EMERGENT ONE-THIRD SCALING LAW AS ATTENTION TRIES TO CONCENTRATE

Anonymous authors
Paper under double-blind review

ABSTRACT

The neural scaling law relating longer training to better performance through a power law is central to today's large language models (LLMs), yet its origin remains debated. One recent proposal is that power laws can emerge from the strong non-linearity of a single softmax head learning peaked distributions. What happens with multiple softmax functions, as in LLMs, is unclear. Here, we show through toy models that any softmax learning peaked distributions, regardless of its position in the model, can develop logit magnitudes that grow in a power law with exponent $1/3$, becoming a training bottleneck whose loss contribution decays as a power law with the same exponent $1/3$. The overall loss therefore obeys $1/3$ scaling whenever at least one softmax learns peaked distributions. We confirm that many softmax functions in LLMs learn peaked distributions and that LLM loss scaling matches this $1/3$ prediction. Moreover, logit growth dynamics reveal that attention heads, rather than the language modeling head, are the bottleneck likely driving the $1/3$ loss scaling in LLMs. Attention trying to concentrate on specific information, which is the heart of Transformers, may therefore also be the heart of the neural scaling law of training.

1 INTRODUCTION

Neural scaling laws (Hestness et al., 2017; Kaplan et al., 2020; Hoffmann et al., 2022) are a key reason large language models (LLMs) are large: loss decreases as a power law with training dataset size and model size, so longer training and more parameters persistently lead to better performance. Despite their empirical power in guiding pre-training, the origins of these power laws remain debated, leaving uncertainty about predicting and improving future scaling.

In this work, we focus on the origin of dataset size scaling. Existing proposals fall into two branches. One attributes the loss to limited information in finite observed samples (Sharma & Kaplan, 2022; Bahri et al., 2024). The other notes that LLMs use online one-epoch training, where the number of training steps is proportional to dataset size, and explains the scaling law from training dynamics (Bordelon et al., 2025a;b). Despite the difference, most works from both branches similarly conclude that power laws in loss inherit from power-law data structures.

Recently, a new mechanism was proposed showing that neural scaling can emerge from non-linearity (Liu et al., 2026b; Kühn et al., 2026) even without power-law structures in the data. When a softmax learns peaked distributions, as happens in LLMs, the non-linearity yields power-law vanishing gradients, ultimately producing power-law training dynamics. Preliminary agreement (Liu et al., 2026b; Liu & Gore, 2026) between the predicted exponent $1/3$ and loss data from open-source models (Biderman et al., 2023; OLMo et al., 2024), as well as the Chinchilla scaling laws (Hoffmann et al., 2022), suggests the mechanism may be relevant to LLMs. However, more detailed theoretical studies and LLM experiments are needed to verify this connection. In particular, the toy modeling in prior works (Liu et al., 2026b; Kühn et al., 2026) considered a single layer without feature learning (Bordelon et al., 2025a), while LLMs have multiple layers and softmax functions (e.g., softmax attention). We try to bridge this gap by asking

Question: How do multiple layers/softmaxes affect the loss landscape and training dynamics?

Answering this enables a closer comparison between theory and LLMs and tests the relevance of the proposed mechanism.

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.

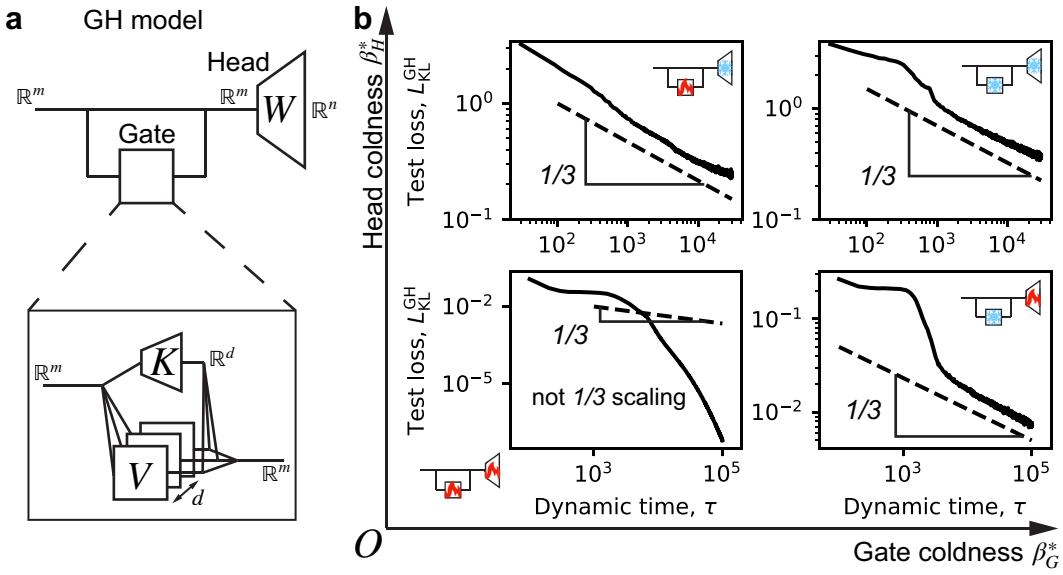

Figure 1: In the gate-head (GH) model, any cold softmax, whether the head or the gate, leads to $1/3$ time scaling of loss. (a) The GH model has two parts: a gate layer and a head layer. The gate layer uses the key matrix K and softmax to weight the value matrices. The head layer uses a matrix W and softmax to project hidden states to distributions. (b) The GH model's loss exhibits $1/3$ time scaling whenever the target coldness β_G^*, β_H^*, or both are large. Experiment details in Appendix A.1.

By experimenting with toy models, we find that as long as one softmax layer produces peaked distributions, the overall loss decays as a power law with exponent $1/3$. Softmax functions driving this power-law loss also have their logit magnitudes growing as a power law with exponent $1/3$. The theoretical contribution is explaining why softmax in middle layers can lead to the same exponent $1/3$ for the final loss, which follows from a new theorem that this scaling is not specific to loss definitions. In studying LLMs, we find that (i) attention heads and the language modeling (LM) head learn peaked distributions, (ii) some attention heads, not the LM head, have logit magnitudes growing as power laws with exponents close to $1/3$, and (iii) losses decay as power laws with exponents close to $1/3$. These findings suggest that power laws emerging from non-linearity are likely relevant to LLMs, with attention exhibiting the key non-linear effects.

To summarize, our contributions are

- Any softmax function learning peaked distributions becomes a bottleneck regardless of its location and leads to power-law loss with exponent $1/3$.
- Loss universality: The $1/3$ power-law scaling is not specific to loss definitions, which explains why softmax in middle layers can lead to $1/3$ loss scaling.
- In LLMs, we identified that attention layers trying to concentrate (learning peaked distributions over the context) may be the non-linear effect leading to $1/3$ loss scaling.

Overall, we first show that multiple softmax functions still lead to $1/3$ loss scaling and find that attention trying to concentrate may be an origin of neural scaling laws. Toy model setups are in Section 2, theoretical analysis in Section 3, and LLM experiments in Section 4. We compare with related works in Section 5 and conclude with discussions in Section 6.

2 Toy Model

To study more than one softmax function as in LLMs while remaining as simple as possible for theoretical and experimental analysis, we propose gate-head (GH) models (Figure 1a). Given input $x \in \mathbb{R}^m$ sampled as i.i.d. standard normal, the GH model outputs a probability distribution $q \in \mathbb{R}^n$,

$$q(x) = \text{Head}(\text{Norm}(h(x))),\ h(x) = \text{Norm}(x) + \text{Gate}(\text{Norm}(x)), \tag{1}$$

where $\mathrm{Norm}(\cdot)$ is root mean square layer normalization. The gate layer contains a key matrix $K \in \mathbb{R}^{d\times m}$ and a value tensor $\mathbf{V} \in \mathbb{R}^{m\times m\times d}$. Each slice of $\mathbf{V}$, $\mathbf{V}_{:,:,i} \in \mathbb{R}^{m\times m}$, is called a value matrix, and we have d value matrices. For any input $x^G \in \mathbb{R}^m$ to the gate, the gate distribution is

$$q^G = \mathrm{Softmax}(Kx^G) \in \mathbb{R}^d, \tag{2}$$

which "selects" the value matrices to use for the input x^G:

$$\mathrm{Gate}(x^G)_i = \sum_k q^G_k \sum_j \mathbf{V}_{ijk} x^G_j,\ i = 1, 2, ..., m. \tag{3}$$

Fixing k, $(\sum_j \mathbf{V}_{ijk} x^G_j)_i \in \mathbb{R}^m$ is called a value vector. The head contains a matrix $W \in \mathbb{R}^{n\times m}$:

$$\mathrm{Head}(x^H) = \mathrm{Softmax}(Wx^H) \in \mathbb{R}^n,\ \forall x^H \in \mathbb{R}^m. \tag{4}$$

To train the GH model, we use a teacher-student setup, where the student is as defined above and is trainable, and the teacher has the same architecture but with fixed parameters. We will use the subscript $*$ to denote teacher parameters (e.g., W^* is the teacher head matrix) and replace q by p for distributions from the teacher (p^G is the teacher distribution in its gate layer). Loss is defined as the Kullback–Leibler (KL) divergence between teacher and student outputs averaged over inputs,

$$L^{\mathrm{GH}}_{\mathrm{KL}} = \left\langle \sum_{i=1}^{n} p_i(x) \ln \frac{p_i(x)}{q_i(x)} \right\rangle_x. \tag{5}$$

Intuitively, an attention layer selects information from the context by a weighted sum of value vectors of previous tokens. Here, value vectors $(\sum_j \mathbf{V}_{ijk} x^G_j)_i$ from different k are weighted and summed similarly. The head layer copies the architecture of the LM head, which selects the next token to output.

The above GH model is not the only toy model that shares a heuristic similarity with LLMs. We tried different gate architectures, including the original attention, which led to the same phenomena (Appendix A). To gain a concrete understanding, we will focus on a single GH model, after which generalizations will be natural.

The most important quantity to determine the degree of non-linearity is the magnitude of logits (i.e., inputs to softmax). For small logits, softmax can be Taylor expanded with respect to logits and is therefore approximately linear (Liu et al., 2026b). To control the scale of logits, we initialize the teacher key matrix and head matrix as

$$K^* = \frac{1}{\sqrt{m}} \hat{K} \beta^*_G,\ W^* = \frac{1}{\sqrt{m}} \hat{W} \beta^*_H, \tag{6}$$

respectively, where entries in $\hat{K}$ and $\hat{W}$ are i.i.d. standard normal. Since we have normalization layers, the gate logits and head logits have standard deviations β^*_G and β^*_H, respectively. The teacher value matrices follow LeCun initialization (LeCun et al., 2002), not affecting the logit magnitude. Logit standard deviations of the student gate and head will be written as β_G and β_H, respectively. For later convenience, we emphasize

Important concepts

- Logit standard deviation will also be called **inverse temperature** and **coldness**.
- A larger logit standard deviation means lower temperature and sharper distributions.

Having introduced the GH model, we next study its training dynamics experimentally. We sweep both β^*_G and β^*_H over a wide range, from 1 to 1000, while fixing $m = 32$, $d = 8$, and $n = 128$. We use stochastic gradient descent (SGD) for optimization (details in Appendix A.1). Following Liu et al. (2026b), we plot the loss against the effective gradient-flow time, the **dynamic time**, $\tau(t) = \sum_{t'=1}^{t} \eta_{t'}$, where t is the training step and $\eta_{t'}$ is the learning rate at step t'. For a single softmax head, sufficiently large coldness leads to the asymptotic loss scaling $\sim \tau^{-1/3}$, i.e., $1/3$ time scaling (Liu et al., 2026b; Kühn et al., 2026). We find that the GH model exhibits $\tau^{-1/3}$ scaling whenever β^*_G, β^*_H, or both are large. In contrast, when both operate at high temperature, the loss can converge substantially faster than the $1/3$ time scaling (Figure 1b). The observation suggests the following picture:

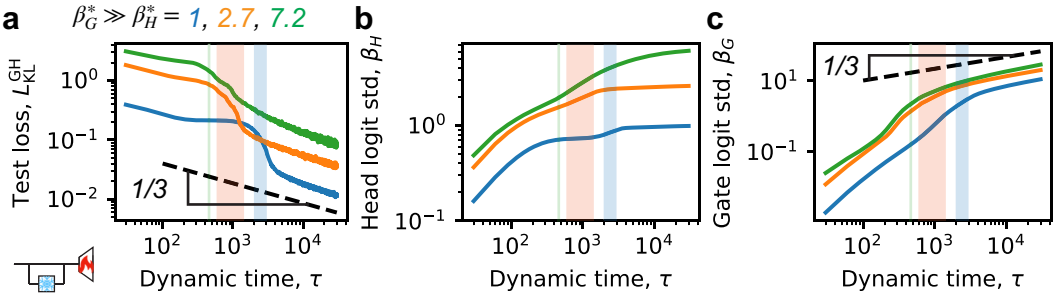

Figure 2: When gate coldness is large but head coldness is low, the gate logit magnitude grows as a $1/3$ power law after the head saturates, sustaining $1/3$ loss scaling. (a) The late-time loss follows $1/3$ time scaling across different β_H^* (represented by colors). Different loss curves enter their $1/3$ time scaling regions at different times, with the transitions marked by the vertical bands. (b) When loss enters the $1/3$ time scaling region, student head coldness saturates. (c) The $1/3$ scaling in loss also corresponds to student gate coldness growing as a $1/3$ power law. Details in Appendix A.1.

Result 1: Any softmax can be a training bottleneck for $1/3$ scaling

Any softmax in the model, whether in the gate or the head, becomes a training bottleneck when it must produce a sufficiently peaked distribution. If at least one such bottleneck is present, the overall loss cannot decay faster than $1/3$ time scaling.

3 Toy model analysis

The key new insight should come from the regime where the gate operates at low temperature while the head remains at high temperature. When the gate coldness β_G^* is small (left of Figure 1b), the gate softmax can be linearized in its logits, reducing the problem to the previously solved single softmax case (Liu et al., 2026b; Kühn et al., 2026). If β_H^* is also small, the entire model is approximately linear, and the loss eventually converges exponentially. If instead β_H^* is large, the student head must grow in norm to increase its coldness, leading to power-law vanishing gradients and the $1/3$ time scaling (Liu et al., 2026b). The remaining puzzle is therefore why a large gate coldness β_G^* can also induce the $1/3$ time scaling (right of Figure 1b). We first study the simplest such regime, with large β_G^* and small β_H^*, before extending the analysis to the case where both are large.

We examine the regime of large β_G^* and small β_H^* in more detail. Using the same setup as in Figure 1, we additionally track the standard deviations of the student logits as proxies for the weight dynamics (Appendix A.1). Fixing $\beta_G^* = 1000$, different small values of β_H^* produce loss curves with the $1/3$ time scaling tails but different crossover times (Figure 2a). Interestingly, the student head coldness β_H saturates over the same time window in which the loss crosses over to the $1/3$ scaling regime (Figure 2, a and b; colored bands). For $\beta_H^* = 7.2$, the crossover is smoother, so we mark only its onset. Because the normalization layer makes β_H depend only on the student head matrix W, saturation of β_H indicates that W has effectively converged. In contrast, the student gate coldness β_G continues to grow after head saturation, approximately following a positive $1/3$ time scaling (Figure 2c). These observations suggest that the $1/3$ power-law loss scaling is governed by a still-evolving gate (targeting large β_G^*) and an already converged head (arriving at small β_H^*).

We next turn to the gate dynamics. Once the student head W has converged to W^* and the student hidden state $h(x)$ (Eq. (1)) is close to the teacher hidden state $h^*(x)$, the leading contribution to the loss $L_{\mathrm{KL}}^{\mathrm{GH}}$ should be quadratic in $h - h^*$. Its asymptotic scaling should therefore match that of the hidden-state mean squared error (MSE),

$$L_{\mathrm{Sh}}^{\mathrm{G}} = \langle \|h(x) - h^*(x)\|^2 / m \rangle_x. \tag{7}$$

Thus, to understand the late-time scaling of $L_{\mathrm{KL}}^{\mathrm{GH}}$ after the head has converged, it is sufficient to study $L_{\mathrm{Sh}}^{\mathrm{G}}$ in a simplified model where only the gate is trained, i.e., the gate model.

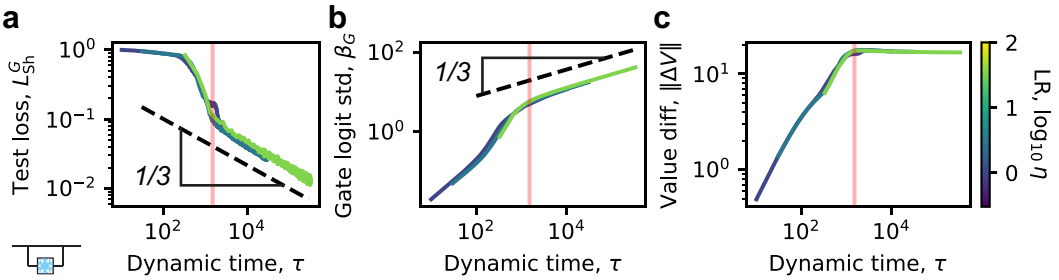

Figure 3: Training only the gate with MSE between hidden states yields $1/3$ scaling in both loss and logit magnitude. (a) The loss enters the asymptotic $1/3$ time scaling regime, with the vertical line marking its onset. (b) The student coldness exhibits positive $1/3$ time scaling, beginning at the same time. (c) The value matrix displacement from initialization saturates as the loss and coldness enter their $1/3$ scaling regimes. Curves from different learning rates collapse when plotted against τ, supporting τ as the fundamental variable governing dynamics. See Appendix A.2 for details.

To test this gate-only reduction, we train the gate using $L_{\mathrm{Sh}}^{\mathrm{G}}$ (Appendix A.2). The loss has the same asymptotic $1/3$ time scaling (Figure 3a), supporting the approximation. Meanwhile, the gate coldness grows with positive $1/3$ time scaling (Figure 3b). To determine which gate parameters remain active in this regime, we track the value tensor through its displacement from initialization, $\|\mathbf{V}(\tau) - \mathbf{V}(\tau = 0)\|$. This quantity saturates at the same time that the loss and gate coldness enter their $1/3$ scaling regimes (Figure 3c). Thus, the late-time dynamics are dominated by the non-convergent key matrix K, which controls β_G, while $\mathbf{V}$ has effectively converged.

Once $\mathbf{V} = \mathbf{V}^*$, the moments of the value vectors are constant, so we expect the MSE between hidden states to have the same scaling as that between gate probabilities,

$$L_{\mathrm{Sh}}^{\mathrm{G}} \propto L_{\mathrm{Sp}}^{\mathrm{G}} = \langle \|q^G(x) - p^G(x)\|^2/d \rangle_x. \tag{8}$$

We can therefore reduce the problem further: the origin of the $1/3$ time scaling in the full GH model can be understood by studying how training K alone under $L_{\mathrm{Sp}}^{\mathrm{G}}$ yields the same scaling.

We next develop a theory for training K with a large target β_G^* (detailed derivations in Appendix B). As explained in Liu et al. (2026b) and Kühn et al. (2026), when the norm is small, the student matrix K rapidly aligns with the teacher, while the late-time dynamics are dominated by norm growth. We therefore consider the simplified aligned student $K(\tau) = \frac{1}{\sqrt{m}}\hat{K}\beta_G(\tau)$ under gradient flow dynamics, where $K^* = \frac{1}{\sqrt{m}}\hat{K}\beta_G^*$ was defined in Eq. (6). In this case, $\beta_G(\tau)$ effectively follows its own negative gradient,

$$\frac{\mathrm{d}\beta_G}{\mathrm{d}\tau} = -\frac{1}{d}\frac{\mathrm{d}L_{\mathrm{Sp}}^{\mathrm{G}}(\beta_G)}{\mathrm{d}\beta_G}. \tag{9}$$

Since $K \propto K^*$, the loss now depends only on β_G. In the limit of large β_G and $\beta_G^* \gg \beta_G$,

$$L_{\mathrm{Sp}}^{\mathrm{G}}(\beta_G) \approx \left\langle \frac{2e^{-2\beta_G\Delta\epsilon_G}}{d(1+e^{-\beta_G\Delta\epsilon_G})^2} \right\rangle = \int_0^\infty \rho_{\Delta\epsilon_G}(\Delta\epsilon_G)\mathrm{d}\Delta\epsilon_G \frac{2e^{-2\beta_G\Delta\epsilon_G}}{d(1+e^{-\beta_G\Delta\epsilon_G})^2}. \tag{10}$$

Here, $\Delta\epsilon_G$ is the gap between the minimum and second-minimum energies, with energy defined as the normalized logit $\epsilon_G = -y_G/\beta_G$ for logits y_G. At large β_G, only $\Delta\epsilon_G$ contributes appreciably, as higher-energy gaps are exponentially suppressed. Each input therefore contributes one gap $\Delta\epsilon_G$, and the gap distribution $\rho_{\Delta\epsilon_G}(\cdot)$ reflects the data distribution. For sufficiently **complex and diverse** data, such that the inputs to K vary continuously over the space, we expect $\rho_{\Delta\epsilon_G}(0) \neq 0$, which leads to the expansion

$$L_{\mathrm{Sp}}^{\mathrm{G}}(\beta_G) = \frac{\rho_{\Delta\epsilon_G}(0)}{\beta_G d}\int_0^\infty \mathrm{d}z \frac{2e^{-2z}}{(1+e^{-z})^2} + o\left(\frac{1}{\beta_G}\right). \tag{11}$$

In our setting, where the logits are i.i.d. Gaussian, $\rho_{\Delta\epsilon_G}(\cdot)$ is approximately an exponential distribution with rate $\sqrt{2\ln d}$. The leading term above therefore dominates when $\beta_G \gg \sqrt{2\ln d}$, making

$\sqrt{2\ln d}$ a rough boundary between the high- and low-temperature regimes. In the low-temperature regime $\beta_G^* \gg \beta_G \gg \sqrt{2\ln d}$, $L_{\rm Sp}^{\rm G}(\beta_G) \sim 1/\beta_G$, which together with Eq. (9) leads to

$$\beta_G \sim \tau^{1/3},\ L_{\rm Sp}^{\rm G} \sim \tau^{-1/3}. \tag{12}$$

In this idealized theory, the dynamics depend only on τ, consistent with the empirical collapse of curves obtained with different learning rates when plotted against τ (Figure 3). We therefore identify the continued growth of the key matrix K as the origin of the $1/3$ time scaling. Specifically, the power-law training dynamics arise because the non-linear gate, together with a complex data distribution, produces a loss that intrinsically vanishes as a power law.

The key to the emergent power law is the expansion, Eq. (11). Previous work (Liu et al., 2026b) indicated that this inverse-coldness scaling arises specifically from the KL divergence between probabilities. Here, we show that MSE between probabilities exhibits the same scaling. More generally, replacing the MSE $\langle \|q^G - p^G\|^2/d \rangle$ with $\langle \|q^G - p^G\|^k/d \rangle$ for $k = 4, 6, 8, \ldots$ changes only the integral coefficient (i.e., $\int_0^\infty {\rm d}z \cdots$ in Eq. (11)), but not the exponent. In fact, the same inverse-coldness scaling holds for all "reasonable" loss functions:

Result 2: Loss universality (informal Theorem 1)

Consider aligned student and teacher with coldness β and β^*, corresponding probabilities q and p, and a non-zero density of zero energy gap. In the regime of large β and $\beta^* \gg \beta$, **any reasonable loss** $L = \langle D(q,p) \rangle$ scales as $L \sim 1/\beta$, **provided that** $D(q,p)$ is non-negative, vanishes iff $q = p$, and remains differentiable with bounded gradient even when p is one-hot.

The conditions for a "reasonable" loss are mild and probably necessary for trainability. The emergent power law is therefore not a special property of KL divergence or MSE. Instead, it comes from generic ingredients: saturation of probabilities at large coldness and the presence of ambiguous cases in complex data (i.e., $\rho_{\Delta\epsilon_G}(0) \neq 0$).

We can now explain the $1/3$ time scaling of the final $L_{\rm KL}^{\rm GH}$ when the gate is at low temperature and the head is at high temperature more directly. The high-temperature part of the GH model converges rapidly, leaving $L_{\rm KL}^{\rm GH}$ effectively as a function of the non-convergent q^G. Although this dependence is complicated, the mapping from q^G to the final probabilities is well-behaved, so $L_{\rm KL}^{\rm GH}$ remains a reasonable loss between q^G and p^G. Loss universality gives $L_{\rm KL}^{\rm GH} \sim 1/\beta_G$, and gradient-based training then leads to the $1/3$ time scaling. Our earlier reduction from $L_{\rm KL}^{\rm GH}$ to $L_{\rm Sp}^{\rm G}$ provided an intuitive explanation but relied on several approximations. Loss universality shows that the precise form of these approximations is irrelevant to the scaling exponent. A low-temperature softmax in a middle layer can therefore yield the same $1/3$ time scaling as a low-temperature softmax at the end.

The case where both β_G^* and β_H^* are large (Figure 1b, upper right) follows naturally. After alignment, the weights other than K and W converge, while K and W continue to grow in norm and dominate the late-time loss. Analogous to a multi-variable Taylor expansion, $L_{\rm KL}^{\rm GH}$ can be decomposed at leading order into a contribution from imperfect K with W converged and a contribution from imperfect W with K converged (cross terms are higher order at late times). By loss universality, both leading contributions scale as $\tau^{-1/3}$, and therefore so does $L_{\rm KL}^{\rm GH}$.

This argument extends easily beyond the specific GH model. In a model with more softmax functions, each softmax that must learn a peaked distribution can contribute a $\tau^{-1/3}$ term to the late-time loss. Thus, the presence of one or more non-convergent softmax modules, regardless of their positions, can lead to the $1/3$ time scaling of the overall loss. Such bottlenecks in late-time training can be identified by coldness that continues to grow with positive $1/3$ time scaling. We therefore provide an answer to the Question in Section 1 on how multiple softmaxes affect the loss scaling.

4 LLM EXPERIMENTS

Guided by the mechanism revealed in our toy models, we next ask whether the same picture can explain neural scaling laws in LLMs. Since we mainly focus on the original scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022), which are reported from dense models with softmax attention (Vaswani et al., 2017), we evaluate such models (Biderman et al., 2023; OLMo et al., 2024).

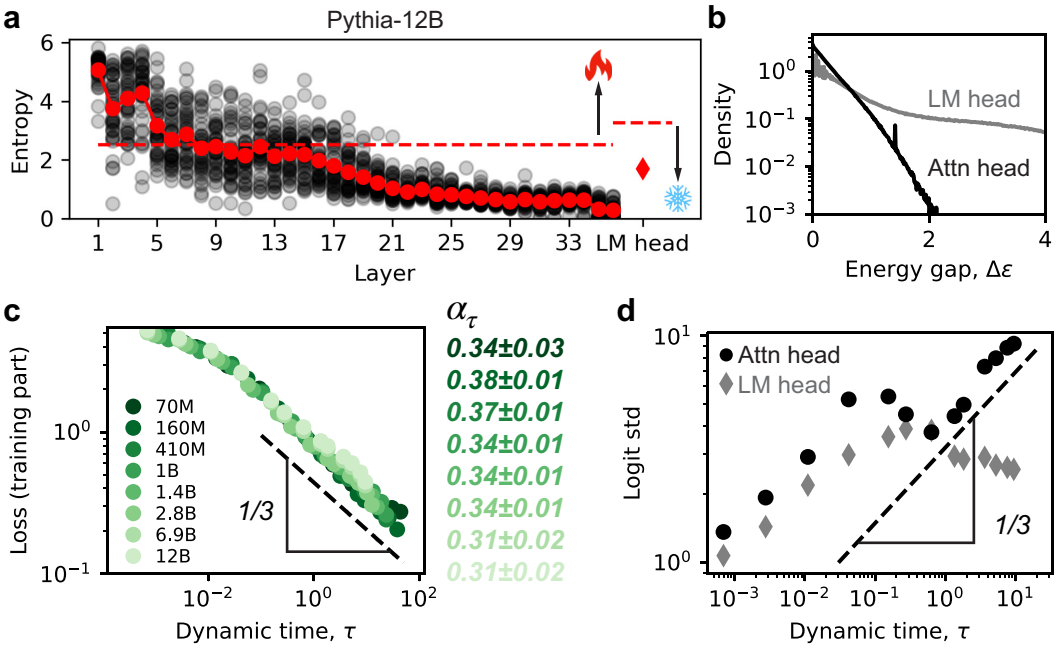

Figure 4: In Pythia LLMs, attention heads are more strongly peaked than the LM head, with logit magnitudes growing as $1/3$ power laws, identifying attention as the bottleneck for $1/3$ loss scaling. (a) Entropy suggests that later-layer attention heads and the LM head are in the low-temperature regime (below the horizontal dashed lines). (b) Probability densities at zero energy gap are non-zero. (c) LLMs have losses following the $1/3$ time scaling. (d) Logit coldness of attention heads rather than the LM head grows as the positive $1/3$ time scaling at late training. Details in Appendix C.1.

We first ask whether softmax functions in LLMs learn peaked distributions. Our theory predicts a rough boundary between high- and low-temperature regimes for i.i.d. Gaussian logits (Appendix B.3). Because LLM logits need not follow this assumption, we convert this boundary into an entropy threshold and compare it directly with LLM distributions (details in Appendix C.1). Figure 4a shows results for Pythia-12B (Biderman et al., 2023), with other models giving similar results (Appendix C.1). Dark dots denote the mean attention entropy of individual heads, while red dots show layer averages. The red diamond attached to "LM head" marks the entropy of next-token prediction (Appendix C.1). The dashed line indicates the corresponding high- or low-temperature boundary, with separate thresholds for attention and the LM head. We find that later-layer attention heads and the LM head lie in the low-temperature regime, with some attention heads reaching extremely low entropies.

We next test the second condition required by our theory for the $1/3$ time scaling: the probability density of the energy gap is non-zero at zero. Each input produces a set of logits and hence one energy gap, giving the gap distribution after evaluating the dataset (Appendix C.1). For both attention heads and the LM head, the observed densities approach non-zero values as the gap tends to zero (Figure 4b). For attention, we show one representative later-layer head from Pythia-12B, with similar behavior across other low-temperature heads (Appendix C.1). We conclude that the ingredients in our toy models that lead to the $1/3$ time scaling are present in LLMs.

We next test whether LLMs exhibit the predicted power-law training dynamics. Following Liu et al. (2026b), we evaluate checkpoints of open-source models and fit the raw loss as

$$L = \frac{c_\tau}{\tau^{\alpha_\tau}} + L_{\setminus\tau}, \tag{13}$$

where $L_{\setminus\tau}$ is the part of the loss irreducible by training (Appendix C.1). For Pythia, we plot the training-dependent component $L - L_{\setminus\tau}$ in Figure 4c, together with the fitted α_τ. Curves from different model sizes collapse when plotted against τ, indicating that τ is the relevant training variable and that c_τ and α_τ are approximately size-independent. The power-law fits are good, with exponents

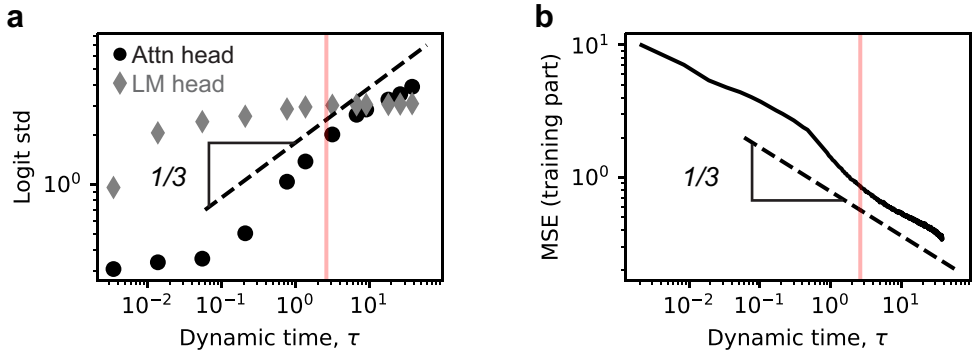

Figure 5: Using MSE between final logits for training, attention logit magnitude grows like a positive $1/3$ scaling while the LM head saturates (panel a), ruling out the LM head as the only bottleneck for $1/3$ loss scaling. (b) MSE loss has a tail following the $1/3$ time scaling, as suggested by loss universality. Details in Appendix C.2.

all close to $1/3$. Our theory also predicts positive $1/3$ time scaling in some logit magnitudes, which is observed (Figure 4d from Pythia-12B; more examples in Appendix C.1): the growing logits occur in attention heads rather than the LM head, suggesting that attention is the late-training bottleneck. OLMo models (OLMo et al., 2024) show the same behaviors (Appendix C.1). Together, all these agreements suggest that our theory may capture the mechanism underlying the neural scaling law of training, with attention as the likely source of the emergent power laws.

Liu et al. (2026b) indicated that the LM head is the bottleneck responsible for the $1/3$ loss scaling. We test this independently by training LLMs with MSE between final logits, which bypasses the output softmax. If that softmax were the only bottleneck, both logits and loss should converge much faster. We train a Pythia-160M student using a trained Pythia-160M model as the teacher providing target logits (Appendix C.2). As expected, the student LM head coldness saturates rather than exhibiting power-law growth (Figure 5a). However, some attention-head coldness continues to grow approximately with positive $1/3$ time scaling, ruling out the LM head as the sole bottleneck. Fitting the MSE loss after LM head saturation (vertical line in Figure 5a) with a power law plus a constant reveals a $1/3$ scaling tail in the training part of the MSE (Figure 5b). These results identify attention as a training bottleneck and directly support loss universality in LLMs.

To summarize, in LLMs we tested the theoretical assumptions and verified predictions for both loss and internal logits. We further ran training experiments with MSE between logits, providing an independent test. The agreement across these analyses supports the following conclusion:

Result 3: Relevance to LLMs

Power laws emerging from non-linearity are likely relevant to neural scaling laws in LLMs. In particular, attention concentrating toward peaked distributions is likely the late-training bottleneck for the $1/3$ power-law loss scaling.

5 Related Works

Empirical studies first established that loss follows approximate power laws (Kaplan et al., 2020; Brown et al., 2020; Henighan et al., 2020; Hoffmann et al., 2022; Achiam et al., 2023), with the Chinchilla scaling laws (Hoffmann et al., 2022) providing more accurate estimates than Kaplan et al. (2020) by accounting for irreducible loss. Following Liu et al. (2026b), in online learning with a fixed learning rate schedule shape, dynamic time is proportional to dataset size. The $1/3$ time scaling can therefore also be viewed as $1/3$ dataset size scaling, consistent with the exponent $0.28 \sim 0.37$ measured from Chinchilla data (Hoffmann et al., 2022; Besiroglu et al., 2024).

One branch of theory studies the optimal test loss achievable with a fixed number of training samples. These works attribute power-law scaling to data properties such as manifold dimension (Spigler et al., 2020; Hutter, 2021; Sharma & Kaplan, 2022) or power-law covariance spectra (Bordelon et al.,

2020; Maloney et al., 2022; Bahri et al., 2024; Brill, 2024), leading to exponents that depend sensitively on data. Such high-level theories are difficult to test directly in LLMs. A more specific theory based on effective context horizons growing with dataset size (Cagnetta et al., 2026) is testable, but its data analysis, like that of Kaplan et al. (2020), ignores irreducible loss. The relevance of these theories to LLMs therefore may benefit from further justification.

For online, one-epoch training, theories of training dynamics are conceptually more relevant. Most works analyze linear models (Lin et al., 2024; Bordelon et al., 2025b), kernels (Bordelon & Pehlevan, 2021; Bordelon et al., 2024; Worschech & Rosenow, 2024; Paquette et al., 2024; Bordelon et al., 2025a; Defilippis et al., 2025), or high-level learning arguments (Michaud et al., 2023; Arora & Goyal, 2023; Liu et al., 2025b) without treating non-linear effects explicitly. Their common picture is that more important modes are learned first, so power-law structures in the data, such as a power-law covariance spectrum, are required for power-law training dynamics and determine the loss exponent.

Recent work on training dynamics (Liu et al., 2026b) instead showed that softmax can yield power-law loss when learning peaked distributions, leading to a robust $1/3$ exponent across diverse data structures even without power laws in data. Our work generalizes this analysis from a single softmax to multiple softmax functions and identifies training attention as the bottleneck for $1/3$ loss scaling in LLMs. Kühn et al. (2026) analyzed the single softmax case more rigorously by incorporating gradient noise and the alignment process, but did not test the theory against LLMs.

Observations of "attention sinks", "massive activations", etc. (Xiao et al., 2024; Sun et al., 2024; Zucchet et al., 2025; Queipo-de-Llano et al., 2026) suggest that attention tries to concentrate, yet to our knowledge, this work first proposes that these observations may connect to neural scaling laws.

6 Discussion

From toy modeling, we identify learning peaked distributions as a mechanism for emergent power-law loss, regardless of where it occurs in the model. LLM loss and logit dynamics follow the predicted $1/3$ time scaling, suggesting that attention concentrating on specific tokens is the bottleneck underlying this power law. Thus, the tendency of attention to concentrate, a key feature of Transformers, may also underlie the neural scaling law of training.

Our work has several limitations. We focus on scaling with training time, leaving possible effects of non-linearity on model-size scaling unexplored (Liu et al., 2025a; 2026a). We analyze gradient-flow dynamics, neglecting gradient noise (Kühn et al., 2026), optimizer-specific effects (although Adam shows similar time scaling; Appendix A.3), and hyperparameter scaling (Bergsma et al., 2025), all of which may affect the mapping between dynamic time and dataset size. We also focus on the original neural scaling laws, while newer architectures such as mixture of experts (MoE) may modify their form (Clark et al., 2022). For training-time scaling, however, we expect MoE to retain the $1/3$ exponent: expert routing introduces another softmax that may itself become another bottleneck, while increased capacity lowers the achievable loss. These limitations, together with our findings, open directions for future work.

Our results suggest a broader principle for improving LLM scaling. The "bitter lesson" (Sutton, 2019) favors flexible mechanisms such as attention, which gives LLMs great freedom in modeling context dependence. Yet language may not require such freedom uniformly: only a small fraction of context is often relevant, forcing attention to concentrate and, in our view, producing slow power-law training. Structural priors might therefore accelerate learning by reducing how sharply attention must focus, without sacrificing expressivity. Alternatively, intelligence may fundamentally require selecting a few relevant elements from a vast space of possibilities. Such non-linear concentration, and the associated $1/3$ scaling, may thus be unavoidable. We anticipate that distinguishing between these possibilities could enable more efficient scaling or reveal deeper principles underlying language and intelligence.

AI use statement

In this work, we used generative AI tools for general brainstorming, formulating mathematical claims, writing code with clear instructions, drafting appendices based on code, and improving

the main text readability. We did not use generative AI tools to develop theoretical models or conceptual frameworks, prove mathematical claims, write proofs, or perform data analysis. The rest of the required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work. We judged AI-generated ideas, verified mathematical claims manually, and wrote proofs manually at the end with our own logic. LLM-generated code was verified and tested for correctness by the authors. Data analysis was done manually by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Reproducibility statement

We explain the logic of experiments and data analysis in appendices, with clear references to the corresponding code files. Code to reproduce all of the results is in the supplementary material. For the theoretical results, clear explanations of assumptions and complete proofs of the claims can be found in the appendices.

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.

Sanjeev Arora and Anirudh Goyal. A theory for emergence of complex skills in language models. *arXiv preprint arXiv:2307.15936*, 2023.

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.

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. *Advances in Neural Information Processing Systems*, 38:125153–125188, 2025.

Tamay Besiroglu, Ege Erdil, Matthew Barnett, and Josh You. Chinchilla scaling: A replication attempt. *arXiv preprint arXiv:2404.10102*, 2024.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023.

Blake Bordelon and Cengiz Pehlevan. Learning curves for sgd on structured features. *arXiv preprint arXiv:2106.02713*, 2021.

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pp. 1024–1034. PMLR, 2020.

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. *arXiv preprint arXiv:2402.01092*, 2024.

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. How feature learning can improve neural scaling laws. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(8):084002, 2025a.

Blake Bordelon, Mary I Letey, and Cengiz Pehlevan. Theory of scaling laws for in-context regression: Depth, width, context and time. *arXiv preprint arXiv:2510.01098*, 2025b.

Ari Brill. Neural scaling laws rooted in the data distribution. *arXiv preprint arXiv:2412.07942*, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Francesco Cagnetta, Allan Raventós, Surya Ganguli, and Matthieu Wyart. Deriving neural scaling laws from the statistics of natural language. *arXiv preprint arXiv:2602.07488*, 2026.

Aidan Clark, Diego de Las Casas, Aurelia Guy, Arthur Mensch, Michela Paganini, Jordan Hoffmann, Bogdan Damoc, Blake Hechtman, Trevor Cai, Sebastian Borgeaud, et al. Unified scaling laws for routed language models. In *International conference on machine learning*, pp. 4057–4086. PMLR, 2022.

Leonardo Defilippis, Yizhou Xu, Julius Girardin, Emanuele Troiani, Vittorio Erba, Lenka Zdeborová, Bruno Loureiro, and Florent Krzakala. Scaling laws and spectra of shallow neural networks in the feature learning regime. *arXiv preprint arXiv:2509.24882*, 2025.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

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.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

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.

Marcus Hutter. Learning curve theory. *arXiv preprint arXiv:2102.04074*, 2021.

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.

Marcel Kühn, Yoon Thelge, and Bernd Rosenow. A boundary-layer mechanism for one-third scaling in online softmax classification. *arXiv preprint arXiv:2605.22341*, 2026.

Yann LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pp. 9–50. Springer, 2002.

Licong Lin, Jingfeng Wu, Sham M Kakade, Peter L Bartlett, and Jason D Lee. Scaling laws in linear regression: Compute, parameters, and data. *Advances in Neural Information Processing Systems*, 37:60556–60606, 2024.

Yizhou Liu and Jeff Gore. Neural scaling universality: If exponents are fixed, time to understand coefficients. *arXiv preprint arXiv:2606.25008*, 2026.

Yizhou Liu, Ziming Liu, and Jeff Gore. Superposition yields robust neural scaling. *Advances in Neural Information Processing Systems*, 38:159269–159305, 2025a.

Yizhou Liu, Sara Kangaslahti, Ziming Liu, and Jeff Gore. Inverse depth scaling from most layers being similar. *arXiv preprint arXiv:2602.05970*, 2026a.

Yizhou Liu, Ziming Liu, Cengiz Pehlevan, and Jeff Gore. Universal one-third time scaling in learning peaked distributions. *arXiv preprint arXiv:2602.03685*, 2026b.

Ziming Liu, Yizhou Liu, Eric J Michaud, Jeff Gore, and Max Tegmark. Physics of skill learning. *arXiv preprint arXiv:2501.12391*, 2025b.

Alexander Maloney, Daniel A Roberts, and James Sully. A solvable model of neural scaling laws. *arXiv preprint arXiv:2210.16859*, 2022.

Eric Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. The quantization model of neural scaling. *Advances in Neural Information Processing Systems*, 36:28699–28722, 2023.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.

Elliot Paquette, Courtney Paquette, Lechao Xiao, and Jeffrey Pennington. 4+ 3 phases of compute-optimal neural scaling laws. *Advances in Neural Information Processing Systems*, 37:16459–16537, 2024.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin A Raffel, Leandro Von Werra, Thomas Wolf, et al. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024.

Enrique Queipo-de-Llano, Álvaro Arroyo, Federico Barbero, Xiaowen Dong, Michael Bronstein, Yann LeCun, and Ravid Shwartz-Ziv. Attention sinks and compression valleys in llms are two sides of the same coin. In *International Conference on Learning Representations*, volume 2026, pp. 13911–13937, 2026.

Sidney I. Resnick. *Extreme Values, Regular Variation, and Point Processes*. Springer New York, NY, 1987. doi: 10.1007/978-0-387-75953-1.

Utkarsh Sharma and Jared Kaplan. Scaling laws from the data manifold dimension. *Journal of Machine Learning Research*, 23(9):1–34, 2022.

Stefano Spigler, Mario Geiger, and Matthieu Wyart. Asymptotic learning curves of kernel methods: empirical data versus teacher–student paradigm. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12):124001, 2020.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*, 2024.

Richard S. Sutton. The bitter lesson. Incomplete Ideas, 2019. March 13, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Roman Worschech and Bernd Rosenow. Analyzing neural scaling laws in two-layer networks with power-law data spectra. *arXiv preprint arXiv:2410.09005*, 2024.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024.

Nicolas Zucchet, Francesco d'Angelo, Andrew Lampinen, and Stephanie Chan. The emergence of sparse attention: impact of data distribution and benefits of repetition. *Advances in Neural Information Processing Systems*, 38:35121–35158, 2025.

A TOY MODEL EXPERIMENTS

In the main text, we introduced the GH model and its key results. We hereafter refer to this model as the type-I GH model, or GH-I. In this appendix, we first provide details of the training and analysis of GH-I. We then introduce two additional variants, GH-II and GH-III, which differ from GH-I in their gate architectures. Analysis of GH-I suggests the origin of the emergent $1/3$ scaling as a saturating non-linear effect not specific to its architecture. Experiments on GH-II and GH-III further support the generality of this mechanism, pointing toward more fundamental mathematical principles.

A.1 GH-I MODEL (THE MAIN-TEXT GH MODEL)

The basic architecture and setup of the GH-I model have been described in the main text (Section 2).

A.1.1 EXPERIMENT METHODS

The full implementation is in `exp-0-1.py`. The architecture is as described in Section 2, with $m = 32$, $d = 8$, and $n = 128$. The teacher key and head matrices are set according to Eq. (6), and the teacher value tensor $\mathbf{V}^*$ follows LeCun initialization. The student key and head matrices are initialized to zero, and the student value tensor is initialized to a smaller scale than LeCun initialization.

In each training step, inputs $x \in \mathbb{R}^m$ are sampled i.i.d. from the standard normal distribution with batch size 2048. The training loss is the KL divergence $L_{\mathrm{KL}}^{\mathrm{GH}}$ (Eq. (5)). The optimizer is SGD with a constant learning rate η. We train for 10,000 steps, sweeping β_G^* and β_H^* each over 8 logarithmically spaced values from 1 to 1000, and η over $\{3, 10, 30\}$, giving 192 runs in total. The dynamic time is $\tau = \eta t$ for a constant learning rate, where t is the training step.

Every 10 steps, the test loss is evaluated on 4 fresh batches of 2048 samples and averaged. Three additional quantities are tracked: the student head coldness $\beta_H = \|W\|/\sqrt{n}$, the student gate coldness $\beta_G = \|K\|/\sqrt{d}$, and the value tensor displacement from initialization $\|\mathbf{V}(t) - \mathbf{V}(0)\|$. These are consistent with the main text definitions, since the normalization layers make the logit standard deviations proportional to the respective matrix norms. All quantities are saved per run to a result `.pt` file.

A.1.2 DATA ANALYSIS

All 64 combinations of (β_G^*, β_H^*) were inspected by plotting 8×8 panels of test loss curves against τ on log-log axes, with three learning rates overlaid per panel (see `exp-0-1.ipynb`). The dynamics of β_H and β_G were similarly scanned across all combinations. These scans confirm that, within the large-β^* regime, behaviors across different parameter values are qualitatively consistent. The curves shown in the figures are therefore representative of their regimes, not particular cases.

For Figure 1b, we selected the four corners of the (β_G^*, β_H^*) sweep, where each axis takes its minimum value 1 or maximum value 1000. Each of the four panels plots $L_{\mathrm{KL}}^{\mathrm{GH}}$ against τ on log-log axes with one representative learning rate (see `exp-0-1.ipynb`) and a dashed reference line of slope $-1/3$.

For Figure 2, we fix $\beta_G^* = 1000$ and vary β_H^* over its three smallest sweep values, $\beta_H^* \approx 1, 2.7, 7.2$, with learning rate $\eta = 3$. Panels a, b, and c plot $L_{\mathrm{KL}}^{\mathrm{GH}}$, β_H, and β_G against τ on log-log axes, respectively. The quantities $\beta_H = \|W\|/\sqrt{n}$ and $\beta_G = \|K\|/\sqrt{d}$ are the saved proxies described in Appendix A.1.1, with no further processing. Dashed reference lines of slope $-1/3$ and $+1/3$ are drawn for comparison in panels a and c, respectively. The vertical bands marking the transition to the $1/3$ scaling regime are drawn consistently across all three panels. For $\beta_H^* \approx 7.2$, the crossover is smoother and longer, so a single vertical line marks only its onset.

A.2 G-I MODEL

The G-I model is the gate layer of GH-I trained in isolation, as described in the main text (Section 3). The full implementation is in `exp-1-1.py`.

A.2.1 EXPERIMENT METHODS

The G-I model uses the gate layer architecture with the same hyperparameters as GH-I ($m = 32$, $d = 8$), but without the head layer. The teacher key matrix K^* is set according to Eq. (6), and the teacher value tensor $\mathbf{V}^*$ follows LeCun initialization. The student K is initialized to zero, and the student $\mathbf{V}$ is initialized to a smaller scale than LeCun initialization.

In each training step, inputs are sampled i.i.d. from the standard normal distribution with batch size 2048. The training loss is $L_{\mathrm{Sh}}^{\mathrm{G}}$, the MSE between student and teacher hidden states (main text, Section 3). The optimizer is SGD with constant learning rate η. We train for 10,000 steps, sweeping β_G^* over 8 logarithmically spaced values from 1 to 1000 and η over $\{0.3, 1, 3, 10, 15, 25, 33, 100\}$, giving 64 runs in total. The dynamic time is $\tau = \eta t$. Every 10 steps, the test loss is evaluated on 4 fresh batches and averaged. The student gate coldness $\beta_G = \|K\|/\sqrt{d}$ and the value tensor displacement $\|\mathbf{V}(t) - \mathbf{V}(0)\|$ are also recorded.

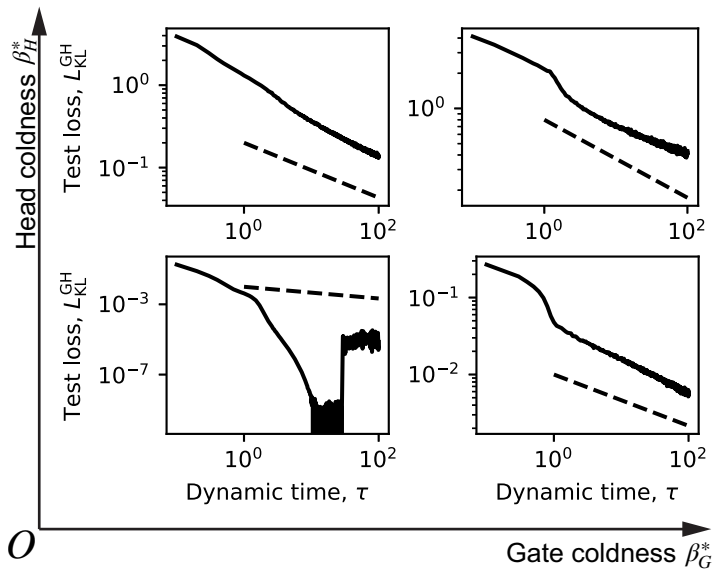

Figure 6: Under Adam, the GH model's loss exhibits approximate $1/3$ time scaling whenever the target coldness β_G^*, β_H^*, or both are large. Each panel shows the test loss $L_{\mathrm{KL}}^{\mathrm{GH}}$ against dynamic time τ on log-log axes, with a dashed reference line of slope $-1/3$. The 2×2 layout arranges the four corners of the (β_G^*, β_H^*) sweep, with gate coldness increasing rightward and head coldness increasing upward. When both coldnesses are small (bottom left), the loss initially descends rapidly as in the SGD case, but Adam becomes unstable at late times. In the remaining three panels, approximate $1/3$ loss scaling is observed, consistent with the SGD results in Figure 1b.

A.2.2 Data analysis

All 8 values of β_G^* were inspected by plotting test loss, β_G, and $\|\mathbf{V}(t) - \mathbf{V}(0)\|$ against τ on log-log axes, with all 8 learning rates overlaid per panel (see `exp-1-1.ipynb`). These scans confirm that behaviors in the large-β_G^* regime are qualitatively consistent, so the selected curves are representative.

For Figure 3, we fix $\beta_G^* = 1000$ and show three representative learning rates $\eta = 1, 3, 33$, color-coded by $\log_{10} \eta$. Panels a, b, and c plot $L_{\mathrm{Sh}}^{\mathrm{G}}$, β_G, and $\|\mathbf{V}(t) - \mathbf{V}(0)\|$ against τ on log-log axes, respectively. Dashed reference lines of slope $-1/3$ and $+1/3$ are drawn in panels a and b; panel c has no reference slope, as it shows saturation rather than a power-law trend. The vertical line at $\tau = 1500$ marking the onset of the $1/3$ scaling regime is consistent across all three panels. Curves from different learning rates collapse when plotted against τ, supporting τ as the fundamental variable governing the dynamics.

A.3 Adam optimizer

Adam is considerably harder to analyze theoretically than gradient flow, the continuous-time limit that underlies the analysis of SGD. The core difficulty is that Adam maintains per-parameter adaptive learning rates based on gradient moments, making the exact dynamics of β analytically intractable. We therefore take an empirical approach and run the same GH-I and G-I experiments under Adam. The results show qualitatively identical phenomena: loss and β follow approximate $1/3$ power laws in both models. The effective dynamics of β appear similar under Adam and SGD, and differences at the level of individual parameter updates appear to be higher-order corrections that do not change the dominant scaling exponent.

A.3.1 GH-I extra experiments

The GH-I Adam experiments use the same architecture, hyperparameters, and sweep design as Appendix A.1.1, with one modification: Adam replaces SGD. The learning rate is swept over

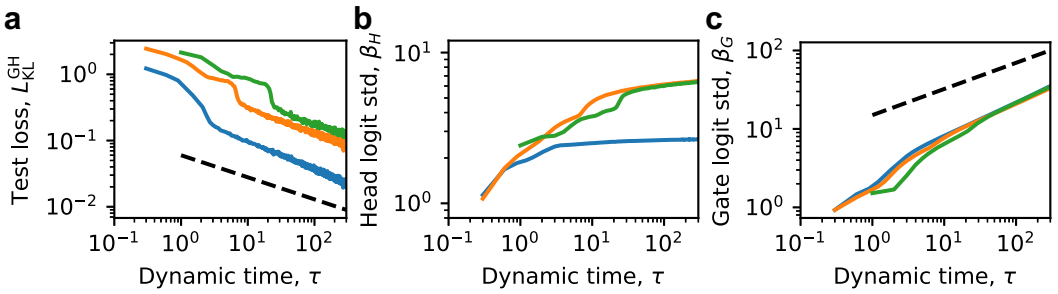

Figure 7: Under Adam, when gate coldness is large but head coldness is small, the head saturates and the gate logit magnitude subsequently grows approximately as $\tau^{1/3}$, sustaining approximate $1/3$ loss scaling. Three representative runs are shown, color-coded by run. (a) Test loss $L_{\mathrm{KL}}^{\mathrm{GH}}$ against τ; the dashed line has slope $-1/3$. (b) Head logit standard deviation β_H against τ; β_H grows initially then saturates. (c) Gate logit standard deviation β_G against τ; the dashed line has slope $+1/3$. The sequence of head saturation followed by gate growth and continued loss decay mirrors the SGD behavior in Figure 2, though the alignment with the reference slopes is less tight.

$\{0.01, 0.03, 0.1\}$, giving 192 runs in total. The dynamic time is defined as $\tau = \eta t$, where η is the nominal learning rate.

All 64 combinations of (β_G^*, β_H^*) were inspected by plotting 8×8 panels of test loss curves against τ on log-log axes, with three learning rates overlaid per panel (see `exp-0.ipynb`). The dynamics of β_H and β_G were similarly scanned across all combinations. Curves from different learning rates approximately collapse when plotted against τ, though less tightly than under SGD. Within the large-β^* regime, behaviors across different parameter values are qualitatively consistent, so the curves shown in the figures are representative.

Figure 6 reproduces the four-corner selection of Figure 1b under Adam, using learning rate $\eta = 0.01$ and a dashed reference line of slope $-1/3$. Approximate $1/3$ loss scaling is observed whenever β_G^*, β_H^*, or both are large, consistent with the SGD results.

Figure 7 shows three panels plotting $L_{\mathrm{KL}}^{\mathrm{GH}}$, β_H, and β_G against τ on log-log axes. Three representative runs are shown: $(\beta_G^*, \beta_H^*, \eta) \in \{(373, 2.7, 0.03), (1000, 7.2, 0.03), (373, 7.2, 0.1)\}$, all with large β_G^* and small β_H^*, color-coded by blue, orange, and green, respectively. Dashed reference lines of slope $-1/3$ and $+1/3$ are drawn in the loss and β_G panels, respectively. As with SGD, β_H saturates early and β_G subsequently grows approximately as a $\tau^{1/3}$ power law, consistent with the gate becoming the bottleneck.

These results show that the phenomena observed under SGD persist under Adam: approximate $1/3$ loss scaling emerges whenever a softmax learns peaked distributions, and the gate-bottleneck mechanism is visible in the logit dynamics. While the theoretical analysis relies on gradient flow, the empirical agreement under Adam suggests the underlying mechanism may not depend sensitively on the optimizer.

A.3.2 G-I EXTRA EXPERIMENTS

The G-I Adam experiments use the same architecture, hyperparameters, and sweep design as the G-I experiment in Appendix A.2.1, with one modification: Adam replaces SGD. The learning rate is swept over $\{0.001, 0.003, 0.01, 0.03, 0.05, 0.08, 0.1, 0.3\}$, giving 64 runs in total. The dynamic time is defined as $\tau = \eta t$.

All 8 values of β_G^* were inspected by plotting test loss, β_G, and $\|\mathbf{V}(t) - \mathbf{V}(0)\|$ against τ on log-log axes, with all 8 learning rates overlaid per panel (see `exp-1.ipynb`). Curves from different learning rates approximately collapse when plotted against τ. Within the large-β_G^* regime, behaviors are qualitatively consistent, so the selected curves are representative.

For Figure 8, we fix $\beta_G^* = 1000$ and show three representative learning rates $\eta = 0.001$, 0.03, and 0.1, color-coded by $\log_{10} \eta$ from dark purple to light green. Panels a, b, and c plot $L_{\mathrm{Sh}}^{\mathrm{G}}$, β_G,

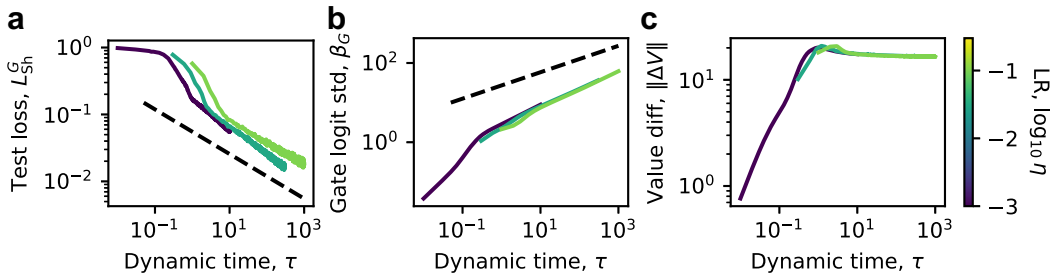

Figure 8: Under Adam, training only the gate with MSE loss yields approximate $1/3$ scaling in both loss and gate logit magnitude, mirroring the SGD results in Figure 3. We fix $\beta_G^* = 1000$ and show three representative learning rates, color-coded by $\log_{10}\eta$ from dark purple to light green. (a) Test loss $L_{\mathrm{Sh}}^{\mathrm{G}}$ against τ; the dashed line has slope $-1/3$. (b) Gate logit standard deviation β_G against τ; the dashed line has slope $+1/3$. (c) Value tensor displacement $\|\mathbf{V}(t) - \mathbf{V}(0)\|$ against τ; $\|\mathbf{V}(t) - \mathbf{V}(0)\|$ rises and then saturates. Curves from different learning rates collapse well when plotted against τ.

and $\|\mathbf{V}(t) - \mathbf{V}(0)\|$ against τ on log-log axes. Dashed reference lines of slope $-1/3$ and $+1/3$ are drawn in panels a and b; panel c has no reference slope. As with SGD, loss decays and β_G grows approximately as $\tau^{1/3}$ power laws, and $\|\mathbf{V}(t) - \mathbf{V}(0)\|$ rises and saturates.

The G-I model under Adam reproduces the same phenomena as under SGD: approximate $1/3$ loss scaling and β_G growth driven by the gate. Together with the GH-I Adam results, this confirms that the $1/3$ scaling phenomena are not specific to SGD.

A.4 GH-II MODEL

The GH-I model (the main-text GH model) and its gate-only reduction established that any low-temperature softmax in the model can become the training bottleneck, producing $1/3$ time scaling in loss. The theoretical explanation (Section 3) hinges on loss universality and the $1/\beta_G$ loss scaling in the low-temperature regime, both of which depend on the non-linear property of the gate, not on its specific form. This raises the question of whether the same phenomena appear when the gate uses a different architecture. GH-II tests this with a modified gate, using the same head and training protocol as GH-I.

GH-II uses the same overall structure as GH-I (Eq. (1)), including the head layer and residual connection. The gate function $\mathrm{Gate}(\cdot)$ is modified: it now contains a key matrix $K \in \mathbb{R}^{d\times m}$, a gate matrix $G \in \mathbb{R}^{d\times m}$, and a value matrix $V \in \mathbb{R}^{m\times d}$, with output

$$\mathrm{Gate}(x^G) = V\big(Kx^G \odot \sigma(Gx^G)\big), \tag{14}$$

where $\odot$ denotes elementwise multiplication and σ is applied elementwise. Two variants are studied: $\sigma = \mathrm{Sigmoid}$ (`test-2-1.ipynb`) and $\sigma = \mathrm{Tanh}$ (`test-2-2.ipynb`), with initialization and learning rate details in the notebooks. The gate coldness is $\beta_G = \|G\|/\sqrt{d}$, and the teacher gate coldness β_G^* controls the scale of the teacher gate matrix G^*. The coldness β_G therefore controls the magnitude of Gx^G before the non-linearity, determining how saturated the outputs of σ will be. We use $m = 32$, $d = 96$, $n = 128$, and fix $\beta_G^* = 1000$. The training loss is $L_{\mathrm{KL}}^{\mathrm{GH}}$, the optimizer is SGD with a constant learning rate η, the dynamic time is $\tau = \eta t$, and we log the loss and β_G.

We run both GH-II variants in the regime of large gate coldness ($\beta_G^* = 1000$) and small head coldness ($\beta_H^* = 1$). Both the Sigmoid and Tanh variants exhibit $\tau^{-1/3}$ loss scaling (Figure 9, panels a and c) and $\tau^{1/3}$ gate coldness growth (panels b and d), reproducing the same phenomena as GH-I. The $1/3$ scaling is therefore more general, extending beyond the specific softmax non-linearity.

The Sigmoid results admit a direct theoretical interpretation: $\mathrm{Sigmoid}(y) = e^y/(e^y+1)$ is exactly the softmax probability of the more likely class in a binary problem, so each of the d gate components is a 2-class softmax. The theory of Section 3 therefore applies. With $\beta_G^* = 1000$ controlling the scale

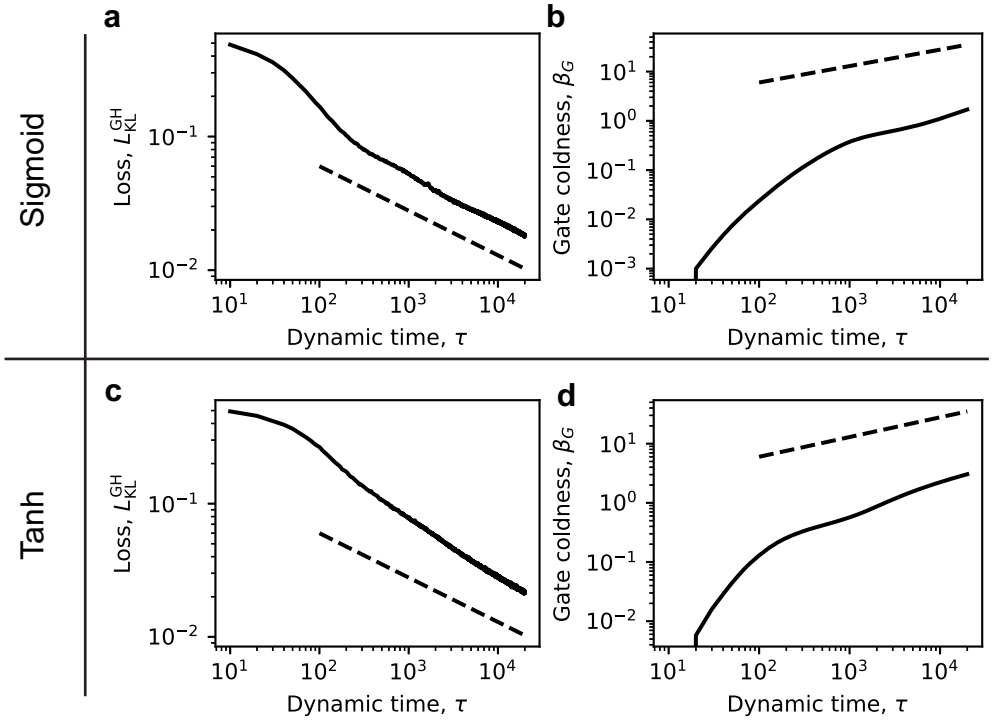

Figure 9: The GH-II model reproduces $1/3$ time scaling in loss and gate coldness for both sigmoid and tanh gate non-linearities ($\beta_G^* = 1000$, $\beta_H^* = 1$). Each pair of panels shows $L_{\mathrm{KL}}^{\mathrm{GH}}$ (left) and β_G (right) against τ on log-log axes. (a,b) Sigmoid variant. (c,d) Tanh variant. Dashed reference lines have slope $-1/3$ (panels a,c) and $+1/3$ (panels b,d).

of all rows of G^*, all d sigmoid gates are in the low-temperature regime. Once K and V converge, the overall loss is a reasonable function of the d gate activations $\mathrm{Sigmoid}(Gx^G)$. Loss universality then gives $L_{\mathrm{KL}}^{\mathrm{GH}} \sim 1/\beta_G$, which together with gradient flow dynamics yields $\tau^{-1/3}$ loss and $\tau^{1/3}$ gate coldness growth, as observed. The Tanh variant follows by the same argument: $\mathrm{Tanh}(y) = 2\mathrm{Sigmoid}(2y) - 1$ is a linear rescaling of sigmoid, so it shares the same saturation behavior as a function of β_G, and the low-temperature expansion and loss universality apply identically.

In summary, GH-II replaces the softmax gate of GH-I with sigmoid and tanh non-linearities and finds that both reproduce $1/3$ time scaling in loss and gate coldness. Theoretical analysis connects these results directly to the softmax case via the binary-softmax equivalence and loss universality. Taken together with GH-I, these three concrete examples (softmax, sigmoid, and tanh) consistently exhibit the same $1/3$ scaling across different gate architectures and non-linearities. This suggests a **non-linearity universality** analogous to loss universality: just as the $1/3$ scaling does not depend on the loss form, it also does not depend on the specific non-linearity or gate architecture, pointing toward a more general mathematical principle whose complete characterization is a direction for future work.

Notably, the GH-II gate is structurally a gated linear unit (GLU), which is widely adopted in modern LLMs as a feed-forward layer variant. If such a GLU gate operates in the low-temperature regime, it would become yet another training bottleneck, making the $1/3$ loss scaling harder to escape. The benefit of GLU over standard feed-forward layers may lie in reducing the irreducible loss through more expressive gating. However, convergence to that lower loss floor may still follow the same $1/3$ power law.

A.5 GH-III MODEL

GH-I and GH-II demonstrated the $1/3$ time scaling with gate architectures that, while realistic, are simpler than the full attention mechanism used in LLMs. GH-III asks whether the same phenomena persist when the gate is replaced by real attention, providing a more direct connection between the toy model and the original Transformers.

GH-III uses the same overall structure as GH-I (the main text GH model), including the head layer and residual connection, but takes $x \in \mathbb{R}^{m\times d}$ as input, where d is the context length ($d = 8$ in our experiments). The gate function $\mathrm{Gate}(\cdot)$ is replaced by single-head attention computing the output for the last token. The input to the gate layer is $x^G = \mathrm{Norm}(x) \in \mathbb{R}^{m\times d}$, and the residual connection and the head act on the last token only, so $h = \mathrm{Norm}(x_d) + \mathrm{Gate}(x^G) \in \mathbb{R}^m$, where x_d is the last token. A combined weight matrix $W^{QKV} \in \mathbb{R}^{3m\times m}$ projects each token of x^G into query, key, and value vectors, giving $Q = [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_d]$, $K = [\boldsymbol{k}_1, \ldots, \boldsymbol{k}_d]$, $V = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d] \in \mathbb{R}^{m\times d}$ (here K and V denote the stacked key and value vectors of the tokens, not the weight matrices of GH-I). The gate output is

$$\mathrm{Gate}(x^G) = V\,\mathrm{Softmax}\left(\frac{K^T\boldsymbol{q}_d}{\sqrt{m}}\right) \in \mathbb{R}^m, \tag{15}$$

where $\boldsymbol{q}_d$ is the last token's query. The gate coldness β_G is the standard deviation of the attention logits, averaged over the batch, and the teacher coldness is $\beta_G^* = 1000$. We use $m = 32$, $d = 8$, $n = 128$, $\beta_H^* = 1$, and train with AdamW with dynamic time $\tau = \eta t$. The loss and β_G are logged, with initialization and learning rate details in `test-3-3.ipynb`.

We run GH-III in the regime of large gate coldness ($\beta_G^* = 1000$) and small head coldness ($\beta_H^* = 1$). The loss exhibits $\tau^{-1/3}$ scaling (Figure 10a) and the gate coldness grows as $\tau^{1/3}$ (Figure 10b), reproducing the same phenomena as GH-I. The $1/3$ time scaling therefore persists with real attention as the gate.

The GH-III gate is conceptually more complex than GH-I: the attention logits involve both K and Q, and the d-dimensional context introduces additional structure absent in GH-I's single-vector input. These complexities are one reason GH-I, with its simpler gate, was chosen as the main-text model for theoretical analysis. However, once the student Q, K, V matrices are aligned with the teacher (ensured here by initialization), the extra structure does not affect the leading-order behavior. The loss reduces to a function of β_G alone, and the low-temperature expansion gives $L_{\mathrm{KL}}^{\mathrm{GH}} \sim 1/\beta_G$ by the same argument as in GH-I. If β_G then follows effective gradient flow dynamics, the same $\tau^{1/3}$ growth follows naturally, explaining the $1/3$ scaling in both β_G and loss.

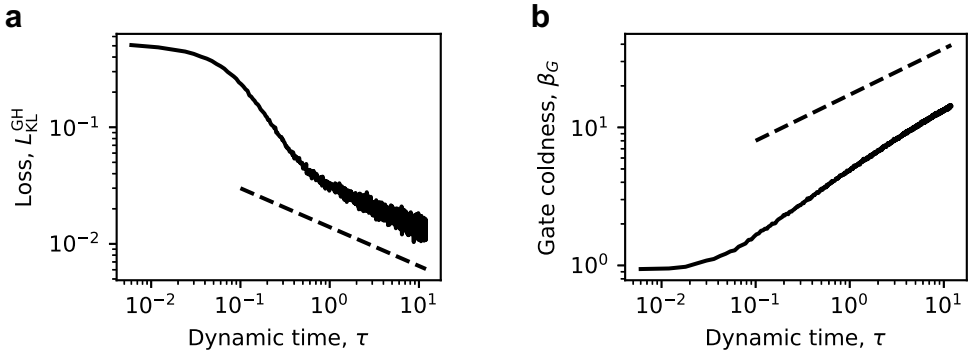

Figure 10: The GH-III model with real attention as the gate reproduces $1/3$ time scaling in loss and gate coldness ($\beta_G^* = 1000$, $\beta_H^* = 1$). (a) The loss $L_{\mathrm{KL}}^{\mathrm{GH}}$ follows $\tau^{-1/3}$ scaling. (b) The gate coldness β_G grows as $\tau^{1/3}$. Dashed reference lines have slope $-1/3$ (panel a) and $+1/3$ (panel b).

In summary, GH-III confirms that the $1/3$ time scaling persists when the gate is replaced by real attention, with the theoretical argument carrying over through the aligned-student assumption. GH-I was presented in the main text because its simpler gate makes the essential mechanism (a low-temperature softmax as a training bottleneck) transparent and analytically tractable. Once this mechanism is understood from GH-I, the extension to GH-III follows naturally, and GH-III then provides the direct experimental link between the toy model and real Transformer attention. The resulting logic chain, from GH-I to GH-III to the LLM experiments in the main text, forms a coherent and increasingly realistic case that attention trying to concentrate is at the origin of the emergent $1/3$ scaling law.

We also note a structural connection between GH-I (main text GH model) and MoE architectures: the GH-I gate uses $q^G = \mathrm{Softmax}(Kx^G)$ to weight d value matrices, in the same way MoE routing selects among experts. If the routing distribution becomes peaked as the model learns to specialize experts, our analysis predicts that the routing gate will become a training bottleneck with a $1/3$ time-scaling contribution to the loss. As with GLU (Appendix A.4), the benefit of MoE is increased model capacity and a lower achievable loss. However, convergence to that lower loss floor may still follow the same $1/3$ power law. This supports the prediction in Section 6 that MoE architectures are likely to retain the $1/3$ scaling exponent.

B Theory

In this section, we provide full details of our theoretical derivations. Appendices B.1, B.2, and B.3 elaborate on the gate dynamics theory in Section 3. Appendix B.4 generalizes the setup and proves loss universality.

B.1 Gradient flow

In Appendices B.1, B.2, and B.3, we analyze the G-I model (gate layer of GH-I or the main text GH model) with all notations used identical to the main text. As described in Section 3, we focus on the training dynamics of the key matrix K using MSE between gate probabilities, $L_{\mathrm{Sp}}^{\mathrm{G}}$, as loss. Essentially, our analysis is on a linear layer followed by a softmax non-linearity, similar to Liu et al. (2026b) and Kühn et al. (2026).

The gradient descent update of the student K is given by

$$K_{t+1} = K_t - \eta_t \nabla_K L_{\mathrm{Sp}}^{\mathrm{G}}, \tag{16}$$

where ∇_K takes the gradient with respect to K. In the continuous limit of small η_t, with $\tau(t) = \sum_{t'=1}^{t} \eta_{t'} \approx \int_0^t \eta_{t'} \mathrm{d}t'$, the step η_t plays the role of $\mathrm{d}\tau$, and we obtain the gradient flow dynamics

$$\frac{\mathrm{d}K}{\mathrm{d}\tau} = -\nabla_K L_{\mathrm{Sp}}^{\mathrm{G}}. \tag{17}$$

B.2 ALIGNED STUDENT

As explained in Liu et al. (2026b), when the norm is small, the student matrix K rapidly aligns with the teacher, while the late-time dynamics are dominated by norm growth. We therefore consider the simplified aligned student $K(\tau) = \frac{1}{\sqrt{m}}\hat{K}\beta_G(\tau)$ under gradient flow dynamics, where $K^* = \frac{1}{\sqrt{m}}\hat{K}\beta_G^*$ was defined in Eq. (6). To derive the effective dynamics of β_G, we substitute this aligned K into the gradient flow of K. We then have

$$\frac{\hat{K}}{\sqrt{m}}\frac{\mathrm{d}\beta_G}{\mathrm{d}\tau} = -\nabla_K L_{\mathrm{Sp}}^{\mathrm{G}}, \tag{18}$$

and then

$$\frac{\mathrm{Tr}(\hat{K}\hat{K}^T)}{m}\frac{\mathrm{d}\beta_G}{\mathrm{d}\tau} = -\mathrm{Tr}\left(\nabla_K L_{\mathrm{Sp}}^{\mathrm{G}}\frac{\hat{K}^T}{\sqrt{m}}\right). \tag{19}$$

Noting that $\frac{\mathrm{d}K}{\mathrm{d}\beta_G} = \frac{\hat{K}}{\sqrt{m}}$, the right side of the above equation is $\nabla_{\beta_G} L_{\mathrm{Sp}}^{\mathrm{G}}$,

$$\frac{\mathrm{Tr}(\hat{K}\hat{K}^T)}{m}\frac{\mathrm{d}\beta_G}{\mathrm{d}\tau} = -\nabla_{\beta_G} L_{\mathrm{Sp}}^{\mathrm{G}}. \tag{20}$$

We consider large d and m, such that $\mathrm{Tr}(\hat{K}\hat{K}^T) \approx md$. We therefore reach

$$\frac{\mathrm{d}\beta_G}{\mathrm{d}\tau} = -\frac{1}{d}\frac{\mathrm{d}L_{\mathrm{Sp}}^{\mathrm{G}}(\beta_G)}{\mathrm{d}\beta_G}. \tag{21}$$

B.3 LOW TEMPERATURE EXPANSION

The key is to understand $L_{\mathrm{Sp}}^{\mathrm{G}}(\beta_G)$ now. We study the limit of large β_G and $\beta_G^* \gg \beta_G$ (low temperatures). Recall that

$$L_{\mathrm{Sp}}^{\mathrm{G}} = \langle \|q^G - p^G\|^2/d\rangle = \Big\langle \frac{1}{d}\sum_{i=1}^{d}(q_i^G - p_i^G)^2\Big\rangle, \tag{22}$$

where $\langle\cdot\rangle$ means averaging over the data distribution. The student's gate probability distribution is

$$q_i^G = \frac{e^{y_{G,i}}}{\sum_j e^{y_{G,j}}} = \frac{e^{-\beta_G\epsilon_{G,i}}}{\sum_j e^{-\beta_G\epsilon_{G,j}}}, \tag{23}$$

and the teacher's probability distribution is

$$p_i^G = \frac{e^{y^*_{G,i}}}{\sum_j e^{y^*_{G,j}}} = \frac{e^{-\beta_G^*\epsilon_{G,i}}}{\sum_j e^{-\beta_G^*\epsilon_{G,j}}}. \tag{24}$$

Here, $y_G = Kx^G \in \mathbb{R}^d$ are the gate logits, with $x^G \in \mathbb{R}^m$ as the gate input. Energies $\epsilon_G = -y_G/\beta_G$ are i.i.d. standard normal. Since the teacher and student are aligned, their energies are the same, and only the standard deviations β_G and β_G^* differ.

For each $\{\epsilon_{G,i}\}_{i=1}^d$ or $\epsilon_G = [\epsilon_{G,1}, ..., \epsilon_{G,d}]^T$, we use $\epsilon_G^{\uparrow}$ for the corresponding sorted energies, such that $\min_i \epsilon_{G,i} = \epsilon_{G,1}^{\uparrow} \le \epsilon_{G,2}^{\uparrow} \le \cdots \le \epsilon_{G,d}^{\uparrow}$. We can define energy gaps as $\Delta\epsilon_{G,i} = \epsilon_{G,i}^{\uparrow} - \epsilon_{G,1}^{\uparrow}$. In particular, we define

$$\Delta\epsilon_G \equiv \Delta\epsilon_{G,2}. \tag{25}$$

With these preparations, we can write

$$\sum_i (q_i^G)^2 = \frac{1 + e^{-2\beta_G\Delta\epsilon_G} + \sum_{i=3}^d e^{-2\beta_G\Delta\epsilon_{G,i}}}{(1 + e^{-\beta_G\Delta\epsilon_G} + \sum_{i=3}^d e^{-\beta_G\Delta\epsilon_{G,i}})^2}. \tag{26}$$

The condition $\beta_G^* \gg \beta_G$ effectively means any term involving β_G^* is higher-order, and we can use $\beta_G^* = \infty$ to only keep the leading terms. We then have $\sum_i (p_i^G)^2 = 1$ and

$$\sum_i p_i^G q_i^G = \frac{1}{1 + e^{-\beta_G\Delta\epsilon_G} + \sum_{i=3}^d e^{-\beta_G\Delta\epsilon_{G,i}}}. \tag{27}$$

The loss can then be written as

$$L_{\mathrm{Sp}}^{\mathrm{G}} = \frac{1}{d}\Big\langle \frac{2e^{-2\beta_G\Delta\epsilon_G}(1+O(e^{-\beta_G(\epsilon^{\uparrow}_{G,3}-\epsilon^{\uparrow}_{G,2})}))}{(1+e^{-\beta_G\Delta\epsilon_G}+\sum_{i=3}^{d}e^{-\beta_G\Delta\epsilon_{G,i}})^2}\Big\rangle \tag{28}$$

$$= \frac{1}{d}\Big\langle \frac{2e^{-2\beta_G\Delta\epsilon_G}}{(1+e^{-\beta_G\Delta\epsilon_G})^2}\Big\rangle + \frac{1}{d}\Big\langle \frac{2e^{-2\beta_G\Delta\epsilon_G}}{(1+e^{-\beta_G\Delta\epsilon_G})^2}O(e^{-\beta_G(\epsilon^{\uparrow}_{G,3}-\epsilon^{\uparrow}_{G,2})})\Big\rangle. \tag{29}$$

The first term in Eq. (29) is the leading term, such that in terms of the scaling with β_G, the second term can only be the same or faster. To find the leading scaling with β_G, it suffices to study the scaling of the first term. As in the main text,

$$\left\langle \frac{2e^{-2\beta_G\Delta\epsilon_G}}{(1+e^{-\beta_G\Delta\epsilon_G})^2}\right\rangle = \int_0^\infty \rho_{\Delta\epsilon_G}(\Delta\epsilon_G)\mathrm{d}\Delta\epsilon_G \frac{2e^{-2\beta_G\Delta\epsilon_G}}{(1+e^{-\beta_G\Delta\epsilon_G})^2}. \tag{30}$$

For sufficiently **complex and diverse** data, such that the inputs to K vary continuously over the space, we expect $\rho_{\Delta\epsilon_G}(0)\neq 0$, which leads to the low-temperature expansion

$$\int_0^\infty \rho_{\Delta\epsilon_G}(\Delta\epsilon_G)\mathrm{d}\Delta\epsilon_G \frac{2e^{-2\beta_G\Delta\epsilon_G}}{(1+e^{-\beta_G\Delta\epsilon_G})^2} = \int_0^\infty \rho_{\Delta\epsilon_G}\Big(\frac{z}{\beta_G}\Big)\frac{\mathrm{d}z}{\beta_G}\frac{2e^{-2z}}{(1+e^{-z})^2} \tag{31}$$

$$= \frac{\rho_{\Delta\epsilon_G}(0)}{\beta_G}\int_0^\infty \mathrm{d}z \frac{2e^{-2z}}{(1+e^{-z})^2} + o\left(\frac{1}{\beta_G}\right). \tag{32}$$

In the specific setup of our toy model, where energies are i.i.d. standard normal, the gap distributions can be obtained via the limiting Poisson point process for extreme order statistics in the Gumbel class (Resnick, 1987). For $\Delta\epsilon_G$, $\rho_{\Delta\epsilon_G}(\cdot)$ is approximately an exponential distribution with rate $\sqrt{2\ln d}$. The first term in Eq. (32) therefore dominates when $\beta_G \gg \sqrt{2\ln d}$, making $\sqrt{2\ln d}$ a rough boundary between the high- and low-temperature regimes.[1]

Finally, we can have the loss as

$$L_{\mathrm{Sp}}^{\mathrm{G}} = \frac{1}{d}\left(\frac{\rho_{\Delta\epsilon_G}(0)}{\beta_G}\int_0^\infty \mathrm{d}z\frac{2e^{-2z}}{(1+e^{-z})^2} + o\left(\frac{1}{\beta_G}\right)\right)(1+O(1)), \tag{33}$$

where the $O(1)$ comes from the second term of Eq. (29). The second term in Eq. (29) is actually $O(\langle e^{-\beta_G\Delta\epsilon_{G,3}}\rangle)$. With extra information from extreme order statistics of i.i.d. standard normal random variables, $\rho_{\Delta\epsilon_{G,k}}(\Delta\epsilon_{G,k}) \sim (\Delta\epsilon_{G,k})^{k-2}$, we can obtain $O(\langle e^{-\beta_G\Delta\epsilon_{G,3}}\rangle) = O(1/\beta_G^2)$, similar to what we did for the first term of Eq. (29). In our specific setup, we can further tighten the loss expression by knowing $O(1)$ above as $O(1/\beta_G)$, yielding

$$L_{\mathrm{Sp}}^{\mathrm{G}} = \frac{\rho_{\Delta\epsilon_G}(0)}{\beta_G d}\int_0^\infty \mathrm{d}z\frac{2e^{-2z}}{(1+e^{-z})^2} + o\left(\frac{1}{\beta_G}\right). \tag{34}$$

With or without the extra information about the extreme value statistics, once $\rho_{\Delta\epsilon_G}(0)\neq 0$, we have the scaling for large β_G:

$$L_{\mathrm{Sp}}^{\mathrm{G}} = \Theta\left(\frac{1}{\beta_G}\right). \tag{35}$$

Combined with the effective gradient flow dynamics of β_G, i.e., Eq. (21), we have

$$\beta_G = \Theta(\tau^{1/3}),\ \text{thus}\ L_{\mathrm{Sp}}^{\mathrm{G}} = \Theta(\tau^{-1/3}). \tag{36}$$

The power-law result above is correct when $\beta_G^* \gg \beta_G$. When β_G approaches β_G^*, one can always expand the loss around β_G^*. The leading term in the loss should then be $\Theta((\beta_G-\beta_G^*)^2)$, where the coefficient is related to the Hessian. We therefore have exponential convergence. When β_G^* is large, the full picture is then: (i) when β_G is small, there is no power-law loss; (ii) when β_G is large while still much smaller than β_G^*, loss enters the $1/3$ power-law regime; (iii) when β_G approaches β_G^*, loss will deviate from the power law and converge exponentially. If $\beta_G^*=\infty$, the $1/3$ scaling will last forever. In Liu et al. (2026b), the $1/3$ power-law regime is called the "intermediate regime".

[1] With the same argument, for the head layer, where the output dimension is n, the rough boundary between the high- and low-temperature regimes is given by $\beta_H = \sqrt{2\ln n}$.

We focus on dynamics at low temperatures in this paper. On the other hand, when both β_G^* and β_G are small, the softmax function can be expanded with respect to the logits directly, i.e., the high-temperature expansion (Liu et al., 2026b). The leading terms in the high-temperature expansion are linear in logits, and the model is therefore effectively linear. In our specific setup, where the inputs are isotropic, the loss converges exponentially in the high-temperature regime.

B.4 Loss universality

In the above appendices, we analyzed a single softmax inside the model. We find that training such a softmax to learn peaked distributions will lead to a part of the loss scaling as $\tau^{-1/3}$. If there are more softmax functions operating at low temperatures, each of them can contribute a $\tau^{-1/3}$ term to the late-time loss. The interaction between softmax functions is higher-order if each is already close to being optimal. The late-time loss therefore scales as $\tau^{-1/3}$ as long as there exist softmax functions approaching peaked distributions.

How the final loss specifically depends on the softmax outputs is affected by the position of the softmax. For the final softmax, the final loss is the KL divergence of output probabilities, which shows $1/3$ scaling (Liu et al., 2026b; Kühn et al., 2026). For intermediate softmax functions, the final loss is more similar to MSE between their outputs (after later layers converge), which also scales as $\tau^{-1/3}$. These examples therefore suggest that $1/3$ time scaling is not specific to how the loss depends on softmax probabilities.

Here, we solidify intuitions into a theorem. We will first introduce the setup and notations valid and self-consistent only in this Appendix B.4.

Definition 1 (Problem setup)**.** *Consider a trainable student module $q = \mathrm{Softmax}(y(x)) \in \mathbb{R}^d$ with logits y depending on model input x. It tries to approach a teacher $p = \mathrm{Softmax}(y^*(x)) \in \mathbb{R}^d$ where $y^*(\cdot)$ is a fixed mapping by minimizing loss $L = \langle D(q,p)\rangle$. Here, $D(\cdot,\cdot)$ is a measure of difference and $\langle\cdot\rangle$ denotes averaging over the data distribution. Without loss of generality, we set $\sum_i y_i = 0$ and $\sum_i y_i^* = 0$.*

This setup is general. If we imagine the module is the final softmax and $D(\cdot,\cdot)$ is KL divergence, it can describe toy models in Liu et al. (2026b). And if we imagine the module as an intermediate softmax, once all later layers converge, the loss only depends on this softmax ($D(\cdot,\cdot)$ will depend on parameters of the later layers and cannot be written as a simple function), which is the gate-layer training that the current paper focuses on.

We next introduce the assumptions facilitating further analysis. As in the main text, we define coldness β as the standard deviation of y and β^* as the standard deviation of y^*. The student is often aligned in early training:

Assumption 1 (Aligned student)**.** *Under the setup (Definition 1), the student has $y \propto y^*$ for all inputs, such that the loss only depends on β and β^*.*

We also define energies $\epsilon = -y/\beta = -y^*/\beta^*$ under aligned student. If the logits are obtained by multiplying a hidden state by a matrix, and the data is complex enough such that hidden states are distributed continuously across the hidden space, the probability density of hidden states on decision boundaries of output classes will be non-zero, which inspires the following generic assumption.

Assumption 2 (Complex and diverse data)**.** *Under the setup (Definition 1) and with the aligned student (Assumption 1), for each $\{\epsilon_i\}_{i=1}^d$ or $\epsilon = [\epsilon_1, ..., \epsilon_d]^T$, we use $\epsilon^\uparrow$ for the corresponding sorted energies, such that $\min_i \epsilon_i = \epsilon_1^\uparrow \le \epsilon_2^\uparrow \le \cdots \le \epsilon_d^\uparrow$. We write energy gaps as $\Delta\epsilon_i = \epsilon_i^\uparrow - \epsilon_1^\uparrow \ge 0$ and denote $\Delta\epsilon \equiv \Delta\epsilon_2$. The data distribution affects the probability density $\rho_{\Delta\epsilon_i}(\cdot)$ of gap $\Delta\epsilon_i$. Data are sufficiently complex and diverse such that $\rho_{\Delta\epsilon}$ is continuous near zero and $\rho_{\Delta\epsilon}(0) \neq 0$.*

We next explain the constraints on the loss functions.

Definition 2 (Reasonable loss)**.** *Under the setup (Definition 1), the loss function $L = \langle D(q,p)\rangle$ is called reasonable if the following conditions are satisfied:*

• (Regular) $D(q,p)$ is permutation invariant (i.e., exchanging any two coordinates of q and p does not change $D(q,p)$), non-negative, and vanishes if and only if the two distributions agree. $D(q,p)$ remains finite and continuous when the teacher p is a one-hot distribution for an aligned student.

• *(Differentiable) $D(q,p)$ is Lipschitz continuous in q on the probability simplex for any p when the student is aligned. $D(q,p)$ is differentiable in q on the probability simplex excluding the one-hot vertices for any p when the student is aligned. The loss L is differentiable in model parameters.*

There are several remarks. The requirement that $D(q,p)=0$ iff $q=p$ is not necessary but convenient. For example, cross-entropy converges to the entropy of p when $q=p$, which may not be zero. But it is easy to define cross-entropy minus the entropy of p, i.e., the KL divergence, to satisfy the condition. The behaviors of how cross-entropy converges to its final value (the entropy) and how KL divergence converges to zero are the same. And we only care about how the reducible part of the loss vanishes. So, it is convenient to set $D(q,p)=0$ iff $q=p$. But any $D(q,p)$ that can be modified to satisfy this condition, like cross-entropy, is also "reasonable". We think the constraints define reasonable losses as they basically require the loss to be finite and to have finite gradients, which are desired by healthy gradient-based optimization. It is straightforward to check that commonly used MSE, KL divergence, and polynomials of $\|p-q\|^2$ are reasonable.

With all the preparations, we can now state the claim.

Theorem 1 (Loss universality)**.** *Under the setup (Definition 1), assuming aligned student (Assumption 1) and complex data (Assumption 2), any reasonable loss L (Definition 2) satisfies $L=\Theta(\beta^{-1})$ and $\nabla_\beta L=\Theta(\beta^{-2})$ for sufficiently large β and $\beta^*\gg\beta$.*

The result $L=\Theta(\beta^{-1})$ is the key to $1/3$ time scaling. If β follows gradient flow effectively, the gradient is $\Theta(\beta^{-2})$ and time integral yields $\beta=\Theta(\tau^{1/3})$, which leads to $L=\Theta(\tau^{-1/3})$.

Proof. We next prove Theorem 1. Introducing

$$q^\uparrow=\mathrm{Softmax}(-\beta\epsilon^\uparrow)\text{ and }p^\uparrow=\mathrm{Softmax}(-\beta^*\epsilon^\uparrow) \tag{37}$$

as the sorted student and teacher probabilities, respectively, we have

$$D(q,p)=D(q^\uparrow,p^\uparrow). \tag{38}$$

We denote

$$q_z^\circ=\left(\frac{1}{1+e^{-z}},\frac{e^{-z}}{1+e^{-z}},0,\dots,0\right)^T\text{ and }\boldsymbol{e}_1=(1,0,0,\dots,0)^T. \tag{39}$$

The condition $\beta^*\gg\beta$ is equivalent to substituting $\beta^*=\infty$, ignoring high-order terms about β^*.

$$D(q,p)=D(q^\uparrow,\boldsymbol{e}_1)=D(q^\circ_{\beta\Delta\epsilon},\boldsymbol{e}_1)+(D(q^\uparrow,\boldsymbol{e}_1)-D(q^\circ_{\beta\Delta\epsilon},\boldsymbol{e}_1)). \tag{40}$$

Since $D(\cdot,\cdot)$ is differentiable,

$$\frac{D(q^\uparrow,\boldsymbol{e}_1)-D(q^\circ_{\beta\Delta\epsilon},\boldsymbol{e}_1)}{D(\boldsymbol{e}_1,\boldsymbol{e}_1)-D(q^\circ_{\beta\Delta\epsilon},\boldsymbol{e}_1)}=O(e^{-\beta(\epsilon_3^\uparrow-\epsilon_2^\uparrow)})=O(1). \tag{41}$$

The $O(1)$ bound is very loose, including the worst case that $\epsilon_3^\uparrow-\epsilon_2^\uparrow=0$. It is then safe to write

$$L=\langle D(q^\uparrow,\boldsymbol{e}_1)\rangle=\langle D(q^\circ_{\beta\Delta\epsilon},\boldsymbol{e}_1)\rangle(1+O(1)). \tag{42}$$

It suffices to study the scaling of $\langle D(q^\circ_{\beta\Delta\epsilon},\boldsymbol{e}_1)\rangle$.

Averaging data is equivalent to averaging over $\Delta\epsilon$ for $D(q^\circ_{\beta\Delta\epsilon},\boldsymbol{e}_1)$,

$$\langle D(q^\circ_{\beta\Delta\epsilon},\boldsymbol{e}_1)\rangle=\int_0^\infty D(q^\circ_{\beta\Delta\epsilon},\boldsymbol{e}_1)\rho_{\Delta\epsilon}(\Delta\epsilon)\mathrm{d}\Delta\epsilon \tag{43}$$

$$=\int_0^\infty D(q_z^\circ,\boldsymbol{e}_1)\rho_{\Delta\epsilon}(z/\beta)\frac{\mathrm{d}z}{\beta}. \tag{44}$$

We next need to study the property of $\lambda(z)\equiv D(q_z^\circ,\boldsymbol{e}_1)$.[2] Since $D(\cdot,\cdot)$ is Lipschitz, we know that there exists some constant C_λ such that

$$\lambda(z)=D(q_z^\circ,\boldsymbol{e}_1)\le C_\lambda e^{-z}. \tag{45}$$

[2]For KL divergence, $\lambda(z)=\ln(1+e^{-z})$; for MSE, $\lambda(z)=\frac{2e^{-2z}}{(1+e^{-z})^2d}$.

This guarantees that

$$\int_0^\infty \lambda(z)\mathrm{d}z < \infty. \tag{46}$$

It is therefore valid to write

$$\int_0^\infty D(q_z^\circ, \boldsymbol{e}_1)\rho_{\Delta\epsilon}(z/\beta)\frac{\mathrm{d}z}{\beta} = \int_0^\infty \lambda(z)(\rho_{\Delta\epsilon}(0) + o(1))\frac{\mathrm{d}z}{\beta} \tag{47}$$

$$= \frac{\rho_{\Delta\epsilon}(0)}{\beta}\int_0^\infty \lambda(z)\mathrm{d}z + o\left(\frac{1}{\beta}\right), \tag{48}$$

where Assumption 2 is also used. We can obtain

$$L = \left(\frac{\rho_{\Delta\epsilon}(0)}{\beta}\int_0^\infty \lambda(z)\mathrm{d}z + o\left(\frac{1}{\beta}\right)\right)(1 + O(1)), \tag{49}$$

which is equivalent to

$$L = \Theta(\beta^{-1}),\ \beta \to \infty\ \&\ \beta/\beta^* \to 0. \tag{50}$$

For an arbitrary function that is $\Theta(\beta^{-1})$, its derivative is not necessarily $\Theta(\beta^{-2})$ as it can be very oscillatory. But $\nabla_\beta L = \Theta(\beta^{-2})$ as will be shown. Intuitively, fast decay of $\lambda(z)$ as a result of Lipschitz continuity of $D(\cdot,\cdot)$ and continuity of $\rho_{\Delta\epsilon}$ prevents strong oscillation of L. Formally,

$$|\lambda'(z)| \le C_\lambda e^{-z}, \tag{51}$$

and

$$\int_0^\infty z\lambda'(z)\mathrm{d}z < \infty. \tag{52}$$

Integrating by parts, we have

$$\int_0^\infty z\lambda'(z)\mathrm{d}z = -\int_0^\infty \lambda(z)\mathrm{d}z. \tag{53}$$

We can then have

$$\nabla_\beta\langle D(q^\circ_{\beta\Delta\epsilon}, \boldsymbol{e}_1)\rangle = \int_0^\infty \nabla_\beta\lambda(\beta\Delta\epsilon)\rho_{\Delta\epsilon}(\Delta\epsilon)\mathrm{d}\Delta\epsilon \tag{54}$$

$$= \int_0^\infty \lambda'(z)z\rho_{\Delta\epsilon}(z/\beta)\frac{\mathrm{d}z}{\beta^2} \tag{55}$$

$$= -\frac{\rho_{\Delta\epsilon}(0)}{\beta^2}\int_0^\infty \lambda(z)\mathrm{d}z + o\left(\frac{1}{\beta^2}\right), \tag{56}$$

which leads to $\nabla_\beta L = \Theta(\beta^{-2}),\ \beta \to \infty\ \&\ \beta/\beta^* \to 0.$ □

C LLM EXPERIMENTS

C.1 EVALUATION OF OPEN-SOURCE MODELS

C.1.1 EVALUATION METHODS

We evaluate the Pythia (Biderman et al., 2023) and OLMo-2 (OLMo et al., 2024) model families, using text from the Pile (Gao et al., 2020) and FineWeb (Penedo et al., 2024) for Pythia and from FineWeb for OLMo-2. For each family, we load checkpoints at training steps that are approximately logarithmically spaced, spanning early training to the final checkpoint. The same procedures are applied uniformly across all model sizes. The checkpoint steps, listed from early to late for each model, are specified in the evaluation scripts referenced below, and the specific models are reported alongside the results in Appendix C.1.3. Three quantities are measured at each checkpoint, described in turn below.

Loss and LM head logits. We compute the cross-entropy loss and the LM head logit coldness at each checkpoint (`pythia-logit-1.py` for Pythia and `olmo-logit-0.py` for OLMo-2). For each checkpoint, we perform a forward pass on batches of text tokenized to a context length of 1024, evaluating on approximately 5×10^6 tokens in total. Valid token positions are those that are non-padding and have a non-padding successor (since the loss is next-token prediction). The cross-entropy loss is averaged over valid tokens. The LM head logit coldness is the standard deviation of the pre-softmax LM head logits over the vocabulary, averaged over valid token positions.

Attention logit coldness and entropy. We compute the per-head pre-softmax attention score standard deviation and softmax entropy across training checkpoints (`eval_attn_logit_std.py` for Pythia and `olmo/eval_attn_logit_std.py` for OLMo-2). For each checkpoint, we evaluate on approximately 2×10^5 tokens with context length 2048. For each attention layer and head, we extract the pre-softmax attention scores $s_{ij} = (\boldsymbol{q}_i \cdot \boldsymbol{k}_j)/\sqrt{d_{\text{head}}}$ with rotary embeddings applied, where i and j denote query and key positions, respectively. Causal masking restricts each query position i to the $i+1$ valid key positions $0, 1, \ldots, i$. For each query position $i \geq 1$, we compute the unbiased standard deviation of the scores over its valid key positions, then average over query positions and evaluation batches to obtain the per-head logit coldness. We also compute the entropy (in nats) of the resulting softmax attention distribution, averaged in the same way.

Energy gap distribution. We compute the energy gap for both attention heads and the LM head at each checkpoint (`eval_energy_gap.py` for Pythia and `olmo/eval_energy_gap.py` for OLMo-2). For a logit vector y, the energy gap is

$$\Delta\epsilon = \frac{y_{(1)} - y_{(2)}}{\text{std}(y)}, \tag{57}$$

where $y_{(1)}$ and $y_{(2)}$ are the largest and second-largest logits. For attention, y consists of the pre-softmax scores over valid causal key positions, using the same masking as above for $i \geq 1$. For the LM head, y is the logit vector over the full vocabulary. Gaps are accumulated as a fixed-bin histogram over approximately 2×10^6 tokens per checkpoint, which yields the distribution needed to check whether the density at $\Delta\epsilon \to 0$ is non-zero, as required by the theory (main text, Section 3).

C.1.2 Analysis methods

Dynamic time. The dynamic time $\tau(t) = \sum_{t'=1}^{t} \eta_{t'}$ (main text) is computed by numerically integrating the learning rate schedule. For Pythia, the schedule uses a linear warmup for the first $0.01\, t_{\max}$ steps, followed by a cosine decay:

$$\eta_{t'} = \eta_{\text{peak}}\big[0.9 \cdot \tfrac{1}{2}(\cos\theta_{t'} + 1) + 0.1\big], \quad \theta_{t'} = \frac{t' - 0.01\, t_{\max}}{0.99\, t_{\max}}\,\pi, \tag{58}$$

where $t_{\max}$ is the total number of training steps and η_{peak} varies by model size. For OLMo-2, the warmup covers the first $t_{\text{warm}} = 2000$ steps, followed by the same cosine form:

$$\eta_{t'} = \eta_{\text{peak}}\big[0.9 \cdot \tfrac{1}{2}(\cos\theta_{t'} + 1) + 0.1\big], \quad \theta_{t'} = \frac{t' - t_{\text{warm}}}{t_{\max} - t_{\text{warm}}}\,\pi. \tag{59}$$

In both cases, $\tau(t)$ is obtained by cumulative summation of $\eta_{t'}$ over steps $t' = 1, \ldots, t$, evaluated at the selected checkpoint steps. Since η_{peak} differs across model sizes, the resulting τ places all model sizes on a common dynamic-time axis, enabling the loss-curve collapse shown in Figure 4c.

Entropy threshold. Our theory predicts a rough boundary between the high- and low-temperature regimes for i.i.d. Gaussian logits: a softmax with n logits drawn from $\mathcal{N}(0, \beta^2)$ crosses the boundary at $\beta = \sqrt{2 \ln n}$ (Appendix B.3). Because LLM logits are not i.i.d. Gaussian, we convert this criterion into an entropy threshold and compare it directly with entropies of LLM distributions. For attention, the number of valid key positions grows from 1 to the context length as the query position advances. For each number of valid key positions i from 2 to 2047, we set $\beta = \sqrt{2 \ln i}$, draw i logits from $\mathcal{N}(0, \beta^2)$, and compute the entropy (in nats) of the resulting softmax distribution. The attention entropy threshold $\bar{H}_{\text{attn}}$ is the mean of these values over i, representing the average boundary entropy at the transition. For the LM head, we apply the same procedure with n equal to the vocabulary size, giving the threshold $\bar{H}_{\text{LM}}$. These two thresholds appear as horizontal dashed

lines in Figure 4a. For attention heads, we compare the measured entropy with $\bar{H}_{\mathrm{attn}}$, and heads below the threshold are in the low-temperature regime. For the LM head, we instead compare the target entropy of next-token prediction with $\bar{H}_{\mathrm{LM}}$: at convergence, the LM head must reproduce the next-token distribution of the data, whose entropy is bounded above by the fitted irreducible loss of 1.69 nats from the Chinchilla scaling laws (Hoffmann et al., 2022). This upper bound is the red diamond in Figure 4a and already lies below $\bar{H}_{\mathrm{LM}}$, so the target entropy is lower still, and the LM head must learn peaked distributions and is in the low-temperature regime. Results for Pythia-410M and Pythia-160M are shown in Figure 11.

Loss fitting. Following Liu et al. (2026b), we fit the raw loss at each model size independently to

$$L(\tau) = c_\tau \, \tau^{-\alpha_\tau} + L_{\setminus\tau}, \tag{60}$$

where c_τ, α_τ, and $L_{\setminus\tau}$ are free parameters (`pythia-logit-0to7-v3.ipynb`). Fitting is performed by nonlinear least squares on $\log L$ (rather than L) to reduce the influence of early-training checkpoints where the loss is large, using `scipy.optimize.curve_fit`. The uncertainties reported with α_τ in Figures 4c and 16c are standard errors, the square roots of the diagonal of the covariance matrix returned by the solver. Early checkpoints (typically the first one or two per model) are excluded from the fit because the model is still in the rapid alignment phase and has not yet entered the power-law regime. The fitted irreducible loss $L_{\setminus\tau}$ is subtracted from all checkpoints to yield the training-dependent component $L - L_{\setminus\tau}$, which is plotted in Figure 4c together with the fitted exponents α_τ. The raw loss and the full fitting procedure for all eight Pythia model sizes are shown in Figure 15.

Attention logit std and energy gap density. The per-head attention logit coldness (standard deviation of pre-softmax scores) is plotted against τ on logarithmic axes (`attn_logit_std_*.ipynb`). A reference line of slope $+1/3$ is overlaid for visual comparison. No fitting of the attention logit std is performed. Figure 4d shows selected heads from Pythia-12B (`attn_logit_std_12b.ipynb`): the logit coldness of certain attention heads grows approximately as $\tau^{1/3}$ while the LM head coldness saturates, identifying attention as the late-training bottleneck. Results for all heads across all layers of Pythia-12B are shown in Figure 13. For the energy gap, the raw histogram counts are divided by the product of the bin width and the total sample count to obtain a probability density estimate (`energy_gap_*.ipynb`). Figure 4b shows results from Pythia-12B (`energy_gap_12b.ipynb`): the density remains non-zero as $\Delta\epsilon \to 0$ for both a representative attention head and the LM head, confirming the theoretical condition (main text, Section 3). Results for all heads across all layers of Pythia-12B are shown in Figure 12. Energy gap and logit coldness results for Pythia-410M and Pythia-160M are shown in Figure 14. The OLMo-2-13B analysis, covering entropy (panel a), energy gap density (panel b), loss scaling (panel c), and attention logit coldness (panel d), is shown in Figure 16.

C.1.3 Additional results

Figure 11 extends the peaked distribution analysis of Figure 4a to Pythia-410M and Pythia-160M. The three models (160M, 410M, and 12B) were chosen because their depths span a wide range: 12, 24, and 36 layers, respectively. The per-head entropy at the final checkpoint is computed by the same procedure as in Appendix C.1.2 (entropy threshold paragraph), using `attn_logit_std_410m.ipynb` for Pythia-410M and `attn_logit_std.ipynb` for Pythia-160M. In both models, low-temperature attention heads (below the entropy threshold) exist across the depth of the network, and the upper bound of the LM head target entropy also falls below its threshold. On average, later layers tend to have lower entropy, but concentrated heads appear in early layers as well. Our theory requires only that at least one low-temperature softmax exists, and this condition is satisfied across all three model sizes, indicating that the peaked distribution pattern is not an artifact of model scale.

Figure 12 extends the energy gap analysis of Figure 4b from a single representative head to all attention heads across all 36 layers of Pythia-12B (`energy_gap_12b.ipynb`; same normalization procedure as Appendix C.1.2). For most heads and layers, the density is non-zero at $\Delta\epsilon \to 0$, indicating that the theoretical condition is not confined to a few exceptional heads. Some heads in early layers show density approaching zero at small $\Delta\epsilon$, consistent with their higher entropies in Figure 4a. This confirms that the energy gap condition required by the theory is broadly satisfied across the depth of Pythia-12B.

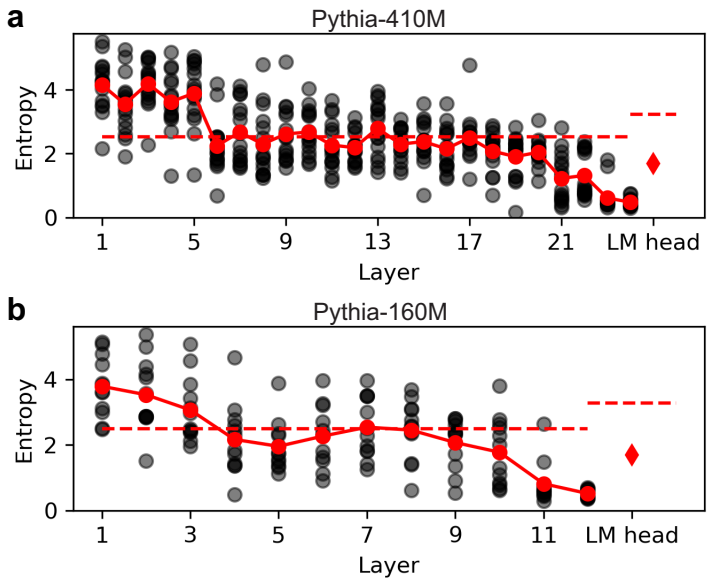

Figure 11: Later-layer attention heads and the LM head are in the low-temperature regime for Pythia-410M (a) and Pythia-160M (b). Same format as Figure 4a: black dots show per-head entropy at the final checkpoint, red dots show layer averages, and the red diamond shows the upper bound of the LM head target entropy (1.69 nats). Dashed lines mark the entropy thresholds for attention heads (left) and the LM head (right). The pattern is consistent with Pythia-12B in the main text.

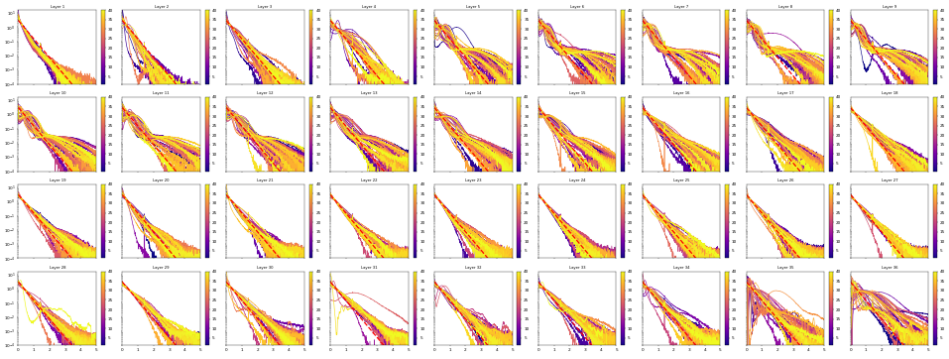

Figure 12: Energy gap density distributions for all attention heads across all 36 layers of Pythia-12B (final checkpoint), extending Figure 4b. Each panel corresponds to one layer; curves are colored by head index. For most heads across all layers, the density is non-zero at $\Delta\epsilon \to 0$, confirming the theoretical condition broadly. Heads with density approaching zero at small $\Delta\epsilon$ are in the high-temperature regime. Analysis follows Appendix C.1.2.

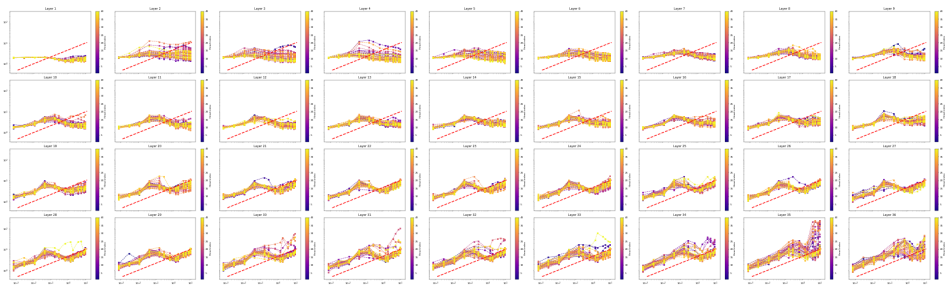

Figure 13: Attention logit coldness versus dynamic time τ for all attention heads across all 36 layers of Pythia-12B, extending Figure 4d. Each panel corresponds to one layer; curves are colored by head index. The dashed red line has slope $+1/3$ for reference. Low-temperature heads broadly track the $+1/3$ slope at late training, while high-temperature heads grow more slowly or saturate. Analysis follows Appendix C.1.2.

Figure 13 extends the logit coldness analysis of Figure 4d to all attention heads across all 36 layers of Pythia-12B (`attn_logit_std_12b.ipynb`; same procedure as Appendix C.1.2). Across the network, the low-temperature heads identified in Figure 4a broadly show logit coldness growing at approximately the $+1/3$ rate at late training, indicating that their logit magnitudes have not converged. This ongoing growth is precisely what the theory predicts for a bottleneck softmax: its logit magnitude must keep growing to track the teacher, producing $1/3$ time scaling in its loss contribution. The behavior is therefore not peculiar to the representative head shown in Figure 4d but is generic among low-temperature heads, further supporting attention as the late-training bottleneck.

Figure 14 replicates the energy gap and logit coldness analyses of Figure 4b and d for Pythia-410M (panels a–b) and Pythia-160M (panels c–d), using `energy_gap_410m.ipynb`, `energy_gap_160m.ipynb`, `attn_logit_std_410m.ipynb`, and `attn_logit_std.ipynb` with the same procedures as Appendix C.1.2. In both models, the energy gap density of representative attention heads is non-zero at $\Delta\epsilon \to 0$, and the logit coldness of attention heads grows approximately as $\tau^{1/3}$ while the LM head coldness saturates. These results reproduce the findings of Figure 4b and d across different model sizes, confirming that the behaviors are not specific to Pythia-12B.

Figure 15 illustrates the loss fitting procedure described in Appendix C.1.2 for all eight Pythia model sizes. The raw cross-entropy loss L (left panel) does not decay to zero as training progresses: each model converges toward a strictly positive floor, the irreducible loss $L_{\setminus\tau}$, which decreases with model size. Fitting L as a plain power law without this constant would therefore confound the training-dependent decay with model-size offsets and yield unreliable exponents. Instead, we fit $L(\tau) = c_\tau\, \tau^{-\alpha_\tau} + L_{\setminus\tau}$ with $L_{\setminus\tau}$ as a free parameter, which correctly separates the two contributions. After subtracting the fitted $L_{\setminus\tau}$ (right panel), the training-dependent component $L - L_{\setminus\tau}$ collapses across all eight model sizes onto a single power law close to $1/3$, validating the fitting procedure and confirming that the exponent is approximately size-independent.

Figure 16 replicates the full analysis of Figure 4 for OLMo-2, a different model family with a distinct architecture and training setup (`olmo/eval_attn_logit_std.py`, `olmo-step-tau.py`, and OLMo-2 variants of the analysis notebooks). Panel a shows the per-head entropy of OLMo-2-13B at the final checkpoint. Unlike Pythia, where many later-layer heads fall well below the entropy threshold, later-layer heads in OLMo-2-13B fall at most slightly below it. Like Pythia, however, the heads deepest in the low-temperature regime appear in early layers. We highlight one such head at layer 1 with a blue square: although the mean entropy trend still decreases with layer depth, it is the early layers that most clearly satisfy the low-temperature condition. This difference may reflect the distinct architectural choices of OLMo-2, including grouped-query attention, RMS normalization applied to query and key projections before RoPE, and larger peak learning rates compared to Pythia. Despite this architectural difference in which layers host the low-temperature heads, the theoretical conditions remain satisfied. Panel b shows that the energy gap density of the highlighted layer-1 head is non-zero at $\Delta\epsilon \to 0$, and panel d shows that its logit coldness grows as $\tau^{1/3}$ at late training, consistent with the theory's prediction for a bottleneck softmax. Based on the theory, the existence of such a head, regardless of its layer, is sufficient to drive $1/3$ scaling in

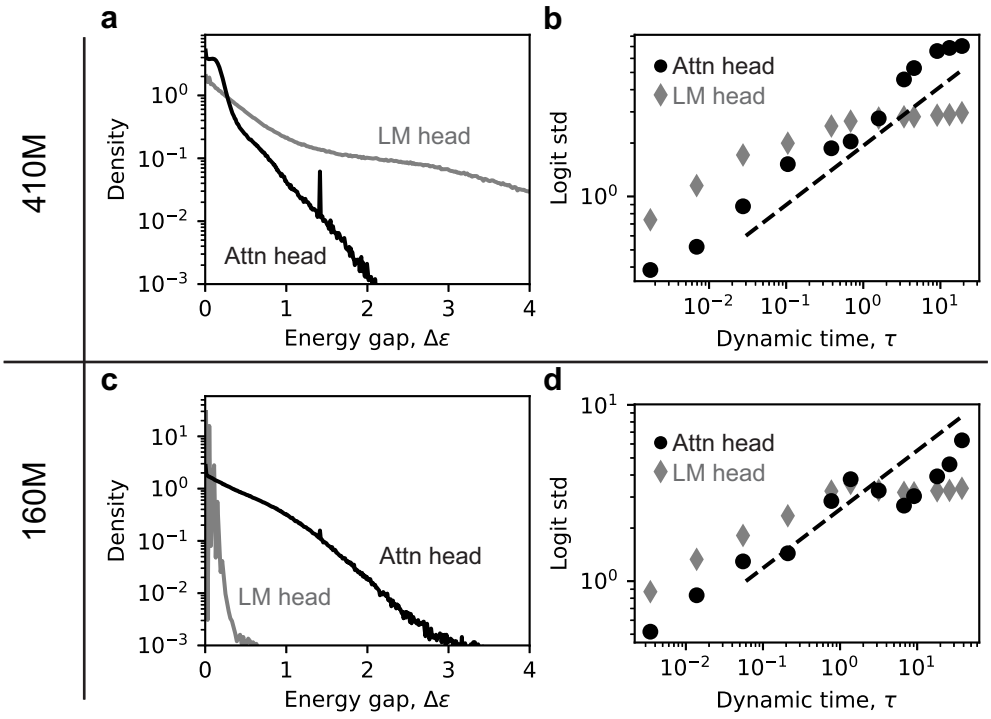

Figure 14: Energy gap density (left column) and attention logit coldness versus τ (right column) for Pythia-410M (a–b) and Pythia-160M (c–d), replicating Figure 4b and d. In each energy gap panel, black shows a representative attention head and gray shows the LM head; both have non-zero density at $\Delta\epsilon \to 0$. In each coldness panel, circles show an attention head and diamonds show the LM head; the dashed line has slope $+1/3$ for reference. Analysis follows Appendix C.1.2.

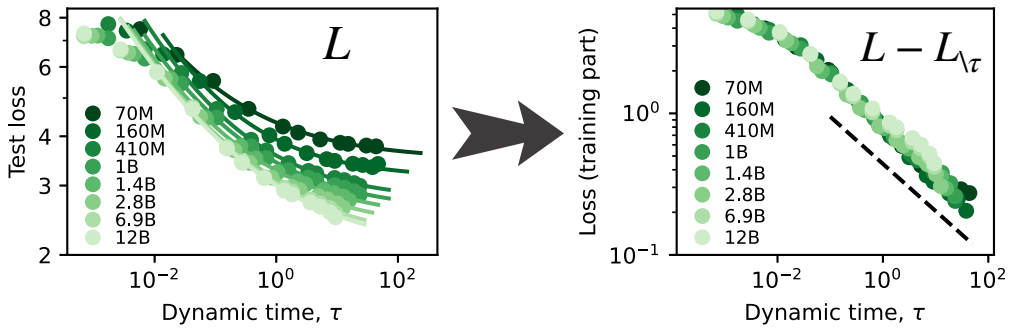

Figure 15: The raw cross-entropy loss and the fitting procedure for all eight Pythia model sizes (70M to 12B), extending Figure 4c. Left: raw loss L versus τ with power-law-plus-constant fits overlaid (solid lines). The loss converges to a strictly positive floor $L_{\setminus\tau}$ that decreases with model size, requiring the constant to be included in the fit. Right: training-dependent component $L - L_{\setminus\tau}$ after subtracting the fitted irreducible loss. Curves from all model sizes collapse and follow a single $\sim \tau^{-1/3}$ power law (dashed reference line). Analysis follows Appendix C.1.2.

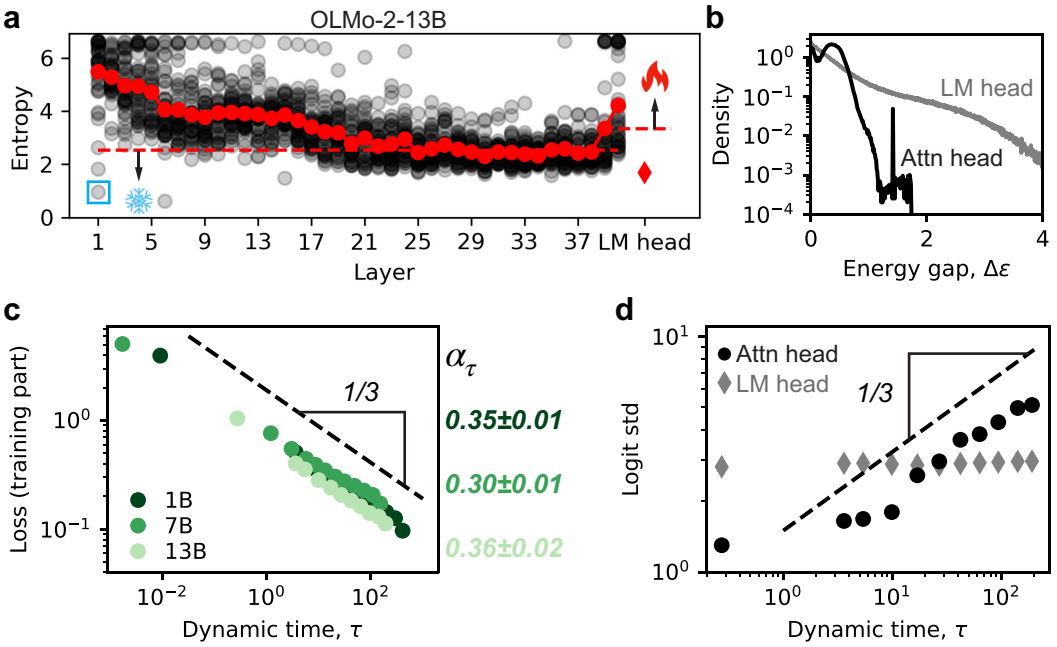

Figure 16: Replication of the Figure 4 analysis for OLMo-2, extending the results to a different model family. (a) Per-head entropy of OLMo-2-13B at the final checkpoint. The red diamond shows the upper bound of the LM head target entropy. Although the mean entropy decreases in later layers, later-layer heads fall at most slightly below the entropy threshold (horizontal dashed line), and the heads deepest in the low-temperature regime appear in early layers. The layer-1 head highlighted by a blue square is in the low-temperature regime. (b) Energy gap density for the highlighted layer-1 head; the density is non-zero at $\Delta\epsilon \to 0$. (c) Loss versus dynamic time τ for OLMo-2-1B, 7B, and 13B, with fitted exponents 0.35, 0.30, and 0.36. (d) Attention logit coldness of the highlighted layer-1 head grows as $\tau^{1/3}$ (dashed reference line), while the LM head saturates. Analysis follows Appendix C.1.2.

the total loss. Panel c confirms this prediction: the fitted loss exponents for OLMo-2-1B, 7B, and 13B are approximately 0.35, 0.30, and 0.36, all close to $1/3$. These results establish that the $1/3$ scaling law generalizes across model families, even though details like which head has the lowest temperature may differ.

C.2 Training with MSE

C.2.1 Training methods

To test whether the LM head is the sole bottleneck for the $1/3$ loss scaling, we train a Pythia-160M student using MSE between final logits, bypassing the output softmax entirely. The student is randomly initialized using the Pythia-160M architecture. The teacher is a pretrained Pythia-160M model at its final checkpoint (step 79,000), held frozen throughout. The training objective is MSE between mean-centered logits of the student and the teacher, averaged over tokens:

$$L_{\text{MSE}} = \left\langle \frac{1}{n} \|\tilde{y}_s - \tilde{y}_t\|^2 \right\rangle, \tag{61}$$

where $\tilde{y}_s = y_s - \bar{y}_s$ and $\tilde{y}_t = y_t - \bar{y}_t$ are the mean-centered logit vectors of the student and teacher ($\bar{y}$ denotes the vocabulary mean), n is the vocabulary size here, $\langle\cdot\rangle$ denotes averaging over tokens, and the teacher temperature is 1 (`loss.py`). Mean-centering makes the loss invariant to global logit shifts, which do not affect any softmax output and are therefore irrelevant to the distribution being learned.

The student is trained for 79,000 steps with peak learning rate $\eta_{\text{peak}} = 6 \times 10^{-4}$, linear warmup over 1430 steps, and cosine decay over 143,000 steps, the same schedule shape as Pythia (`train.py`).

Training uses the Pile (deduped) dataset with a batch size of 256 sequences of 2048 tokens (approximately 0.5×10^6 tokens per step) on 8 H200 GPUs. Checkpoints are saved at the same 12 steps as the Pythia evaluation in Appendix C.1.1: $128, 256, \ldots, 79{,}000$.

C.2.2 Analysis methods

Dynamic time. Dynamic time τ is computed by the same cumulative-sum procedure as Appendix C.1.2, using the Pythia learning rate schedule with $\eta_{\text{peak}} = 6 \times 10^{-4}$ and $t_{\max} = 143{,}000$.

Logit coldness. The per-head attention logit coldness is computed by the same procedure as in Appendix C.1.2 (`attn_logit_std_mse1.ipynb`). A representative attention head (layer 1, head 4) is selected for display in Figure 5a as filled circles. The LM head logit coldness is the standard deviation of the pre-softmax LM logits over the vocabulary, evaluated at each checkpoint and plotted as diamonds in Figure 5a. A vertical line marks the dynamic time at which the LM head coldness saturates, chosen by eye as the point after which later checkpoints no longer increase, separating the early-training phase from the late-training plateau where the MSE loss fit is performed.

MSE loss fitting. The MSE training loss after the LM head saturation point is fit using the same power-law-plus-constant form as Appendix C.1.2:

$$L_{\text{MSE}}(\tau) = c_\tau \, \tau^{-\alpha_\tau} + L_{\setminus\tau}, \tag{62}$$

with free parameters c_τ, α_τ, and $L_{\setminus\tau}$ estimated by nonlinear least squares on $\log L_{\text{MSE}}$ (`mse_1.ipynb`). The fitted $L_{\setminus\tau}$ is subtracted to yield the training-dependent component $L_{\text{MSE}} - L_{\setminus\tau}$, plotted in Figure 5b together with a reference line of slope $-1/3$.