ATTENTION REGULATES RESOLUTION, NOT SHARPNESS: A DIFFRACTION LIMIT FOR THE SOFTMAX SCALE

Anonymous authors
Paper under double-blind review

ABSTRACT

While the canonical softmax scale $1/\sqrt{D}$ preserves variance at initialization, it fails to capture the operating regime of trained attention. With normalized queries and keys, logits factor exactly into a cosine similarity and an explicit scale s, where the canonical choice corresponds to setting $s = \sqrt{D}$. Because the learned geometry can spread its cosines to offset an imposed scale, the scale and the geometry are redundant; we sweep s on a logarithmic grid to examine how autoregressive training resolves this redundancy. Training absorbs much of the imposed scale into the learned geometry, yielding a rank correlation of -0.97 across a 32-fold sweep, and conserves a logit margin rather than a scale. Indeed, eight public models with learned scales ranging from 11 to 73 maintain nearly the same margin. At the optimum, this margin sits at or just past the diffraction limit, defined as the threshold $\ln N$ at which a single key claims half the attention weight against N equal competitors, a property we formalize and verify in Lean 4. Thus, while the scale dictates sharpness, the margin normalized by the diffraction limit dictates resolution, and next-token prediction regulates resolution rather than sharpness. Empirically, the loss-optimal scale sits at $s = 16$ from 25M to 393M block parameters, across context lengths from 512 to 2048 tokens and vocabularies from 6 to 32,768 tokens, on text and RNA, with one slight exception. At this fixed scale, 64-dimensional heads nearly match 128-dimensional heads, and learned global, per-layer, or per-head scales fail to outperform the constant value. Degradation under the canonical $s = \sqrt{D}$ grows quadratically with its log-distance from this optimum, a distance set by the head dimension alone, and steepens as the context lengthens. These findings show that comparisons across head sizes under the canonical scale conflate a systematic scale artifact with architectural capacity, and that freezing $s = 16$ decouples attention resolution from head size.

1 INTRODUCTION

Scaled dot-product attention divides every query–key inner product by $\sqrt{D}$, the square root of the head dimension (Vaswani et al., 2017), so that the logits of an untrained network have unit variance under an isotropic assumption. In unnormalized attention this is only a starting point, since the norms of queries and keys are learned. Modern architectures increasingly remove this freedom: to keep attention logits from growing without bound (Dehghani et al., 2023; Wortsman et al., 2024), they normalize queries and keys before the inner product (Henry et al., 2020; OLMo Team, 2025; Qwen Team, 2025). Under query–key normalization every logit is exactly an explicit scale s times a cosine similarity, so this scale alone sets how sharply attention can select, yet it stays anchored to the historical default $s = \sqrt{D}$, tying attainable sharpness to head size. Because the scale multiplies a learned cosine, optimization could in principle absorb any imposed value by adjusting the query–key representations. Does training absorb the scale, what does it preserve instead, and does $s = \sqrt{D}$ sit at that operating point for every head size? Decoupling the scale from the head dimension lets autoregressive training answer directly (§2).

The optimum does not follow the head dimension. We sweep the scale on a $\sqrt{2}$ grid that contains the canonical point and read the held-out loss. The optimum stays at $s = 16$ for both 64- and 128-

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.

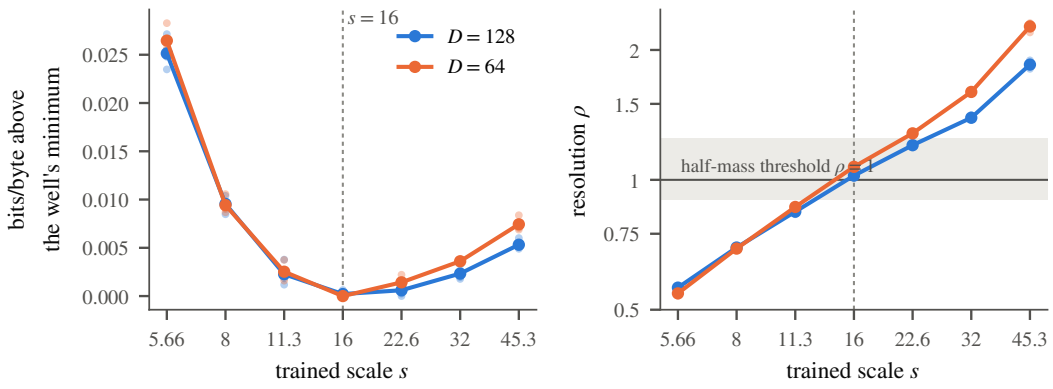


Figure 1: The loss well in the scale s on text at $D = 64$ and $D = 128$ (the reference model; lines are three-seed means, dots single seeds). Left: the loss above each well's minimum; both wells bottom out at $s = 16$ while the isotropic scale moves from 8 to 11.3. Right: the resolution ρ; the line is the half-mass threshold and the band the pre-registered $[0.90, 1.25]$.

dimensional heads (Figure 1), and holds there from 25M to 393M block parameters, across shapes at a fixed budget, contexts of 512 to 2048 tokens, vocabularies of 6 to 32,768 tokens, and on text and RNA, with one exception at most one grid step blurrier. The canonical scale therefore pays a penalty set by the head dimension alone: it fits one parameter times the squared log distance to the optimum, is a fifth of a percent of the loss at $D = 128$ and one percent at $D = 64$, steepens as the context lengthens, and does not diminish with model size. At the operating scale the head size nearly stops mattering: 64-dimensional heads match 128-dimensional heads on three seeds (§4).

Training regulates resolution, not sharpness: it conserves a logit margin rather than a scale. A result we machine-check in Lean 4 (§5, Supplement A) identifies the critical margin: a key competing against $T - 1$ equal competitors claims half the attention exactly when its logit exceeds theirs by $\ln(T - 1)$, the diffraction limit. Models trained at their optimum sit at or just past this threshold, with the resolution ρ, the margin in units of the diffraction limit, between 1.00 and 1.08 on every text well whose optimum is 16. They reach it by absorbing much of the imposed scale into their query–key geometry, which neither shrinks with the head, as the isotropic prior predicts, nor follows the vocabulary, so the scale that reaches the threshold is the same at every head size. Eight public models with learned scales from 11 to 73 hold nearly the same margin, and a scale learned during training does not improve on the fixed optimum.

For practitioners, the scale can be set once: found on a small model of the right architecture and data, it transfers to larger ones, and for next-token prediction the answer is $s = 16$ on text and RNA, a single-line change with no added parameters or compute. A fixed scale also corrects a confounded comparison: under $s = \sqrt{D}$ a wide network with large heads beats an equally sized deep network with small heads by 2.2 times the gap between them at the operating scale, because the smaller heads pay a larger penalty, and freezing the scale removes this head-size term (§6).

2 Normalization turns a learned magnitude into a fixed scale

The canonical scale reflects an initialization heuristic. Scaled dot-product attention computes $\mathrm{softmax}(QK^\top/\sqrt{D})\,V$ for heads of dimension D (Vaswani et al., 2017): if the elements of queries and keys are independent with zero mean and unit variance, their inner product has variance D, and dividing by $\sqrt{D}$ gives unit-variance logits at initialization. Each logit decomposes into magnitudes and a cosine,

$$\ell_{ij} \;=\; \frac{q_i^\top k_j}{\sqrt{D}} \;=\; \frac{\|q_i\|\,\|k_j\|}{\sqrt{D}}\,\cos\theta_{ij}, \tag{1}$$

where the prefactor sets how sharply the softmax concentrates on the winning keys. In unnormalized attention this prefactor varies for every pair and is learned through the projection weights; because nothing bounds the norms, the logits can grow without limit, a known instability (Dehghani et al.,

2023; Wortsman et al., 2024), and containing this growth is why large models normalize queries and keys.

Normalization isolates dynamic range into a single scale. Root-mean-square normalization (Zhang & Sennrich, 2019) without learnable gains maps an input x to $\sqrt{D}\,\hat{x}$, with $\hat{x} = x/\|x\|$. Applied to queries and keys, the two factors of $\sqrt{D}$ multiply to D, dividing by the conventional $\sqrt{D}$ leaves one of them, and equation 1 becomes

$$\mathrm{softmax}\Big(\frac{Q'K'^{\top}}{\sqrt{D}}\Big) \;=\; \mathrm{softmax}\big(s\,\hat{Q}\hat{K}^{\top}\big), \qquad s=\sqrt{D}, \tag{2}$$

an exact equality (Theorem 1, Supplement A) that rotary position embeddings (Su et al., 2021) leave intact. Magnitude leaves the projections, which learn directions alone; the prefactor collapses into one global hyperparameter s, set by the designer; and the logits become bounded, $\ell_{ij} \in [-s, s]$, so s alone sets how far toward hard selection attention can go. Learnable gains on the normalization (OLMo Team, 2025; Qwen Team, 2025) amount to a learned scale starting at $\sqrt{D}$ (§5).

The inherited convention targets a regime discarded by training. After normalization, $s = \sqrt{D}$ retains only its original justification: for unit vectors drawn independently and uniformly from the sphere, $\mathrm{Var}(\cos\theta) = 1/D$, so $s = \sqrt{D}$ gives unit-variance logits before training (Theorem 2). We call this the *isotropic scale*, $s_{\mathrm{iso}} = \sqrt{D}$. Trained queries and keys are not uniform on the sphere, yet the convention persists, locking attention's dynamic range to the head dimension. Because the scale multiplies a learned cosine, equation 2 carries a redundancy: for any imposed s the network can in principle adjust its angular geometry to compensate. We therefore decouple the scale from the head dimension, keeping $s = s_{\mathrm{iso}}$ as an exact baseline (§3), and let training answer. A mean-field count suggests why the head dimension could matter: for one key to take half the attention against $T-1$ equal competitors its logit must exceed theirs by $\ln(T-1)$, which at $T = 2048$ asks the geometry for a cosine gap of 0.95 at $s = 8$ ($D = 64$) against 0.48 at 16. The count is an idealization, since at the loss optimum the top key takes about 29% of the attention rather than half; what training holds fixed, we measure.

3 Method and measurement

Decoupled-scale attention. We implement attention with an ℓ_2 normalization prepass on queries and keys, decoupling the softmax scale s from the head dimension:

$$\mathrm{Attention}(Q,K,V) \;=\; \mathrm{softmax}\big(s\,\hat{Q}\hat{K}^{\top}\big)\,V, \tag{3}$$

where the rows of $\hat{Q}$ and $\hat{K}$ are unit vectors. By equation 2 this is a change of coordinate, not a new mechanism: canonical $1/\sqrt{D}$ attention on gainless RMS-normalized queries and keys is identically $s = s_{\mathrm{iso}} = \sqrt{D}$, with no added parameters or compute, so every comparison of scales pairs identical architectures with everything else held fixed. Unless noted otherwise, s is one number shared by all layers and heads.

Grid, wells and penalties. We train otherwise identical models on the logarithmic grid $s = 16 \cdot 2^{k/2}$, $k = -3, \ldots, 3$, from 5.66 to 45.25, and read the held-out loss over 100 batches. The baseline is one of the cells, since the grid contains $\sqrt{D}$ for $D = 32$, 64 and 128, and the aim is not to beat $\sqrt{D}$, which a swept hyperparameter routinely does, but to locate the optimum and see whether it moves. A *well* is the family trained across the grid with everything else fixed, and its argmin the *operating scale* s^*, resolved to one grid step. The *isotropic penalty* is the loss the canonical scale pays against the fixed scale 16,

$$\Delta\mathcal{L}(D) \;=\; \mathcal{L}(s_{\mathrm{iso}}) - \mathcal{L}(16), \tag{4}$$

which is what a practitioner gains by switching and avoids the upward bias of a minimum over noisy cells; on text, 16 is the argmin of every well but three, and in those it sits within a bar of the argmin; smaller wells are marked where used.

Setup and disclosure. Text models are decoder-only transformers with rotary position embeddings (Su et al., 2021) and a 32,768-token byte-pair vocabulary, trained on ClimbMix (Diao et al., 2025) at a context of 2048 tokens and 12 tokens per parameter under one scaling-law recipe; the reference model has 8 layers of width 512 with four heads of dimension $D = 128$ (25M block parameters). RNA models are causal language models over the spliced transcripts of 16 animal and 16 plant species in Ensembl (Yates et al., 2016; 2026), 12 layers (85M block parameters) at 11.7 tokens per parameter, over three tokenizations at the same context. Every family is trained at one commit with one recipe and, where seeds are paired, on the same seeds. The noise bar of one text cell is 0.0024 bits/byte, the spread of one model's held-out loss across evaluation windows; it falls to 0.0017 for a two-seed mean and 0.0014 for three (Supplement G). Every experiment at the current commit was pre-registered, its rules and predictions committed before the first cell ran. Two sweeps, a head-dimension ladder and a context scan, come from an earlier version of the same code on one seed each, were not pre-registered, and are marked † (Supplement G); §5 also reads earlier held-out cells and eight public checkpoints with pre-registered probes.

4 THE OPTIMUM DOES NOT FOLLOW THE HEAD DIMENSION

Stability across head dimensions. If the canonical scale identified the operating regime of trained attention, the optimum would track the head dimension, from 5.66 at $D = 32$ to 11.3 at $D = 128$. It does not. On text, on three paired seeds, the loss minimum for $D = 64$ lands at $s^* = 16$ on all three seeds, and for $D = 128$ at 16 on two seeds and one step sharper, at 22.63, on the third (Table S6 in Supplement E); neither moves toward its isotropic scale, two steps blurrier at 8 for $D = 64$ and one step at 11.3 for $D = 128$, and a finer grid† from 16 to 22 at D from 32 to 256 finds no dependence on D while $\sqrt{D}$ runs from 5.66 to 16. We state $s^* = 16$ at the resolution of the $\sqrt{2}$ grid: the claim is not a universal constant, but an optimum that does not track the head dimension. The same operating point holds across depth and size, shape, context, vocabulary and modality (Table S9 in Supplement F), with one exception.

Stability across depth, size, shape and context. At $D = 128$, from 8 layers of width 512 (25M block parameters) through 12×768 (85M) and 16×1024 (201M) to 20×1280 (393M), the scale 16 beats both adjacent grid points at every size, on one seed above 25M, at 393M by only 0.0004 bits/byte over 22.63 (Supplement F). At a matched 50M block parameters, three shapes at two head dimensions on three seeds (§6) give 18 wells, 16 with the optimum at 16 and two one step sharper, inside a bar. In the earlier context scan† at $D = 64$ the minimum stays at 16 at $T = 512$, 1024 and 2048, flattening as the context shortens but not drifting toward 8.

Stability across vocabularies, and the one exception. RNA changes the vocabulary while the corpus, the trunk and the number of keys stay fixed: single nucleotides, 3-mers and 5-mers give 6, 89 and 1,371 tokens at the same 2048 positions. At $D = 128$ the optimum is 16 on eight of nine seeds; the ninth, over single nucleotides, lands one step sharper in a well nearly flat between 16 and 22.63, where the three-seed means differ by 0.0009 bits per base. The one exception is RNA with $D = 64$: over single nucleotides the optimum is 16 on the three-seed mean, over 3-mers one step blurrier, at 11.3 on all three seeds but within 0.0007 bits per base of 16, and over 5-mers, on a finer grid, a quarter step blurrier, at 13.45 on all three seeds and 0.0018 below 16. Even here the optimum moves at most one of the two steps toward 8, shows no trend in the vocabulary, and the fixed 16 still beats $\sqrt{64}$ at every rung on every seed.

A one-parameter scale penalty. Because s^* stays put while $s_{\mathrm{iso}} = \sqrt{D}$ varies with head size, the log-distance between them,

$$\delta(D) \;=\; \ln \frac{s^*}{\sqrt{D}} \;=\; \ln s^* - \tfrac{1}{2} \ln D, \tag{5}$$

is set by the head dimension alone. Near its minimum the loss is locally quadratic in $\ln s$, so the isotropic penalty should be one curvature parameter w times the squared log-distance:

$$\mathrm{penalty}(D) \;\approx\; \tfrac{1}{2}\, w\, \delta(D)^2. \tag{6}$$

It vanishes at $D = s^{*2} = 256$, where the two models are the same model. On the earlier ladder†, fitting w at $D = 128$ alone predicts the penalties at $D = 192$, 96, 64 and 32 to within 10%, and

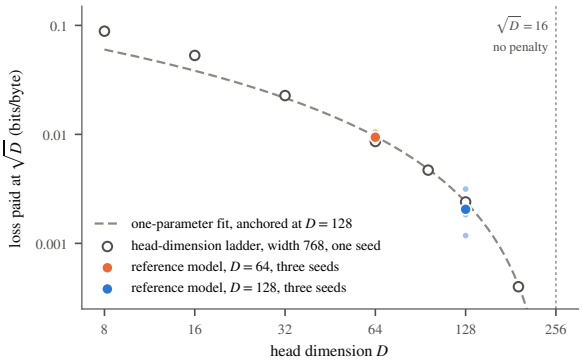


Figure 2: The loss paid at the isotropic scale, $\mathcal{L}(\sqrt{D}) - \mathcal{L}(16)$, against the head dimension. Open circles: the head-dimension ladder at width 768, one seed†. Dashed: the fit equation 6 anchored at $D = 128$, zero at $D = 256$ where $\sqrt{D} = 16$. Filled: the reference model at $D = 64$ and $D = 128$, three-seed means with single seeds as dots.

below $D = 32$ the well steepens past the parabola. Across the ladder the penalty spans a 220-fold range, from 0.0004 bits/byte at $D = 192$ to 0.0885 at $D = 8$ (Figure 2).

Narrower heads and longer contexts pay more, and size does not help. On text, on three matched seeds, the isotropic scale pays 0.0094 bits/byte at $D = 64$ against 0.0021 at $D = 128$, a ratio of 4.6 against the 4.0 that equation 6 predicts; on RNA the ratio is 2.2 to 4.3 across the three vocabularies, and the sign is the same on every seed of both modalities. In the context scan† at $D = 64$ the isotropic scale pays 0.0036, 0.0057 and 0.0086 bits/byte at $T = 512$, 1024 and 2048, about $1.5\times$ per doubling, while the optimum stays at 16, so the growth lies in the curvature w. Nor does the curvature flatten with size: at $D = 128$ the isotropic scale pays $+0.0018$, $+0.0017$, $+0.0013$ and $+0.0016$ bits/byte at 25M, 85M, 201M and 393M, inside a factor of two with no monotone trend, and for narrower heads $+0.0106$ to $+0.0068$ from 25M to 201M (Supplement F).

Architectural parity across head dimensions at the optimal scale. At the operating scale the choice between $D = 64$ and $D = 128$ nearly stops mattering. On text, on three seeds, the narrow head trails by 0.0000 bits/byte at $s = 16$ and by 0.0002 between their own optima, the parity holds at 85M and 201M ($+0.0009$ at both), and in the three shapes at 50M it trails by 0.0013–0.0024, within two bars (Supplement E). The tuned $D = 64$ model beats the untuned $D = 128$ baseline by $+0.0020$ on every seed, a margin that is entirely the wide baseline's own penalty: the narrow head is nearly equal, not better, and what made it look worse was the scale. On RNA over 5-mers the narrow head ties the wide head within 0.0001 bits per base at their own optima but trails by 0.0017 at the fixed 16, about one bar; its one outright win, over single nucleotides, is reported in Supplement C and not extrapolated.

5 Training conserves a margin, not a scale

This section identifies the quantity that training regulates instead of the scale: results machine-checked in Lean 4 (Supplement A) fix the units, and measurement puts the optimum on them.

The logit margin and the diffraction limit. For a causal window of T tokens we define

$$\Lambda = \ln(T-1), \qquad \Delta_{QK} = \frac{1}{|Q|}\sum_{i\in Q}\Big(\max_{j\le i}\cos\theta_{ij} - \mathrm{mean}_{j\le i,\, j\neq j^*}\cos\theta_{ij}\Big), \qquad \rho = \frac{s\,\Delta_{QK}}{\Lambda}, \tag{7}$$

where Λ, the log of the number of competing keys, is the unit we call the *diffraction limit*; Δ_{QK} is the trained query–key margin, the top cosine minus the mean of the rest, averaged over the queries of the longest row of a batch, whose T sets Λ; and ρ, the *resolution*, is the top-key logit margin in units of Λ. The unit is fixed by theory: where one key stands a cosine gap Δ above $T-1$ equal

competitors, its attention weight is $p_+ = 1/\big(1 + (T-1)e^{-s\Delta}\big)$, so

$$p_+ \geq \tfrac{1}{2} \iff \rho \geq 1, \qquad -\ln p_+ = \mathrm{softplus}\big(\Lambda\,(1-\rho)\big), \tag{8}$$

the half-mass threshold of Proposition 4 and the same retrieval cross-entropy, steep below the threshold and flat above it. Two caveats: Δ_{QK} measures dispersion rather than retrieval, since shuffling the keys reproduces 96–98% of it, and real competitors are spread, so at the operating scale the top key takes 0.29 of the attention, not half. The theorem fixes what the unit means; measurement, reading Δ_{QK} and ρ from the pre-softmax logits of every layer of each model at its own trained scale on held-out data (public checkpoints with an adapted probe, Supplement C), determines where the margin settles. Resolution is a name for a unit, and we claim no optics beyond it.

Training absorbs the scale, so the scale is a coordinate. When s is forced over a 32-fold range on a fixed architecture, corpus and horizon, the learned geometry gives back much of it: the rank correlation between s and Δ_{QK} is -0.973 on text. The give-back is partial, so the cells run between the horizontal line of no absorption and the iso-ρ diagonal of full absorption (Figure 3): on the $D = 64$ well, as s rises from 5.66 to 32, ρ rises only from 0.55 to 1.61, roughly as $s^{0.6}$, so the resolution is regulated nearly twice as tightly as the scale. The scale is therefore a coordinate, not an evaluation temperature. For a fixed geometry it is an ideal sharpness dial (Theorem 3); but freeze a trained model and sweep s at evaluation, and the loss well re-centres on the scale the model was trained at, in 11 of 11 cells across three head dimensions. The operating scale is a property of the family trained across scales, not of any one model swept after training.

Operating margins settle at the diffraction threshold. At the reference context of 2048 tokens, the operating scale puts the top key at or just past the diffraction limit. Across the 24 text wells whose optimum is 16, spanning both head dimensions, three shapes, 25M to 393M block parameters and three seeds, ρ^* falls within $[1.00, 1.08]$; the three wells whose optimum sits one step sharper read 1.19–1.21. RNA at $D = 128$ sits on the threshold at every vocabulary, 0.98–1.05 where the optimum is 16, and the narrow-head exception reads lower, 0.93 at the 5-mer quarter step and 0.83–0.84 over 3-mers. Against the band fixed before the first gate, $\rho \in [0.90, 1.25]$, every text optimum falls inside and five of the 18 RNA optima outside (Figure 3; per seed in Supplement C); the band is asymmetric about $\rho = 1$ because the well is, as equation 8 predicts. Eight public models, four OLMo-2 and four Qwen3 sizes, carry their scale in learned normalization gains: the effective scale runs from 11 to 17 for OLMo-2 at a context of 4096 and from 37 to 73 for Qwen3 at 8192, below its native 32,768, yet the resolution reads 0.89–1.13 and 1.24–1.29, none more than 0.04 outside the band (Supplement C). The reading is observational: a scale six times larger was met by a gap a third to a half as wide. Across contexts the threshold moves and the optimum does not: in the context scan the optimum stays at 16 from $T = 512$ to 2048 while the diffraction limit falls with the log of the context, so ρ at $s = 16$ drifts upward as the context shortens, 1.07, 1.11 and 1.16 at $T = 2048$, 1024 and 512, in the direction Proposition 5 gives and inside the band (Supplement B).

The trained gap follows neither the head nor the vocabulary. This is why the optimum does not follow the head dimension. Under the isotropic prior every cosine statistic shrinks as $1/\sqrt{D}$ (Theorem 2), so holding $s\Delta_{QK}$ at the threshold would require $s \propto \sqrt{D}$; trained queries and keys break the prior. At $s = 16$ the trained gap is about 0.51 at $D = 64$ against 0.49 at $D = 128$, where the prior predicts the narrow head's gap 41% wider, and an earlier probe from $D = 32$ to 128 reads a 7% change where the prior predicts a factor of two (Supplement G). With the gap pinned, the threshold scale Λ/Δ_{QK} is the same at every head dimension, and so is the optimum; why training pins the gap is open (Conjecture 7). The canonical scale leaves the margin below the threshold, $\rho = 0.84$ at $\sqrt{128}$ and 0.70 at $\sqrt{64}$, on the steep side of equation 8, where its penalty is paid. Nor does the gap follow the vocabulary: the threshold counts the keys a query weighs, $T - 1$, not the tokens of the output softmax, and on the RNA ladder, 2048 keys at every rung while the vocabulary grows 230-fold, the gap at $s = 16$ moves under 5% at $D = 128$ and about 10% at $D = 64$, so $\rho(16)$ stays at the threshold on all 18 grids (Supplement C); the number of keys moves the threshold, the number of words moves nothing.

Learnable scales drift to the sharp side. Learning s by gradient descent does not remove the need to set it: because the representations absorb the scale, little gradient is left along s, and on the trunk's optimizer a learnable global scale lands where its learning rate takes it. Given its own

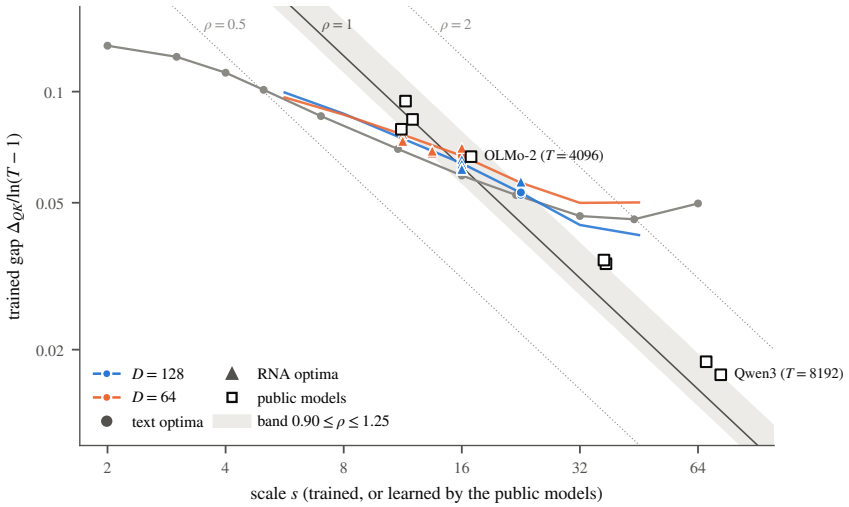


Figure 3: The margin plane: each model is placed by its scale s and its trained gap in units of the diffraction limit, $\Delta_{QK}/\ln(T-1) = \rho/s$, so constant ρ is a diagonal; the line is the threshold $\rho = 1$ and the strip the band $[0.90, 1.25]$. The forced-scale sweeps (grey: the 32-fold text sweep; colour: the $D = 64$ and $D = 128$ wells) lie between. Their optima (27 text and 18 RNA wells, coloured by head dimension) collect in the band; the five RNA optima outside it are listed in Supplement C. The eight public models (squares) sit at its edges from $s = 11$ to 73.

coordinate and schedule it becomes independent of its initialization, with runs from 8, 16 and 32 at two rates agreeing within 4%, but on $s \approx 22$, one grid step sharper than the operating scale (Supplement D). Over the nine learnable runs that reached this basin the loss is $+0.0013$ bits/byte above fixed 16, comparable to the $+0.0016$ of a fixed model at 22.63: a learnable scale neither locates the operating scale nor improves on it. Per-layer and per-head scales add nothing either (Supplement D): by Proposition 6 separate scales gain only half the curvature-weighted spread of the units' optima, near zero on text, where the per-layer optima coincide.

6 Practical implications and recipes

Calibrate on a small proxy, then freeze. Because the loss-optimal scale does not move over the ranges we trained and a learned scale settles elsewhere, it can be found once on a compact proxy and frozen for the large run. A seven-point sweep at 25M block parameters took about two hours on two GPUs and returned the operating scale of every larger model we trained, up to 393M. The proxy must share what the operating scale may depend on: the objective and its readout, the query–key normalization and the data domain. It need not match the parameter count, shape, head dimension, context or tokenizer, since re-tokenizing one RNA corpus 230-fold leaves $s^* = 16$; a new domain still needs its one sweep, and RNA at $D = 64$ needs a quarter-step sweep around the proxy's optimum.

Default constant for next-token prediction. For transformers with normalized queries and keys trained by next-token prediction, the proxy sweep returns $s^* = 16$ on text and on RNA, with the one exception above, so it can be skipped: set $s = 16$, and neither sweep it nor learn it. The change is one line and adds no parameters or compute: under unit ℓ_2 normalization set the scale to 16; under gainless RMSNorm, where the inner products equal $D\cos\theta$, multiply by $16/D$ in place of $1/\sqrt{D}$. The default carries the scope of the results behind it: head dimensions 64 and 128, 25M to 393M block parameters, contexts of 512 to 2048 tokens, vocabularies of 6 to 32,768 tokens, two modalities; per-channel learnable gains were not trained here, and a readout that pools over keys holds a different operating point (§8) and needs its sweep. The gain is the isotropic penalty removed, about 0.002 bits/byte at $D = 128$, a fifth of a percent of the loss, and 0.009 at $D = 64$; we do not argue for replacing the default on the $D = 128$ improvement alone, and the case rests on the change being free, removing a sweep, and making the head dimension a nearly free choice between

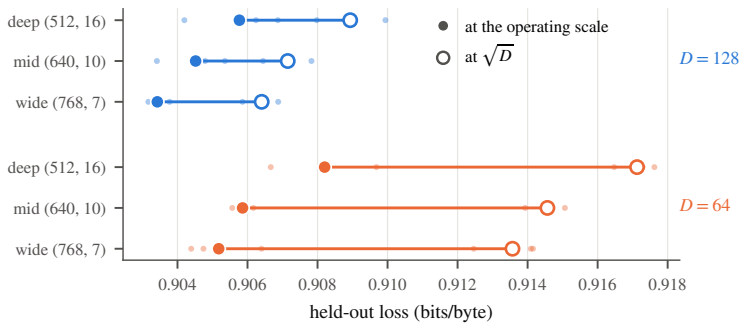


Figure 4: Width against depth at 50M block parameters, three shapes at two head dimensions: held-out loss at the operating scale (filled) and at $\sqrt{D}$ (open), three-seed means with single seeds as dots. Within a head dimension $\sqrt{D}$ moves every shape alike; across head dimensions it widens the wide/128 against deep/64 gap from 0.0048 to 0.0107 bits/byte.

64 and 128, with no saving in memory or speed claimed. The resolution ρ, read on one forward pass, also serves as a diagnostic: on text a reading below the band flags a scale too low for its heads (Supplement I).

Isolating the shape effect in architectural comparisons. Width against depth at a matched budget is conventionally compared at the isotropic scale, where the canonical ties between width, head count and head dimension give each arm its own head size. We trained three shapes at 50M block parameters, $(512, 16)$, $(640, 10)$ and $(768, 7)$, at $D = 128$ and $D = 64$, on the full grid and three seeds. The shapes themselves are close, within two bars at the operating scale, and a seed moves an arm's whole well by up to 0.003 bits/byte, so we make no pairwise shape claim finer than that (Supplement E); within one head dimension the isotropic penalty is nearly uniform, so the shapes keep their order. The distortion arises in the comparison a practitioner actually meets, a deep model with narrow heads against a wide model with broad heads: under the canonical scale the wide/128 model beats the deep/64 model by 0.0107 bits/byte, and with both at the operating scale by 0.0048. The difference is the difference between the two arms' isotropic penalties, $0.0089 - 0.0032 = 0.0057$, as large as the shape effect itself, and the coupled scale inflates the wide model's advantage $2.2\times$ on three seeds (Figure 4).

With the scale decoupled, every shape reaches the same invariant at the same fixed scale, so width-versus-depth comparisons no longer inherit a head-size term from $\sqrt{D}$; under $\sqrt{D}$ that term is as large as the shape effect it obscures.

7 Related Work

Attention scaling rules. Scaled dot-product attention divides by $\sqrt{D}$ so that the logits of independent, unit-variance queries and keys have unit variance (Vaswani et al., 2017), our isotropic scale; the maximal-update parameterization divides by D so that hyperparameters transfer across width (Yang et al., 2021). Both tie the scale to the head dimension; under query–key normalization we find the loss-optimal scale tied to neither (§4), and we do not test the maximal-update rule.

Query–key normalization and learned scales. Query–key normalization began as a guard against saturated attention: Henry et al. (2020) apply ℓ_2 normalization per head and replace $1/\sqrt{D}$ by a learnable scalar initialized by a sequence-length rule of thumb, and Dehghani et al. (2023) normalize queries and keys against logit growth at 22B parameters, an instability Wortsman et al. (2024) reproduce at small scale. Variants divide a cosine by a learnable per-head temperature (Liu et al., 2022) or normalize with trainable per-head factors at the isotropic scale (Loshchilov et al., 2025); recent open-weight language models normalize with learnable per-channel gains (OLMo Team, 2025; Qwen Team, 2025). Prior work keeps the isotropic scale or learns one; we find that learned scales span 11 to 73 at nearly the same margin and that a learned scale does not find the operating scale (§5).

Selectivity, context scaling and model shape. Zhai et al. (2023) tie training instability to collapsing attention entropy, attention that is too sharp; Chiang & Cholak (2022) and Nakanishi (2025) multiply the logits by the log of the number of keys, with a learnable factor in the second, against attention that is too flat over long contexts. Our half-mass threshold carries the same logarithm, but in the margin rather than the scale (§5); we do not test position-dependent scales or longer contexts. Scaling laws find shape a weak effect at a fixed parameter count (Kaplan et al., 2020) and size models and data together (Hoffmann et al., 2022); later work finds that shape matters for deep and narrow models (Tay et al., 2022; Liu et al., 2024), and Bhojanapalli et al. (2020) argue that tying the head size to the width over the head count creates a low-rank bottleneck and propose setting it independently. These studies use unnormalized attention, where the effective scale is learned with the lengths of queries and keys; the head-size term we measure applies where queries and keys are normalized and the scale is $\sqrt{D}$, and there it distorts width-versus-depth comparisons (§6).

8 Limitations

Explicit boundaries of our claims. Three negative claims are explicit. *Training does not choose the operating scale*: a learnable scale settles one grid step sharper than it (§5), and the operating scale is a property of the family of fixed-scale models, not recoverable from inside one run. *Shapes are not interchangeable*: at the operating scale the three shapes lie within two bars at $D = 128$, the deep shape is last at $D = 64$, and the seed swing sets a floor of about 0.003 bits/byte on any pairwise shape claim. *The diffraction limit is a unit, not a conservation law*: ρ at the operating scale drifts across contexts (§5), and we do not explain the size of the drift.

Scale, seeds, resolution and RNA. Our causal evidence tops out at 393M block parameters, on one corpus per modality, with the text runs above 50M on one seed; the public-model reading is observational. Every operating scale is resolved to the $\sqrt{2}$ grid, and at the one quarter step we ran, on RNA, even the $D = 128$ well reads 13.45 as 0.0005 bits per base below 16, within a bar. The penalty is a one-parameter fit with no mechanism, and its size at $D = 128$, 0.002 bits/byte, under one bar on a single seed, is not our reason to adopt the fixed scale; the head-size term inside every cross-head comparison is. At $D = 64$ the RNA optimum sits blurrier, at 13.45 ($\rho^* = 0.93$) over 5-mers and 11.3 (0.84) over 3-mers, the latter within a bar of 16; we report this boundary as measured and assert no cause, on one corpus at one depth (Supplement C).

Downstream evaluation and the causal regime. We report validation loss and claim nothing downstream: on CORE (Li et al., 2024), 22 zero- and few-shot tasks, a pre-registered comparison at 201M and 393M block parameters reads the fixed scale and $\sqrt{D}$ as tied on all three pairs, within two bars of a bar set mostly by training-seed noise, so the benchmark does not resolve loss gains of 0.0013–0.0068 bits/byte at these sizes (Supplement H). Every conclusion is established for causal models trained by next-token cross-entropy; a pooled classification readout holds a different operating point, and bidirectional attention we did not test.

9 Conclusion

The constant attention divides by was chosen for a network nobody trains: derived for independent queries and keys with unit-variance entries, $s = \sqrt{D}$ cancels a shrinkage of cosine statistics with the head dimension that trained networks barely exhibit. Once queries and keys are normalized, the scale is the coordinate that sets attention's dynamic range, and across head dimensions, shapes, model sizes, context lengths, vocabularies and two modalities its loss-optimal value does not follow the head dimension: it stays at $s^* = 16$, with one slight RNA exception, while $\sqrt{D}$ doubles from $D = 32$ to 128. The reason is a margin. Training regulates resolution, not sharpness: at the optimum the top key sits just past the diffraction limit while the query–key gap stays nearly fixed across head dimensions, so a scale designed for isotropic vectors is off by a distance the head dimension alone sets. Training absorbs much of any imposed scale, so gradient descent does not recover the operating scale: it is a coordinate fixed at training. Calibrate it once on a small proxy of the same architecture and data and freeze it, or, for causal next-token prediction, set $s = 16$. Decoupled from the head dimension, the scale drops out of the width-versus-depth comparisons that $\sqrt{D}$ had quietly entered, and the head dimension becomes a nearly free choice between 64 and 128.

AI USE STATEMENT

Generative AI tools were used in four ways. First, to aid and polish the writing: the authors drafted and revised the text, and AI tools proposed edits for clarity, concision and consistency of terminology, which the authors accepted or rejected sentence by sentence. Second, to draft sections of the paper from the authors' outline, results and decisions; every draft was rewritten and checked by the authors against the committed experimental record, and the research questions, the experimental design, the pre-registered rules and predictions, the interpretation of the results and the claims are the authors' own. Third, for retrieval and discovery, to locate related work and to check citation details; every cited work was read and verified against its source by the authors. Fourth, to assist in proving mathematical claims: AI tools helped write and debug the Lean 4 formalizations of Supplement A, whose statements were specified by the authors and whose proofs are checked by the Lean kernel against Mathlib. We did not use generative AI tools for research ideation or execution, or for generating synthetic datasets. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Every number in the paper and its supplement is read from a committed score file by a deterministic scorer, and at the current commit each scorer re-derives its file byte for byte from the per-cell records. Every experiment at the current commit was pre-registered, its rules, predictions and falsifiers committed before the first cell ran; the sweeps from the earlier version of the code, marked †, were not, and their records are released as well. One recipe sizes every model from its parameter count (§3). Text trains on ClimbMix (Diao et al., 2025) with a 32,768-token byte-pair tokenizer; RNA trains on Ensembl complementary DNA, whose species, genome assemblies and splits are listed in Supplement C. Training seeds are 0, 1 and 2 on text and 42, 43 and 44 on RNA, and each cell records its own. At the current commit every cell's loss is read on the same 100 held-out batches, the probe on 16 batches of four sequences, and CORE on every item of its 22 tasks (Supplement H); the noise bar, and the guard rule that voids a cell that trained unstably, are defined in Supplement G. The Lean proofs build with Lean and Mathlib v4.30.0, and each proved result has a numerical check (Supplement A). Every model we trained ran on one node with two NVIDIA H800 GPUs; the 25M sweep that finds the operating scale takes about two hours there. An anonymized copy of the code, the pre-registrations and the per-cell records accompanies the submission.

REFERENCES

Srinadh Bhojanapalli, Chulhee Yun, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Low-rank bottleneck in multi-head attention models. In *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 864–873, 2020.

David Chiang and Peter Cholak. Overcoming a theoretical limitation of self-attention. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7654–7664, 2022. doi: 10.18653/v1/2022.acl-long.527.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 7480–7512, 2023.

Shizhe Diao, Yu Yang, Yonggan Fu, Xin Dong, Dan Su, et al. Nemotron-CLIMB: Clustering-based iterative data mixture bootstrapping for language model pre-training. In *Advances in Neural Information Processing Systems*, volume 38, 2025. Datasets and Benchmarks Track.

Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4246–4253, 2020. doi: 10.18653/v1/2020.findings-emnlp.379.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

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.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, et al. DataComp-LM: In search of the next generation of training sets for language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. Datasets and Benchmarks Track.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer V2: Scaling up capacity and resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12009–12019, 2022.

Zechun Liu, Changsheng Zhao, Forrest Iandola, Chen Lai, Yuandong Tian, Igor Fedorov, Yunyang Xiong, Ernie Chang, Yangyang Shi, Raghuraman Krishnamoorthi, Liangzhen Lai, and Vikas Chandra. MobileLLM: Optimizing sub-billion parameter language models for on-device use cases. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32431–32454, 2024.

Ilya Loshchilov, Cheng-Ping Hsieh, Simeng Sun, and Boris Ginsburg. nGPT: Normalized transformer with representation learning on the hypersphere. In *International Conference on Learning Representations*, 2025.

Ken M. Nakanishi. Scalable-softmax is superior for attention. *arXiv preprint arXiv:2501.19399*, 2025.

OLMo Team. 2 OLMo 2 Furious. *arXiv preprint arXiv:2501.00656*, 2025.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

Yi Tay, Mostafa Dehghani, Jinfeng Rao, William Fedus, Samira Abnar, Hyung Won Chung, Sharan Narang, Dani Yogatama, Ashish Vaswani, and Donald Metzler. Scale efficiently: Insights from pre-training and fine-tuning transformers. In *International Conference on Learning Representations*, 2022.

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

Mitchell Wortsman, Peter J. Liu, Lechao Xiao, Katie Everett, et al. Small-scale proxies for large-scale transformer training instabilities. In *International Conference on Learning Representations*, 2024.

Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tuning large neural networks via zero-shot hyperparameter transfer. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17084–17097, 2021.

Andrew Yates, Wasiu Akanni, M. Ridwan Amode, Daniel Barrell, Konstantinos Billis, Denise Carvalho-Silva, et al. Ensembl 2016. *Nucleic Acids Research*, 44(D1):D710–D716, 2016. doi: 10.1093/nar/gkv1157.

Andrew D. Yates, Olanrewaju Austine-Orimoloye, Andrey G. Azov, et al. Ensembl 2026. *Nucleic Acids Research*, 54(D1):D1053–D1060, 2026. doi: 10.1093/nar/gkaf1239.

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Josh Susskind. Stabilizing transformer training by preventing attention entropy collapse. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 40770–40803, 2023.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.