RADIUS LEVEL AND RADIUS VARIATION IN PRE-NORMALIZED DECODERS: SHARP NORMALIZATION DISTORTION AND WIDTH DEPENDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the worst-case cost of RMS normalization in the single-sign vector Rademacher process—one Rademacher sign per example, a Euclidean norm outside the signed sum—that arises when a bounded linear readout is peeled from a hidden-state class. We separate out two effects : variation of hidden-state radii across hypotheses, and distance of those states from zero. For fixed $\nu > 0$ and $\rho_0 > 0$, on the annulus $1 \le \|h\|_\circ \le 1 + 2\rho$ with $0 \le \rho \le \rho_0$, the dimension-unconstrained sharp (with matching lower bound) order is $\Theta_{\nu,\rho_0}(1 + \rho\sqrt{m})$; at every fixed positive ρ the width-aware sharp order is $\Theta_{\rho,\nu}(1 + \sqrt{\min\{m, d\}})$, so distortion saturates at $\Theta_{\rho,\nu}(1 + \sqrt{d})$ in a fixed-width architecture. A uniform joint envelope for bounded ρ identifies the remaining gap; it matches throughout $0 < \rho \le 1$ and $m \le d/\rho$. Conditionally on a bound for the raw hidden-state complexity, these factors enter a standard bounded-linear-readout generalization inequality. For radius level, we derive sufficient residual-amplitude conditions under which an embedding floor propagates to a population hidden-state floor at every depth, and give explicit attention and SwiGLU amplitude bounds. A five-checkpoint audit on public pre-RMS decoders shows that the tested uniform certificates fail at native scale, although sampled radii remain positive throughout depth.

1 INTRODUCTION

Normalization appears at every sublayer of a pre-normalized transformer, but its role in capacity analysis is easy to misidentify. A small Lipschitz constant controls the distance between two normalized vectors; a Rademacher supremum asks which hypothesis makes a signed sum largest. Because the hypothesis determines both direction and radius, normalization can alter that choice. Two properties of the hidden-state radii therefore play distinct roles: *radius variation*—how much the radius at one example can differ across hypotheses—and *radius level*—how far the radii sit from zero. We find that they resolve differently.

We isolate the variation question for the Euclidean norm of a signed vector sum, with one sign per example. On a common annulus of half-width ρ, our sharp characterizing results yield a *dimension-unconstrained* worst-case cost of order $1+\rho\sqrt{m}$, so an arbitrarily thin annulus alone cannot guarantee a contraction uniform in both sample size and dimension. The threshold is $\rho = O(m^{-1/2})$. Our matching construction uses $d = m + 1$ and therefore does not assert unavoidable $\sqrt{m}$ growth inside a fixed-width architecture.

Three constructions organize the lower bounds. One with a shared direction and independently variable orthogonal components gives the dimension-unconstrained bound; it can be realized by a restricted decoder class and also with fixed embeddings via a single attention output projection. A grouped construction makes the role of available orthogonal directions explicit. A spherical-average contraction argument gives an $O(\sqrt{d})$ upper bound uniform in m; balanced groups match it to yield $\Theta_{\rho,\nu}(1 + \sqrt{\min\{m, d\}})$ at fixed $\rho > 0$. The same lower construction combined with both upper bounds gives a uniform joint envelope whose lower term is $\min\{\rho\sqrt{m}, \sqrt{\rho \min(m, d)}\}$ and whose upper term is $\min\{\rho\sqrt{m}, \sqrt{d}\}$. In particular, for $0 < \rho \le 1$ and $m \le d/\rho$ the sharp order remains

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

$1 + \rho\sqrt{m}$; in the remaining corner the envelopes differ by at most $\rho^{-1/2}$. The practically important specialization at fixed positive annulus width is $\Theta_{\rho,\nu}(1 + \sqrt{d})$.

For transformers, an annulus assumption alone leaves a practical gap: what parameter constraints guarantee the required radius control? We give a constructive answer that needs no reference profile. A branch has an absolute amplitude—bounding how far one residual update moves the stream—and an input sensitivity. On normalized inputs, the amplitude admits a bound independent of the stream's size and, for attention, of the query and key caps. These quantities yield a cumulative budget under which a population radius floor follows from an embedding-floor condition, while a separate sensitivity budget controls propagation. With a common reference embedding, the same budget supplies an across-hypothesis radius profile that composes the architectural and normalization results. We also carry the normalization result through a bounded linear readout and Lipschitz loss, connecting the single-sign vector process to an explicit generalization inequality.

Our main contributions are as follows.

1. In §3, we establish first-of-its-kind three complementary matching lower bounds for the normalization distortion: the dimension-unconstrained order $1 + \rho\sqrt{m}$; the joint lower envelope $1 + \min\{\rho\sqrt{m}, \sqrt{\rho \min(m, d)}\}$; and, at fixed positive ρ, the matching width-aware order $1 + \sqrt{\min\{m, d\}}$. A linear-readout corollary carries these factors into a scalar generalization inequality conditional on the raw hidden-state complexity.
2. In §4, we derive radius-level mechanisms from architecture: explicit attention and SwiGLU amplitude bounds, an isotropic residual-amplitude theorem deriving a population floor from an embedding floor, and its composition with the normalization results. Sensitivity, propagation, and a directional refinement are deferred to the appendix.
3. In §5, we verify the sharp constructions deterministically, audit the grouped family across available directions, and compare certificates with realized quantities on five public checkpoints. The sufficient budgets fail at native scale while sampled radii remain positive.

Table 1 summarizes the sharp regimes where we have matching lower bound alongside upper bounds.

1.1 Related literature

RMSNorm (Zhang & Sennrich, 2019) and LayerNorm (Ba et al., 2016) normalize different radial quantities; the obstruction below applies to the former and, after restricting to the mean-zero subspace, to the latter. The zero-bias CenterNorm map of LipsFormer (Qi et al., 2023) is a fixed linear map and contracts immediately (Proposition C.1), which isolates the role of division by a class-dependent radius. Maurer (2016) rules out a generic dimension-free single-sign contraction (Section 6 of that paper). Our contribution is the RMS-specific annulus and width dependence, , including the first explicit lower-bound constructions for this process. Maurer's positive vector-contraction theorem instead uses independent signs across sample and coordinate indices, so it does not directly bound the process in (2). Truong's high-dimensional contraction lemmas (Truong, 2025) likewise treat coordinatewise vector maps; RMS normalization is non-separable because every coordinate is divided by a common radial functional. The peeling technique of Golowich et al. (2018) uses degree-one positive homogeneity; at $\nu = 0$, RMS normalization is homogeneous of degree zero, and at fixed $\nu > 0$ exact scale invariance also fails. Covering-based analyses (Edelman et al., 2022; Trauger & Tewari, 2024) sidestep the contraction question. Spectrum-adaptive analyses (Sakai & Imaizumi, 2026) emphasize checkpoint-dependent spectral structure rather than fixed norm budgets. Residual-scaling schemes (Qi et al., 2023) are an architectural analogue of the amplitude budget in §4. For residual parameterization and initialization, see Bachlechner et al. (2021); De & Smith (2020); Wang et al. (2024); Noci et al. (2022); for attention Lipschitz behavior, see Kim et al. (2021); Castin et al. (2024). Our normalization theorem concerns a single-sign complexity ratio, which is a different quantity. An extended discussion of these connections appears in Appendix J.

2 Setup and the contraction question

For row vectors $h \in \mathbb{R}^d$ and streams $H \in \mathbb{R}^{T\times d}$, write $\|h\|_\circ = \|h\|_2 / \sqrt{d}$ (the *RMS radius* of h) and $\|H\|_{\infty,\circ} = \max_t \|H_t\|_\circ$. The mixed norm is $\|W\|_{2,1} = \sum_j \|W_{\cdot j}\|_2$; it sums *columns* in our

row-vector convention. Calligraphic letters ($\mathcal{H}$, $\mathcal{Q}$) denote evaluation classes, roman H with stage superscripts denotes a stream, and the attention head count is written n_h throughout. Fix $\nu > 0$ and define

$$\mathcal{N}(h) = \frac{h}{\sqrt{\|h\|_\circ^2 + \nu}}, \qquad \mathrm{RMS}_\gamma(h) = \mathcal{N}(h)\,\mathrm{diag}(\gamma). \tag{1}$$

For a nonempty class $\mathcal{H} = \{(h_i(\theta))_{i=1}^m : \theta \in \Theta\}$, let

$$\mathfrak{R}_m^{\mathrm{vec}}(\mathcal{H}) = \frac{1}{m}\mathbb{E}_\varepsilon \sup_{\theta\in\Theta} \left\| \sum_{i=1}^m \varepsilon_i h_i(\theta) \right\|_2 . \tag{2}$$

There is one independent sign per evaluation, and scalar empirical Rademacher complexity $\widehat{\mathfrak{R}}_m$ uses $m^{-1}\mathbb{E}_\varepsilon \sup_f \sum_i \varepsilon_i f_i$ without an absolute value. For $\mathcal{F}_B := \{i \mapsto \langle w, h_i(\theta)\rangle : \theta \in \Theta,\ \|w\|_2 \le B\}$, Euclidean duality gives

$$\widehat{\mathfrak{R}}_m(\mathcal{F}_B) = B\,\mathfrak{R}_m^{\mathrm{vec}}(\mathcal{H}). \tag{3}$$

The single sign per example is thus the ordinary scalar sign after optimizing a linear readout; the norm outside the sum is not an auxiliary choice.

Regime	Distortion	Lower construction
$\rho = 0$	exactly 1	singleton baseline
Width unrestricted, $0 \le \rho \le \rho_0$	$\Theta_{\nu,\rho_0}(1 + \rho\sqrt{m})$	orthogonal directions, $d = m+1$; fixed-embedding attention realization
Bounded ρ, arbitrary m, d	$1+\min\{\rho\sqrt{m}, \sqrt{\rho\min(m,d)}\}$ (lower) $1+\min\{\rho\sqrt{m}, \sqrt{d}\}$ (upper)	balanced groups, $k = \min\{m, d-1\}$; matching when $m \le d/\rho$
Fixed $\rho > 0$, arbitrary m, d	$\Theta_{\rho,\nu}(1 + \sqrt{\min\{m,d\}})$	balanced groups
Fixed d, fixed $\rho > 0$, $m \to \infty$	$\Theta_{\rho,\nu}(1 + \sqrt{d})$	specialization of balanced groups

Table 1: Sharp regimes for the normalization distortion. The unrestricted-width and fixed-positive-ρ rows are matching theorems. Gap remains only when $0 < \rho \le \min\{1, \rho_0\}$ and $m > d/\rho$; see §6.

If all radii are at least $c \ge 0$, the gain-free normalizer is $b_c = (c^2 + \nu)^{-1/2}$-Lipschitz on that region. This follows from Lemma A.1 below. The tempting inference is

$$\mathfrak{R}_m^{\mathrm{vec}}(\mathcal{N} \circ \mathcal{H}) \ \le\ C b_c \mathfrak{R}_m^{\mathrm{vec}}(\mathcal{H}) \quad \text{with a constant } C \text{ independent of } m, d, \mathcal{H}. \tag{4}$$

Section 3 characterizes precisely how this inference fails on common annuli, at the level of the evaluation-class process (2).

3 The sharp cost of radius variation

Theorem 3.1 (Contraction with a reference radius profile)**.** *Suppose $\|h_i(\theta)\|_\circ \ge c$, $\|h_i(\theta)\|_2 \le M$, and $\|\gamma(\theta)\|_\infty \le G$. If numbers r_i^0, fixed across θ, satisfy $|\,\|h_i(\theta)\|_\circ - r_i^0| \le \rho$, then*

$$\mathfrak{R}_m^{\mathrm{vec}}(\mathrm{RMS}_\gamma \circ \mathcal{H}) \le G\{b_c \mathfrak{R}_m^{\mathrm{vec}}(\mathcal{H}) + M\rho b_c^2\}. \tag{5}$$

In particular, fixed radii across hypotheses give the constant-one bound $G b_c \mathfrak{R}_m^{\mathrm{vec}}(\mathcal{H})$.

Proof. See Appendix C.

The profile may vary between examples; the condition controls variation *between hypotheses at the same example*.

For $\rho \ge 0$, define

$$C^\star_{m,\nu}(\rho) = \sup_{\substack{d\ge 2,\ \mathcal{H}\neq\varnothing \\ 1\le\|h_i(\theta)\|_\circ\le 1+2\rho\ \forall i,\theta}} \frac{\sqrt{1+\nu}\,\mathfrak{R}_m^{\mathrm{vec}}(\mathcal{N}\circ\mathcal{H})}{\mathfrak{R}_m^{\mathrm{vec}}(\mathcal{H})}. \tag{6}$$

The denominator is positive under the stated floor.

Theorem 3.2 (Sharp dimension-unconstrained annulus scaling). *Fix $\nu > 0$ and $\rho_0 > 0$. There are explicit positive constants c_{ν,ρ_0} and C_{ν,ρ_0} such that, for every $m \geq 1$ and $0 \leq \rho \leq \rho_0$,*

$$\max\{1, c_{\nu,\rho_0}\rho\sqrt{m}\} \leq C^\star_{m,\nu}(\rho) \leq 1 + C_{\nu,\rho_0}\rho\sqrt{m}. \tag{7}$$

Consequently $C^\star_{m,\nu}(\rho) = \Theta_{\nu,\rho_0}(1 + \rho\sqrt{m})$. For bounded sequences ρ_m, the worst-case distortion is uniformly bounded exactly when $\sup_m \rho_m\sqrt{m} < \infty$. The proof also gives the sharper ρ-specific upper bound

$$C^\star_{m,\nu}(\rho) \leq 1 + \frac{\sqrt{2}(1+2\rho)\rho\sqrt{m}}{\sqrt{1+\nu}}, \tag{8}$$

before replacing $1 + 2\rho$ by $1 + 2\rho_0$ to obtain the uniform constant in (7).

The supremum in (6) ranges over d, and the matching construction takes $d = m + 1$; Proposition 3.7 and Theorem 3.6 together give the matching fixed-width order.

Proof sketch. Upper bound. Lemma C.4 gives $\mathfrak{R}^{\mathrm{vec}}_m(\mathcal{H}) \geq \sqrt{d/(2m)}$; apply Theorem 3.1 with $r_i^0 = 1 + \rho$ and $M = \sqrt{d}(1 + 2\rho)$. *Lower bound.* Set $d = m + 1$, choose an orthonormal basis $e_0, \ldots, e_m$, and with $t^2 = 4\rho(1 + \rho)$ let

$$h_i(a) = \sqrt{d}(e_0 + a_i e_i), \qquad a \in \{0, t\}^m. \tag{9}$$

Every radius is 1 or $1 + 2\rho$. Before normalization the shared component e_0 is independent of a; after normalization its coefficient can respond to the signs, because the orthogonal coordinates shift the denominator. Setting $a_i = 0$ for positive signs and $a_i = t$ for negative signs produces a lower bound of order $\rho\sqrt{m}$; a singleton radius-one class gives the separate lower bound one. See Appendix C for explicit constants. □

Figure 4 in Appendix C illustrates the mechanism in the (e_0, e_i) plane.

3.1 A LINEAR-READOUT GENERALIZATION CONSEQUENCE

Corollary 3.3 (Generalization after RMS normalization). *Let $S = ((X_i, Y_i))_{i=1}^m$ be i.i.d., fix one decoder output position, and let $\mathcal{H}_S$ be the hidden-state evaluation class. For $f_{\theta,w}(X) = \langle w, \mathcal{N}(h_\theta(X))\rangle$ with $\|w\|_2 \leq B$, if $\ell(y, \cdot)$ is L_ℓ-Lipschitz and $0 \leq \ell \leq U$, then for any $0 < \delta < 1$, with probability at least $1 - \delta$,* simultaneously *for all $\theta \in \Theta$ and $\|w\|_2 \leq B$,*

$$\mathbb{E}\,\ell(Y, f_{\theta,w}(X)) \leq \frac{1}{m}\sum_{i=1}^m \ell(Y_i, f_{\theta,w}(X_i)) + 2L_\ell B\,\mathfrak{R}^{\mathrm{vec}}_m(\mathcal{N} \circ \mathcal{H}_S) + 3U\sqrt{\frac{\log(2/\delta)}{2m}}. \tag{10}$$

Under the reference-profile assumptions of Theorem 3.1 with gain one, the complexity term is at most $2L_\ell B\{b_c\mathfrak{R}^{\mathrm{vec}}_m(\mathcal{H}_S) + M\rho b_c^2\}$; under the annulus $[1, 1 + 2\rho]$, it is at most $\frac{2L_\ell B}{\sqrt{1+\nu}}(1 + \frac{\sqrt{2}(1+2\rho)\rho\sqrt{m}}{\sqrt{1+\nu}})\mathfrak{R}^{\mathrm{vec}}_m(\mathcal{H}_S)$. The factor in parentheses is replaced by the smaller bound in (13).

Proof. The L_ℓ-Lipschitz contraction and (3) give loss-class complexity at most $L_\ell B\mathfrak{R}^{\mathrm{vec}}_m(\mathcal{N} \circ \mathcal{H}_S)$; a standard Rademacher bound (Bartlett & Mendelson, 2002) gives (10). The two specializations follow from Theorem 3.1 and (8); see Appendix C.1. □

The corollary identifies the proof step affected by radius variation: treating RMS normalization as an ordinary contraction retains only $b_c\mathfrak{R}^{\mathrm{vec}}_m(\mathcal{H}_S)$, and the profile-width term is the extra price for a hypothesis-dependent denominator. At fixed d, c, R, a raw complexity $\mathfrak{R}^{\mathrm{vec}}_m(\mathcal{H}_S) = O(m^{-1/2})$ gives a decaying generalization penalty; neither the width theorem nor a radius floor alone proves this assumption. Here m counts independent examples, not correlated tokens.

Proposition 3.4 (Fixed-embedding attention realization). *Fix $m \geq 1$, put $d = m+1$, $t^2 = 4\rho(1+\rho)$, and $s = \sqrt{1 + t^2 + \nu}$. There is a one-head, length-one pre-RMS attention residual block with fixed token embeddings, fixed query/key/value maps, residual coefficient one, and only the attention output projection varying, whose post-residual evaluation class on m distinct tokens is exactly the class in (9). The varying output projection obeys $\|W_O\|_{\mathrm{op}} \leq s$.*

Proof. See Appendix C.7.

Remark 3.5 (Realization and scope). The class (9) is also realized by a learned embedding table (Proposition C.5) and, in the mean-zero subspace of $\mathbb{R}^{m+2}$, applies to LayerNorm (Corollary C.6). The RMS realization uses $d = m+1$; LayerNorm uses $d = m+2$. In each case width grows with m, so fixed-width behavior is a separate question. Sharpness for the full norm ball and under Frobenius or $(2,1)$ budgets remains open; see Appendix C.7.

To expose width explicitly, define the dimension-restricted distortion

$$C^{\star}_{m,d,\nu}(\rho) = \sup_{\substack{\mathcal{H}\subset(\mathbb{R}^d)^m,\ \mathcal{H}\neq\varnothing \\ 1\le\|h_i(\theta)\|_\circ\le 1+2\rho}} \frac{\sqrt{1+\nu}\,\mathfrak{R}^{\mathrm{vec}}_m(\mathcal{N}\circ\mathcal{H})}{\mathfrak{R}^{\mathrm{vec}}_m(\mathcal{H})}.$$

3.2 Uniform contraction at fixed width

Let $\kappa_d = \Gamma(d/2)/(\sqrt{\pi}\Gamma((d+1)/2))$. The next bound replaces dependence on the sample size by dependence on width.

Theorem 3.6 (Fixed-width normalization). *Suppose $0 \le c \le \|h_i(\theta)\|_\circ \le R < \infty$ for all i, θ, and let $b_c = (c^2+\nu)^{-1/2}$. If a common gain for each hypothesis obeys $\|\gamma(\theta)\|_\infty \le G_0$, then*

$$\mathfrak{R}^{\mathrm{vec}}_m(\mathrm{RMS}_\gamma \circ \mathcal{H}) \le 2G_0\left(b_c + \frac{Rb_c^2}{\kappa_d}\right)\mathfrak{R}^{\mathrm{vec}}_m(\mathcal{H}). \tag{11}$$

In particular, uniformly in m,

$$C^{\star}_{m,d,\nu}(\rho) \le 2\left[1 + \frac{1+2\rho}{\kappa_d\sqrt{1+\nu}}\right] \le 2 + (1+2\rho)\sqrt{\frac{2\pi d}{1+\nu}}. \tag{12}$$

Combining this with Theorem 3.2 gives

$$C^{\star}_{m,d,\nu}(\rho) \le \min\left\{1 + \frac{\sqrt{2}(1+2\rho)\rho\sqrt{m}}{\sqrt{1+\nu}},\quad 2 + (1+2\rho)\sqrt{\frac{2\pi d}{1+\nu}}\right\}. \tag{13}$$

Proof sketch. For a uniform sphere direction U, we have $\|x\|_2 = \kappa_d^{-1}\mathbb{E}_U|\langle U, x\rangle|$ with $\kappa_d^{-1} \asymp \sqrt{d}$. Scalar contraction bounds the RMS-radius complexity by $\sqrt{\pi/2}\,\mathfrak{R}^{\mathrm{vec}}_m(\mathcal{H})$; a second contraction handles the inverse radius, and the scalar product inequality (Lemma C.2) controls normalized projections. This explains why the bound is m-free and where the $\sqrt{d}$ enters. If all vectors lie in a fixed r-dimensional subspace, κ_d may be replaced by κ_r (Remark G.2). See Appendix G. □

Proposition 3.7 (Balanced groups). *Let $1 \le k \le \min\{m, d-1\}$, $t^2 = 4\rho(1+\rho)$, and $a_{\rho,\nu} = \sqrt{(1+\nu)/(1+t^2+\nu)}$. For balanced groups $n_j \in \{\lfloor m/k\rfloor, \lceil m/k\rceil\}$,*

$$C^{\star}_{m,d,\nu}(\rho) \ge \max\left\{1,\ \frac{1-a_{\rho,\nu}}{2\sqrt{m^{-1}+2t^2/k}}\right\}. \tag{14}$$

A general-group formula and the shrinking-width lower rate are in Appendix C.8.

Theorem 3.8 (Uniform joint radius–width envelope). *Fix $\nu > 0$ and $\rho_0 > 0$. There are constants $0 < c_{\nu,\rho_0} \le C_{\nu,\rho_0} < \infty$ such that, for every $m \ge 1$, $d \ge 2$, and $0 \le \rho \le \rho_0$,*

$$\max\left\{1, c_{\nu,\rho_0}\min\left(\rho\sqrt{m}, \sqrt{\rho\min\{m,d\}}\right)\right\} \le C^{\star}_{m,d,\nu}(\rho) \tag{15}$$

$$\le 1 + C_{\nu,\rho_0}\min\left(\rho\sqrt{m}, \sqrt{d}\right). \tag{16}$$

Consequently, for $0 < \rho \le 1$ and $m \le d/\rho$, $C^{\star}_{m,d,\nu}(\rho) = \Theta_{\nu,\rho_0}(1+\rho\sqrt{m})$. If $0 < \rho \le \min\{1, \rho_0\}$ and $m > d/\rho$, the upper and lower nonconstant envelope terms differ by at most a factor of order $\rho^{-1/2}$, with constants depending only on ν and ρ_0.

Proof. Take $k = \min\{m, d-1\}$ balanced groups. See Appendix G.

Corollary 3.9 (Sharp width order at fixed ρ)**.** *For every fixed $\rho > 0$ and $\nu > 0$, $C^{\star}_{m,d,\nu}(\rho) = \Theta_{\rho,\nu}(1 + \sqrt{\min\{m, d\}})$. At $\rho = 0$, $C^{\star}_{m,d,\nu}(\rho) = 1$. In particular, at fixed d and $\rho > 0$, $C^{\star}_{m,d,\nu}(\rho) = \Theta(1 + \sqrt{d})$ as $m \to \infty$.*

Proof. See Appendix G.

This settles the fixed-positive-width order, including $m \gg d$.

4 Certifiable radius floors from an amplitude budget

The normalization results require a common annulus. We now derive such control from architecture. A branch has an *amplitude*, bounding how far one residual update moves the stream, and a *sensitivity*, bounding how a perturbation propagates. On normalized inputs, the amplitude is bounded independently of the stream's size and, for attention, of the query and key caps. Throughout, the stream evolves by

$$H^s = H^{s-1} + \lambda_s F_s\big(\mathcal{N}(H^{s-1})\big), \qquad s = 1, \dots, D, \tag{17}$$

with fixed $\lambda_s \geq 0$ and $\mathcal{N}$ the gain-free RMS normalizer. Here D denotes the number of residual sublayers; for a standard decoder block with one attention and one feedforward residual branch, $D = 2L$ when L is the number of transformer blocks.

4.1 Absolute branch-magnitude certificates

Proposition 4.1 (Attention and SwiGLU magnitude bounds)**.** *Suppose $\|Z\|_{\infty,\circ} \leq 1$.* Attention: *with $n_h = gJ$ heads of per-head dimension $d_h = d/n_h$, arranged in J key/value groups of g query heads each (grouped-query attention; standard multi-head attention is $J = n_h$, $g = 1$), value caps $\|A_{V,\mu}\|_{\mathrm{op}} \leq v_\mu$, output cap $\|W_O\|_{\mathrm{op}} \leq S_O$, and $\Sigma_V := (g \sum_\mu v_\mu^2)^{1/2}$,*

$$\|F_{\mathrm{att}}(Z)\|_{\infty,\circ} \leq M_{\mathrm{att}} := S_O \Sigma_V, \tag{18}$$

independently of the query and key norm caps. SwiGLU: *with spectral caps S_g, S_u, S_d, column caps C_g, C_u, $\beta_a := \sqrt{d}\, C_g$, $\beta_b := \sqrt{d}\, C_u$,*

$$\|F_{\mathrm{ff}}(Z)\|_{\infty,\circ} \leq M_{\mathrm{ff}} := S_d \min\{\beta_a S_u,\ \beta_b S_g\}. \tag{19}$$

Proof. See Appendix A.2.

Remark 4.2 (Sensitivities and learned gains). Input-sensitivity constants $\Lambda_{\mathrm{att}}, \Lambda_{\mathrm{ff}}$ appear in Proposition A.2 (Appendix A.3); query/key terms enter only the sensitivity, so the amplitude is typically much smaller. Learned RMSNorm gains γ are absorbed into the first linear maps of the following branch; all formulas are read for these effective matrices.

4.2 From an amplitude budget to a population radius floor

Theorem 4.3 (Residual-amplitude budget implies a population radius floor)**.** *Assume that for each residual branch there is a deterministic constant $M_s \geq 0$ such that*

$$\sup_{W_s} \sup_{\|Z\|_{\infty,\circ} \leq 1} \|F_s(Z; W_s)\|_{\infty,\circ} \leq M_s. \tag{20}$$

Suppose the embedding stream has a uniform population floor $a_0 := \inf_\theta \inf_{X \in \mathcal{X}} \min_t \big\|H_t^0(X;\theta)\big\|_\circ > 0$, and define $A_s := \sum_{u=1}^{s} \lambda_u M_u$. If $A_D < a_0$, then every model in the class satisfies

$$\inf_{X \in \mathcal{X}} \min_t \|H_t^s(X;\theta)\|_\circ \geq c_s := a_0 - A_s > 0. \tag{21}$$

Proof. See Appendix A.4.

We see that the induction is a reverse-triangle argument connecting architecture-specific branch amplitudes to a uniform population statement.

Proposition 4.4 (Optimality from amplitude information alone)**.** *At the abstraction level of Theorem 4.3, the floor $a_0 - A_s$ cannot be improved using only a_0, the residual coefficients $\{\lambda_u\}$, and the branch-magnitude bounds $\{M_u\}$. In fact there exists a one-dimensional residual system satisfying the magnitude assumptions for which*

$$\|H^s\|_\circ = a_0 - A_s \qquad \textit{for every } s.$$

Proof. See Appendix A.5.

Remark 4.5 (Directional and local-reference variants). A directional variant (Theorem A.3, Appendix A.6) replaces S_O, S_d by directional caps $S_{O,w}, S_{d,w}$; a checkable local-reference version (Corollary A.4, Appendix A.7) uses a reference embedding E^0 and an embedding variation bound e_E. We keep the three support minima—full vocabulary, population, observed—separate in Experiment E3.

4.3 Composing architecture with normalization

Proposition 4.6 (Embedding reference profile)**.** *Fix a reference embedding table E^0 common to the class. Suppose $H_t^0 = E_{X_t}$ with $\max_v \left\|E_v - E_v^0\right\|_\circ \le e_E$, and let $a_E = \inf_v \left\|E_v^0\right\|_\circ$, $b_E = \sup_v \left\|E_v^0\right\|_\circ$, $\delta_s = e_E + A_s$. Then, uniformly over the class, inputs, and positions,*

$$\left\|H_t^s - E_{X_t}^0\right\|_\circ \le \delta_s, \qquad \left|\|H_t^s\|_\circ - \left\|E_{X_t}^0\right\|_\circ\right| \le \delta_s. \tag{22}$$

If $\delta_D < a_E$, set $c = a_E - \delta_D$ and $R = b_E + \delta_D$. For the final-layer evaluation class $\mathcal{H}_S$ and gain one,

$$\mathfrak{R}_m^{\mathrm{vec}}(\mathcal{N} \circ \mathcal{H}_S) \le b_c \mathfrak{R}_m^{\mathrm{vec}}(\mathcal{H}_S) + \sqrt{d}\, R \delta_D b_c^2, \tag{23}$$

$$\mathfrak{R}_m^{\mathrm{vec}}(\mathcal{N} \circ \mathcal{H}_S) \le 2\left(b_c + \frac{R b_c^2}{\kappa_d}\right) \mathfrak{R}_m^{\mathrm{vec}}(\mathcal{H}_S). \tag{24}$$

Both bounds hold simultaneously, so their minimum may be used.

Proof. Telescoping from E^0 gives (22). Since $\delta_D < a_E$, every final state lies in $[c, R]$, and Theorems 3.1 and 3.6 apply. Substituting the minimum into Corollary 3.3 gives the full architectural composition. □

4.4 Propagation, interpretation, and scope

The amplitude budget $\sum_s \lambda_s M_s$ prevents residual motion from consuming the embedding radius. A separate sensitivity budget $B_D := \sum_s \lambda_s \Lambda_s$, where Λ_s are input-Lipschitz constants (Proposition A.2), yields a propagation bound $\Pi_0 \le \exp\{B_D/\sqrt{c_D^2 + \nu}\}$ that is depth-independent whenever both budgets are uniformly bounded (Corollary A.6). A hybrid formula sharpens this for SwiGLU stages (Proposition H.1). The full propagation analysis, two-model perturbation recurrence, and finite-cover consumer appear in Appendix H.

5 Verification and targeted audits

We illustrate the proved constructions and audit the sufficient conditions on differentcheckpoints.

E1: deterministic sharp-scaling verification. The exact 80-point binomial sweep agrees with exhaustive enumeration for $m \le 10$ and satisfies both analytic bounds at every point. At $m = 2048$ the largest exact family ratio is 5.405; the mirrored assignment (Proposition C.7) lies within 0.02 of all six ratios in Table 3. See Appendix E for the protocol.

E2: fixed-width grouped-family audit. We instantiate Proposition 3.7's explicit family at $k \in \{4, 16, 64, 256, 1024\}$, $\rho \in \{0.025, 0.1, (\sqrt{2}-1)/2\}$, and $\nu \in \{10^{-3}, 0.1\}$, producing 198 configurations with values as large as $m = 1{,}048{,}576$. The class supremum is computed exactly by the group-count reduction of Appendix F.1; only the outer expectation is Monte Carlo (65,536 paired draws per row; largest SE 4.23×10^{-3}). No row violates either analytic bound, even after a

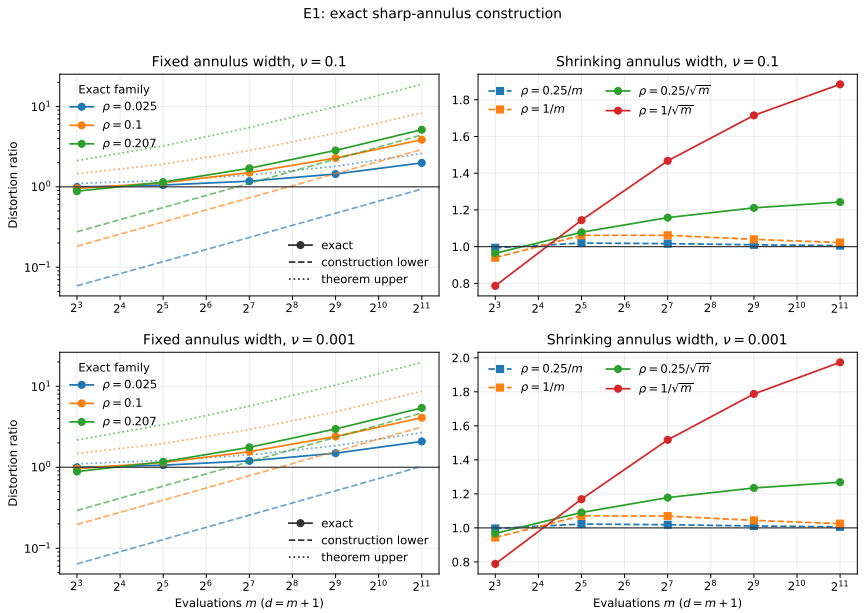

Figure 1: Exact E1 scaling. Fixed-annulus panels compare exact finite sums with construction lower and theorem upper bounds. Shrinking-annulus panels illustrate the threshold regimes. Every construction uses $d = m + 1$.

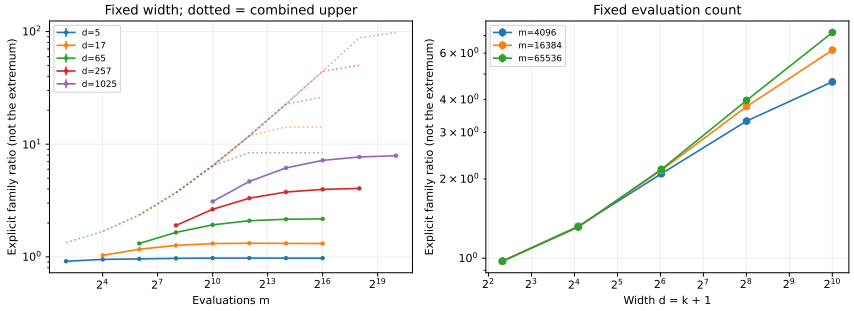

Figure 2: Fixed-width grouped-family audit at $\rho = 0.1$, $\nu = 10^{-3}$. Left: evaluation count varies at fixed width; dotted curves are the combined upper bounds. Right: width varies at three fixed evaluation counts. Solid points are Monte Carlo estimates with ± 2 SE bars.

six-standard-error allowance. For the 30 cases with $n = 1$, the grouped calculation agrees with the exact binomial formula from E1 within 2.17 Monte Carlo standard errors. We see in Figure 2 the two limits cleanly separated: holding d fixed and increasing m yields a bounded curve (Theorem 3.6), while increasing width at $m = 65{,}536$ recovers the $\sqrt{d}$ dependence of Corollary 3.9.

E3: pretrained audit of certificates and realized radii. We evaluate five public pre-RMS/SwiGLU checkpoints: base and instruct variants of SmolLM2-135M and SmolLM2-360M, and TinyLlama-1.1B-Chat (exact revisions and WikiText-103 validation data (Merity et al., 2017) archived; protocol in Appendix F).

The cumulative amplitude diagnostic $\widehat{G}_M^{\mathrm{all}}$ (power-iteration estimates) reaches 3.53×10^6–7.60×10^6 for the SmolLM2 checkpoints; TinyLlama, with near-zero full-vocabulary embedding rows, reaches 8.97×10^{23}. The induced-norm route is still larger, so no checkpoint yields a positive native-scale amplitude floor. All directional variants (Appendix A.6) are likewise negative.

We therefore run the checkpoints themselves. Twenty-eight evaluation chunks of 256 tokens give 7,168 sampled tokens per sublayer. Table 2 reports the results: all sampled radii remain positive, with minima 0.091–0.113 for SmolLM2 and 0.003 for TinyLlama.

The paired trajectory diagnostic $L_{\text{paired}}(D)$ (Appendix I.1) remains positive through depth and ends between 0.080 and 2.70. The earlier pooled diagnostic $L_{\text{old}}(D)$ reached 10^{-10}–10^{-6}; this comparison changes both the retained information and the aggregation across trajectories, so it does not isolate the benefit of magnitude information alone.

Model	a_E^{all}	a_E^{obs}	$\widehat{G}_M^{\text{all}}$	sampled $r_{\min}$
SmolLM2-135M	0.073	0.073	3.53×10^{6}	0.091
SmolLM2-135M-Instr.	0.075	0.075	3.54×10^{6}	0.096
SmolLM2-360M	0.073	0.078	7.60×10^{6}	0.108
SmolLM2-360M-Instr.	0.075	0.078	7.49×10^{6}	0.113
TinyLlama-1.1B-Chat	7.82×10^{-21}	0.005	8.97×10^{23}	0.003

Table 2: Checkpoint audit (compact). $a_E^{\text{all}}/a_E^{\text{obs}}$: minimum embedding RMS radius over the full vocabulary / observed tokens. $\widehat{G}_M^{\text{all}}$: cumulative amplitude diagnostic. Sampled $r_{\min}$: minimum post-update RMS radius over 7,168 tokens per sublayer. The complete table appears as Table 5 in Appendix F.

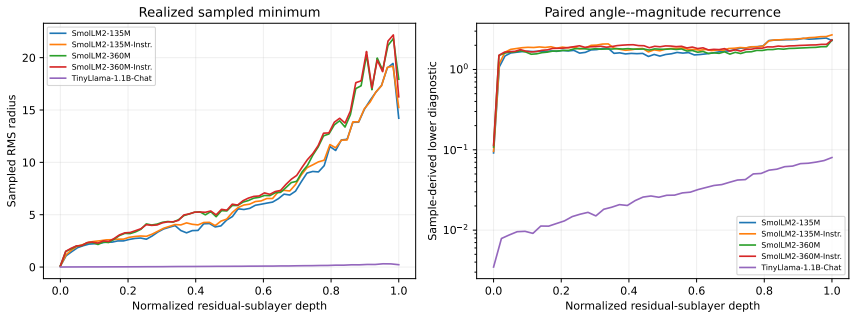

Figure 3: Realized-radius and alignment diagnostics on WikiText-103 evaluation chunks. Left: minimum post-update RMS radius per sublayer. Right: the paired trajectory diagnostic L_{paired} (Appendix I.1). Both panels are sampled diagnostics.

The archived attention audit (3,949,568 held-out rows) confirms that the largest layerwise minimum of $\max_j p_j$ is only 0.121, so the segment-stable concentration refinement (Appendix D) is inactive. Using (11) with gain one at the sampled final stage, the readout factor $K(c, R, d, \nu)$ evaluates to 99–232 (Appendix Table 6).

Together, this auditing deliver the same practical message: the sufficient conditions are conservative for native checkpoints, but the forward passes exhibit noncollapsed radii on the sampled text.

6 Discussion and Limitations

The two radius questions resolve differently. Radius *variation* creates a sharp worst-case cost of $\Theta(1+\rho\sqrt{m})$ when width is unrestricted and $\Theta(1+\sqrt{\min\{m,d\}})$ at fixed positive annulus width; the m-dependent divergence is a growing-width phenomenon. Radius *level* can be certified by cumulative branch amplitude, but the resulting certificates are far from tight on current decoders.

For construction (9) with $d = m + 1$, we see raw complexity of order $m^{-1/2}$ while normalized complexity stays bounded below at fixed $\rho > 0$ and ν. That conclusion concerns the growing-width construction, not every hidden-state class; Corollary 3.9 bounds the fixed-width factor.

On the audit side, the isotropic floor is optimal from amplitude information alone, yet E3 shows its estimate is of order 10^6–10^7 on SmolLM2. The directional variant's cumulative budget is the active failure for every direction tested. Complementary local analyses that exploit checkpoint-dependent structure (Sakai & Imaizumi, 2026; Sun et al., 2024; Xiao et al., 2024) are a natural next step; a concentration refinement for attention sensitivity (Appendix D) is also inactive at every audited stage.

Open problems. Three questions remain. (i) Sharpness of the annulus lower bound for the complete output-projection norm ball and under Frobenius or $(2, 1)$ budgets (Remark 3.5). (ii) The exact joint rate in the corner $0 < \rho \le \min\{1, \rho_0\}$, $m > d/\rho$, where the envelopes of Theorem 3.8 differ by a factor of order $\rho^{-1/2}$. (iii) Closing the gap between the sufficient amplitude condition and native checkpoint scales, which E3 places at more than six orders of magnitude on every audited checkpoint.

AI USE STATEMENT

The authors used an AI assistant for ideation and literature search, for checking mathematical arguments, for preparing and revising the LaTeX sources and restructuring the manuscript, for designing and implementing the numerical audits, and for processing the archived outputs. The authors ran the deterministic E1 calculations and the checkpoint evaluations themselves, and take full responsibility for all mathematical statements, numerical claims, and interpretations in the paper.

REPRODUCIBILITY STATEMENT

The supplementary archive separates three components. E1 includes 80 deterministic values and exhaustive enumeration through $m \leq 10$. E2 contains all 198 rows with 65,536 draws per row. E3 archives exact commit SHAs, revision-pinned WikiText-103 data, embedding radii, matrix-norm diagnostics, activation envelopes, attention-row maxima, per-layer hidden-state radii, directional budgets, state–update angles, and the forward-pass data underlying Table 2 and Figure 3. Appendix F gives the full protocol and model revisions.

REFERENCES

Noga Alon and Assaf Naor. Approximating the cut-norm via Grothendieck's inequality. *SIAM Journal on Computing*, 35(4):787–803, 2006. doi: 10.1137/S0097539704441629.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. arXiv:1607.06450, 2016. URL https://arxiv.org/abs/1607.06450.

Thomas Bachlechner, Bodhisattwa Prasad Majumder, Henry Mao, Gary Cottrell, and Julian McAuley. ReZero is all you need: Fast convergence at large depth. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 1352–1361, 2021. URL https://proceedings.mlr.press/v161/bachlechner21a.html.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002. URL https://www.jmlr.org/papers/v3/bartlett02a.html.

Valérie Castin, Pierre Ablin, and Gabriel Peyré. How smooth is attention? In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5817–5840, 2024. URL https://proceedings.mlr.press/v235/castin24a.html.

Soham De and Samuel L. Smith. Batch normalization biases residual blocks towards the identity function in deep networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19964–19975, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/e6b738eca0e6792ba8a9cbcba6c1881d-Abstract.html.

Benjamin L. Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5793–5831, 2022. URL https://proceedings.mlr.press/v162/edelman22a.html.

Seyed Morteza Emadi. Exact attention sensitivity and the geometry of transformer stability. arXiv:2602.18849, version 2, 17 August 2026, 2026. URL https://arxiv.org/abs/2602.18849.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 297–299, 2018. URL https://proceedings.mlr.press/v75/golowich18a.html.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The Lipschitz constant of self-attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5562–5571, 2021. URL https://proceedings.mlr.press/v139/kim21i.html.

Rafał Latała and Krzysztof Oleszkiewicz. On the best constant in the Khinchin–Kahane inequality. *Studia Mathematica*, 109(1):101–104, 1994. doi: 10.4064/sm-109-1-101-104.

Andreas Maurer. A vector-contraction inequality for Rademacher complexities. In *Algorithmic Learning Theory*, pp. 3–17, 2016. doi: 10.1007/978-3-319-46379-7_1. URL https://arxiv.org/abs/1605.00251.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurelien Lucchi. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27198–27211, 2022. doi: 10.52202/068431-1972. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/ae0cba715b60c4052359b3d52a2cff7f-Abstract-Conference.html.

Xianbiao Qi, Jianan Wang, Yihao Chen, Yukai Shi, and Lei Zhang. LipsFormer: Introducing Lipschitz continuity to vision transformers. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=cHf1DcCwcH3.

Mana Sakai and Masaaki Imaizumi. Spectrum-adaptive generalization bounds for trained deep transformers. arXiv:2605.07297, 2026. URL https://arxiv.org/abs/2605.07297.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *First Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2402.17762.

Jacob Trauger and Ambuj Tewari. Sequence length independent norm-based generalization bounds for transformers. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1405–1413, 2024. URL https://proceedings.mlr.press/v238/trauger24a.html.

Lan V. Truong. On rademacher complexity-based generalization bounds for deep learning. arXiv:2208.04284, version 5, 28 February 2025, 2025. URL https://arxiv.org/abs/2208.04284.

Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, and Furu Wei. DeepNet: Scaling transformers to 1,000 layers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10):6761–6774, 2024. doi: 10.1109/TPAMI.2024.3386927. URL https://doi.org/10.1109/TPAMI.2024.3386927.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NG7sS51zVF.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, volume 32, pp. 12360–12371, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/1e8a19426224ca89e83cef47f1e7f53b-Abstract.html.

A Proofs deferred from the main text

The proof ideas for the two central normalization results remain in the main text; the complete proofs below collect the remaining arguments that were deferred for readability.

A.1 Global extension of normalization

Lemma A.1 (Global extension of normalization). *For $c \geq 0$, the map $\overline{\mathcal{N}}_c(h) = h/\sqrt{\max\{\|h\|_\circ^2, c^2\} + \nu}$ agrees with $\mathcal{N}$ above radius c, has output RMS radius at most one, and is globally $(c^2+\nu)^{-1/2}$-Lipschitz.*

Proof. Write $\phi(r) = r/\sqrt{\max\{r^2, c^2\} + \nu}$, so that $\overline{\mathcal{N}}_c(h) = \phi(\|h\|_\circ)\, h/\|h\|_\circ$ for $h \neq 0$. Agreement above c and the bound $\phi \leq 1$ are immediate. The function ϕ is continuous, vanishes at zero, equals $b_c r$ for $r \leq c$ with $b_c = (c^2+\nu)^{-1/2}$, and for $r > c$ has $\phi'(r) = \nu(r^2+\nu)^{-3/2} \leq (r^2+\nu)^{-1/2} \leq b_c$; hence $|\phi(r) - \phi(s)| \leq b_c|r-s|$, and $\phi(r) \leq b_c r$ for every $r > 0$. Because $\overline{\mathcal{N}}_c$ is radial it preserves directions, so if two inputs have RMS radii r, s and angular cosine u,

$$\left\|\overline{\mathcal{N}}_c(h) - \overline{\mathcal{N}}_c(h')\right\|_\circ^2 = [\phi(r)-\phi(s)]^2 + 2\phi(r)\phi(s)(1-u) \leq b_c^2\big[(r-s)^2 + 2rs(1-u)\big] = b_c^2\,\|h-h'\|_\circ^2.$$

using the slope bound on the first term and $\phi(r)\phi(s) \leq b_c^2 rs$ on the second. Zero inputs follow by continuity. □

A.2 Proof of Proposition 4.1

Proof. For attention, each row of Z has Euclidean norm at most $\sqrt{d}$. Hence every value row in group μ has Euclidean norm at most $\sqrt{d}\, v_\mu$. A head output is a convex combination of visible value rows, so its Euclidean norm is bounded by the same quantity. Concatenating the $n_h = gJ$ heads therefore gives, at every position,

$$\|[U_1|\cdots|U_{n_h}]\|_2^2 \leq d\, g \sum_{\mu=1}^{J} v_\mu^2 = d\Sigma_V^2.$$

Multiplication by W_O and division by $\sqrt{d}$ yield (18).

For SwiGLU, fix a row z of Z and write $a = zA_g$, $b = zA_u$. Then

$$\|a\|_\infty \leq \beta_a, \qquad \|b\|_\infty \leq \beta_b,$$

while

$$\|a\|_2 \leq \sqrt{d}\, S_g, \qquad \|b\|_2 \leq \sqrt{d}\, S_u.$$

Since $|\operatorname{swish}(x)| \leq |x|$,

$$\|\operatorname{swish}(a) \odot b\|_2 \leq \|\operatorname{swish}(a)\|_\infty \|b\|_2 \leq \beta_a \sqrt{d}\, S_u.$$

Alternatively,

$$\|\operatorname{swish}(a) \odot b\|_2 \leq \|b\|_\infty \|\operatorname{swish}(a)\|_2 \leq \beta_b \sqrt{d}\, S_g.$$

Taking the better of the two bounds, multiplying by W_d, and converting Euclidean norm to RMS norm gives (19). □

A.3 Input sensitivities

Proposition A.2 (Concrete attention and SwiGLU input sensitivities). *Suppose $\|Z\|_{\infty,\circ}, \|Z'\|_{\infty,\circ} \leq 1$. For attention, in addition to the value/output caps of Proposition 4.1, suppose $\max_h \|A_{Q,h}\|_{\mathrm{op}} \leq S_Q$ and $\max_\mu \|A_{K,\mu}\|_{\mathrm{op}} \leq S_K$, and define $\beta_Q = \sqrt{n_h}\, S_Q$, $\beta_K = \sqrt{n_h}\, S_K$, and $\beta_V = \sqrt{n_h}\max_\mu v_\mu$. Then a valid input-sensitivity constant is*

$$\Lambda_{\mathrm{att}} = S_O\left[\Sigma_V + 2\sqrt{d_h}\,\beta_Q\beta_K\beta_V\right]. \tag{25}$$

For the SwiGLU branch, let $\kappa_{\mathrm{sw}} = 1 + e^{-1}$. Then

$$\Lambda_{\mathrm{ff}} := S_d(1+\kappa_{\mathrm{sw}})\min\{\beta_a S_u,\, \beta_b S_g\} = (1+\kappa_{\mathrm{sw}})M_{\mathrm{ff}} \tag{26}$$

is valid. Fixed orthogonal rotary maps and fixed scalar 1-Lipschitz score caps do not increase (25).

Proof. Write $\delta = \|Z - Z'\|_{\infty,\circ}$. For attention, a query or key head changes in Euclidean row norm by at most $\sqrt{d_h}\beta_Q\delta$ or $\sqrt{d_h}\beta_K\delta$, respectively, while every corresponding query/key row itself has the same bound without δ. Hence the scaled query–key score changes by at most $2\sqrt{d_h}\beta_Q\beta_K\delta$ in maximum norm. Lemma B.1 then bounds the change in an attention-probability row in ℓ_1 by the same quantity. Since a value row has Euclidean norm at most $\sqrt{d_h}\beta_V$, the resulting weight-change contribution to a head output is at most $2d_h\beta_Q\beta_K\beta_V\delta$. Concatenating the n_h heads, converting to RMS norm, and applying W_O gives the second term in (25). With attention weights fixed, the group-specific value changes give the sharper RMS bound $\Sigma_V\delta$, producing the first term. Orthogonal rotary maps preserve Euclidean norms, and a fixed scalar 1-Lipschitz score cap cannot increase the score difference.

For SwiGLU, fix a row and write $a = zA_g$, $a' = z'A_g$, $b = zA_u$, and $b' = z'A_u$. Then

$$\mathrm{swish}(a) \odot b - \mathrm{swish}(a') \odot b' = [\mathrm{swish}(a) - \mathrm{swish}(a')] \odot b + \mathrm{swish}(a') \odot (b - b').$$

The first term obeys both

$$\|[\mathrm{swish}(a) - \mathrm{swish}(a')] \odot b\|_2 \le \kappa_{\mathrm{sw}}\sqrt{d}\,\delta\, S_g\beta_b$$

and

$$\|[\mathrm{swish}(a) - \mathrm{swish}(a')] \odot b\|_2 \le \kappa_{\mathrm{sw}}\sqrt{d}\,\delta\,\beta_a S_u,$$

using, respectively, the ℓ_2/ℓ_∞ and ℓ_∞/ℓ_2 product bounds. Therefore it is at most $\kappa_{\mathrm{sw}}\sqrt{d}\,\delta\min\{S_g\beta_b, \beta_a S_u\}$. Similarly, because $|\mathrm{swish}(x)| \le |x|$, the second term is at most $\sqrt{d}\,\delta\min\{S_g\beta_b, \beta_a S_u\}$. After multiplication by W_d and division by $\sqrt{d}$, this gives (26) rowwise; taking the maximum over rows completes the proof. □

A.4 Proof of Theorem 4.3

Proof. By the normalization lemma,

$$\left\|\mathcal{N}(H^{s-1})\right\|_{\infty,\circ} \le 1.$$

Hence (20) gives, at every position t,

$$\left\|F_s(\mathcal{N}(H^{s-1}); W_s)_t\right\|_\circ \le M_s.$$

Using the residual recursion (17) and the reverse triangle inequality,

$$\begin{aligned}\|H_t^s\|_\circ &\ge \left\|H_t^{s-1}\right\|_\circ - \lambda_s \left\|F_s(\mathcal{N}(H^{s-1}); W_s)_t\right\|_\circ \\ &\ge \left\|H_t^{s-1}\right\|_\circ - \lambda_s M_s.\end{aligned}$$

Iterating from the embedding layer and using the embedding-floor assumption $a_0 > 0$,

$$\|H_t^s\|_\circ \ge a_0 - \sum_{u=1}^{s} \lambda_u M_u = a_0 - A_s = c_s.$$

Since $A_s \le A_D < a_0$, every c_s is strictly positive. The argument is uniform in parameters, inputs, and positions, so (21) follows. □

A.5 Proof of Proposition 4.4

Proof. Take $H^0 = a_0 > 0$ and constant branch maps $F_s(Z) \equiv -M_s$. They satisfy the magnitude assumptions with equality, and the recursion becomes $H^s = H^{s-1} - \lambda_s M_s$, hence $H^s = a_0 - A_s > 0$. □

A.6 Directional amplitude budget

Theorem A.3 (Directional amplitude budget implies a population radius floor)**.** *Fix a unit vector $w \in \mathbb{R}^d$. Suppose the embedding stream has the uniform directional floor*

$$a_w := \inf_\theta \inf_{X\in\mathcal{X}} \min_t \frac{|\langle H_t^0(X;\theta), w\rangle|}{\sqrt{d}} > 0, \tag{27}$$

and for every branch there is $\mu_s \geq 0$ such that

$$\sup_{W_s} \sup_{\|Z\|_{\infty,\circ} \leq 1} \max_t \frac{|\langle F_s(Z; W_s)_t, w\rangle|}{\sqrt{d}} \leq \mu_s. \tag{28}$$

Define $A_s^w := \sum_{u=1}^{s} \lambda_u \mu_u$ and assume $A_D^w < a_w$. Then $\|H_t^s(X;\theta)\|_\circ \geq a_w - A_s^w > 0$ uniformly. For the branches in Proposition 4.1, valid directional amplitudes are

$$\mu_{\text{att}} = \Sigma_V S_{O,w}, \qquad \mu_{\text{ff}} = S_{d,w} \min\{\beta_a S_u, \beta_b S_g\}, \tag{29}$$

where $S_{O,w} := \sup_{W_O} \|W_O w\|_2 \leq S_O$ and $S_{d,w} := \sup_{W_d} \|W_d w\|_2 \leq S_d$.

The refinement trades starting floor against cumulative motion. Since $|\langle h, w\rangle|/\sqrt{d} \leq \|h\|_\circ$ for unit w, the directional floor satisfies $a_w \leq a_* := \inf_\theta \inf_{X \in \mathcal{X}} \min_t \left\|H_t^0(X;\theta)\right\|_\circ$, the actual isotropic infimum. It need not satisfy $a_w \leq a_0$ for an arbitrary certified lower bound a_0 in Theorem 4.3: if every initial vector has RMS projection 2 along w, then $a_w = 2$, while $a_0 = 1$ remains a valid certificate. Taking $a_0 = a_*$ recovers the comparison. The directional budget can be much smaller than the isotropic one when the residual branches have weak projection along w. This is a structural refinement, not an empirically validated remedy; over a large token support a single embedding with near-zero projection onto w makes a_w small. The directional route is complementary to the isotropic certificate; failure of one route does not logically determine the other.

Proof. For unit w, $|\langle h, w\rangle| \leq \|h\|_2 = \sqrt{d}\, \|h\|_\circ$. Define

$$\gamma_t^s := \frac{\langle H_t^s, w\rangle}{\sqrt{d}}.$$

By (1), $\left\|\mathcal{N}(H^{s-1})\right\|_{\infty,\circ} \leq 1$, so (28) applies to the branch input. The residual recursion gives

$$\gamma_t^s = \gamma_t^{s-1} + \lambda_s \frac{\langle F_s(\mathcal{N}(H^{s-1}); W_s)_t, w\rangle}{\sqrt{d}}.$$

The scalar reverse triangle inequality therefore yields $|\gamma_t^s| \geq |\gamma_t^{s-1}| - \lambda_s \mu_s$. Iterating from (27) proves $|\gamma_t^s| \geq a_w - A_s^w > 0$, uniformly in parameters, inputs and positions.

For attention, the proof of Proposition 4.1 gives $\|[U_1|\cdots|U_{n_h}]_t\|_2 \leq \sqrt{d}\,\Sigma_V$. Hence

$$\frac{|\langle [U_1|\cdots|U_{n_h}]_t W_O, w\rangle|}{\sqrt{d}} \leq \Sigma_V \|W_O w\|_2 \leq \Sigma_V S_{O,w}.$$

For SwiGLU, the same proof bounds the pre-down-projection row by $\sqrt{d}\min\{\beta_a S_u, \beta_b S_g\}$, so projection onto w after W_d gives the second formula in (29). Finally, $S_{O,w} \leq S_O$ and $S_{d,w} \leq S_d$ for unit w, hence $\mu_s \leq M_s$. □

A.7 Checkable local-reference version

Corollary A.4 (A checkable local-reference version)**.** *Suppose $H_t^0 = E_{X_t,\cdot}$ and $\max_v \left\|E_v - E_v^0\right\|_\circ \leq e_E$. Let $a_E^0 := \min_{v \in \mathcal{V}_{\text{supp}}} \left\|E_v^0\right\|_\circ$. If $e_E < a_E^0$ and $\sum_{s=1}^{D} \lambda_s M_s < a_E^0 - e_E$, then $\|H_t^s(X;\theta)\|_\circ \geq a_E^0 - e_E - \sum_{u=1}^{s} \lambda_u M_u > 0$ for every population input, position, and stage.*

Remark A.5 (Full vocabulary, population support, and observed support). Three minima have different meanings. A full-vocabulary guarantee takes the minimum over every embedding row. Theorem 4.3 and Corollary A.4 need only the population support. A finite held-out sample supplies only an observed-support diagnostic. An observed token with a very small radius can falsify a proposed lower bound, but a positive sample minimum cannot certify the population infimum.

Proof. For every supported token v, the reverse triangle inequality gives

$$\|E_v\|_\circ \geq \left\|E_v^0\right\|_\circ - \left\|E_v - E_v^0\right\|_\circ \geq a_E^0 - e_E.$$

Thus Theorem 4.3 applies with $a_0 = a_E^0 - e_E$. □

A.8 PROPAGATION CONSEQUENCES

When stage $s-1$ has a certified floor c_{s-1}, the RMS normalizer is Lipschitz with coefficient $\omega_{s-1} := (c_{s-1}^2 + \nu)^{-1/2}$, giving the residual propagation multiplier

$$L_s := 1 + \lambda_s \Lambda_s \omega_{s-1}, \tag{30}$$

and the tail propagation product $\Pi_s := \prod_{u=s+1}^{D} L_u$ with $\Pi_D = 1$.

Corollary A.6 (Derived floor plus cumulative sensitivity controls propagation)**.** *Under Theorem 4.3, define* $B_s := \sum_{u=1}^{s} \lambda_u \Lambda_u$. *Then*

$$\Pi_s \le \exp\left\{ \frac{B_D - B_s}{\sqrt{c_D^2 + \nu}} \right\}, \qquad \Pi_0 \le \exp\left\{ \frac{B_D}{\sqrt{(a_0 - A_D)^2 + \nu}} \right\}. \tag{31}$$

If $A_D \le (1-\kappa) a_0$ *and* $B_D \le B$ *uniformly in* D, *then* $\Pi_0 \le \exp\{B/\sqrt{\kappa^2 a_0^2 + \nu}\}$, *a bound independent of depth.*

Proof. Using $\log(1+x) \le x$ in (30),

$$\log \Pi_s \le \sum_{u=s+1}^{D} \frac{\lambda_u \Lambda_u}{\sqrt{c_{u-1}^2 + \nu}}.$$

The sequence $c_u = a_0 - A_u$ is nonincreasing, so $c_{u-1} \ge c_D$. Hence

$$\log \Pi_s \le \frac{\sum_{u=s+1}^{D} \lambda_u \Lambda_u}{\sqrt{c_D^2 + \nu}} = \frac{B_D - B_s}{\sqrt{c_D^2 + \nu}}.$$

Exponentiating proves (31); the remaining statements are immediate. □

B THE ATTENTION MODEL AND THE SOFTMAX STEP

The width factors as $d = n_h d_h$, with $n_h = gJ$ heads arranged in J key/value groups of g query heads each; $\mu(h)$ denotes the group of head h, and $J = n_h$, $g = 1$ recovers standard multi-head attention. On an input $Z \in \mathbb{R}^{T \times d}$ with head maps $A_{Q,h}, A_{K,\mu}, A_{V,\mu} \in \mathbb{R}^{d \times d_h}$, head h forms the scaled scores $s^{(h)} = (ZA_{Q,h})(ZA_{K,\mu(h)})^\top / \sqrt{d_h}$, optionally after a fixed orthogonal rotary map on the query/key rows and a fixed scalar 1-Lipschitz score cap; a row softmax over the visible (masked) set gives the probability matrix $P^{(h)}$, and the head output is $U_h = P^{(h)}(ZA_{V,\mu(h)})$. The branch output is $F_{\mathrm{att}}(Z) = [U_1 | \cdots | U_{n_h}]\, W_O$ with $W_O \in \mathbb{R}^{d \times d}$.

Fix class-wide bounds

$$\max_h \|A_{Q,h}\|_{\mathrm{op}} \le S_Q, \quad \max_\mu \|A_{K,\mu}\|_{\mathrm{op}} \le S_K, \quad \|A_{V,\mu}\|_{\mathrm{op}} \le v_\mu, \quad \|W_O\|_{\mathrm{op}} \le S_O.$$

Define

$$\beta_Q = \sqrt{n_h}\, S_Q, \quad \beta_K = \sqrt{n_h}\, S_K, \quad \beta_V = \sqrt{n_h} \max_\mu v_\mu, \quad \Sigma_V = \left(g \sum_{\mu=1}^{J} v_\mu^2 \right)^{1/2}. \tag{32}$$

For $\|Z\|_{\infty,\circ} \le 1$, a query or key head has Euclidean row radius at most $\sqrt{d_h}\beta_Q$ or $\sqrt{d_h}\beta_K$; a value head has radius at most $\sqrt{d_h}\beta_V$. All quantities can depend on the sublayer index.

B.1 GATED FEEDFORWARD BRANCHES

For hidden width d_{ff} the branch is

$$F_{\mathrm{ff}}(Z) = [\mathrm{swish}(ZA_g) \odot (ZA_u)] W_d, \qquad \mathrm{swish}(x) = \frac{x}{1 + e^{-x}},$$

with $A_g, A_u \in \mathbb{R}^{d\times d_{\mathrm{ff}}}$ and $W_d \in \mathbb{R}^{d_{\mathrm{ff}}\times d}$. Fix spectral caps S_g, S_u, S_d and column caps $C_g \geq \max_j \|(A_g)_{\cdot j}\|_2, C_u \geq \max_j \|(A_u)_{\cdot j}\|_2$. One may always take $C_g = S_g, C_u = S_u$. Put

$$\beta_a = \sqrt{d}C_g, \qquad \beta_b = \sqrt{d}C_u, \qquad \kappa_{\mathrm{sw}} = 1 + e^{-1}.$$

This explicit choice bounds $\sup_x |\mathrm{swish}'(x)|$: with $\sigma(x) = (1+e^{-x})^{-1}$, $\mathrm{swish}'(x) = \sigma(x) + x\sigma(x)(1-\sigma(x))$ and $|x|\sigma(x)(1-\sigma(x)) \leq |x|e^{-|x|} \leq e^{-1}$. We do not need the optimal scalar Lipschitz constant.

Lemma B.1 (Softmax on a fixed visible set)**.** *For any two finite score vectors on the same nonempty visible set,*

$$\|\mathrm{softmax}(a) - \mathrm{softmax}(b)\|_1 \leq \|a-b\|_\infty .$$

Proof. At a probability vector p, the derivative applied to u is $p \odot (u - \langle p, u\rangle \mathbf{1})$. Its ℓ_1 norm is the mean absolute deviation of a random variable taking values u_j with probabilities p_j. If $\|u\|_\infty \leq 1$, this is at most its standard deviation, which is at most one. Integrate the derivative along the segment between the score vectors. □

C Proofs for normalization and sharpness

C.1 Proof of Corollary 3.3

Proof. Let $\mathcal{F}_N = \{X \mapsto \langle w, \mathcal{N}(h_\theta(X))\rangle : \theta \in \Theta, \|w\|_2 \leq B\}$. By Euclidean duality exactly as in (3),

$$\widehat{\mathfrak{R}}_S(\mathcal{F}_N) = B\mathfrak{R}_m^{\mathrm{vec}}(\mathcal{N} \circ \mathcal{H}_S).$$

For each i, set $\phi_i(t) = \ell(Y_i, t) - \ell(Y_i, 0)$. Then $\phi_i(0) = 0$ and ϕ_i is L_ℓ-Lipschitz. Scalar Rademacher contraction gives

$$\widehat{\mathfrak{R}}_S(\ell \circ \mathcal{F}_N) = \widehat{\mathfrak{R}}_S(\phi \circ \mathcal{F}_N) \leq L_\ell B\mathfrak{R}_m^{\mathrm{vec}}(\mathcal{N} \circ \mathcal{H}_S),$$

where subtracting $\ell(Y_i, 0)$ changes the signed sum by a term independent of the hypothesis and hence has zero expectation over the signs. The standard empirical Rademacher generalization inequality for a $[0, U]$-valued loss class then yields (10); see, e.g., Bartlett & Mendelson (2002). Theorem 3.1 with gain one gives the profile bound

$$2L_\ell B\{b_c\mathfrak{R}_m^{\mathrm{vec}}(\mathcal{H}_S) + M\rho b_c^2\}. \tag{33}$$

Under the annulus condition, the definition of $C^\star_{m,\nu}(\rho)$ and (8) give

$$\mathfrak{R}_m^{\mathrm{vec}}(\mathcal{N} \circ \mathcal{H}_S) \leq \frac{1}{\sqrt{1+\nu}}\left(1 + \frac{\sqrt{2}(1+2\rho)\rho\sqrt{m}}{\sqrt{1+\nu}}\right)\mathfrak{R}_m^{\mathrm{vec}}(\mathcal{H}_S), \tag{34}$$

completing the proof. □

C.2 Proof of Theorem 3.1

Proof. Let $\widetilde{r}_i = \max\{r_i^0, c\}$ and $q_i^0 = (\widetilde{r}_i^2 + \nu)^{-1/2}$. Projection onto $[c, \infty)$ cannot increase the distance from an admissible radius. Moreover

$$\left|\frac{d}{dr}(r^2+\nu)^{-1/2}\right| = \frac{r}{(r^2+\nu)^{3/2}} \leq b_c^2 \quad (r \geq c).$$

Consequently $0 \leq q_i^0 \leq b_c$ and $|q_i(\theta) - q_i^0| \leq \rho b_c^2$. Decompose $q_i(\theta)h_i(\theta) = q_i^0 h_i(\theta) + [q_i(\theta) - q_i^0]h_i(\theta)$. Fixed-coordinate scalar contraction in the dual projection class bounds the first signed-sum complexity by $b_c\mathfrak{R}_m^{\mathrm{vec}}(\mathcal{H})$. The triangle inequality bounds the second by $M\rho b_c^2$. The gain contributes at most G. □

For clarity, the fixed-coordinate contraction used here is scalar contraction on the class of projections $\{(\langle u, h_i(\theta)\rangle)_i : \|u\|_2 \leq 1, \theta \in \Theta\}$. Its scalar complexity is exactly (2). Coefficients q_i^0/b_c are fixed and lie in $[0, 1]$, so they cannot increase this complexity. The reference profile may depend on the fixed evaluation inputs; it must be common across hypotheses. This deterministic statement alone supplies no high-probability claim for a data-selected class.

C.3 Linear normalization, and the general inverse-radius bound

Theorem 3.2 concerns normalizers that divide by a θ-dependent radius. The obstruction is specific to that dependence rather than to normalization as such, and the following records the contrast.

Proposition C.1 (A fixed linear normalization contracts). *Let $N_\gamma(h) = hP\,\mathrm{diag}(\gamma)$ for a fixed matrix P and a gain with $\|\gamma\|_\infty \le G$, with no additive bias. Then for every class $\mathcal{H}$,*

$$\mathfrak{R}_m^{\mathrm{vec}}(N_\gamma \circ \mathcal{H}) \;\le\; \|P\,\mathrm{diag}(\gamma)\|_{\mathrm{op}}\; \mathfrak{R}_m^{\mathrm{vec}}(\mathcal{H}) \;\le\; G\;\|P\|_{\mathrm{op}}\; \mathfrak{R}_m^{\mathrm{vec}}(\mathcal{H}), \tag{35}$$

and the first inequality is attained. For the zero-bias CenterNorm map $P = \frac{d}{d-1}(I - \mathbf{1}\mathbf{1}^\top/d)$ with $d \ge 2$ one has $\|P\|_{\mathrm{op}} = \frac{d}{d-1}$, so the constant is at most $G\,d/(d-1)$.

Proof. P and $\mathrm{diag}(\gamma)$ do not depend on the evaluation index, so at each fixed θ we may write $\sum_i \varepsilon_i N_\gamma(h_i(\theta)) = \big(\sum_i \varepsilon_i h_i(\theta)\big)P\,\mathrm{diag}(\gamma)$. Taking Euclidean norms and then the supremum over θ gives (35); equality holds for a class supported on a top left singular direction of $P\,\mathrm{diag}(\gamma)$. If the gain varies with θ, bound $\|\mathrm{diag}(\gamma(\theta))\|_{\mathrm{op}} \le G$ before taking the supremum. Finally $I - \mathbf{1}\mathbf{1}^\top/d$ is an orthogonal projection, with eigenvalues one and zero, so $\|P\|_{\mathrm{op}} = d/(d-1)$. □

No radius appears in (35), hence no class-dependent normalizing factor, and Theorem 3.2 has nothing to act on. The distinction the theorem turns on is therefore division by a norm of the input, not centering or rescaling by fixed quantities. This is an observation about the normalizer alone: it says nothing about contraction through the nonlinear attention and gated branches of any architecture that uses such a map.

We next record the general bound behind the diagnostic quantity used in Remark C.8. It is stated for the deterministic evaluation setting of Section 3.

Lemma C.2 (Bounded scalar products). *Let $F, G \subset \mathbb{R}^m$ satisfy $|f_i| \le A$ and $|g_i| \le B$ for all $f \in F$, $g \in G$ and i, with $A, B > 0$. For the coordinatewise product class FG,*

$$\widehat{\mathfrak{R}}_m(FG) \;\le\; 2B\,\widehat{\mathfrak{R}}_m(F) + 2A\,\widehat{\mathfrak{R}}_m(G). \tag{36}$$

Proof. The claim is immediate if $A = 0$ or $B = 0$. Otherwise write, pointwise,

$$f_i g_i = \frac{AB}{4}\Big[\Big(\frac{f_i}{A} + \frac{g_i}{B}\Big)^2 - \Big(\frac{f_i}{A} - \frac{g_i}{B}\Big)^2\Big].$$

Each bracketed argument lies in $[-2, 2]$, where $t \mapsto t^2$ is 4-Lipschitz. Scalar Rademacher contraction in the scalar convention of Section 3, which carries no absolute value and therefore has contraction constant one, gives $\widehat{\mathfrak{R}}_m\big((F/A \pm G/B)^2\big) \le 4\,\widehat{\mathfrak{R}}_m(F/A \pm G/B)$. Subadditivity of the supremum over a sum class and $\widehat{\mathfrak{R}}_m(-G) = \widehat{\mathfrak{R}}_m(G)$ then yield $\widehat{\mathfrak{R}}_m(FG) \le AB\big[2\widehat{\mathfrak{R}}_m(F)/A + 2\widehat{\mathfrak{R}}_m(G)/B\big]$, which is (36). □

Proposition C.3 (Inverse-radius complexity). *Suppose $\|h_i(\theta)\|_\circ \ge c$ and $\|h_i(\theta)\|_2 \le M$ for all i and θ, and $\|\gamma(\theta)\|_\infty \le G$. With $b_c = (c^2 + \nu)^{-1/2}$ and $\mathcal{Q} = \{((\|h_i(\theta)\|_\circ^2 + \nu)^{-1/2})_{i=1}^m : \theta \in \Theta\}$,*

$$\mathfrak{R}_m^{\mathrm{vec}}(\mathrm{RMS}_\gamma \circ \mathcal{H}) \;\le\; 2G\Big[b_c\,\mathfrak{R}_m^{\mathrm{vec}}(\mathcal{H}) + M\,\widehat{\mathfrak{R}}_m(\mathcal{Q})\Big]. \tag{37}$$

Proof. Remove the gain by its operator norm. Let $\mathcal{P} = \{(\langle u, h_i(\theta)\rangle)_{i=1}^m : \|u\|_2 \le 1,\ \theta \in \Theta\}$; by Euclidean duality $\widehat{\mathfrak{R}}_m(\mathcal{P}) = \mathfrak{R}_m^{\mathrm{vec}}(\mathcal{H})$, and every coordinate of $\mathcal{P}$ has magnitude at most M. Writing $q_i(\theta)$ for the inverse radii, we have $\mathcal{N}(h_i(\theta)) = q_i(\theta)h_i(\theta)$, so duality again gives $\mathfrak{R}_m^{\mathrm{vec}}(\mathcal{N} \circ \mathcal{H}) \le \widehat{\mathfrak{R}}_m(\mathcal{P}\mathcal{Q})$, where the product class is enlarged by letting its two scalar factors vary independently. Every coordinate of $\mathcal{Q}$ lies in $[0, b_c]$, so Lemma C.2 with $A = M$ and $B = b_c$ completes the proof. □

Two remarks on how to read (37). It is an upper bound carrying $\widehat{\mathfrak{R}}_m(\mathcal{Q})$, not a characterization: it does not establish that this quantity decides when contraction is possible, and Remark C.8 exhibits a class where $\widehat{\mathfrak{R}}_m(\mathcal{Q}) > 0$ while the distortion stays below one. And when every sample radius is

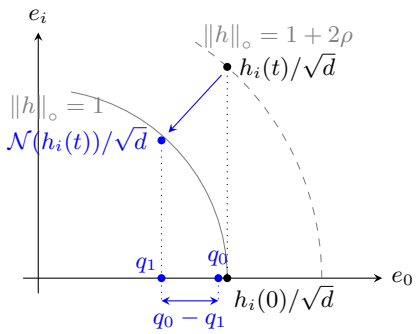

Figure 4: Mechanism of the lower-bound construction (9), drawn in RMS units in the (e_0, e_i) plane. Before normalization both class members share the e_0 coordinate (dotted vertical line), so this coordinate contributes the cancelling sum $\sum_i \varepsilon_i$. Normalization divides by a radius the hypothesis controls, moving the e_0 coordinate to q_0 or q_1. Choosing $a_i = 0$ on positive signs and $a_i = t$ on negative signs converts the gap $q_0 - q_1$ (double arrow) into a coherent e_0 drift of size $m(q_0 - q_1)/2$ in expectation, while the orthogonal e_i components keep the *raw* supremum of order $\sqrt{m}$.

constant across the hypothesis class, $\widehat{\mathfrak{R}}_m(\mathcal{Q}) = 0$ and (37) reduces to a contraction with constant $2Gb_c$, which Theorem 3.1 sharpens to Gb_c by treating the fixed coefficients directly. The difference reflects a real source of slack: Proposition C.3 enlarges the product class by allowing inverse radii and projected hidden states to vary independently, thereby discarding their phase/alignment within a common hypothesis. The profile theorem preserves that coupling through a fixed reference profile and can therefore be strictly sharper.

C.4 A SHARP MOMENT INEQUALITY

Lemma C.4 (Sharp Hilbert-space first/second moment comparison)**.** *For deterministic vectors $v_1, \dots, v_m$ in a Hilbert space,*

$$\frac{1}{\sqrt{2}} \left(\sum_i \|v_i\|_2^2 \right)^{1/2} \le \mathbb{E}_\varepsilon \left\| \sum_i \varepsilon_i v_i \right\|_2 \le \left(\sum_i \|v_i\|_2^2 \right)^{1/2}.$$

The lower constant $1/\sqrt{2}$ is optimal.

Proof. The upper bound is Jensen's inequality. The lower bound is the sharp Khintchine–Kahane comparison of Latała & Oleszkiewicz (1994), applied to the Hilbert norm and using $(\mathbb{E}\| \sum_i \epsilon_i v_i\|_2^2)^{1/2} = (\sum_i \|v_i\|_2^2)^{1/2}$. Sharpness already occurs in one dimension with $m = 2$ and $v_1 = v_2 \neq 0$: the first moment equals $\|v_1\|_2$ while the second-moment scale is $\sqrt{2}\|v_1\|_2$. □

C.5 PROOF OF THEOREM 3.2

One valid choice of constants is

$$C_{\nu,\rho_0} = \frac{\sqrt{2}(1 + 2\rho_0)}{\sqrt{1+\nu}}, \qquad c_{\nu,\rho_0} = \frac{\sqrt{1+\nu}}{(1+2\rho_0)[1+\nu+4\rho_0(1+\rho_0)]^{3/2}}. \tag{38}$$

Proof. Fix any class in (6). For a fixed member of the class, $\sum_i \|h_i\|_2^2 \ge md$, and Lemma C.4 yields $\mathfrak{R}_m^{\mathrm{vec}}(\mathcal{H}) \ge \sqrt{d/(2m)}$. The profile inequality at reference radius $1 + \rho$, radius floor one, and Euclidean envelope $\sqrt{d}(1 + 2\rho)$ gives

$$\frac{\sqrt{1+\nu}\,\mathfrak{R}_m^{\mathrm{vec}}(\mathcal{N} \circ \mathcal{H})}{\mathfrak{R}_m^{\mathrm{vec}}(\mathcal{H})} \le 1 + \frac{\sqrt{2}(1+2\rho)\rho\sqrt{m}}{\sqrt{1+\nu}}.$$

This proves the upper bound.

For the binary class (9), orthogonality gives

$$\left\| \sum_i \varepsilon_i h_i(a) \right\|_2^2 = d \left(S^2 + \sum_i a_i^2 \right).$$

The unnormalized supremum is attained by taking every $a_i = t$, which gives

$$\mathfrak{R}_m^{\mathrm{vec}}(\mathcal{H}_{m,t}) = \frac{\sqrt{d}}{m} \mathbb{E}\sqrt{S^2 + t^2 m} \le \sqrt{\frac{d(1+t^2)}{m}}; \tag{39}$$

Jensen's inequality uses $\mathbb{E}S^2 = m$. For each sign vector choose $a_i = 0$ if $\varepsilon_i = 1$ and $a_i = t$ otherwise. The normalized sum has e_0 projection $\sqrt{d}(q_0 n_+ - q_1 n_-)$, where $q_0 = (1+\nu)^{-1/2}$ and $q_1 = (1+t^2+\nu)^{-1/2}$. For every sign vector the Euclidean norm is at least its signed e_0 coordinate. Taking expectations first and using $\mathbb{E}n_+ = \mathbb{E}n_- = m/2$ therefore gives the positive lower bound $m\sqrt{d}(q_0 - q_1)/2$ for the expected supremum norm. Hence the normalized-to-raw ratio in (6) is at least

$$\frac{\sqrt{m}}{2\sqrt{1+t^2}} \left(1 - \sqrt{\frac{1+\nu}{1+t^2+\nu}} \right). \tag{40}$$

Take $t^2 = 4\rho(1+\rho)$, so $\sqrt{1+t^2} = 1+2\rho$. Writing $A = 1+\nu$ and $T_0 = 4\rho_0(1+\rho_0)$, integration gives

$$1 - \sqrt{\frac{A}{A+t^2}} = \int_0^{t^2} \frac{\sqrt{A}}{2(A+u)^{3/2}}\, du \ge \frac{t^2\sqrt{A}}{2(A+T_0)^{3/2}}.$$

Use $t^2 \ge 4\rho$ and $\sqrt{1+t^2} \le 1+2\rho_0$ in (40) to obtain the stated lower constant. A singleton class with every radius one gives distortion exactly one for all m. This also proves the result when $\rho = 0$. Finally, $\max\{1, cx\} \ge \min\{1, c\}(1+x)/2$ for $x \ge 0$, which gives the claimed order. The uniform boundedness equivalence follows directly from the two inequalities. □

C.6 Decoder realization and LayerNorm

Proposition C.5 (Realization by learned token embeddings)**.** *For $V \ge \max\{m, 2\}$, the class in (9) is exactly the hidden-state evaluation class of a length-one decoder subclass on m distinct input tokens, with embedding RMS row budget $\sqrt{1+t^2}$, gain one, and all residual branches set to zero. Every hidden-state radius on the support of these tokens is at least one.*

Proof. Restrict embedding row i to the values $\sqrt{d}(e_0 + a_i e_i)$, where $a_i \in \{0, t\}$, and fix any unused rows to zero. Each residual layer is the identity when its branch weights are zero, so the hidden states remain these embedding rows at every depth. All upper bounds on branch matrix norms admit the zero matrices. □

The realization allows d and V to grow with m and constrains embedding rows in RMS units. Their Euclidean budget is consequently $\sqrt{d}\sqrt{1+t^2}$. It does not establish the same obstruction for fixed embeddings, a dimension-independent Euclidean row budget, or the capped loss of a complete decoder. In particular, the realization rules out a contraction asserted uniformly for every such decoder subclass. It does not prove failure for the entire class defined only by norm upper bounds: the ratio of two Rademacher complexities is not monotone when the class is enlarged.

Corollary C.6 (Mean-zero LayerNorm realization)**.** *The construction in (9) applies to gain-one LayerNorm with regularizer ν when $e_0, \dots, e_m$ are chosen in the mean-zero subspace of $\mathbb{R}^{m+2}$ and the RMS scaling uses $d = m+2$.*

Proof. The subspace has dimension $m+1$. All hidden vectors have zero coordinate mean, so centering leaves them unchanged. LayerNorm and RMS normalization therefore coincide on every constructed vector. □

C.7 FIXED-EMBEDDING AND GROUPED CONSTRUCTIONS

Proof of Proposition 3.4. Take fixed embeddings

$$E_i = \sqrt{d}(e_0 + te_i), \qquad i = 1, \dots, m.$$

Every embedding has RMS radius $\sqrt{1+t^2} = 1 + 2\rho$. Use one attention head, set $A_Q = A_K = 0$ and $A_V = I_d$. For a length-one input the masked attention probability is exactly one, so the attention branch on a normalized row Z is simply ZW_O.

For $w = (w_1, \dots, w_m) \in \{0, -s\}^m$ set

$$W_O(w) = \mathrm{diag}(0, w_1, \dots, w_m).$$

Since

$$\mathcal{N}(E_i) = \frac{\sqrt{d}}{s}(e_0 + te_i),$$

the post-residual state is

$$H_i^1(w) = E_i + \mathcal{N}(E_i)W_O(w) = \sqrt{d}\left(e_0 + t\left(1 + \frac{w_i}{s}\right)e_i\right).$$

Thus the coefficient of e_i ranges independently over $\{0, t\}$, giving exactly (9). Finally $\|W_O(w)\|_{\mathrm{op}} \le s$. □

C.8 PROOF OF PROPOSITION 3.7 AND THE GENERAL-GROUP BOUND

The general-group formula, stated separately from the balanced specialization in the main text, is as follows. For arbitrary positive group sizes $n_1, \dots, n_k$ summing to m,

$$C^\star_{m,d,\nu}(\rho) \ge \max\left\{1,\ \frac{1 - a_{\rho,\nu}}{2\sqrt{m^{-1} + t^2 \sum_{j=1}^k n_j^2/m^2}}\right\}. \tag{41}$$

For equal groups $n_j = m/k$, the denominator becomes $2\sqrt{m^{-1} + t^2/k}$, recovering the formula used in E2.

Proof of Proposition 3.7. Choose orthonormal $e_0, \dots, e_k$ in $\mathbb{R}^d$ and use

$$h_i(a) = \sqrt{d}\,(e_0 + a_i e_{g(i)}), \qquad a_i \in \{0, t\},$$

where $g(i)$ is the group of i. Let $S = \sum_i \varepsilon_i$ and $U_j = \sum_{i:g(i)=j} \varepsilon_i a_i$. Since $|U_j| \le t n_j$ and $\mathbb{E}S^2 = m$,

$$\mathfrak{R}_m^{\mathrm{vec}}(\mathcal{H}) \le \frac{\sqrt{d}}{m}\mathbb{E}\sqrt{S^2 + t^2\sum_j n_j^2} \le \sqrt{d}\sqrt{m^{-1} + t^2 \sum_j n_j^2/m^2}.$$

Choose $a_i = 0$ on positive signs and $a_i = t$ on negative signs. With $q_0 = (1+\nu)^{-1/2}$ and $q_1 = (1 + t^2 + \nu)^{-1/2}$, the expected normalized signed sum has e_0 projection $m\sqrt{d}(q_0 - q_1)/2$. The expected supremum norm is at least this quantity. Dividing gives the general formula (41). The separate lower bound one comes from a singleton class of radius one. Finally,

$$\sum_j n_j^2 \le m\lceil m/k \rceil \le 2m^2/k,$$

which proves the balanced specialization (14).

For the rate statement, put $A = 1 + \nu$ and $T_0 = 4\rho_0(1 + \rho_0)$. Then

$$1 - \sqrt{\frac{A}{A + t^2}} = \int_0^{t^2} \frac{\sqrt{A}}{2(A+u)^{3/2}}\,du \ge \frac{t^2\sqrt{A}}{2(A+T_0)^{3/2}} \gtrsim_{\nu,\rho_0} \rho,$$

while $t^2 \le 4(1+\rho_0)\rho$. Hence the explicit term is bounded below by a constant multiple of

$$\frac{\rho}{\sqrt{m^{-1} + \rho/k}} \gtrsim \min\{\rho\sqrt{m}, \sqrt{\rho k}\}.$$

□

C.9 AN IMPROVED FINITE-SAMPLE LOWER BOUND

For the original growing-dimension construction, let $S_m = \sum_i \varepsilon_i$, $t^2 = 4\rho(1+\rho)$, and $a = a_{\rho,\nu}$ in Proposition 3.7. Its unnormalized expected supremum, after canceling the common $\sqrt{d}$ factor, is exactly $\mathbb{E}\sqrt{S_m^2 + t^2 m}$.

Proposition C.7 (Mirrored assignment)**.** *The distortion of the original constructed family is at least*

$$\frac{\frac{m}{2}(1-a) + \frac{1+a}{2}\mathbb{E}|S_m|}{\mathbb{E}\sqrt{S_m^2 + t^2 m}}. \tag{42}$$

At $\rho = 0$ this expression is one.

Proof. The assignment $a_i = 0$ for positive signs and $a_i = t$ for negative signs produces a shared-direction projection, with the $\sqrt{d}$ factor omitted, of

$$\frac{m}{2}(q_0 - q_1) + \frac{q_0 + q_1}{2} S_m.$$

The mirrored assignment produces $-m(q_0 - q_1)/2 + (q_0 + q_1)S_m/2$. The larger absolute projection is $m(q_0 - q_1)/2 + (q_0 + q_1)|S_m|/2$. Take expectations, multiply by $\sqrt{1+\nu}$, and divide by the exact raw supremum. At $t = 0$, both numerator and denominator equal $\mathbb{E}|S_m| > 0$. □

This improves the finite-sample constant; it does not change the order in the main theorem. It also gives a useful analytic curve near the shrinking-annulus threshold.

C.10 WHY VARIATION ALONE DOES NOT FORCE LARGE DISTORTION

Remark C.8 (Radius variation need not cause large distortion). For this remark only, let $\mathcal{Q} = \{((\|h_i(\theta)\|_\circ^2 + \nu)^{-1/2})_{i=1}^m : \theta \in \Theta\}$. Let $d \geq m$ and $h_i(a) = \sqrt{d}\, a_i e_i$, with independent $a_i \in [1, 2]$. Then

$$\mathfrak{R}_m^{\mathrm{vec}}(\mathcal{H}) = 2\sqrt{d/m}, \qquad \mathfrak{R}_m^{\mathrm{vec}}(\mathcal{N} \circ \mathcal{H}) = \frac{2}{\sqrt{4+\nu}}\sqrt{d/m},$$

whereas $\widehat{\mathfrak{R}}_m(\mathcal{Q}) = \frac{1}{2}[(1+\nu)^{-1/2} - (4+\nu)^{-1/2}] > 0$. Indeed the norm is maximized by $a_i = 2$, and the scalar inverse-radius supremum selects its larger value for positive signs and its smaller value for negative signs. Hence a nonvanishing inverse-radius complexity does not itself force a large contraction factor.

D A CONCENTRATION-DEPENDENT SOFTMAX CERTIFICATE

The attention sensitivity certificate contains a value path $S_O \Sigma_V$ and a weight path $2S_O\sqrt{d_h}\beta_Q\beta_K\beta_V$. The pointwise $\ell_\infty \to \ell_1$ norm of the softmax Jacobian is an instance of classical matrix-norm geometry (Alon & Naor, 2006) and was recently stated explicitly for attention sensitivity by Emadi (2026). The remaining issue for a Lipschitz bound is different: the Jacobian must be controlled along the whole interpolation segment between two score vectors. Recent bounded-domain attention bounds therefore revert to the generic inequality $\theta \leq 1$ when integrating along that path (Emadi, 2026). We show that endpoint concentration survives a bounded score perturbation quantitatively, which retains a sub-one coefficient throughout the segment whenever the surviving mass stays above one half. The checkpoint audit in §5 evaluates the uniformity requirement of that refinement.

Lemma D.1 (Softmax Jacobian norm; recalled)**.** *Let p be a probability vector on $[n]$ and $J(p) = \mathrm{diag}(p) - p^\top p$. Then*

$$\|J(p)\|_{\infty\to 1} \;=\; \max_{S\subseteq[n]} 4\,p(S)\big(1-p(S)\big), \qquad p(S) := \sum_{k\in S} p_k. \tag{43}$$

The identity is included for completeness; see also Emadi (2026). It can also be viewed as a specialization of classical $\ell_\infty \to \ell_1$ matrix-norm geometry (Alon & Naor, 2006).

Proof. For $\|v\|_\infty \le 1$ we have $\|vJ(p)\|_1 = \sum_k p_k |v_k - \langle p, v\rangle|$, a sum of absolute values of affine functions of v with nonnegative weights, hence convex in v. Its maximum over the cube $[-1,1]^n$ is therefore attained at a vertex. Let $v \in \{\pm 1\}^n$, put $S = \{k : v_k = 1\}$ and $s = p(S)$. Then $\langle p, v\rangle = 2s - 1$, so $|v_k - \langle p, v\rangle|$ equals $2(1-s)$ for $k \in S$ and $2s$ for $k \notin S$, giving $\|vJ(p)\|_1 = s \cdot 2(1-s) + (1-s) \cdot 2s = 4s(1-s)$. Maximizing over vertices is maximizing over subsets. □

Corollary D.2 (Concentrated rows). *If $M := \max_j p_j \ge \frac{1}{2}$ then $\|J(p)\|_{\infty\to 1} \le 4M(1-M)$, and the bound is attained at $n = 2$, $p = (M, 1-M)$.*

Proof. Any S either contains a maximizing index, in which case $p(S) \ge M$, or does not, in which case $p(S) \le 1 - M$. The map $s \mapsto 4s(1-s)$ is symmetric about $s = \frac{1}{2}$ and increasing on $[0, \frac{1}{2}]$, and both M and $1 - M$ lie at distance $M - \frac{1}{2}$ from $\frac{1}{2}$, so (43) is at most $4M(1-M)$. □

A Lipschitz statement requires the bound along the whole segment between two score vectors, not only at an endpoint, so concentration must survive the perturbation.

Lemma D.3 (Concentration is stable under score perturbation). *Let a, b satisfy $\|a - b\|_\infty \le \eta$ and put $M = \max_j \mathrm{softmax}(a)_j$. Then for every $t \in [0,1]$,*

$$\max_j \mathrm{softmax}\big(a + t(b-a)\big)_j \;\ge\; M_\eta := \frac{M}{M + e^{2\eta}(1-M)}. \tag{44}$$

Proof. Let $j^\star$ maximize $\mathrm{softmax}(a)$ and write $\delta = t(b-a)$, so $\|\delta\|_\infty \le \eta$. Then

$$\mathrm{softmax}(a+\delta)_{j^\star} = \Big(1 + \sum_{k \ne j^\star} e^{(a_k+\delta_k)-(a_{j^\star}+\delta_{j^\star})}\Big)^{-1} \;\ge\; \Big(1 + e^{2\eta} \sum_{k \ne j^\star} e^{a_k - a_{j^\star}}\Big)^{-1},$$

and $\sum_{k \ne j^\star} e^{a_k - a_{j^\star}} = (1-M)/M$ by definition of softmax. Rearranging gives (44). □

Proposition D.4 (Segment-stable concentration-dependent softmax certificate). *Let a, b be score vectors on the same nonempty visible set with $\|a - b\|_\infty \le \eta$, and let $M = \max_j \mathrm{softmax}(a)_j$. With M_η as in (44),*

$$\|\mathrm{softmax}(a) - \mathrm{softmax}(b)\|_1 \le C(M,\eta)\ \|a - b\|_\infty,$$
$$C(M,\eta) := \begin{cases} 4M_\eta(1-M_\eta), & M_\eta \ge \frac{1}{2}, \\ 1, & M_\eta < \frac{1}{2}. \end{cases} \tag{45}$$

The case split is necessary: $4m(1-m)$ is below one for $m < \frac{1}{2}$ as well, but Corollary D.2 certifies the improvement only above one half. Consequently the weight path of Λ_{att} may be multiplied by $C(M,\eta)$.

Proof. Integrate the derivative along the segment from a to b, bounding $\|J(p_t)\|_{\infty\to 1}$ at each t by Corollary D.2 applied to $p_t = \mathrm{softmax}(a + t(b-a))$, whose maximum entry is at least M_η by Lemma D.3; the map $m \mapsto 4m(1-m)$ is decreasing on $[\frac{1}{2}, 1]$. The entry 1 is Lemma B.1. □

Corollary D.5 (Per-stage attention certificate). *Fix a residual stage ℓ. Suppose every attention row arising at that stage in the class satisfies $\max_j P_h^{(\ell)}(i,j) \ge M_\star^{(\ell)}$ over every head, query position, and population input, and every score pair entering the Lipschitz comparison at that stage satisfies $\|a - b\|_\infty \le \eta_\star^{(\ell)}$. Then*

$$\Lambda_{\mathrm{att},\ell}^{\mathrm{conc}} = S_O^{(\ell)} \Big[\Sigma_V^{(\ell)} + 2\sqrt{d_h}\,\beta_Q^{(\ell)}\beta_K^{(\ell)}\beta_V^{(\ell)} C(M_\star^{(\ell)}, \eta_\star^{(\ell)})\Big] \tag{46}$$

is a valid stage-ℓ attention Lipschitz certificate. Substituting the corresponding stagewise branch certificates into (30) yields

$$\Pi_0 \le \prod_{\ell=1}^{D} (1 + \lambda_\ell \Lambda_\ell^{\mathrm{conc}} \omega_{\ell-1}). \tag{47}$$

Thus a sub-one coefficient at one stage improves the product even if other stages use the generic softmax constant one.

Proof. Apply Proposition D.4 separately at stage ℓ and substitute the resulting coefficient into the attention sensitivity bound. The residual propagation product already factors stage by stage, so no common concentration floor across distinct stages is required. □

The distinction is between a pairwise fact and a uniform stagewise statement. Proposition D.4 uses the concentration of one endpoint plus a score-distance bound to control the full interpolation path. Corollary D.5 is conditional on the specified comparison domain: both the concentration floor and the score-distance bound must hold for every relevant input and every model pair in that stage's class (or, if the class is explicitly local, throughout that local parameter neighborhood). It is not a global Lipschitz statement over all normalized inputs unless those hypotheses are verified globally. Sampled minima can falsify an ambitious value of $M_\star^{(\ell)}$, but cannot by themselves certify a population lower bound or a class-wide score-distance bound.

Remark D.6 (Pairwise gain versus uniformity). At $\eta = 0.05$ the segment-stable pairwise coefficient is 0.390 at $M = 0.9$, 0.0437 at $M = 0.99$, and 0.00442 at $M = 0.999$, compared with the generic constant one. Thus concentration can materially sharpen sensitivity when it persists along the segment. Corollary D.5 asks for the least concentrated row within each stage of the specified class. Across 3,949,568 sampled rows from five pretrained checkpoints, the checkpointwise minima of $M = \max_j p_j$ are 0.0083–0.0147. Moreover, the largest layerwise minimum over all audited model–layer pairs is only 0.121, so at $\eta = 0.05$ every sampled stage falls back to the generic coefficient one even though median rows are substantially more concentrated.

E Exact-construction protocol and provenance

The E1 component of §5 is a deterministic finite-sum evaluation, with all 80 values archived together with exhaustive small-m checks.

E.1 E1: exact construction

We evaluate $m \in \{8, 32, 128, 512, 2048\}$, $\nu \in \{0.1, 10^{-3}\}$, fixed $\rho \in \{0, 0.025, 0.1, (\sqrt{2}-1)/2\}$, and shrinking schedules $\rho = a/\sqrt{m}$ and $\rho = a/m$ with $a \in \{0.25, 1\}$. This gives 80 deterministic values. For n_+ positive signs, put $S = 2n_+ - m$ and let k_+, k_- count coordinates with $a_i = t$ among positive and negative signs. With $q_0 = (1+\nu)^{-1/2}$ and $q_1 = (1 + t^2 + \nu)^{-1/2}$, the squared normalized sum divided by d equals

$$V(n_+, k_+, k_-) = [q_0 S - (q_0 - q_1)(k_+ - k_-)]^2 + t^2 q_1^2 (k_+ + k_-). \tag{48}$$

It is convex on $[0, n_+] \times [0, m - n_+]$, so its maximum is attained at one of the four corners, all of which are admissible integer choices. The exact distortion is therefore

$$\frac{\sqrt{1+\nu}\,\mathbb{E}_{n_+ \sim \mathrm{Bin}(m,1/2)} \sqrt{\max_{\text{four corners}} V(n_+, k_+, k_-)}}{\mathbb{E}_{n_+ \sim \mathrm{Bin}(m,1/2)} \sqrt{(2n_+ - m)^2 + t^2 m}}. \tag{49}$$

Binomial weights are computed stably. The notebook checks this reduction against exhaustive enumeration for $m \le 10$. All archived exact values lie above the construction lower bound and below the theorem upper bound. An individual finite-m family ratio can be below one: the separate lower bound one in Theorem 3.2 is for the supremum over all admissible classes.

F Targeted-audit protocol and provenance

F.1 E2: fixed-width grouped-family calculation

For the grouped family in Proposition 3.7, the final run uses $k \in \{4, 16, 64, 256, 1024\}$, $d = k + 1$, $\rho \in \{0.025, 0.1, (\sqrt{2}-1)/2\}$, and $\nu \in \{10^{-3}, 0.1\}$. It combines the six equal group sizes $n \in \{1, 4, 16, 64, 256, 1024\}$ with all available pairs having $m = kn \in \{4096, 16384, 65536\}$, leaving 198 unique configurations.

ν	ρ	Exact family	Construction lower	Mirrored lower	Theorem upper
0.1	0.025	1.986	0.939	1.970	2.602
0.1	0.1	3.857	2.920	3.840	8.323
0.1	0.2071	5.136	4.420	5.122	18.873
0.001	0.025	2.084	1.025	2.069	2.679
0.001	0.1	4.091	3.140	4.076	8.676
0.001	0.2071	5.405	4.683	5.393	19.736

Table 3: E1 at $m = 2048$ and $d = m+1$. The theorem upper column uses the sharp $1/\sqrt{2}$ first/second-moment constant of Lemma C.4 and evaluates the pre-uniformization upper bound at each row's own ρ; it coincides with the ρ_0-uniform bound when $\rho = \rho_0$. The exact constructed-family ratio lies below the worst-case upper bound and above both analytic lower bounds. The mirrored assignment tracks the exact family within 0.02 in all six displayed cases.

Exact inner supremum. Fix a sign vector and let p_j be the number of positive signs in group j (all groups have size $n = m/k$), $S = \sum_j (2p_j - n)$, $q_0 = (1+\nu)^{-1/2}$, and $q_1 = (1 + t^2 + \nu)^{-1/2}$. By orthogonality, $\|\sum_i \varepsilon_i h_i(a)\|_2^2 / d = S^2 + \sum_j U_j^2$ with $U_j = \sum_{i:g(i)=j} \varepsilon_i a_i$, so the raw supremum is attained by setting $a_i = t$ on the majority sign of each group:

$$\sup_a \|\textstyle\sum_i \varepsilon_i h_i(a)\|_2 = \sqrt{d}\left(S^2 + t^2 \sum_j \max\{p_j, n - p_j\}^2\right)^{1/2}.$$

After normalization, $\mathcal{N}(h_i(a)) = q(a_i) h_i(a)$ with $q(0) = q_0$ and $q(t) = q_1$. If k_{+j} positive-sign and k_{-j} negative-sign indices of group j take $a_i = t$, then with $D_j = k_{+j} - k_{-j} \in [-(n - p_j), p_j]$,

$$\|\textstyle\sum_i \varepsilon_i \mathcal{N}(h_i(a))\|_2^2 / d = V(D) := \left[q_0 S - (q_0 - q_1) \sum_j D_j\right]^2 + t^2 q_1^2 \sum_j D_j^2.$$

V is convex, so its maximum over the box is attained with every $D_j \in \{p_j, -(n - p_j)\}$. If group j takes the upper endpoint and group j' the lower one with $p_j < p_{j'}$, exchanging their choices leaves $\sum_j D_j$ unchanged and increases $\sum_j D_j^2$ by $2n(p_{j'} - p_j)$. Hence some maximizer assigns the upper endpoint to the ℓ groups with the largest counts, for some $\ell \in \{0, \dots, k\}$. Each draw is therefore processed as follows.

1. Draw the counts $p_1, \dots, p_k$, which are independent $\mathrm{Bin}(n, 1/2)$ (equivalently, draw the signs), and form S.
2. Sort the counts in decreasing order and form prefix sums of p_j, p_j^2, $n - p_j$, and $(n - p_j)^2$.
3. For $\ell = 0, \dots, k$, obtain $\sum_j D_j$ and $\sum_j D_j^2$ from the prefix sums and evaluate V_ℓ; set $X = (\max_\ell V_\ell)^{1/2}$.
4. Set $Y = (S^2 + t^2 \sum_j \max\{p_j, n - p_j\}^2)^{1/2}$.

The cost is $O(k \log k)$ per draw. For $n = 1$ the scan reduces to the four-corner E1 formula (48)–(49). Exhaustive enumeration over all signs and all $a \in \{0, t\}^m$ for small (m, k) agrees with the scan to machine precision.

Outer expectation. The common factor $\sqrt{d}$ cancels. The reported family ratio is $\sqrt{1+\nu}\,\bar{X}/\bar{Y}$, computed from the same $B = 65{,}536$ paired draws of (X, Y); it estimates a ratio of expectations, not the mean of per-draw ratios. Its standard error is the delta-method standard error for a ratio of paired means, $\sqrt{1+\nu}\,\widehat{\mathrm{sd}}(X - \hat{R}Y)/(\sqrt{B}\,\bar{Y})$ with $\hat{R} = \bar{X}/\bar{Y}$. The draws are split into eight atomic checkpoint files. The largest standard error is 4.23×10^{-3}, and the largest relative standard error is 1.12×10^{-3}. Every value is between the construction lower term and the minimum of the sample-size and fixed-width upper bounds with more than six standard errors of slack. An independent exact binomial calculation for all 30 $n = 1$ cases differs by at most 2.17 estimated standard errors. The raw archive contains all 1,584 checkpoint chunks, the complete 198-row table, and the plotting source.

F.2 E3: PRETRAINED CHECKPOINT AUDIT

Five compatible Llama-family pre-RMS/SwiGLU checkpoints are loaded sequentially and evaluated on WikiText-103 validation text. The exact model revisions are listed in Table 7. The dataset is

`Salesforce/wikitext`, configuration `wikitext-103-raw-v1`, validation split. Its exact revision, fingerprint, and raw-text SHA256 are stored in the archived dataset manifest. Thirty-two text chunks of length 256 are used for activation-radius measurements; 16 chunks of length 128 are used for attention probabilities, excluding the first 16 query positions to reduce the trivial short-prefix effect.

The E3 results combine two separate runs. The original audit supplies the embedding floors and the $\widehat{G}_M^{\mathrm{all}}$ column in Table 2, together with the attention-concentration summary. The induced-norm upper bound $G_{M,\mathrm{up}}^{\mathrm{all}}$ and the paired angle–magnitude diagnostic are reported model by model in Table 5. The later forward-pass run supplies the sampled-radius column, the paired diagnostic, and the directional decomposition in Table 4.

For each checkpoint we save the RMS radius of every embedding row, the minimum radius on the observed validation-token support, per-layer query/key/value and gated activation maxima, matrix statistics, and every sampled attention-row maximum $M = \max_j p_j$. Before computing matrix statistics, each learned RMSNorm gain is absorbed into the following query/key/value or gate/up projection, as described after Proposition 4.1. Operator norms in the main audit table use 80-step power iteration; on the SmolLM2-135M anchor, exact singular-value calculations on the calibration subset differ by at most 1.21×10^{-3} relatively. The archive also stores the generic valid upper bound $\|W\|_{\mathrm{op}} \le \sqrt{\|W\|_1 \|W\|_\infty}$ for every matrix. The main text uses the power estimates only as numerical diagnostics and does not promote sampled activation maxima or sampled attention minima to population bounds.

A separate forward-pass run uses the same five resolved model revisions and the same revision-pinned validation text. Four 8,192-character chunks are used only to choose three directions without seeing the evaluation chunks: the mean embedding direction, the leading uncentered embedding second-moment direction, and a coordinate direction selected by the largest calibration activation. The following 28 chunks are evaluation chunks; each is independently tokenized to length 256 and passed through one model at a time in FP32 with eager attention on a CUDA device. Hooks record the input, branch update, and reconstructed output at both the attention and FFN residual additions. Each model therefore contributes 7,168 token rows per stage. Completed chunks are written atomically; full hidden-state tensors are not retained. For every stage the archive stores radius minima, medians, 5th and 95th percentiles, update radii, inner products, negative cosines, three axis projections, and the exact energy-identity discrepancy. The largest such discrepancy across all models is 3.05×10^{-7}, and residual reconstruction has zero recorded discrepancy at output precision. Direction selection uses no evaluation chunk. These are correlated token-level diagnostics rather than independent observations or population certificates.

	Mean embedding		Leading moment		Selected coordinate	
Model	a_w	A_D^w	a_w	A_D^w	a_w	A_D^w
SmolLM2-135M	6.37×10^{-5}	8.88×10^{4}	3.72×10^{-5}	8.86×10^{4}	4.31×10^{-6}	1.99×10^{5}
SmolLM2-135M-Instr.	1.58×10^{-5}	9.10×10^{4}	9.67×10^{-5}	9.07×10^{4}	5.68×10^{-6}	2.04×10^{5}
SmolLM2-360M	2.37×10^{-5}	2.96×10^{5}	3.22×10^{-6}	2.97×10^{5}	1.02×10^{-5}	4.77×10^{5}
SmolLM2-360M-Instr.	1.02×10^{-4}	3.00×10^{5}	1.02×10^{-5}	3.01×10^{5}	2.31×10^{-5}	4.84×10^{5}
TinyLlama-1.1B-Chat	6.69×10^{-5}	3.13×10^{3}	3.03×10^{-6}	3.00×10^{3}	1.63×10^{-7}	4.47×10^{3}

Table 4: Directional certificate decomposition on evaluation-token support, using the power-iteration norm route. In every cell pair, $A_D^w \gg a_w$, so $a_w - A_D^w < 0$. The induced-norm route is still larger. The axes are chosen on separate calibration chunks.

The five runs contain $3{,}949{,}568$ sampled attention rows in total. Exact repository, dataset, software, and run manifests, together with compressed raw row maxima and the original Colab result archives, are included in the supplementary repository.

G FIXED-WIDTH NORMALIZATION: FULL PROOFS

Let U be uniform on the Euclidean unit sphere S^{d-1} and put

$$\kappa_d = \mathbb{E}|U_1| = \frac{\Gamma(d/2)}{\sqrt{\pi}\,\Gamma((d+1)/2)}. \tag{50}$$

Model	a_E^{all}	a_E^{obs}	$\widehat{G}_M^{\mathrm{all}}$	$G_{M,\mathrm{up}}^{\mathrm{all}}$	sampled $r_{\min}$	paired final
SmolLM2-135M	0.0732	0.0732	3.53×10^{6}	5.41×10^{8}	0.0913	2.293
SmolLM2-135M-Instr.	0.0747	0.0747	3.54×10^{6}	5.43×10^{8}	0.0956	2.696
SmolLM2-360M	0.0727	0.0776	7.60×10^{6}	1.59×10^{9}	0.1083	2.323
SmolLM2-360M-Instr.	0.0749	0.0783	7.49×10^{6}	1.57×10^{9}	0.1134	2.330
TinyLlama-1.1B-Chat	7.82×10^{-21}	0.00466	8.97×10^{23}	2.97×10^{26}	0.00347	0.080

Table 5: Complete checkpoint audit. a_E^{all} and a_E^{obs} separate full-vocabulary and observed-token support. The two G_M columns use, respectively, power-iteration estimates and induced-norm upper bounds $\sqrt{\|W\|_1\|W\|_\infty}$. The sampled radius is the minimum after any residual sublayer over all 7,168 evaluation tokens per stage. The paired final column is the cumulative paired angle–magnitude lower diagnostic at the last stage; it is sample-derived and is not a population lower bound. All tested isotropic and directional sufficient floors are negative at native scale; a failed sufficient test does not show that the population floor is absent. Even on TinyLlama's observed-token support the amplitude diagnostic is 1.51×10^6.

Model	d	final $r_{.05}$	final $r_{.95}$	sampled factor
SmolLM2-135M	576	28.404	45.950	99.3
SmolLM2-135M-Instr.	576	29.056	49.902	105.3
SmolLM2-360M	960	35.614	60.575	134.1
SmolLM2-360M-Instr.	960	34.615	63.211	143.8
TinyLlama-1.1B-Chat	2048	0.511	1.034	231.5

Table 6: Sampled readout-stage instantiation of the fixed-width factor $2[1 + R/(\kappa_d\sqrt{c^2+\nu})]$, using archived 5th and 95th radius percentiles for (c, R). Percentiles are not uniform class bounds, so this is a scale diagnostic rather than a theorem certificate.

Model	resolved Hugging Face revision
SmolLM2-135M	`93efa2f097d58c2a74874c7e644dbc9b0cee75a2`
SmolLM2-135M-Instr.	`12fd25f77366fa6b3b4b768ec3050bf629380bac`
SmolLM2-360M	`f8027fd0eaeea54caa13c31d31b9fdc459c38b49`
SmolLM2-360M-Instr.	`a10cc1512eabd3dde888204e902eca88bddb4951`
TinyLlama-1.1B-Chat	`fe8a4ea1ffedaf415f4da2f062534de366a451e6`

Table 7: Exact model revisions used in the pretrained audit.

Rotational invariance implies $\|x\|_2 = \kappa_d^{-1}\mathbb{E}_U|\langle U, x\rangle|$. Moreover,

$$\frac{1}{\sqrt{d}\,\kappa_d} \le \sqrt{\frac{\pi}{2}}. \tag{51}$$

Indeed, for $G \sim N(0, I_d)$, independence of its radius and direction gives $\kappa_d\mathbb{E}\|G\|_2 = \mathbb{E}|G_1| = \sqrt{2/\pi}$, while $\mathbb{E}\|G\|_2 \le \sqrt{d}$.

Lemma G.1 (Scalar complexity of the RMS radii)**.** *For the scalar radius class $\mathscr{R} = \{(\|h_i(\theta)\|_\circ)_i : \theta \in \Theta\}$,*

$$\widehat{\mathfrak{R}}_m(\mathscr{R}) \le \frac{1}{\sqrt{d}\,\kappa_d}\mathfrak{R}_m^{\mathrm{vec}}(\mathcal{H}) \le \sqrt{\frac{\pi}{2}}\mathfrak{R}_m^{\mathrm{vec}}(\mathcal{H}). \tag{52}$$

Proof. For each fixed sign vector, move the supremum inside the spherical expectation. Then apply scalar contraction to $z \mapsto |z|$, with the convention that scalar complexity has no outside absolute value:

$$\begin{aligned}
m\widehat{\mathfrak{R}}_m(\mathscr{R}) &= \frac{1}{\sqrt{d}\,\kappa_d}\mathbb{E}_\varepsilon \sup_\theta \mathbb{E}_U \sum_i \varepsilon_i |\langle U, h_i(\theta)\rangle| \\
&\le \frac{1}{\sqrt{d}\,\kappa_d}\mathbb{E}_U\mathbb{E}_\varepsilon \sup_\theta \sum_i \varepsilon_i |\langle U, h_i(\theta)\rangle| \\
&\le \frac{1}{\sqrt{d}\,\kappa_d}\mathbb{E}_U\mathbb{E}_\varepsilon \sup_\theta \sum_i \varepsilon_i \langle U, h_i(\theta)\rangle \\
&\le \frac{m}{\sqrt{d}\,\kappa_d}\mathfrak{R}_m^{\mathrm{vec}}(\mathcal{H}).
\end{aligned}$$

The last step uses $\|U\|_2 = 1$. Boundedness and separability justify the expectations and suprema; finite classes suffice first, followed by separable approximation. The scalar contraction convention is recalled in Maurer (2016). □

G.1 Proof of Theorem 3.6

Proof. The inverse-radius function $q(r) = (r^2 + \nu)^{-1/2}$ satisfies

$$|q'(r)| = r(r^2 + \nu)^{-3/2} \le b_c^2 \qquad (r \ge c).$$

Scalar contraction and Lemma G.1 therefore imply, for $\mathscr{Q} = \{(q(\|h_i(\theta)\|_\circ))_i : \theta \in \Theta\}$,

$$\widehat{\mathfrak{R}}_m(\mathscr{Q}) \le \frac{b_c^2}{\sqrt{d}\,\kappa_d}\mathfrak{R}_m^{\mathrm{vec}}(\mathcal{H}).$$

One may extend q from $[c, R]$ to a function on the line with the same Lipschitz constant. A constant shift has zero scalar complexity, so no zero-at-origin issue occurs.

The projection class $\mathscr{P} = \{(\langle u, h_i(\theta)\rangle)_i : \|u\|_2 \le 1, \theta \in \Theta\}$ has $\widehat{\mathfrak{R}}_m(\mathscr{P}) = \mathfrak{R}_m^{\mathrm{vec}}(\mathcal{H})$ and envelope $\sqrt{d}R$. Enlarging the coupled product to $\mathscr{P}\mathscr{Q}$ and applying Lemma C.2 yields

$$\mathfrak{R}_m^{\mathrm{vec}}(\mathcal{N} \circ \mathcal{H}) \le 2b_c\mathfrak{R}_m^{\mathrm{vec}}(\mathcal{H}) + 2\sqrt{d}R\widehat{\mathfrak{R}}_m(\mathscr{Q}),$$

which proves (11) for gain one. For a hypothesis-dependent but sample-common gain, bound its diagonal operator norm by G_0 before the supremum. Set $c = 1, R = 1 + 2\rho$, divide by $b_1\mathfrak{R}_m^{\mathrm{vec}}(\mathcal{H})$, and apply (51) to obtain (12). The floor ensures the denominator is positive. The last display is the minimum of two valid upper bounds. □

Remark G.2 (Actual dimension versus class-wide subspace dimension). If every $h_i(\theta)$ belongs to one fixed r-dimensional linear subspace of $\mathbb{R}^d$, apply the spherical identity inside that subspace. Formula (11) then holds with κ_r in place of κ_d (the RMS scaling is still $\sqrt{d}$). Thus the factor can depend on the class-wide span dimension. A low effective rank measured at one checkpoint is not, by itself, such a uniform subspace assumption.

G.2 Proof of Theorem 3.8

Proof. Let $k = \min\{m, d-1\}$ and use balanced groups in Proposition 3.7. Put $A = 1 + \nu$ and $T_0 = 4\rho_0(1 + \rho_0)$. Since $t^2 = 4\rho(1 + \rho)$,

$$1 - a_{\rho,\nu} = \int_0^{t^2} \frac{\sqrt{A}}{2(A + u)^{3/2}}\,du \ge \frac{t^2\sqrt{A}}{2(A + T_0)^{3/2}} \ge c_0\rho,$$

where $c_0 > 0$ depends only on (ν, ρ_0). Also $t^2 \le 4(1 + \rho_0)\rho$. Hence

$$C^\star_{m,d,\nu}(\rho) \ge \max\left\{1, c_1\frac{\rho}{\sqrt{m^{-1} + \rho/k}}\right\} \ge \max\left\{1, c_2\min(\rho\sqrt{m}, \sqrt{\rho k})\right\}.$$

Because $d \ge 2$, $k = \min\{m, d-1\} \ge \min\{m, d\}/2$, proving (15).

For the upper bound, take the minimum of (8) and (12). Since $0 \le \rho \le \rho_0$ and $d \ge 2$, their constants can be absorbed to give (16). If $0 < \rho \le 1$ and $m \le d/\rho$, then $\rho\sqrt{m} \le \sqrt{\rho\min\{m, d\}}$, so the lower and upper displays match at order $1 + \rho\sqrt{m}$. If instead $0 < \rho \le \min\{1, \rho_0\}$ and $m > d/\rho$, the lower nonconstant term is of order $\sqrt{\rho d}$, while the upper is at most $\min\{\rho\sqrt{m}, \sqrt{d}\}$; their ratio is at most of order $\rho^{-1/2}$, with constants depending only on ν and ρ_0. This bounds the gap between the two envelopes; it is not an exact rate in that regime. This completes the proof. □

G.3 Proof of Corollary 3.9

Proof. The upper bound follows from (13). For the lower bound, take $k = \min\{m, d-1\}$ and balanced groups. Since $k \ge \min\{m, d\}/2$ and $m^{-1} \le k^{-1}$,

$$C^\star_{m,d,\nu}(\rho) \ge \max\left\{1, \frac{1 - a_{\rho,\nu}}{2\sqrt{1 + 2t^2}}\sqrt{k}\right\}.$$

For fixed $\rho > 0$, the coefficient is strictly positive. Combining the constant and square-root terms proves the result. At $\rho = 0$, every state is scaled by exactly $(1+\nu)^{-1/2}$, so the distortion is one for every class. □

H Propagation refinements and an explicit consumer

H.1 Hybrid propagation product

Let $\mathcal{F}$ be the FFN stages, $\mathcal{A}$ the attention stages, $L_s = 1 + \lambda_s \Lambda_s \omega_{s-1}$, and $\chi = 1 + \kappa_{\mathrm{sw}}$.

Proposition H.1 (Integrated and hybrid propagation product)**.** *Let $J(x) = \operatorname{arsinh}(x/\sqrt{\nu})$. Then*

$$\Pi_s \le \left[\prod_{\substack{u>s \\ u\in\mathcal{A}}} L_u\right] \exp\left\{\chi \sum_{\substack{u>s \\ u\in\mathcal{F}}} [J(c_{u-1}) - J(c_u)]\right\} \tag{53}$$

$$\le \left[\prod_{\substack{u>s \\ u\in\mathcal{A}}} L_u\right] \left(\frac{c_s + \sqrt{c_s^2+\nu}}{c_D + \sqrt{c_D^2+\nu}}\right)^{\chi}. \tag{54}$$

If $\mathcal{A} = \varnothing$, this reduces to $\Pi_s \le [(c_s + \sqrt{c_s^2+\nu})/(c_D + \sqrt{c_D^2+\nu})]^{\chi}$.

Proof. At $u \in \mathcal{F}$, $\log L_u \le \chi \lambda_u M_u \omega_{u-1}$. Because $c_{u-1} - c_u = \lambda_u M_u$ and $x \mapsto (x^2+\nu)^{-1/2}$ is decreasing for $x \ge 0$,

$$\lambda_u M_u \omega_{u-1} \le \int_{c_u}^{c_{u-1}} \frac{dx}{\sqrt{x^2+\nu}} = J(c_{u-1}) - J(c_u).$$

Keep the attention factors unchanged and sum these inequalities to obtain (53). All intervals have nonnegative integrals, so filling in the omitted intervals bounds the exponent by $\chi[J(c_s) - J(c_D)]$, proving (54). □

This avoids choosing one enormous χ based on attention. It does *not* make a model with an uncertified radius floor satisfy the hypotheses; the floor requirement remains explicit.

H.2 Parameter perturbations and finite covers

Proposition H.2 (Two-model perturbation recurrence)**.** *Consider two models with the same residual coefficients and common certified floors c_s. Suppose the normalized-input branch sensitivity is at most Λ_s uniformly over the class, and define*

$$D_E = \left\|H^0 - H'^0\right\|_{\infty,\circ}, \qquad D_s = \sup_{\|Z\|_{\infty,\circ}\le 1} \|F_s(Z;W_s) - F_s(Z;W_s')\|_{\infty,\circ}.$$

For the same input sequence,

$$\left\|H^D - H'^D\right\|_{\infty,\circ} \le \Pi_0 D_E + \sum_{s=1}^{D} \lambda_s \Pi_s D_s. \tag{55}$$

For gain-free normalized readouts $f_{\theta,w} = \langle w, \mathcal{N}(h_\theta)\rangle$, with $\|w\|_2, \|w'\|_2 \le B$,

$$|f_{\theta,w} - f_{\theta',w'}| \le B\sqrt{d}\,\omega_D \left(\Pi_0 D_E + \sum_s \lambda_s \Pi_s D_s\right) + \sqrt{d}\,\|w - w'\|_2. \tag{56}$$

Proof. Add and subtract $F_s(\mathcal{N}(H'^{s-1}); W_s)$ in the branch difference. Lipschitz continuity of normalization above the common floor gives

$$\Delta_s \le L_s \Delta_{s-1} + \lambda_s D_s, \qquad \Delta_s = \|H^s - H'^s\|_{\infty,\circ}.$$

Unroll this recursion. For the readout, separate the hidden-state and readout perturbations, use the Lipschitz coefficient ω_D, convert RMS to Euclidean norm, and use $\|\mathcal{N}(h)\|_2 \le \sqrt{d}$. □

Corollary H.3 (A finite-cover consumer)**.** *Suppose an admissible product parameter class has an internal embedding net of size N_E and radius e_E, and internal branch nets of sizes N_s and radii e_s in the branch metrics D_s above. Assume every combination of net centers remains in the uniformly certified class. Let a Euclidean e_w-net of the readout ball have size N_w. Put*

$$K = N_E N_w \prod_s N_s, \qquad \eta = \Pi_0 e_E + \sum_s \lambda_s \Pi_s e_s, \qquad \delta = \sqrt{d}\,(B\omega_D \eta + e_w).$$

On any fixed sample, the normalized scalar readout class satisfies

$$\widehat{\mathcal{R}}_m(\mathscr{F}_N) \le \delta + B\sqrt{\frac{2d\log K}{m}}. \tag{57}$$

Proof. Proposition H.2 makes the product net a prediction-space δ-net in the supremum metric. Replacing a predictor by its net representative adds at most δ to the empirical complexity. Each net predictor is bounded by $B\sqrt{d}$. For K such vectors, the Rademacher moment generating function bound and the log-sum-exp inequality give expected maximum signed sum at most $B\sqrt{2dm\log K}$. Divide by m. The $K = 1$ case follows since a singleton scalar class has zero empirical complexity. □

For a coupled parameter class, use an internal net of that full class in the weighted metric in (55), rather than a product net whose centers may violate the constraints. This prevents a hidden admissibility gap. Equation (57) is a transparent consumer of propagation, not a claim of a state-of-the-art transformer bound. No numerical usefulness at the audited checkpoints is inferred.

I Optional structural and finite-sample refinements

I.1 A conditional paired angle–magnitude floor

This alternative adds a uniform directional assumption. It is not implied by operator-norm caps or by a sampled angle diagnostic.

Proposition I.1 (Combined non-cancellation and amplitude floor)**.** *At stage s, uniformly over the specified class, population inputs, and positions, write $h = H_t^{s-1}$ and $f = F_s(\mathcal{N}(H^{s-1}); W_s)_t$. Suppose*

$$\|h\|_\circ \ge c_{s-1} > 0, \qquad \|f\|_\circ \le M_s, \qquad \langle h, f\rangle \ge -\zeta_s \|h\|_2 \|f\|_2, \qquad 0 \le \zeta_s < 1. \tag{58}$$

For $u_s = \lambda_s M_s$, define

$$c_s := \Phi(c_{s-1}, u_s, \zeta_s) = \begin{cases} \sqrt{c_{s-1}^2 - 2\zeta_s u_s c_{s-1} + u_s^2}, & u_s < \zeta_s c_{s-1}, \\ c_{s-1}\sqrt{1-\zeta_s^2}, & u_s \ge \zeta_s c_{s-1}. \end{cases} \tag{59}$$

Then $\|h + \lambda_s f\|_\circ \ge c_s$. Recursing from $c_0 = a_0$ gives a population radius floor through depth. This floor dominates both the reverse-triangle update $(c_{s-1} - u_s)_+$ and the angle-only update $c_{s-1}\sqrt{1-\zeta_s^2}$.

Proof. Set $r = \|h\|_\circ$ and $v = \|f\|_\circ$. The energy identity gives

$$\|h + \lambda_s f\|_\circ^2 \ge r^2 - 2\lambda_s \zeta_s r v + \lambda_s^2 v^2.$$

Minimize the right-hand side over $r \ge c_{s-1}$ and $0 \le v \le M_s$. If $u_s < \zeta_s c_{s-1}$, the minimizer is $(r, v) = (c_{s-1}, M_s)$; otherwise, when $\lambda_s > 0$, $(r, v) = (c_{s-1}, \zeta_s c_{s-1}/\lambda_s)$ is feasible. The cases $\lambda_s = 0$ or $\zeta_s = 0$ are immediate. These two cases give (59). The displayed quadratic also shows directly that the result dominates the two separate floors. If $\langle h, f\rangle \ge 0$, the exact energy identity implies that radii are nondecreasing. □

If $\zeta_s \le \bar{\zeta} < 1$ and $\sum_s \zeta_s^2 \le V < \infty$, then

$$\inf_{D,t,X,\theta} \left\|H_t^D\right\|_\circ \ge a_0 \exp\left(-\frac{V}{2(1-\bar{\zeta}^2)}\right) > 0,$$

by $\log(1-x) \geq -x/(1-\bar{\zeta}^2)$ for $0 \leq x \leq \bar{\zeta}^2$.

This is an optional structural refinement, not an unconditional theorem about trained decoders. Operator-norm upper bounds alone do not imply (58). A sampled paired value $\sqrt{1-2\zeta_s x_s + x_s^2}$, where $x_s = \lambda_s \|f\|_\circ / \|h\|_\circ$, is a diagnostic and cannot certify a population or across-hypothesis condition. Pairing ζ_s and x_s from the same token avoids the artificial cancellation produced by combining separate worst-case marginals. The hybrid integral formula in Proposition H.1 specifically uses the amplitude-floor trajectory $a_0 - A_s$; it should not be transplanted to an arbitrary floor sequence.

Sampled diagnostic constructions. Let i index evaluated token trajectories and $s = 1, \ldots, D$ index residual additions. Write $r_{s,i} = \|h_{s,i}\|_\circ$, with $r_{0,i}$ the embedding radius; every $r_{s-1,i}$ is positive on the evaluated support. Let $\Delta h_{s,i} = h_{s,i} - h_{s-1,i}$ be the actual residual update, including the residual multiplier λ_s. Define $x_{s,i} = \|\Delta h_{s,i}\|_\circ / r_{s-1,i}$ and $\zeta_{s,i} = \max\{0, -\cos(h_{s-1,i}, \Delta h_{s,i})\}$, with $\zeta_{s,i} = 0$ when $\Delta h_{s,i} = 0$, and set $b_{s,i} = \sqrt{1-2\zeta_{s,i}x_{s,i} + x_{s,i}^2}$ and $a_{s,i} = \sqrt{1-\zeta_{s,i}^2}$. The two reported constructions are

$$L_{\mathrm{old}}(D) = \min_i r_{0,i} \prod_{s=1}^{D} \min_i a_{s,i}, \qquad L_{\mathrm{paired}}(D) = \min_i \left[r_{0,i} \prod_{s=1}^{D} b_{s,i} \right].$$

The first takes separate adverse tokens at each stage and discards observed magnitudes. The second retains one trajectory and its magnitudes. Both changes can improve the bound. Positive state–update cosines are clipped to zero, so the paired curve is still generally below exact energy reconstruction. Depth curves (Figure 3, right) replace D by the current stage. Both quantities are computed from observed trajectories; neither is the uniform recursion of Proposition I.1 evaluated with certified class-wide constants.

J Extended positioning and related work

This appendix collects the extended discussion of prior work and positioning that was summarized in §1.1 for space.

Positioning. RMSNorm (Zhang & Sennrich, 2019) and LayerNorm (Ba et al., 2016) normalize different radial quantities; the obstruction in Section 3 applies to the former and, after restricting to the mean-zero subspace, to the latter. The zero-bias CenterNorm map of LipsFormer (Qi et al., 2023) is a fixed linear map and contracts immediately (Proposition C.1), which isolates the role of division by a class-dependent radius. Maurer's vector-contraction inequality (Maurer, 2016) already shows in Section 6 that generic dimension-free single-sign contraction fails. Our novelty is the RMS-specific annulus and width characterization, not that generic obstruction. Maurer's positive contraction inequality uses a different process: vector outputs are paired with independent signs ϵ_{ik} across sample and coordinate indices, with the familiar $\sqrt{2}$ contraction factor, whereas (2) has one sign per sample and a norm outside the sum. Truong's high-dimensional contraction lemmas (Truong, 2025) likewise treat coordinatewise vector maps; RMS normalization is non-separable because every coordinate is divided by a common radial functional. The peeling technique and the role of positive homogeneity are due to Golowich et al. (2018); RMS normalization is homogeneous of degree zero as $\nu \to 0$, so that argument does not transfer directly. Covering-based transformer analyses (Edelman et al., 2022; Trauger & Tewari, 2024) sidestep this contraction question entirely. Recent spectrum-adaptive post-hoc analyses also emphasize the value of checkpoint-dependent spectral structure rather than fixed global norm budgets (Sakai & Imaizumi, 2026). Residual-scaling schemes with small shortcut coefficients (Qi et al., 2023) are an architectural analogue of the amplitude budget of §4. The obstruction proved here is specific to arguments that pass through the single-sign process in (2); it is not a claim about covering-number, PAC-Bayesian, stability, compression, or other analyses that use different proof objects.

Conversely, the spherical-average identity used for our fixed-width upper bound is the finite-dimensional positive counterpart of Maurer's obstruction: $\kappa_d^{-1} \asymp \sqrt{d}$, and the balanced-group construction shows that this dimension dependence is sharp at each fixed positive ρ, up to constants depending on the fixed ρ and ν (Corollary 3.9).

Relation to residual scaling and attention regularity. ReZero (Bachlechner et al., 2021), SkipInit (De & Smith, 2020), and DeepNorm (Wang et al., 2024) study residual parameterizations and initialization for optimization. Signal-propagation analyses such as Noci et al. (2022) study rank collapse and depth scaling at initialization. Our amplitude induction is a deterministic sufficient condition uniform over a constrained class; it is not an initialization or training-dynamics result. Likewise, Kim et al. (2021); Castin et al. (2024) study attention Lipschitz behavior, whereas our normalization theorem concerns a single-sign complexity ratio. These are different quantities and assumptions; the architectural bounds in §4 are supporting certificates rather than the main contribution.

K Supplementary proofs, scope, and audit details

This appendix collects proofs, scope qualifications, and audit details deferred from the main text for space.

K.1 Scope and proof details for Section 3

Proof idea for Theorem 3.1. Clip the reference radii from below at c and write $q_i(\theta) = (\|h_i(\theta)\|_\circ^2 + \nu)^{-1/2} = q_i^0 + \Delta q_i(\theta)$. Then $q_i^0 \le b_c$ and $|\Delta q_i| \le \rho b_c^2$. Fixed-coordinate contraction controls the $q_i^0 h_i$ term; the triangle inequality controls the remaining signed sum by $mM\rho b_c^2$. The gain is a common operator for each hypothesis.

The profile may vary arbitrarily between examples; the condition controls variation *between hypotheses at the same example*. Measuring the radii of one checkpoint across tokens does not quantify this.

Additional displays for Theorem 3.2. If $S = \sum_i \varepsilon_i$, then

$$\mathfrak{R}_m^{\mathrm{vec}}(\mathcal{H}_{m,t}) = \frac{\sqrt{d}}{m}\mathbb{E}\sqrt{S^2 + t^2 m} \le \sqrt{\frac{d(1+t^2)}{m}}.$$

The expected projection of the normalized signed sum onto e_0, divided by m, is

$$\frac{\sqrt{d}}{2}\left\{(1+\nu)^{-1/2} - (1+t^2+\nu)^{-1/2}\right\}.$$

The construction illustrates the obstruction: before normalization the shared component is independent of a, whereas after normalization its coefficient can respond to the signs. The orthogonal components keep the unnormalized supremum small enough for the ratio to grow.

Corollary 3.3 context. The process in (2) is exactly the empirical complexity exposed by a bounded linear readout. Corollary 3.3 is intentionally a scalar-readout statement rather than a claim of a state-of-the-art language-model bound. The amplitude results supply both a floor and a profile-width bound when embedding variation around one common reference is specified; see Proposition 4.6. In (34), the factor in parentheses may be replaced by the smaller bound in (13). Neither the width theorem nor a radius floor alone proves the raw-complexity assumption. Here m counts independent examples, not correlated tokens treated as independent observations.

Proposition 3.4 scope. The result removes hypothesis-dependent embeddings from the realization, but not the growing-width feature: $d = m + 1$. For the extremal choices in the proof, $\|W_O\|_{\mathrm{op}} = s$, whereas $\|W_O\|_F = s\sqrt{m}$ and $\|W_O\|_{2,1} = sm$. Thus this decoder realization establishes sharpness for a combinatorially restricted subclass obeying an operator/spectral-norm cap. It does not establish sharpness for the complete norm ball: that ball contains collapsing output projections, and complexity ratios are not monotone under class enlargement. Nor does this realization establish a matching lower bound under sample-independent Frobenius or $(2,1)$ budgets. That question remains open.

Proof idea for Theorem 3.6. For a uniform sphere direction U, $\|x\|_2 = \kappa_d^{-1}\mathbb{E}|\langle U, x\rangle|$. Scalar contraction therefore bounds the scalar complexity of the RMS radii by $(\sqrt{d}\kappa_d)^{-1}\mathfrak{R}_m^{\mathrm{vec}}(\mathcal{H})$, with $(\sqrt{d}\kappa_d)^{-1} \le \sqrt{\pi/2}$. Apply scalar contraction again to the inverse radius and the product bound of Lemma C.2.

General-group formula for Proposition 3.7. The bound for arbitrary positive group sizes is (41), stated and proved in Appendix C.8; the balanced case (14) follows from $\sum_j n_j^2 \le 2m^2/k$.

Remark K.1 (Primitive floor and later profile composition). A reference-centered route can assume a known hidden-state radius profile and propagate it to nearby parameters, but this relocates the floor hypothesis. The approach of Section 4 instead derives the primitive population floor directly from an embedding floor and a residual-amplitude budget. A common reference embedding appears later only to convert that primitive control into an across-hypothesis radius profile for the complexity composition; no deeper hidden-state profile is assumed there.

K.2 Scope and remarks for Section 4

The mechanism is a distinction the Lipschitz analysis obscures. A branch has a *sensitivity*, bounding how an input perturbation propagates, and an *amplitude*, bounding how far one residual update moves the stream. Because the branch reads a normalized input, its amplitude admits a bound valid regardless of the size of the stream, and for attention regardless of the query and key caps. These bounds are deliberately separated from branch Lipschitz constants: the former control how far a residual update can move the hidden state, while the latter control how perturbations propagate through the network.

Remark K.2 (Origin of the $\sqrt{d}$ factor in the SwiGLU bound). The factors $\beta_a = \sqrt{d}\, C_g$ and $\beta_b = \sqrt{d}\, C_u$ come from combining the normalized-input bound $\|z\|_2 \le \sqrt{d}$ with a column-norm bound such as $|zA_{g,\cdot j}| \le \|z\|_2 \|A_{g,\cdot j}\|_2$. Thus the displayed $\sqrt{d}$ is a consequence of this uniform RMS-plus-column-norm route. It is not asserted to be necessary for a trained branch; a direct $2\to\infty$ bound or other structure could replace β_a, β_b by smaller quantities.

Remark K.3 (Learned RMSNorm gains). Practical RMSNorm layers include a learned gain γ. In the row-vector convention, $\mathrm{RMS}_\gamma(h) = \mathcal{N}(h)\,\mathrm{diag}(\gamma)$, so the gain can be absorbed into the first linear maps of the following branch: replace A_Q, A_K, A_V by $\mathrm{diag}(\gamma)A_Q, \mathrm{diag}(\gamma)A_K, \mathrm{diag}(\gamma)A_V$ and replace A_g, A_u analogously in SwiGLU. The output maps W_O and W_d are unchanged. If $\|\gamma\|_\infty \le G$, operator, column, and mixed-norm caps of these effective first maps increase by at most a factor G. All amplitude and sensitivity formulas in this paper are to be read for these effective matrices. The checkpoint audit performs this absorption before computing the reported matrix statistics.

Remark K.4 (Why these constants differ from Lipschitz constants). For attention, query/key terms are needed to control how the softmax weights change when the input changes, but they are not needed to control the absolute size of a probability-weighted average of the values. Thus M_{att} can be much smaller than Λ_{att}. For SwiGLU the same norm ingredients control both quantities, but the product rule contributes both the gate and up-projection perturbations; Proposition A.2 gives the explicit comparison $\Lambda_{\mathrm{ff}} \le (1+\kappa_{\mathrm{sw}})M_{\mathrm{ff}}$.

Propagation setup. Suppose now that each branch additionally has an input-Lipschitz certificate Λ_s on normalized inputs. Proposition A.2 supplies explicit choices for the attention and SwiGLU branches considered here:

$$\|F_s(Z;W_s) - F_s(Z';W_s)\|_{\infty,\circ} \le \Lambda_s \|Z - Z'\|_{\infty,\circ}.$$

When stage $s-1$ has a certified floor c_{s-1}, the RMS normalizer is Lipschitz with coefficient $\omega_{s-1} := (c_{s-1}^2 + \nu)^{-1/2}$.

The product Π_0 first controls perturbations under the same parameters. For two models, branch-parameter differences introduce forcing terms: with D_E the RMS embedding difference and $D_s = \sup_{\|Z\|_{\infty,\circ}\le 1} \|F_s(Z;W_s) - F_s(Z;W_s')\|_{\infty,\circ}$,

$$\left\|H^D - H'^D\right\|_{\infty,\circ} \le \Pi_0 D_E + \sum_{s=1}^{D} \lambda_s \Pi_s D_s.$$

Remark K.5 (Two budgets with different meanings). The theorem separates

$$\sum_s \lambda_s M_s \qquad \text{and} \qquad \sum_s \lambda_s \Lambda_s.$$

The first is an *amplitude budget*: it prevents cumulative residual motion from consuming the initial embedding radius. The second is a *sensitivity budget*: once a positive radius floor is available, it

prevents the product of perturbation multipliers from exploding. A single branch Lipschitz constant can upper-bound both effects, but doing so can be much looser than necessary.

Remark K.6 (Nonemptiness and numerical interpretation). The theorem defines a nonempty sufficient subclass whenever the admissible embedding set contains an embedding table with the required positive floor and the branch constraint sets contain, for example, the zero matrices. Its numerical scale is determined by the sharper branch amplitudes M_s together with the two cumulative budgets; Experiment E3 reports those quantities directly for public checkpoints.

Remark K.7 (One-sided nature of the certificate). The reverse-triangle argument uses only an upper bound on the magnitude of each residual update, so its floor is one-sided: equality need not hold when updates cancel or when the state and update share coherent directions. The directional Theorem A.3 records one way to retain such structure without changing the residual recursion.

K.3 Additional checkpoint audit details

The complete model-by-model audit table with all seven columns (both norm routes, sampled minima, and paired final values) is Table 5 in Appendix F.

Certificate details. The hatted route ($\widehat{G}_M^{\mathrm{all}}$) is a numerical diagnostic; the induced-norm route ($G_{M,\mathrm{up}}^{\mathrm{all}}$) is a mathematically valid matrix-norm bound evaluated in floating point rather than outward-rounded interval arithmetic. Failure of either sufficient condition does not predict hidden-state collapse.

Four calibration chunks choose three directions without seeing the evaluation chunks: the mean embedding direction, the leading uncentered embedding second-moment direction, and a coordinate direction selected by the largest calibration activation.

All three directional amplitude calculations are negative for all five checkpoints under both norm routes. The three natural directions do not repair the uniform certificate at native scale. Appendix Table 4 reports a_w and A_D^w separately and shows that the cumulative budget is the active failure. This negative result does not rule out a useful learned direction or a complementary local analysis around trained weights.

The paired trajectory diagnostic $L_{\mathrm{paired}}(D)$ (Appendix I.1) remains positive through depth and ends between 0.080 and 2.70. The earlier diagnostic $L_{\mathrm{old}}(D)$, which pooled the most adverse angle separately at each stage and discarded relative update magnitudes, reached 10^{-10}–10^{-6}. This comparison changes both the information retained and the aggregation across token trajectories; it illustrates the value of retaining state–update information, consistent with the motivation of Proposition I.1, but does not isolate the benefit of magnitude information alone. Neither diagnostic is a population certificate.

Readout-stage instantiation. We instantiate Theorem 3.6 at the sampled final readout stage using c and R equal to the archived 5th and 95th radius percentiles. The resulting dimensionless factors $2[1 + R/(\kappa_d\sqrt{c^2+\nu})]$ range from 99.3 to 231.5 (Table 6). These are percentile-based diagnostics, not class-wide certificates, but they make clear that the fixed-width theorem controls growth in m without promising a small numerical constant at transformer widths.

K.4 Sharp laws and softmax refinement

Sharp laws and the remaining joint-rate corner. Theorem 3.2 has a matching lower bound uniform over bounded, possibly shrinking annuli when width is unrestricted. Independently, Theorem 3.6 and Proposition 3.7 match for every fixed positive ρ: Corollary 3.9 gives the order $1 + \sqrt{\min\{m, d\}}$. This rules out divergence in m at fixed width and identifies the saturation level $1 + \sqrt{d}$. Theorem 3.8 additionally gives a uniform joint lower envelope $1+\min\{\rho\sqrt{m}, \sqrt{\rho\min(m,d)}\}$ and upper envelope $1 + \min\{\rho\sqrt{m}, \sqrt{d}\}$. These match throughout $m \le d/\rho$. For $0 < \rho \le \min\{1, \rho_0\}$ and $m > d/\rho$, the remaining gap between the envelopes is at most of order $\rho^{-1/2}$; we do not claim a matching joint rate in that corner. Nor do evaluation-class lower bounds automatically establish sharpness for a fixed-vocabulary, fixed-parameter-budget transformer family.

Supporting attention-sensitivity refinement. The exact pointwise softmax Jacobian norm is classical matrix-norm geometry and has also been stated explicitly in recent attention-sensitivity work (Alon & Naor, 2006; Emadi, 2026). The contribution used here is the interpolation step: Lemma D.3 preserves a dominant mass under bounded score perturbations, and Proposition D.4 converts that fact into a sub-one Lipschitz coefficient along the entire segment whenever $M_\eta \geq 1/2$. This directly complements bounded-domain analyses that revert to the generic $\theta \leq 1$ along an interpolation path (Emadi, 2026). The stagewise Corollary D.5 localizes this requirement instead of imposing one global concentration floor across all stages. The checkpoint audit nevertheless finds a sampled diffuse row in every stage: the largest layerwise minimum is only 0.121, so at $\eta = 0.05$ the sampled-support coefficient remains one stage by stage. This separates the mathematical gain from the stronger population condition needed to use it uniformly.

K.5 Extended reproducibility details

E2 computes each grouped-family inner supremum exactly and estimates only the outer Rademacher expectation; the final archive contains all 198 rows, 65,536 draws per row, 1,584 atomic checkpoint chunks, Monte Carlo standard errors, seeds, the exact $n = 1$ regression check, and the code that regenerates Figure 2. E3: the main table uses power-iteration spectral estimates as diagnostics; a separate induced-norm upper-bound calculation and an exact-vs-power calibration on the anchor model are also archived. A second forward-pass archive contains per-layer hidden-state radii, three held-out-selected directional budgets, state–update angles, token identifiers, calibration axes, and residual-identity validation errors for all five models. It uses four calibration and 28 evaluation chunks, with 256 tokens per chunk; all completed batch files, the paired angle–magnitude post-processing code, and the aggregate tables underlying Figure 3 are included.