.

Fitting-run design	Log-range positions / rule	Mean regret	Median family	Worst family
Default triplet	Main-experiment runs	2.51	1.67	8.88
Wide coverage	0, 0.5, 1	2.15	1.47	6.80
Short-end cluster	0, 0.1, 0.2	2.32	1.47	12.34
Middle cluster	0.4, 0.5, 0.6	3.55	2.13	13.44
Long-end cluster	0.8, 0.9, 1	5.19	1.67	49.95
Random triplets	Up to 8 admissible triplets	2.63	2.21	6.72

E.4 SENSITIVITY TO THE NUMBER OF FITTING RUNS

The three-run experiment asks how little fitting data can support useful extrapolation. We vary the number of observed warmup trajectories at a fixed 32k fit horizon. Unlike the main three-run protocol (Section E.1), which uses only checkpoints shared by all three fitting runs, this sweep uses all post-warmup observations from each selected trajectory, so the number of fitting runs can vary across the available grid. Figure 21 shows a clear benefit from moving beyond very sparse coverage, followed by smaller, nonmonotonic changes as the observed grid becomes denser. The effect depends on which trajectories are added, but the same pattern appears across model sizes. Three fitting runs are thus a useful low-budget operating point. Table 21 records the corresponding reference values.

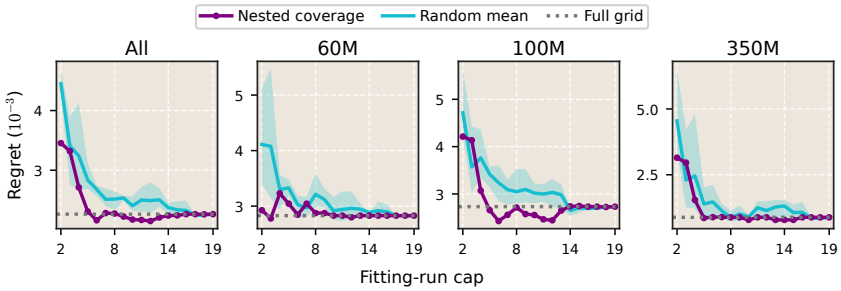


Figure 21: **Selection as the number of fitting runs increases.** At a cap n, each family uses up to n available trajectories. Fitting uses checkpoints through 32k. Purple shows the nested coverage order; cyan shows the mean of three fixed random orders, with shading spanning their aggregate curves. Dotted lines show the full-grid reference. The same 33 families are used throughout; their available counts range from 3 to 19. Regret averages the four later target horizons.

Table 21: **Reference fits for varying the number of runs.** The main three-run protocol uses shared post-warmup checkpoints, while the count sweep uses all post-warmup observations from each selected trajectory. The full-grid row reproduces the corresponding full-prefix benchmark. Regret is in 10^{-3} loss units.

Reference	Fitting runs	Checkpoints used	Mean regret
Main three-run fit	3	Shared post-warmup	2.51
Count-sweep fit	3	All selected post-warmup	3.32
Full-grid fit	All available	All post-warmup	2.26

E.5 FIT HORIZON AND SHORT-RUN BUDGET

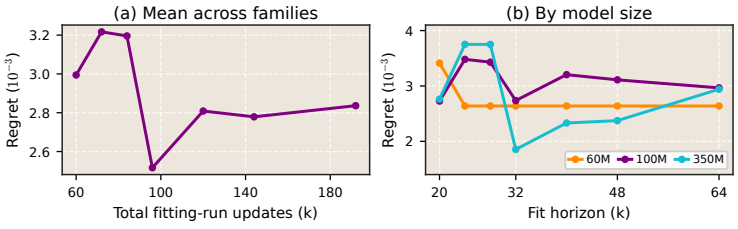


Figure 22: **Sensitivity to fit horizon and short-run budget.** The same three fitting trajectories are used at every fit horizon, from 20k to 64k. All comparisons use the same 29 families and the 128k target horizon. Total fitting-run updates equal three times the fit horizon, excluding the target run. Left: mean final-horizon regret against the total updates observed across the fitting runs. Right: the same comparison by model size as the fit horizon increases.

We next vary how long the same three fitting trajectories are observed before the fitted coefficients are frozen. All fit horizons are evaluated on the same families and at the same final target horizon, so the comparison changes only the amount of short-run data available to the fit.

Figure 22 shows that a longer fit horizon does not produce a monotonic improvement. The error changes most when the available prefix is still short and then remains in a similar range across later fit horizons. Simply extending every fitting trajectory is therefore not guaranteed to improve the final selection.

E.6 Long-horizon extrapolation protocol

The long-horizon experiment repeats the same procedure with the 100M model at batch size 128 and longer target horizons. For each peak rate, we fit the loss law independently using three fitting runs; the frozen absolute fit is minimized continuously over warmup duration at each target horizon.

Table 22: **Long-horizon protocol.**

Item	Setting
Model / batch / sequence length	100M / 128 sequences / 256 tokens
Fit horizons	50k and 75k updates
Target horizons	250k, 500k, and 1M updates
Fitting warmups, LR 0.5×10^{-3}	0, 512, 8k warmup updates
Fitting warmups, LR 2×10^{-3}	8k, 16k, 32k warmup updates
Fitting warmups, LR 5×10^{-3}	12k, 16k, 32k warmup updates
Short-warmup baseline	1k warmup updates at every peak rate
Evaluation	validation loss vs. 1k and 10% warmup baselines

The predicted schedule is trained from initialization. Because these targets use selected schedules instead of a dense warmup sweep, we report validation loss directly, not regret to an oracle warmup.

E.7 Model-scale extrapolation

Our focus is the horizon dependence of warmup and extrapolation to longer training runs at a fixed model size. We also observe similar behavior across model scales: a larger model at a lower peak learning rate can exhibit a similar dependence of preferred warmup on the training horizon to a smaller model at a higher rate.

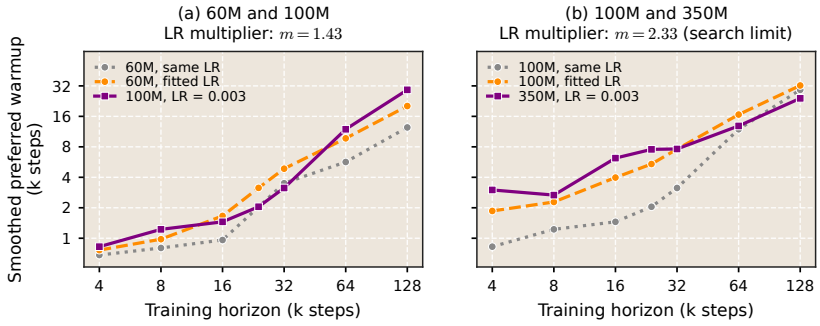


Figure 23: **Warmup behavior across model scales.** The larger model uses peak LR 0.003; the smaller model uses the same LR (gray) or a fitted higher LR (orange). We average $\log(W + 32)$, with W in steps, over seed-0 warmup sweeps with weights proportional to $\exp(-L/0.002)$, then fit monotonically in log LR. The multiplier minimizes weighted squared log-warmup differences across horizons. The 100M-to-350M fit reaches the LR limit, 0.007.

Figure 23 illustrates this correspondence. We compare smoothed estimates of the preferred warmup and fit one learning-rate multiplier m per model pair, shared across horizons, with $\eta_{\text{small}} = m\eta_{\text{large}}$. This shift improves the overall agreement, although differences remain at individual horizons. This

correspondence suggests that a model-size correction, such as a rescaling of the effective peak learning rate, could extend horizon-based warmup predictions across scales.

F SENSITIVITY OF THE LOSS LAW

F.1 WARMUP EXPONENT

We compare the free warmup exponent with a fixed-s ablation at $s = 1/2$.

Table 23: **Sensitivity to the warmup exponent.** RMSE and regret are in 10^{-3} loss units. Lower values are better.

Metric	Free s	Fixed $s = 1/2$
Full-data absolute-fit median RMSE	**4.92**	5.94
Held-out absolute-fit median RMSE	**4.98**	5.96
Four-horizon absolute-fit regret	2.26 ± 0.66	$\mathbf{2.00} \pm 0.35$
Three-run absolute-fit regret, 20k (29 families)	$\mathbf{2.83} \pm 0.16$	2.87 ± 0.32
Three-run absolute-fit regret, 32k	$\mathbf{2.51} \pm 0.27$	2.64 ± 0.23
Four-horizon difference-fit regret	$\mathbf{1.58} \pm 0.77$	$\mathbf{1.58} \pm 0.67$
Three-run difference-fit regret, 20k (29 families)	$\mathbf{4.22} \pm 0.29$	4.85 ± 0.32
Three-run difference-fit regret, 32k	$\mathbf{2.62} \pm 0.36$	2.86 ± 0.21

The free exponent improves fit quality, while selection performance remains similar across the reported comparisons.

F.2 ZERO-WARMUP OFFSET

We refit the model over a range of fixed offsets in $W + w_0$.

Table 24: **Sensitivity to the fixed warmup offset.** Each row refits all families with the indicated global offset. Here, later-horizon regret averages the dense 33k–128k grid, rather than the four target horizons used in the main comparison. Regret is in 10^{-3} loss units.

	Difference fit		Absolute fit	
w_0 (updates)	Later-horizon regret	Final-horizon regret	Later-horizon regret	Final-horizon regret
8	1.50	2.58	2.32	3.54
16	1.84	2.54	2.09	2.83
32	1.85	2.61	2.49	3.05
64	1.61	2.67	2.41	3.16
128	1.68	2.92	2.28	3.06

The offset w_0 only keeps the law finite at $W = 0$, so we fix it rather than fit it. Table 24 refits every family with offsets from 8 to 128 updates. Across this 16-fold range, later-horizon regret changes by at most 0.4×10^{-3} for either fitting form, so the fits are insensitive to the choice; we use $w_0 = 32$ updates throughout.

F.3 SHARING THE PERSISTENT EXPONENT SUM ACROSS PEAK RATES

The local product construction in Appendix B motivates $q_\eta + s_\eta = \gamma$ at a fixed reference point, while the empirical law fits q and s independently in every peak-rate family. We test an intermediate constraint by jointly fitting all peak-rate families at each model size with

$$q_\eta + s_\eta = \gamma_N,$$

where the shared exponent sum γ_N is learned from the same fitting data. The remaining parameters stay family-specific. The joint objective averages the mean Huber loss equally across peak-rate families and retains the original parameter bounds. For J peak-rate families, this reduces the parameter count from $6J$ to $5J + 1$.

We evaluate the constraint in two settings. Full-data fits measure descriptive fit quality. Separately, fits restricted to the early training window learn both the family-specific parameters and the shared exponent sum from those prefixes, freeze them, and select warmup at the later target horizons.

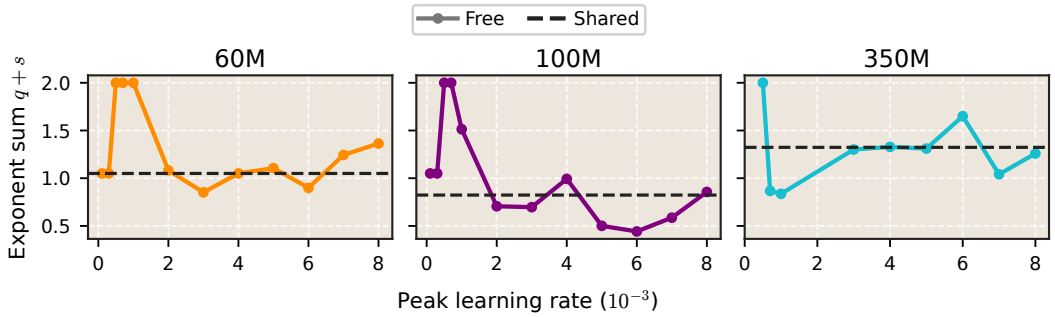


Figure 24: **Freely fitted and shared persistent exponent sums.** Points show $q_\eta + s_\eta$ from independent full-data fits at each peak learning rate. Dashed lines show the single sum fitted jointly across all peak-rate families of the same model size. Points at 2.0 lie on the parameter bound ($q = s = 1$).

Figure 24 compares the freely fitted sums $q_\eta + s_\eta$ with the shared value γ_N. For most moderate and large peak rates the free sums already lie near the shared value. The largest departures occur at the smallest peak rates, where several free fits reach the upper bound $q_\eta + s_\eta = 2$, that is, $q_\eta = s_\eta = 1$. These exponents are set by the bounds rather than determined by the data, so we do not read them as a change in the underlying relaxation exponent.

Figure 25 and Table 25 show that sharing the sum preserves most of the descriptive fit quality while reducing the exponent parameters. Its effect on later warmup selection is mixed across model sizes: it improves some groups and degrades others. We therefore retain independently fitted exponents in the primary predictor, while the ablation shows that a much more constrained cross-rate parameterization captures nearly the same observed loss structure.

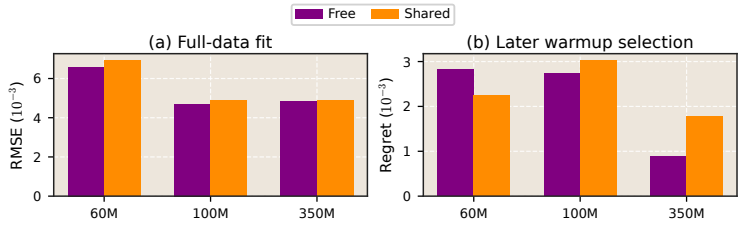


Figure 25: **Effect of sharing $q + s$ across peak learning rates.** Left: median descriptive RMSE for free and shared-sum fits at each model size. Right: later-horizon selection regret after fitting only through the 32k fit window. The shared sum is learned from the same fitting data and is not transferred from the full-data fit.

Table 25: **Shared-exponent-sum ablation.** RMSE and regret are in 10^{-3} loss units. RMSE and R^2 are medians over peak-rate families. Full-data fit quality is descriptive; later regret averages families and the four later target horizons after fitting through 32k.

Model	Fit	Full RMSE	Full R^2	Later regret
60M	Free	**6.59**	**0.993**	2.83
60M	Shared	6.92	**0.993**	**2.24**
100M	Free	**4.71**	**0.997**	**2.73**
100M	Shared	4.89	**0.997**	3.02
350M	Free	**4.81**	**0.997**	**0.88**
350M	Shared	4.88	**0.997**	1.77

G COMPLETE FIT GALLERY

The plots below show the full-data absolute fit for every analyzed family. Points are measured validation losses and curves are model predictions at the displayed horizons.

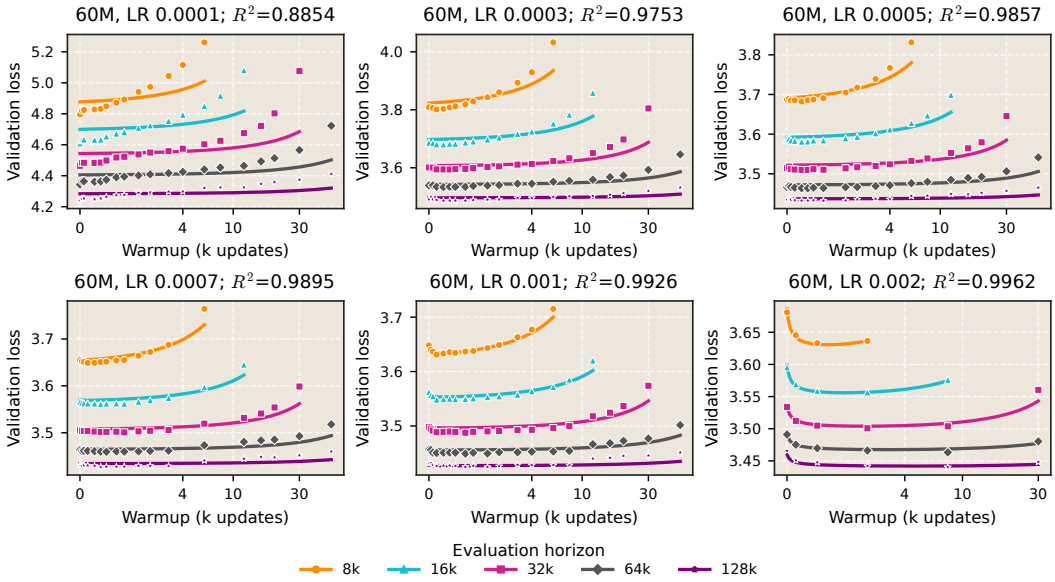


Figure 26: Full-data absolute fits: 60M, lower peak learning rates.

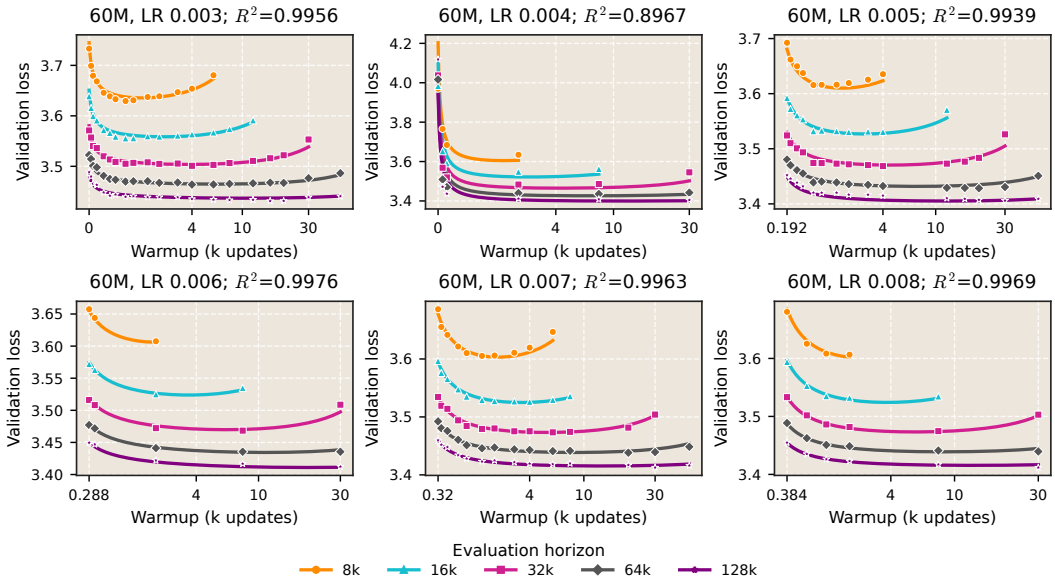


Figure 27: Full-data absolute fits: 60M, higher peak learning rates.

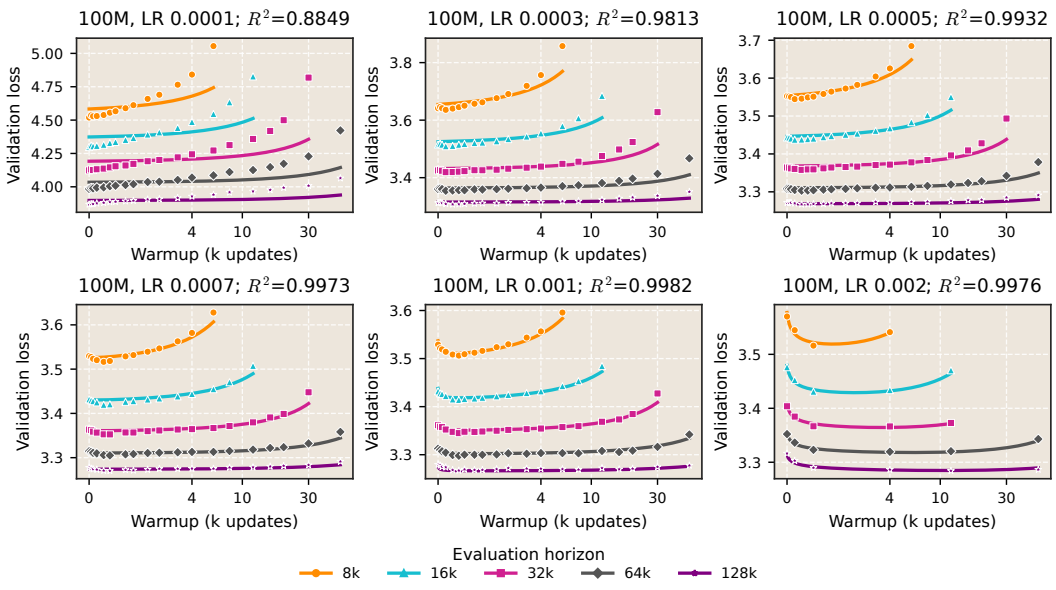


Figure 28: Full-data absolute fits: 100M, lower peak learning rates.

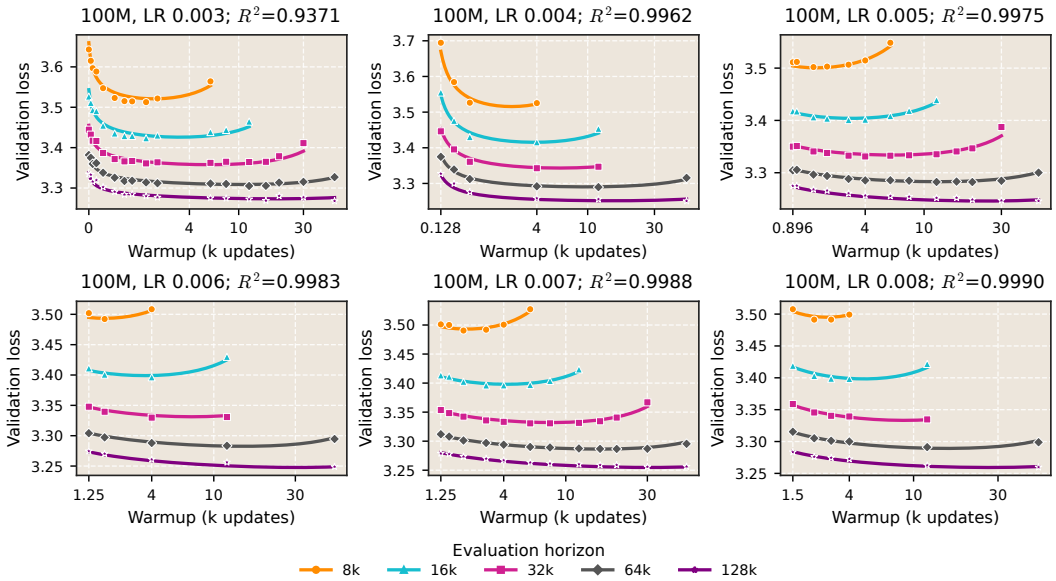


Figure 29: Full-data absolute fits: 100M, higher peak learning rates.

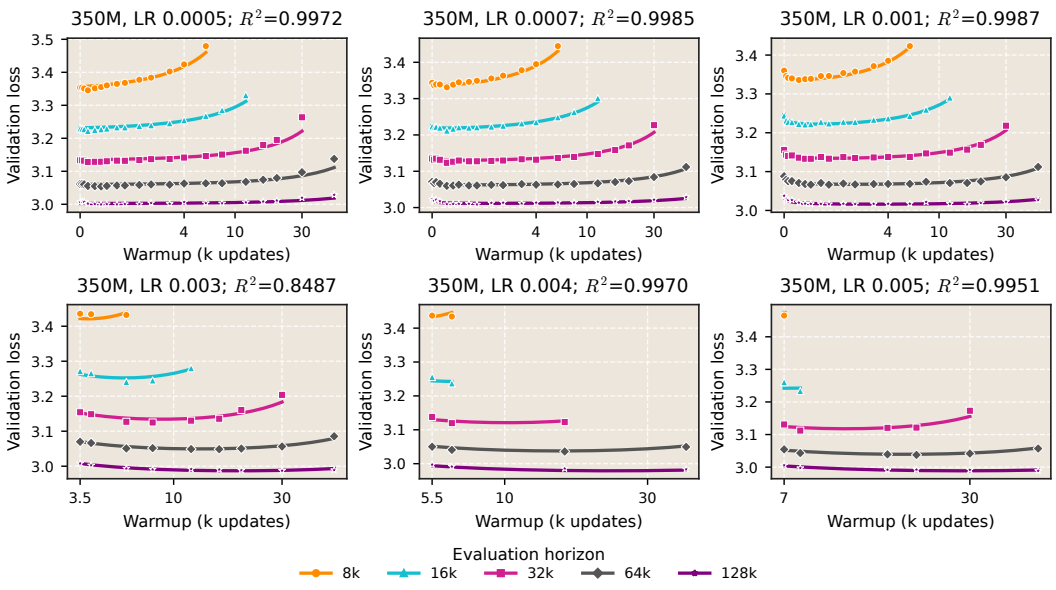


Figure 30: Full-data absolute fits: 350M, lower peak learning rates.

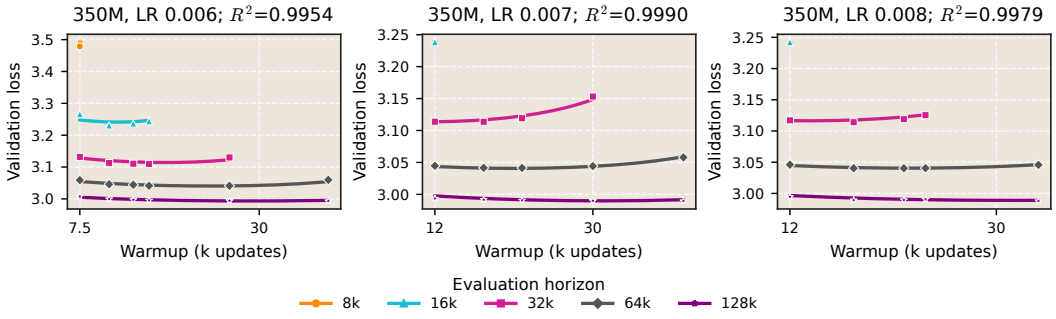


Figure 31: Full-data absolute fits: 350M, higher peak learning rates.

H Related work

Warmup. Learning-rate warmup is widely used in large-batch training and language models (Goyal et al., 2017; Black et al., 2022; Groeneveld et al., 2024). Recent large-model recipes use warmups ranging from 500 to 8,000 optimizer steps (Grattafiori et al., 2024; DeepSeek-AI, 2024; Kimi Team, 2025). Prior work has connected warmup to adaptive-optimizer statistics, update size, curvature, and early optimization dynamics (Liu et al., 2020; Ma & Yarats, 2021; Gilmer et al., 2022; Kalra & Barkeshli, 2024; Kosson et al., 2024; Alimisis et al., 2026; Liu et al., 2025b). Our focus is complementary: we study how the loss-optimal duration changes with the training horizon and peak learning rate.

Hyperparameter scaling. Prior work predicts or transfers hyperparameters such as the learning rate, batch size, and weight decay across model width, training horizon, and compute (Yang et al., 2021; Bjorck et al., 2025; Filatov et al., 2024; Li et al., 2025; Zhang et al., 2025; Bergsma et al., 2025). Warmup can also shift the learning-rate optimum (Filatov et al., 2024) and affect compute-optimal comparisons (Porian et al., 2024). Our experiments keep model size and peak learning rate fixed within each family and ask how warmup should change as the horizon grows.

Loss models and quadratic theory. Foundational scaling laws relate loss to model size, data, and compute (Kaplan et al., 2020; Hoffmann et al., 2022). Schedule-aware learning-curve models additionally describe accumulated learning rate and annealing effects (Tissue et al., 2025; Luo et al., 2025). Quadratic and least-squares models provide explicit analyses of learning-rate, batch-size, and spectral effects (Jain et al., 2018b;a; Zou et al., 2023; Zhang et al., 2019; Meterez et al., 2026a;b). Power-law spectral assumptions also lead naturally to algebraic learning curves (Caponnetto & De Vito, 2007; Bordelon et al., 2020; Canatar et al., 2021; Bahri et al., 2024; Bordelon et al., 2024; Velikanov & Yarotsky, 2024). These ideas motivate the quadratic analysis and fitted horizon-dependent loss law used in this paper.