A MODEL-AGNOSTIC SCALING RULE BEHIND THE CRITICAL INVERSE TEMPERATURE IN SOFTMAX SELF-ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Length-dependent logit rescaling is widely used to stabilize long-context self-attention, but existing analyses and methods suggest different inverse-temperature laws for the context length n, ranging from $(\log n)^{1/2}$ to $\log n$ and $(\log n)^2$. We provide a general theory showing that the critical scale is determined by the gap-counting function N_n of each attention row, without specifying the details of the model structure. Counting how many competitors lie within each gap from the maximum, we define an upper-tail accumulation scale and prove that it gives the critical inverse-temperature scale for softmax concentration: below this scale, the top competitors remain unseparated, whereas above it, the attention entropy collapses. This framework provides a model-agnostic, higher-level explanation of how scaling laws vary with N_n, and yields a direct diagnostic for attention-score families, from idealized theoretical models to practical transformers.

1 INTRODUCTION

Self-attention (Vaswani et al., 2017) is a central component of modern deep learning systems. Theoretical studies seek to understand its behavior through large-degree-of-freedom limits, including large embedding dimensions and head counts (Hron et al., 2020; Bordelon et al., 2024; Sakai et al., 2025). The practical importance of long-context inference also motivates the infinite-context-length limit. Related analyses reveal token clustering and rank collapse (Geshkovski et al., 2025; Bruno et al., 2025; Nait Saada et al., 2025). Model-specific analyses show that increasing the inverse temperature with context length changes the phase from rank collapse to entropy collapse, with a critical scaling separating the two regimes (Giorlandino & Goldt, 2026; Chen et al., 2026).

Despite this progress, what determines the inverse-temperature scaling for long contexts remains unclear. Existing analyses derive particular scaling exponents under particular model assumptions. In practice, inverse temperatures that increase with context length are also widely used to avoid instability in long-context inference (Bai et al., 2023; Nakanishi, 2025; Peng et al., 2024). As reviewed in Section 2.2, these theoretical scales and practical rules yield different logarithmic exponents. This suggests that the exponent is not determined by context length alone. What is missing is a principle that identifies, from the attention scores themselves, the scale at which softmax changes regime, without specifying how those scores are generated.

Our framework provides this relation directly from the structure of softmax. We call this approach model-agnostic because it characterizes the critical inverse-temperature scale from attention scores without prescribing the query–key structure, data structure, positional encoding, or training dynamics that generate them. These model-specific details determine the attention scores, while the same relation maps those scores to the critical scale.

Our contributions are briefly summarized as follows.

- We reveal how the critical inverse-temperature scale in self-attention varies with the score gaps and how many tokens lie within each gap of the maximum, through a common, model-agnostic relation. As a technical contribution, we construct this scale by counting competitors within each score gap and prove that it governs softmax concentration for score families with a unique maximum (Section 3).

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.

- We show that our model-agnostic relation recovers the critical inverse-temperature scale of an existing solvable attention model as a special case (Chen et al., 2026), connecting model-specific results to a common score-based description (Section 3.4).
- Compared to model-specific results that each yield a single exponent under their assumptions, we explain the diversity of exponents observed experimentally. We apply the same relation to measured attention scores, obtain estimates consistent with observed temperature scaling in training, and predict appropriate temperature scaling in an inference task (Section 4).

Thus, our work reveals a general scaling rule for softmax self-attention that explains how different temperature scalings emerge from differences in attention scores. By connecting model-specific theory with empirical observations through a common relation, it provides a theoretical foundation for future research on understanding and designing long-context transformers.

2 Preliminaries

2.1 Notation

Let the context length be n. For $n \geq 2$, let $z_{n,1}, \ldots, z_{n,n} \in \mathbb{R}$. The softmax probability of token j at inverse temperature $\beta > 0$ is $p_{n,j}(\beta) := e^{\beta z_{n,j}} / \sum_{\ell=1}^{n} e^{\beta z_{n,\ell}}$. The concrete instance we have in mind is a fixed-query self-attention row of the standard dot-product form (Vaswani et al., 2017),

$$z_{n,j} = \frac{\langle W_n^Q x_{n,i_n},\, W_n^K x_{n,j} \rangle}{\sqrt{d_{\mathrm{qk},n}}} = \frac{\langle q_{n,i_n}, k_{n,j} \rangle}{\sqrt{d_{\mathrm{qk},n}}}, \tag{2.1}$$

with token embeddings $x_{n,j} \in \mathbb{R}^{d_n}$, weight matrices $W_n^Q, W_n^K \in \mathbb{R}^{d_{\mathrm{qk},n} \times d_n}$, query/key vectors $q_{n,i} := W_n^Q x_{n,i}$, $k_{n,j} := W_n^K x_{n,j}$, and a fixed query index i_n. The theory below uses only the resulting score sequence $(z_{n,j})_j$, so the embedding-weight factorization is absorbed into q, k from Section 3 onward. For positive sequences, $f_n \asymp g_n$ means $c_1 g_n \leq f_n \leq c_2 g_n$ for some constants $0 < c_1 \leq c_2$ and all sufficiently large n. We study the asymptotic regime $n \to \infty$ with $\beta = \beta_n$, and parametrise scaling laws by the exponent $\xi_\beta \geq 0$ in $\beta_n \asymp (\log n)^{\xi_\beta}$.

2.2 Related work

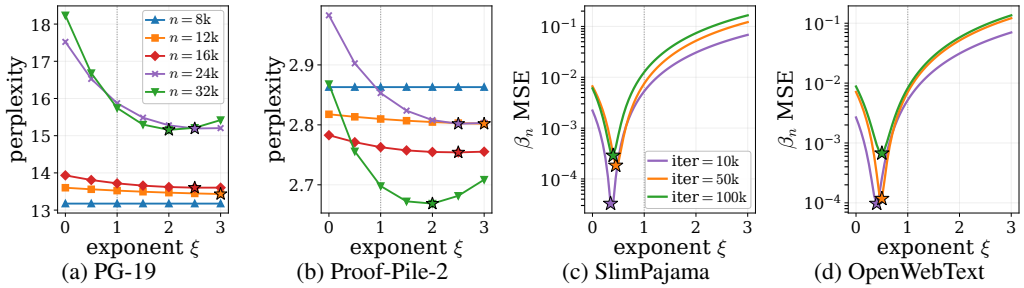
Practical inverse-temperature rules and model-specific theoretical scales lead to the exponents summarized in Table 1. In practice, SSMax (Nakanishi, 2025) multiplies logits by $s \log n$ with s learned per head, and Qwen (Bai et al., 2023) uses a LogN multiplier of order $\log n / \log n_{\mathrm{train}}$ beyond the training window n_{train}. YaRN (Peng et al., 2024) applies a common logarithmic multiplier to both query and key sides, yielding $\xi_\beta = 2$ (Appendix E).

ξ_β	Source and role
$1/2$	Gaussian-logit theory (Giorlandino & Goldt, 2026; Anson et al., 2025)
1	SSMax / Qwen rules (Nakanishi, 2025; Bai et al., 2023); equicorrelated theory (Chen et al., 2026)
2	YaRN rule (Peng et al., 2024)

Table 1: Theoretical scales and practical rules, grouped by their logarithmic exponent. Their criteria differ; see text.

In theory, Giorlandino & Goldt (2026) use a connection with the Random Energy Model (Derrida, 1981) to propose a $\sqrt{\log n}$ scale. Under marginally Gaussian logits, Anson et al. (2025) derive a position-dependent multiplier growing as the square root of the logarithm of the query–key distance. With an additive bias, it keeps the expected sum of unnormalized attention weights of constant order across fixed-ratio distance ranges (scale-invariant total attention). With independent Gaussian queries and keys, Hilton-Jones et al. (2026) establish a necessary $\Omega(\sqrt{\log n})$ scaling to keep the expected log maximum attention probability bounded below independently of context length. In the equicorrelated configuration, Chen et al. (2026) instead obtain a $\log n$ scale and analyze token contraction and gradients. Our simplex calculation recovers this scale directly from its constant score gap and competitor count (Section 3.4).

These analyses derive exponents under particular model assumptions. For independent standard Gaussian scores, classical extreme-value theory gives a scale of $1/\sqrt{2 \log n}$ for the gap between the two largest scores (Resnick, 1987). This motivates our gap-based perspective, which counts



(a) PG-19 (b) Proof-Pile-2 (c) SlimPajama (d) OpenWebText

Figure 1: Sweeping ξ_β shows that optima need not occur at $\xi_\beta = 1$. (a, b) Inference-time perplexity (Appendix A.2) is minimized around $\xi^*_\beta \in [2, 2.5]$. (c, d) The bias-free least-squares fit of the learned β_n to $a_1(\log n)^{\xi_\beta}$ ($n \geq 64$) has its MSE minimum below $\xi_\beta = 1$ on both datasets. The dotted line marks $\xi_\beta = 1$. Doc-level SE on PPL is at most 10% of the cell mean in (a, b); residual-bootstrap half-IQR on MSE is at most 8% of the point-estimate MSE in (c, d).

all competitors within each gap through $N_n(t)$. Our framework characterizes how the exponent depends on the score gaps and how many tokens lie within each gap of the maximum through a common, model-agnostic relation, and applies that relation to observed scores during training and inference. The theoretical orders and practical choices in Table 1 concern different criteria; SSMax also combines general score bounds with learned-temperature and performance experiments. The sweeps in Section 2.3 sample different values of the exponent ξ_β.

2.3 A MOTIVATING EMPIRICAL CHECK: $\xi_\beta = 1$ IS NOT UNIVERSAL

We use the parameterization $\beta_n = c(\log n)^{\xi_\beta}$ to test whether the practical logarithmic choice $\xi_\beta = 1$ is preferred across inference and training settings.

Figure 1 gives two complementary checks. Three context-length quantities appear repeatedly below: the per-token context length n of Section 2.1, the *evaluation context length* n_{eval} (the inference sliding-window size, fixed per evaluation, with $n \leq n_{\text{eval}}$), and the *training context length* n_{train} (the model's training window). First, we evaluate the pretrained Qwen-7B-Chat at extrapolation lengths $n_{\text{eval}} > n_{\text{train}}$ ($n_{\text{train}} = 8192$ for Qwen-7B-Chat), sweeping the inverse-temperature exponent ξ_β in the length-dependent multiplier $\beta_n = c(\log n)^{\xi_\beta}$, applied per token, while dynamic-NTK RoPE (Appendix A.2) is held fixed (the RoPE base is rescaled per n_{eval}). The sliding-window perplexity sweep (Appendix A.2) is reported per n_{eval}. On both PG-19 (PG19) (Rae et al., 2020) and Proof-Pile-2 (PP2) (Azerbayev et al., 2024), $\xi_\beta = 1$ is not a local minimum: the perplexity minima on the tested grid occur around $\xi^*_\beta \in [2, 2.5]$, near the $\xi_\beta = 2$ regime of YaRN (Peng et al., 2024). Qwen's default $\log n$ scaling is therefore not optimal on the tested datasets and extrapolation lengths.

Second, the SSMax design of Nakanishi (2025) was motivated by a probe experiment in which a learnable per-context-length inverse-temperature vector trained from scratch was fit well by $a_1 \log n + a_2$. We replay this probe on GPT-124M (Radford et al., 2019) trained from scratch with the nanoGPT codebase (Karpathy, 2022), with a length-$n_{\text{eval}} = 1024$ learnable vector β_n (one parameter per context position; protocol in Appendix B.1) but generalize the fit to $\beta_n = a_1(\log n)^{\xi_\beta}$ and sweep ξ_β. On both SlimPajama (SP) (Soboleva et al., 2023) and OpenWebText (OWT) (Gokaslan et al., 2019), across the reported training snapshots, the least-squares MSE for $n \geq 64$ has its minimum at $\xi^*_\beta \in \{0.40, 0.50\}$ across SlimPajama and OpenWebText; the logarithmic fit is therefore not the best fit in these runs.

The two probes favor exponents on opposite sides of one, under different models and objectives (full protocols and numerics in Appendix B). They motivate a score-based characterization of critical scaling in Section 3; Section 4 compares its finite-length exponent estimates with these observations.

3 MODEL-AGNOSTIC THEORY

For any attention-score row, increasing a competitor's gap from the maximum by δ suppresses its relative weight by $e^{-\beta\delta}$. The balance between this suppression and the number of competitors within each gap determines the relevant inverse-temperature scale. We express this balance through the cumulative gap-counting curve $N_n(t)$ and its exponential envelope, without specifying how the scores are generated. We first define the collapse observables used to characterize the scale.

3.1 COLLAPSE OBSERVABLES AND THE CRITICAL SCALE

Definition 3.1 (Collapse observables)**.** Adopt the softmax probabilities $p_{n,j}(\beta)$ from Section 2.1. The *Shannon entropy* is $H_n(\beta) := -\sum_{j=1}^{n} p_{n,j}(\beta) \log p_{n,j}(\beta)$, and the *top-two weight gap* $D_n(\beta)$ is the difference between the largest and second-largest entries of $(p_{n,j}(\beta))_{j=1}^{n}$. Along a positive deterministic sequence (β_n), we say *top-two collapse* holds if $D_n(\beta_n) \to 0$, *entropy collapse* if $H_n(\beta_n) \to 0$, and *rank collapse* if $\max_{i,j} \left|p_{n,i}(\beta_n) - p_{n,j}(\beta_n)\right| \to 0$.

Definition 3.2 (Critical scale)**.** A positive deterministic sequence (X_n) is a *rank–entropy scale* (*RE scale*) if, for every positive deterministic sequence (β_n),

$$\beta_n/X_n \to 0 \implies \text{rank collapse}, \qquad \beta_n/X_n \to \infty \implies \text{entropy collapse},$$

in the sense of Definition 3.1, and a *top-two–entropy scale* (*TE scale*) if the subcritical conclusion is replaced by top-two collapse. Each is unique up to $\asymp$ when it exists.

We refer to TE and RE scales collectively as *critical scales*: each is a deterministic sequence that separates a subcritical regime, in which a collapse observable vanishes, from a supercritical regime in which entropy collapses. Rank collapse is strictly stronger than top-two collapse.

The primary scale is Λ_n of Section 3.2, identified as a TE scale by Theorem 3.6. Within the logarithmic-power class, avoiding both collapses requires matching its growth exponent (Corollary 3.8). Corollary 3.7 additionally upgrades Λ_n to the rank-collapse boundary under $C_n := \Lambda_n \Delta_n = \alpha_n \log n \to \infty$ (Appendix D for the contrary case).

3.2 COUNTING FUNCTION, UPPER-TAIL ACCUMULATION SCALE, AND CONTACT POINT

Definition 3.3 (Cumulative gap-counting function)**.** Set $z_n^* := \max_{1\le j\le n} z_{n,j}$. The *cumulative gap-counting function*

$$N_n(t) := \#\{j \mid z_n^* - z_{n,j} \le t\}, \qquad t \ge 0,$$

counts the indices j with $z_n^* - z_{n,j} \le t$; such an index j is called a *t-competitor* (the prefix t is dropped when unambiguous). In the physics analogy with $-z_{n,j}$ as energies, the gap $z_n^* - z_{n,j}$ is the excitation above the ground state $z = z_n^*$, and $N_n(t)$ counts states with excitation energy at most t.

The softmax probabilities of Section 2.1 factor through the normalised partition function $Z_n(\beta) := \sum_{j=1}^{n} e^{-\beta(z_n^* - z_{n,j})}$, $p_{n,j}(\beta) = \exp[-\beta(z_n^* - z_{n,j})]/Z_n(\beta)$, and integration by parts (with $N_n(0^-) = 0$ and $e^{-\beta t} N_n(t) \to 0$ as $t \to \infty$) gives

$$Z_n(\beta) = \beta \int_0^\infty e^{-\beta t}\, N_n(t)\, dt. \tag{3.1}$$

The integrand of (3.1) is $e^{-\beta t} N_n(t)$. Whether the exponential damping $e^{-\beta t}$ outpaces the growth of $N_n(t)$ is captured by the envelope $N_n(t) \le e^{\beta t}$, and the smallest exponent for which this envelope holds is the next definition.

Definition 3.4 (Upper-tail accumulation scale)**.** The *upper-tail accumulation scale* Λ_n is the smallest exponent for which the envelope

$$N_n(t) \le e^{\Lambda_n t} \qquad (t > 0) \tag{3.2}$$

holds, equivalently

$$\Lambda_n := \sup_{t>0} \frac{\log N_n(t)}{t}. \tag{3.3}$$

Throughout the remainder of this section, we assume $\Lambda_n < +\infty$ eventually in n; the boundary case $\Lambda_n = \infty$ (ties at the maximum, which forces $\log N_n(t)/t \to \infty$ as $t \downarrow 0$) is deferred to Corollary C.2. Under this finite-Λ_n assumption the map $t \mapsto \log N_n(t)/t$ is non-increasing between consecutive competitor gaps and the supremum is attained as a maximum at some positive competitor gap $z_n^* - z_{n,j} > 0$ (Lemma C.1).

Definition 3.5 (Contact point)**.** Call a positive competitor gap $u = z_n^* - z_{n,j} > 0$ a *contact gap* if $\log N_n(u)/u = \Lambda_n$ (the envelope $N \le e^{\Lambda_n t}$ is attained at $t = u$). Let Δ_n be the largest such gap, and define the *contact accumulation exponent* $\alpha_n := \log N_n(\Delta_n)/\log n$. The pair (α_n, Δ_n) is the *(largest) contact point* and satisfies the *contact formula*

$$\Lambda_n = \frac{\alpha_n \log n}{\Delta_n}. \tag{3.4}$$

Figure 2 shows the gap-counting curve, the exponential envelope, and the contact point. The coordinate α_n records the finite-n contact count while Δ_n records the corresponding gap. The contact formula (3.4) is the only relation between these coordinates used below.

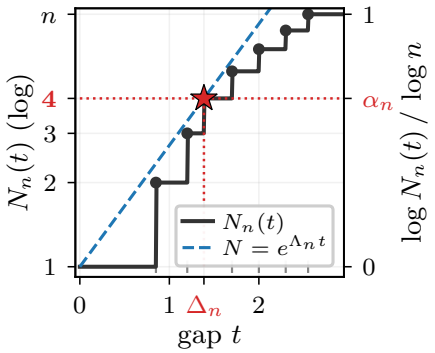


Figure 2: $N_n(t)$ and envelope $N = e^{\Lambda_n t}$; contact point (red star). Sorted-rank version: Figure C.1.

Theorem 3.6. (Λ_n) *is the TE scale.*

Proof. Set $s_n := \beta_n/\Lambda_n$. For the subcritical direction, let $\eta_n > 0$ be the gap to the closest competitor, so $D_n(\beta) = (1 - e^{-\beta\eta_n})/Z_n(\beta)$. Suppose $s_n \to 0$. By (3.4), $N_n(\Delta_n) = n^{\alpha_n}$ and $\beta_n\Delta_n = s_n\alpha_n \log n$, so monotonicity of $e^{-\beta_n t}$ on the $N_n(\Delta_n)$ tokens with $z_n^* - z_{n,j} \le \Delta_n$ gives

$$Z_n(\beta_n) \ge N_n(\Delta_n)\, e^{-\beta_n\Delta_n} = n^{\alpha_n(1-s_n)} \ge n^{\alpha_n/2}$$

for all large n since $s_n \le 1/2$ eventually. Using $1 - e^{-x} \le x$ and $\eta_n \le \Delta_n$, it holds that $D_n(\beta_n) \le \beta_n\eta_n/Z_n(\beta_n) \le \beta_n\Delta_n/n^{\alpha_n/2} = s_n\alpha_n \log n/n^{\alpha_n/2} \le (2/e)\, s_n \to 0$, where the last step uses $\sup_{a>0} a \log n/n^{a/2} = 2/e$.

For the supercritical direction, use $H_n(\beta_n) = \log Z_n(\beta_n) - \beta_n Z_n'(\beta_n)/Z_n(\beta_n)$. Suppose $s_n \to \infty$. The winner ($z_{n,j^*} = z_n^*$) contributes 1, so $Z_n(\beta_n) \ge 1$. Then, Equations (3.2) and (3.1) give

$$Z_n(\beta_n) \le \beta_n \int_0^\infty e^{-(\beta_n - \Lambda_n)t}\, dt = \frac{\beta_n}{\beta_n - \Lambda_n}, \tag{3.5}$$

which $\to 1$, so $Z_n(\beta_n) \to 1$. Now, differentiating (3.1) gives $-\beta Z_n'(\beta) = \beta^2 \int_0^\infty t e^{-\beta t} N_n(t)\, dt - Z_n(\beta)$. Since $0 \le -\beta Z_n'(\beta)$, $N_n(t) \le e^{\Lambda_n t}$, and $Z_n(\beta) \ge 1$, we have $0 \le -\beta Z_n'(\beta) \le (\beta/(\beta - \Lambda_n))^2 - 1$ for $\beta > \Lambda_n$. At $\beta = \beta_n$ this is $o(1)$, so $H_n(\beta_n) \to 0$ follows from $Z_n(\beta_n) \to 1$. □

Corollary 3.7. (Λ_n) *is the RE scale if the contact-count entropy* $C_n := \Lambda_n\Delta_n = \alpha_n \log n$ *diverges.*

A diverging contact-count entropy is what lets the rank-collapse argument absorb the lower-order tail of N_n and upgrade the TE scale to the RE scale (proof and the rank-resolved boundary in the contrary case in Appendix D).

3.3 Logarithmic critical-scaling exponent

Having established that Λ_n is a critical scale, we now record how its logarithmic growth rate is read from the contact coordinates of Definition 3.5.

For any positive sequence (X_n), if $X_n \asymp (\log n)^\xi$, then the exponent is unique when it exists: if ξ and ξ' both work, then $(\log n)^{\xi - \xi'}$ is bounded above and below, forcing $\xi = \xi'$. We call this ξ the *logarithmic growth exponent* of (X_n) and denote it ξ_X, so that

$$X_n \asymp (\log n)^{\xi_X}. \tag{3.6}$$

This single rule produces $\xi_\Lambda, \xi_\alpha, \xi_\Delta, \xi_\beta$ as instances of one family. The principal exponent of the system is ξ_Λ, the growth exponent of the upper-tail accumulation scale Λ_n.

Corollary 3.8 (Exponent-selection rule)**.** *Under $\Lambda_n \asymp (\log n)^{\xi_\Lambda}$ and $\beta_n \asymp (\log n)^{\xi_\beta}$, any schedule avoiding both top-two and entropy collapse has $\xi_\beta = \xi_\Lambda$ (otherwise $\beta_n/\Lambda_n \to 0$ or $\to \infty$, contradicting Theorem 3.6).*

The contact formula (3.4) expresses ξ_Λ in the contact coordinates of Definition 3.5 as a linear combination.

Corollary 3.9 (Coordinate decomposition of ξ_Λ)**.** *If any two of $\Lambda_n, \alpha_n, \Delta_n$ admit logarithmic growth exponents in the sense of (3.6), then so does the third, and they satisfy*

$$\xi_\Lambda \;=\; \xi_\alpha - \xi_\Delta + 1. \tag{3.7}$$

The proof is given in Appendix C.3. The corollary makes ξ_α, ξ_Δ first-class coordinate exponents on equal footing with ξ_Λ, and the side of 1 on which ξ_Λ lands is determined by the sign of $\xi_\alpha - \xi_\Delta$. Section 4.2 uses this decomposition empirically.

3.4 Gap-counting on the equicorrelated configuration and a block extension

The contact formula (3.4) constrains the one-row N_n but not the configuration behind it. The deterministic equicorrelated ("simplex") configuration of Cowsik et al. (2025), whose critical $\beta_n \asymp \log n$ was established by Chen et al. (2026), is the simplest configuration to which the framework applies; gap-counting reads ξ_Λ off it directly. The block extension varies the normalized log-count α_n. Both body constructions take the score matrix as a symmetric Gram matrix Σ_n, equivalently the self-attention score matrix (2.1) with $W_n^Q = W_n^K$ (Proposition C.4); this $Q = K$ restriction is standard in theoretical analyses of attention (Chen et al., 2026; Arnaboldi et al., 2025). Chen et al. (2026), in particular, work with the equicorrelated Gram of identity-coupled tokens, equivalent in our notation to the $W_n^Q = W_n^K = I$ specialization of (2.1). In both body constructions, $N_n(t)$ has at most two positive jumps. We write δ_n for the minimum positive gap and reserve Δ_n for the contact gap of Definition 3.5; they coincide ($\Delta_n = \delta_n$) in the simplex case and in the within-block-dominant case of the extension.

Equicorrelated (simplex) configuration. With common norm $\|x_i\|^2 = q$ and pairwise inner product $\langle x_i, x_j\rangle = \rho < q$, set $J_n := \mathbf{1}\mathbf{1}^\top$, $r_n^2 := q$, $\Delta_n := q - \rho$ to obtain

$$\Sigma_n^{\mathrm{simp}} := (r_n^2 - \Delta_n)J_n + \Delta_n I_n \geq 0.$$

Every query has flat competitor gap Δ_n, so N_n jumps from 1 to n at $t = \Delta_n$ and $\Lambda_n = \log n/\Delta_n$, $\alpha_n = 1$, the full-count endpoint. The $\beta_n \asymp \log n$ scaling of Chen et al. (2026) is the $\xi_\Lambda = 1$ slice ($\Delta_n = \Theta(1)$); $\Delta_n = (\log n)^{1-\xi_\Lambda}$ realizes any $\xi_\Lambda \geq 0$. Reading off the contact coordinates gives $\xi_\alpha = 0$ and $\xi_\Delta = 1 - \xi_\Lambda$, which satisfy the decomposition (3.7) of Corollary 3.9: this configuration sits on the $\xi_\alpha = 0$ axis with ξ_Λ entirely carried by ξ_Δ.

Two-level block-constant extension. A balanced block-constant analogue of Σ_n^{simp} shows how changes in block size and within- and between-block score gaps affect the scale and its exponent. Fix $m_n \in \{2, \ldots, n\}$ with $m_n \mid n$, and partition $\{1, \ldots, n\}$ into blocks of size m_n. Let B_n be the block-membership matrix, equivalently the block-diagonal matrix with n/m_n diagonal blocks equal to J_{m_n}. For $0 < \delta_n < \tau_n \leq r_n^2$, define

$$\Sigma_n^{\mathrm{block}} := \delta_n I_n + (\tau_n - \delta_n)B_n + (r_n^2 - \tau_n)J_n \;\geq 0. \tag{3.8}$$

Positivity follows from $I_n, B_n, J_n \geq 0$ and the nonnegative coefficients. The entries are r_n^2 on the diagonal, $r_n^2 - \delta_n$ inside a block, and $r_n^2 - \tau_n$ across blocks; thus r_n^2 is a common score shift, while δ_n and τ_n are the two relevant row gaps. Consequently $N_n(t) = 1$ for $0 \leq t < \delta_n$, $N_n(t) = m_n$ for $\delta_n \leq t < \tau_n$, and $N_n(t) = n$ for $t \geq \tau_n$ (Figure 3). Hence

$$\Lambda_n = \max\left\{\frac{\log m_n}{\delta_n}, \frac{\log n}{\tau_n}\right\}. \tag{3.9}$$

The two terms correspond to the within-block and full-row competitor groups. Shrinking either gap, or increasing its competitor count while holding the other quantities fixed, increases the corresponding

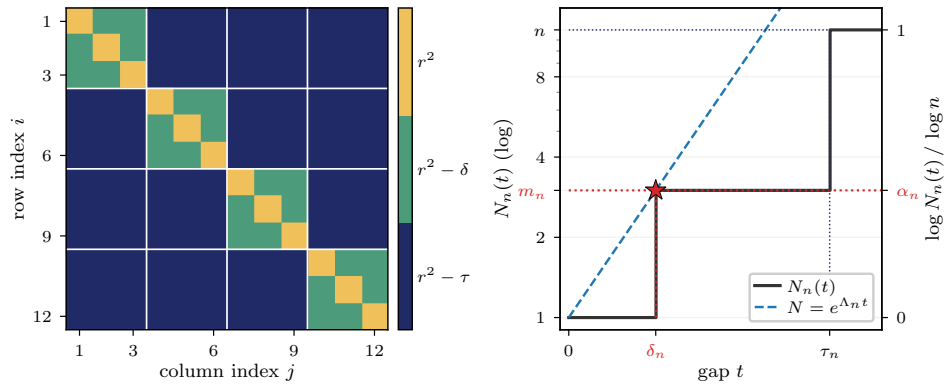


Figure 3: The two-level block-constant configuration reads Λ_n off the within-block gap δ_n in the dominant regime $\log m_n/\delta_n > \log n/\tau_n$. *Left:* the three-valued block-constant Gram matrix Σ_n^{block} (3.8) for $n = 12$, $m_n = 3$. *Right:* $N_n(t)$ jumps from 1 to m_n at $t = \delta_n$ and from m_n to n at $t = \tau_n$; the dashed envelope $N = e^{\Lambda_n t}$ is tangent at the contact point $(\Delta_n, N_n(\Delta_n)) = (\delta_n, m_n)$ (red star), with secondary axis $\log N_n(t)/\log n$.

term. The larger term determines the contact: $\Delta_n = \delta_n$ and $\alpha_n = \log m_n/\log n$ when the first is larger; otherwise $\Delta_n = \tau_n$ and $\alpha_n = 1$ (including a tie).

A prescribed $\Lambda_n = (\log n)^{\xi_\Lambda}$ is realized by $\delta_n = \log m_n/(\log n)^{\xi_\Lambda}$ and $\tau_n = C \log n/(\log n)^{\xi_\Lambda}$ with $C > 1$ and $r_n^2 \geq \tau_n$. On block-compatible subsequences, $m_n = n^a$ gives $\alpha_n = a$ and $\delta_n = a(\log n)^{1-\xi_\Lambda}$, hence $(\xi_\alpha, \xi_\Delta) = (0, 1-\xi_\Lambda)$. For bounded $m_n = m \geq 2$, $\alpha_n = \log m/\log n \asymp (\log n)^{-1}$ and $\delta_n = \log m/(\log n)^{\xi_\Lambda}$ instead give $(\xi_\alpha, \xi_\Delta) = (-1, -\xi_\Lambda)$. Both satisfy (3.7), with the latter illustrating how the normalized contact count can contribute to the scale's growth exponent.

The block configuration represents a hierarchical cluster structure: tokens within a block share the smaller score gap δ_n, while distinct blocks are separated by the larger gap τ_n. Li et al. (2023) give a mechanistic account of topic clustering inside trained transformers, and Garnier-Brun et al. (2025) analyze attention block structure under hierarchical filtering of the input. Such analyses address how the structure arises; the gap-counting relation describes how the score gaps and how many tokens lie within each gap of the maximum determine the inverse-temperature scale.

4 Empirical gap-counting analysis

We apply the gap-counting construction of Definitions 3.3 to 3.5 on two representative settings of Section 2.3: pretrained Qwen-7B-Chat at extrapolation context lengths (inference-time) and GPT-124M trained from scratch with a learnable per-position β_n vector (training-time). The aim is to read $\hat{\xi}_\Lambda$ from attention scores alone, align it with the ξ_β^* optima of Section 2.3, and interpret the coordinate decomposition of Corollary 3.9 on real attention rows.

4.1 Aggregation protocol

Each tuple (layer, head, document or validation batch, query position) determines an attention row: a length-n vector of scores $(z_{n,1}, \ldots, z_{n,n})$. From every row, we read the contact triple $(\Lambda_n, \alpha_n, \Delta_n)$ via Definitions 3.3 to 3.5 at numerical tolerance $\epsilon = 10^{-6}$. Rows with $\log_{10} \Lambda_n > 5$ are treated as ties for aggregation and excluded from every aggregate below.

Let $\mathcal{C}$ index the experimental conditions: each *cell* $c \in \mathcal{C}$ is a fixed (setting, dataset, n_{eval}) on Qwen and a fixed (setting, dataset, training iteration) on nanoGPT. Each cell carries a finite *n-grid* $\mathcal{N}(c) \subset \mathbb{N}$, the set of context lengths at which scores are sampled in that cell (values specified below). For $c \in \mathcal{C}$ and $n \in \mathcal{N}(c)$, let $\mathcal{R}(c, n) := \{$ retained attention rows of cell c at context length n $\}$ and write $K(c, n) := |\mathcal{R}(c, n)|$ for its cardinality.

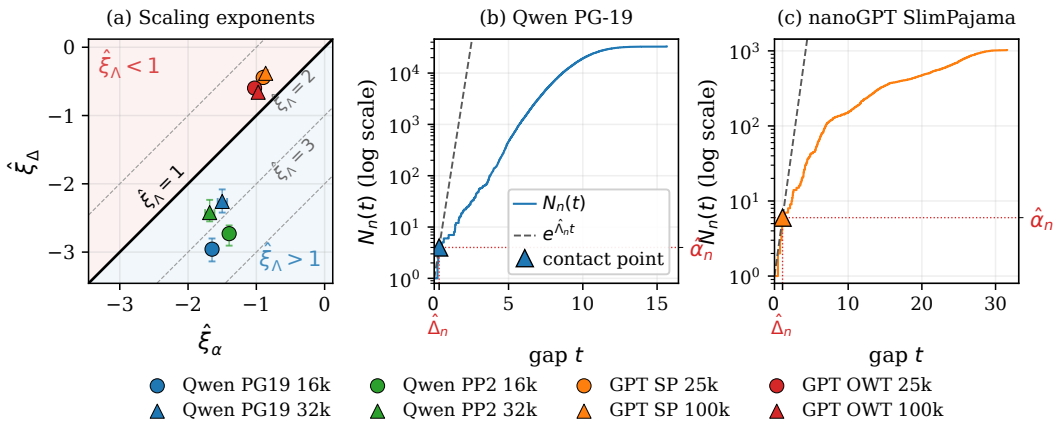


Figure 4: (a) The $\hat{\xi}_\Lambda = 1$ boundary $\hat{\xi}_\alpha = \hat{\xi}_\Delta$ separates Qwen (below) from nanoGPT (above), by (3.7). (b,c) Representative $N_n(t)$ curves: Qwen-7B PG-19 at $n_{\text{eval}} = 32768$ ($\hat{\alpha}_n = 0.13$, $\hat{\Delta}_n = 0.33$), and nanoGPT SlimPajama at iter $= 100$k ($\hat{\alpha}_n = 0.26$, $\hat{\Delta}_n = 1.1$). Qwen's smaller contact gap lies on a steeper Λ_n-envelope.

For $X \in \{\Lambda, \alpha, \Delta\}$ and a row $i \in \mathcal{R}(c, n)$, write $X_n^{(i)}$ for the value of the contact-triple coordinate X read off row i at context length n. The *bucket mean* of $\log X_n$ in cell c at grid point $n \in \mathcal{N}(c)$ is $\overline{\log X}(c, n) := K(c, n)^{-1} \sum_{i \in \mathcal{R}(c,n)} \log X_n^{(i)}$. We estimate $\hat{\xi}_X(c)$ by fitting the slope of $\overline{\log X}(c, n)$ against $\log \log n$ over $n \in \mathcal{N}(c)$ (details in Appendices F and G.1).

We apply this aggregation protocol to Qwen-7B-Chat over PG-19 (Rae et al., 2020) and Proof-Pile-2 (Azerbayev et al., 2024) at the two extrapolation context lengths $n_{\text{eval}} \in \{16384, 32768\}$ of Section 2.3 (4-point n-grid per n_{eval}, fit per-n_{eval} to keep the dynamic-NTK RoPE rescale fixed within a fit; full Qwen acquisition in Appendix G). On nanoGPT, GPT-124M is trained from scratch with a learnable per-position β_n at $n_{\text{eval}} = 1024$ (6-point grid $\{64, 128, 256, 512, 768, 1024\}$, three evaluation seeds pooled into $\mathcal{R}(c, n)$; configurations in Appendix A.3, checkpoint selection and the per-(layer, head) and rescaled-gap variants in Appendix F). Uncertainty estimates in Table 2 are specified in Appendix F.

4.2 Score-based exponents and temperature sweeps

Cell-by-cell agreement. Table 2 aligns $\hat{\xi}_\Lambda$ with the Section 2.3 optimum ξ_β^* on eight (setting, run, iter) cells. On Qwen, the per-n_{eval} PPL optima $\xi_{\beta,\text{PPL}}^*$ and $\hat{\xi}_\Lambda$ agree within the Section 2.3 0.5-grid (absolute gap ≤ 0.26), with both decreasing from $\xi_\beta^* \in [2.31, 2.50]$ at $n_{\text{eval}} = 16384$ to $[1.74, 2.00]$ at $n_{\text{eval}} = 32768$. On nanoGPT, $\hat{\xi}_\Lambda$ matches the multiplier-side $\hat{\xi}_\beta$ on the same six-point domain to $|\hat{\xi}_\Lambda - \hat{\xi}_\beta| \in \{0.12, 0.14\}$ at iter $= 25$k and $\{0.17, 0.27\}$ at iter $= 100$k. The two columns are read off disjoint inputs (attention scores vs. the trained β_n vector); Theorem 3.6 predicts equality only asymptotically, consistent with the residual on the nanoGPT $n \leq 1024$ window.

Coordinate decomposition. The decomposition (3.7) of Corollary 3.9 carries through to the bucket-mean slopes, and the right two columns of Table 2 confirm $\hat{\xi}_\Lambda = \hat{\xi}_\alpha - \hat{\xi}_\Delta + 1$ row by row at the table's precision; the side of 1 on which $\hat{\xi}_\Lambda$ lands is fixed by the sign of $\hat{\xi}_\alpha - \hat{\xi}_\Delta$. Figure 4 (left) plots the eight $(\hat{\xi}_\alpha, \hat{\xi}_\Delta)$ pairs of Table 2: the four Qwen extrapolation-time points sit below the diagonal $\hat{\xi}_\alpha = \hat{\xi}_\Delta$ ($|\hat{\xi}_\Delta| > |\hat{\xi}_\alpha|$, contact gap shrinks faster with $\log \log n$), and the four nanoGPT training-time points sit above it ($|\hat{\xi}_\alpha| > |\hat{\xi}_\Delta|$, contact count shrinks faster). The middle and right panels recast one representative row per contrast cell (Qwen PG-19 $n_{\text{eval}} = 32768$ at $n = 32768$; nanoGPT SlimPajama iter $= 100$k at $n = 1024$) as $N_n(t)$.

Calibration character of Λ_n. Table 2 evaluates Λ_n as a calibration scale for inverse temperature: agreement with ξ_β^* relies only on Corollary 3.8 and does not require TE and RE scales to coincide.

setting	dataset	n_{eval}/iter	tie %	ξ^*_β	$\hat{\xi}_\Lambda$	$\hat{\xi}_\alpha$	$\hat{\xi}_\Delta$
Qwen7B	PG19	16384	7.4 ($\pm$0.2)	2.5	2.31 ($\pm$0.10)	-1.65 ($\pm$0.08)	-2.96 ($\pm$0.17)
Qwen7B	PG19	32768	8.8 ($\pm$0.2)	2.0	1.76 ($\pm$0.10)	-1.50 ($\pm$0.07)	-2.26 ($\pm$0.17)
Qwen7B	PP2	16384	6.8 ($\pm$0.2)	2.5	2.33 ($\pm$0.11)	-1.40 ($\pm$0.05)	-2.73 ($\pm$0.15)
Qwen7B	PP2	32768	7.2 ($\pm$0.2)	2.0	1.74 ($\pm$0.11)	-1.68 ($\pm$0.07)	-2.42 ($\pm$0.16)
nanoGPT	SP	25k	2.8 ($\pm$0.1)	0.40 ($\pm$0.03)	0.54 ($\pm$0.03)	-0.90 ($\pm$0.01)	-0.44 ($\pm$0.04)
nanoGPT	SP	100k	6.0 ($\pm$0.4)	0.35 ($\pm$0.04)	0.52 ($\pm$0.03)	-0.86 ($\pm$0.01)	-0.38 ($\pm$0.04)
nanoGPT	OWT	25k	4.0 ($\pm$0.4)	0.46 ($\pm$0.05)	0.57 ($\pm$0.04)	-1.03 ($\pm$0.02)	-0.60 ($\pm$0.05)
nanoGPT	OWT	100k	5.7 ($\pm$0.3)	0.42 ($\pm$0.05)	0.69 ($\pm$0.05)	-0.97 ($\pm$0.02)	-0.66 ($\pm$0.05)

Table 2: $\hat{\xi}_\Lambda$, read off attention scores alone, lands on the same side of $\xi = 1$ as the sweep optimum ξ^*_β in every cell: both exceed 1 on the four Qwen cells, both fall below 1 on the four nanoGPT cells. Each $\hat{\xi}_X$ ($X \in \{\Lambda, \alpha, \Delta\}$) is the fitted slope of the bucket-mean $\overline{\log X}(c, n)$ against $\log \log n$ (Section 4.1); ξ^*_β is the per-cell optimum from Section 2.3. Bracketed uncertainties are specified in Appendix F; tie % is the fraction of rows treated as ties by $\log_{10} \Lambda_n > 5$.

The rank-side checks in Appendix D.1 examine the contact-count condition and the rank-resolved boundary, including a direct p^*_n sweep on the four Qwen cells.

5 Conclusion and Discussion

This paper asks what determines the critical scaling of inverse temperature for self-attention in the long-context limit. Our answer is that it is determined by $(\Lambda_n, \alpha_n, \Delta_n)$. The cumulative gap-counting function $N_n(t)$ of Definition 3.3 records how many competitors lie within each gap of the winning score, and the upper-tail accumulation scale Λ_n of Definition 3.4 gives the relevant scale. Theorem 3.6 identifies Λ_n as a critical scale for arbitrary score rows with a unique maximum; Corollary 3.7 sharpens this to the standard rank-collapse boundary whenever the contact-count entropy $\alpha_n \log n = \Lambda_n \Delta_n$ diverges, with empirical verification on the bounded-C_n cells via Appendix D.1. Thus the exponent ξ_β in $\beta_n \asymp (\log n)^{\xi_\beta}$ is not universal: when the logarithmic growth exponents exist, any schedule avoiding both top-two and entropy collapse has $\xi_\beta = \xi_\Lambda$ (Corollary 3.8), the logarithmic growth rate of Λ_n in the sense of (3.6). Equivalently, the contact formula (3.4) shows that ξ_β is determined by both the normalized log-contact count α_n and the contact gap Δ_n of Definition 3.5, namely $\xi_\beta = \xi_\alpha - \xi_\Delta + 1$ (Corollary 3.9). The simplex and block examples show how the score gaps and the number of tokens within each gap of the maximum determine this exponent, while Section 4 applies the same relation to observed scores from inference and training.

Limitations and open problems. Our framework explains the diversity of scaling exponents observed experimentally through the measured contact quantities $(\Lambda_n, \alpha_n, \Delta_n)$ across context lengths (Section 4). Existing model-specific results each yield a single exponent under their respective assumptions and therefore cannot explain this diversity. The complementary limitation of our theory is that it does not explain how architecture, positional encoding, data, or training determine the score gaps and how many tokens lie within each gap of the maximum, which together determine the exponents. Connecting these two levels of description through logit distributions is an open problem. Inference-time theories describe how token representations form clusters, exhibit metastability, or develop lower-dimensional structure across depth or loop time (Geshkovski et al., 2023; 2024; Karagodin et al., 2024; Bruno et al., 2025; Tomihari & Karakida, 2025; Rigollet, 2026), while learning-time analyses explain semantic, positional, or hierarchical structure in embeddings and attention maps (Li et al., 2023; Garnier-Brun et al., 2025; Arnaboldi et al., 2025). Coupling these dynamics with gap-counting could connect changes in $N_n(t)$, Λ_n, and (α_n, Δ_n) across depth, loop time, and training time to the mechanisms shaping the critical scale.

AI use statement

LLMs were used only for writing, editing, and formatting assistance, which does not impact the core methodology, scientific rigor, or originality of the research. The pretrained Qwen LLMs that appear

in the paper are the experimental subject of the analysis (Appendix A.2), not a methodological tool used to produce the results.

Reproducibility statement

The assumptions and proofs for the theoretical results are provided in Section 3 and Appendices C and D. Appendix A describes the models, datasets, training and evaluation settings, implementation details, and computational resources. The temperature-fitting, attention-score acquisition, gap-counting, and statistical estimation procedures are documented in Appendices B, B.2, F and G.

References

Ben Anson, Xi Wang, and Laurence Aitchison. Scale-invariant attention. In *Advances in Neural Information Processing Systems*, 2025.

Luca Arnaboldi, Bruno Loureiro, Ludovic Stephan, Florent Krzakala, and Lenka Zdeborová. Asymptotics of SGD in sequence-single index models and single-layer attention networks. In *Advances in Neural Information Processing Systems*, 2025.

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. Llemma: An open language model for mathematics. In *International Conference on Learning Representations*, 2024.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint*, 2023. URL https://arxiv.org/abs/2309.16609.

Blake Bordelon, Hamza Chaudhry, and Cengiz Pehlevan. Infinite limits of multi-head transformer dynamics. In *Advances in Neural Information Processing Systems*, 2024.

Giuseppe Bruno, Federico Pasqualotto, and Andrea Agazzi. A multiscale analysis of mean-field transformers in the moderate interaction regime. In *Advances in Neural Information Processing Systems*, 2025.

Shi Chen, Zhengjiang Lin, Yury Polyanskiy, and Philippe Rigollet. Critical attention scaling in long-context transformers. In *International Conference on Learning Representations*, 2026.

Aditya Cowsik, Tamra Nebabu, Xiaoliang Qi, and Surya Ganguli. Geometric dynamics of signal propagation predict trainability of transformers. *Physical Review E*, 112(5):055301, 2025. doi: 10.1103/gjsm-l642.

B. Derrida. Random-energy model: An exactly solvable model of disordered systems. *Physical Review B*, 24(5):2613–2626, 1981. doi: 10.1103/PhysRevB.24.2613.

Jérôme Garnier-Brun, Marc Mézard, Emanuele Moscato, and Luca Saglietti. How transformers learn structured data: Insights from hierarchical filtering. In *International Conference on Machine Learning*, 2025.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. In *Advances in Neural Information Processing Systems*, 2023.

Borjan Geshkovski, Hugo Koubbi, Yury Polyanskiy, and Philippe Rigollet. Dynamic metastability in the self-attention model. *arXiv preprint*, 2024. URL https://arxiv.org/abs/2410.06833.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on Transformers. *Bulletin of the American Mathematical Society*, 62(3):427–479, 2025. doi: 10.1090/bull/1863.

Alessio Giorlandino and Sebastian Goldt. Two failure modes of deep transformers and how to avoid them: a unified theory of signal propagation at initialisation. In *International Conference on Learning Representations*, 2026.

Aaron Gokaslan, Vanya Cohen, Ellie Pavlick, and Stefanie Tellex. OpenWebText corpus. https://skylion007.github.io/OpenWebTextCorpus/, 2019.

Sam Hilton-Jones, Timothy J. Norman, and Zhanxing Zhu. Modelling attention with Aitchison geometry: Token distinguishability and temperature scaling. In *International Conference on Machine Learning*, 2026.

Jiri Hron, Yasaman Bahri, Jascha Sohl-Dickstein, and Roman Novak. Infinite attention: NNGP and NTK for deep attention networks. In *International Conference on Machine Learning*, 2020.

Nikita Karagodin, Yury Polyanskiy, and Philippe Rigollet. Clustering in causal attention masking. In *Advances in Neural Information Processing Systems*, 2024.

Andrej Karpathy. nanoGPT: The simplest, fastest repository for training/finetuning medium-sized GPTs. https://github.com/karpathy/nanoGPT, 2022.

Yuchen Li, Yuanzhi Li, and Andrej Risteski. How do transformers learn topic structure: Towards a mechanistic understanding. In *International Conference on Machine Learning*, 2023.

Thiziri Nait Saada, Alireza Naderi, and Jared Tanner. Mind the gap: a spectral analysis of rank collapse and signal propagation in attention layers. In *International Conference on Machine Learning*, 2025.

Ken M. Nakanishi. Scalable-softmax is superior for attention. *arXiv preprint*, 2025. URL https://arxiv.org/abs/2501.19399.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models. In *International Conference on Learning Representations*, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations*, 2020.

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

Philippe Rigollet. The mean-field dynamics of transformers. In *Proceedings of the International Congress of Mathematicians*, 2026.

Mana Sakai, Ryo Karakida, and Masaaki Imaizumi. Infinite-width limit of a single attention layer: Analysis via Tensor Programs. In *Advances in Neural Information Processing Systems*, 2025.

Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R. Steeves, Joel Hestness, and Nolan Dey. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama. https://www.cerebras.ai/blog/slimpajama-a-627b-token-cleaned-and-deduplicated-version-of-redpajama, 2023.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

Together Computer. RedPajama: An open source recipe to reproduce LLaMA training dataset. https://github.com/togethercomputer/RedPajama-Data, 2023.

Akiyoshi Tomihari and Ryo Karakida. Recurrent self-attention dynamics: An energy-agnostic perspective from Jacobians. In *Advances in Neural Information Processing Systems*, 2025.

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*, 2017.

A EXPERIMENTAL SETUP

This appendix collects the experimental basic setup shared across the inference-time ξ-sweep of Section 2.3 and the training-time gap-exponent trajectory of Section 4.1. The per-experiment grids, seeds, acquisition exponents, and numerical tolerances are recorded in the corresponding protocol appendices (Appendices B, F and G.1).

A.1 DATASETS

The Qwen long-context ξ-sweep (Section 2.3) draws documents from the PG-19 (Rae et al., 2020) test split and the Proof-Pile-2 (Azerbayev et al., 2024) arXiv test split, sourced from RedPajama (Together Computer, 2023) (up to 20 per dataset per n_{eval}; the per-cell counts entering the perplexity averages are listed in Appendix B.2). The Qwen-7B-Chat gap-counting analysis (Section 4) uses a smaller fixed subset of these documents because the per-row analysis is heavier: 4 from PG-19 and 5 from Proof-Pile-2 enter the headline statistics of Table 2, with acquisition and aggregation conditions specified in Appendix G. The GPT-124M training-time runs of Sections 2.3 and 4.1 are pretrained on SlimPajama (Soboleva et al., 2023) and OpenWebText (Gokaslan et al., 2019), one run per dataset.

A.2 QWEN LONG-CONTEXT EVALUATION

We probe two pretrained checkpoints: Qwen-1.8B-Chat with training window $n_{\text{train}} = 8192$, and Qwen-7B-Chat (32 transformer layers, 32 attention heads per layer) also with $n_{\text{train}} = 8192$. Both keep rotary position embeddings (Su et al., 2024) on with the dynamic-NTK RoPE base schedule of Qwen (Bai et al., 2023) defined just below. On top of dynamic-NTK, we apply the legacy length-dependent multiplier

$$m(n;\xi) = \big(\max(1, \log n / \log n_{\text{train}})\big)^{\xi}$$

to the attention logits at extrapolation lengths $n > n_{\text{train}}$, equivalent to the parameterization $\beta_n \propto (\log n)^{\xi}$ for $n > n_{\text{train}}$ and to the no-multiplier baseline at $\xi = 0$; the choice $\xi = 2$ has the same asymptotic exponent as the YaRN schedule (Peng et al., 2024), as derived in Appendix E. The contexts n and ξ-grids vary by experiment and are recorded with the respective figures.

Dynamic-NTK RoPE. Standard RoPE (Su et al., 2024) rotates the per-head query/key pair in $\mathbb{R}^d$ using a fixed base θ_0 (typically $\theta_0 = 10000$): the k-th coordinate pair at position i is multiplied by $R(i\,\theta_0^{-2k/d})$. The *dynamic-NTK* schedule implemented in Qwen (Bai et al., 2023) replaces θ_0 by a per-evaluation base $\theta_0\, s(n_{\text{eval}})$ where

$$s(n_{\text{eval}}) \;:=\; \Big[\max\Big(1, 2^{\lceil \log_2(n_{\text{eval}}/n_{\text{train}})+1 \rceil} - 1\Big)\Big]^{d/(d-2)}.$$

The base is recomputed once per inference window of size n_{eval} and then held fixed across positions $i \le n_{\text{eval}}$ inside the window. The high-frequency rotations are nearly unchanged while the low-frequency ones are stretched. Throughout this paper, "dynamic-NTK RoPE" refers to this schedule, with the legacy multiplier $m(n;\xi)$ above applied as a separate inverse-temperature scalar on the attention logits.

Sliding-window perplexity. We evaluate complete windows of length n_{eval} with stride $n_{\text{eval}}/2$. In the first window, all tokens except the first are scored; subsequent windows score only tokens not already scored, so each token contributes at most once. For each document, we exponentiate the mean cross-entropy over its scored tokens, then report the arithmetic mean of these per-document perplexities. This is the quantity denoted *strict perplexity* in Appendix B.2.

A.3 GPT-124M TRAINING-TIME SETUP

GPT-124M (Radford et al., 2019) (12 transformer layers, 12 attention heads per layer, context length $n_{\text{eval}} = 1024$) is trained from scratch using the nanoGPT codebase (Karpathy, 2022) in four configurations: two positional encodings (learned absolute, as in GPT-2 (Radford et al., 2019); rotary (Su et al., 2024)) $\times$ two pretraining datasets (SlimPajama, OpenWebText). Each configuration runs for 100,000 iterations with checkpoints saved every 5000 iterations; the seven milestones $\text{iter} \in \{0, 5\text{k}, 10\text{k}, 25\text{k}, 50\text{k}, 75\text{k}, 100\text{k}\}$ are the snapshots used by the body and appendix figures.

Learnable inverse-temperature vector β_n. Each model carries an additional length-n_{eval} vector β_n (one parameter per context position n) tied across layers and heads; per (layer, head) the actual attention multiplier is

$$\beta_{n,\ell,h} = s_{\ell,h}\,\beta_n + b_{\ell,h}, \tag{A.1}$$

with per-head scale $s_{\ell,h}$ and bias $b_{\ell,h}$ trained alongside β_n. The n-dependence is carried entirely by β_n; the body's β_n regression fit (Section 2.3) reads β_n off after training and fits it to $a_1(\log n)^{\xi}$, while the gap-exponent diagnostic (Section 4.1) compares raw and effective gap exponents row-wise. Unlike SSMax, which fixes $\beta_{n,\ell,h} = s_{\ell,h}\log n$ with a single scalar $s_{\ell,h}$ per head, here β_n is unconstrained at initialization and is shaped only by the training loss. The original SSMax design-rationale experiment of Nakanishi (2025) used the same unconstrained β_n vector and fit it to $a_1 \log n + a_2$, motivating the logarithmic constraint.

Hyperparameters. Table A.1 records the architecture, β_n realization, and optimization settings shared across the four configurations, together with the four (positional encoding, dataset) pairs that distinguish them. All values are read from the training configs of the runs (`n_layer`, `n_head`, `n_embd`, optimizer schedule, etc.).

Architecture (4 runs shared, GPT-2 124M)	
Layers L	12
Attention heads H	12
Embedding dim d_{model}	768
Per-head dim $d_{qk} = d_v$	64
FFN dim d_{ff}	3072 ($= 4\,d_{\text{model}}$)
Vocabulary size	50,257 (GPT-2 BPE; padded to 50,304)
Context length n_{eval}	1024
Linear bias	enabled
Dropout	0
Total parameters	$\sim$124M
Inverse-temperature β_n realization	
Per-position factor β_n	learnable vector of length n_{eval}, init. $\beta_n[n] = 1$
Per-(layer,head) realization	$\beta_{n,\ell,h} = s_{\ell,h}\,\beta_n + b_{\ell,h}$ (Equation (A.1))
$s_{\ell,h}, b_{\ell,h}$	per-head, learnable; $s_{\ell,h}$ init. 1, $b_{\ell,h}$ init. 0
Optimization (4 runs shared)	
Optimizer	AdamW
Learning-rate schedule	cosine: $6\times10^{-4} \to 6\times10^{-5}$
Warmup	1000 iters
Weight decay	0.1
(β_1, β_2)	$(0.9, 0.95)$
Gradient clip	1.0
Per-step batch	32 sequences $\times$ grad-accum 64 (effective 2048)
Precision	bfloat16
Iterations	100,000
Checkpoint cadence	every 5000 iters (milestone keep)
Configurations (4 runs)	
1: learned absolute PE	SlimPajama
2: learned absolute PE	OpenWebText
3: RoPE ($\theta_0 = 10000$)	SlimPajama
4: RoPE	OpenWebText

Table A.1: GPT-124M training-time configuration. The four runs share the architecture, β_n realization, and optimization settings listed above; they differ only in the positional encoding (learned absolute as in Radford et al. 2019, or RoPE Su et al., 2024) and the pretraining dataset (SlimPajama Soboleva et al., 2023, or OpenWebText Gokaslan et al., 2019). Three evaluation seeds produce independent validation batches at every checkpoint.

A.4 Compute resources

The reported experiments were run on NVIDIA H100 GPUs. Each GPT-124M training run used four GPUs, while Qwen evaluation and attention-score extraction used one GPU per job.

The four GPT-124M training runs in Table A.1 each completed 100,000 iterations in approximately 38–45 hours, requiring approximately 670 H100 GPU-hours in total.

The Qwen-7B perplexity sweep required approximately 5 GPU-hours of runtime, with a few additional GPU-hours for attention-score extraction and post-processing across the Qwen and GPT-124M experiments. These estimates cover the reported experiments; preliminary and unsuccessful runs required additional compute.

A.5 Implementation details

The experiments use PyTorch 2.6, with Transformers 4.32.0 for Qwen evaluation. The pretrained checkpoints are the public `Qwen/Qwen-7B-Chat` and `Qwen/Qwen-1_8B-Chat` Hugging Face releases, used without fine-tuning.

The datasets are the `deepmind/pg19`, `EleutherAI/proof-pile-2` (arXiv subset), `cerebras/SlimPajama-627B`, and `Skylion007/openwebtext` Hugging Face releases, with splits and selection described in Appendix A.1. Qwen evaluation uses its bundled BPE tokenizer; GPT-124M training and evaluation use GPT-2 BPE via `tiktoken`.

For Qwen, the length-dependent multiplier defined in Appendix A.2 is applied to the per-head dot-product score before the softmax. For the gap-counting results in Table 2, the multiplier's exponent during score acquisition, denoted by ξ_0, is set to 2 in every layer. The gap-counting analysis reads the unrescaled score before this multiplier is applied. Dynamic-NTK RoPE uses the Qwen schedule in Appendix A.2, controlled by Qwen's existing `use_dynamic_ntk` configuration switch. The GPT-124M inverse-temperature parameterization and initialization are specified in Appendix A.3 and Table A.1.

B Protocols and results for the ξ_β-sweeps

This appendix records the body β_n regression-fit protocol and the training-time trajectory deferred from Section 2.3.

B.1 Regression fits and trajectories of β_n

The four-configuration training schedule and the per-(layer, head) realization (A.1) of the learnable vector β_n are as in Appendix A.3. At each of the seven milestone checkpoints, we extract β_n (a length-1024 vector) and fit the bias-free regression model $\beta_n = a_1(\log n)^{\xi_\beta}$ in two settings:

	Domain	Bias a_2	ξ_β search
SSMax-inspired fit (Nakanishi, 2025) (= M1)	$n \in [1, 1024]$	free	grid
Ours, grid (= M2)	$n \in \{64, \dots, 1024\}$	$\equiv 0$	grid (step 0.05)
Ours, OLS slope	$n \in \{64, 128, \dots, 1024\}$ (six points)	$\equiv 0$	continuous

The body figures Figures 1c and 1d use the grid version of Ours. Figure B.1 reports the M2 loss landscape on the same fine grid for all four runs at three training milestones; the rotary runs are added here only as a reference, since the body intentionally does not contrast positional encodings.

Why $a_2 \equiv 0$. The per-(layer, head) realization (A.1) already places a layer-side scalar bias inside the inverse-temperature parameterization. A free a_2 in the post-hoc fit would re-introduce that same bias on top, leaving ξ_β under-identified by the residual curvature of β_n. Empirically, on the trained checkpoints, M1 (a_2 free) selects $\xi_\beta = 1$ in only $3/28$ (11%) snapshots and otherwise wanders the grid with no stable trend across iters; the bias absorbs the signal. Forcing $a_2 \equiv 0$ removes this absorption and lets ξ_β identify the residual log-log slope of β_n.

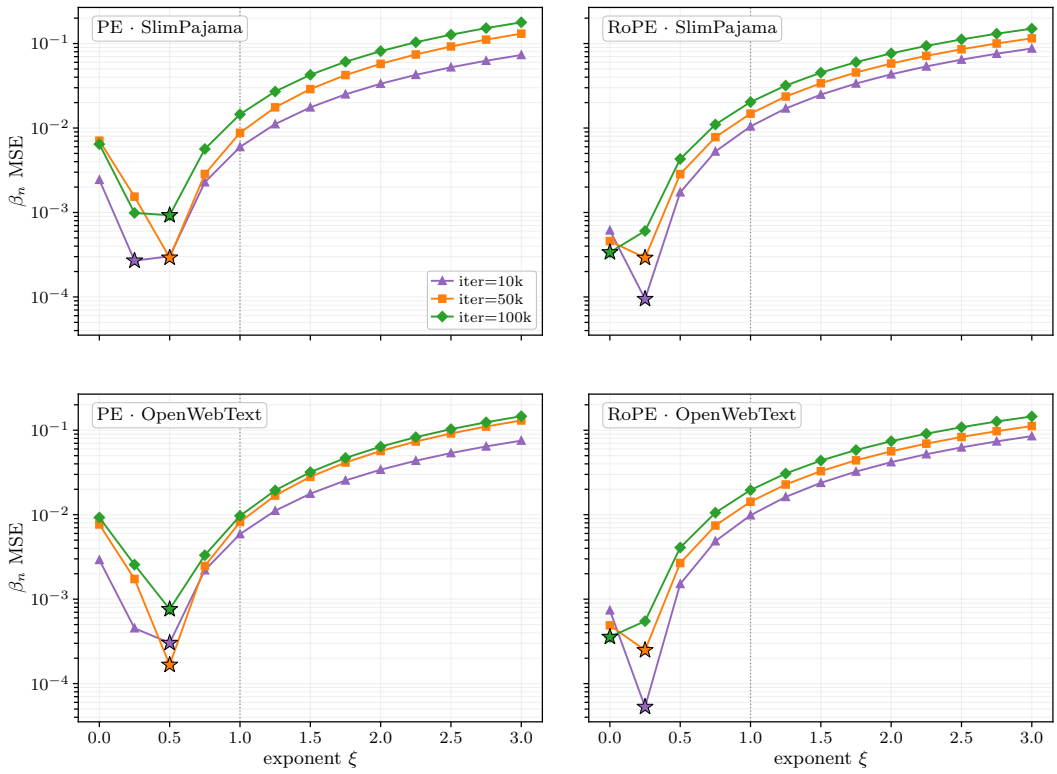


Figure B.1: M2 loss landscape on all four training-time GPT-124M runs. The y-axis is the per-point MSE of the bias-free post-hoc least-squares fit of β_n to $a_1(\log n)^{\xi_\beta}$ on $n \geq 64$, on a log scale; one curve per snapshot iter $\in \{10\text{k}, 50\text{k}, 100\text{k}\}$, with stars at per-curve argmin and a dotted vertical reference at the logarithmic value $\xi_\beta = 1$. PE and RoPE runs settle at distinct exponents within each dataset.

Why $n \geq 64$. The trimming threshold matches the smallest n used by the body's gap-counting estimator $\hat{\xi}_\Lambda$ of Definitions 3.3 and 3.5, namely $n = 64$ (the smallest of the six n values in Appendix F). Defining $\hat{\xi}_\beta$ on the same domain makes the comparison with $\hat{\xi}_\Lambda$ direct rather than a domain-mismatch artefact; the residual finite-n gap $|\hat{\xi}_\Lambda - \hat{\xi}_\beta|$ is the value reported per cell in Table 2. The first half of the context length $n \leq 50$ also contains the small-n dynamics of β_n that deviate from the asymptotic power family; trimming there isolates the scaling regime that Corollary 3.9 characterizes.

Trim sensitivity. Table B.1 reports $\hat{\xi}_\beta$ as the trim threshold varies, on the two absolute-PE runs at iter $= 100\text{k}$. The $n \geq 64$ choice sits in a stable plateau (within ± 0.04 of the $n \geq 51$ value), and pushing the threshold to $n \geq 200$ shifts $\hat{\xi}_\beta$ by ≤ 0.16 on the more sensitive dataset (OpenWebText), consistent with β_n being well approximated by the power family on the asymptotic tail.

trim	$n \geq 10$	$n \geq 25$	$n \geq 51$	$n \geq 64$	$n \geq 100$	$n \geq 200$
SlimPajama	0.305	0.336	0.368	0.380	0.410	0.477
OpenWebText	0.375	0.414	0.457	0.474	0.517	0.618

Table B.1: $\hat{\xi}_\beta$ from the continuous OLS fit of β_n to $a_1(\log n)^{\xi_\beta}$ on $n \geq$ trim, two absolute-PE training-time GPT-124M runs at iter $= 100\text{k}$. The body uses the discrete six-point variant on $n \in \{64, 128, 256, 512, 768, 1024\}$, which gives $\hat{\xi}_\beta = 0.346$ (SP) and 0.418 (OWT).

B.2 Error-bar widths for the body four-panel figure

The cell-wise relative upper bounds quoted in Figure 1 (SE on PPL $\leq 10\%$ of the cell mean in (a, b); half-IQR on MSE $\leq 8\%$ of the point-estimate MSE in (c, d)) are derived by the protocol below. Each width is the natural sampling uncertainty of the y-axis quantity at that panel type, reduced to a single per-panel number by taking the maximum across the cells (Qwen) or ξ_β grid points (nanoGPT) plotted in the panel.

Panels (a, b): doc-level standard error on PPL. At each $(\xi_\beta, n_{\text{eval}})$ cell, the body figure plots the per-document strict perplexity $\text{PPL}^{(d)}_{\text{strict}}$ averaged over the n_{doc} documents that fit at that n_{eval}, with $n_{\text{doc}} \in \{14, 15, \ldots, 19\}$ on PG-19 and $n_{\text{doc}} = 20$ on Proof-Pile-2 (Appendix A.2). The reported quantity is the doc-level standard error of the cell mean $\text{SE} = s/\sqrt{n_{\text{doc}}}$, where s is the sample standard deviation of $\{\text{PPL}^{(d)}_{\text{strict}}\}_{d=1}^{n_{\text{doc}}}$; the cell-wise maxima of SE/mean across the 35 cells per dataset are 9.52% on PG-19 and 8.79% on Proof-Pile-2 (both attained at $\xi_\beta = 0$, the length-multiplier-off endpoint). The cell-wise medians are 7.49% and 8.04% respectively, so the body's 10% envelope is loose by less than a factor of 1.3. A per-cell document-resampling bootstrap (B=200, percentile half-IQR of the cell mean) gives a tighter 4.95% / 5.32% but is dominated by the more conservative SE-based number quoted in the body caption.

Panels (c, d): residual bootstrap of the M2 MSE curve. At each ξ_β on the body grid, the MSE plotted in (c, d) is a deterministic function of a single trained β_n vector (Appendix B.1); its sampling uncertainty arises from the regression residuals on positions $n \in [64, 1024]$. Let $\hat{a}_1$ denote the M2 estimate at the given ξ_β and $e_n = \beta_n - \hat{a}_1(\log n)^{\xi_\beta}$ the residual. We resample e_n with replacement across positions $B = 200$ times; for each resample e^*_n we form $y^*_n = \hat{a}_1(\log n)^{\xi_\beta} + e^*_n$, refit a_1, and recompute the per-point MSE. The width is the half-IQR of these B MSE values divided by the point-estimate MSE. Across all six (dataset, iter) snapshots (Figure B.1), the per-grid maximum of this ratio peaks at 7.29% (SlimPajama, $\text{iter} = 50\text{k}$, near $\xi_\beta \approx 0.55$); medians stay between 2.7% and 2.9%.

C Auxiliary results for the model-agnostic theory

This appendix collects strands deferred from Section 3. Appendix C.1 reduces the supremum (3.3) defining Λ_n to a finite maximum over competitor gaps. Appendix C.2 handles the degenerate case of ties at the maximum. Appendix C.3 proves the coordinate decomposition of Corollary 3.9. Appendix C.4 gives the score-realizability reduction used throughout Section 3.4: any positive score matrix arises from self-attention with $Q_n = K_n$.

Throughout this appendix we write $\mathcal{U}_n := \{z^*_n - z_{n,j} : 1 \leq j \leq n,\ z_{n,j} < z^*_n\}$ for the set of positive competitor gaps.

C.1 Discrete representation of Λ_n

Lemma C.1 (Discrete representation of Λ_n)**.** *On the finite-Λ_n domain ($N_n(0) = 1$), the upper-tail accumulation scale of Definition 3.4 satisfies*

$$\Lambda_n = \sup_{t>0} \frac{\log N_n(t)}{t} = \max_{u \in \mathcal{U}_n} \frac{\log N_n(u)}{u}.$$

If $N_n(0) \geq 2$ (ties at the maximum), then $\Lambda_n = \infty$.

Proof. $t \mapsto N_n(t)$ is right-continuous and piecewise constant with jumps only at points of $\{0\} \cup \mathcal{U}_n$, so $\log N_n(t)$ is constant on each interval $[u, u')$ between consecutive points of $\{0\} \cup \mathcal{U}_n \cup \{+\infty\}$, while the denominator t is increasing. Hence $t \mapsto \log N_n(t)/t$ is nonincreasing on each such interval, and the supremum is attained at some $u \in \mathcal{U}_n$. If $N_n(0) \geq 2$, $N_n(t) \geq 2$ for every $t > 0$ and $\log N_n(t)/t \to \infty$ as $t \downarrow 0$. □

Sorted-rank rendering. The body figure Figure 2 plots the counting-function view $t \mapsto N_n(t)$. The same data admits an equivalent rendering as the sorted competitor gaps $j \mapsto g_{n,(j)}$, obtained

by swapping the two axes $j \leftrightarrow N$: the right-continuous step $N_n(t)$ on log y maps to the ascending staircase $g_{n,(j)}$ on log x, and the exponential envelope $N = e^{\Lambda_n t}$ inverts to its logarithmic form $t = \log j/\Lambda_n$. The contact point $(\Delta_n, N_n(\Delta_n))$ on the right is the same point (j^*, Δ_n) with $j^* = N_n(\Delta_n)$ on the left, and Lemma C.1's claim that the supremum is attained at a competitor gap is the statement that the sorted-rank staircase touches the envelope on its riser, not its tread. Figure C.1 shows both renderings together.

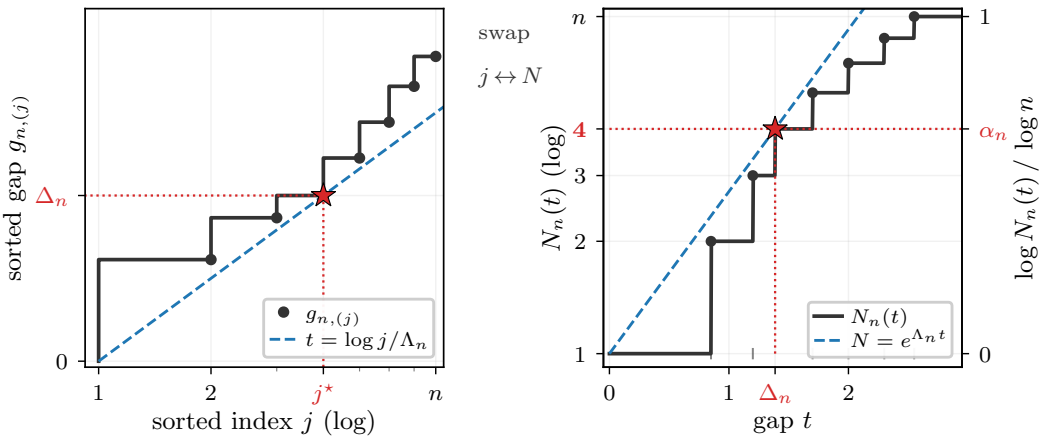


Figure C.1: Two equivalent renderings of the same N_n. *Right:* the cumulative gap-count curve $t \mapsto N_n(t)$ on the gap axis $t \geq 0$ and a log y-axis, with the dashed exponential envelope $N = e^{\Lambda_n t}$ (3.2); the red star marks the contact point $(\Delta_n, N_n(\Delta_n))$, and the secondary right axis reads $\alpha_n = \log N_n(\Delta_n)/\log n$. This is the panel reproduced in the body as Figure 2. *Left:* the equivalent sorted-competitor-gap rendering $j \mapsto g_{n,(j)}$ on a log j-axis, with $t = \log j/\Lambda_n$ as the inverse form of the accumulation envelope; the red star sits at the same contact point with horizontal coordinate $j^* = N_n(\Delta_n)$. The two panels are related by the axis swap $j \leftrightarrow N$.

C.2 Ties at the maximum

Corollary C.2 (Ties at the maximum)**.** *Let*

$$n_{\max} := N_n(0) = \#\{j \in \{1, \dots, n\} : z_{n,j} = z_n^*\}.$$

Then:

1. *if $n_{\max} \geq 2$, then $\Lambda_n = \infty$;*
2. *for every $\beta > 0$,*

$$\max_j p_{n,j}(\beta) \leq \frac{1}{n_{\max}}, \qquad H_n(\beta) \geq \log n_{\max}.$$

In particular, if $n_{\max} \geq 2$ for infinitely many n, then no finite TE scale exists.

Proof. If $n_{\max} \geq 2$, then $z_n^* - z_{n,j} = 0$ for the $n_{\max}$ tokens at the maximum, so $N_n(t) \geq n_{\max}$ for every $t > 0$. Consequently

$$\frac{\log N_n(t)}{t} \geq \frac{\log n_{\max}}{t} \to \infty \qquad (t \downarrow 0),$$

so $\Lambda_n = \infty$ by Definition 3.4. The top $n_{\max}$ softmax coordinates are equal, and their total mass is at most 1, so each of them is at most $1/n_{\max}$. Thus $\max_j p_{n,j}(\beta) \leq 1/n_{\max}$. Since every softmax coordinate is at most $1/n_{\max}$,

$$H_n(\beta) = -\sum_{j=1}^{n} p_{n,j}(\beta) \log p_{n,j}(\beta) \geq -\sum_{j=1}^{n} p_{n,j}(\beta) \log(1/n_{\max}) = \log n_{\max}.$$

For the final assertion, suppose $(X_n) \subset (0, \infty)$ is a finite TE scale and set $\beta_n := nX_n$, so $\beta_n/X_n = n \to \infty$. Then Definition 3.2 forces $H_n(\beta_n) \to 0$, contradicting $H_n(\beta_n) \geq \log n_{\max} \geq \log 2$ on the infinite subsequence with $n_{\max} \geq 2$. □

C.3 Coordinate decomposition

Proof of Corollary 3.9. Multiplying $\asymp$-classes in the contact formula (3.4) gives

$$\begin{aligned}\Lambda_n &= \alpha_n \,(\log n)\, \Delta_n^{-1} \\ &\asymp (\log n)^{\xi_\alpha}\,(\log n)\,(\log n)^{-\xi_\Delta} = (\log n)^{\xi_\alpha - \xi_\Delta + 1},\end{aligned}$$

and uniqueness of the exponent yields (3.7). Solving the contact formula for α_n or Δ_n covers the symmetric directions. □

C.4 Score realizability via positive Gram matrices

Definition C.3 (Positive matrix)**.** A matrix $A \in \mathbb{R}^{n\times n}$ is *positive semidefinite* if $x^\top A x \geq 0$ for every $x \in \mathbb{R}^n$. We call A *positive*, and write $A \geq 0$, if A is symmetric and positive semidefinite. Every Gram matrix considered in the sequel is positive in this sense, so we use the predicate $A \geq 0$ throughout.

Proposition C.4 (Gram reduction for positive score matrices)**.** *Let $\Sigma_n \in \mathbb{R}^{n\times n}$ satisfy $\Sigma_n \geq 0$ with $\operatorname{rank}\Sigma_n \leq d_{\mathrm{qk},n}$. Then there exist matrices $Q_n, K_n \in \mathbb{R}^{d_{\mathrm{qk},n}\times n}$ with $Q_n = K_n$ such that the self-attention score matrix equals Σ_n:*

$$\frac{\langle q_{n,i}, k_{n,j}\rangle}{\sqrt{d_{\mathrm{qk},n}}} = (\Sigma_n)_{ij}, \qquad 1 \leq i,j \leq n.$$

In particular, every row of Σ_n is realized as a fixed-query self-attention score vector.

Proof. Choose $B_n \in \mathbb{R}^{d_{\mathrm{qk},n}\times n}$ with $B_n^\top B_n = \Sigma_n$, padding by zero rows if necessary. Setting $Q_n = K_n = d_{\mathrm{qk},n}^{1/4} B_n$ gives $Q_n^\top K_n/\sqrt{d_{\mathrm{qk},n}} = B_n^\top B_n = \Sigma_n$. □

D Rank-collapse boundary and resolved accumulation

The body's TE scale Λ_n controls the top-two/entropy pair via Theorem 3.6 but does not by itself imply coordinate-wise rank collapse (Section 3.1). This appendix records the exact rank-collapse criterion, defines the corresponding *rank boundary*, identifies the positive-resolution accumulation proxy that sandwiches it, and proves Corollary 3.7.

Define the *rank free energy*

$$\mathcal{F}_n(\beta) := \log Z_n(\beta). \tag{D.1}$$

Proposition D.1 (Exact coordinate-wise rank-collapse criterion)**.** *Let*

$$\mathcal{G}_n(\beta) := \max_{i,j} \big| p_{n,i}(\beta) - p_{n,j}(\beta) \big|.$$

For any positive deterministic sequence (β_n),

$$\mathcal{G}_n(\beta_n) \to 0 \quad \Longleftrightarrow \quad \mathcal{F}_n(\beta_n) \to \infty.$$

Proof. The largest unnormalised softmax weight equals 1, so

$$p_n^*(\beta) := \max_j p_{n,j}(\beta) = \frac{1}{Z_n(\beta)} = e^{-\mathcal{F}_n(\beta)}.$$

Since $\mathcal{G}_n(\beta) \leq p_n^*(\beta)$, the condition $\mathcal{F}_n(\beta_n) \to \infty$ implies $\mathcal{G}_n(\beta_n) \to 0$. Conversely, $\min_j p_{n,j}(\beta) \leq 1/n$, hence

$$p_n^*(\beta) \leq \mathcal{G}_n(\beta) + \frac{1}{n}.$$

Thus $\mathcal{G}_n(\beta_n) \to 0$ implies $p_n^*(\beta_n) \to 0$, which is equivalent to $\mathcal{F}_n(\beta_n) \to \infty$. □

Definition D.2 (Rank boundary)**.** For $0 \leq r \leq \log n$, define

$$B_n^{\mathrm{rank}}(r) := \sup\big\{\beta \geq 0 : \mathcal{F}_n(\beta) \geq r\big\}.$$

Proposition D.1 implies that coordinate-wise rank collapse holds along (β_n) if and only if there exists $r_n \to \infty$, $r_n \le \log n$, such that

$$\beta_n \le B_n^{\mathrm{rank}}(r_n)$$

eventually.

Definition D.3 (Resolved accumulation scale)**.** For $r \ge 0$, define

$$\Lambda_n^{(r)} := \sup_{t>0:\, \log N_n(t) > r} \frac{\log N_n(t) - r}{t},$$

with the convention $\sup \emptyset = 0$.

At zero resolution,

$$\Lambda_n^{(0)} = \Lambda_n.$$

Thus the upper-tail accumulation scale of the body is the *zero-resolution accumulation scale*. The rank-collapse boundary is obtained by taking a positive diverging resolution $r_n \to \infty$, which removes finite near-tie clusters from the envelope.

Proposition D.4 (Resolved accumulation implies rank-gap collapse)**.** *Let $r_n \to \infty$. If*

$$\beta_n < \Lambda_n^{(r_n)}$$

eventually, then $\mathcal{G}_n(\beta_n) \to 0$.

Proof. By the definition of $\Lambda_n^{(r_n)}$, for all large n there exists $t_n > 0$ such that

$$\beta_n t_n < \log N_n(t_n) - r_n.$$

The $N_n(t_n)$ tokens within gap t_n each contribute at least $e^{-\beta_n t_n}$ to the partition function. Hence

$$Z_n(\beta_n) \ge N_n(t_n)\, e^{-\beta_n t_n} > e^{r_n}.$$

Since $r_n \to \infty$, Proposition D.1 gives $\mathcal{G}_n(\beta_n) \to 0$. □

Define the Laplace envelope

$$S_n(\beta) := \sup_{t \ge 0} \big\{ \log N_n(t) - \beta t \big\}. \tag{D.2}$$

Lemma D.5 (Laplace-envelope sandwich)**.** *For every $\beta > 0$,*

$$S_n(\beta) \le \mathcal{F}_n(\beta) \le S_n(\beta) + \log(1 + \log n).$$

Proof. For the lower bound, fix $t \ge 0$. The $N_n(t)$ tokens with gap at most t each contribute at least $e^{-\beta t}$, hence

$$Z_n(\beta) \ge N_n(t)\, e^{-\beta t}.$$

Taking the supremum over t gives $S_n(\beta) \le \mathcal{F}_n(\beta)$.

For the upper bound, write the ordered competitor gaps as $0 = t_{n,1} \le t_{n,2} \le \cdots \le t_{n,n}$. Since $N_n(t_{n,j}) \ge j$, the definition of S_n implies

$$\log j - \beta t_{n,j} \le S_n(\beta).$$

Thus

$$e^{-\beta t_{n,j}} \le \frac{e^{S_n(\beta)}}{j}.$$

Summing over j gives

$$Z_n(\beta) \le e^{S_n(\beta)} \sum_{j=1}^{n} \frac{1}{j} \le e^{S_n(\beta)}\, (1 + \log n).$$

Taking logarithms proves the upper bound. □

Consequently, for $r > \log(1 + \log n)$,

$$\Lambda_n^{(r)} \le B_n^{\mathrm{rank}}(r) \le \Lambda_n^{(r-\log(1+\log n))}.$$

Therefore, at resolutions $r_n \gg \log\log n$, the exact rank boundary is described by the resolved accumulation scale up to the envelope discretisation error $\log(1 + \log n)$.

Corollary D.6 (Contact-count entropy criterion for rank collapse)**.** *Assume $\beta_n/\Lambda_n \to 0$. If the contact-count entropy $C_n := \log N_n(\Delta_n) = \Lambda_n \Delta_n$ satisfies $C_n \to \infty$, then $\mathcal{G}_n(\beta_n) \to 0$.*

Proof. Set $s_n = \beta_n/\Lambda_n$. The contact gap Δ_n gives

$$Z_n(\beta_n) \ge N_n(\Delta_n)\, e^{-\beta_n \Delta_n} = \exp\{(1 - s_n)\, C_n\}.$$

Since $s_n \to 0$ and $C_n \to \infty$, the right-hand side diverges. Proposition D.1 applies. □

Proof of Corollary 3.7. The supercritical implication is the supercritical direction of Theorem 3.6 and uses no contact-count hypothesis. The subcritical implication is exactly Corollary D.6. □

Example D.7 (Finite contact count: TE and RE scales separate)**.** The condition $C_n \to \infty$ in Corollary 3.7 cannot be dropped: the score family below has $C_n = \log 2$ (bounded) and rank collapse fails, even though (Λ_n) is the TE scale. Set

$$z_{n,1} = 0, \qquad z_{n,2} = -\frac{\log 2}{\log n}, \qquad z_{n,j} = -\log n \ \ (j \ge 3),$$

so that $z_n^* = 0$ and the competitor gaps are $z_n^* - z_{n,2} = (\log 2)/\log n$ and $z_n^* - z_{n,j} = \log n$ for $j \ge 3$. Then $\Lambda_n = \log n$, $\Delta_n = (\log 2)/\log n$, and $\alpha_n = (\log 2)/\log n \to 0$. With $\beta_n = \sqrt{\log n}$ so that $\beta_n/\Lambda_n \to 0$, the softmax row converges to $(\frac{1}{2}, \frac{1}{2}, 0, \ldots, 0)$: top-two collapse holds, but rank collapse fails. A rank-resolved analysis is given in Example D.8.

Example D.8 (Rank-resolved analysis of the finite-contact case)**.** For the score family of Example D.7, $\Lambda_n = \log n$ because the contact gap $\Delta_n = (\log 2)/\log n$ attains the zero-resolution envelope. The two competitor gap levels satisfy $\log N_n((\log 2)/\log n) = \log 2$ and $\log N_n(\log n) = \log n$. For any $r_n \to \infty$ with $r_n = o(\log n)$, eventually $r_n > \log 2$, so the contact gap Δ_n is excluded by the constraint $\log N_n(t) > r_n$ in Definition D.3, leaving only $t = \log n$:

$$\Lambda_n^{(r_n)} = \frac{\log n - r_n}{\log n} \sim 1,$$

where $f_n \sim g_n$ means $f_n/g_n \to 1$ for positive sequences. Thus the TE scale is order $\log n$, while the rank-resolved accumulation scale is order one. With $\beta_n = \sqrt{\log n}$, $Z_n(\beta_n) \to 2$ and $\mathcal{G}_n(\beta_n) \to \frac{1}{2}$.

D.1 Empirical instances on the verification cells

We examine the verification cells of Table 2 in light of the condition in Corollary 3.7.

TE-to-RE upgrade. Theorem 3.6 identifies Λ_n as the TE scale unconditionally; the upgrade to RE scale via Corollary 3.7 requires the contact-count entropy $C_n = \alpha_n \log n$ to diverge, equivalently $\xi_\alpha > -1$. Three of the four nanoGPT cells (SlimPajama at 25k and 100k, and OpenWebText at 100k) have $1 + \hat{\xi}_\alpha > 0$, consistent with slow C_n divergence. OpenWebText at 25k instead has $\hat{\xi}_\alpha = -1.03$, so this estimate does not support the RE-scale upgrade. The four Qwen cells have $\hat{\xi}_\alpha \in (-1.69, -1.40)$ on the observed window; since $C_n \ge \log 2$ always, the bucket-mean slope $\hat{\xi}_\alpha < -1$ cannot persist asymptotically, and the analytical RE-scale upgrade of Corollary 3.7 does not directly apply, with the rank-resolved boundary of Proposition D.1 as the formal substitute. The empirical boundary character of Λ_n on these cells is then verified directly below.

Boundary check on Λ_n. Independently of the analytical sufficient condition, the boundary character of Λ_n can be tested by reading the per-row maximum softmax weight $p_n^*(\beta) = 1/Z_n(\beta)$ at $\beta = \gamma\Lambda_n$ on a sweep around criticality. Table D.1 reports the median p_n^* at $\gamma \in \{0.5, 1.0, 1.5\}$ and the per-cell fraction of rows with $p_n^* \le 1/\log n$ at $\gamma = 0.1$ (strict subcritical) on the four Qwen cells. Across all cells the median traces a smooth monotone curve from $\sim 10^{-3}$ at $\gamma = 0.1$ to ~ 1 at $\gamma = 4$, crossing

		median p_n^* at $\beta = \gamma\Lambda_n$			frac. $p_n^* \le 1/\log n$
dataset	n_{eval}	$\gamma = 0.5$	$\gamma = 1.0$	$\gamma = 1.5$	at $\gamma = 0.1$
PG-19	16384	0.20	0.45	0.64	89%
PG-19	32768	0.17	0.42	0.63	91%
Proof-Pile-2	16384	0.18	0.42	0.63	89%
Proof-Pile-2	32768	0.15	0.40	0.62	93%

Table D.1: Boundary check via $p_n^*(\gamma\Lambda_n)$ on Qwen-7B-Chat attention rows at $\xi_0 = 2$ (45,467 rows total across the four cells, rows with $\log_{10}\Lambda_n > 5$ excluded). Median p_n^* at $\gamma \in \{0.5, 1.0, 1.5\}$ and per-cell fraction with $p_n^* \le 1/\log n$ at $\gamma = 0.1$. On the wider sweep $\gamma \in \{0.1, 0.25, 0.5, 0.75, 1.0, 1.5, 2.0, 4.0\}$ the median p_n^* is monotone in γ on every cell, ranging from $\sim$0.001 at $\gamma = 0.1$ to $\sim$0.96 at $\gamma = 4$, and the same monotonicity holds on the nanoGPT cells.

~ 0.42 at $\gamma = 1$, with $\sim 90\%$ of rows below $1/\log n$ at $\gamma = 0.1$ and 0% at $\gamma \ge 1$. The framework's Λ_n separates the subcritical-collapse regime from the supercritical-concentration regime empirically, including on the Qwen cells where the contact- count entropy is bounded and Corollary 3.7 does not directly apply. Example D.7 shows that this analytical upgrade is sharp in the worst case (a contrived score family with C_n bounded fails rank collapse), but the Qwen rows in the bounded- C_n regime do not realize that pathology at the operating β.

E Derivation of the YaRN inverse-temperature scaling

This appendix spells out why the YaRN context-window extension of Peng et al. (2024) produces an effective inverse-temperature scaling $\beta_n \asymp (\log n)^2$, i.e. implicit exponent $\xi = 2$ in the parameterisation of Section 2.2.

Peng et al. (2024) replace the standard softmax logit $q_m^\top k_n/\sqrt{d}$ with $q_m^\top k_n/(t\sqrt{d})$ for a length-dependent temperature $t = t(s)$, where $s = n/n_{\text{train}}$ is the ratio of the target context length n to the training window n_{train}. The recommended schedule (their §3.3) is

$$\tfrac{1}{\sqrt{t(s)}} \;=\; 1 + 0.1\,\ln s,$$

and the multiplier $1/\sqrt{t(s)}$ is *multiplied onto the rotary embeddings of both the query and the key*. Substituting into the logit gives a rescaling

$$\beta_n \;\propto\; 1/t(n/n_{\text{train}}) \;=\; \big(1 + 0.1\ln(n/n_{\text{train}})\big)^2 \;\asymp\; (\log n)^2,$$

i.e. a scalar multiplier $\asymp (\log n)^2$ asymptotically. The fact that the multiplier is applied to both sides of the inner product, rather than once, is what doubles the exponent relative to a $\log n$-only temperature, and places YaRN in the $\xi = 2$ row of Table 1.

F Training-time exponent measurement

This appendix gives the per-row gap-counting acquisition protocol, the β_n regression-fit protocol, and limitations for the training-time consistency check of Section 4.1. The upstream training setup that produces the learnable inverse-temperature vector β_n being measured here, namely the models, datasets, four-configuration schedule, and the per-(layer, head) realization (A.1), is documented in Appendix A.3; the body β_n regression fit itself is in Appendix B.1.

Gap-counting estimator $\hat{\xi}_\Lambda$. We analyse the four training-time GPT-124M runs (two PE datasets, two RoPE) of Appendix A.3 at the seven milestone checkpoints. For each checkpoint, three evaluation seeds produce independent validation batches ($B = 16$ sequences, context lengths $n \in \{64, 128, 256, 512, 768, 1024\}$, block size 1024). For each (layer, head, n, batch) row we record the current-layer dot product $z_{n,j} = \langle q, k_j\rangle/\sqrt{d}$ with $z_n^* := \max_j z_{n,j}$. Per-row Λ_n and the contact coordinates (α_n, Δ_n) are computed from Definition 3.5 via the competitor gaps $z_n^* - z_{n,j}$ at numerical

tolerance $\epsilon = 10^{-6}$. The body's $\hat{\xi}_\Lambda$ is the bucket-mean slope estimator of Section 4.1: rows are pooled across evaluation seeds into the index sets $\mathcal{R}(c, n)$, the bucket mean is taken over those pooled rows at each of the six n values, and $\hat{\xi}_\Lambda$ is the weighted least-squares slope of those bucket means against $\log \log n$, with weights equal to the retained row counts $K(c, n)$. The same fit gives $\hat{\xi}_\alpha$ and $\hat{\xi}_\Delta$. The body's headline half-width is the bootstrap-IQR/2 over (s, ℓ, h) tuples of Section 4.1. Rows with $\log_{10} \Lambda_n > 5$ are dropped before averaging, matching the Qwen protocol of Appendix G.1.

Uncertainty estimates. For each cell c let $\mathcal{T}(c)$ be the cell's bootstrap-unit set: $\{(\ell, h)\}$ on Qwen and $\{(s, \ell, h)\}$ on nanoGPT, the latter pooled across the three evaluation seeds. Each draw selects $|\mathcal{T}(c)|$ tuples uniformly with replacement from $\mathcal{T}(c)$, recomputes the bucket-mean slope $\hat{\xi}_X(c)$ from the rows in $\mathcal{R}(c, n)$ that belong to the drawn tuples (separately at each $n \in \mathcal{N}(c)$), and the half-IQR over $B = 200$ such draws is the half-width reported in Table 2; tie % follows the same convention. The ξ^*_β column carries no interval on Qwen (0.5-grid PPL minimum) and the six-point OLS standard error of $\hat{\xi}_\beta$ on nanoGPT.

β_n-fit estimator $\hat{\xi}_\beta$. At every checkpoint β_n is read off the model state and fitted to the bias-free power family $\beta_n = a_1 (\log n)^{\xi_\beta}$ at the same six n that the gap-counting uses. The OLS slope of $\log \beta_n$ on $\log \log n$ at those six points yields $\hat{\xi}_\beta$ (closed-form least squares, no grid). The shared domain matters: Theorem 3.6 predicts equality of $\hat{\xi}_\Lambda$ and $\hat{\xi}_\beta$ in the limit, but only when both estimators are read on the same n set; this is the choice we make throughout. The bias-free constraint $a_2 \equiv 0$ is justified in Appendix B.1.

Limitations. The check is limited in two respects. (i) Each configuration uses a single training seed. The $\hat{\xi}_\Lambda$ band is over three evaluation batches at fixed training seed; variation from initialisation and data ordering is not measured. $\hat{\xi}_\beta$ has no eval-seed variance by construction. (ii) Contexts are short: with $n \leq 1024$, the log-log dynamic range covers half a decade, which constrains the bucket-mean slope but leaves higher-order curvature unconstrained; this is part of why $\hat{\xi}_\Lambda$ and $\hat{\xi}_\beta$ retain a finite-n gap.

G Attention-score acquisition on Qwen-7B-Chat

This appendix records the per-row attention acquisition pipeline that backs the verification analysis of Section 4 and the body figures it feeds.

G.1 Acquisition protocol

We use Qwen-7B-Chat with dynamic-NTK RoPE enabled (Appendix A.2). Attention rows are collected from all 32 layers and 32 heads on PG19 and Proof-Pile-2 at context lengths $n_{\text{eval}} \in \{8192, 16384, 32768\}$, with four context-length values per document at the ratios $n/n_{\text{eval}} \in \{1/4, 1/2, 3/4, 1\}$ and one query position per (layer, head, document, n) tuple. The body's per-n_{eval} fit of Section 4.1 skips $n_{\text{eval}} = n_{\text{train}} = 8192$ (training window, no extrapolation). The acquisition pass applies the inverse-temperature multiplier $m(n; \xi_0)$ of Appendix A.2 in every layer at the seven values $\xi_0 \in \{0, 0.5, 1, 1.5, 2, 2.5, 3\}$, where $\xi_0 = 0$ disables the multiplier; ξ_0 perturbs the representation entering a layer while the measured object is always the unrescaled current-layer score $z_{n,j} = \langle \mathbf{q}_{n,i_n}, \mathbf{k}_{n,j} \rangle / \sqrt{d_{qk}}$. Across the three context lengths and four ratios, this yields eight values $n \in \{2048, 4096, 6144, 8192, 12288, 16384, 24576, 32768\}$. The published-paper headline aggregation (Section 4.1 and Table 2) uses only $\xi_0 = 2$ at $n_{\text{eval}} \in \{16384, 32768\}$, with 4 documents from PG-19 and 5 from Proof-Pile-2. This gives 73,728 attention rows before the tie filter below and 68,212 after it. The empirical contact gap is the largest competitor gap u satisfying $\log N_n(u)/u \geq (1 - \epsilon)\, \Lambda_n$ at $\epsilon = 10^{-6}$, a floating-point relaxation of the exact maximizer in Definition 3.5.

Per-n_{eval} fit. The body $\hat{\xi}_{\text{pool}}(n_{\text{eval}})$ is the weighted least-squares slope of the means of $\log \Lambda_n$ at the four n values within an n_{eval}, regressed against $\log \log n$ with weights equal to the retained row counts $K(c, n)$. The same fit gives $\hat{\xi}_\alpha$ and $\hat{\xi}_\Delta$. Attention rows with $\log_{10} \Lambda_n > 5$ are excluded from

the bulk before averaging. Aggregating across n_{eval} at shared n mixes three dynamic-NTK RoPE rescale conditions (one per source n_{eval}) and biases the slope; per-n_{eval} fits avoid this contamination.

Representative rows shown in body figures. The single attention rows recast as $N_n(t)$ in Figure 4 (middle, right) are selected deterministically from the protocol above: for each contrast cell c, take the row whose $(\log \alpha_n, \log \Delta_n)$ is closest to the cell's bucket mean $(\overline{\log \alpha}(c,n), \overline{\log \Delta}(c,n))$ in $(\log \alpha_n, \log \Delta_n)$ at the largest $n \in \mathcal{N}(c)$. The two cells used in body Figure 4 are Qwen-7B-Chat PG-19 at $n_{\text{eval}} = 32768$ ($n = 32768$) and nanoGPT SlimPajama at $\text{iter} = 100\text{k}$ ($n = 1024$); the contact statistics of the selected rows are $(\hat{\Lambda}_n, \hat{\alpha}_n, \hat{\Delta}_n) \approx (4.22, 0.13, 0.33)$ and $(1.64, 0.26, 1.09)$ respectively.