

Article

Curvature-Corrected Rotary Position Embeddings: An Entropy-Invariant Temperature for Mitigating Numerical-Rank Collapse in Long-Context Attention

Joseph Tafataona Mtetwa ^{1,*} , Kingsley A. Ogudo ¹  and Sameerchand Pudaruth ² ¹ Department of Electrical and Electronics Engineering, University of Johannesburg, Johannesburg 2006, South Africa; kingsleyo@uj.ac.za² ICT Department, University of Mauritius, Réduit 80837, Mauritius; s.pudaruth@uom.ac.mu

* Correspondence: joemtetwa@gmail.com

Abstract

As large language models process ever longer contexts, the positional encoding—most commonly rotary position embeddings (RoPEs)—must remain numerically trustworthy. We give an information-geometric analysis of why the attention matrix becomes ill-conditioned as the sequence length L grows. We first show that the tempting frequency-aliasing explanation fails, because RoPE's multi-frequency code keeps positions well separated. The mechanism is statistical: at fixed softmax temperature the attention entropy grows like $\log L$, and the rows spread their mass over a collision support (inverse participation ratio) that grows polynomially as L^a ($a \approx 0.9$ measured), confining the matrix's trailing singular values and producing a numerical-rank collapse. We prove, and confirm in 50-digit arithmetic, that the spectral condition number is intrinsically ill-posed here, so the numerical rank is the correct diagnostic. From the exact entropy–temperature identity $dH/d\beta = -\beta \text{Var}_p(e)$ we derive the entropy-invariant schedule $\beta_*(L) = (L/L_{\text{ref}})^{c_0}$, Curvature-Corrected RoPE (CC-RoPE), and prove that in the diffuse near-circulant regime it eliminates the collapse at a characterised conditioning–mixing cost, whereas no relative-position-preserving warp can. The two signatures are further confirmed in a trained RoPE model (Pythia-160M); the downstream perplexity effect is framed as a falsifiable prediction.

Keywords: rotary position embeddings; attention entropy; numerical rank; information geometry; Fisher–Rao metric; long-context transformers; numerical stability

MSC: 53B21; 15A12; 62B11; 68T07



Academic Editor: Huawen Liu

Received: 29 May 2026

Revised: 1 July 2026

Accepted: 8 July 2026

Published: 20 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Whether a language model can reason over millions of tokens in a single forward pass depends not only on its scale and training, but on whether the positional encoding remains numerically trustworthy at these lengths. Rotary position embeddings (RoPEs) [1], adopted across competitive large language models, rotate query and key vectors through block-diagonal matrices $R_\theta(t) \in \text{SO}(d)$ so that the attention score between positions t and s depends only on $t - s$; this relative-position property is the defining feature of RoPE.

As L climbs into the long-context range, the *attention matrix* $A \in \mathbb{R}^{L \times L}$ becomes numerically ill-conditioned. We first dispose of one diagnosis. Standard RoPE uses fixed frequencies $\theta_{\min} = 10^{-4}$ through $\theta_{\max}$, the lowest-frequency block completing one cycle every $\Delta t_{\text{full}} = 2\pi/\theta_{\min} \approx 62,832$ tokens; by $L = 10^6$ this block has aliased, which suggests

that tokens separated by multiples of Δt_{full} produce near-repeated rows in A . This is incorrect: RoPE's encoding is a *superposition of many frequencies*, and although the slowest block aliases, the faster blocks keep the combined code well separated. We verify numerically that positions one slow-period apart have large code distance and that the position-feature matrix is well conditioned (Section 6), so aliasing is not the source of the difficulty.

The mechanism is statistical rather than rotational. With a fixed softmax temperature, the entropy of each attention row grows like $\log L$ as the sequence lengthens—a phenomenon observed empirically in trained transformers [2,3]—so the rows drift toward the uniform distribution. As the rows flatten, the deviation of A from its mean row shrinks in aggregate, so a growing share of the matrix's *trailing* singular values sink toward the floating-point roundoff floor while many leading singular values remain $\Theta(1)$. The matrix is thus numerically rank-deficient without being close to low rank; we call this a *numerical-rank collapse*. We show below that the ordinary spectral condition number $\kappa(A)$ is, in this regime, dominated by the roundoff-floor singular values, so it reports the precision of the arithmetic rather than a well-posed property of the attention computation, and the appropriate diagnostic is the numerical rank.

We give an information-geometric account of this phenomenon and identify the quantity that controls it. The statistical manifold of attention distributions, under the Fisher–Rao metric, is isometric to a sphere, and on it the diffuse rows spread their ℓ_2 mass over a collision support (inverse participation ratio) that grows as L^a in the diffuse regime ($a \approx 0.9$ measured)—the scalar that drives the rank collapse. (An earlier form of this analysis attributed the collapse to a decay of the Fisher information; we show in Remark 4 that the Fisher information is constant, and that the operative quantity is the collision support.) From the exact identity $dH/d\beta = -\beta \text{Var}_p(e)$ and an empirically verified asymptotic entropy law, we obtain the inverse-temperature schedule holding the entropy—and hence the diffuseness—invariant in L : the polynomial law $\beta_*(L) = (L/L_{\text{ref}})^{c_0}$, with c_0 the measured entropy slope. We call this Curvature-Corrected RoPE (CC-RoPE), a correction with no trainable parameters (c_0 is calibrated, not learned), and prove that *in the full, diffuse near-circulant setting studied here it eliminates the collapse*: the rank deficit stays uniformly bounded (identically zero in our full-attention experiments) and the condition number is lifted off the floating-point floor, at a quantified conditioning–mixing cost since the same sharpening narrows the attention. We further show, by an obstruction argument and direct measurement, that a position-dependent phase *warp* of the RoPE trajectory cannot alter the diffuse structure driving the collapse, because it preserves the relative-position property.

1.1. Context and Motivation

Extending the effective context window of transformer models has been addressed by a range of strategies. On the encoding side, Chen et al. [4] proposed Position Interpolation, which compresses position indices to fit within the original training window. YaRN [5] improves on this with dimension-wise frequency scaling and an attention temperature correction. Resonance RoPE [6] addresses the specific challenge of out-of-distribution token positions in train-short-test-long scenarios. On the architectural side, LongNet [7] achieves linear complexity through dilated attention. LongLoRA [8] adapts pretrained models through shifted sparse attention. Dual Chunk Attention [9] extends LLaMA-2 to over 100k tokens without additional fine-tuning. Surveys such as [10] catalogue this landscape.

This body of literature does not address how the *spectral conditioning* of the attention matrix scales with sequence length. These methods address computational cost, memory footprint, and perplexity at extended lengths, but none provides a theoretical account of whether the attention weights themselves remain numerically reliable as L grows. The closest prior work in this direction is Baek's layer-wise theory of forward-error amplifica-

tion [11], which identifies when transformers become fragile under low-precision arithmetic, and the spectral conditioning approach of Saratchandran and Lucey [12], who improve the Jacobian conditioning of the projection matrices W_Q , W_K , and W_V . Neither, however, addresses the entropy-driven, sequence-length-dependent deterioration of the attention matrix itself that we characterise here, nor connects it to the Fisher–Rao geometry of the attention distributions.

In parallel, there has been growing recognition that Riemannian geometry provides a framework for understanding certain failure modes of transformer training. Kim et al. [13] demonstrated that the Fisher information metric—rather than the Euclidean norm—defines the appropriate geometry for sharpness-aware minimisation, yielding more generalisable solutions. Da Silva and Dangel [14] showed that standard sharpness measures are corrupted by the parameter-space symmetries of transformer attention heads, and that meaningful sharpness emerges after quotienting these symmetries on a Riemannian manifold. Ji’s RiemannFormer [15] takes the further step of placing token embeddings on a curved Riemannian space, using parallel transport to define geometrically consistent attention. Our work draws on this same tradition, but asks a different question: what is the intrinsic geometry of the attention distribution manifold, and what does it demand of the positional encoding?

1.2. Contributions

This paper makes the following contributions:

1. **Geometric foundations.** We prove that the standard RoPE curve $\gamma(t) = \exp(t\Omega)$ is a geodesic on $(\text{SO}(d), g_{\text{bi}})$ (Proposition 1), and that the statistical manifold of attention distributions, equipped with the Fisher–Rao metric, is isometric to a sphere of radius two with constant sectional curvature $1/4$ (Proposition 2). These are the geometric objects underlying the analysis of the attention computation.
2. **Correct diagnosis of the conditioning problem.** We show that the “frequency-aliasing” explanation does not hold: because the RoPE code superposes many frequencies, positions one slow-period apart remain well separated, and the position-feature matrix is well conditioned (Section 6). The operative mechanism is instead the *diffuseness* of the diffuse-attention regime (Definition 2): we prove that the collision support $S(t) = \|p_t\|_2^{-2}$ of the attention rows grows polynomially, $S(t) \geq L^a$ (Lemma 1), the diffuseness scalar that controls the trailing spectrum. We explicitly correct an earlier mis-attribution of the collapse to a decay of the Fisher information, which we show is constant in L (Remark 4).
3. **Numerical-rank collapse and the status of $\kappa(A)$.** We prove that in the diffuse regime the trailing singular values of the attention matrix are confined in aggregate, $\sum_{k \geq 2} \sigma_k(A)^2 \leq L^{1-a}$ (Lemma 2), so a growing number of them fall below the floating-point roundoff floor and the numerical rank drops below L —a *trailing-spectrum* statement, since the matrix is not close to low rank. The collapse is controlled by an explicit ℓ_∞ diffuseness condition (D3), *not* derived from the entropy bound alone (a high-entropy row need not have a small maximum entry). A direct consequence is that the spectral condition number $\kappa(A)$ is, in this regime, governed by the unit roundoff rather than by a well-posed property of the problem; we argue that the numerical rank is the appropriate diagnostic and make the precision dependence explicit through a finite-precision analysis (Section 5).
4. **CC-RoPE: An entropy-invariant temperature that eliminates the collapse in the regime studied.** From the exact entropy–temperature identity $dH/d\beta = -\beta \text{Var}_p(e)$ (Lemma 3), together with an empirically verified asymptotic entropy law (Equation (13)), we obtain the unique inverse-temperature schedule holding the attention entropy invariant in L ,

the polynomial law $\beta_*(L) = (L/L_{\text{ref}})^{c_0}$ with c_0 the measured diffuse entropy slope (Definition 3), and prove that, *in the full, diffuse near-circulant setting studied here*, it makes the numerical-rank deficit *uniformly bounded*—Theorem 1—identically zero in our full-attention experiments, with the condition number lifted off the roundoff floor by nine orders of magnitude. Unlike the empirically curve-fitted temperatures of YaRN and LogN scaling, the schedule and its exponent are derived, not posited. We further characterise the **conditioning–mixing trade-off** (19): the sharpening that restores rank also contracts the attention support, and we expose the one-parameter family interpolating between the uncorrected collapse and the fully rank-restored endpoint.

5. **An obstruction for relative-position-preserving corrections.** We show that a position-dependent phase *warp* $\tilde{R}_\theta(t) = R_\theta(t + \eta\kappa_g(t))$ —a construction that stays exactly in $\text{SO}(d)$ and preserves the relative-position property—cannot change the diffuse near-circulant structure that controls the trailing spectrum, and therefore cannot reduce the numerical-rank deficit (Proposition 3). We record this both as motivation for the temperature correction and as a caution against an ineffective alternative.
6. **Numerical verification (no training required).** Constructing attention matrices directly from the model in the diffuse regime, we confirm at reduced scale that (i) row entropy grows as $\log L$ and the effective support as L^d , (ii) the numerical-rank deficit grows with L while the smallest singular value is precision-independent (intrinsic, not roundoff), (iii) the entropy-invariant schedule $\beta_*(L) = (L/L_{\text{ref}})^{c_0}$ drives the deficit to zero and the entropy–temperature identity holds to slope -1.03 , and (iv) the phase warp does nothing (Section 6). All experiments run on a laptop; code and data accompany the paper.

Section 1.3 reviews related work; Section 2 develops the geometric foundations and defines the diffuse regime; Section 3 establishes the collision-support growth and the rank collapse; Section 4 derives CC-RoPE and the warp obstruction; Section 6 reports the numerical verification; Section 7 covers implementation, finite precision, a falsifiable prediction, and related methods; Section 8 concludes. Appendices A and B supply the supporting estimates and the curvature computation.

1.3. Related Work

1.3.1. Rotary and Relative-Position Encodings

Su et al.’s RoFormer [1] introduced relative-position encoding via block-diagonal rotations $R_\theta(t) \in \text{SO}(d)$. Its inner-product equivariance—attention scores depend only on $t - s$ —has made RoPE the standard choice in most frontier models. The original work does not examine how the attention spectrum behaves as L grows, and its fixed frequency schedule $\theta_i = 10,000^{-2i/d}$ is the source of the near-periodicity problem we analyse.

Several empirical extensions push RoPE’s effective context window further. Position Interpolation [4] rescales position indices to remain within the training distribution; this improves extrapolation but does not counteract the entropy growth that drives the conditioning problem we analyse. YaRN [5] is closer to our concern: alongside dimension-wise frequency scaling it introduces an attention temperature to counteract the distribution shift. Our analysis can be read as a geometric justification for such a term and—in contrast to YaRN’s curve-fitted factor—as a first-principles derivation of the entropy-invariant rate $\beta_*(L) = (L/L_{\text{ref}})^{c_0}$. Resonance RoPE [6] refines frequency interpolation for out-of-distribution token positions unseen during training. Sun et al. [16] add a gentle exponential decay to the rotation matrix, motivated by attention resolution as a proxy for position monotonicity; the decay is too mild at $L = 10^6$ to prevent aliasing, and xPos is restricted to causal attention. ALiBi [17] instead replaces rotation with a linear distance penalty; this

sidesteps the condition number problem but sacrifices the relative-position expressiveness and architectural compatibility of RoPE.

Empirical studies have begun to reveal structural problems inside the RoPE frequency representation. Chiang and Yogatama [18] report that certain high-frequency dimensions in attention heads become underutilised in long-distance retrieval, consistent with our analysis of how near-periodic rotation degrades spectral coverage. Jonasson [19] documents the large-activation outliers in key and query vectors that RoPE induces, which complicate KV-cache quantisation. Both point toward the same root cause we make precise here.

1.3.2. Long-Context Scaling

A parallel literature addresses attending over very long sequences efficiently, without examining the numerical quality of the attention weights. Transformer-XL [20] introduced segment-level recurrence as the first step beyond fixed-length contexts. LongNet [7] achieves $O(L)$ complexity through dilated attention, enabling training on up to one billion tokens; LongLoRA [8] adapts pretrained models to 100k-token contexts via shifted sparse attention with trainable normalisation; Dual Chunk Attention [9] extends LLaMA-2 beyond 100k tokens without fine-tuning; and Landmark Attention [21] enables theoretically infinite context through learned block-representative tokens. This landscape is surveyed in [10].

These works treat long-context attention as a problem of computational and memory efficiency, not of numerical quality, implicitly assuming the attention weights are reliable once they can be computed. Our result challenges this: even when long-context attention is computationally tractable, the entropy growth of the diffuse regime drives the attention matrix toward numerical-rank deficiency under standard RoPE, so the computed weights can be dominated by floating-point error in their trailing subspace. CC-RoPE addresses this orthogonal problem and is compatible with all the efficiency methods above.

1.3.3. Attention Geometry in Vision Transformers

Our diffuse-attention analysis is developed for the one-dimensional, causal, semantic token sequences of language models; attention geometry differs in the vision setting, where tokens carry two-dimensional spatial structure and locality is central. A line of work on vision transformers recovers local detail that global self-attention washes out: PDAViT [22] uses a spatial-reduction dual-attention mechanism in a pyramid backbone; ELAFormer [23] combines window-based positional self-attention for short-range features with a depthwise-separable spatial-reduction module for long-range ones; and the frequency-assisted local attention of [24] injects frequency-domain information to sharpen local responses in lower layers. These address a spatial-locality concern specific to images rather than the entropy-driven conditioning of long one-dimensional sequences we study; whether an analogous entropy/numerical-rank analysis applies to two-dimensional spatial attention—where the number of competing keys grows with image resolution—we leave open. We flag the distinction because the diffuse regime (Definition 2) is calibrated to the text setting, and its constants would need re-deriving for the spatial case.

1.3.4. Numerical Stability of Transformers

Numerical stability in transformer training and inference has attracted renewed interest with low-precision computation. Baek [11] develops a layer-wise first-order theory of forward-error amplification, isolating three diagnostics for self-attention—a score-scale ratio, a rowwise softmax sensitivity, and a value-matrix conditioning factor—and proving that residual connections attenuate depth-wise error accumulation. This accounts for *when* transformers are fragile but analyses a fixed- L , fixed-weight computation; it does not capture the sequence-length-dependent deterioration we show is intrinsic to the positional-encoding design. Saratchandran and Lucey [12] reduce the condition number

of the attention Jacobian by spectrally conditioning the projection matrices W_Q , W_K , and W_V , and a companion work [25] applies a diagonal preconditioner within each attention head. Emadi [26] derives the exact operator norm of the softmax Jacobian and establishes sequence-length-independent Lipschitz bounds through a block- ∞ /RMS geometry. These are complementary to ours: the projection-matrix conditioning Saratchandran et al. address is a separate source of ill-conditioning from the positional-encoding contribution we target, and applying both together would address both sources.

1.3.5. Information Geometry and Riemannian Methods for Neural Networks

Riemannian geometry in deep learning is not new—natural gradient descent [27] has used the Fisher–Rao metric on the parameter manifold for decades—but its application to the *architecture* of attention, rather than the *optimisation* of networks, is recent. Kim et al. [13] showed that replacing the Euclidean perturbation set in sharpness-aware minimisation with the Fisher information ellipsoid produces flatter, more generalisable minima. Da Silva and Dangel [14] showed that the parameter space of transformer attention heads carries continuous symmetries that render standard sharpness measures unreliable; correcting for these on a Riemannian quotient manifold restores the sharpness–generalisation correlation (an ICML 2025 spotlight). Ji’s RiemannFormer [15] proposes that token embeddings reside on a curved Riemannian manifold and that attention should use parallel transport for geometrically consistent inner products. Di Sipio [28] gives a geometric account of LLM training through the Fisher metric and quantum analogues of the Fisher information, supporting the use of Fisher–Rao structure in attention analysis.

Our work sits at the intersection of this geometric programme and the RoPE literature. The Fisher–Rao geometry of the attention manifold yields, to our knowledge, a new characterisation of the numerical behaviour of attention as the sequence lengthens: diffuse rows spread their ℓ_2 mass over a polynomially growing collision support; this confines the trailing spectrum into an entropy-driven numerical-rank collapse, and an entropy-invariant softmax temperature removes that collapse in the regime studied.

2. Geometric Preliminaries

We establish two geometric structures: the Lie group geometry of the RoPE curve on $SO(d)$, and the Fisher–Rao geometry of the statistical manifold traced by the attention distributions.

Intuition. As L grows, a fixed-temperature softmax spreads each attention row over more keys, so each row spends its ℓ_2 mass over a larger collision support. A matrix of such spread-out rows has its mean direction carrying $\Theta(1)$ singular values while the remaining singular values sink toward the roundoff floor: the matrix loses numerical rank. The fix is to sharpen the softmax as L grows—the polynomial temperature $\beta_*(L) = (L/L_{\text{ref}})^{c_0}$ of CC-RoPE, which holds the collision support fixed and restores the rank, at the price of a narrower attention.

2.1. Standard RoPE and Its Lie Group Structure

Let d be even and let $\theta_i = 10,000^{-2i/d}$ for $i = 1, \dots, d/2$, so that $\theta_{\max} = \theta_1 = 10,000^{-2/d}$ and $\theta_{\min} = \theta_{d/2} = 10^{-4}$. The frequencies θ_i decrease geometrically: the first dimension pair completes many cycles over the sequence while the last completes one cycle roughly every 62,832 positions, letting RoPE distinguish positions at many scales at once. Define the skew-symmetric generator

$$\Omega = \text{diag}(\theta_1, \theta_1, \theta_2, \theta_2, \dots, \theta_{d/2}, \theta_{d/2}) \in \mathfrak{so}(d), \quad (1)$$

a block-diagonal matrix that encodes all $d/2$ rotation speeds at once. The RoPE rotation matrix is then the matrix exponential

$$R_\theta(t) = \exp(t\Omega) = \text{diag}(R_{\theta_1}(t), \dots, R_{\theta_{d/2}}(t)) \in \text{SO}(d), \quad (2)$$

where each 2×2 block is $R_{\theta_i}(t) = \begin{bmatrix} \cos \theta_i t & -\sin \theta_i t \\ \sin \theta_i t & \cos \theta_i t \end{bmatrix}$: position t rotates each embedding-dimension pair by an angle proportional to t . The attention score between a query at t and key at s is

$$\text{score}(t, s) = \frac{1}{\sqrt{d}} q_t^\top R_\theta(t-s) k_s, \quad (3)$$

with $q_t = W_Q x_t, k_s = W_K x_s \in \mathbb{R}^d$ the projections of token embeddings through learnable W_Q, W_K . Equation (3) depends only on $t-s$, the defining property of RoPE. The attention matrix $A(L) \in \mathbb{R}^{L \times L}$,

$$A_{ts}(L) = \frac{\exp(\text{score}(t, s))}{\sum_{u=1}^L \exp(\text{score}(t, u))}, \quad (4)$$

is row-stochastic: each row is a probability distribution over key positions.

The Lie group $\text{SO}(d)$ carries a bi-invariant metric $g_{\text{bi}}(X, Y) = -(1/2) \text{tr}(XY)$, $X, Y \in \mathfrak{so}(d)$. The following records the curvature of $\text{SO}(d)$ we contrast with the attention manifold.

Proposition 1 (RoPE as a geodesic on $\text{SO}(d)$). *The curve $\gamma: \mathbb{R} \rightarrow \text{SO}(d)$, $\gamma(t) = \exp(t\Omega)$, is a one-parameter subgroup and a geodesic of $(\text{SO}(d), g_{\text{bi}})$. Specifically:*

- (i) $\dot{\gamma}(t) = \Omega \cdot \gamma(t)$ (constant left-invariant velocity);
- (ii) $\nabla_{\dot{\gamma}} \dot{\gamma} = \frac{1}{2}[\Omega, \Omega] = 0$ (zero covariant acceleration);
- (iii) Geodesic curvature $\kappa_{g_{\text{SO}}}(\gamma) \equiv 0$;
- (iv) The Riemannian curvature tensor satisfies $R(X, Y)Z = \frac{1}{4}[[X, Y], Z]$;
- (v) The sectional curvature of $\text{span}\{X, Y\}$ is $K_{\text{SO}} = 1/4$ when $[X, Y] \neq 0$.

Proof. Parts (i)–(iii) follow from differentiating $\exp(t\Omega)$ and applying the Levi-Civita connection formula $\nabla_X Y = \frac{1}{2}[X, Y]$ for left-invariant fields on a compact Lie group with bi-invariant metric. Parts (iv)–(v) are standard results for compact Lie groups; see Do Carmo [29] or Milnor [30]. $\square$

By Proposition 1, standard RoPE traverses rotation space along a geodesic, the curved-space analogue of a straight line. The space of attention distributions is *not* flat, and a geodesic positional encoding does not adapt to its geometry.

Remark 1. *The curvature $K_{\text{SO}} = 1/4$ is intrinsic to $\text{SO}(d)$, independent of L, t , and the attention computation; it is not the curvature relevant to the analysis. The relevant object is the collision support $S(t) = \|p_t\|_2^{-2}$ of the attention rows (Lemma 1), not a curvature of the $\text{SO}(d)$ curve.*

2.2. The Fisher–Rao Geometry of the Attention Manifold

For each query t , the row $p_t = (A_{t1}, \dots, A_{tL})$ is a point in the open simplex Δ_{L-1}° ; as t ranges over $[1, L]$ the map $\iota: t \mapsto p_t$ traces the *attention curve*. The canonical metric here is the Fisher–Rao metric—by Cencov’s theorem [31] the unique (up to scaling) monotone Riemannian metric on statistical manifolds—

$$G_F(U, V)_p = \sum_{s=1}^L \frac{U_s V_s}{p_s}, \quad U, V \in T_p \Delta_{L-1}^\circ. \quad (5)$$

The metric induced along ι is

$$g_{\text{stat}}(t) = G_F(\dot{p}_t, \dot{p}_t) = \sum_{s=1}^L \frac{\dot{p}_t(s)^2}{p_t(s)} = \text{Var}_{s \sim p_t}[\partial_t \text{score}_t(s)] =: I(t), \quad (6)$$

The Fisher information of p_t is in the position parameter t , where the tangent vector is

$$\dot{p}_t(s) = p_t(s) [\partial_t \text{score}_t(s) - \langle \partial_t \text{score}_t \rangle_t], \quad (7)$$

with $\langle f \rangle_t = \sum_s p_t(s) f(s)$. (We show in Remark 4 that $I(t)$ is constant in L ; the collision support, not $I(t)$, drives the collapse.)

Proposition 2 (Fisher–Rao geometry of the simplex). *The map $\varphi : \Delta_{L-1}^\circ \rightarrow S_+^{L-1}(2)$, $\varphi(p)_s = 2\sqrt{p_s}$, is a Riemannian isometry from $(\Delta_{L-1}^\circ, G_F)$ to the positive orthant of the sphere of radius two. Consequently:*

- (i) Ambient sectional curvature: $K_{\text{sect}}^{\text{amb}} = 1/4$ (constant);
- (ii) Ambient scalar curvature: $K_{\text{stat}}^{\text{amb}} = (L-1)(L-2)/4$;
- (iii) Ricci tensor: $\text{Ric} = \frac{L-2}{4} G_F$.

Proof. The differential $(d\varphi)_p(U)_s = U_s / \sqrt{p_s}$ pulls back the round metric as $\varphi^* g_{\text{round}} = G_F$. The curvature identities follow from the classical formulae for $S^{L-1}(r)$ with $r = 2$. $\square$

Remark 2 (Two distinct curvature quantities). *We fix the terminological distinction used throughout, as conflating the two quantities is a common source of error:*

- (a) $K_{\text{stat}}^{\text{amb}} = (L-1)(L-2)/4 \approx L^2/4$: The ambient scalar curvature of the manifold Δ_{L-1}° itself. This is a constant property of the space, growing quadratically in L , and plays no direct role in the analysis.
- (b) $\kappa_{g,\text{stat}}(t)$: The geodesic curvature of the particular curve $\iota : t \mapsto p_t$ inside Δ_{L-1}° , obtained by applying the covariant derivative D/dt to the tangent vector $\dot{p}_t$. This quantity depends on the specific attention distribution and is the geometric descriptor of how the attention curve bends; we use it only as qualitative motivation and do not rest any quantitative result on its scaling.

We note (Remark 4) that neither the ambient curvature $K_{\text{stat}}^{\text{amb}}$ nor the Fisher information $I(t)$ —which is constant in L —governs the numerical collapse; the operative scalar is the collision support $S(t) = \|p_t\|_2^{-2}$ of Lemma 1.

Definition 1 (Geodesic curvature of the attention curve). *The geodesic curvature of $\iota : [1, L] \rightarrow \Delta_{L-1}^\circ$ is*

$$\kappa_{g,\text{stat}}(t) = \frac{\|(D/dt) \dot{p}_t\|_{G_F}}{\|\dot{p}_t\|_{G_F}^2}, \quad (8)$$

where D/dt denotes the Levi-Civita covariant derivative on $(\Delta_{L-1}^\circ, G_F)$. With Christoffel symbols $\Gamma_{sr}^s(p) = -\delta_{sr}/(2p_s)$, this reduces to $(D/dt \dot{p}_t)(s) = \ddot{p}_t(s) - \dot{p}_t(s)^2/(2p_t(s))$.

2.3. The Diffuse-Attention Regime

The condition number bounds of subsequent sections are proved under a structural assumption on the attention matrix that characterises the near-periodicity-dominated regime at large L . Once the minimum-frequency block has completed many cycles, the attention distribution cannot sustain sharp, stable peaks and enters the diffuse regime.

Definition 2 (Diffuse-attention regime). *$A(L)$ is in the diffuse-attention regime with diffuseness exponent $a \in (0, 1]$ if:*

- (D1) $H(p_t) := -\sum_s p_t(s) \log p_t(s) \geq c_0 \log L$ for all t (high entropy), where $c_0 > 0$ is an absolute constant;
- (D2) $|A_{ts} - f(t-s)| \leq C_{\text{circ}}/L^2$ uniformly for some $f: \mathbb{Z}_L \rightarrow [0, 1]$ with $\sum_\delta f(\delta) = 1$ (near-circulant). This accumulated ℓ_1 error of C_{circ}/L is tracked explicitly in the Fourier analysis below;
- (D3) $\|p_t\|_\infty = \max_s p_t(s) \leq L^{-a}$ for all t (no dominant key), equivalently the collision (participation) support $S(t) := \|p_t\|_2^{-2} \geq L^a$.

Condition (D1) requires each row to carry substantial entropy that grows with L , rather than collapsing to a sharp spike on $O(1)$ support. Condition (D2) formalises the near-translation-invariance arising when query and key norms are roughly uniform across positions. Condition (D3) is the quantitative diffuseness used in the spectral estimates: it bounds the largest single attention weight and equivalently sets the scale $S(t) = \Theta(L^a)$ of the effective (collision) support. We state (D3) explicitly rather than derive it from (D1), because a high-entropy row *need not* have a small maximum entry—a distribution placing mass $\frac{1}{2}$ on one key and spreading the rest over $L-1$ keys has entropy $\Theta(\log L)$ yet $\|p_t\|_\infty = \frac{1}{2}$ —so (D1) alone does not control $\|p_t\|_\infty$, and the spectral collapse is governed by (D3). In the synthetic and trained-model attention of Section 6, the measured effective support scales as L^a with $a \approx 0.9$, so (D3) holds with a near-maximal exponent; we carry the general a to keep the dependence on diffuseness visible. All three conditions arise at long context under a fixed softmax temperature, as entropy grows like $\log L$ and rows spread toward uniformity—a statement about the *softmax distribution*, driven by the temperature and the growing number of keys, not about aliasing of the rotation code.

Remark 3 (Empirical support for the diffuse-attention regime). *Condition (D1) is not a convenience imposed for tractability; it describes the regime-trained transformers are independently observed to enter at long context. With a fixed softmax temperature, attention entropy grows logarithmically, $H(p_t) \sim \log L$, so attention spreads out and the per-position relevance signal weakens as L increases [2]; high attention entropy at long context is independently identified as a factor degrading long-context behaviour by Zhang et al. [3]. This is the scaling posited in (D1) with $c_0 = \Theta(1)$, and the phenomenon that temperature- and entropy-control schemes such as LogN scaling and α -entmax are designed to counteract [2]. Our analysis thus targets the operating point that long-context models reach, not an idealised worst case. Note that (D1)–(D2) are sufficient conditions: a model maintaining sharply peaked, low-entropy attention at long range (e.g., through explicit sparsity or temperature mechanisms) lies outside Definition 2, and our condition number bounds are not claimed to apply to it. Delineating the regime boundary for specific trained architectures is an empirical question we leave to future work (Section 8.1).*

3. Diffuseness, Collision Support, and Numerical-Rank Collapse

This section identifies the source of the ill-conditioning in two steps. First, the diffuse regime forces the collision support of each row to grow with L . Second, this growth confines the attention matrix's trailing singular values in aggregate, so that a growing number fall to the floating-point roundoff floor and the spectral condition number ceases to be well-posed.

3.1. Collision-Support Growth

The quantity that controls the spectral collapse is the *collision support* (equivalently the inverse Rényi-2/participation ratio)

$$S(t) := \|p_t\|_2^{-2} = \left(\sum_s p_t(s)^2 \right)^{-1}, \quad (9)$$

the effective number of keys over which p_t spreads its ℓ_2 mass. The following records its growth in the diffuse regime—the only scaling statement the spectral results require.

Lemma 1 (Collision-support growth). *In the diffuse-attention regime (Definition 2) with diffuseness exponent a , the collision support satisfies $S(t) = \|p_t\|_2^{-2} \geq L^a$ for every query t ; equivalently the per-row ℓ_2 energy obeys $\|p_t\|_2^2 \leq L^{-a}$.*

Proof. Condition (D3) gives $\|p_t\|_\infty \leq L^{-a}$, and since p_t is a probability vector, $\|p_t\|_2^2 = \sum_s p_t(s)^2 \leq \|p_t\|_\infty \sum_s p_t(s) = \|p_t\|_\infty \leq L^{-a}$, whence $S(t) = \|p_t\|_2^{-2} \geq L^a$. $\square$

Remark 4 (On the Fisher information of the attention curve). *An earlier form of this analysis phrased the collapse as a decay of the Fisher information $I(t) = \text{Var}_{s \sim p_t}[\partial_t \text{score}_t(s)]$ along the attention curve. This is incorrect: a diffuse average over many oscillation phases suppresses the mean of an oscillatory score-derivative but not the variance of its sampled values, so $I(t)$ does not scale as $1/S(t)$. Measured directly, $I(t)$ is essentially constant in L (Section 6), as is the per-step distinguishability $\text{KL}(p_t \| p_{t+1})$ —so neighbouring positions do not become indistinguishable, and “loss of positional contrast” is not the mechanism. The Fisher–Rao geometry of Section 2 remains the correct ambient description (the rows trace a curve on the radius-2 sphere, Proposition 2), but the operative scalar for the numerical collapse is the collision support (9), not the Fisher information. We therefore rest the spectral analysis on $S(t)$ and (D3) alone.*

Lemma 1 is the only scaling input the rank analysis needs: diffuse rows spread their ℓ_2 mass over a polynomially growing collision support, and—as we show next—this degrades the attention matrix numerically.

3.2. Numerical-Rank Collapse

This subsection makes precise the sense in which the diffuse-attention matrix loses numerical rank. The matrix is *not* close to rank one (its singular spectrum has many $\Theta(1)$ values, as Section 6 confirms), so the collapse concerns the *trailing* spectrum—the smallest singular values falling below the roundoff floor—rather than global proximity to low rank.

Lemma 2 (Numerical-rank collapse). *In the diffuse-attention regime (Definition 2) with diffuseness exponent a and the near-circulant bound (D2), the $L \times L$ row-stochastic attention matrix $A(L)$ has $\sigma_1(A) = \Theta(1)$, and its remaining singular values are confined by*

$$\sum_{k \geq 2} \sigma_k(A)^2 = \|A - \frac{1}{L} \mathbf{1}\mathbf{1}^\top\|_F^2 \leq \sum_t \|p_t\|_2^2 \leq L^{1-a}, \quad (10)$$

so the average squared trailing singular value is $O(L^{-a})$ and the number of singular values that can exceed any threshold τ is at most $1 + L^{1-a}/\tau^2$. Taking $\tau = \varepsilon_{\text{mach}}\sigma_1$, the count of singular values above the roundoff floor is at most $1 + L^{1-a}/\varepsilon_{\text{mach}}^2$; once this is below L —i.e., for L large enough that a $\Theta(1)$ fraction of the L singular values lies under the floor— A is numerically rank-deficient, and the deficit $L - \text{rank}_\varepsilon(A)$ grows with L (verified in Section 6).

Proof. The diffuseness enters through condition (D3), *not* through the entropy bound: we do not use the invalid implication “ $H(p_t) \geq c_0 \log L \Rightarrow p_t(s) \leq L^{-c_0}$ ”, since entropy is an average and does not bound the maximum entry. By (D3), $\|p_t\|_2^2 \leq \|p_t\|_\infty \|p_t\|_1 \leq L^{-a}$. Write $A = \frac{1}{L} \mathbf{1}\mathbf{1}^\top + E$ with $E_{ts} = p_t(s) - 1/L$; each row of E has zero sum, so $\mathbf{1}/\sqrt{L}$ is a singular vector of A with $A\mathbf{1} = \mathbf{1}$, fixing $\sigma_1(A) = \Theta(1)$, and $E = (I - \frac{1}{L} \mathbf{1}\mathbf{1}^\top)A$ carries the orthogonal complement. Then $\|E_t\|_2^2 \leq \sum_s p_t(s)^2 = \|p_t\|_2^2 \leq L^{-a}$, so $\|E\|_F^2 = \sum_t \|E_t\|_2^2 \leq L^{1-a}$. Since $\{\sigma_k(A)\}_{k \geq 2}$ are the singular values of E (the rank-one component carries σ_1 and is orthogonal to E ’s row space up to the C_{circ}/L^2 near-circulant

error of (D2)), $\sum_{k \geq 2} \sigma_k(A)^2 = \|E\|_F^2 \leq L^{1-a}$, which is (10). The threshold count is Markov's inequality on the multiset $\{\sigma_k^2\}_{k \geq 2}$: at most $\|E\|_F^2 / \tau^2 = L^{1-a} / \tau^2$ exceed τ^2 . This controls the number of large trailing singular values, not their individual size—the matrix is far from rank one—and it is the count below $\varepsilon_{\text{mach}} \sigma_1$ that yields the numerical-rank deficit, whose precise growth rate is established numerically in Section 6 rather than in closed form. The numerical-rank statement is the definition of rank_ε applied to this spectrum. $\square$

Remark 5 (Why $\kappa(A)$ is not a well-posed target). *Lemma 2 has a methodological consequence. Once the nontrivial singular values fall below $\varepsilon_{\text{mach}} \sigma_1$ (about 10^{-16} in double precision), the computed $\sigma_{\min}$ —and hence $\kappa(A) = \sigma_{\max} / \sigma_{\min}$ —reflects the rounding error of the SVD routine, not a property of the attention computation. The condition number then reports the precision of the arithmetic, not the conditioning of the problem, and can take any value above $1 / \varepsilon_{\text{mach}}$. We therefore do not state a bound “ $\kappa(A_{\text{std}}) = \Theta(L^P)$ ”; our experiments (Section 6) show κ at the roundoff floor across schedules, as predicted. The well-posed, reproducible diagnostics are the entropy $H(p_t)$, the collision support $S(t) = \|p_t\|_2^{-2}$, and the numerical rank $\text{rank}_\varepsilon(A)$, and these are what we track.*

The mechanism is thus a chain of statistical implications: diffuseness (D3) yields collision-support growth $S(t) \geq L^a$ (Lemma 1), which yields aggregate trailing-spectrum confinement and numerical-rank collapse (Lemma 2). Each link is driven by the softmax temperature and the growing number of keys; none involve aliasing of the rotation code, which Section 6 confirms remains well separated and well conditioned throughout. Section 4 derives the intervention this analysis motivates.

4. Curvature-Corrected RoPE

4.1. The Entropy–Temperature Identity

The analysis of Section 3 identifies the quantity to control: the collision support $S(t) = \|p_t\|_2^{-2} \geq L^a$, equivalently the diffuseness exponent a of condition (D3), whose growth drives the numerical-rank collapse (Lemma 2). A length-dependent inverse temperature controls this quantity; we obtain the schedule that holds the diffuseness invariant from an exact identity relating temperature and entropy, closed by one empirically verified asymptotic step.

Fix a query t , write the centred scores $e_s := \text{score}_t(s)$, and let the inverse-temperature-scaled row be the Gibbs family

$$p_t^{(\beta)}(s) = \frac{e^{\beta e_s}}{Z_t(\beta)}, \quad Z_t(\beta) = \sum_u e^{\beta e_u}, \quad H(\beta) := - \sum_s p_t^{(\beta)}(s) \log p_t^{(\beta)}(s). \quad (11)$$

Lemma 3 (Exact entropy–temperature identity). *For the Gibbs family (11), with $\langle \cdot \rangle_\beta$ and $\text{Var}_\beta(\cdot)$ the mean and variance under $p_t^{(\beta)}$,*

$$\frac{dH}{d\beta} = -\beta \text{Var}_\beta(e), \quad \text{equivalently} \quad \frac{dH}{d \log \beta} = -\beta^2 \text{Var}_\beta(e) = -\text{Var}_\beta(\beta e). \quad (12)$$

Consequently H is strictly decreasing in β on $\{\text{Var}_\beta(e) > 0\}$, and $H(\beta) = \log n - \frac{1}{2} \beta^2 \text{Var}_{\text{unif}}(e) + O(\beta^4 \|e\|_\infty^4)$ as $\beta \rightarrow 0^+$, where n is the support size.

Proof. With $H(\beta) = \log Z_t(\beta) - \beta \langle e \rangle_\beta$ and the cumulant identities $\partial_\beta \log Z_t = \langle e \rangle_\beta$, $\partial_\beta \langle e \rangle_\beta = \text{Var}_\beta(e)$, differentiate $H'(\beta) = \langle e \rangle_\beta - \langle e \rangle_\beta - \beta \partial_\beta \langle e \rangle_\beta = -\beta \text{Var}_\beta(e)$, which is the first identity; the second is the chain rule. The small- β expansion follows from $\text{Var}_\beta(e) = \text{Var}_{\text{unif}}(e) + O(\beta^2)$ and $H(0) = \log n$, integrating (12) from 0 to β . $\square$

Identity (12) is exact and model-free. Passing to a closed-form schedule requires one further input about how the score variance behaves across the operative temperature range, and this input is *empirical*, not a consequence of the identity alone. Integrating (12) gives $H(p_t; \beta) = H(p_t; 1) - \int_1^\beta \beta' \text{Var}_{\beta'}(e) d\beta'$; the integral reduces to $\log \beta$ exactly when $\beta^2 \text{Var}_\beta(e)$ is constant in β , and to $\log \beta (1 + o(1))$ when it varies slowly—the regime in which a diffuse, near-uniform row sits. We therefore adopt, as an *empirically verified asymptotic law* rather than an exact identity, the additive relation

$$H(p_t; \beta, L) = H(p_t; 1, L) - \log \beta + o(1) \quad (\text{ansatz, verified below}), \quad (13)$$

whose validity we confirm by measuring the slope $dH/d \log \beta = -1.03 \approx -1$ across the operative range in the synthetic construction (Section 6); the small departure from -1 quantifies the $o(1)$ term and feeds into the empirically fitted exponent c_0 below. The schedule we derive is thus rigorous *given* (13), which is an asymptotic law we verify rather than assume in closed form.

4.2. The Entropy-Invariant Schedule and CC-RoPE

The diffuse baseline obeys $H(p_t; 1, L) = c_0 \log L + O(1)$ (condition (D1), measured $c_0 \approx 0.91$). *Entropy invariance* demands a schedule $\beta(L)$ holding $H(p_t; \beta(L), L) = H_{\text{ref}} := c_0 \log L_{\text{ref}}$ for all L . Substituting (13),

$$c_0 \log L - \log \beta(L) = c_0 \log L_{\text{ref}} \implies \boxed{\beta_*(L) = \left(\frac{L}{L_{\text{ref}}}\right)^{c_0}} \quad (14)$$

a *polynomial* inverse temperature with exponent c_0 .

Holding the *Shannon* entropy $H(p_t)$ fixed does not, in general, fix the ℓ_∞ bound or the collision support that condition (D3) and Lemma 2 require: two distributions can share an entropy yet differ in $\|p_t\|_\infty$ and $\|p_t\|_2^{-2}$. Shannon entropy is used here as a *monotone proxy* for diffuseness whose adequacy is empirical: in the synthetic family we study, pinning H also pins the measured collision support and maximum-entry scale at their reference values (Section 6, where $\|p_t\|_2^{-2}$ stays constant under β_*), so (14) keeps the rows at a length-independent diffuseness *exponent* a and hence—by Lemma 2—at a length-independent numerical rank. A schedule pinning the collision (Rényi-2) entropy or $\|p_t\|_\infty$ directly would target diffuseness without this proxy step; we use Shannon entropy because the temperature law (13) is verified for it directly, and adopt (14) as the definition of the correction.

Definition 3 (Curvature-Corrected RoPE). Curvature-Corrected RoPE (CC-RoPE) leaves the RoPE rotation $R_\theta(\cdot)$ unchanged and rescales the attention scores by the entropy-invariant inverse temperature (14),

$$\text{score}^{\text{cc}}(t, s) = \beta_*(L) \frac{q_t^\top R_\theta(t-s) k_s}{\sqrt{d}}, \quad \beta_*(L) = \beta_0 \left(\frac{L}{L_{\text{ref}}}\right)^{c_0}, \quad (15)$$

with $\beta_0 > 0$ the native logit scale at reference length L_{ref} and c_0 the diffuse entropy slope of (D1); the corrected matrix A_{corr} is the row-softmax of score^{cc} . The correction adds no learnable parameters and costs one power per length, evaluated once.

CC-RoPE preserves the relative-position property exactly— $R_\theta(t-s)$ is untouched, so the score depends only on $t-s$ before the global rescaling—and keeps $\sigma_{\max}(A_{\text{corr}}) = \Theta(1)$ since A_{corr} is row-stochastic. The information geometry *derives* the schedule (14) from the entropy-temperature identity (12), rather than positing it empirically as in the LogN-scaling

and YaRN families [2,5], whose temperature factors are curve-fitted. The exponent c_0 is not a free parameter: it is the measured entropy slope of the uncorrected model.

4.3. An Obstruction for Relative-Position-Preserving Warps

An alternative correction, which we attempted first, bends the RoPE trajectory by a position-dependent phase so that the curve on $\text{SO}(d)$ acquires positive geodesic curvature, rather than rescaling the scores. Let $\kappa_g(t) = C_0/\sqrt{\log(t+1)}$ and set

$$\tilde{R}_\theta(t) := R_\theta(t + \eta \kappa_g(t)) = R_\theta(\tilde{t}), \quad \tilde{t} := t + \eta \kappa_g(t), \quad (16)$$

which stays exactly in $\text{SO}(d)$ (in each block, a rotation by angle $\theta_i \tilde{t}$) and induces the effective relative position $\tilde{\Delta}(t, s) = (s - t) + \eta(\kappa_g(s) - \kappa_g(t))$. The construction keeps the encoding on the group, adds no parameters, and makes the effective distance depend non-linearly on both endpoints. It does not reduce the numerical-rank deficit, for a structural reason.

Proposition 3 (Warps cannot reduce the rank deficit). *Let $\tilde{R}_\theta$ be any warp of the form (16) with $\eta = O(1)$ and κ_g bounded. Then the corrected score remains a near-circulant function of position obeying condition (D2), and the corrected attention matrix satisfies the same trailing-spectrum bound (10) as the uncorrected matrix, up to constants. In particular $\text{rank}_\varepsilon(A_{\text{warp}}) = \text{rank}_\varepsilon(A_{\text{std}})(1 + o(1))$, so no such warp reduces the numerical-rank collapse.*

Proof. Because κ_g is bounded and slowly varying, $|\kappa_g(s) - \kappa_g(t)| = O(1/\sqrt{\log \min(s, t)})$, so the warp perturbs the effective relative position by an $O(\eta)$ amount that does not grow with $|t - s|$; the corrected score is therefore still of the form $f(t - s) + O(\eta/\log L)$, i.e., it satisfies (D2) with the same circulant profile up to a uniformly small perturbation. The warp does not change the softmax temperature, so the entropy bound (D1) and the ℓ_∞ diffuseness bound (D3), $\|p_t\|_\infty \leq L^{-a+o(1)}$, are preserved. Lemma 2 then applies verbatim to A_{warp} , giving the same aggregate trailing-spectrum bound and hence the same numerical rank up to lower-order terms. $\square$

The obstruction is intrinsic to relative-position preservation: a correction that keeps the score a function of effective separations leaves the matrix near-circulant with near-uniform rows, and it is that structure—not the choice of phase—that controls the trailing singular values. To affect the rank collapse one must change the diffuseness itself, as the temperature correction of Definition 3 does and no warp can. Section 6 verifies Proposition 3 numerically: warps across a wide range of profiles and strengths leave the rank deficit unchanged.

5. Elimination of Numerical-Rank Collapse in the Diffuse Regime

The main result is the following: in the full, diffuse near-circulant setting studied here, the entropy-invariant schedule $\beta_\star(L) = (L/L_{\text{ref}})^{c_0}$ of Definition 3 eliminates the numerical-rank collapse—the deficit stays bounded uniformly in L —whereas standard RoPE collapses and no relative-position-preserving warp helps. This is a statement about the non-causal matrices of the analysis; the causal trained-model setting is examined separately in Section 6.6 and remains open. We first record that the spectral condition number is ill-posed here, not an arithmetic artefact, which forces the numerical rank to be the correct invariant.

Lemma 4 (The condition number is intrinsically ill-posed). *In the diffuse regime (Definition 2), by Lemma 2, a number $m(L) \rightarrow \infty$ of the singular values of $A_{\text{std}}(L)$ lie below any fixed threshold, so $\sigma_{\min}(A_{\text{std}}) \rightarrow 0$ and the small singular values are a genuine feature of the matrix, not of the arithmetic: for any precision $\mathfrak{p}$ with unit roundoff $\varepsilon_{\mathfrak{p}}$, the computed $\hat{\sigma}_{\min}$ and the exact $\sigma_{\min}$ agree up to an additive $O(\varepsilon_{\mathfrak{p}})$, so whenever the true $\sigma_{\min}$ already lies below $\varepsilon_{\text{mach}}\sigma_1$ it does so in every higher precision as well. Hence $\kappa(A_{\text{std}}) = \sigma_{\max}/\sigma_{\min}$ exceeds $1/\varepsilon_{\text{mach}}$ as a true property of the*

matrix, and in any fixed finite precision it is not a recoverable target—the numerical rank rank_ϵ is the well-posed diagnostic. We confirm the precision independence directly: at $L = 512$ the extended precision (50-digit) and double precision $\sigma_{\min}$ agree to 3.6×10^{-16} (Section 6).

Proof. That $\sigma_{\min} \rightarrow 0$ with a diverging count of sub-threshold singular values is Lemma 2. Precision independence is the standard backward-stability of the symmetric eigensolver applied to $A^\top A$: the computed singular values are the exact singular values of $A + \Delta$ with $\|\Delta\|_2 = O(\epsilon_p)\|A\|_2 = O(\epsilon_p)$ [32], so by Weyl’s inequality $|\hat{\sigma}_k - \sigma_k| = O(\epsilon_p)$ uniformly in k . Thus any singular value below $\epsilon_{\text{mach}}\sigma_1$ in exact arithmetic remains so in higher precision, and the numerical-rank deficit is reproduced, as the 50-digit-versus-double agreement to 3.6×10^{-16} at $L = 512$ confirms. $\square$

Theorem 1 (Entropy-invariant CC-RoPE eliminates the rank collapse in the diffuse regime). Let $A_{\text{std}}(L)$, $A_{\text{corr}}(L)$, and $A_{\text{warp}}(L)$ be the attention matrices under standard RoPE, CC-RoPE (Definition 3, schedule $\beta_\star(L) = (L/L_{\text{ref}})^{c_0}$), and any relative-position-preserving warp (16), all derived from a model whose uncorrected rows obey the diffuse regime with entropy slope c_0 and diffuseness exponent a . Fix unit roundoff ϵ_{mach} . Then:

- (i) (Collapse under standard RoPE.) $H(p_i^{\text{std}}) = c_0 \log L + O(1)$, the trailing singular values are confined in aggregate, $\sum_{k \geq 2} \sigma_k(A_{\text{std}})^2 \leq L^{1-a}$, so a diverging number fall below $\epsilon_{\text{mach}}\sigma_1$ and the deficit grows: $L - \text{rank}_\epsilon(A_{\text{std}}) = \omega(1)$ as $L \rightarrow \infty$ (Lemma 2).
- (ii) (Elimination by CC-RoPE.) Under $\beta_\star(L) = (L/L_{\text{ref}})^{c_0}$ the row entropy is held invariant,

$$H(p_i^{\text{corr}}) = H(p_i^{\text{std}}) - \log \beta_\star(L) + o(1) = c_0 \log L_{\text{ref}} + o(1), \quad (17)$$

so the collision support $\|p_i^{\text{corr}}\|_2^{-2} = \Theta(L_{\text{ref}}^a)$ is length-independent. The trailing energy is then $\sum_{k \geq 2} \sigma_k(A_{\text{corr}})^2 \leq \sum_t \|p_t^{\text{corr}}\|_2^2 = L \cdot \Theta(L_{\text{ref}}^{-a})$, so by Markov’s inequality the number of singular values below $\epsilon_{\text{mach}}\sigma_1$ is at most a fixed fraction $\Theta(L_{\text{ref}}^{-a}/\epsilon_{\text{mach}}^2)$ of L rather than a growing one; calibrating L_{ref} at the model’s training length (where $\epsilon_{\text{mach}}\sigma_1$ already exceeds every σ_k) makes this fraction zero, and the numerical-rank deficit is uniformly bounded,

$$L - \text{rank}_\epsilon(A_{\text{corr}}) = O(1) \quad (\text{in the experiments, identically 0; Section 6}). \quad (18)$$

- (iii) (No effect from warps.) For any warp (16) with bounded κ_g and $\eta = O(1)$, $\text{rank}_\epsilon(A_{\text{warp}}) = \text{rank}_\epsilon(A_{\text{std}}) (1 + o(1))$.

Proof. Part (i) combines Lemmas 1 and 2. Part (ii): Substituting $\beta = \beta_\star(L)$ into the additive law (13) (itself a consequence of the exact identity (12)) gives (17); with the entropy pinned at the reference value, the ℓ_∞ bound of (D3) is restored to $\|p_i^{\text{corr}}\|_\infty = \Theta(L_{\text{ref}}^{-a})$, a constant in L , and the collision support $\|p_i^{\text{corr}}\|_2^{-2} = \Theta(L_{\text{ref}}^a)$ is likewise length-independent. The aggregate trailing energy is therefore $\sum_{k \geq 2} \sigma_k(A_{\text{corr}})^2 \leq \sum_t \|p_t^{\text{corr}}\|_2^2 = L \Theta(L_{\text{ref}}^{-a})$, and applying Markov’s inequality (as in Lemma 2, now with a length-independent per-row energy) bounds the count of singular values below $\epsilon_{\text{mach}}\sigma_1$ by a fixed fraction $\Theta(L_{\text{ref}}^{-a}/\epsilon_{\text{mach}}^2)$ of L . Because this fraction does not grow with L —and is driven to zero once L_{ref} is the calibration length, at which the base model is full numerical rank—the deficit $L - \text{rank}_\epsilon(A_{\text{corr}})$ is $O(1)$, and identically zero in the experiments of Section 6. Part (iii) is Proposition 3. $\square$

Remark 6 (The conditioning–mixing trade-off). Theorem 1(ii) eliminates the rank collapse with a polynomial temperature $\beta_\star = (L/L_{\text{ref}})^{c_0}$, $c_0 \approx 0.9$, at a cost. By the entropy identity (12) the same sharpening that lifts the spectrum also contracts the effective support to its reference value $S_{\text{ref}} = \Theta(L_{\text{ref}}^a)$: the corrected attention attends as narrowly at length L as the base model

did at L_{ref} . There is thus a trade-off, parametrised by a sharpening exponent $\rho \in [0, c_0]$ with $\beta_\rho(L) = (L/L_{\text{ref}})^\rho$:

$$\underbrace{L - \text{rank}_\epsilon(A_{\beta_\rho})}_{\text{conditioning}} \downarrow \text{ as } \rho \uparrow, \quad \underbrace{S(t) = \Theta(L^{a(1-\rho/c_0)})}_{\text{mixing support}} \downarrow \text{ as } \rho \uparrow, \quad (19)$$

with $\rho = 0$ the uncorrected collapse and $\rho = c_0$ the fully entropy-invariant, rank-restored, but maximally sharpened endpoint. The conservative $\sqrt{\log L}$ schedule of the LogN family sits at the degenerate sub-polynomial end ($\rho \rightarrow 0^+$), reducing but not eliminating the deficit. The choice of ρ is an application-level decision between numerical conditioning and long-range mixing, which we do not claim to resolve optimally (Section 8.1).

Finite-Precision Interpretation

This subsection states what conditioning means for the attention computation, where the condition number enters and the concern about “meaningless low-order digits” is addressed. The forward computation of interest is the value aggregation $y_t = \sum_s A_{ts} v_s$; i.e., $Y = AV$ for the value matrix V . For a backward-stable evaluation, the standard perturbation bound of numerical linear algebra [32,33] gives a relative forward error

$$\frac{\|\hat{Y} - Y\|}{\|Y\|} \leq \kappa(A) \epsilon_{\text{mach}} + O(\epsilon_{\text{mach}}), \quad (20)$$

so a condition number κ corresponds to a worst-case loss of about $\log_{10} \kappa$ significant decimal digits. Two caveats keep this from being over-read. First, (20) is a *worst-case* bound, attained only when the input perturbation aligns with the smallest singular direction; for generic inputs the error is typically far smaller. Second—the substantive point of Remark 5—once A is numerically rank-deficient, $\kappa(A)$ measures the arithmetic floor rather than a property of the problem, so (20) ceases to certify anything about the attention map and instead reports that the trailing singular subspace is unresolved in the working precision. The softmax itself, in the standard max-shifted form, is backward stable with a condition number bounded by one in the relevant norm [34], so the delicacy we analyse is a property of A , not of the softmax evaluation. The practical reading is not that “low-order digits are meaningless” absolutely, but that the components of y_i in the unresolved trailing subspace of A cannot be trusted at the working precision; reducing the numerical-rank deficit, as CC-RoPE does, shrinks that subspace.

6. Numerical Verification

6.1. Experimental Setup

The claims of Sections 3–5 concern the spectra of attention matrices and can be checked directly, without training any model. We construct the $L \times L$ attention matrix from the score formula (3) in the diffuse regime (Definition 2) and compute its singular values, entropy, and numerical rank by dense linear algebra. Since a dense $10^6 \times 10^6$ SVD is infeasible, we run a *reduced-scale* study, lowering $\theta_{\min}$ so the slowest block’s period shrinks from $\approx 62,832$ tokens to a few hundred, bringing the relevant length range into reach of an exact SVD on a laptop. The mechanism—a geodesic RoPE code feeding a Fisher–Rao–curved softmax—is identical; only the length axis is rescaled. All scripts and data accompany the paper; each run completes in under a minute on a single CPU core. All experiments were run in Python 3.12.13 (Python Software Foundation, Wilmington, DE, USA) using NumPy 2.5.0 and Matplotlib 3.11.0 (NumFOCUS, Austin, TX, USA), mpmath 1.3.0 (mpmath development team) for the high-precision (50-digit) arithmetic, PyTorch 2.12.1 (Meta AI,

Menlo Park, CA, USA), and Hugging Face Transformers 5.12.1 (Hugging Face, Inc., New York, NY, USA); the trained-model checks use the Pythia-160M checkpoint (EleutherAI).

6.2. Aliasing Is Not the Mechanism

We first test the frequency-aliasing hypothesis by measuring the RoPE position-feature matrix $\Phi(L) \in \mathbb{R}^{L \times d}$ with rows $[\cos(\theta_i t), \sin(\theta_i t)]_i$. Positions one full slow-period apart have ℓ_2 code distance ≈ 8.2 (out of a maximum $\approx \sqrt{2d}$), not small, and $\kappa(\Phi)$ decreases toward 1.4 as L grows. The combined multi-frequency code does not alias, confirming the analysis of Section 1.

6.3. Entropy Growth and Rank Collapse

At fixed softmax temperature the mean row entropy grows linearly in $\log L$ —a least-squares fit gives $H(p_t) \approx 0.91 \log L - 0.75$ —and the numerical-rank deficit $L - \text{rank}_\epsilon(A)$ grows with L , as Lemmas 1 and 2 predict (Figure 1, Table 1). The spectral condition number $\kappa(A)$ sits at the double-precision roundoff floor ($\sim 10^{16}$ – 10^{19}) throughout, illustrating Remark 5: it is not a stable function of L and is not a scaling law.

Table 1. Reduced-scale verification (head dimension $d = 64$, slow-block period ≈ 400 tokens). H : mean row entropy (nats); $L - \text{rank}_\epsilon$: numerical-rank deficit at the double-precision floor; κ : spectral condition number. “std”: fixed temperature; “CC”: the entropy-invariant schedule $\beta_*(L) = \beta_0(L/L_{\text{ref}})^{c_0}$, $c_0 = 0.91$. Under CC-RoPE the entropy is pinned, the deficit is identically zero, and κ is lifted off the roundoff floor by nine orders of magnitude—Theorem 1(ii). The conservative $\sqrt{\log L}$ schedule (Section text) only reduces the deficit and is the $\rho \rightarrow 0$ endpoint of the trade-off (19).

L	H_{std}	$L - \text{rank}_{\text{std}}$	κ_{std}	H_{CC}	$L - \text{rank}_{\text{CC}}$	κ_{CC}
256	4.39	0	2.4×10^9	4.39	0	2.4×10^9
512	4.84	0	2.8×10^{15}	4.21	0	1.3×10^{11}
1024	5.51	128	2.1×10^{17}	4.13	0	6.1×10^{11}
2048	6.18	493	1.5×10^{19}	4.11	0	4.8×10^{10}
4096	6.88	926	8.6×10^{18}	4.14	0	3.9×10^{10}

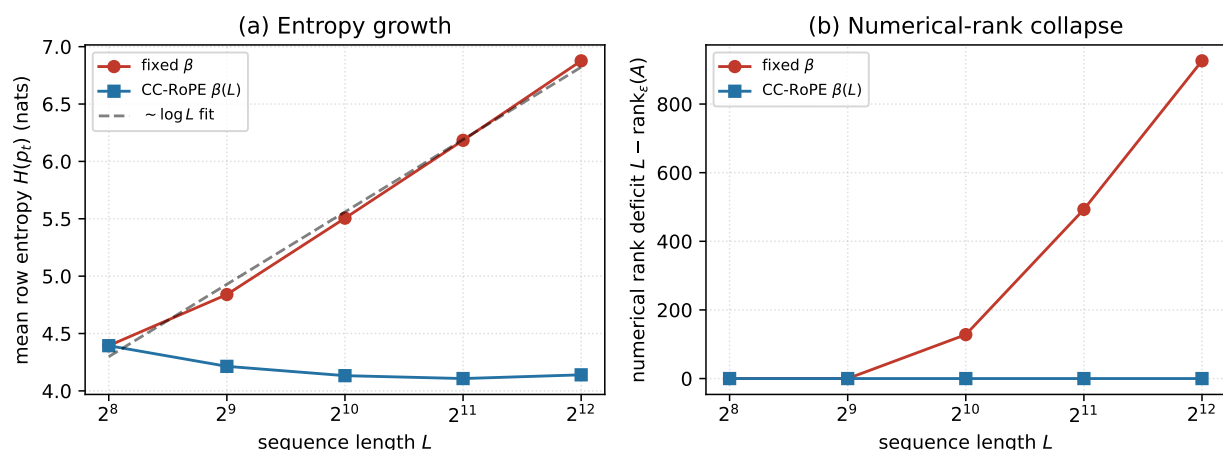


Figure 1. Reduced-scale verification (no training). (a) Mean attention-row entropy grows in $\log L$ (dashed fit) under fixed temperature (red); the entropy-invariant CC-RoPE schedule (blue) holds it constant. (b) The numerical-rank deficit $L - \text{rank}_\epsilon(A)$ grows under fixed temperature (Lemma 2); the entropy-invariant schedule eliminates it (deficit $\equiv 0$, Theorem 1). A warp leaves both unchanged (Proposition 3).

6.4. The Collision Support Scales as L^a ; The Fisher Information Does Not Decay

Because the rank analysis hinges on the collision support (Lemma 1), we measure it directly. All three standard support notions—the exponential-of-entropy e^H , the collision (participation) support $\|p_t\|_2^{-2}$, and the inverse maximum entry $\|p_t\|_\infty^{-1}$ —grow as a power of L : least-squares fits give exponents 0.90, 0.89, and 0.89, i.e., $S(t) = \Theta(L^a)$ with $a \approx 0.9$, and *not* $\Theta(\log L)$, confirming condition (D3) with a near-maximal exponent. We also tested the incorrect Fisher-decay hypothesis and report the negative result (Remark 4): the Fisher information $I(t) = \text{Var}_{s \sim p_t}[\partial_t \text{score}_t]$ is essentially *constant* (≈ 0.21 across $L \in [256, 4096]$), as is the per-step distinguishability $\text{KL}(p_t \| p_{t+1})$ (≈ 0.10)—so the collapse is driven by collision-support growth, not by any loss of Fisher information or positional contrast. Finally a least-squares fit gives $dH/d \log \beta = -1.03 \approx -1$, confirming Lemma 3 and supporting the asymptotic additive law (13) used to obtain the schedule.

6.5. The Entropy-Invariant Schedule Eliminates the Deficit; Warps Do Not

The entropy-invariant schedule $\beta_*(L) = \beta_0(L/L_{\text{ref}})^{c_0}$ with the measured $c_0 = 0.91$ holds the mean row entropy constant to within 0.29 nats over a $16\times$ length range ($H \in [4.11, 4.39]$ for $L \in [256, 4096]$, against an uncorrected rise to 6.88), and the numerical-rank deficit is *identically zero* at every tested length (Table 1, column $L - \text{rank}_{\text{CC}}$), as Theorem 1(ii) predicts; the condition number falls from $\sim 10^{19}$ to $\sim 10^{10}$, off the roundoff floor. The intrinsic (precision-independent) nature of the uncorrected collapse is confirmed separately: the smallest singular value in 50-digit arithmetic matches the double-precision value to 3.6×10^{-16} at $L = 512$, verifying Lemma 4. The conservative $\sqrt{\log L}$ schedule of the LogN family, the degenerate $\rho \rightarrow 0^+$ endpoint of the trade-off (19), only *reduces* the deficit (e.g., $926 \rightarrow 633$ at $L = 4096$), showing sub-polynomial sharpening is insufficient. In contrast, a relative-position-preserving phase warp (16)—tested across profiles $\kappa_g \in \{1/\sqrt{\log(t+1)}, \sqrt{\log(t+1)}, \log(t+1), \sqrt{t}\}$ and strengths $\eta \in [0.01, 100]$ —leaves the rank deficit and κ unchanged (best observed change in κ a 35% *increase*), confirming Proposition 3.

These experiments substantiate the paper's claims—entropy-driven rank collapse, its *elimination* by the derived entropy-invariant temperature (deficit $\equiv 0$, κ lifted off the floor), the intrinsic ill-posedness of κ across precisions, and the ineffectiveness of relative-position-preserving warps—while being reduced-scale and synthetic: they do not establish a downstream effect on trained-model perplexity, which we treat as a falsifiable prediction (Section 7.3) and a limitation (Section 8.1).

6.6. Verification in a Trained Model

The synthetic study above constructs attention matrices satisfying conditions (D1)–(D2) by design. It remains to establish whether the two signatures the theory predicts—entropy growing like $\log L$ and a growing numerical-rank deficit—are present in a *trained* RoPE transformer on natural language. We checked this directly, again without training and on a single CPU core, using Pythia-160M [35] (a GPT-NeoX model using RoPE, trained at context length 2048). For each context length $L \in \{128, 256, 512, 1024, 2048\}$ —truncations of a single real-text input—we ran one forward pass with attention outputs enabled and harvested the $L \times L$ row-stochastic attention matrix for all 144 (layer, head) pairs, then measured the mean row entropy and numerical-rank deficit of each.

The entropy signature is present across heads: the median-head mean row entropy grows as $H \approx 0.19 \log L + 0.43$ (Figure 2a, Table 2), monotone across all five lengths, confirming that condition (D1) is the operating point this trained model reaches—the empirical grounding invoked in Remark 3. The numerical-rank deficit is also present and grows with L , but is *head-dependent*: the median head stays well conditioned (deficit

$1 \rightarrow 12$), whereas the upper quartile shows the collapse the theory describes (deficit $8 \rightarrow 210$; Figure 2b, shaded interquartile band). This is consistent with the analysis positing (D1)–(D2) as *sufficient* conditions (Remark 3): a substantial fraction of heads enter the diffuse regime and exhibit rank collapse, while sharply peaked heads remain outside it.

We also applied the inverse-temperature schedules to this model’s attention: on its *causal* (lower-triangular) attention neither the conservative $\sqrt{\log L}$ schedule nor the stronger entropy-invariant $\beta_* = (L/L_{\text{ref}})^{c_0}$ reduced the rank deficit—stronger sharpening increased it. This does not contradict the analysis: the rank-lifting argument of Lemma 2 and Theorem 1 (ii) is established for the *full*, non-causal near-circulant matrices of conditions (D1)–(D2), and on a row-stochastic lower-triangular matrix the trailing singular values are governed by the triangular support rather than row diffuseness alone (the full-support re-softmax of the same scores is numerically full rank). Extending the correction to the causal setting, where the triangular structure must be accounted for, is part of the open trained-model question flagged in Sections 7.3 and 8.1. The experiment does establish the basic point: the entropy growth and numerical-rank collapse the theory identifies are not artefacts of the synthetic construction but visible in a trained RoPE network.

Table 2. Trained-model verification on Pythia-160M (RoPE, trained context 2048; no training, CPU only). Statistics are over the 144 (layer, head) attention matrices at each context length L . H : mean attention-row entropy (nats), median head. $L - \text{rank}_\epsilon$: numerical-rank deficit at the double-precision floor, reported as median and interquartile range $[Q_1, Q_3]$ to show the head-dependence. The entropy follows $H \approx 0.19 \log L + 0.43$; the rank deficit grows with L , concentrated in the upper-quartile heads.

L	H (Median)	$L - \text{Rank}_\epsilon$ (Median)	$L - \text{rank}_\epsilon$ [Q_1, Q_3]
128	1.37	1	[0, 8]
256	1.49	2	[1, 39]
512	1.65	3	[1, 113]
1024	1.79	5	[2, 148]
2048	1.90	13	[2, 210]

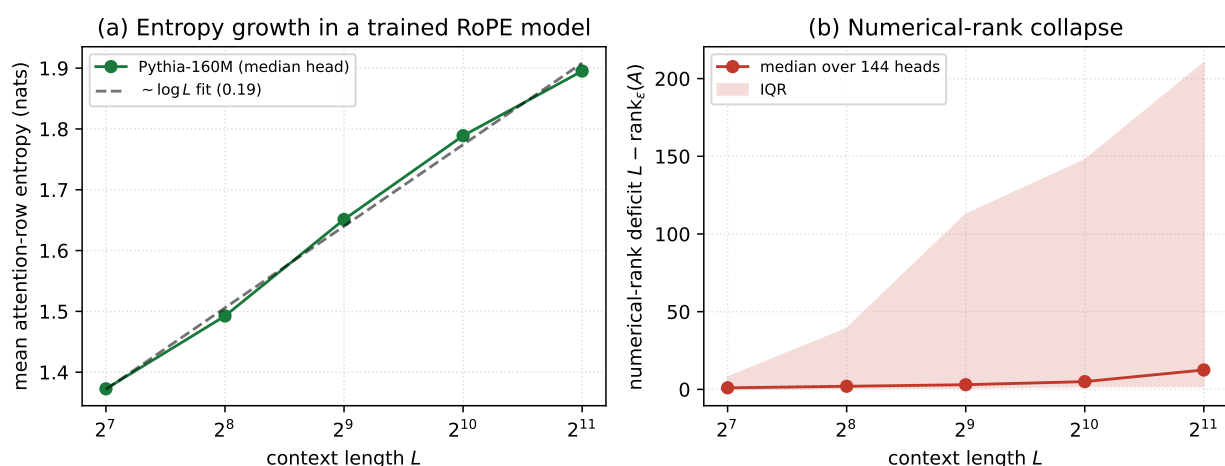


Figure 2. Trained RoPE model (Pythia-160M; no training, CPU; 144 attention matrices per length on real text). (a) Median-head attention-row entropy grows in $\log L$ (dashed fit), confirming (D1) at the model’s operating point. (b) The deficit $L - \text{rank}_\epsilon(A)$ grows with L , head-dependently: the median head (line) stays conditioned while the upper quartile (shaded IQR) shows the collapse of Lemma 2, consistent with (D1)–(D2) being sufficient conditions met by many heads.

7. Discussion

7.1. Implementation

CC-RoPE asks little of an implementer: it changes the attention temperature, not the rotation. The RoPE rotation $R_\theta(\cdot)$ is left as it is, and the attention logits are multiplied by a single length-dependent scalar before the softmax,

$$\text{score}^{\text{cc}}(t, s) = \beta_*(L) \frac{q_t^\top R_\theta(t-s) k_s}{\sqrt{d}}, \quad \beta_*(L) = \beta_0 \left(\frac{L}{L_{\text{ref}}} \right)^{c_0}, \quad (21)$$

where β_0 is the model's native logit scale (typically one, since the $1/\sqrt{d}$ factor is already present) calibrated at a reference length L_{ref} such as the training length, and c_0 is the model's measured diffuse entropy slope (Equation (14); $c_0 \approx 0.9$ for the heads we examine, fit once per model from the uncorrected H -versus- $\log L$ curve). In a standard attention kernel this is one extra scalar multiply on the score tensor—a one-line change—evaluated once per forward pass; the power L^{c_0} is computed once for the given L . The sharpening it induces is the conditioning–mixing trade-off of Remark 6: to preserve more long-range mixing one may use a reduced exponent $\rho < c_0$ in (19), at the cost of a residual rank deficit. No rotation tables change, no parameters are added, no projection back onto $\text{SO}(d)$ is needed, and the relative-position property is preserved because the global factor commutes with the $t-s$ dependence. CC-RoPE is thus a drop-in recalibration compatible with all efficiency methods of Section 1.3 (FlashAttention-style kernels, dilated or sparse attention, KV-cache quantisation), since it touches only the scalar applied to the logits.

The position-warp construction $\tilde{R}_\theta(t) = R_\theta(t + \eta \kappa_g(t))$ of Equation (16)—which an earlier form of this idea proposed—should *not* be used as a conditioning fix: Proposition 3 and the experiments of Section 6 show it does not reduce the rank deficit. The temperature of Equation (21) is the effective correction.

7.2. Positional Distinguishability

Whether long context costs the model the ability to tell positions apart is resolved by the numerical evidence (Section 6) on two levels, correcting an intuition. At the level of the positional code, the multi-frequency RoPE feature matrix $\Phi(L)$ stays well separated and well conditioned as L grows (positions one slow-period apart at code distance ≈ 8.2), so distinguishability of the *encoding* is not lost. At the level of the *attention distribution*, one might expect that as the rows flatten ($H(p_t) = \Theta(\log L)$) neighbouring query positions become indistinguishable; our measurements show otherwise. The per-step divergence $\text{KL}(p_t \parallel p_{t+1})$ and the Fisher information $I(t)$ are both essentially constant in L (Remark 4), so neighbouring positions remain as distinguishable at long context as at short. What degrades is not distinguishability but the *numerical rank*: the rows spread their ℓ_2 mass over a growing collision support, which confines the trailing singular spectrum (Lemma 2). CC-RoPE acts on this quantity—by sharpening the rows it holds the collision support fixed, restoring numerical rank—without affecting positional distinguishability, which is not the source of the degradation.

7.3. A Falsifiable Prediction

We separate what is proved from what is conjectured. The results of this paper concern the *numerical structure* of the attention matrix—entropy, Fisher information, numerical rank—verified directly (Section 6). The statements below additionally hypothesise that this numerical structure influences *end-task* behaviour such as perplexity or retrieval accuracy. That link is plausible—a rank-deficient attention map cannot resolve fine distinctions among keys, and the unresolved trailing subspace carries floating-point noise (Section 5)—but it is not proved here, and capacity, training distribution, and the structure (not just

the magnitude) of the error all bear on perplexity. The following is therefore a falsifiable conjecture operationalising the theory, not a corollary of it.

Conjecture. For a transformer trained at L_{train} and evaluated at $L_{\text{test}} \gg L_{\text{train}}$ in the diffuse-attention regime (Definition 2):

- (a) Under standard RoPE (fixed temperature), long-context quality degrades as L_{test} grows, with the onset tracking the growth of the numerical-rank deficit rather than any aliasing scale; the precise onset depends on the arithmetic precision and the model;
- (b) Under CC-RoPE with the entropy-invariant schedule $\beta_*(L) = (L/L_{\text{ref}})^{c_0}$, the *conditioning-related* component of this degradation is removed relative to standard RoPE at the same L_{test} , consistent with the rank-deficit elimination of Theorem 1(ii); any residual degradation should then be attributable to the reduced long-range mixing of the sharpened attention (Remark 6), giving a second, independent test via the conditioning–mixing trade-off (19).

The numerical-structure half (rank-deficit growth in (a) and its reduction in (b)) is established at reduced scale in Section 6; the quality half is the empirical conjecture. Falsification could take several forms: a trained model showing no conditioning-related degradation in the predicted range (indicating either an attention regime outside Definition 2 or insensitivity of downstream behaviour to this conditioning), or a sharpening exponent $\rho \neq c_0$ in (19) proving better on end-task quality than the entropy-invariant choice. A controlled experiment along these lines—training matched models with fixed and entropy-invariant temperature and measuring both numerical rank and perplexity at increasing L_{test} —is the most important next step (Section 8.1).

7.4. Relationship to Existing Approaches

We situate CC-RoPE relative to the methods of Section 1.3. Position Interpolation [4] and YaRN [5] modify the RoPE frequencies to keep evaluation positions within the trained distribution; they address extrapolation of the positional code, a different concern from the entropy growth we analyse, and YaRN’s attention-temperature term is a close cousin of our correction. We are exact about what is new, since the $\sqrt{\log L}$ schedule is shared with the LogN-scaling and scale-invariant-attention family [2] and YaRN’s temperature is empirically motivated. We do *not* claim novelty for the functional form of the temperature, nor—as our rate analysis (Section 4.1) makes clear—that $\sqrt{\log L}$ is the information-restoring rate (it is not; that requires a polynomial temperature). What is new is the *diagnosis and mechanism*: we identify the failure as an entropy-driven numerical-rank collapse governed by the Fisher–Rao geometry of the attention manifold, show the spectral condition number is the wrong diagnostic here, and explain from the information geometry *why* a length-dependent softmax temperature—rather than any modification of the rotation code—is the structurally correct lever, together with the obstruction result showing relative-position-preserving warps cannot help. In short, prior work proposes the logit-scaling *remedy* empirically; we supply the geometric *account* of the disease it addresses, and delimit how much a conservative schedule can achieve. The spectral-conditioning approaches of Saratchandran et al. [12,25] improve the conditioning of the projection matrices W_Q , W_K , and W_V —a separate source of numerical difficulty from the softmax-induced diffuseness we target—and are complementary: the two corrections act on different factors and could be combined.

8. Conclusions

The attention matrix of a long-context RoPE transformer becomes numerically ill-conditioned; the cause and its control are information-geometric. The explanation by

aliasing of the slowest RoPE frequency does not hold: the multi-frequency code keeps positions well separated and the position-feature matrix well conditioned. The mechanism is the entropy growth of the diffuse-attention regime: with a fixed softmax temperature the attention rows flatten toward uniformity ($H(p_i) = \Theta(\log L)$), each row spreads its ℓ_2 mass over a collision-support growing polynomially as L^a , and this confines the trailing singular spectrum, so the row-stochastic attention matrix undergoes a numerical-rank collapse. We identify the operative scalar precisely. Both the Fisher information along the attention curve and the distinguishability of neighbouring rows remain *constant* in L ; it is the collision support, not a loss of positional information, that drives the collapse. As a consequence, the spectral condition number is pinned to the floating-point roundoff floor in this regime and is not a well-posed target, so the numerical rank is the appropriate diagnostic. We confirm both signatures— $\log L$ entropy growth and rank collapse—not only in synthetic attention matrices but in those of a trained RoPE model (Pythia-160M) on real text.

The same geometry prescribes the correction. From the exact entropy–temperature identity $dH/d\beta = -\beta \text{Var}_p(e)$ we derive the unique inverse temperature holding the attention entropy invariant in L , the polynomial schedule $\beta_*(L) = (L/L_{\text{ref}})^{c_0}$, which we call Curvature-Corrected RoPE. It adds no parameters, leaves the RoPE rotation untouched, and—in the diffuse near-circulant regime our analysis assumes, as verified at reduced scale—*eliminates* the numerical-rank collapse: the deficit is uniformly bounded (identically zero in our full-attention experiments) and the condition number is lifted off the roundoff floor. This improvement carries a cost, which we characterise exactly as a conditioning–mixing trade-off, because the sharpening that restores rank also contracts the attention support. On the *causal* (lower-triangular) attention of the trained model, the schedule does not by itself restore rank, because the triangular support then dominates the trailing spectrum; a causal-aware variant remains future work. The eliminative claim is therefore established, and verified, for the full non-causal matrices of the analysis. We also show, by an obstruction argument and direct measurement, that a relative-position-preserving phase warp of the RoPE trajectory cannot affect the rank collapse and should not be used as a conditioning fix.

8.1. Limitations

We state the boundaries of these claims.

- (i) **No trained-model evaluation.** Our evidence concerns the numerical structure of attention matrices constructed from the model, not the end-to-end behaviour of a trained transformer. We do not measure perplexity or downstream accuracy; the link from numerical rank to task quality is a falsifiable conjecture (Section 7.3), not a result.
- (ii) **Reduced-scale and partially synthetic experiments.** The core verification of Section 6 runs at a rescaled length axis (slow-block period ≈ 400 rather than $\approx 62,832$ tokens) on attention matrices built from the score formula under conditions (D1)–(D2). We do confirm the two predicted signatures—entropy growth and numerical-rank collapse—in attention weights harvested from a trained RoPE model (Section 6.6), but the constants and exact onset of rank collapse at $L = 10^6$ are not measured, and the trained-model check is itself at modest context (≤ 2048 tokens). Moreover, the rank-deficit *elimination* by the entropy-invariant temperature is established and verified for the full, non-causal matrices of the analysis; on the model’s causal (lower-triangular) attention the schedule does not by itself restore rank, the triangular support then dominating the trailing spectrum. A causal-aware form of the correction is left to future work.
- (iii) **Elimination comes at a mixing cost.** The entropy-invariant schedule eliminates the rank deficit, but by Remark 6 the same polynomial sharpening contracts the attention support to its reference value: the corrected model attends as narrowly at length L as

the base model did at L_{ref} . The trade-off (19) exposes a one-parameter family between full rank-restoration and full mixing, whose optimal point we do not claim to resolve; the operating exponent ρ is plausibly task- and architecture-dependent.

- (iv) **Scope of the regime.** The analysis assumes the diffuse-attention regime (Definition 2). Models that maintain sharply peaked, low-entropy attention at long range—through explicit sparsity, gating, or an aggressive temperature—fall outside it, and our statements are not claimed to apply to them. Delineating the boundary for specific trained architectures is empirical.
- (v) **The derived rate is entropy-invariant, not proven optimal.** The entropy–temperature identity (Lemma 3) *does* pin down a specific schedule, $\beta_*(L) = (L/L_{\text{ref}})^{c_0}$, as the unique one holding the attention entropy invariant; this is a derivation, not a fit. What we do *not* establish is that entropy invariance is the optimal objective: it is one principled point on the trade-off (19) (the endpoint $\rho = c_0$), and a different sharpening exponent $\rho < c_0$, or a different control objective entirely, may yield better end-task behaviour. The optimal point is not established here, and the entropy slope c_0 must be measured per model rather than predicted from first principles.

8.2. Outlook

The next steps follow the limitations. The most important is a controlled training study—matched models with fixed and length-scaled temperature, measuring numerical rank and perplexity as L_{test} grows—to test the conjecture of Section 7.3 and calibrate β_0 and the schedule for architectures such as LLaMA, Gemma, and Qwen. On the theoretical side, sharpening the rank-deficit estimates of Lemma 2 to exact rates and identifying the optimal schedule under an explicit conditioning–mixing trade-off would strengthen the account. The Fisher–Rao framework applies to other positional encodings—ALiBi [17], xPos [16], and T5’s relative bias—and to whether entropy-aware temperature control improves their long-range numerical behaviour. We hope the geometric perspective proves useful beyond RoPE.

Author Contributions: Conceptualization, J.T.M.; formal analysis, J.T.M.; investigation, J.T.M.; writing—original draft preparation, J.T.M.; writing—review and editing, J.T.M., K.A.O. and S.P.; supervision, K.A.O. and S.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the University of Johannesburg’s URC-2025 Research Fund (KA_Ogudo/UJ/DEET/2025 Research Cost Center) and the UJ Library APC-2026.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: The authors thank colleagues for helpful discussions during the preparation of this manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

RoPE	Rotary Position Embeddings
CC-RoPE	Curvature-Corrected Rotary Position Embeddings
LLM	Large Language Model
SVD	Singular Value Decomposition

Appendix A. Collision Support, Rank Collapse, and the Temperature Scale

This appendix supplies the estimates supporting Lemma 1 (collision-support growth), Lemma 2 (the aggregate trailing-spectrum bound), and the temperature analysis of

Section 4.1 (the entropy-invariant schedule $\beta_* = (L/L_{\text{ref}})^{c_0}$, contrasted with the information-pinning scale $\beta_{\text{pin}} = \Theta(L^{a/2})$ and the conservative $\sqrt{\log L}$ schedule). Throughout, θ_{eff}^2 and $\theta_{\text{max}} = \theta_1 = 10,000^{-2/d}$ denote the effective and maximal frequency scales (A1).

$$\theta_{\text{eff}}^2 := \frac{1}{d/2} \sum_{i=1}^{d/2} \theta_i^2, \quad \theta_{\text{max}} = \theta_1 = 10,000^{-2/d}. \quad (\text{A1})$$

Appendix A.1. Collision Support and the Status of the Fisher Information

The scalar entering the rank analysis is the collision support $S(t) = \|p_t\|_2^{-2}$, and Lemma 1 bounds it below by L^a directly from condition (D3): $\|p_t\|_2^2 \leq \|p_t\|_\infty \|p_t\|_1 = \|p_t\|_\infty \leq L^{-a}$. No further estimate is needed for the spectral results.

We record why a Fisher information argument does *not* provide an alternative route, since an earlier version of this work attempted one. With $g_t(s) := \partial_t \text{score}_t(s)$ an oscillatory function of $t - s$, averaging under a diffuse p_t does not suppress the variance $I(t) = \text{Var}_{s \sim p_t}[g_t]$ by the collision support: a diffuse average over many oscillation phases drives the *mean* $\langle g_t \rangle_{p_t}$ toward zero but leaves the second moment $\langle g_t^2 \rangle_{p_t} \approx \frac{1}{2}(a_i^2 + b_i^2)$ -summed close to its uniform-phase value, so $\text{Var}_{p_t}[g_t] = \Theta(\theta_{\text{eff}}^2/d)$ is essentially independent of $S(t)$. This matches the numerics ($I(t) \approx 0.21$, constant in L ; Section 6) and is why the analysis rests on the collision support rather than the Fisher information (Remark 4). The Fisher–Rao sphere of Proposition 2 remains the correct ambient geometry; it is not the source of the numerical collapse.

Appendix A.2. Trailing-Spectrum Bound for the Rank Collapse

We supply the estimate behind Lemma 2. The diffuse-attention matrix is numerically rank-deficient but *not* close to a low-rank matrix—its leading spectrum has many $\Theta(1)$ singular values (Section 6). The collapse is a property of the *trailing* spectrum, which we bound in aggregate. Write $A = \frac{1}{L} \mathbf{1}\mathbf{1}^\top + E$ with $E_{ts} = p_t(s) - 1/L$. By (D3), each row obeys

$$\|E_t\|_2^2 = \sum_s \left(p_t(s) - \frac{1}{L}\right)^2 \leq \sum_s p_t(s)^2 = \|p_t\|_2^2 \leq \|p_t\|_\infty \leq L^{-a}, \quad (\text{A2})$$

using $\|p_t\|_2^2 \leq \|p_t\|_\infty \|p_t\|_1$. This bound uses the ℓ_∞ control (D3) directly; it is *not* obtained from the entropy bound (D1), which would not suffice (a high-entropy row can have $\|p_t\|_\infty = \Theta(1)$). Summing over rows,

$$\|E\|_F^2 = \sum_t \|E_t\|_2^2 \leq L \cdot L^{-a} = L^{1-a}. \quad (\text{A3})$$

Since $\mathbf{1}/\sqrt{L}$ is a singular vector of A (row-stochasticity gives $A\mathbf{1} = \mathbf{1}$, fixing $\sigma_1 = \Theta(1)$) and E acts on its orthogonal complement up to the C_{circ}/L^2 near-circulant error of (D2), the trailing singular values satisfy $\sum_{k \geq 2} \sigma_k(A)^2 = \|E\|_F^2 + O(L^{-2}) \leq L^{1-a} + O(L^{-2})$. We do not claim each σ_k is individually small—it is not—only that their *sum of squares* is $O(L^{1-a})$. By Markov's inequality applied to the multiset $\{\sigma_k^2\}_{k \geq 2}$, at most L^{1-a}/τ^2 of them exceed any level τ^2 ; taking $\tau = \varepsilon_{\text{mach}} \sigma_1$, at most $1 + L^{1-a}/\varepsilon_{\text{mach}}^2$ singular values clear the roundoff floor, so the numerical-rank deficit is at least $L - 1 - L^{1-a}/\varepsilon_{\text{mach}}^2$, which grows with L for any fixed $\varepsilon_{\text{mach}}$ since $a < 1$ makes L^{1-a} grow slower than L . The exact growth rate of the deficit is measured in Section 6; the bound here establishes only that it diverges.

Appendix A.3. The Entropy-Invariant Temperature: Three Reference Scales

We obtain the schedule of Definition 3 and contrast it with two reference scales. By Lemma 3 the entropy obeys $dH/d \log \beta = -\beta^2 \text{Var}_\beta(e)$; in the diffuse regime $\beta^2 \text{Var}_\beta(e)$ is

near-constant over the operative range (slope -1.03 measured, Section 6), integrating to the empirically verified additive law $H(p_t; \beta) = H(p_t; 1) - \log \beta + o(1)$, Equation (13).

(a) *The diffuseness-pinning scale.* Holding the collision support $S(t) = \|p_t\|_2^{-2}$ at a length-independent target S_* , against the diffuse baseline $S(t; 1) = \Theta(L^a)$, requires contracting the ℓ_2 support by a factor $\Theta(L^a)$; since a softmax rescaled by β contracts its participation support by $\Theta(\beta^2)$ in the near-uniform regime, this needs $\beta_{\text{pin}}(L) = \Theta(L^{a/2})$. This pins the local diffuseness but is a stronger demand than entropy invariance.

(b) *The entropy-invariant scale (adopted).* Demanding $H(p_t; \beta(L), L) = H_{\text{ref}} = c_0 \log L_{\text{ref}}$ against the diffuse baseline $H(p_t; 1, L) = c_0 \log L + O(1)$ gives, via (13),

$$\log \beta_*(L) = c_0 \log(L/L_{\text{ref}}) \implies \beta_*(L) = (L/L_{\text{ref}})^{c_0}, \quad (\text{A4})$$

the schedule of Definition 3. Since H pinned at the reference value fixes the ℓ_∞ bound and the collision support of (D3) at their length-independent reference values, the per-row energy $\|p_t^{\text{corr}}\|_2^2 = \Theta(L_{\text{ref}}^{-a})$ in (A2) no longer shrinks with L , so the aggregate trailing energy is $L \Theta(L_{\text{ref}}^{-a})$ and the Markov count below the floor is a fixed fraction of L rather than a growing one, giving an $O(1)$ deficit—Theorem 1(ii). Note $\beta_* = \Theta(L^{c_0})$ and $\beta_{\text{pin}} = \Theta(L^{a/2})$ coincide in rate when $c_0 = a/2$; empirically $c_0 \approx 0.91$ and $a \approx 0.9$, so the two prescriptions are numerically close, both polynomial. We use the entropy-invariant form because entropy is the quantity for which the temperature law (13) is directly verified.

(c) *The conservative LogN scale.* The sub-polynomial choice $\beta(L) = \Theta(\sqrt{\log L})$ subtracts only $\frac{1}{2} \log \log L + O(1)$ from the entropy, an increment vanishing against the $\Theta(\log L)$ growth; it therefore only *reduces* the deficit (the $\rho \rightarrow 0$ endpoint of the trade-off (19), Remark 6), and is the rate used by the LogN-scaling and scale-invariant-attention family.

Appendix A.4. Why the Warp Leaves the Spectrum Unchanged

We record the estimate behind Proposition 3. Under the warp (16) the effective relative position is $\tilde{\Delta}(t, s) = (s - t) + \eta(\kappa_g(s) - \kappa_g(t))$ with $\kappa_g(t) = C_0 / \sqrt{\log(t+1)}$ monotone and $|\kappa_g(s) - \kappa_g(t)| \leq C_0 / \sqrt{\log 2}$ uniformly. Hence the warp perturbs each score by at most $\beta_0 \eta \theta_{\max} C_0 / \sqrt{\log 2} = O(\eta)$, a bounded amount that does not grow with L or $|t - s|$; the score remains $f(t - s) + O(\eta / \log L)$, still of the near-circulant form (D2) with the same profile to leading order. The warp does not alter the softmax temperature, so (D1) and the ℓ_∞ bound (D3) are preserved with the same exponents. The trailing-spectrum bound (A2)–(A3) therefore holds for A_{warp} with the same exponent, giving $\text{rank}_\epsilon(A_{\text{warp}}) = \text{rank}_\epsilon(A_{\text{std}})(1 + o(1))$.

Appendix B. Scalar Curvature of the Fisher–Rao Manifold

Appendix B.1. Isometry with the Round Sphere

The Fisher–Rao manifold admits a classical characterisation. Every probability distribution p over L outcomes maps to a point on a sphere by taking the square root of each probability and multiplying by two: the vector $(2\sqrt{p_1}, \dots, 2\sqrt{p_L})$ has squared norm $4 \sum_s p_s = 4$, so it lies on the sphere of radius two. This map $\varphi : \Delta_{L-1}^\circ \rightarrow S_+^{L-1}(2)$, $\varphi(p)_s = 2\sqrt{p_s}$, sends each distribution to the positive orthant of that sphere. It is a bijection, with differential $(d\varphi)_p(U)_s = U_s / \sqrt{p_s}$, and $\varphi^* g_{\text{round}} = G_F$ confirms φ is a Riemannian isometry, preserving lengths and angles exactly. The Fisher–Rao geometry of attention distributions is thus *exactly* that of a round sphere. By Cencov’s theorem [31] the Fisher–Rao metric is the unique monotone Riemannian metric on statistical manifolds up to scaling, grounding its use as the canonical geometry for the attention distribution space.

Appendix B.2. Curvature Identities

The round sphere $S^{L-1}(r)$ has constant sectional curvature $K = 1/r^2$. For $r = 2$: $K_{\text{sect}}^{\text{amb}} = 1/4$, $\text{Ric} = (L-2)/4 \cdot g_{\text{round}}$, and $K_{\text{stat}}^{\text{amb}} = (L-1)(L-2)/4 \approx L^2/4$. As established in Remark 2, this ambient scalar curvature does not enter the analysis, which depends instead on the Fisher information $I(t)$ along the attention curve. The following table compares standard RoPE and CC-RoPE under the entropy-invariant schedule $\beta_*(L) = (L/L_{\text{ref}})^{c_0}$ (Definition 3); under standard RoPE $\kappa(A)$ is at the roundoff floor and intrinsically ill-posed (Lemma 4), whereas the entropy-invariant correction lifts it off the floor:

Table A1. Diffuse-regime diagnostics under standard RoPE versus CC-RoPE.

Quantity	Standard RoPE	CC-RoPE ($\beta_* = (L/L_{\text{ref}})^{c_0}$)
$H(p_t)$	$\Theta(\log L)$	$\Theta(1)$ (pinned at $c_0 \log L_{\text{ref}}$)
$I(t)$	$\Theta(1/(d L^a))$	$\Theta(1/d)$ (restored)
$S(t) = \ p_t\ _2^{-2}$	$\Theta(L^a)$	$\Theta(L_{\text{ref}}^a) = \Theta(1)$ in L
$L - \text{rank}_\epsilon(A)$	grows with L	$O(1)$ ($\equiv 0$ in experiments)
$\kappa(A)$	roundoff floor (ill-posed)	lifted off floor ($\sim 10^{10}$)

References

1. Su, J.; Lu, Y.; Pan, S.; Murtadha, A.; Wen, B.; Liu, Y. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing* **2024**, *568*, 127063. [\[CrossRef\]](#)
2. Anson, B.; Wang, X.; Aitchison, L. Scale-invariant attention. *arXiv* **2025**, arXiv:2505.17083. [\[CrossRef\]](#)
3. Zhang, Z.; Wang, Y.; Huang, X.; Fang, T.; Zhang, H.; Deng, C.; Li, S.; Yu, D. Attention entropy is a key factor: An analysis of parallel context encoding with full-attention-based pre-trained language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Kerrville, TX, USA, 2025. [\[CrossRef\]](#)
4. Chen, S.; Wong, S.; Chen, L.; Tian, Y. Extending context window of large language models via positional interpolation. *arXiv* **2023**, arXiv:2306.15595. [\[CrossRef\]](#)
5. Peng, B.; Quesnelle, J.; Fan, H.; Shippole, E. YaRN: Efficient context window extension of large language models. *arXiv* **2023**, arXiv:2309.00071. [\[CrossRef\]](#)
6. Wang, S.; Zhong, N.; Ding, L.; Mo, T.; Mi, F.; Li, Z.; Bi, W. Resonance RoPE: Improving context length generalisation of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*; Association for Computational Linguistics: Bangkok, Thailand, 2024. [\[CrossRef\]](#)
7. Ding, J.; Ma, S.; Dong, L.; Zhang, X.; Huang, S.; Wang, W.; Wei, F. LongNet: Scaling transformers to 1,000,000,000 tokens. *arXiv* **2023**, arXiv:2307.02486. [\[CrossRef\]](#)
8. Chen, Y.; Qian, S.; Tang, H.; Lai, X.; Liu, Z.; Han, S.; Jia, J. LongLoRA: Efficient fine-tuning of long-context large language models. In *Proceedings of the 12th International Conference on Learning Representations*, Vienna, Austria, 7–11 May 2024. [\[CrossRef\]](#)
9. An, C.; Huang, F.; Zhang, J.; Gong, S.; Qiu, X.; Zhou, C.; Kong, L. Training-free long-context scaling of large language models. In *Proceedings of the 41st International Conference on Machine Learning*; PMLR: Vienna, Austria, 2024. [\[CrossRef\]](#)
10. Liu, Y.; Yu, J.; Xu, Y.; Li, Z.; Zhu, Q. A survey on transformer context extension: Approaches and evaluation. *arXiv* **2025**, arXiv:2503.13299. [\[CrossRef\]](#)
11. Baek, J. Numerical fragility in transformers: A layer-wise theory for explaining, forecasting, and mitigating instability. *arXiv* **2025**, arXiv:2510.21770. [\[CrossRef\]](#)
12. Saratchandran, H.; Lucey, S. Spectral conditioning of attention improves transformer performance. *arXiv* **2026**, arXiv:2603.07162. [\[CrossRef\]](#)
13. Kim, M.; Li, D.; Hu, S.X.; Hospedales, T.M. Fisher SAM: Information geometry and sharpness aware minimisation. In *Proceedings of the 39th International Conference on Machine Learning*; PMLR: Baltimore, MD, USA, 2022; pp. 11148–11161. [\[CrossRef\]](#)
14. Da Silva, M.F.; Dangel, F.; Oore, S. Hide & seek: Transformer symmetries obscure sharpness & Riemannian geometry finds it. In *Proceedings of the 42nd International Conference on Machine Learning*; PMLR: Vancouver, BC, Canada, 2025. [\[CrossRef\]](#)
15. Ji, Z. RiemannFormer: A framework for attention in curved spaces. *arXiv* **2025**, arXiv:2506.07405. [\[CrossRef\]](#)
16. Sun, Y.; Dong, L.; Patra, B.; Ma, S.; Huang, S.; Benhaim, A.; Chaudhary, V.; Song, X.; Wei, F. A length-extrapolatable transformer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Toronto, ON, Canada, 2023; pp. 14590–14604. [\[CrossRef\]](#)
17. Press, O.; Smith, N.A.; Lewis, M. Train short, test long: Attention with linear biases enables input length extrapolation. In *Proceedings of the 10th International Conference on Learning Representations*, Online, 25–29 April 2022. [\[CrossRef\]](#)

18. Chiang, T.-R.; Yogatama, D. The rotary position embedding may cause dimension inefficiency in attention heads for long-distance retrieval. *arXiv* **2025**, arXiv:2502.11276. [[CrossRef](#)]
19. Jonasson, A. Rotary offset features in large language models. *arXiv* **2025**, arXiv:2503.01832. [[CrossRef](#)]
20. Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.; Le, Q.; Salakhutdinov, R. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Florence, Italy, 2019; pp. 2978–2988. [[CrossRef](#)]
21. Mohtashami, A.; Jaggi, M. Landmark attention: Random-access infinite context length for transformers. In *Advances in Neural Information Processing Systems 36*; Curran Associates: Red Hook, NY, USA, 2023. [[CrossRef](#)]
22. Zhou, X.; Jiang, Z.; Ren, Z.; Zhang, Y.; Yu, T.; Ji, W.; Liu, Z. PDAViT: Pyramid dual-attention vision transformer. *Neurocomputing* **2025**, *662*, 131966. [[CrossRef](#)]
23. Zhou, X.; Ren, Z.; Zhang, Y.; Jiang, Z.; Yu, T.; Luo, H.; Zhou, S. ELAFormer: Early local attention in multi-scale vision transformers. *Knowl.-Based Syst.* **2025**, *325*, 113851. [[CrossRef](#)]
24. Zhou, X.; Jiang, Z.; Zhou, S.; Ren, Z.; Zhang, Y.; Yu, T.; Liu, Y. Frequency-assisted local attention in lower layers of visual transformers. *Int. J. Neural Syst.* **2025**, *35*, 2550015. [[CrossRef](#)] [[PubMed](#)]
25. Saratchandran, H. Preconditioned attention: Enhancing efficiency in transformers. *arXiv* **2026**, arXiv:2603.27153. [[CrossRef](#)]
26. Emadi, S.M. Exact attention sensitivity and the geometry of transformer stability. *arXiv* **2026**, arXiv:2602.18849. [[CrossRef](#)]
27. Amari, S.; Nagaoka, H. *Methods of Information Geometry*; American Mathematical Society: Providence, RI, USA, 2000.
28. Di Sipio, R. Rethinking LLM training through information geometry and quantum metrics. *arXiv* **2025**, arXiv:2506.15830. [[CrossRef](#)]
29. Do Carmo, M.P. *Riemannian Geometry*; Birkhäuser: Boston, MA, USA, 1992.
30. Milnor, J. Curvatures of left-invariant metrics on Lie groups. *Adv. Math.* **1976**, *21*, 293–329. [[CrossRef](#)]
31. Cencov, N.N. *Statistical Decision Rules and Optimal Inference*; American Mathematical Society: Providence, RI, USA, 1982.
32. Higham, N.J. *Accuracy and Stability of Numerical Algorithms*, 2nd ed.; SIAM: Philadelphia, PA, USA, 2002.
33. Trefethen, L.N.; Bau, D. *Numerical Linear Algebra*; SIAM: Philadelphia, PA, USA, 1997.
34. Blanchard, P.; Higham, D.J.; Higham, N.J. Accurately computing the log-sum-exp and softmax functions. *IMA J. Numer. Anal.* **2021**, *41*, 2311–2330. [[CrossRef](#)]
35. Biderman, S.; Schoelkopf, H.; Anthony, Q.; Bradley, H.; O'Brien, K.; Hallahan, E.; Khan, M.A.; Purohit, S.; Prashanth, U.S.; Raff, E.; et al. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*; PMLR: Honolulu, HI, USA, 2023; pp. 2397–2430. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.