THE ART AND PRINCIPLE OF UNIVERSAL SCALING

Anonymous authors

Paper under double-blind review

ABSTRACT

Scaling laws have become the bedrock for foundation model training, yet standard scaling laws are largely restricted to evaluating cross-entropy loss at terminal checkpoints – leaving the continuous dynamics of general observables under-explored. In this work, we establish universal scaling laws across large language models (LLM) and multi-modal models (MMLM) that accurately characterize the any-time-any-scale trajectories of losses, evaluation metrics, and parameter norms. To establish this unified framework, we integrate three key pillars: conventional end-point scaling laws, convex optimization theory, and an enhanced collapse transformation. We systematically validate this framework across diverse model architectures, optimizers, training hyperparameters, compute regimes, and modalities, providing a principled recipe for universal scaling.

1 INTRODUCTION

Scaling laws are central to foundation model training, precisely characterizing the loss by simple functions (Kaplan et al., 2020; Hoffmann et al., 2022) across wide ranges of model size, dataset size, and compute. These laws enable extrapolation from small-scale experiments to expensive large-scale ones, and their applicability is broad: from language modeling to multi-modal training (Tong et al., 2026a; Aghajanyan et al., 2023; Bai et al., 2025), across various datasets, model architectures, and optimization choices.

A conventional scaling law, however, provides only a compressed description of training: it predicts a single quantity, the loss value at a single time, the terminal checkpoint. A broad line of work has therefore sought to enrich scaling laws along two complementary axes. The first broadens the prediction from the final loss alone to the entire training trajectory (Bu et al.; Defazio et al., 2023; Tissue et al., 2026), though so far largely restricted to the loss itself. The second broadens the predicted observables: from pretraining loss to downstream evaluation and emergent capabilities (Achiam et al., 2023; Gadre et al., 2025; Grattafiori et al., 2024; Krajewski et al., 2026).

In this work, we advance both axes jointly: we characterize the full training trajectory for generalized observables, not only loss but also evaluation metrics and parameter norms. We show that end-point scaling laws and scaling collapse (Qiu et al., 2025; Bergsma et al., 2025) are complementary rather than competing. A scaling law determines the absolute loss scale, while collapse gives the relative evolution of loss. Their combination therefore converts an end-point prediction into an absolute prediction of the full training trajectory:

$$\text{scaling law} + \text{collapse} \quad \Longrightarrow \quad \text{universal scaling law.} \tag{1.1}$$

We further show convex optimization can be used as a tractable model of training dynamics, not only characterizing the trajectory in closed form, but also identifying the scaling of hyperparameters for trajectories at different training scales to collapse. Together, these connections link scaling laws, collapse, and convex optimization theory into a unified description of training dynamics.

1.1 SCALING LAWS — ABSOLUTE PREDICTION OF END-POINTS

The standard scaling laws (Kaplan et al., 2020; Henighan et al., 2020; Hoffmann et al., 2022; Achiam et al., 2023; Dey et al., 2023; Grattafiori et al., 2024; Bi et al., 2024) characterize the last-iteration losses: with N denoting the model size, T the total number of training iterations, and B the batch

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.

size (tokens per iteration), the scaling law under fixed token-per-parameter ratio (TPP) has a common form as

$$L_{T,N}(T) - L_* = \alpha \cdot C(N,T)^{\beta} = \alpha \cdot (6BNT)^{\beta} \tag{1.2}$$

Here $L_{T,N}(t)$ measures the intermediate losses after consuming t/T of training budget for a model of size N, $C(N,T) = 6BNT$ is the training compute, L_* is the irreducible loss, and α, β are the coefficients to be determined by multiple small-scale runs.

Scaling laws have been generally observed across data qualities and mixtures (Ye et al., 2025; Shukor et al., 2026), model architectures (Grattafiori et al., 2024; Achiam et al., 2023; Bi et al., 2024; Olmo et al., 2025; Yang et al., 2025; Ludziejewski et al., 2024; Wang et al., 2024; Liang et al., 2026), and optimization choices (Liu et al., 2025; Zhang et al., 2025; Bjorck et al., 2025; Volkova et al., 2026; Wen et al., 2026; Bu, 2026).

Furthermore, scaling laws extend beyond the cross-entropy loss to downstream and discrete evaluation metrics, including the log-scale pass rate (Achiam et al., 2023; Khatri et al., 2026) and the log-scale task accuracy (Gadre et al., 2025; Grattafiori et al., 2024; Krajewski et al., 2026). These extensions significantly enhance the capability of scaling laws, allowing robust and broader prediction of model performance at large scale.

However, conventional scaling laws are limited to the terminal loss at $t = T$ and do not characterize intermediate $L_{T,N}(t)$ at $t < T$ for unseen model scale N. We establish that the scaling-law structure itself extends from terminal checkpoints to the entire training trajectory: for a fixed-TPP scale family,

$$L_{T,N}(xT) \approx L_* + \Phi(x) \cdot \alpha \cdot C(N,T)^{\beta},$$

where $\Phi(x)$ is the universal trajectory to be introduced in the next subsection. Thus, the end-point law is the $x = 1$ slice of our universal scaling law, and every fixed-x slice yields an intermediate-point scaling law with the same irreducible loss L_* and compute exponent β. This shared structure also allows all checkpoints (not only the terminal ones) across scale and training progress to jointly constrain the scaling coefficients.

1.2 COLLAPSE — RELATIVE PREDICTION OF ENTIRE TRAJECTORY

The collapse phenomenon (Qiu et al., 2025; Bergsma et al., 2025) shows that, under fixed TPP $= BT/N$ and appropriate training configurations, the loss curves of different model sizes and training budgets can collapse onto a universal trajectory. To be specific, one can obtain the relative trajectory $\Phi : [0,1] \to (0,\infty)$ by normalizing the training progress and the loss value, which is approximately scale-invariant:

$$\frac{L_{T,N}(xT;\eta^*) - L_*}{L_{T,N}(T;\eta^*) - L_*} \approx \Phi(x; \text{TPP}) \tag{1.3}$$

Here $x := t/T$ is the normalized training progress, $\eta^* = \text{argmin}_\eta L_{T,N}(T;\eta)$ is the optimal learning rate, and $L_{T,N}(t;\eta)$ is the loss obtained after t out of T iterations with:

- the peak learning rate is equal to η and learning rate schedule shape is fixed;
- the weight decay is dependent on learning rate via $\lambda \propto \frac{1}{\eta T}$, i.e. fixing the optimization timescale $\frac{1}{\lambda \eta T}$ as suggested by (Wang & Aitchison, 2025)[1].

However, prior collapse results leave $\Phi(x)$ as an unparameterized empirical curve and lacks a principled recipe for preserving collapse across scales. We address these limitations from the convex-optimization perspective developed in the next subsection, which motivates $1/\sqrt{T}$-scaling of the learning rate and weight decay to preserve collapse, together with an offset T_* and a schedule-dependent parameterization of $\Phi(x)$.

1.3 CONVEX DOMINANCE — AN OPTIMIZATION PRINCIPLE OF ENTIRE TRAJECTORY

Convex dominance (Bu et al.; Li et al., 2026) shows that convex optimization theory by (Defazio et al., 2023) can accurately characterize the training dynamics and loss curves for deep learning (c.f.

[1] Qiu et al. (2025) fixes $\lambda = 0$ as a special case by not using the weight decay.

Figure 2-5 in (Bu et al.), Figure 2 in (Li et al., 2026), and Schaipp et al. (2025)), indicating that deep learning is convex-like although the optimization of neural networks is highly non-convex and non-smooth.

Concretely, for any learning rate sequence and optimizers (e.g. AdamW (Loshchilov & Hutter, 2019) and Muon (Jordan et al., 2024)) following

$$w_{t+1} = (1 - \lambda\eta_t)w_t - \eta_t u_t, \tag{1.4}$$

the any-iteration loss can be characterized by Theorem 2 in Bu et al. as

$$L_{T,N}(t;\eta) - L_N \approx \frac{D_N^2}{2\sum_{\tau=1}^{t}\eta_\tau} + \frac{1}{2}\sum_{\tau=1}^{t-1}\frac{\eta_\tau^2 G_N^2}{\sum_{k=\tau+1}^{t}\eta_k} + \frac{1}{2}\eta_t G_N^2 \tag{1.5}$$

where (L_N, D_N, G_N) are coefficients that depend on model size. Moreover, by optimizing the last-iteration loss in (1.5), the asymptotically optimal convergence can be achieved when

- The peak learning rate is $1/\sqrt{T}$-scaled: $\eta(T) = \frac{\eta_{ref}}{\sqrt{T}}$ where η_{ref} is tuned at small scale.
- The learning rate schedule shape is qualified, which includes cosine, warmup-stable-decay, and linear schedules, but not constant learning rate (c.f. Condition 2.5 of (Bu et al.)).
- The weight decay λ is $1/\sqrt{T}$-scaled: $\lambda(T) = \frac{\lambda_{ref}}{\sqrt{T}}$ where λ_{ref} is tuned at small scale.

We leverage this convex perspective as a principle for collapse: $1/\sqrt{T}$-scaling keeps the optimization timescale $1/(\lambda\eta T)$ invariant and empirically preserves collapse across scales. Moreover, convex analysis yields an explicit parameterized form for $\Phi(x)$ in Equation (2.4), and further motivates a training-progress offset T_*.

1.4 Contributions

Universal scaling of entire loss trajectory. We combine the end-point scaling law and universal trajectory into an any-time, any-scale law, indicating that all intermediate-point scaling laws share the same irreducible loss and compute exponent. This structure enables joint estimation from all checkpoints for robust trajectory forecasting at unseen scales.

Universal scaling beyond loss. We generalize universal scaling to evaluation metrics – such as continuous LAMBADA perplexity and discrete HellaSwag accuracy – and parameter norms at different granularities.

Convexity-informed collapse. We advance collapse through the lens of convex dominance. Specifically, we employ $1/\sqrt{T}$-scaling of hyperparameters to preserve collapse, introduce a new transformation with offset to sharpen the collapse, and derive a closed-form parameterization for the universal trajectory.

Extensive experiments across settings and scales. We validate universal scaling on a variety of optimizers (AdamW and Muon), model architectures (LlaMA3 (Grattafiori et al., 2024), Qwen3-MoE (Yang et al., 2025), Transfusion (Zhou et al., 2025)), TPP notions, and modalities (text and image). We train up to 7B MoE (mixture of experts) and 14B MMLM with trillions of tokens, and evaluate over 20 vision and language metrics.

2 Convexity-informed collapse

We connect the convex optimization to the collapse phenomenon, transferring explicit hyperparameter scaling and loss characterization to the deep learning regime. This connection allows collapse to be more principled: it provides guidance on how to induce collapse, improve the quality of collapse, and explicitly characterize the resulting trajectory.

2.1 SCALED HYPERPARAMETERS PRESERVE COLLAPSE

Prior work establishes collapse under a fixed AdamW timescale (Qiu et al., 2025; Bergsma et al., 2025), and does not uniquely determine how the hyperparameters should scale to preserve such timescale. We extend this phenomenon to general optimizers, including Muon (Jordan et al., 2024; Boreiko et al., 2025), and specifically derive the $1/\sqrt{T}$-scaling for hyperparameters, which attains the optimal asymptotic convergence rate in convex optimization analysis. We show this in Theorem 1, with a proof and an alternative derivation via Robbins-Monro conditions in Appendix B.

Theorem 1. *For convex loss $L_{T,N}$, bounded gradient $\|g_t\| \le G_N$ and bounded parameters $\|w_t\| \le W$, the gradient descent $w_{t+1} = (1 - \lambda\eta_t)w_t - \eta_t g_t$ satisfies*

$$L_{T,N}(T) - L_N = O(1/\sqrt{T})$$

under $\lambda = \lambda_{ref}/\sqrt{T}, \eta_t = s(t/T) \cdot \eta_{ref}/\sqrt{T}$ for any choice of $(\lambda_{ref}, \eta_{ref})$ and any qualified learning rate schedule $s : [0,1] \to [0,1]$ (c.f. condition 2.5 of Bu et al.).

We highlight that $1/\sqrt{T}$-scaling preserves collapse for **any** $(\lambda_{\text{ref}}, \eta_{\text{ref}})$, including the optimal and non-optimal choices. This can be observed by comparing first and second row in Figure 1, where the optimal $\eta_{\text{ref}} = \lambda_{\text{ref}} = 2$.

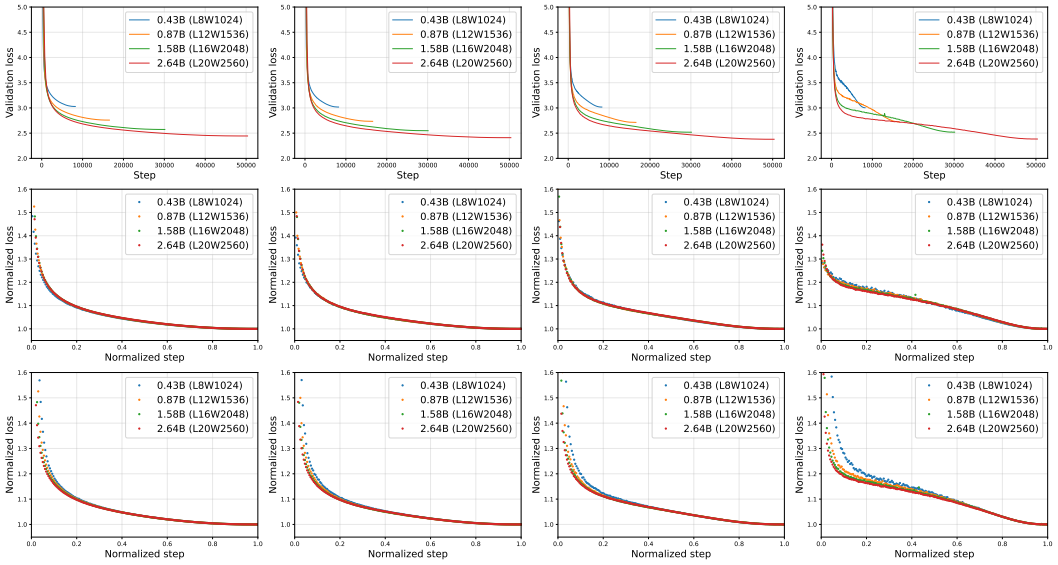


Figure 1 Validation loss for varying model sizes, raw (top row) and normalized (middle row with fitted T_*; bottom row with $T_* = 0$). We train Qwen3-MoE models with Muon-NSGD optimizer under TPP $= 20$. Each column is a different $(\lambda_{\text{ref}}, \eta_{\text{ref}})$ pair: $\eta_{\text{ref}} = \lambda_{\text{ref}} = \{0.5, 1, 2, 5\}$ from left to right.

2.2 TRAINING-PROGRESS OFFSET SHARPENS COLLAPSE

Convex dominance reveals that convex-like behavior emerges only after an initial burn-in period (Bu et al.; Li et al., 2026). This motivates introducing a training-progress offset T_* into the collapse transformation: defining

$$t(x; T_*) = T_* + x(T - T_*), \tag{2.1}$$

we obtain a new transformation that sharpens collapse in Figure 1:

$$\frac{L(t(x); \eta) - L_*}{L(T; \eta) - L_*} := F(x, T_*, L_*) \approx \Phi(x; \text{TPP}). \tag{2.2}$$

To estimate (T_*, L_*) in (2.2), we minimize the discrepancy among the transformed trajectories so that all trajectories are as close as possible to one unknown $\Phi(x)$: denoting $F_i(x; T_*, L_*)$ as the

transformed loss curve of i-th run at a given TPP and $(\eta_{\text{ref}}, \lambda_{\text{ref}})$ values, we apply the Nelder-Mead algorithm (Nelder & Mead, 1965) to solve (2.3):

$$\min_{T_*, L_*} \sum_{i<j} \int_0^1 |F_i(x; T_*, L_*) - F_j(x; T_*, L_*)|^2 \, dx. \tag{2.3}$$

2.3 Closed-form structure of universal trajectory

Collapse establishes that normalized loss trajectories converge to a universal trajectory $\Phi(x)$, but leaves its functional form unspecified[2]. We now leverage the convex loss characterization in (1.5) to unravel the structure of this universal trajectory: as derived in Appendix B.3, we have

$$L_{T,N}(xT; \eta) \approx L_N + \frac{D_N^2}{2\eta T \int_0^x s(u)\, du} + \frac{\eta G_N^2}{2} s(x) \log N,$$

which informs a closed-form parameterization as visualized in Figure 2:

$$\Phi(x) \approx \frac{A + \frac{1}{\int_0^x s(u)\, du + Q} + R\, s(x)}{A + \frac{1}{\int_0^1 s(u)\, du + Q} + R\, s(1)}. \tag{2.4}$$

Here A, R, and Q are fitted coefficients and $s(x)$ is the normalized learning rate schedule, e.g. linear schedule is $s(x) = 1 - x$. Importantly, collapse is refined from an empirical statement that a universal trajectory exists to an explicit characterization of the trajectory, whose shape strongly depends on the shape of learning rate schedule.

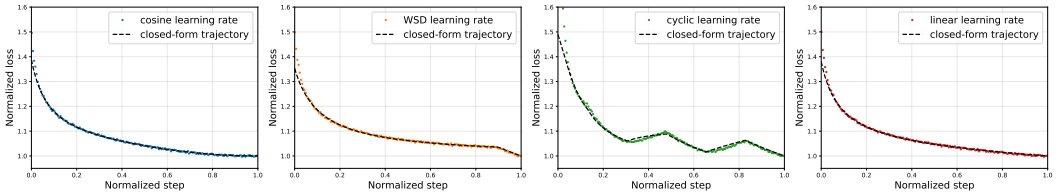


Figure 2 Empirical normalized loss curves (colored dots) compared to analytically derived closed-form trajectories (dashed black lines) for different learning rate schedules.

3 Universal scaling of loss trajectory

We connect scaling laws with the collapse phenomenon, extending scaling from end-point performance to the entire training trajectory. Remarkably, the two are perfectly complementary and indispensible: collapse reveals a universal but normalized trajectory, providing the relative shape; the end-point scaling law predicts the terminal loss, providing the absolute magnitude. As a result, we propose the universal scaling law, to upgrade point-wise scaling to trajectory-wise scaling that enables **any-time, any-scale** loss prediction:

$$L_{T,N}(xT + (1-x)T_*) - L_* = \Phi(x) \cdot \alpha \cdot C(N,T)^{\beta} \tag{3.1}$$

In words, universal scaling law completely factorizes the reducible loss into independent contributions from training progress and scale.

3.1 Scaling law at every intermediate point

Even without knowing the functional form of $\Phi(x)$ nor estimating (T_*, L_*), Equation (3.1) already yields an important consequence. For every fixed $0 < x \leq 1$, $\alpha\Phi(x)$ and L_* are constants with respect to model and data scale. Hence, every intermediate-point scaling law has exactly the same exponent β and floor L_*; only its amplitude changes with x. We visualize these intermediate-point scaling laws in Figure 3. The conventional end-point law is the special case $x = 1$.

[2]Eq. (21) of (Qiu et al., 2025) derives a schedule-transfer relation for the universal trajectory that predicts a target trajectory from a reference trajectory, together with an empirically measured function $h(x) \approx \text{Tr}(\widetilde{\Sigma}(xT))/L(xT)$, where $\widetilde{\Sigma}$ is the preconditioned gradient-noise covariance. Thus, their formulation is not fully explicit or readily applicable in practice.

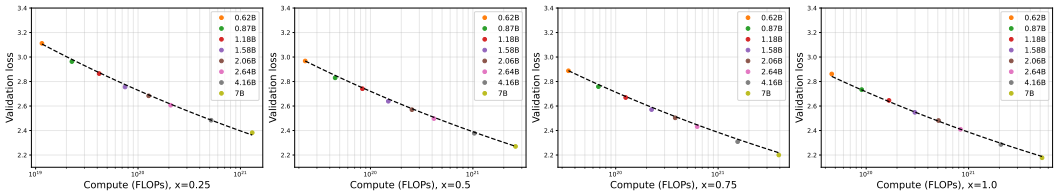


Figure 3 Intermediate-point scaling laws have the common formula $L_{T,N}(xT) \approx L_* + \alpha\Phi(x) \cdot C(N,T)^{\beta}$. We derive $L_* = 0.91, \alpha = 102, \beta = -0.088$. All fits have $R^2 > 0.998$.

3.2 Joint estimation of scaling coefficients

To apply the universal scaling law, we need to estimate $(\alpha, \beta, T_*, L_*)$. In contrast, the end-point scaling law estimates (α, β, L_*) in (1.2), while collapse estimates (T_*, L_*) in (2.3). Notice that L_* is estimated twice, and the two estimates may disagree. To yield more precise estimates than relying on end-points alone, we jointly estimate the coefficients by fitting (3.1) to the full loss trajectories across runs, thus allowing intermediate checkpoints to jointly constrain the coefficients. To be specific (see full derivation in Appendix C), we define an error term for the i-th run as

$$E_i(x; L_*, T_*) := L_{T_i,N_i}(xT_i + (1-x)T_*) - L_*, \qquad C_i := 6BN_iT_i. \tag{3.2}$$

We estimate (β, T_*, L_*) by

$$\min_{T_*, L_*, \beta} \int_0^1 \left[\sum_i E_i(x)^2 - \frac{\left(\sum_i C_i^{\beta} E_i(x)\right)^2}{\sum_i C_i^{2\beta}} \right] dx. \tag{3.3}$$

and then recover

$$\alpha = \frac{\sum_i C_i^{\beta} E_i(1)}{\sum_i C_i^{2\beta}}, \qquad \Phi(x) = \frac{\sum_i C_i^{\beta} E_i(x)}{\sum_i C_i^{\beta} E_i(1)}. \tag{3.4}$$

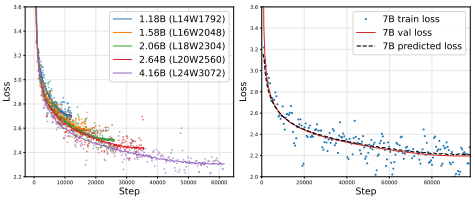


Figure 4 Joint estimation of scaling coefficients on smaller runs can forecast the loss trajectory of larger runs.

In Figure 4, our joint estimation on smaller runs from 1.18B to 4.16B models precisely forecasts the entire loss trajectory of the 7B model. The relative difference between validation loss and predicted loss is $< 0.5\%$ along the trajectory.

4 Enhanced recipe for universal scaling

We now illustrate that universal scaling is robust to modeling and optimization choices. Table 1 summarizes the scope of our validation relative to prior collapse studies. We focus on TPP notions, optimizers, and model architectures, deferring the complete recommended configuration, including all prescribed choices, to Appendix A.

Table 1 Summary of collapse in the literature.

	TPP	modality	model arch	optimizer	hyperparameters	T_*	observable
Qiu et al. (2025)	total	language	dense	Adam	fixed	none	loss
Bergsma et al. (2025)	total	language	dense	AdamW	fixed	none	loss
This work	general	language/vision	general	Muon/AdamW	scaled	fitted	loss/metric/norm

4.1 TPP notions.

To preserve collapse, we require a fixed TPP and explore its notions. Decomposing model parameters into {`embedding`, `body(active)`, `body(stale)`, `head`}, we define $\text{TPP}_{\text{active}}$ on `body(active)`, TPP_{body} on `body(active)` + `body(stale)`, and $\text{TPP}_{\text{total}}$ over the entire parameter set. We note that $\text{TPP}_{\text{active}}$ is proportional to TPP_{body} – a condition that is always true for

dense models and readily preserved in standard MoE scaling suites[3]; in addition, $\text{TPP}_{\text{total}}$ is asymptotically equal to TPP_{body} as parameter count scales and non-body parameter overhead becomes negligible.

Consequently, we can preserve collapse using any of these TPP notions, without strictly adhering to $\text{TPP}_{\text{total}}$ as in prior work (Qiu et al., 2025; Bergsma et al., 2025). In particular, one can naturally adopt TPP_{body} or $\text{TPP}_{\text{active}}$ for MoEs.

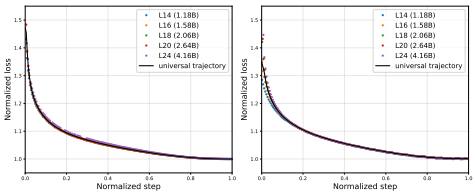


Figure 5 Universal trajectory holds under $\text{TPP}_{\text{total}} = 20$ (left) and $\text{TPP}_{\text{body}} = 20$ (right) in sufficiently scaled regimes.

4.2 Optimizers and hyperparameters.

We observe collapse using various optimizers under various hyperparameters : (1) Muon-NSGD is our default optimizer for LLM training and preserves collapse across sections. (2) AdamW preserves collapse in Qiu et al. (2025); Bergsma et al. (2025) as well as in Section 6.

4.3 Model architectures.

We observe collapse under various model architectures. While our previous experiments are conducted on Qwen3-MoE without weight tying, we consistently observe the same pattern on LLaMA-3 dense models (left two plots, no weight tying) and Qwen3-MoE with **weight tying** ((Press & Wolf, 2017); middle two plots) in Figure 6. Furthermore, we validate the universal scaling on Transfusion architecture in Section 6. In addition, we vary the **aspect ratio** from 32 to 320 in Figure 6 (right two plots), and observe a universal trajectory when the model size ranges from 0.8B to 4B.

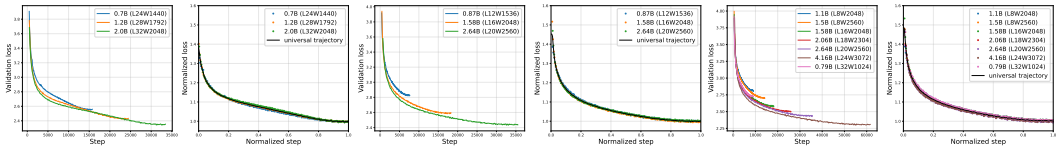


Figure 6 Universal trajectory holds under $\text{TPP}_{\text{body}} = 20$. Left two: LlaMA-3 model, no weight tying, aspect ratio≈ 64. Middle two: Qwen3-MoE model, weight tying, aspect ratio$= 128$. Right two: Qwen3-MoE model, no weight tying, various aspect ratios.

5 Generalized universal scaling

We generalize universal scaling to evaluation metrics and parameter norms, beyond the optimization loss. We use M to denote a metric or norm, and the end-point scaling law as $g(N, T; \theta) := M_{T,N}(T)$. The generalized universal scaling law follows

$$M_{T,N}(xT) = M_{\text{offset}} + \Phi(x)\left[g(N, T; \theta) - M_{\text{offset}}\right]. \tag{5.1}$$

5.1 Universal scaling of evaluation metrics

We analyze a range of continuous and discrete metrics, including but not limited to LAMBADA (Paperno et al., 2016) and HellaSwag (Zellers et al., 2019). We fit g with a power law to LAMBADA perplexity and a sigmoid law to HellaSwag accuracy in Figure 7. For these metrics, the end-point laws are precise and trajectories are universal, allowing the construction of entire metric trajectory across scale. Additionally, we observe the collapse of continuous and discrete metrics on multi-modal training in Section 6.

[3] $\text{TPP}_{\text{body}}/\text{TPP}_{\text{active}}$ ratio is scale-invariant when expert-intermediate/hidden dimension is fixed, number of experts and top-k routing is fixed, and attention parameters scale as width^2. Our setup keeps this ratio ≈ 4.

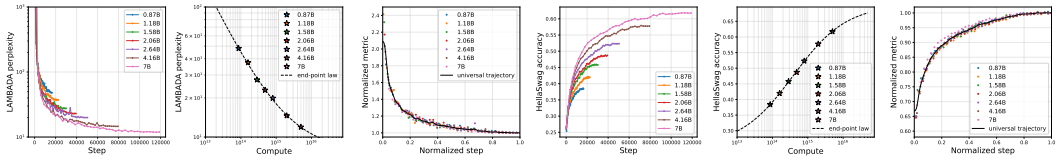


Figure 7 Universal scaling of continuous and discrete evaluation metrics on Qwen3-MoE, under TPP $= 20$. Left three: LAMBADA. Right three: HellaSwag. Each group shows raw metric, end-point scaling law, and generalized collapse. More metrics are presented in Appendix E.

5.2 Universal scaling of parameter norms

For universal scaling of parameter norm, we apply an initialization-aware transformation:

$$\frac{M_{T,N}(xT) - M_{T,N}(0)}{M_{T,N}(T) - M_{T,N}(0)} \approx \Phi(x; \text{TPP}). \tag{5.2}$$

We analyze Frobenius norms $\sqrt{\sum_l \|\mathbf{W}_l\|^2}$ at various levels, agnostic to model depth and width: (1) *Model-level* parameter norm. (2) *Module-level* parameter norms, including the embedding layer, output layer, and hidden layers. (3) *Matrix-level* parameter norms, including Q/K/V matrices, SWiGLU, experts, router, and normalization.

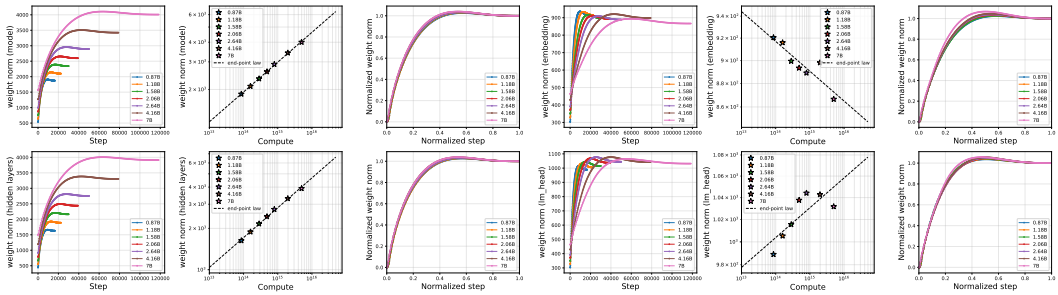


Figure 8 Universal scaling of paramater norms on Qwen3-MoE, under TPP $= 20$. Top-left: whole model norm. Top-right: embedding norm. Bottom-left: hidden-layer norm. Bottom-right: language model head norm. Each group shows raw norm, end-point scaling law, and generalized collapse. More norms are presented in Appendix E.

We fit a power law of parameter norms in Figure 8, which is sufficiently precise for the whole model and hidden layers (attention, MLP, and layer-norm), with an outlier of MoE router (`mlp.gate`). In particular, compared to the model norm in Figure 8, different norms can have significantly different universal trajectories, e.g. `self_attn.q_norm/k_norm`.

6 Universal scaling in native multi-modal training

We further generalize universal scaling to multi-modal training and multi-task optimization, strictly following (Tong et al., 2026a). Details on model architecture, data, and optimization are listed in Appendix F, as well as fine-grained evaluation for individual tasks.

Models. We use the Transfusion framework (Zhou et al., 2025) to natively unify discrete next-token prediction for text and continuous flow matching for visual generation. We adopt LLaMA-3 architecture (Grattafiori et al., 2024) and Representation Autoencoder (RAE; (Zheng et al., 2026; Tong et al., 2026b)) in this framework. We employ both dense networks and modality-specific split feed-forward networks (split-FFNs).

Training. We train models with 1.5B, 2.5B, 3.6B, and 8B active parameters, i.e. up to 14B total parameters for split-FFN. We fix TPP $= 100$ based on active parameters. We primarily use DCLM (Li et al., 2024) as text data, and Shutterstock-Image as text-image paired data for both image-to-text understanding and text-to-image conditional generation.

Optimization. Our optimization is multi-task in a single pass:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{text}} + 3\mathcal{L}_{\text{image}} \tag{6.1}$$

where $\mathcal{L}_{\text{text}}$ is the cross-entropy loss and $\mathcal{L}_{\text{image}}$ is the diffusion flow-matching loss. We optimize using AdamW with $1/\sqrt{T}$ hyperparameters scaling, anchored at learning rate=2e-4 and weight decay=0.1 for 1.5B model.

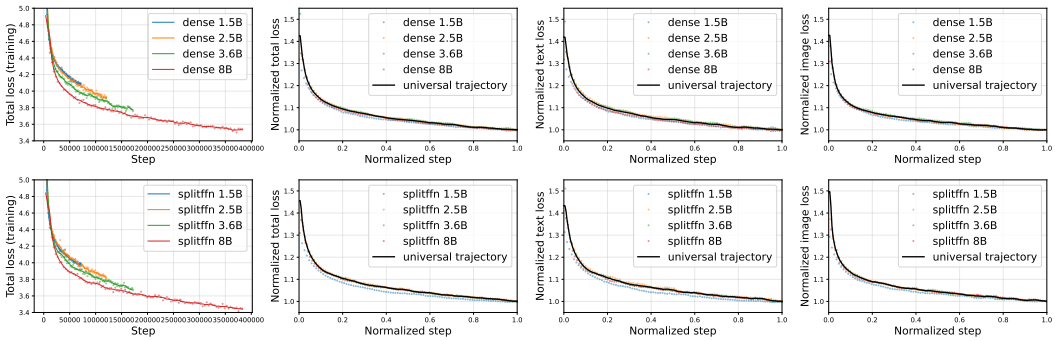


Figure 9 Universal scaling for multi-modal training in terms of losses. The pattern is consistent across loss type (total loss and component losses), scale (1.5-8B), modality (text and image), and architecture (dense and split-FFN).

Universal scaling consistently holds for dense and split-FFN models across losses and evaluation metrics. To be specific, universal scaling is observed not only for the directly optimized loss $\mathcal{L}_{\text{total}}$ but also for component losses $\mathcal{L}_{\text{text}}$ and $\mathcal{L}_{\text{image}}$ in Figure 9. This holds across scale in each sub-plot, and across modality by comparing columns. Furthermore, it is observed in Figure 10 and Figure 15 for multiple types of evaluation metrics: (1) **Discriminative text accuracy** – multiple-choice accuracy on 8 commonsense and knowledge benchmarks. (2) **Generative text score** – F1 on the NQ (Kwiatkowski et al., 2019) and TriviaQA (Joshi et al., 2017). (3) **Continuous text perplexity** on C4 dataset. (4) **Generative image score**, such as DPG (Hu et al., 2024), GenEval (Ghosh et al., 2023), and FID (Heusel et al., 2017).

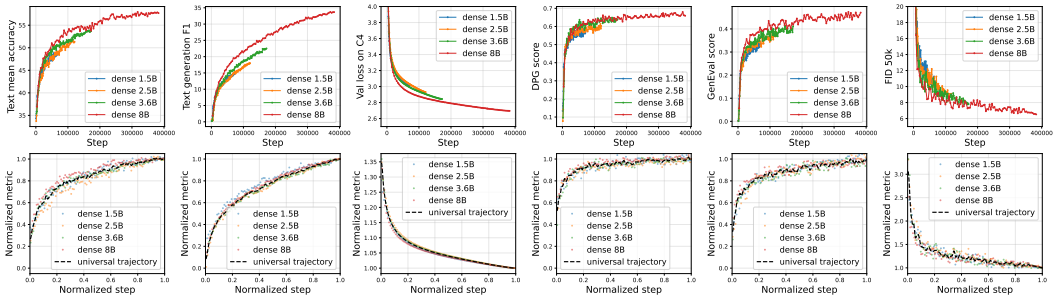


Figure 10 Universal scaling for multi-modal training in terms of metrics. The pattern is consistent across metric type, scale (1.5-8B), modality (text and image), and architecture (dense and split-FFN; the latter is visualized in Figure 15).

7 Conclusion

In this work, we study universal scaling laws across LLM and MMLM to enable the precise prediction of observable trajectories. To establish our unified framework, we advance the collapse transformation with a closed-form parameterization of the universal trajectory, and leverage convex optimization theory to consolidate the hyperparameter choices. We demonstrate that universal scaling extends the standard end-point scaling to encompass full dynamics across loss values, parameter norms, and evaluation metrics, which is validated through a series of ablations of model architectures, optimizers, and training configurations. Together, these contributions yield a practical recipe towards universal scaling.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for [assisting in the writing of proofs, designing or providing feedback on research methodology or experiments, implementing methods, interpreting results]. We have not used generative AI tools for [generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, proposing or refining hypotheses, cleaning and reformatting dataset, supporting qualitative and thematic data analysis], and [generating synthetic data sets, assisting with translation] are not applicable to this work. Additionally, we used generative AI tools for [creating or modifying scientific figures or images, drafting parts of a research paper, summarizing or analysing existing literature, editing a research paper to improve readability, identifying relevant literature, suggesting a structure for a research paper]. We have reviewed all AI-assisted work. LLM-generated content was verified and tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 2

Armen Aghajanyan, Lili Yu, Alexis Conneau, Wei-Ning Hsu, Karen Hambardzumyan, Susan Zhang, Stephen Roller, Naman Goyal, Omer Levy, and Luke Zettlemoyer. Scaling laws for generative mixed-modal language models. In *International Conference on Machine Learning*, pp. 265–279. PMLR, 2023. 1

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 1

Shane Bergsma, Bin Claire Zhang, Nolan Dey, Shaheer Muhammad, Gurpreet Gosal, and Joel Hestness. Scaling with collapse: Efficient and predictable training of llm families. *arXiv preprint arXiv:2509.25087*, 2025. 1, 2, 4, 6, 7

Akshita Bhagia, Jiacheng Liu, Alexander Wettig, David Heineman, Oyvind Tafjord, Ananya Harsh Jha, Luca Soldaini, Noah A Smith, Dirk Groeneveld, Pang Wei Koh, et al. Establishing task scaling laws via compute-efficient model ladders. *arXiv preprint arXiv:2412.04403*, 2024. 18

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024. 1, 2

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 7432–7439, 2020. 18, 22

Johan Bjorck, Alon Benhaim, Vishrav Chaudhary, Furu Wei, and Xia Song. Scaling optimal lr across token horizons. In *International Conference on Learning Representations*, volume 2025, pp. 83640–83657, 2025. 2

Valentyn Boreiko, Zhiqi Bu, and Sheng Zha. Towards understanding of orthogonalization in muon. In *Tiny Titans: The next wave of On-Device Learning for Foundational Models (TTODLer-FM)*, 2025. 4, 18

Zhiqi Bu. Scaling depth capacity via zero/one-layer model expansion. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=WP5q6Hn3Cv. 2

Zhiqi Bu, Shiyun Xu, and Jialin Mao. Convex dominance in deep learning i: A scaling law of loss and learning rate. In *The Fourteenth International Conference on Learning Representations*. 1, 2, 3, 4

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. 18, 22

Aaron Defazio, Ashok Cutkosky, Harsh Mehta, and Konstantin Mishchenko. Optimal linear decay learning rate schedules and further refinements. *arXiv preprint arXiv:2310.07831*, 2023. 1, 2

Nolan Dey, Gurpreet Gosal, Hemant Khachane, William Marshall, Ribhu Pathria, Marvin Tom, Joel Hestness, et al. Cerebras-gpt: Open compute-optimal language models trained on the cerebras wafer-scale cluster. *arXiv preprint arXiv:2304.03208*, 2023. 1

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Licheng Zhang, Xiaorui Wang, and Zhendong Mao. Deepresearch bench: A comprehensive benchmark for deep research agents. In *International Conference on Learning Representations*, volume 2026, pp. 42414–42448, 2026. 18

Samir Yitzhak Gadre, Georgios Smyrnis, Vaishaal Shankar, Suchin Gururangan, Mitchell Wortsman, Rulin Shao, Jean Mercat, Alex Fang, Jeffrey Li, Sedrick Keh, et al. Language models scale reliably with over-training and on downstream tasks. In *International Conference on Learning Representations*, volume 2025, pp. 67661–67682, 2025. 1, 2

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023. 9

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 1, 2, 3, 8, 18

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020. 1

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 9

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022. 1

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024. 9

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL `https://kellerjordan.github.io/posts/muon/`. 3, 4

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017. 9

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. 1

Devvrit Khatri, Lovish Madaan, Rishabh Tiwari, Rachit Bansal, Venkata Sai Surya Subramanyam Duvvuri, Manzil Zaheer, Inderjit Dhillon, David Brandfonbrener, and Rishabh Agarwal. The art of scaling reinforcement learning compute for llms. In *International Conference on Learning Representations*, volume 2026, pp. 72438–72467, 2026. 2, 18

Jakub Krajewski, Amitis Shidani, Dan Busbridge, Sam Wiseman, and Jason Ramapuram. Revisiting the scaling properties of downstream metrics in large language model training. In *International Conference on Learning Representations*, volume 2026, pp. 123493–123533, 2026. 1, 2

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019. 9, 18

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, et al. Datacomp-lm: In search of the next generation of training sets for language models. *Advances in Neural Information Processing Systems*, 37:14200–14282, 2024. 8

Jiaxiang Li, Zhiqi Bu, and Shiyun Xu. Towards joint scaling laws with optimal batch size schedules. *arXiv preprint arXiv:2607.27731*, 2026. 2, 3, 4, 15

Zhengyang Liang, Hao He, Ceyuan Yang, and Bo Dai. Scaling laws for diffusion transformers. In *International Conference on Learning Representations*, volume 2026, pp. 84488–84521, 2026. 2

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for llm training. *arXiv preprint arXiv:2502.16982*, 2025. 2

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7. 3

Jan Ludziejewski, Jakub Krajewski, Kamil Adamczewski, Maciej Pióro, Michal Krutul, Szymon Antoniak, Kamil Ciebiera, Krystian Król, Tomasz Odrzygózdz, Piotr Sankowski, et al. Scaling laws for fine-grained mixture of experts. In *ICML*, pp. 33270–33288, 2024. 2

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016. 18

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2381–2391, 2018. 18, 22

J. A. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, 7 (4):308–313, 1965. doi: 10.1093/comjnl/7.4.308. 5

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, et al. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025. 2

David Owen. How predictable is language model benchmark performance? *arXiv preprint arXiv:2401.04757*, 2024. 18

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc-Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The lambada dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 1525–1534, 2016. 7

Ofir Press and Lior Wolf. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 157–163, 2017. 7

Shikai Qiu, Lechao Xiao, Andrew Gordon Wilson, Jeffrey Pennington, and Atish Agarwala. Scaling collapse reveals universal dynamics in compute-optimally trained neural networks. In *International Conference on Machine Learning*, pp. 50697–50720. PMLR, 2025. 1, 2, 4, 5, 6, 7

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI spring symposium: logical formalizations of commonsense reasoning*, pp. 90–95, 2011. 22

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021. 22

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 4463–4473, 2019. 22

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*, pp. 53267–53294. PMLR, 2025. 3

Mustafa Shukor, Louis Bethune, Dan Busbridge, David Grangier, Enrico Fini, Alaaeldin El-Nouby, and Pierre Ablin. Scaling laws for optimal data mixtures. *Advances in Neural Information Processing Systems*, 38:129554–129579, 2026. 2

Howe Tissue, Venus Wang, and Lu Wang. Scaling law with learning rate annealing. *Advances in Neural Information Processing Systems*, 38:91017–91053, 2026. 1

Shengbang Tong, David Fan, John Nguyen, Ellis Brown, Gaoyue Zhou, Shengyi Qian, Boyang Zheng, Théophane Vallaeys, Junlin Han, Rob Fergus, et al. Beyond language modeling: An exploration of multimodal pretraining. *arXiv preprint arXiv:2603.03276*, 2026a. 1, 8

Shengbang Tong, Boyang Zheng, Ziteng Wang, Bingda Tang, Nanye Ma, Ellis Brown, Jihan Yang, Rob Fergus, Yann LeCun, and Saining Xie. Scaling text-to-image diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2601.16208*, 2026b. 8

Alexandra Volkova, Mher Safaryan, Christoph H Lampert, and Dan Alistarh. Towards robust scaling laws for optimizers. *arXiv preprint arXiv:2602.07712*, 2026. 2

Siqi Wang, Zhengyu Chen, Bei Li, Keqing He, Min Zhang, and Jingang Wang. Scaling laws across model architectures: A comparative analysis of dense and moe models in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5583–5595, 2024. 2

Xi Wang and Laurence Aitchison. How to set adamw's weight decay as you scale model and dataset size. In *International Conference on Machine Learning*, pp. 62222–62250. PMLR, 2025. 2

Johannes Welbl, Nelson F Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106, 2017. 18

Kaiyue Wen, David Hall, Tengyu Ma, and Percy Liang. Fantastic pretraining optimizers and where to find them. In *International Conference on Learning Representations*, volume 2026, pp. 144731–144838, 2026. 2

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 2, 3

Jiasheng Ye, Peiju Liu, Tianxiang Sun, Jun Zhan, Yunhua Zhou, and Xipeng Qiu. Data mixing laws: Optimizing data mixtures by predicting language modeling performance. In *International Conference on Learning Representations*, volume 2025, pp. 82263–82287, 2025. 2

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4791–4800, 2019. 7, 22

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*, volume 2025, pp. 66756–66782, 2025. 2

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *International Conference on Learning Representations*, volume 2026, pp. 35791–35820, 2026. 8

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *International Conference on Learning Representations*, volume 2025, pp. 6446–6469, 2025. 3, 8

A FULL RECIPE OF UNIVERSAL SCALING

We summarize the recipe for multiple runs to follow the universal scaling.

1. TPP is fixed. The notion of TPP can be $\text{TPP}_{\text{active}}$, TPP_{body}, or $\text{TPP}_{\text{total}}$.
2. Choice of model family is flexible, e.g. dense model and sparse MoE, or LLM and MMLM.
3. Model size or TPP is sufficiently large, e.g. model $\geq 2B$ parameters for TPP $= 20$, or model $\geq 1B$ parameters for TPP $= 100$.
4. Learning rate schedule shape is fixed, e.g. cosine/WSD/linear schedule that decays to a fixed fraction of peak learning rate. We can use fixed-ratio warmup or fixed-step warmup, because the learning rate schedule shape is asymptotically identical.
5. The optimizer is fixed with decoupled weight decay or no weight decay.
6. $1/\sqrt{T}$-scaling on learning rate and weight decay.
7. We recommend that the aspect ratio be fixed, though universal scaling holds when it varies moderately.
8. We recommend that, for MoE, the number of experts and top-k routing be fixed, which is necessary to lock the ratio $\text{TPP}_{\text{body}}/\text{TPP}_{\text{active}}$. Consequently, sweeping models at a fixed $\text{TPP}_{\text{active}} = 80$ is mathematically equivalent to sweeping them at a fixed $\text{TPP}_{\text{body}} = 20$, if $\text{TPP}_{\text{body}}/\text{TPP}_{\text{active}} = 4$.
9. Weight tying can be used or not.

B THEORETICAL ANALYSIS

B.1 PROOF OF THEOREM 1

Denote $L^{\lambda}_{T,N}(\mathbf{w}) = L_{T,N}(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w}\|^2$, then weight-decayed SGD on $L_{T,N}$ is equivalent to vanilla SGD on $L^{\lambda}_{T,N}$. It is clear that $\|\nabla L^{\lambda}_{T,N}\| \leq \|\nabla L_{T,N}\| + \lambda\|\mathbf{w}\| \leq G_N + \lambda W$. Via Theorem 1 in Li et al. (2026) that holds for vanilla SGD:

$$\mathbb{E}L^{\lambda}_{T,N}(\mathbf{w}_T) - L^{\lambda}_N \leq \frac{D_N^2}{2\sum_{t=1}^{T}\eta_t} + \frac{1}{2}\sum_{t=1}^{T-1}\frac{\eta_t^2\|\nabla L^{\lambda}_{T,N}\|^2}{\sum_{k=t+1}^{T}\eta_k},$$

as long as the learning rate decays to zero for any fixed T.

Using the bounded gradient, we obtain

$$\mathbb{E}L^{\lambda}_{T,N}(\mathbf{w}_T) - L^{\lambda}_N \leq \frac{D_N^2}{2\sum_{t=1}^{T}\eta_t} + \frac{1}{2}\sum_{t=1}^{T-1}\frac{\eta_t^2(G_N + \lambda W)^2}{\sum_{k=t+1}^{T}\eta_k}$$

For $1/\sqrt{T}$-scaled peak learning rate, we get the following upper bound:

$$\mathbb{E}L^{\lambda}_{T,N}(\mathbf{w}_T) - L^{\lambda}_N = O(D_N G_N/\sqrt{T})$$

To translate the upper bound from $L^{\lambda}_{T,N}$ to $L_{T,N}$, we see that, with $\mathbf{w}_*$ minimizing $L_{T,N}$ not $L^{\lambda}_{T,N}$,

$$L^{\lambda}_N - L_N \leq \frac{\lambda}{2}\|\mathbf{w}_*\|^2 \leq \frac{\lambda}{2}W^2$$

All in all, we need $\lambda = O(1/\sqrt{T})$ so that

$$\mathbb{E}L^{\lambda}_{T,N}(\mathbf{w}_T) - L_N = O(1/\sqrt{T})$$

B.2 Robbins–Monro conditions

Alternatively, we view gradient as signal and $\lambda \mathbf{w}_t$ as noise. We require

$$\sum_{t=1}^{T} \eta_t = \Theta(\sqrt{T}), \qquad \sum_{t=1}^{T} \eta_t^2 = O(1), \qquad \sum_{t=1}^{T} \eta_t \lambda \|\mathbf{w}_t\| = O(1),$$

which are satisfied by $1/\sqrt{T}$-scaling. Therefore, under the standard assumptions for stochastic convex optimization, the weight-decay perturbation is asymptotically negligible and the optimization error converges to zero.

B.3 Closed-form structure of Φ

We start from Eq. (1.5):

$$L_{T,N}(t;\eta) - L_N \approx \frac{D_N^2}{2\sum_{\tau=1}^{t} \eta_\tau} + \frac{G_N^2}{2} \sum_{\tau=1}^{t-1} \frac{\eta_\tau^2}{\sum_{k=\tau+1}^{t} \eta_k} + \frac{G_N^2}{2} \eta_t.$$

Let

$$\eta_\tau = \eta\, s(\tau/T), \qquad t = \lfloor xT \rfloor, \qquad x \in (0,1],$$

where s denotes the normalized learning rate schedule.

For the first term, Riemann-sum approximation gives

$$\sum_{\tau=1}^{xT} \eta_\tau = \eta \sum_{\tau=1}^{xT} s(\tau/T) = \eta T \int_0^x s(u)\, du + o(\eta T),$$

and hence

$$\frac{D_N^2}{2\sum_{\tau=1}^{xT} \eta_\tau} = \frac{D_N^2}{2\eta T \int_0^x s(u)\, du} + o\!\left(\frac{1}{\eta T}\right).$$

For the second term,

$$\sum_{\tau=1}^{xT-1} \frac{\eta_\tau^2}{\sum_{k=\tau+1}^{xT} \eta_k} = \eta \sum_{\tau=1}^{xT-1} \frac{s^2(\tau/T)}{\sum_{k=\tau+1}^{xT} s(k/T)}.$$

Near $\tau = xT$, the summand behaves as $s(x)/(xT - \tau)$, producing a harmonic sum. Therefore, under mild regularity conditions on s,

$$\sum_{\tau=1}^{xT-1} \frac{s^2(\tau/T)}{\sum_{k=\tau+1}^{xT} s(k/T)} = s(x) \log T + C_s(x) + o(1),$$

where $C_s(x)$ is a finite schedule-dependent correction. In particular, one may write

$$C_s(x) = s(x)\big(\log x + 0.577215\big) + \int_0^x \left[\frac{s^2(u)}{\int_u^x s(v)\, dv} - \frac{s(x)}{x-u} \right] du,$$

with 0.577215 denoting the Euler–Mascheroni constant. Thus,

$$\frac{G_N^2}{2} \sum_{\tau=1}^{xT-1} \frac{\eta_\tau^2}{\sum_{k=\tau+1}^{xT} \eta_k} = \frac{\eta G_N^2}{2} \left[s(x) \log T + C_s(x) + o(1) \right].$$

Finally,

$$\frac{G_N^2}{2} \eta_{xT} = \frac{\eta G_N^2}{2} s(x) + o(\eta).$$

Combining the three terms yields

$$L_{T,N}(xT;\eta) - L_N \approx \frac{D_N^2}{2\eta T \int_0^x s(u)\, du} + \frac{\eta G_N^2}{2} \left[s(x) \log T + C_s(x) + s(x) \right],$$

Retaining only the leading dependence on T,

$$L_{T,N}(xT;\eta) \approx L_N + \frac{D_N^2}{2\eta T \int_0^x s(u)\,du} + \frac{\eta G_N^2}{2} s(x)\log T + O(\eta). \tag{B.1}$$

Under fixed TPP $= \frac{T}{N}$, we have

$$\log T = \log N + \log \text{TPP}.$$

The last term is independent of scale and can be absorbed into the $O(\eta)$ schedule-dependent correction. Therefore,

$$L_{T,N}(xT;\eta) \approx L_N + \frac{D_N^2}{2\eta T \int_0^x s(u)\,du} + \frac{\eta G_N^2}{2} s(x)\log N$$

C Coefficient optimization

Objective. For the i-th run (trained to T_i steps with N_i parameters, so total compute $C_i := 6N_iT_i$), we model

$$L(x) \;=\; L_* \;+\; \Phi(x)\,\alpha\,(6NT)^{\beta}, \qquad x = \frac{s - T_*}{T - T_*} \in [0,1]. \tag{C.1}$$

The scalars $(L_*,\, T_*,\, \beta,\, \alpha)$ and the universal normalized trajectory Φ are common to all runs; only C_i varies. At the end-point $x = 1$ this reduces to the usual scaling law $L = L_* + \alpha C^{\beta}$, while the interior points pin down Φ.

We resample each run onto the shared grid $x \in [0,1]$ and define an error term

$$E_i(x; L_*, T_*) := L_{T_i,N_i}(xT_i + (1-x)T_*) - L_*, \qquad C_i := 6N_iT_i. \tag{C.2}$$

Under the model (C.1), $E_i(x) \approx \alpha\, C_i^{\beta}\, \Phi(x)$, our primary objective is

$$\min_{L_*,T_*,\beta,\alpha,\Phi(\cdot)} \int_0^1 \sum_i \Big(E_i(x) - \alpha\, C_i^{\beta}\, \Phi(x)\Big)^2 dx, \qquad \text{s.t. } \Phi(1) = 1. \tag{C.3}$$

Profiling out α and $\Phi(x)$. The primary objective is of 4 scalars and 1 function, thus difficult to solve due to functional optimization. We can profile out α and $\Phi(x)$.

Write $\Psi(x) := \alpha\,\Phi(x)$ for the un-normalized shape. The key observation is that at any fixed (L_*, T_*, β), the objective (C.3) is a *separable* least-squares problem: it decouples across x and is quadratic in $\Psi(x)$. For each x we minimize $\sum_i \big(E_i(x) - C_i^{\beta}\Psi(x)\big)^2$ over the scalar $\Psi(x)$, whose closed-form solution is the compute-weighted regression coefficient

$$\Psi^*(x) = \frac{\sum_i C_i^{\beta} E_i(x)}{\sum_i C_i^{2\beta}}. \tag{C.4}$$

Substituting Ψ^* back eliminates both α and the entire function Φ, leaving a *profiled* objective in only 3 scalars:

$$\min_{T_*,L_*,\beta} \int_0^1 \left[\sum_i E_i(x)^2 - \frac{\Big(\sum_i C_i^{\beta} E_i(x)\Big)^2}{\sum_i C_i^{2\beta}}\right] dx. \tag{C.5}$$

Minimization and recovery. We minimize (C.5) over the three scalars (T_*, L_*, β) directly which is fast and low-dimensional. Next, given (T_*^*, L_*^*, β^*), we recover the amplitude and shape from Ψ^* using the normalization $\Phi(1) = 1$:

$$\alpha = \Psi^*(1) = \frac{\sum_i C_i^{\beta} E_i(1)}{\sum_i C_i^{2\beta}}, \qquad \Phi(x) = \frac{\Psi^*(x)}{\alpha} = \frac{\sum_i C_i^{\beta} E_i(x)}{\sum_i C_i^{\beta} E_i(1)}. \tag{C.6}$$

D EXPERIMENTS

Model. Our experiments primarily use Qwen3-MoE architecture (Qwen3-30B-A3B) with fixed aspect ratio=128. The intermediate size is $3\times$ hidden size, and MoE intermediate size is $1/4\times$ hidden size. At every scale, we use 16 experts and top-2 routing. We use "L8W1024" to denote 8-layer 1024-hidden model.

Notably, our $\text{TPP}_{\text{body}}/\text{TPP}_{\text{active}} \approx 4$ at all scales.

Data and Optimization. We use Fineweb-Edu dataset to train models under Fully Sharded Data Parallel (FSDP-2). We use a global batch size 512, and sequence length 4096, thus 2M tokens/step. The optimizated objective is cross-entropy plus 0.001 auxiliary loss, while the validation loss is purely cross-entropy. We employ Muon-NSGD optimizer similar to Boreiko et al. (2025) but routing all 2D tensors (including embedding) to Muon component and all 1D tensors to NSGD component:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \text{Ortho}(\mathbf{g}_t)$$

where Ortho is orthogonalization with Newton-Schulz iteration for 2D tensors, and normalization $\text{Ortho}(\mathbf{g}_t) = \frac{\mathbf{g}_t}{\|\mathbf{g}_t\|_F}$ for 1D tensors. We use gradient clipping at 1.0, and momentum at 0.95.

Our default learning rate is cosine decay to zero with $\eta_{\text{ref}} = \lambda_{\text{ref}} = 1$, in order to keep hyperparameters $1/\sqrt{T}$-scaled. Learning rate warmup is $\min(500, 0.02T)$.

E GENERALIZED UNIVERSAL SCALING

E.1 EVALUATION METRICS

In Figure 7 and Figure 11, we establish the universal scaling on LAMBADA, HellaSwag, WikiText (Merity et al., 2016), ARC (Clark et al., 2018), OpenBookQA (Mihaylov et al., 2018), PIQA (Bisk et al., 2020), RACE (Du et al., 2026), SCIQ (Welbl et al., 2017), and NQ (Kwiatkowski et al., 2019).

For benchmark measured by discrete accuracy, we fit a sigmoid law to capture the S-shaped scaling,

$$g(N, T) = 0.25 + (E - 0.25) \cdot \text{sigmoid}\left(a + b \ln(6NT)\right). \tag{E.1}$$

We note that sigmoid laws relating training scale to benchmark accuracy have been observed both in pretraining (Owen, 2024; Grattafiori et al., 2024; Bhagia et al., 2024) and in reinforcement learning (Khatri et al., 2026).

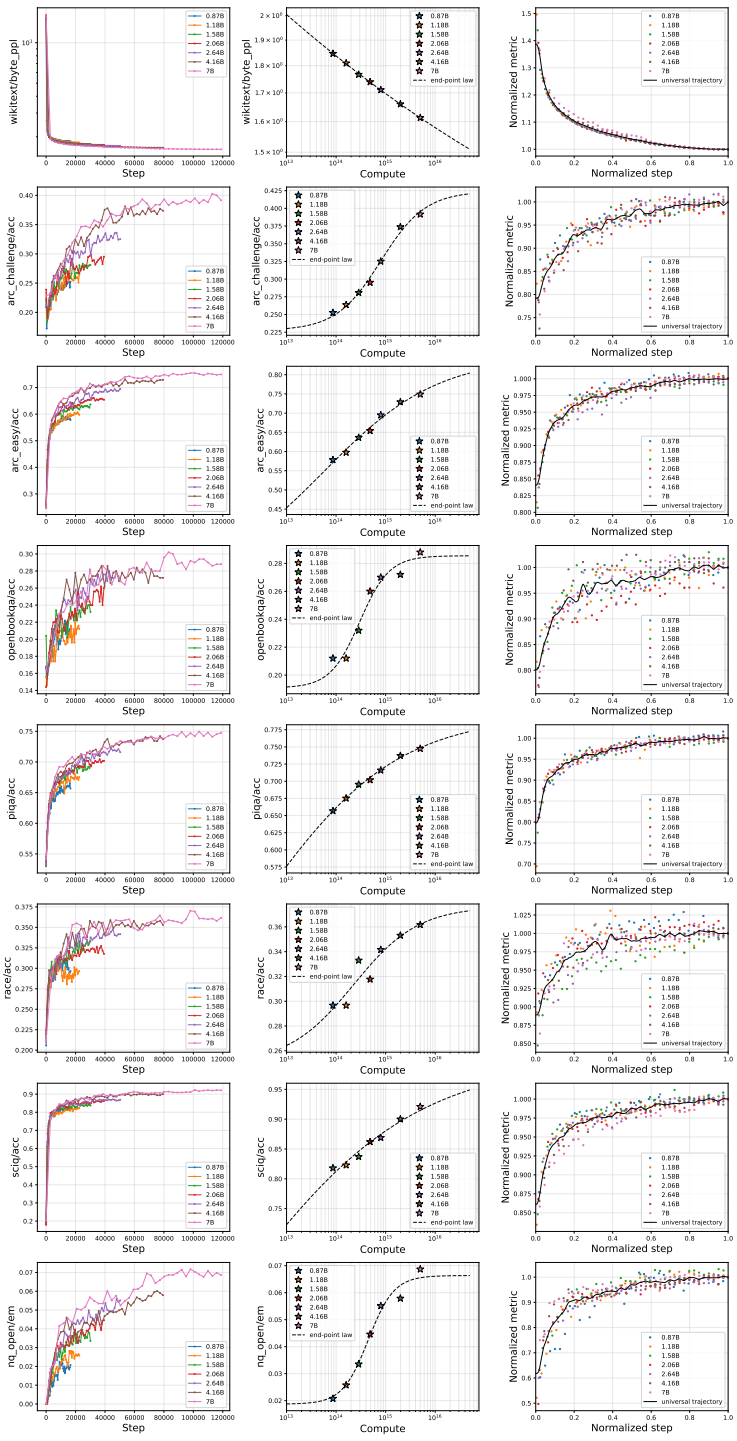


Figure 11 Universal scaling of continuous and discrete evaluation metrics on Qwen3-MoE, under TPP $= 20$. The tasks include WikiText, ARC Easy/Challenge, OpenBookQA, PIQA, RACE, SCIQ, and NQ. Each row shows raw metric, end-point scaling law, and generalized supercollapse.

E.2 Parameter norms

In Figure 8 and Figure 12, we establish the universal scaling on different levels of parameter groups.

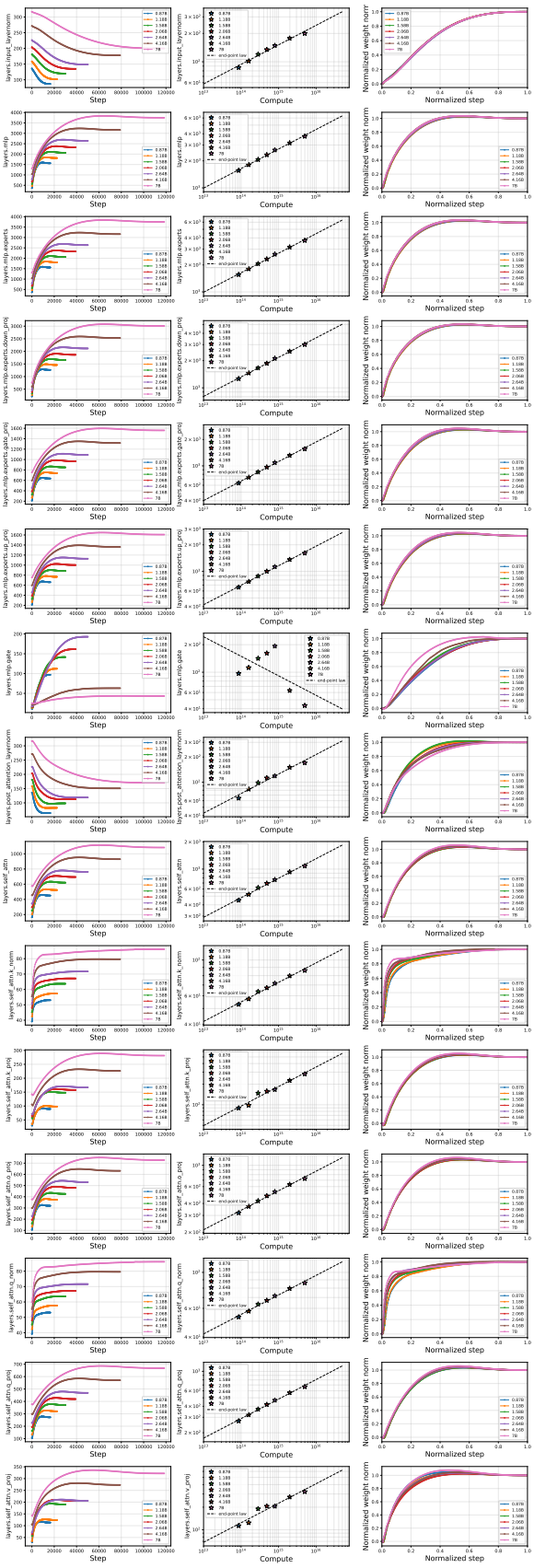


Figure 12 Universal scaling of parameter norms on Qwen3 MoE, under TPP $= 20$. Each group shows raw norm, end-point scaling law, and generalized supercollapse.

F MULTI-MODAL TRAINING

F.1 MODELS.

We adopt the Transfusion framework to natively unify discrete next-token prediction for text and continuous flow matching for visual generation within a single autoregressive sequence. We use two variants: dense model in Figure 13 and split-FFN in Figure 14.

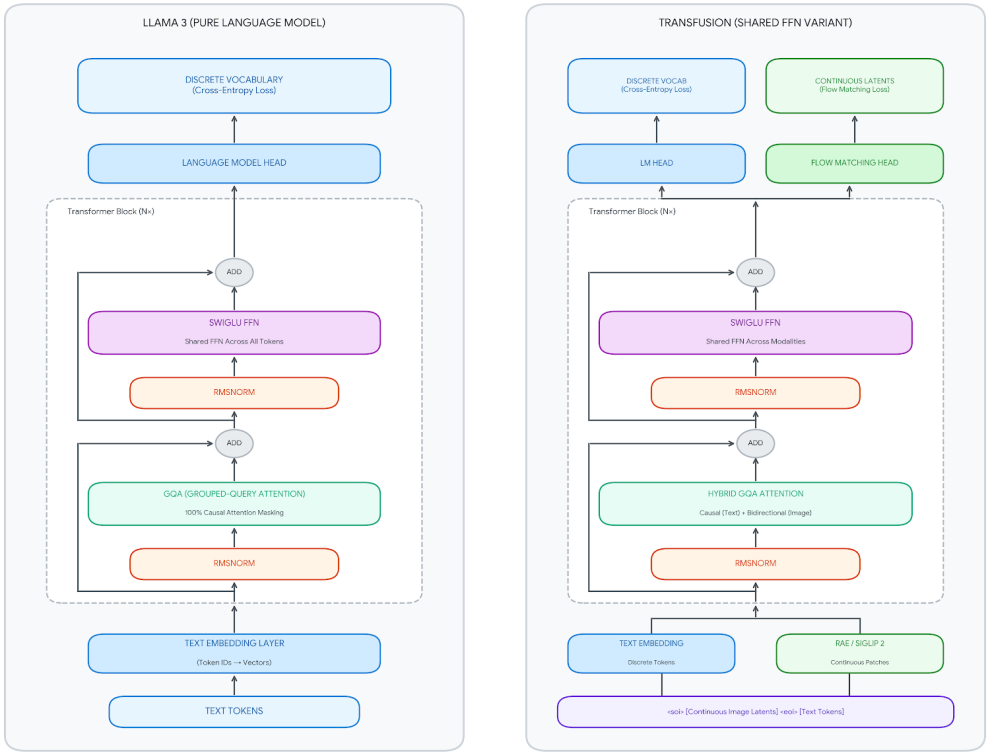


Figure 13 Dense MMLM (right) compared to LLaMA-3 (left).

Model configurations are listed in Table 2.

Table 2 Model configurations across different scales for split-FFN.

Active Params	Layers	Hidden Dim	Query Heads	KV Heads	Head Dim Params	Total Params
1.5B	16	2048	32	8	64	2.3B
2.5B	27	2560	20	5	128	4.0B
3.6B	28	3072	24	8	128	5.7B
8B	32	4096	32	8	128	14B

All configurations incorporate SwiGLU activation, Rotary Position Embeddings (RoPE) with a base frequency $\theta = 500,000$, pre-RMSNorm, grouped-query attention (GQA), QK-norm, and FlashAttention.

F.2 OPTIMIZATION.

Training is optimized using AdamW with $\beta_1 = 0.9, \beta_2 = 0.95$, and gradient clipping at 1.0. All models are trained with a context length of 4096 tokens in `bfloat16` precision using FSDP-2. The global batch size is 512 sequences $\times$ 4096 tokens = 2M tokens/step. We adopt the x-prediction

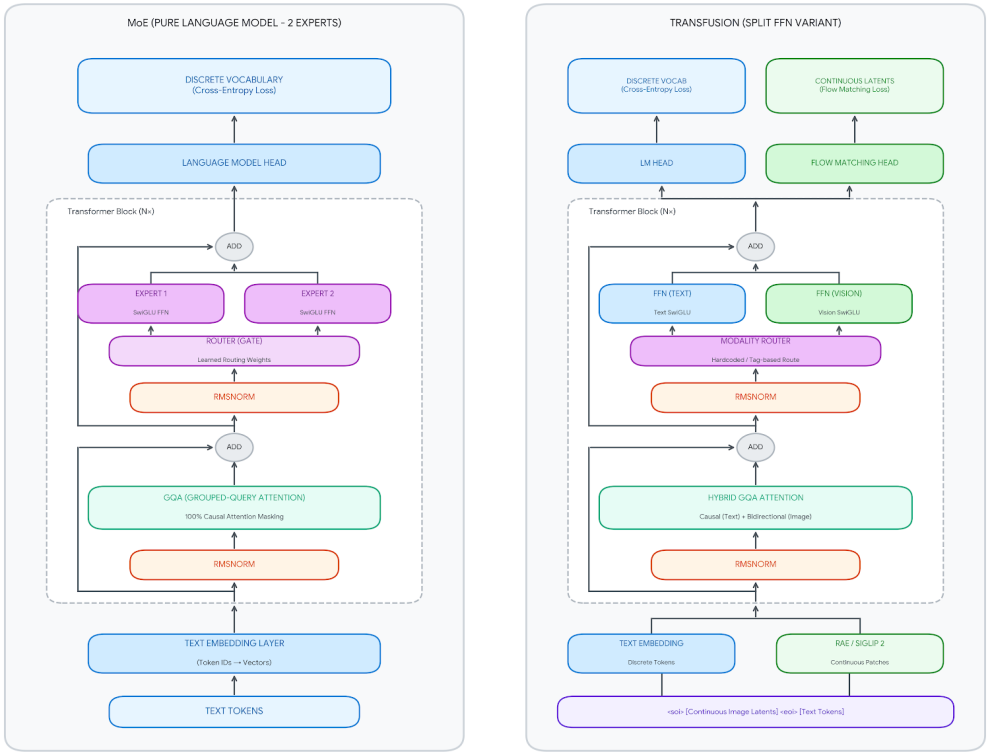


Figure 14 Split-FFN MMLM (right) compared to MoE (left).

formulation where the network outputs the clean sample x_0, which is converted on-the-fly to a velocity $v = (x_0 - x_t)/(1 - t)$ for each rectified flow matching step.

F.3 Tokenization.

- **Text Tokenizer:** We use the LLaMA-3 BPE tokenizer with a vocabulary size of approximately 128,000, augmented with a small set of special multimodal control tokens (e.g., <BOI>, <EOI>).
- **Visual Representation (RAE):** We employ a frozen SigLIP-2 ViT-400m/14 encoder to process 224×224 images into a 16×16 grid of 256 semantic tokens. Visual generation is performed via flow matching within the SigLIP latent space and decoded back to pixels by Representation Autoencoder (RAE).

F.4 Metrics.

We visualize the universal scaling of split-FFN in Figure 15, which is consistent to that of dense models in Figure 10.

Specifically, the text mean accuracy is averaged over 8 benchmarks: ARC-Easy/Challenge (Clark et al., 2018), COPA (Roemmele et al., 2011), HellaSwag (Zellers et al., 2019), OpenBookQA (Mihaylov et al., 2018), PIQA (Bisk et al., 2020), SIQA (Sap et al., 2019), and WinoGrande (Sakaguchi et al., 2021). We observe universal scaling in individual benchmark for dense MMLM in Figure 16, and for split-FFN MMLM in Figure 17, with a few outliers such as split-FFN 1.5B on ARC-Easy.

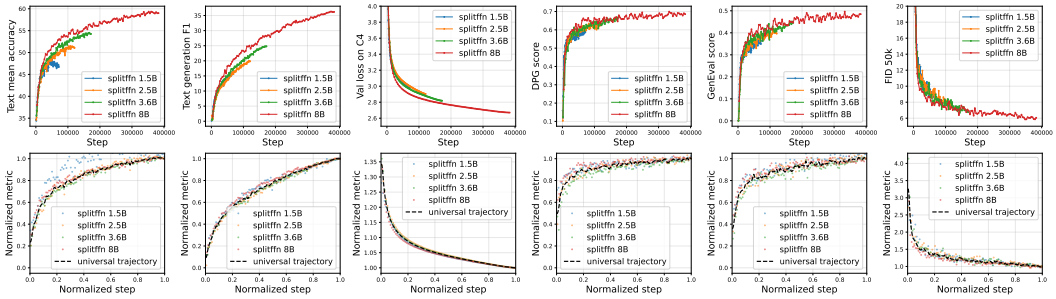


Figure 15 Universal scaling for multi-modal training on split-FFN. The pattern is consistent across metric type, scale (1.5-8B), and modality (text and image). The visualization on dense models is Figure 10.

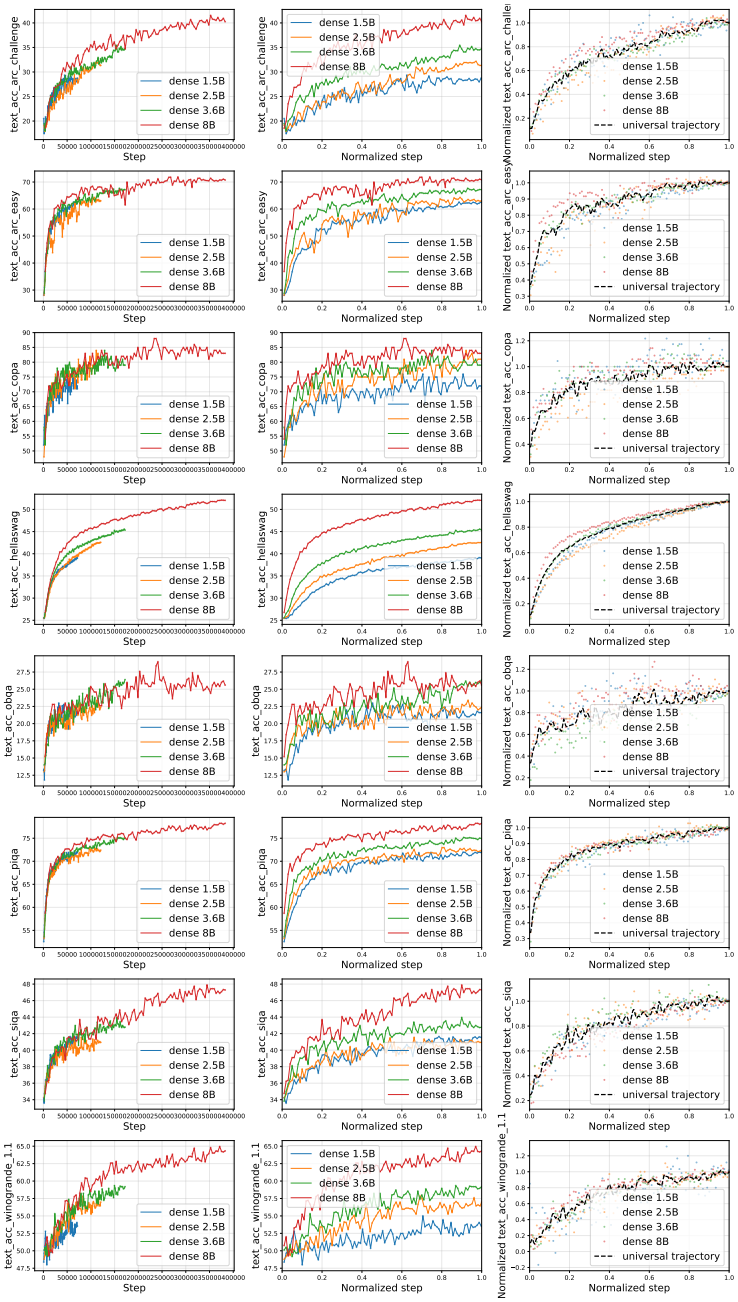


Figure 16 Individual metrics of text mean accuracy for multi-modal training on dense models.

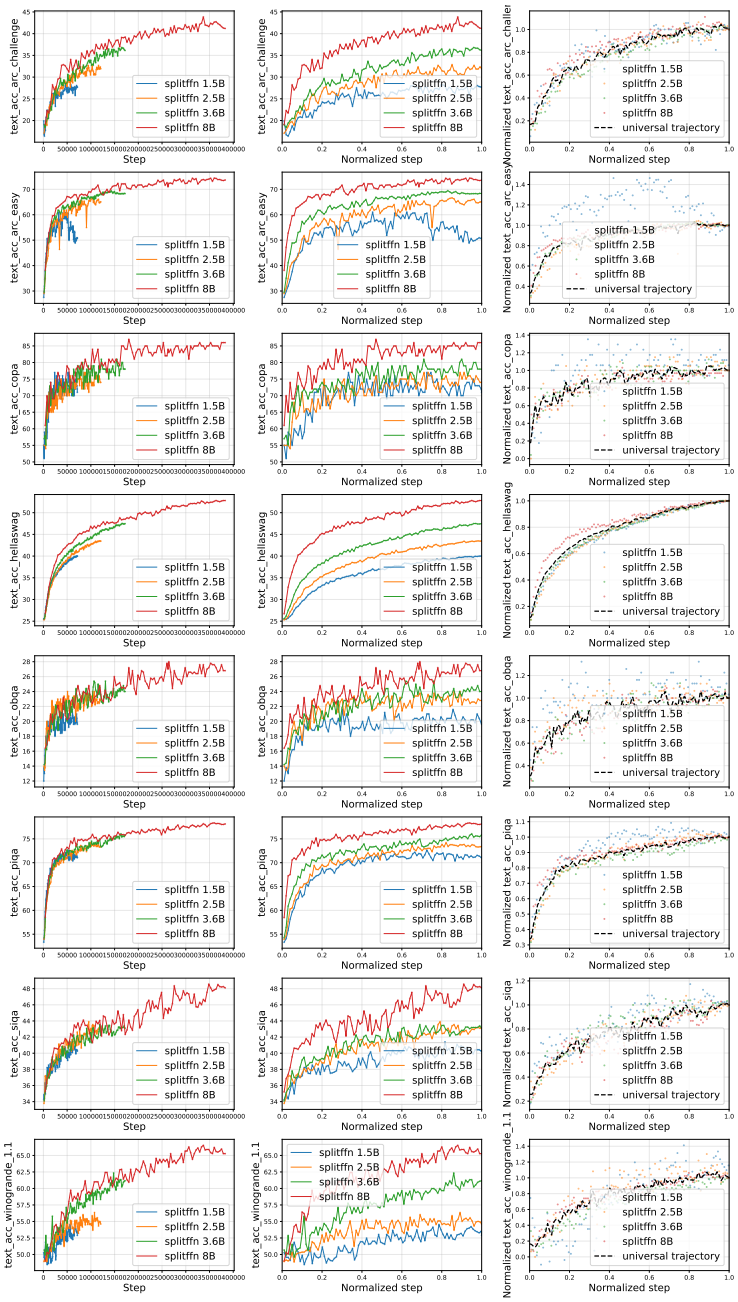


Figure 17 Individual metrics of text mean accuracy for multi-modal training on split-FFN.