REFERENCE-CENTERED GENERALIZATION BOUNDS FOR MODERN PRE-NORMALIZED DECODERS: EXPLICIT CONSTANTS AND A VACUITY THRESHOLD

Anonymous authors
Paper under double-blind review

ABSTRACT

We derive an explicit reference-centered uniform generalization bound for pre-normalized decoders, including grouped-query attention, rotary maps, RMS normalization, and SwiGLU, and ask when such a bound can be nontrivial at all. The bound has two moving parts: a *transport* coefficient, which carries an early parameter perturbation through the residual stack, and a covering-entropy term. We prove a sharp threshold $\kappa_\star = 0.1085$ for the normalized stage coefficient of the truncated-Dudley objective: at or above it, the optimized certificate must equal its range cap, and in a stronger regime improvements in stage-cover constants drop out of the objective altogether. On a frozen public 1.1B checkpoint, a norm-matched Jacobian audit gives a median $\log_{10}$ transport of 40.58 for the product of separately estimated sublayer norms but only 3.66 for the directly estimated composite. Three SmolLM2 checkpoints (135M–1.7B parameters) reproduce the separation, with median paired gaps of 46.17–54.23 decades against 36.84 on the 1.1B reference; all operator norms in this audit are finite-search lower estimates. We find that, within this covering-and-transport proof structure, the obstruction sits in stage-wise transport rather than in the cover constants, and our results point to direct, direction-aware composition bounds rather than further tightening of stagewise constants.

1 INTRODUCTION

Generalization theory for transformers has developed along several complementary routes. Covering-number and norm-based analyses can remove explicit sequence-length dependence under suitable constraints (Edelman et al., 2022; Trauger & Tewari, 2024; Truong, 2024); offset-complexity arguments sharpen rates and architecture dependence (Li et al., 2026); spectrum-adaptive post-hoc bounds exploit learned singular-value profiles (Sakai & Imaizumi, 2026); and recent work analyzes next-token prediction under explicit language-model distributions (Kong et al., 2026). Reference-centered norm control has a longer lineage in margin and PAC-Bayes analyses (Bartlett et al., 2017; Neyshabur et al., 2018; Long & Sedghi, 2020).

We build a reference-centered uniform-convergence framework whose constants are explicit from end to end, and ask: *once a bound is centered around a trained reference, when can the resulting certificate be nontrivial at all, and which structural condition blocks or enables that possibility?* The hypothesis class is a neighborhood of a reference model fixed independently of the data. Spectral perturbation budgets control hidden-state displacement and mixed-norm budgets control stage entropy. The decoder class is the block used by widely deployed open-weight language models such as the Llama and Mistral families (Touvron et al., 2023; Jiang et al., 2023): causal or sliding-window masking, grouped-query attention, rotary maps, bias-free RMS normalization, SwiGLU, and learned embeddings and readout, with a bounded logit map as the only surrogate.

Our framework separates three mechanisms that are usually folded into a single constant: (i) reference-centering, which shrinks the covering radii (Theorem 2.4, Proposition 2.5); (ii) propagation, which transports early perturbations through the residual stack via stagewise Lipschitz products (Theorem 2.2, Corollary 3.5); and (iii) entropy optimization, which determines whether cover improvements

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.

survive at the optimizer of the Dudley-type objective (Lemma 3.1, Theorem 3.2). Because these mechanisms are algebraically distinct, Section 4 can audit each one separately.

We make five contributions organized as the following:

1. In Section 2, we prove a fully explicit reference-centered parameter-to-state and covering theorem that contains every component of this deployed block simultaneously. A globally Lipschitz extension of RMS normalization (Lemma 2.1) removes the circularity between hidden-state floors and cover centers, and proper mixed-norm covers keep every hybrid matrix in the telescoping argument admissible.
2. In Section 3, instead of reporting that a bound is loose, we prove exactly when it must be vacuous: a sharp threshold $\kappa_\star = 0.1085\ldots$ for the normalized stage coefficient, where $\kappa_N \geq \kappa_\star$ forces the certificate to its range cap, and the stronger condition $\kappa_N \geq 4$ makes K_{loc} algebraically inactive. We also show that the transport coefficient is monotone in the radius-floor profile, so even the most optimistic admissible floor leaves the stage obstruction in place.
3. In Section 4, we instantiate the bound on controlled decoders and on a publicly released 1.1B checkpoint of this architecture. A norm-matched trajectory-Jacobian audit shows that the product of per-sublayer induced-norm lower estimates is still large even after removing zero-floor conditioning and ambient-cap slack.
4. In Section 4.4, we directly estimate the composed 44-stage Jacobian and introduce the alignment defect $\mathfrak{A}_D$, whose magnitude determines where a non-product transport proof could improve the certificate. The median ratio of the two finite-search lower estimates is about 7×10^{36} (36.84 decades) on the 1.1B checkpoint.
5. In Section 4.5, we repeat the audit on three SmolLM2 checkpoints (135M–1.7B), finding median paired gaps of 49.21, 54.23, and 46.17 decades. Their recurrence across two independently pretrained families identifies cross-layer directional information as a concrete target for non-product transport theory.

1.1 Related literature

We use *reference-centered* for the parameter class in (3), in contrast to local Rademacher complexity in the sense of Bartlett et al. (2005) and Koltchinskii (2006), where localization is by variance or empirical-risk level; our generalization rate remains $N^{-1/2}$.

The stage obstruction in Theorem 3.2 applies to any truncated-Dudley objective with an ϵ^{-2} logarithmic stage term. Comparable ϵ^{-2} mechanisms appear in Edelman et al. (2022) and Trauger & Tewari (2024); the spectrum-adaptive framework of Sakai & Imaizumi (2026) allows other entropy exponents. The numerical value of $\kappa_\star$ belongs to the optimized expression (10), so transferring it to another paper requires matching that paper's objective to the form (13). Our result is narrower than the uniform-convergence limitation of Nagarajan & Kolter (2019): it concerns one explicit covering-and-transport proof family, but it is quantitative and can be checked on a given checkpoint. Nonvacuous PAC-Bayes and compression certificates (Dziugaite & Roy, 2017; Zhou et al., 2019; Lotfi et al., 2024a;b) use qualitatively different proof technology and lie outside this family; once a covering proof is in the algebraically inactive regime, further effort on stage-cover constants cannot help until the transport step itself changes. Appendix K gives a detailed positioning table.

Our audit also connects to Lipschitz estimation for deep networks. Self-attention on unbounded domains lacks a global Lipschitz constant (Kim et al., 2021; Castin et al., 2024), which motivates certificates on bounded, normalized streams such as ours. Tight end-to-end Lipschitz constants are hard to compute (Scaman & Virmaux, 2018; Jordan & Dimakis, 2020; Fazlyab et al., 2019), and Khromov & Singh (2024) document a large gap between layerwise products and measured Jacobian norms for image classifiers. Section 4.4 measures the analogous gap on pretrained language models, and Theorem 3.2 shows why it decides whether a covering certificate can be nontrivial.

1.2 Notation

For a row vector $h \in \mathbb{R}^d$, write $\|h\|_\circ = \|h\|_2/\sqrt{d}$ and for a stream $H \in \mathbb{R}^{T\times d}$ write $\|H\|_{\infty,\circ} = \max_t \|H_t\|_\circ$. We use $\|\cdot\|_{\mathrm{op}}$ for spectral norms, $\|\cdot\|_{2,1}$ for the sum of column Euclidean norms, and $\|\cdot\|_{\infty,2\to\infty,2}$ for the induced max-row Euclidean stream norm. We report multiplicative gaps

in *decades*: a gap of x decades is a factor 10^x. Appendix K gives a symbol guide and a mechanism ledger.

2 REFERENCE-CENTERED CLASSES AND A COMPUTABLE BOUND

2.1 DECODER AND NEIGHBORHOOD

A depth-L decoder has $D = 2L$ residual sublayers with RMS normalization

$$\mathcal{N}(h) = \frac{h}{\sqrt{\|h\|_\circ^2 + \nu}}, \qquad \nu > 0, \tag{1}$$

applied before each branch:

$$H^s = H^{s-1} + \lambda_s F_s(\mathcal{N}(H^{s-1}); W_s), \qquad s = 1, \dots, D, \tag{2}$$

with learned embeddings $H_t^0 = E_{X_t,\cdot}$. Branches alternate masked grouped-query attention and SwiGLU. Fixed rotary maps are orthogonal and therefore norm preserving. Normalization gains can be absorbed once into the first linear maps of each branch; Appendix B states the exact effective constraints.

Fix a reference $\theta^0 = (E^0, (W_j^0)_j, C^0)$ inside ambient matrix caps. Its reference-centered class satisfies

$$\begin{aligned} \max_v \|E_v - E_v^0\|_\circ \le e_E, \qquad & \|W_j - W_j^0\|_{\mathrm{op}} \le \tau_j, \\ \|W_j - W_j^0\|_{2,1} \le R_j^{\mathrm{loc}}, \qquad & \max_v \|C_{\cdot v} - C_{\cdot v}^0\|_2 \le e_C. \end{aligned} \tag{3}$$

The subscript "loc" marks reference-centered quantities throughout. The reference and budgets are fixed independently of the N training sequences; an independently trained pilot reference is therefore admissible. For each branch, Λ_s is an ambient input-Lipschitz bound on normalized streams; for every linear stage j in sublayer $s(j)$, b_j bounds the stage-input Euclidean row norm and ℓ_j converts a stage-output perturbation into branch RMS error (Appendix C).

2.2 RADIUS RECURSION WITHOUT CIRCULARITY

A covering center need not itself obey an assumed hidden-state floor. The following extension removes that circularity.

Lemma 2.1 (Global extension of RMS normalization)**.** *For $c \ge 0$, define*

$$\overline{\mathcal{N}}_c(h) = \frac{h}{\sqrt{\max\{\|h\|_\circ^2, c^2\} + \nu}}. \tag{4}$$

It agrees with $\mathcal{N}$ above radius c, has output RMS radius at most one, and is globally $(c^2+\nu)^{-1/2}$-Lipschitz.

Theorem 2.2 (Reference-centered parameter-to-radius control)**.** *Suppose the reference satisfies $\|H_t^s(X;\theta^0)\|_\circ \ge a_s$ for every X in a specified set $\mathcal{X}$, every position t, and $s = 0, \dots, D$. Choose $0 \le c_s \le a_s$ and define*

$$\omega_s = (c_s^2 + \nu)^{-1/2}, \quad L_s = 1 + \lambda_s \omega_{s-1} \Lambda_s, \quad q_s = \sum_{j:s(j)=s} \ell_j b_j \tau_j. \tag{5}$$

Set $\Delta_0 = e_E$ and recursively $\Delta_s = L_s \Delta_{s-1} + \lambda_s q_s$. If $\Delta_s \le a_s - c_s$ for every s, then uniformly over the reference-centered class and $\mathcal{X}$,

$$\|H^s(X;\theta) - H^s(X;\theta^0)\|_{\infty,\circ} \le \Delta_s, \qquad \|H_t^s(X;\theta)\|_\circ \ge c_s. \tag{6}$$

If instead every $c_s = 0$, the displacement bound always holds and $\max\{a_s - \Delta_s, 0\}$ is a valid induced floor on $\mathcal{X}$.

The statement is deterministic; it becomes a population statement when $\mathcal{X}$ contains the population support. Proofs are in Appendix A.

2.3 REFERENCE-CENTERED LOSS ENTROPY AND GENERALIZATION

Let final logits be $z_t = \mathcal{N}(H_t^D)C$, apply the fixed cap $\psi(z) = c_{\log} \tanh(z/c_{\log})$, and define the sequence-averaged cross-entropy

$$g_\theta(X, Y) = \frac{1}{T} \sum_{t=1}^{T} \left[\log \sum_{v=1}^{V} e^{\psi(z_{t,v})} - \psi(z_{t,Y_t}) \right]. \tag{7}$$

It lies in $[0, B]$, where $B = \log(1 + (V-1)e^{2c_{\log}})$.

Remark 2.3 (Bounded surrogate loss). The cap makes the loss globally bounded; the theorem concerns this surrogate, and the same tanh soft-cap on final logits is used in deployed models such as Gemma 2 (Gemma Team et al., 2024). On the pretrained reference, ordinary cross-entropy is 2.746 versus 3.819 at $c_{\log} = 10$, and varying $c_{\log}$ in $\{5, 10, 20, 40\}$ moves the conditional obstruction by less than one decade (Appendix N).

Let $\Pi_s = \prod_{u=s+1}^{D} L_u$ and define

$$\Pi_{\text{out}} = 2r_{\text{out}}\sqrt{d}\,\omega_D, A_{\text{loc}} = 4e_E\Pi_{\text{out}}\Pi_0, w_j = \Pi_{\text{out}}\Pi_{s(j)}\lambda_{s(j)}\ell_j b_j R_j^{\text{loc}},$$
$$w_C = 2\sqrt{V}\,e_C\sqrt{d}, K_{\text{loc}} = \left(\sum_j w_j^{2/3} + w_C^{2/3} \right)^{3/2}. \tag{8}$$

Let V_{eff} be the number of distinct input tokens in the sample, $P = \sqrt{V_{\text{eff}}\, d}$, and $\mathcal{L} = \log(2k_*) + \log(2NT + 1)$, where k_* bounds every relevant linear output width and V.

Theorem 2.4 (Reference-centered loss entropy and generalization)**.** *For the fixed reference-centered class, at every $\epsilon > 0$,*

$$\log \mathcal{N}_\infty(\mathcal{G}_{\text{loc}}, \epsilon) \le P^2 \log(1 + A_{\text{loc}}/\epsilon) + \frac{1024 K_{\text{loc}}^2 \mathcal{L}}{\epsilon^2}. \tag{9}$$

Define

$$\Phi_N = \min\left\{ \frac{B}{2}, \inf_{0<\alpha\le B} \left[4\alpha + \frac{12}{\sqrt{N}} \left\{ BP\sqrt{\log(1 + A_{\text{loc}}/\alpha)} + 32K_{\text{loc}}\sqrt{\mathcal{L}}\log(B/\alpha) \right\} \right] \right\}. \tag{10}$$

For N independent sequences, with probability at least $1 - \delta$, simultaneously for all θ in the fixed reference-centered class,

$$\mathbb{E}g_\theta \le \frac{1}{N} \sum_{i=1}^{N} g_\theta(X_i, Y_i) + \min\left\{ B,\ 2\Phi_N + 3B\sqrt{\frac{\log(2/\delta)}{2N}} \right\}. \tag{11}$$

Within-sequence positions may be dependent; the statistical sample size is the number of independent sequences N, not NT.

The proof uses proper mixed-norm linear covers and optimized stage-accuracy allocation producing the $2/3$-power aggregation in (8); Appendix D gives the full argument and the matched global specialization.

Proposition 2.5 (What reference-centering changes)**.** *Hold the ambient matrix caps, floor profile, and hence all L_s, Π_s fixed. If the reference-centered budgets $(e_E, (R_j^{\text{loc}})_j, e_C)$ are all multiplied by $\eta \in [0, 1]$, then $A_{\text{loc}}(\eta) = \eta A_{\text{loc}}(1)$ and $K_{\text{loc}}(\eta) = \eta K_{\text{loc}}(1)$, while every Π_s is unchanged.*

Proof. The first identity is immediate from the definition of A_{loc}. Every w_j and w_C scales by η, so homogeneity of the $2/3$-power aggregate gives the second. □

We use this separation throughout the audit: reference-centered neighborhoods shrink the covering radii without touching the multiplicative mechanism that transports perturbations to the output, so a large improvement in K can coexist with an unchanged final objective whenever Π_0 dominates or the optimizer sits at its endpoint.

Threshold	Condition	Consequence for the optimized certificate	Result
Pinning	$\kappa_N \geq 4$	optimizer at $\alpha = B$; K_{loc} drops out of the objective	Lemma 3.1
Vacuity	$\kappa_N \geq \kappa_\star = 0.1085\ldots$	certificate equals its range cap B	Theorem 3.2
Sample size	$N \leq N_\star^{\mathrm{stage}}$	same as $\kappa_N \geq \kappa_\star$; explicit lower bound on $N_\star^{\mathrm{stage}}$	Theorem 3.4

Table 1: The three thresholds of Section 3, from the strongest condition to the most permissive one. A nonvacuous certificate requires $\kappa_N < \kappa_\star$, equivalently $N > N_\star^{\mathrm{stage}}$; this condition is necessary but not sufficient.

3 Vacuity thresholds: pinning, obstruction, and floor ceilings

The optimized objective (10) contains two regimes. A strong one pins the optimizer to $\alpha = B$, making the stage term exactly zero; a weaker one already forces vacuity. Write $\mathcal{L}_N = \log(2k_*) + \log(2NT+1)$ and $F_N(\alpha)$ for the bracketed expression in (10). Table 1 previews the three thresholds of this section.

Lemma 3.1 (Endpoint pinning and algebraic inactivity)**.** *If*

$$\kappa_N := \frac{384 K_{\mathrm{loc}}\sqrt{\mathcal{L}_N}}{B\sqrt{N}} \geq 4, \tag{12}$$

then the minimum of F_N over $(0, B]$ is attained at $\alpha = B$, $\Phi_N = B/2$, the final correction in (11) equals B, and the entire stage term containing K_{loc} is algebraically inactive.

The next theorem is deliberately stated without transformer notation. It is a property of any truncated-Dudley objective with the same logarithmic stage term.

Theorem 3.2 (Universal stage-term obstruction)**.** *Let $B > 0$, let $U : (0, B] \to [0, \infty)$ be arbitrary, and define*

$$\Psi(\kappa) = \min\left\{\frac{B}{2}, \inf_{0<\alpha\leq B}\left[4\alpha + U(\alpha) + B\kappa\log(B/\alpha)\right]\right\}. \tag{13}$$

Let $\kappa_\star \in (0, 4)$ be the unique solution of

$$\kappa_\star\left(1 + \log\frac{4}{\kappa_\star}\right) = \frac{1}{2}, \qquad \kappa_\star = 0.1085299\ldots. \tag{14}$$

Then $\kappa \geq \kappa_\star$ implies $\Psi(\kappa) = B/2$. Conversely, $\Psi(\kappa) < B/2$ implies $\kappa < \kappa_\star$.

Proofs of Lemma 3.1 and Theorem 3.2 are in Appendix G. Applied to Theorem 2.4 with U the embedding term and $\kappa = \kappa_N = 384K_{\mathrm{loc}}\sqrt{\mathcal{L}_N}/(B\sqrt{N})$, Theorem 3.2 shows that a strictly nonvacuous final correction requires $\kappa_N < \kappa_\star$. Lemma 3.1 is the stronger subregime $\kappa_N \geq 4$, in which improvements in K_{loc} disappear from the optimized objective altogether.

Remark 3.3 (Sharpness of $\kappa_\star$). The threshold $\kappa_\star$ is sharp for the scalar objective (13) with $U \equiv 0$. The converse in Theorem 3.2 is only necessary; a nonvacuous regime does exist ($U \equiv 0$, $B = 1$, $\kappa = 0.05$ gives $\Psi \leq 0.269$; Appendix L.4). Since κ_N scales as the square root of the entropy constant in (9), halving that deliberately unoptimized constant would lower every $\log_{10}\kappa_N$ by only 0.15 (Appendix G).

Theorem 3.4 (Sample-size stage-obstruction threshold)**.** *Fix a reference-centered class and therefore B, T, k_* and K_{loc}. Define*

$$N_\star^{\mathrm{stage}} = \sup\{N \geq 1 : \kappa_N \geq \kappa_\star\}, \tag{15}$$

with N treated as a real variable and $\sup\emptyset = 0$. Every integer $1 \leq N \leq N_\star^{\mathrm{stage}}$ has a range-capped final certificate. Moreover, with $\mathcal{L}_1 = \log(2k_*) + \log(2T+1)$,

$$N_\star^{\mathrm{stage}} \geq \frac{384^2\mathcal{L}_1}{\kappa_\star^2 B^2}K_{\mathrm{loc}}^2. \tag{16}$$

The vacuity region in N is larger than the endpoint-pinning region by a factor of at least $(4/\kappa_\star)^2 \approx 1358$.

The proof is in Appendix G; a neighborhood-radius form and an effective stage-count identity appear in Appendix L.

Configuration	$K_{\mathrm{glob}}/K_{\mathrm{loc}}$	$U_N/(B/2)$	gain (%)
Baseline	87.2 ± 1.8	339.6 ± 0.5	2.279
Depth 2	79.3 ± 2.1	243.1 ± 0.7	4.347
Depth 8	90.5 ± 1.5	477.4 ± 0.1	1.165
Width 64	92.8 ± 3.8	235.8 ± 0.5	2.385
Width 256	65.1 ± 1.2	497.1 ± 0.5	2.117
Scaled $L = 8$	88.3 ± 2.8	449.5 ± 0.1	1.313
$\nu = 10^{-3}$	84.6 ± 1.8	301.6 ± 0.7	2.873

Table 2: Reference-centered versus matched global decomposition at $N = 512$ and perturbation scale 10^{-2}; mean $\pm$ SD over three seeds. U_N is the raw objective in (10) before the $B/2$ cap, and gain (%) is the percentage reduction in U_N when the matched global class is replaced by the reference-centered class. All 60 archived final bounds (Appendix I) equal the range cap $B = 21.946$.

3.1 What the zero-floor obstruction is actually measuring

Corollary 3.5 (Zero-floor conditioning and downstream depth)**.** *Take the unconditional zero-floor profile $c_s = 0$. Suppose an active stage j_0 lies in sublayer s_0, with $\kappa_{j_0} = \lambda_{s_0}\ell_{j_0}b_{j_0}R_{j_0}^{\mathrm{loc}} > 0$, and $\lambda_u\Lambda_u \geq \rho > 0$ for all $u > s_0$. Let $m = D - s_0$. Then*

$$K_{\mathrm{loc}} \geq \frac{2r_{\mathrm{out}}\sqrt{d}}{\sqrt{\nu}}\,\kappa_{j_0}\left(1 + \frac{\rho}{\sqrt{\nu}}\right)^m. \tag{17}$$

If $\rho \geq \sqrt{\nu}$, then

$$N_\star^{\mathrm{stage}} \geq \frac{384^2\mathcal{L}_1}{\kappa_\star^2 B^2}\left(\frac{2r_{\mathrm{out}}\sqrt{d}}{\sqrt{\nu}}\kappa_{j_0}\right)^2\left(1 + \frac{\rho}{\sqrt{\nu}}\right)^{2m}, \tag{18}$$

which is at least a constant times $\rho^{2m}\nu^{-(m+1)}$: depth supplies the number of repetitions of a poorly conditioned normalizer near zero radius, while $\nu^{-1/2}$ sets their scale.

We see this distinction experimentally: increasing ν from 10^{-5} to 10^{-3} lowers $\log_{10}\Pi_0$ by about eight decades across the eight residual sublayers of the controlled baseline, as expected from a change in normalization conditioning rather than a new depth phenomenon.

3.2 Floor profiles and monotonicity ceilings

Proposition 3.6 (Coordinatewise floor monotonicity)**.** *If $0 \leq c_s \leq c_s'$ for every stage, then $K_{\mathrm{loc}}(c) \geq K_{\mathrm{loc}}(c')$. Let $\mathcal{S}$ be any set of inputs on which the reference is evaluated and let $a_s^{\mathcal{S}} = \min_{X\in\mathcal{S},t}\|H_t^s(X;\theta^0)\|_\circ$. Any floor profile valid for the entire reference-centered class on a set containing $\mathcal{S}$ must satisfy $c_s \leq a_s^{\mathcal{S}}$. Consequently*

$$K_{\mathrm{loc}}(c) \geq K_{\mathrm{loc}}(a^{\mathcal{S}}). \tag{19}$$

If the right-hand side still has $\kappa_N \geq \kappa_\star$, then no valid population floor profile that also applies to $\mathcal{S}$ can remove the stage obstruction at that sample size.

The proof is in Appendix G. Growing floors can alter the depth scaling; Appendix L proves a boundary for linear growing floors, and our pretrained reference grows far too slowly for the only potentially benign regime.

4 Experimental audit

4.1 Controlled decoders

The controlled source has vocabulary eight and context length 32; the next token follows an order-two Markov rule with probability 0.9 and is uniform otherwise. Seven decoder configurations are tested over three independent seeds. The baseline has $L = 4$, $d = 128$, and $\nu = 10^{-5}$; variations change depth, width, residual scaling, or ν. Appendix I gives all seeds, protocols, and design details.

We see in Table 2 that the global-to-reference-centered K ratio is 65–93 at scale 10^{-2} (and up to 957 at scale 10^{-3}), yet every raw objective exceeds its cap by two orders of magnitude and all 60 archived

Group	η	max $\widehat{D}$	$K_{\mathrm{glob}}/K_{\mathrm{loc}}$	$\log_{10}\Pi_0$ (power / upper)	bound
First Q/K	10^{-3}	0.671	1.32×10^{3}	262.46 / 408.87	B
First Q/K	10^{-2}	0.671	1.33×10^{2}	262.47 / 408.87	B
Last FF	10^{-3}	0.132	8.72×10^{3}	262.46 / 408.87	B
Last FF	10^{-2}	0.133	8.72×10^{2}	262.47 / 408.87	B

Table 3: Pretrained-model replication ($B = 30.373$). $\max \widehat{D}$ is the finite-family normalization diagnostic of Appendix M; $K_{\mathrm{glob}}/K_{\mathrm{loc}}$ is reported in the power channel. Both channels of $\log_{10}\Pi_0$ are shown; every optimized bound equals B.

Group	channel	$\log_{10} K(a^{\mathrm{eval}})$	$\log_{10}(\kappa_{10^6}/\kappa_\star)$	obstructed?
First Q/K	induced upper	351.31	351.12	yes
First Q/K	power estimate	205.48	205.29	yes
Last FF	induced upper	7.73	7.54	yes
Last FF	power estimate	6.44	6.25	yes

Table 4: Most optimistic reference-profile ceiling at $N = 10^6$, $\eta = 10^{-3}$: the candidate floor equals the observed evaluation minima. The last numerical column is the number of decades by which κ_N still exceeds $\kappa_\star$; any positive value means the stage obstruction is active.

final bounds equal B. The endpoint criterion explains: in every condition $\kappa_N/4 > 8 \times 10^{20}$, so the stage-cover term is algebraically inactive and reductions in K cannot move the optimized certificate.

Depth isolates the propagation mechanism: mean $\log_{10}\Pi_0$ is 20.80, 41.43, and 82.99 at $L = 2, 4, 8$, growing approximately exponentially. Residual scaling at $L = 8$ lowers the product by about 9.6 decades; raising ν to 10^{-3} lowers it by about 8 decades. Every candidate floor at 0.25, 0.5, or 0.75 times the empirical reference profile returns the zero-floor case, giving $N_\star^{\mathrm{stage}} > 4.45 \times 10^{47}$. Appendix M gives further diagnostics.

4.2 Pretrained reference on natural text

We instantiate the same framework around the fixed public checkpoint `TinyLlama/TinyLlama-1.1B-Chat-v1.0` (Zhang et al., 2024) ($L = 22$, $d = 2048$, 32 query heads, 4 key/value groups, SwiGLU width 5632). No model training is performed. The protocol fixes 32 WikiText-103 (Merity et al., 2017) validation records for calibration and 64 disjoint records for evaluation, with sequence length 128. Only the first-layer attention Q/K pair or the final-layer feedforward gate/up pair is perturbed; all other weights remain fixed. Appendix J gives the exact revisions and numerical-audit protocol. We report two channels: the *induced-upper* channel uses the certified bound $\|M\|_{\mathrm{op}} \le (\|M\|_1\|M\|_\infty)^{1/2}$, while the *power* channel substitutes power-iteration estimates that favor the certificate.

Our results yield a covering-coefficient reduction of two to three orders of magnitude without changing the endpoint regime (Table 3). The normalization diagnostic stays well below one (largest value 0.671), and $B = 30.373$ is $7.95\times$ the measured reference loss of 3.819. At $\eta = 10^{-3}$, the induced-norm channel gives $\log_{10} N_\star^{\mathrm{stage}} > 823$ for first-layer Q/K and > 23.9 for the last feedforward class. At $N = 10^6$, breaking the stage obstruction would require further budget contraction below $10^{-408.7}$ for first-layer Q/K.

We can now test the floor-profile route while using empirical minima only as ceilings. Let a_s^{eval} be the minimum reference RMS radius at stage s over the 64 evaluation sequences. By Proposition 3.6, any valid population floor must lie coordinatewise below a^{eval}, so the certificate recomputed at the optimistic ceiling $c = a^{\mathrm{eval}}$ can only underestimate the obstruction.

We see in Table 4 that even this optimistic ceiling leaves the obstruction far above $\kappa_\star$ in every channel, ruling out the floor-profile route for these classes. The observed reference profile also grows far too slowly for the linear-floor regime of Appendix L: the largest admissible slope from the stage-zero radius 0.0055 is 1.57×10^{-3}, versus a branch Lipschitz bound of about 1.45×10^{5}. The remaining question is therefore transport.

$\log_{10}$ cumulative product	min	median	max
Zero-floor ambient caps (induced upper)	408.87	408.87	408.87
Zero-floor ambient caps (power)	262.46	262.46	262.46
Ambient caps + realized radii (power)	201.95	203.35	204.96
Product of stagewise lower estimates	38.25	40.58	42.36
Direct composite lower estimate	3.03	3.66	4.39
Paired log-gap (recordwise)	34.85	36.84	38.65

Table 5: Trajectory transport on 64 evaluation records, accumulated over all 44 residual sublayers. The last two rows compare the stagewise product with the directly estimated composed Jacobian (Section 4.4). The paired log-gap is computed record by record, so its median (36.84) need not equal the difference of the two medians ($40.5751 - 3.6615 = 36.91$).

For the early Q/K group, the final displacement is 44–96$\times$ the immediate post-attention displacement (median ≈ 70); for the final feedforward group the two coincide. This amplification is tiny relative to $\Pi_0 \approx 10^{262}$ in the power channel, localizing a major source of conservatism in the present certificate. Further finite-family diagnostics appear in Appendix J.

4.3 Norm-matched trajectory transport

On the same checkpoint and records, we write each residual sublayer as $\Phi_s(H) = H + F_s(\mathcal{N}(H))$ and estimate $\|D\Phi_s\|_{\infty,2\to\infty,2}$, the induced stream norm appearing in the theorem ($\sqrt{d}$ cancels between $\|\cdot\|_{\infty,\circ}$ and $\|\cdot\|_{\infty,2}$). A deterministic block-power iteration at the reference trajectory gives finite-restart lower estimates; Appendix J.5 gives the protocol.

We see three successive reductions in Table 5. Replacing the zero-floor normalization factor by the actual reference radii removes about 59 median decades. The radius-plug-in product still exceeds the block-Jacobian lower estimate by about 163 decades, reflecting ambient-cap conservatism. Yet the stagewise realized product remains enormous (median $10^{40.58}$, largest single-stage factor 421.54 at layer-2 feedforward), so replacing each ambient constant by the exact per-sublayer Jacobian norm cannot rescue a proof that multiplies those norms independently.

4.4 Directly composed transport and the alignment defect

We now also estimate the composed operator $J^{(r)}_{44:1} = J^{(r)}_{44} \cdots J^{(r)}_{1}$ by propagating JVP and VJP through all 44 stages with logarithmic rescaling (Appendix J.6). Our results yield a median direct composite of $10^{3.66}$ versus a stagewise product of $10^{40.58}$; the median per-record gap is 36.84 decades (Table 5; Figure 2 in Appendix J.6). Even the optimistic tail substitution of Appendix N leaves first-layer Q/K roughly 46–49 decades above the sharp threshold, so the composite estimate marks where a better proof could gain while the existing certificate remains obstructed.

Definition 4.1 (Alignment defect)**.** For the derivative operators $J_s = D\Phi_s(H^{s-1}_{\mathrm{ref}})$ acting on the max-row Euclidean stream space, define

$$\mathfrak{A}_D = \frac{\prod_{s=1}^{D} \|J_s\|_{\infty,2\to\infty,2}}{\|J_D \cdots J_1\|_{\infty,2\to\infty,2}} \geq 1. \tag{20}$$

A large alignment defect can arise when the separately maximizing input/output directions at each stage cannot be realized by a single path through the composition.

The 36.84-decade gap compares two lower estimates; it estimates $\mathfrak{A}_D$ rather than bounding it. Appendix J.6 (Remark J.1) discusses the distinction between the full-stream composite $J_{44:1}$ and the selected-stage tail $J_{44:2}$ that would enter K_{loc}.

4.5 Cross-checkpoint robustness of the composition gap

We repeated the same norm-matched stagewise-versus-composite audit on three public SmolLM2 base checkpoints (Ben Allal et al., 2025) using 32 frozen WikiText records per model. No training or weight perturbation is performed; all operator quantities are finite-search lower estimates.

Checkpoint	D	records	stagewise product	composite / paired gap
TinyLlama-1.1B	44	64	40.58	3.66 / 36.84
SmolLM2-135M	60	32	53.66	4.33 / 49.21
SmolLM2-360M	64	32	58.75	4.49 / 54.23
SmolLM2-1.7B	48	32	51.27	5.04 / 46.17

Table 6: Cross-checkpoint robustness of the estimated stagewise-versus-composite separation. The last two columns report medians of per-record $\log_{10}$ quantities; the paired gap is computed recordwise. All operator quantities are finite-search lower estimates. Appendix O gives checkpoint revisions, convergence traces, and short-window outputs.

We see in Table 6 that the median paired gaps are 49.21, 54.23, and 46.17 decades for the 135M, 360M, and 1.7B checkpoints, compared with 36.84 on TinyLlama. The separation holds record by record (smallest gaps 48.30, 52.47, 41.99 decades) and already appears within the first eight sublayers (6.76, 7.15, 7.33 decades on the diagnostic record; Appendix O). We read this as a cross-checkpoint robustness audit rather than a scaling law.

5 Concluding remarks

Theorem 3.2 and Lemma 3.1 together explain why order-of-magnitude improvements in covering coefficients can leave the final certificate unchanged: once $\kappa_N \geq \kappa_\star$, the certificate equals its range cap regardless of K_{loc}. Corollary 3.5 locates the source of the explosion in repeated use of $\nu^{-1/2}$ at small hidden-state radii, compounded by depth; this is why changing ν has a multi-decade effect on Π_0 without altering the qualitative regime. A growing radius floor can in principle convert exponential transport into polynomial transport (Appendix L), but Proposition 3.6 closes that route for every pretrained class we study.

In both the controlled and pretrained experiments, Π_0 is enormous, the floor route is closed, and the product of per-sublayer Jacobian lower estimates is still $10^{40.6}$ in median. Across four checkpoints from two independently pretrained families, the estimated direct composite sits 37–54 decades below the stagewise product, showing where direction-aware transport may gain what stagewise tightening cannot. The theoretical target is therefore a non-product or direction-aware transport inequality. For a fixed operator product, $\mathfrak{A}_D$ (Definition 4.1) is the alignment question; families of products lead to joint-spectral-radius questions (Jungers, 2009), and random products invite Lyapunov-growth viewpoints (Hanin & Nica, 2020). Rigorous exploitation requires a certified upper bound on the composed downstream derivatives over the parameter neighborhood that tracks how perturbation directions evolve.

Table 7 distills the audit across all three mechanisms.

Mechanism	Controls	Audit finding	Verdict
Reference-centering	K_{loc}	65–957× reduction	inactive at optimizer
Stagewise transport	Π_0	$10^{262} \to 10^{40.6}$ (realized)	dominates certificate
Composed transport	$\mathfrak{A}_D$	37–54 decade gap	room for direction-aware theory

Table 7: Mechanism-by-mechanism verdict. Reference-centering shrinks covering radii but the stage-cover term is pinned. Stagewise transport is the binding constraint. The stagewise-to-composite gap quantifies how much a non-product proof could gain.

Within this proof family, improving constants gains nothing because the optimizer is pinned; improving floor profiles gains nothing because the observed radii already cap them; and improving per-layer Lipschitz factors gains nothing because composition stays multiplicative. The only route that can work is one that tracks how perturbation directions evolve across layers, and the 37–54 decade gap quantifies the room available to such a theory. Our result concerns the truncated-Dudley proof family of Theorem 3.2; PAC-Bayes and compression certificates use different proof technology. The theorem covers the capped cross-entropy surrogate; unbounded cross-entropy would require further tail arguments. All operator quantities in Section 4 are finite-search lower estimates on frozen trajectories, so the stagewise-to-composite ratios estimate rather than certify $\mathfrak{A}_D$. The certificate instantiation uses one checkpoint and one frozen corpus; the SmolLM2 audit extends only the transport diagnostic.

REPRODUCIBILITY STATEMENT

The supplementary repository contains generating notebooks, machine-readable versions of all tables, and scripts that regenerate the principal figures and verify all archived claims. The controlled study is backed by a 105-run archival rerun; the pretrained study by a frozen FP32 archive around the public checkpoint, including protocol and sequence hashes, all calibration and evaluation records, realized perturbation audits, sign-bootstrap outputs, and bound summaries. The trajectory-transport archive contains all 2,816 record–sublayer Jacobian evaluations, and the composite-Jacobian archive contains the separately locked 64-record full-chain audit with 85 SHA256-verified files. The SmolLM2 breadth audit covers three additional checkpoints, 32 records per model, 96 composite evaluations, 5,504 residual-sublayer evaluations, and a manifest with 117 verified entries; Appendix O gives exact revisions, hashes, convergence checks, and short-window outputs.

AI USE STATEMENT

Generative AI tools were used to restructure the manuscript, review mathematical statements, assist with proof organization, develop and debug experimental code, audit archived outputs, and prepare the reproducibility package. Training and experimental execution were performed by the authors in the recorded notebook environments. The authors reviewed all AI-assisted content and take full responsibility for all mathematical claims, numerical values, and citations.

REFERENCES

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4895–4901, 2023.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005. doi: 10.1214/009053605000000282.

Peter L. Bartlett, Dylan J. Foster, and Matus J. Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6240–6249, 2017.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Vaibhav Srivastav, Joshua Lochner, Clémentine Fourrier, Hugo Larcher, Ben Burtenshaw, Haojun Zhao, Caleb Fahlgren, Mathieu Morlon, Agustín Piqueres Lajarín, Xuan-Son Nguyen, Cyril Zakka, Colin Raffel, Leandro von Werra, and Thomas Wolf. SmolLM2: When smol goes big – data-centric training of a fully open small language model. In *Conference on Language Modeling*, 2025.

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.

Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence*, 2017.

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.

Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George J. Pappas. Efficient and accurate estimation of Lipschitz constants for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, pp. 11423–11434, 2019.

Gemma Team, Morgane Riviere, Shreya Pathak, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Boris Hanin and Mihai Nica. Products of many large random matrices and gradients in deep neural networks. *Communications in Mathematical Physics*, 376(1):287–322, 2020. doi: 10.1007/s00220-019-03624-z.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Matt Jordan and Alexandros G. Dimakis. Exactly computing the local Lipschitz constant of ReLU networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7344–7353, 2020.

Raphaël Jungers. *The Joint Spectral Radius: Theory and Applications*, volume 385 of *Lecture Notes in Control and Information Sciences*. Springer, 2009.

Grigory Khromov and Sidak Pal Singh. Some fundamental aspects about Lipschitz continuity of neural networks. In *International Conference on Learning Representations*, 2024.

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.

Vladimir Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006. doi: 10.1214/009053606000001019.

Insung Kong, Niklas Dexheimer, and Johannes Schmidt-Hieber. Generalization bounds for transformer-based next-token prediction in a language model. *arXiv preprint arXiv:2606.13280*, 2026.

Yawen Li, Tao Hu, Zhouhui Lian, Wan Tian, Yijie Peng, Huiming Zhang, and Zhongyi Li. Sharper generalization bounds for transformer. *arXiv preprint arXiv:2603.21541*, 2026.

Philip M. Long and Hanie Sedghi. Generalization bounds for deep convolutional neural networks. In *International Conference on Learning Representations*, 2020.

Sanae Lotfi, Marc Anton Finzi, Yilun Kuang, Tim G. J. Rudner, Micah Goldblum, and Andrew Gordon Wilson. Non-vacuous generalization bounds for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32801–32818, 2024a.

Sanae Lotfi, Yilun Kuang, Brandon Amos, Micah Goldblum, Marc Finzi, and Andrew Gordon Wilson. Unlocking tokens as data points for generalization bounds on larger language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 9229–9256, 2024b. doi: 10.52202/079017-0293.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017.

Vaishnavh Nagarajan and J. Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. In *Advances in Neural Information Processing Systems*, volume 32, pp. 11611–11622, 2019.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018.

Mana Sakai and Masaaki Imaizumi. Spectrum-adaptive generalization bounds for trained deep transformers. *arXiv preprint arXiv:2605.07297*, 2026.

Kevin Scaman and Aladin Virmaux. Lipschitz regularity of deep neural networks: Analysis and efficient estimation. In *Advances in Neural Information Processing Systems*, volume 31, pp. 3839–3848, 2018.

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

Hugo Touvron, Louis Martin, Kevin Stone, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

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.

Lan V. Truong. On rank-dependent generalisation error bounds for transformers. *arXiv preprint arXiv:2410.11500*, 2024.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. TinyLlama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024.

Tong Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2:527–550, 2002.

Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P. Adams, and Peter Orbanz. Non-vacuous generalization bounds at the ImageNet scale: A PAC-bayesian compression approach. In *International Conference on Learning Representations*, 2019.

APPENDIX OVERVIEW

Appendices A–H contain the theory. Appendix A proves Lemma 2.1 and Theorem 2.2; Appendices B and C define the decoder class and its branch and stage certificates; Appendices D and E give the proper covers, the truncated entropy bound, and the proof of Theorem 2.4; Appendix F treats depth and radius growth; Appendix G proves the threshold results of Section 3; and Appendix H details the stage-tail counterfactual and the logit-cap caveats. Appendices I, J, and O give protocols and provenance for the controlled study, the TinyLlama audits, and the SmolLM2 audit. Appendix K collects a positioning table, a mechanism ledger, and a symbol guide; Appendix L contains supplementary theory; and Appendices M and N contain additional controlled and pretrained results.

A PROOFS FOR THE REFERENCE-CENTERED RADIUS CERTIFICATE

A.1 PROOF OF LEMMA 2.1

Proof. The output radius is $\phi(r) = r/\sqrt{\max\{r^2, c^2\} + \nu}$. The function is continuous, vanishes at zero, and has derivative b_c below c and $\nu(r^2+\nu)^{-3/2} \le b_c$ above c. Also $\phi(r) \le b_c r$. If two inputs have radii r, s and angular cosine u, their squared output distance in RMS units is

$$[\phi(r) - \phi(s)]^2 + 2\phi(r)\phi(s)(1-u) \le b_c^2[(r-s)^2 + 2rs(1-u)].$$

This proves the global Lipschitz bound, including zero inputs by continuity. The agreement and output-radius assertions follow directly from the formula. □

The argument proves a two-point Lipschitz estimate on the whole space. A derivative bound restricted to the exterior of a ball would not by itself justify integration along a segment crossing that ball. The extension removes precisely this issue, and $\nu > 0$ handles $c = 0$ without a singular limit.

A.2 PROOF OF THEOREM 2.2

Proof. Fix the proposed floor sequence and replace the normalizer before branch s by $\overline{\mathcal{N}}_{c_{s-1}}$. Denote the resulting streams by $\overline{H}^s(\theta)$. The reference satisfies $a_s \ge c_s$ on $\mathcal{X}$, so induction gives $\overline{H}^s(\theta^0) = H^s(\theta^0)$ there. Write

$$e_s = \sup_{\theta \in \Theta_{\mathrm{loc}}, X \in \mathcal{X}} \left\| \overline{H}^s(X;\theta) - H^s(X;\theta^0) \right\|_{\infty,\circ}.$$

At initialization $e_0 \le e_E$. Both normalized streams have RMS radius at most one. For fixed θ, X, split the branch difference as

$$\begin{aligned} &F_s(\overline{\mathcal{N}}_{c_{s-1}}(\overline{H}^{s-1}(\theta)); W_s) - F_s(\overline{\mathcal{N}}_{c_{s-1}}(H^{s-1}(\theta^0)); W_s^0) \\ &= F_s(\overline{\mathcal{N}}_{c_{s-1}}(\overline{H}^{s-1}(\theta)); W_s) - F_s(\overline{\mathcal{N}}_{c_{s-1}}(H^{s-1}(\theta^0)); W_s) \\ &\quad + F_s(\overline{\mathcal{N}}_{c_{s-1}}(H^{s-1}(\theta^0)); W_s) - F_s(\overline{\mathcal{N}}_{c_{s-1}}(H^{s-1}(\theta^0)); W_s^0). \end{aligned}$$

The first difference is at most $\Lambda_s \omega_{s-1} e_{s-1}$. For the second, telescope one matrix at a time. At every replacement the linear input has Euclidean row norm at most b_j, so its linear output changes by at most $b_j \tau_j$. The stage certificate gives branch error at most $\ell_j b_j \tau_j$. The ambient caps hold for all the hybrid matrices, even when a coupling among the original parameters has been relaxed. Summing the errors yields q_s. Including the residual connection,

$$e_s \le (1 + \lambda_s \Lambda_s \omega_{s-1}) e_{s-1} + \lambda_s q_s,$$

whence $e_s \le \Delta_s$ by induction.

For every position, reverse triangle now gives

$$\left\| \overline{H}_t^s(X; \theta) \right\|_\circ \ge \left\| H_t^s(X; \theta^0) \right\|_\circ - \Delta_s \ge a_s - \Delta_s \ge c_s.$$

This is established for the extended network before any claim of agreement for the candidate. Agreement with the original candidate then follows by induction, beginning with the shared embedding stream. Thus the stream estimate holds for original networks. A second application of reverse triangle gives the radius-profile estimate.

If every $c_s = 0$, then $\overline{\mathcal{N}}_0 = \mathcal{N}$ everywhere. The error recursion is valid without any feasibility test, and reverse triangle with nonnegativity of the radius gives the stated floor. □

The recursion has the explicit solution

$$\Delta_s = e_E \prod_{u=1}^{s} L_u + \sum_{v=1}^{s} \lambda_v q_v \prod_{u=v+1}^{s} L_u. \tag{21}$$

The empty product equals one. A proposed positive-floor sufficient condition can be checked by finite arithmetic once its ambient caps and reference floors are known. When that sufficient condition is not met, the zero-floor calculation remains valid. One may obtain floors from the zero-floor calculation and recompute a positive-floor candidate, with each candidate checked against its own displayed feasibility inequalities.

A.3 A NONEMPTY NEIGHBORHOOD WITH POPULATION RADIUS CONTROL

The result is not limited to an assumed trained-model floor. Fix any embedding table E^0 with every row of RMS radius $a > 0$, set all reference branch matrices to zero, and take any reference readout within its cap. Every reference stream has radius a on every token sequence, so $a_s = a$ is known without a distributional assumption. Take $B_0 \ge a$ and choose the candidate floors $c_s = a/2$. Put $b = ((a/2)^2 + \nu)^{-1/2}$ and $Q = \sum_s \lambda_s \Lambda_s$. Equation (21) and $1 + x \le e^x$ imply

$$\Delta_s \le e^{bQ} \left(e_E + \sum_{v=1}^{D} \lambda_v q_v \right) \quad (0 \le s \le D).$$

Consequently the explicit sufficient condition

$$e_E + \sum_{v=1}^{D} \lambda_v \sum_{j : s(j) = v} \ell_j b_j \tau_j \le \frac{a}{2} e^{-bQ} \tag{22}$$

certifies all floors $a/2$. With positive ambient budgets this permits nonzero branch perturbations, and the reference itself is always included. This example proves that the positive-floor reference-centered theorem has concrete nonempty instances. It does not establish a useful certificate size for trained language models; the right side can be small at large depth.

A.4 From raw weights and gains to effective budgets

Suppose an effective first matrix is $A = \mathrm{diag}(\gamma)W$, with reference $A^0 = \mathrm{diag}(\gamma^0)W^0$. If $\|\gamma\|_\infty \le G$ and $\|\gamma - \gamma^0\|_\infty \le e_\gamma$, then

$$\begin{aligned}\|A - A^0\|_{\mathrm{op}} &\le G\|W - W^0\|_{\mathrm{op}} + e_\gamma\|W^0\|_{\mathrm{op}},\\ \|A - A^0\|_{2,1} &\le G\|W - W^0\|_{2,1} + e_\gamma\|W^0\|_{2,1}.\end{aligned} \tag{23}$$

Indeed, $A - A^0 = \mathrm{diag}(\gamma)(W - W^0) + \mathrm{diag}(\gamma - \gamma^0)W^0$; apply the triangle inequality and the operator norm of a diagonal matrix, column by column for the mixed norm. Thus raw parameter neighborhoods imply explicit effective budgets in (3). Output matrices do not absorb a second gain.

For a tied readout $C = \mathrm{diag}(\gamma_f)E^\top$ with a correspondingly tied reference, bounds $\|\gamma_f\|_\infty \le G_f$, $\|\gamma_f - \gamma_f^0\|_\infty \le e_f$, and $\max_v \left\|E_v^0\right\|_\circ \le B_0^0$ imply

$$\max_v \|C_{\cdot v} - C^0_{\cdot v}\|_2 \le G_f\sqrt{d}\, e_E + e_f\sqrt{d}B_0^0. \tag{24}$$

Use this as e_C and the total cap $r_{\mathrm{out}} = G_f\sqrt{d}B_0$. The independent readout cover contains the tied class and is an upper-bound relaxation; it is not a claim of a sharp capacity cost for tying.

B The decoder class and its constraints

This section defines the ambient global class. Reference-centered classes intersect its individual matrix and embedding constraints with the reference-centered budgets in (3); the translated reference-centered sets need not contain zero.

B.1 Residual stream, gains, and the output loss

The vocabulary has size $V \ge 2$. Inputs $X = (X_1, \dots, X_T)$ are token sequences, and targets $Y = (Y_1, \dots, Y_T)$ lie in the same vocabulary. An embedding matrix $E \in \mathbb{R}^{V\times d}$ initializes $H_t^0 = E_{X_t,\cdot}$, with

$$\max_{v\le V} \|E_{v,\cdot}\|_\circ \le B_0.$$

There are $D = 2L$ residual sublayers, alternating attention and gated feedforward branches:

$$H^s = H^{s-1} + \lambda_s F_s(\mathcal{N}(H^{s-1})), \qquad s = 1, \dots, D. \tag{25}$$

The coefficients $\lambda_s \ge 0$ are fixed, with $\lambda_s = 1$ giving the usual unscaled residual update. Maps act rowwise except attention. There are no additive biases, dropout variables, or learned position maps in the class analyzed here.

Learned normalization gains are included by an exact reparameterization. For an attention gain γ_s replace the first matrices by

$$A_Q = \mathrm{diag}(\gamma_s)W_Q, \quad A_K = \mathrm{diag}(\gamma_s)W_K, \quad A_V = \mathrm{diag}(\gamma_s)W_V.$$

For a gated branch use $A_g = \mathrm{diag}(\gamma_s)W_g$ and $A_u = \mathrm{diag}(\gamma_s)W_u$. The output matrices W_O, W_d are unchanged. A raw bound $\|W\|_{2,1} \le R$, $\|W\|_{\mathrm{op}} \le S$ and $\|\gamma_s\|_\infty \le G_s$ implies effective bounds G_sR, G_sS on these first matrices. The gain is absorbed once.

The final logits are

$$z_t = \mathcal{N}(H_t^D)C, \qquad \max_{v\le V} \|C_{\cdot v}\|_2 \le r_{\mathrm{out}}. \tag{26}$$

For a tied readout, $C = \mathrm{diag}(\gamma_f)E^\top$; thus $r_{\mathrm{out}} = G_f\sqrt{d}B_0$ is valid when $\|\gamma_f\|_\infty \le G_f$. We cover the larger class in which effective matrices and C vary independently within their constraints. This relaxation contains the original gained and tied model, so upper bounds apply to it. It does not preserve lower bounds for independently varying readouts.

Fix a cap parameter $c_{\log} > 0$ and apply $\psi(x) = c_{\log}\tanh(x/c_{\log})$ to every logit. The sequence loss is

$$g_\theta(X, Y) = \frac{1}{T}\sum_{t=1}^{T}\left[\log\sum_{v=1}^{V} e^{\psi(z_{t,v})} - \psi(z_{t,Y_t})\right]. \tag{27}$$

The cap is fixed and 1-Lipschitz. Cross-entropy is 2-Lipschitz in the maximum logit norm, since its gradient has ℓ_1 norm at most two. In particular

$$0 \le g_\theta \le B, \qquad B = \log\bigl(1 + (V-1)e^{2c_{\log}}\bigr). \tag{28}$$

No independence between the T summands is assumed.

Assumption B.1 (Fixed constraints and population floors)**.** All effective matrix norm bounds, B_0, r_{out}, and λ_s are deterministic and fixed before the training sample is observed. Each matrix belongs to a nonempty bounded constraint set containing zero, with a specified mixed-norm budget R_j and the spectral or column bounds stated below. For a deterministic profile $c_0, \dots, c_D \ge 0$, the original model class is restricted to parameters satisfying

$$\|H_t^s(X;\theta)\|_\circ \ge c_s \quad \text{for all } s, t, \theta \text{ and every } X \text{ in the population input support.} \tag{29}$$

The choice $c_s = 0$ is always valid and covers the complete bounded parameter class. Positive floors give a potentially smaller class and stronger constants. They are additional assumptions, not consequences of normalization or empirical minimum radii.

For the proof, replace the normalizer before branch s by $\overline{\mathcal{N}}_{c_{s-1}}$ and the final normalizer by $\overline{\mathcal{N}}_{c_D}$. Remove the hidden-state floor restrictions from this *extended* parameter class. Induction on s shows that every original model satisfying (29) agrees with its extended counterpart on the population support. Every extended normalized stream has RMS radius at most one.

B.2 Grouped-query attention

We use the grouped-query parameterization of Ainslie et al. (2023) and permit fixed orthogonal rotary maps as in Su et al. (2024).

There are H query heads, J key/value groups, $H = gJ$, and $d = Hd_h$. The group of head h is $\mu(h)$, with exactly g heads in each group. Write

$$A_Q \in \mathbb{R}^{d\times d}, \quad A_K, A_V \in \mathbb{R}^{d\times Jd_h}, \quad W_O \in \mathbb{R}^{d\times d}.$$

Subscripts h or μ indicate column blocks of width d_h. At a normalized input Z, head h has queries $ZA_{Q,h}$, keys $ZA_{K,\mu(h)}$, and values $ZA_{V,\mu(h)}$. Fixed orthogonal rotary maps may be applied to queries and keys at each position.

For each query position i, a deterministic nonempty set $\mathcal{J}_i \subseteq \{1, \dots, T\}$ specifies visible keys. The logits on that set are the rotary query–key inner products divided by $\sqrt{d_h}$, optionally followed by a fixed scalar 1-Lipschitz score cap. Rowwise softmax gives probabilities $P_h(i,j)$. With $U_h = P_h ZA_{V,\mu(h)}$, the branch is

$$F_{\text{att}}(Z) = [U_1|\cdots|U_H]W_O.$$

The causal mask is a special case. The normalizer in each attention row is taken only over its visible keys.

Fix class-wide bounds

$$\max_h \|A_{Q,h}\|_{\text{op}} \le q, \quad \max_\mu \|A_{K,\mu}\|_{\text{op}} \le k, \quad \|A_{V,\mu}\|_{\text{op}} \le v_\mu, \quad \|W_O\|_{\text{op}} \le S_O.$$

Define

$$\beta_Q = \sqrt{H}q, \quad \beta_K = \sqrt{H}k, \quad \beta_V = \sqrt{H}\max_\mu v_\mu, \quad \Sigma_V = \left(g\sum_{\mu=1}^{J} v_\mu^2\right)^{1/2}. \tag{30}$$

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.3 Gated feedforward branches

For width w the branch is

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

with $A_g, A_u \in \mathbb{R}^{d\times w}$ and $W_d \in \mathbb{R}^{w\times d}$. 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 $\sup_x |\operatorname{swish}'(x)|$: with $\sigma(x) = (1+e^{-x})^{-1}$, $\operatorname{swish}'(x) = \sigma(x) + x\sigma(x)(1-\sigma(x))$ and $|x|\sigma(x)(1-\sigma(x)) \leq |x|e^{-|x|} \leq e^{-1}$. We do not need the optimal scalar Lipschitz constant.

C Branch and parameter perturbation certificates

Two different constants are needed. A branch Lipschitz constant propagates an error in its input stream. A stage sensitivity propagates an approximation error in the output of one linear map, with that map's input held fixed. The latter is measured in maximum Euclidean row norm, whereas branch outputs are measured in maximum RMS row norm.

Lemma C.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. □

Proposition C.2 (Branch Lipschitz constants)**.** *On $\{Z : \|Z\|_{\infty,\circ} \leq 1\}$ the following are valid Lipschitz certificates in maximum RMS row norm:*

$$\Lambda_{\mathrm{att}} = S_O\left[\Sigma_V + 2\sqrt{d_h}\,\beta_Q\beta_K\beta_V\right], \tag{31}$$

$$\Lambda_{\mathrm{ff}} = S_d\left[\kappa_{\mathrm{sw}}\beta_b S_g + \beta_a S_u\right]. \tag{32}$$

Both branches vanish at zero, so their output RMS radius on this domain is at most the corresponding Λ.

Proof. Let $\delta = \|Z - Z'\|_{\infty,\circ}$. For an attention head, query and key perturbations have Euclidean row norm at most $\sqrt{d_h}\beta_Q\delta$ and $\sqrt{d_h}\beta_K\delta$. The score perturbation is therefore at most $2\sqrt{d_h}\beta_Q\beta_K\delta$ in maximum norm. Rotary maps preserve these norms and a 1-Lipschitz score cap cannot increase the difference. By Lemma C.1, changing the probabilities changes a head output by at most $2d_h\beta_Q\beta_K\beta_V\delta$ in Euclidean norm. Concatenating H heads and dividing by $\sqrt{d}$ gives $2\sqrt{d_h}\beta_Q\beta_K\beta_V\delta$.

When the probabilities are held fixed, the value perturbation of a head in group μ is at most $\sqrt{d}\,v_\mu\delta$. The concatenated RMS perturbation is at most $(g\sum_\mu v_\mu^2)^{1/2}\delta = \Sigma_V\delta$. Multiplication by W_O completes (31).

For a gated branch write $a = ZA_g$, $b = ZA_u$, and similarly a', b'. At each row,

$$\|\mathrm{swish}(a)\odot b - \mathrm{swish}(a')\odot b'\|_2 \leq \kappa_{\mathrm{sw}}\beta_b\|a-a'\|_2 + \beta_a\|b-b'\|_2 .$$

Here $|\operatorname{swish}(x)| \leq |x|$, and the activation bounds hold for both inputs. Apply the spectral caps and divide by $\sqrt{d}$. Zero input gives zero values and zero gate outputs. □

Proposition C.3 (Linear stage certificates)**.** *Consider replacing one matrix by another matrix in the same constraint set, with the input to that linear map fixed. Suppose its linear outputs differ by at most η_j in maximum Euclidean row norm. Then the branch output differs by at most $\ell_j\eta_j$ in maximum RMS row norm, with b_j and ℓ_j in Table 8. The number b_j bounds the Euclidean row radius of that stage's fixed input, for every admissible choice of the other matrices.*

Stage matrix	Input radius b_j	Sensitivity ℓ_j
A_Q	$\sqrt{d}$	$S_O\beta_K\beta_V/\sqrt{H}$
A_K	$\sqrt{d}$	$S_O\beta_Q\beta_V$
A_V	$\sqrt{d}$	$S_O/\sqrt{d_h}$
W_O	$\sqrt{d}\,\Sigma_V$	$1/\sqrt{d}$
A_g	$\sqrt{d}$	$S_d\kappa_{\mathrm{sw}}\beta_b/\sqrt{d}$
A_u	$\sqrt{d}$	$S_d\beta_a/\sqrt{d}$
W_d	$\sqrt{d}\,\beta_a S_u$	$1/\sqrt{d}$

Table 8: Linear-stage certificates before multiplication by the residual coefficient λ_s. All quantities are ambient class-wide bounds for the containing sublayer.

Proof. For a query replacement, let the per-head query errors at a fixed query position be η_h. Since all query heads share this position, $\sum_h \eta_h^2 \le \eta_Q^2$. The score error in head h is at most $\beta_K\eta_h$, so its value-weighting error is at most $\sqrt{d_h}\beta_V\beta_K\eta_h$. Concatenation, RMS conversion, and W_O give $S_O\beta_K\beta_V\eta_Q/\sqrt{H}$.

For a key replacement, each head score error is at most $\beta_Q\eta_K$. A head's probability-weighted value error is at most $\sqrt{d_h}\beta_V\beta_Q\eta_K$. Concatenation gives $S_O\beta_Q\beta_V\eta_K$ in RMS norm. In contrast to the query case, different heads can attain their maximum key error at different positions. No interchange of a maximum over positions and a sum over groups is used.

For a value replacement, every head's weighted value error has Euclidean norm at most η_V. The concatenated RMS error is $\sqrt{H/d}\eta_V = \eta_V/\sqrt{d_h}$ before W_O. For the output projection itself the RMS conversion contributes $1/\sqrt{d}$. Its input obeys $\|[U_1|\cdots|U_H]_i\|_2 \le \sqrt{d}\,\Sigma_V$.

The two gate sensitivities follow from the product estimate in the proof of Proposition C.2. The input of W_d has norm at most $\|\mathrm{swish}(a)\|_\infty \|b\|_2 \le \beta_a\sqrt{d}S_u$. Again, its output error is converted from Euclidean to RMS units by $1/\sqrt{d}$. □

The sensitivities hold for hybrid choices of matrices: one matrix may already be a covering center while another is an original parameter. This is why the centers in the next section are selected inside the matrix constraint sets.

D Proper linear covers and propagation coefficients

D.1 A Euclidean-input linear covering lemma

We include a direct proof so that no change of input norm is hidden in an imported entropy estimate. Its greedy approximation step is closely related to the online-learning approach to linear covers Zhang (2002); mixed-norm covers also appear in Edelman et al. (2022).

Lemma D.1 (Proper scalar linear cover). *Fix $x_1,\dots,x_n \in \mathbb{R}^p$ with $\|x_i\|_2 \le b$. For any nonempty set $\mathcal{W} \subseteq \{w : \|w\|_2 \le R\}$, there is a proper η-cover in the metric $\max_i |\langle x_i, w - w'\rangle|$ satisfying, for every $\eta > 0$,*

$$\log|\mathcal{N}| \le 8\left(\frac{Rb}{\eta}\right)^2 \log(2n+1). \tag{33}$$

Proof. The case $Rb = 0$ is trivial, and an arbitrary admissible center works when $\eta \ge 2Rb$. Otherwise set $\delta = \eta/2$. For a target w in the ball, start at $v_0 = 0$. Whenever some i satisfies $|\langle x_i, w - v_t\rangle| > \delta$, update

$$v_{t+1} = v_t + \frac{\delta}{b^2}\,\mathrm{sign}(\langle x_i, w - v_t\rangle)x_i.$$

The squared distance to w decreases by more than δ^2/b^2:

$$\|w - v_{t+1}\|_2^2 = \|w - v_t\|_2^2 - \frac{2\delta}{b^2}|\langle x_i, w - v_t\rangle| + \frac{\delta^2}{b^4}\|x_i\|_2^2 < \|w - v_t\|_2^2 - \frac{\delta^2}{b^2}.$$

The procedure stops after at most $M = \lfloor R^2b^2/\delta^2 \rfloor$ updates. Its final vector approximates every evaluation within δ. All possible final vectors are encoded by at most M choices of an index and a sign. Their number is at most $\sum_{j=0}^{M}(2n)^j \le (2n+1)^M$.

These centers need not belong to $\mathcal{W}$. For every center whose δ-ball meets $\mathcal{W}$, choose one representative from that intersection. The resulting proper cover has radius $2\delta = \eta$ and no more centers. The displayed bound follows, with the stated constant allowing harmless endpoint rounding. □

Lemma D.2 (Proper column mixed-norm cover)**.** *Fix inputs as in Lemma D.1 and a nonempty constraint set $\mathcal{M} \subseteq \{W \in \mathbb{R}^{p\times k} : \|W\|_{2,1} \le R\}$. For every $\eta > 0$ there is a proper cover, in maximum Euclidean output-row norm, with*

$$\log|\mathcal{N}| \le 256\left(\frac{Rb}{\eta}\right)^2[\log(2k) + \log(2n+1)]. \tag{34}$$

Additional spectral, column, or head-block constraints in $\mathcal{M}$ are preserved by its centers.

Proof. As before, $Rb = 0$ or $\eta \ge 2Rb$ is immediate. Write $q = Rb/\eta > 1/2$. The vector of column norms of a matrix in the mixed-norm ball lies in $\Delta_R = \{t \in \mathbb{R}_+^k : \sum_j t_j \le R\}$. For any $t \in \Delta_R$, take a random vector equal to Re_j with probability t_j/R and to zero otherwise. An average of $M = \lceil 16q^2 \rceil$ independent such vectors has mean t and expected squared distance at most R^2/M. Thus the set of all these averages is an $\eta/(4b)$-net of Δ_R in Euclidean norm. Its logarithmic cardinality is at most $M\log(k+1) \le 20q^2\log(k+1)$, and all its members v still satisfy $\sum_j v_j \le R$.

Given a matrix with column norms t, replace each column's radius by the selected v_j while keeping its direction. If a column is zero, choose any fixed unit direction. The resulting change in any output row is at most $b\|t - v\|_2 \le \eta/4$. For this fixed v, cover each column ball of radius v_j by Lemma D.1 at accuracy

$$\eta_j = \frac{\eta}{4}\sqrt{\frac{v_j}{R}} \quad (v_j > 0).$$

The zero columns need no cover. The squared output errors sum to at most $\sum_j \eta_j^2 \le \eta^2/16$, and the logarithmic cardinalities sum to at most $128q^2\log(2n+1)$. All column centers have norm at most v_j, so the resulting matrix lies in the mixed-norm ball.

We have constructed an $\eta/2$-cover of the full mixed-norm ball. For each center whose $\eta/2$-ball meets $\mathcal{M}$, choose a member of that intersection. This doubles the covering radius and preserves every constraint in $\mathcal{M}$. The two entropy costs above are bounded by (34). □

The factor $\log(2n+1)$ is retained. In particular, this proof does not claim sequence-length independence under a Euclidean input bound. Conversely, it needs no extra logarithm of η^{-1}. All constants used below are deliberately unoptimized.

D.2 Residual propagation and optimized stage allocation

For sublayer s, let Λ_s be its certificate in Proposition C.2. Define

$$\omega_s = (c_s^2 + \nu)^{-1/2}, \qquad L_s = 1 + \lambda_s\omega_{s-1}\Lambda_s, \qquad \Pi_s = \prod_{u=s+1}^{D} L_u, \quad 0 \le s \le D. \tag{35}$$

The empty product is one, so $\Pi_D = 1$ and $\Pi_0 = \prod_{u=1}^{D} L_u$. Here ω_s is a normalization Lipschitz constant, while b_j in Table 8 is a linear-stage input radius.

Index all linear stages of all branches by j, and let $s(j)$ denote the containing sublayer. Put

$$\kappa_j = \lambda_{s(j)}\ell_j R_j b_j, \qquad K_{\mathrm{B}} = \left[\sum_j (\Pi_{s(j)}\kappa_j)^{2/3}\right]^{3/2}. \tag{36}$$

Zero terms can be omitted. Let $k_* \geq 2$ be at least the output width of every linear stage, and define

$$\mathcal{L}_{n,k_*} = \log(2k_*) + \log(2n+1).$$

For the complete loss analysis we also require $k_* \geq V$. Let V_{eff} be the number of distinct input tokens in the sample and $P = \sqrt{V_{\text{eff}} d}$.

E Entropy integration and the global specialization

E.1 Truncated entropy integration

Lemma E.1 (Truncated entropy bound)**.** *For a scalar evaluation class $\mathcal{F} \subseteq [0,B]^N$ with external maximum-norm entropy $H(\epsilon) = \log \mathcal{N}_\infty(\mathcal{F}, \epsilon)$, and any $0 < \alpha \leq B$,*

$$\widehat{\mathfrak{R}}_N(\mathcal{F}) \leq 4\alpha + \frac{12}{\sqrt{N}} \int_\alpha^B \sqrt{H(\epsilon)}\, d\epsilon. \tag{37}$$

Also $\widehat{\mathfrak{R}}_N(\mathcal{F}) \leq B/2$.

Proof. Clip covering centers to $[0,B]^N$, which cannot increase their errors. Set $\epsilon_j = B2^{-j}$ and use zero as the single level-zero center. Choose maximum-norm covers at the subsequent scales. For each evaluation vector choose its closest center at each level. Successive centers differ by at most $3\epsilon_j$ in maximum norm, and the number of possible increments has log-cardinality at most $2H(\epsilon_j)$. The exponential-moment bound for a finite set of Rademacher sums gives a contribution at most $6\epsilon_j\sqrt{H(\epsilon_j)/N}$ at level j.

If $\alpha < B/4$, stop at the largest J with $\epsilon_J \geq 2\alpha$. The remaining error is below 4α. Since $\epsilon_j\sqrt{H(\epsilon_j)} \leq 2\int_{\epsilon_{j+1}}^{\epsilon_j} \sqrt{H(u)}\, du$, summing the levels proves (37). For $\alpha \geq B/4$, its 4α term alone suffices. Finally, for every sign vector, $\sup_f \sum_i \varepsilon_i f_i \leq B|\{i : \varepsilon_i = 1\}|$; take expectations and divide by N. □

E.2 The global norm-constrained class

For the class in Assumption B.1, define

$$\Pi_{\text{out}} = 2r_{\text{out}}\sqrt{d}\,\omega_D, \qquad A = 4B_0\Pi_{\text{out}}\Pi_0, \tag{38}$$

$$\kappa_{\text{R}} = 2\sqrt{V}\, r_{\text{out}}\sqrt{d}, \qquad K = \left[\sum_j (\Pi_{\text{out}}\Pi_{s(j)}\kappa_j)^{2/3} + \kappa_{\text{R}}^{2/3}\right]^{3/2}. \tag{39}$$

These coefficients specialize (8) to neighborhoods centered at zero with absolute, rather than local, budgets.

Corollary E.2 (Global norm-class generalization)**.** *For the population-floor class in Assumption B.1, the entropy bound (9), the complexity bound (10), and the generalization bound (11) hold with $A_{\text{loc}} = A$ and $K_{\text{loc}} = K$. If $r_{\text{out}} = 0$, set $A = K = 0$. Fixed embeddings have no embedding entropy term.*

Proof. Set all reference matrices to zero and choose $e_E = B_0$, $R_j^{\text{loc}} = R_j$, and $e_C = r_{\text{out}}$. Make the spectral perturbation constraints redundant with the ambient caps. This is exactly the global parameter class, restricted to any prescribed population floors. Theorem 2.4 allows imposed floors as well as certified ones, so it applies without a positive reference floor. Its proof uses the extended covering class. Substitution gives (38)–(39); the zero cases follow directly. □

For $c_s = 0$ at every stage this covers the complete bounded parameter class. Positive population floors describe a smaller class. They are not consequences of the zero reference. The corollary is a specialization of the same proof, not an additional contraction argument.

E.3 PREDETERMINED CLASS SELECTION

Corollary E.3 (Predetermined countable stratification)**.** *Let $\{\Theta_a : a \in \mathcal{A}\}$ be a countable collection of classes specified before observing the sample, with weights $\pi_a > 0$ and $\sum_a \pi_a \le 1$. Apply Theorem 2.4 to each class with confidence parameter $\delta\pi_a$. With probability at least $1-\delta$, all the resulting bounds hold simultaneously.*

Proof. The probability that at least one class bound fails is at most $\sum_a \delta\pi_a \le \delta$. □

This permits selection among predetermined norm budgets or floor profiles. If each of $D+1$ floor coordinates is chosen from an M-point grid, a uniform prior costs $(D+1)\log M$ in the confidence logarithm. Stratifying a floor alone does not account for arbitrary selection of a reference radius profile or an unrestricted data-dependent neighborhood. Membership in a positive-floor class still concerns the population support.

E.4 PROOF OF THEOREM 2.4

Proof. Fix the observed N sequences and condition on the independently selected reference and all prescribed budgets. We construct a finite external cover of the corresponding loss evaluations.

Embedding step. Only embedding rows appearing in the observed inputs matter. Translate by E^0. Each observed perturbation row lies in an RMS ball of radius e_E, equivalently a Euclidean ball of radius $\sqrt{d}\,e_E$. A proper volumetric cover at RMS accuracy η_0 has logarithmic cardinality at most

$$P^2 \log(1 + 2e_E/\eta_0), \qquad P^2 = V_{\text{eff}}\, d.$$

Unobserved rows may be fixed at their reference values.

Conditional stage covers. Proceed through the residual stack. Once a previously covered stream is fixed, the extended normalizer produces a fixed design with Euclidean row radius at most $\sqrt{d}$. For a linear stage j, translate its local class by W_j^0. The perturbation matrices lie in the mixed-norm ball $\|\Delta W_j\|_{2,1} \le R_j^{\text{loc}}$ intersected with the translated ambient constraints. Lemma D.2, applied to the current fixed stage input of row radius b_j, gives a proper output cover at Euclidean row accuracy η_j with

$$\log|\mathcal{C}_j| \le 256 \left(\frac{R_j^{\text{loc}} b_j}{\eta_j}\right)^2 \mathcal{L}, \qquad \mathcal{L} = \log(2k_*) + \log(2NT+1).$$

Properness means that after adding back W_j^0, every selected center is an admissible matrix. Therefore the branch sensitivity ℓ_j and every downstream ambient bound remain valid for all hybrids created by telescoping one stage at a time.

Let E_s be the maximum RMS error between an original stream and its covered stream after sublayer s. The extension (4) is globally ω_{s-1}-Lipschitz, so the same-parameter input error contributes at most $L_s E_{s-1}$; telescoping the stage covers contributes $\lambda_s \sum_{j:s(j)=s} \ell_j \eta_j$. Hence

$$E_s \le L_s E_{s-1} + \lambda_s \sum_{j:s(j)=s} \ell_j \eta_j,$$

and unrolling gives

$$E_D \le \Pi_0 \eta_0 + \sum_j \Pi_{s(j)} \lambda_{s(j)} \ell_j \eta_j. \tag{40}$$

No positive floor is required for a covering center: the extended network supplies the Lipschitz bound at every scale, while original models in the stated class agree with their extensions on the population support.

Readout and loss. A body RMS error e produces normalized Euclidean row error at most $\sqrt{d}\,\omega_D e$. With readout column norm at most r_{out} and the 2-Lipschitz capped cross-entropy, its loss contribution is at most $\Pi_{\text{out}} e$. Cover the translated readout columns at Euclidean accuracy chosen so that their combined maximum logit error contributes a scalar budget η_C; a proper columnwise cover gives the

same quadratic entropy form with effective radius $w_C = 2\sqrt{V}\, e_C \sqrt{d}$. Thus, if the total loss-cover accuracy allocated to the body stages and readout is $\epsilon/2$, the active coefficients are precisely

$$w_j = \Pi_{\mathrm{out}} \Pi_{s(j)} \lambda_{s(j)} \ell_j b_j R_j^{\mathrm{loc}}, \qquad w_C = 2\sqrt{V}\, e_C \sqrt{d}.$$

Let $S = \sum_j w_j^{2/3} + w_C^{2/3} = K_{\mathrm{loc}}^{2/3}$. Allocate the non-embedding accuracies in proportion to $w^{2/3}$. The standard Lagrange-multiplier calculation gives total quadratic entropy at most

$$\frac{1024 K_{\mathrm{loc}}^2 \mathcal{L}}{\epsilon^2}.$$

Choose the embedding accuracy so that its propagated loss contribution is $\epsilon/2$, namely $\eta_0 = \epsilon/(2\Pi_{\mathrm{out}}\Pi_0)$. Its entropy becomes

$$P^2 \log\left(1 + \frac{4 e_E \Pi_{\mathrm{out}} \Pi_0}{\epsilon}\right) = P^2 \log(1 + A_{\mathrm{loc}}/\epsilon).$$

Adding the conditional log-cardinalities proves (9). If a coefficient is zero, omit that stage; if the entire loss class is constant, a single center suffices.

From entropy to risk. Lemma E.1 applied to the bounded scalar loss class gives, for every $0 < \alpha \le B$,

$$\widehat{\mathfrak{R}}_N(\mathcal{G}_{\mathrm{loc}}) \le 4\alpha + \frac{12}{\sqrt{N}} \int_\alpha^B \sqrt{P^2 \log(1 + A_{\mathrm{loc}}/\epsilon) + 1024 K_{\mathrm{loc}}^2 \mathcal{L}/\epsilon^2}\, d\epsilon.$$

Use $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$, monotonicity of the first logarithm, and $\int_\alpha^B \epsilon^{-1} d\epsilon = \log(B/\alpha)$ to obtain the objective in (10); combine it with the universal $B/2$ Rademacher cap. Standard symmetrization for N independent sequences followed by bounded differences yields

$$\sup_{\theta \in \Theta_{\mathrm{loc}}} \left(\mathbb{E} g_\theta - \frac{1}{N} \sum_{i=1}^N g_\theta(X_i, Y_i) \right) \le 2\Phi_N + 3B \sqrt{\frac{\log(2/\delta)}{2N}}$$

with probability at least $1 - \delta$. Capping by the loss range B gives (11). Dependence among positions inside each sequence is immaterial because symmetrization and concentration are performed at the sequence level. □

F Depth dependence and the role of radius growth

The coefficient K combines two effects: downstream propagation, through the products Π_s, and the number of independently variable stages, through a sum of $2/3$ powers. Controlling the first effect does not automatically control the second.

Proposition F.1 (A uniform upper growth bound)**.** *Every original or extended network in the stated parameter class satisfies*

$$\|H^s\|_{\infty,\mathrm{o}} \le B_0 + \sum_{u=1}^s \lambda_u \Lambda_u. \tag{41}$$

Proof. Every branch receives an input with maximum RMS radius at most one and has output radius at most Λ_u. Apply the triangle inequality to each residual update and iterate. □

With fixed branch bounds and unscaled residuals, this rules out a geometrically growing *lower* radius floor at arbitrarily large depth. It does not prove a linear lower floor: branches may cancel, vanish, or point against the stream. A linear lower floor must be assumed or established by an additional argument.

Proposition F.2 (Propagation under a linear floor)**.** *Suppose $\lambda_s = 1$, $\Lambda_s \le \Lambda$, and $c_s \ge a + \zeta s$ for fixed $a, \zeta > 0$. Writing $p = \Lambda/\zeta$, for $0 \le s \le D$,*

$$\Pi_s \le e^{\Lambda/a} \left(\frac{a + \zeta D}{a + \zeta s} \right)^p. \tag{42}$$

Thus residual propagation is polynomial in depth.

Proof. Use $\log(1+x) \le x$ and $(c_{u-1}^2+\nu)^{-1/2} \le (a+\zeta(u-1))^{-1}$. The decreasing harmonic sum from $u = s+1$ to D is at most $a^{-1} + \zeta^{-1}\log((a+\zeta D)/(a+\zeta s))$. Exponentiate after multiplying by Λ. □

Corollary F.3 (Accumulated loss-cover coefficient)**.** *Under Proposition F.2, suppose d, r_{out} and all other non-depth constants are fixed, there are at most J_0 linear stages per sublayer, and $\kappa_j \le u_0$ uniformly. Let*

$$K_{\text{B,loss}} = \left[\sum_j (\Pi_{\text{out}}\Pi_{s(j)}\kappa_j)^{2/3}\right]^{3/2}.$$

As $D \to \infty$, the following upper bounds hold:

$$K_{\text{B,loss}} = \begin{cases} O(D^{1/2}), & p < 3/2, \\ O(D^{1/2}(\log D)^{3/2}), & p = 3/2, \\ O(D^{p-1}), & p > 3/2. \end{cases} \tag{43}$$

Also $A = O(D^{p-1})$, and $K \le \sqrt{2}(K_{\text{B,loss}} + \kappa_{\text{R}})$.

Proof. The final normalizer gives $\Pi_{\text{out}} \le 2r_{\text{out}}\sqrt{d}/(a+\zeta D)$. Combining this with (42) and the stage bounds yields

$$K_{\text{B,loss}} \le C(a+\zeta D)^{p-1}\left[\sum_{s=1}^{D}(a+\zeta s)^{-2p/3}\right]^{3/2},$$

where C is independent of D. Integral comparison gives (43). The formula for A follows directly from (38); the last inequality follows from $(x^{2/3}+y^{2/3})^{3/2} \le \sqrt{2}(x+y)$ for $x, y \ge 0$. □

The exponents in Corollary F.3 describe coefficients in this bound, conditional on the entire list of fixed-budget and floor assumptions. They are not assertions about trained networks. For an arbitrarily deep nonempty family with these fixed upper growth bounds, (41) also requires $\zeta \le \Lambda$. There is no guarantee that a given family satisfies any positive linear floor.

Proposition F.4 (A residual coefficient budget)**.** *Suppose $c_s \ge c \ge 0$, $\Lambda_s \le \Lambda$, and $\sum_s \lambda_s \le A_0$. Then*

$$\Pi_0 \le \exp\left(\frac{A_0\Lambda}{\sqrt{c^2+\nu}}\right).$$

If each sublayer has at most J_0 stages and $\ell_j R_j b_j \le u_0$ uniformly, then, with all other constants fixed, $K_{\text{B,loss}} = O(A_0\sqrt{D})$.

Proof. The propagation estimate follows by summing $\log L_s \le \lambda_s\Lambda/\sqrt{c^2+\nu}$. For the stage sum, $\kappa_j \le \lambda_{s(j)}u_0$ and

$$\left(\sum_{s=1}^{D}\lambda_s^{2/3}\right)^{3/2} \le \sqrt{D}\sum_{s=1}^{D}\lambda_s \le A_0\sqrt{D}$$

by Hölder's inequality. Bound all tail products and the final output factor by the constants just obtained. □

The last proposition controls propagation independently of depth while leaving a possible $\sqrt{D}$ accumulation of capacity. A depth-free bound would need additional structure, for example shared parameters, a stronger summability constraint, or an argument exploiting cancellations that this product cover ignores. If the coefficients λ_s are learned rather than fixed, their admissible variation must also be covered or incorporated into the output-matrix constraint sets.

G ENDPOINT SATURATION, STAGE-OBSTRUCTION THRESHOLDS, AND FLOOR CEILINGS

Proof of Lemma 3.1. Recall that $F_N(\alpha)$ is the expression inside the infimum in (10). The embedding term is nonincreasing in α. For $0 < \alpha < B$, differentiating the remaining terms gives

$$F'_N(\alpha) \le 4 - \frac{384K_{\mathrm{loc}}\sqrt{\mathcal{L}_N}}{\sqrt{N}\,\alpha} \le 4 - \frac{384K_{\mathrm{loc}}\sqrt{\mathcal{L}_N}}{\sqrt{N}\,B}.$$

Under (12), the last display is nonpositive. Hence F_N is nonincreasing and its minimum over $(0, B]$ occurs at B. At that endpoint the logarithmic stage term vanishes and the remaining terms give $F_N(B) \ge 4B > B/2$. Therefore $\Phi_N = B/2$, and the outer range cap in (11) returns B. □

Proof of Theorem 3.2. Put $\beta = \alpha/B \in (0, 1]$. Since $U(\alpha) \ge 0$,

$$\frac{1}{B}\left[4\alpha + U(\alpha) + B\kappa \log(B/\alpha)\right] \ge h_\kappa(\beta) := 4\beta + \kappa \log(1/\beta).$$

For $0 < \kappa < 4$, the derivative is $h'_\kappa(\beta) = 4 - \kappa/\beta$, so the unique minimizer is $\beta = \kappa/4$ and

$$\min_{0<\beta\le 1} h_\kappa(\beta) = g(\kappa) := \kappa\left(1 + \log\frac{4}{\kappa}\right).$$

For $\kappa \ge 4$, h_κ is nonincreasing on $(0, 1]$ and its minimum is $h_\kappa(1) = 4$. On $(0, 4)$, $g'(\kappa) = \log(4/\kappa) > 0$, while $g(\kappa) \downarrow 0$ as $\kappa \downarrow 0$ and $g(4) = 4$. Hence there is a unique $\kappa_\star \in (0, 4)$ satisfying $g(\kappa_\star) = 1/2$. If $\kappa \ge \kappa_\star$, the infimum in (13) is at least $B/2$, so $\Psi(\kappa) = B/2$. Conversely, if $\Psi(\kappa) < B/2$, then the infimum is below $B/2$, which forces $\min h_\kappa < 1/2$ and therefore $\kappa < \kappa_\star$. Numerical root finding gives $\kappa_\star = 0.1085299426\ldots$. □

Proof of Theorem 3.4. Write $C = 384K_{\mathrm{loc}}/B$. Then $\kappa_N^2 = C^2\mathcal{L}_N/N$. Treating N as a positive real,

$$\frac{d}{dN}\frac{\mathcal{L}_N}{N} = \frac{\frac{2NT}{2NT+1} - \mathcal{L}_N}{N^2} < 0,$$

because $2NT/(2NT+1) < 1$ while $\mathcal{L}_N \ge \log 2 + \log 3 > 1$. Hence κ_N is strictly decreasing and $\{N : \kappa_N \ge \kappa_\star\}$ is an interval. Theorem 3.2 gives the first claim. Since $\mathcal{L}_N \ge \mathcal{L}_1$, every

$$N \le \frac{C^2\mathcal{L}_1}{\kappa_\star^2}$$

satisfies $\kappa_N \ge \kappa_\star$, proving (16). □

Proof of Corollary 3.5. At the zero-floor profile, $\omega_s = \nu^{-1/2}$ for every s and

$$\Pi_{s_0} = \prod_{u=s_0+1}^{D}\left(1 + \frac{\lambda_u\Lambda_u}{\sqrt{\nu}}\right) \ge \left(1 + \frac{\rho}{\sqrt{\nu}}\right)^m.$$

Also $\Pi_{\mathrm{out}} = 2r_{\mathrm{out}}\sqrt{d}/\sqrt{\nu}$. Since the $2/3$ aggregate K_{loc} dominates each individual stage weight, $K_{\mathrm{loc}} \ge \Pi_{\mathrm{out}}\Pi_{s_0}\kappa_{j_0}$, giving (17). Substitution into (16) yields (18). If $\rho \ge \sqrt{\nu}$, then $1 + \rho/\sqrt{\nu} \ge \rho/\sqrt{\nu}$; after squaring, the explicit Π_{out}^2 contributes one additional factor ν^{-1}, giving the stated $\rho^{2m}\nu^{-(m+1)}$ dependence up to constants independent of m and ν. □

Proof of Proposition 3.6. If $c_s \le c'_s$, then $\omega_s(c) = (c_s^2 + \nu)^{-1/2} \ge \omega_s(c')$. Thus every residual factor $L_{s+1} = 1 + \lambda_{s+1}\Lambda_{s+1}\omega_s$, every tail product Π_s, and the output factor $\Pi_{\mathrm{out}} = 2r_{\mathrm{out}}\sqrt{d}\,\omega_D$ are coordinatewise nonincreasing as the floor profile increases. Every stage weight w_j and hence their $2/3$ aggregate K_{loc} inherit the same monotonicity.

If a floor profile is valid for the full reference-centered class on a set containing $\mathcal{S}$, it is in particular valid for the reference θ^0 on every $X \in \mathcal{S}$. Therefore $c_s \le a_s^{\mathcal{S}}$ for every s. Apply the coordinatewise monotonicity to obtain (19). The final obstruction claim then follows from Theorem 3.2. □

Constants ledger for the proper-cover term. Lemma D.1 contributes the constant 8 for a proper scalar linear cover. The column-radius discretization, per-column scalar covers, and properization in Lemma D.2 give 256 for a proper mixed $2,1$ matrix cover. In the loss cover, the body/readout portion is assigned accuracy $\epsilon/2$; squaring the resulting factor of two turns 256 into 1024 in (9). The separate embedding half of the accuracy budget produces $A_{\mathrm{loc}} = 4e_E\Pi_{\mathrm{out}}\Pi_0$ rather than another quadratic term. Thus the numerical constants in Theorem 2.4 are inherited directly from the proper-cover lemmas and the explicit $\epsilon/2$ split. The constants are deliberately unoptimized. For instance, the argument of Lemma D.2 already gives $20q^2\log(k+1) + 128q^2\log(2n+1) \le 128q^2[\log(2k) + \log(2n+1)]$, so 256 could be replaced by 128 and 1024 by 512. Since κ_N scales with the square root of this constant, the change would lower every $\log_{10}\kappa_N$ by $\log_{10}\sqrt{2} \approx 0.15$ and leave every conclusion of Section 4 unchanged; we keep the stated constants, which match all archived computations.

H Exact downstream-tail counterfactual and logit-cap caveats

For a fixed reference record r, let $\widehat{L}_u^{(r)}$ be the finite-restart lower estimate of the induced Jacobian norm of residual sublayer u. The archived ambient zero-floor factors are $L_u^{(0)}$. In the two actual perturbation families, the embedding/readout budgets vanish and the two varying matrices belong to one selected sublayer s. Consequently, with all other constants fixed, homogeneity gives exactly

$$\frac{K_{\mathrm{tail}}^{(r)}}{K_{\mathrm{loc}}^{(0)}} = \frac{\prod_{u>s}\widehat{L}_u^{(r)}}{\prod_{u>s}L_u^{(0)}}.$$

This is specific to the studied two-matrix same-stage perturbation families: no corresponding single-tail identity is claimed for a class varying multiple distinct stages or the readout. Writing $N_0 = 64$ and $N = 10^6$, the archived endpoint ratio yields

$$\log_{10}\Gamma_N^{(r)} = \log_{10}\Gamma_{N_0}^{(0)} + \sum_{u>s}(\log_{10}\widehat{L}_u^{(r)} - \log_{10}L_u^{(0)}) + \frac{1}{2}\log_{10}\frac{\mathcal{L}_N}{\mathcal{L}_{N_0}} - \frac{1}{2}\log_{10}\frac{N}{N_0}.$$

The two sample-size adjustments total -2.0140 decades with $V = k_* = 32{,}000$ and $T = 128$. Our archived per-record stagewise values and this identity produce Table 14, including its distinction between the first Q/K stage (tail $u = 2,\dots,44$) and the last FF stage (empty tail). The per-record CSV and the exact recomputation script are included in the supplementary repository.

Limits of this counterfactual. The $\widehat{L}_u^{(r)}$ are finite-restart lower estimates and are not upper Lipschitz certificates for an entire parameter class. The counterfactual does not replace Π_{out} or establish an admissible population floor, and it must not be inserted into the original high-probability statement. Its value is diagnostic: even after making this optimistic stage-tail substitution, the scalar stage obstruction remains enormously above the threshold. A genuinely direction-aware bound would need a certified, uniform treatment of admissible perturbation trajectories and possibly an additional complexity cost.

Sharpness and ordinary cross-entropy. The root $\kappa_\star$ is sharp only for the scalar truncated-Dudley lower envelope with its ϵ^{-2} logarithmic term and $U = 0$; no matching lower bound on the decoder class's true entropy is asserted. The bounded-logit map $\psi_c(z) = c\tanh(z/c)$ produces loss range $B(c) = \log[1 + (V-1)e^{2c}]$, with $B'(c) = 2(V-1)e^{2c}/[1 + (V-1)e^{2c}]$. Holding the architecture coefficient fixed makes $\kappa_N(c) \propto 1/B(c)$; nevertheless, varying c also changes the empirical risk and the target bounded loss. Ordinary unbounded cross-entropy would require different tail/smoothing assumptions and is not covered by simply taking $c \to \infty$ in the stated high-probability bound. The completed same-record follow-up measures unclipped cross-entropy 2.7462 and clipped cross-entropy 3.8191 at $c = 10$ over 8,192 next-token outcomes. In particular, the bounded surrogate is not being presented as a numerically identical proxy for ordinary cross-entropy on this reference. These empirical losses remain finite-set diagnostics and do not alter the theorem's loss assumptions.

I EXPERIMENTAL PROTOCOL AND PROVENANCE

I.1 CONTROLLED SOURCE, MODEL, AND SPLITS

The controlled source has vocabulary $V = 8$ and context length $T = 32$. The first two tokens are uniform. Thereafter, conditional on the previous two tokens, the next token equals $(x_{t-1} + 2x_{t-2}) \bmod 8$ with probability 0.9 and is uniform on the vocabulary otherwise. Each training/data seed independently regenerates pilot, fit, and test sequences. The pilot set has 12,000 independent sequences and is used only to train the reference; the fit and test sets have 512 and 2,000 independent sequences. Reference training uses 12 epochs. The baseline decoder has $L = 4$, $d = 128$, head dimension 32, $d_{\text{ff}} = 512$, residual coefficient one, and $\nu = 10^{-5}$. The architecture grid changes one axis at a time: $L = 2$, $L = 8$, $d = 64$, $d = 256$, scaled $L = 8$ with $\lambda = (2L)^{-1/2}$, and $\nu = 10^{-3}$.

The main architecture study uses seeds 20260904, 20260905, and 20260906. Within a seed, sequences and Rademacher draws are paired across architectures when the implementation permits; between seeds the data, initialization, direction bank, and sign draws are regenerated. The baseline notebook and architecture extension in the repository are the recovered generating code for the training and E2–E4 pipeline.

I.2 RANDOM-FAMILY DIAGNOSTIC

For the three-seed architecture audit, each reference-centered family uses 32 presampled paired directions plus the reference, at scales $\eta \in \{10^{-3}, 10^{-2}\}$. The direction bank is fixed before evaluation signs are drawn. E2 uses one position per independent sequence, 1,024 Rademacher draws, and 400 paired sign-bootstrap resamples. Bootstrap intervals quantify only the Monte Carlo sign variation conditional on the chosen family. They do not include training-seed, family-selection, or population uncertainty.

The finite-family minimum radius $\widehat{c}_s$ in (49) is a measurement on the evaluated set, not a population floor. The frozen-denominator channel holds each example's denominator at its reference value. The main text's 612 rows are stage-level diagnostics and are not treated as independent experimental replications.

I.3 RADIUS RECURSION AND E4 BOOKKEEPING

E3 evaluates positive candidate floors at 0.25, 0.5, and 0.75 times the empirical reference profile and separately records the unconditional zero-floor recursion. A candidate is reported as certified only when the displayed sufficient inequalities hold; otherwise the experiment records the zero-floor result without making a claim about the realized hidden-state floor.

E4 uses the same decoder, capped loss, $\delta = 0.05$, and independent-sequence sample unit. Reference-centered and matched-global budgets share the same ambient propagation constants. The numerical optimizer searches $\log(\alpha/B) \in [-40, 0]$ and checks $\alpha = B$ explicitly. We report the raw objective U_N, then $\Phi_N = \min\{B/2, U_N\}$, then the final range-capped bound. For all 60 archived conditions, the endpoint sufficient ratio $96K_{\text{loc}}\sqrt{\mathcal{L}}/(B\sqrt{N})$ exceeds 8.00×10^{20}.

I.4 NO-RETRAINING FOLLOW-UP CONTROLS

The follow-up notebook reuses saved reference checkpoints and performs no additional reference training. It exports per-sign maxima and full search histories. The matched study evaluates trained and initialization references, nested direction counts 32 and 128, centered and uncentered channels, and common-sign paired contrasts. The adversarial study optimizes ℓ_1-bounded mixtures of 16 fixed scaled directions, uses two restarts, selects on search signs only, freezes the selected family, and evaluates it on fresh signs and separately on fresh sequences. The complete follow-up archive and its internal SHA256 inventory are stored in `data/followup`.

I.5 FOLLOW-UP CLAIMS RECOMPUTED FROM THE EXPORTED ARCHIVE

The repository script `analysis/verify_claims.py` recomputes the main follow-up values directly from the exported CSV files. Across trained-reference uncentered rows with nonzero direction

count, the singleton raw share ranges from 0.993780 to 0.999800 with median 0.998671. Across trained centered families with nonzero direction count, the maximum centered distortion is below 0.954. For trained references, the uncentered nested 32-to-128 contrast has mean -5.89×10^{-4} and is negative in every archived row, with largest change -5.44×10^{-6}. In the adversarial same-fit fresh-sign evaluation, the maximum distortion is 0.953677 uncentered and 0.954537 centered, with maximum bootstrap upper endpoints 0.954788 and 0.955671.

I.6 F3 ARCHIVAL RERUN AND ROBUSTNESS

The depth–sample-size diagnostic was regenerated under a frozen protocol after the earlier per-run F3 archive could not be recovered. The scientific grid is $L \in \{2, 4, 8\}$ and $N \in \{128, 256, 512, 1024, 2048, 4096, 8192\}$. The original seeds 20260904, 20260905, and 20260906 define the primary analysis; two additional seeds, 20260920 and 20260921, were fixed before the full run and are used only as a robustness extension. For each fixed seed and depth, the initialization is shared across sample sizes and the training sets are nested prefixes as N increases. Reference training uses 12 epochs; the held-out test set contains 2,000 independent sequences. The descriptive gap is the capped test loss minus the capped loss on the training sample used for that reference. Because the reference is retrained as N changes, F3 is not used as a fixed-reference theorem-coverage experiment.

The complete rerun contains $3 \times 7 \times 5 = 105$ cells. Each cell exports a raw JSON record with data hashes, initialization and final-model fingerprints, checkpoint hash, training and test losses, K, Π_0, the raw entropy objective, endpoint status, and wall time. The primary manuscript table first averages the original three seeds within each (L, N) cell and then computes Spearman correlations across the resulting 21 cells. The two extra seeds do not alter that primary definition.

Candidate	pooled	$L = 2$	$L = 4$	$L = 8$
K	$+0.147$	-1.00	-1.00	-1.00
$\log \Pi_0$	$+0.147$	-1.00	-1.00	-1.00
$\log K/\sqrt{N}$	$+0.987$	$+1.00$	$+1.00$	$+1.00$
U_N before cap	$+0.990$	$+1.00$	$+1.00$	$+1.00$
$1/\sqrt{N}$	$+0.877$	$+1.00$	$+1.00$	$+1.00$
training loss	$+0.301$	$+0.64$	$+0.64$	$+0.54$

Table 9: Prespecified five-seed robustness version of the F3 rank-correlation table. The qualitative separation is unchanged: within depth the $N^{-1/2}$ factor fixes the rank ordering, while pooling depths reveals additional ordering associated with the computed coefficient.

All 105 final bounds are attained at the loss-range cap. Thus the coefficient–gap correlations are descriptive diagnostics of how the computable ingredients co-vary across trained models, not evidence that the reported range-capped bound is quantitatively calibrated. The earlier `results2_legacy_synthetic_debug.zip` archive remains under `data/legacy` only and is not used as evidence for the regenerated table.

J PRETRAINED-MODEL NATURAL-TEXT REPLICATION

J.1 FROZEN SCIENTIFIC PROTOCOL

The external replication uses the public checkpoint `TinyLlama/TinyLlama-1.1B-Chat-v1.0` at exact revision `fe8a4ea1ffedaf415f4da2f062534de366a451e6`. The tokenizer is resolved to the same revision. The architecture has $L = 22$, $d = 2048$, $H = 32$ query heads, $J = 4$ key/value groups, head dimension 64, SwiGLU width 5632, vocabulary size 32,000, and RMS regularizer $\nu = 10^{-5}$. No model training or fine-tuning is performed.

Natural text comes from `Salesforce/wikitext`, configuration `wikitext-103-raw-v1`, validation split, exact dataset revision `b08601e04326c79dfdd32d625aee71d232d685c3`. Before certificate outcomes are computed, a deterministic seed fixes 32 eligible records for calibration and 64 disjoint records for evaluation; each record contributes the first 129 tokenizer outputs so that the predictor/target sequence length is 128. The selected token archive has SHA256

56f268d9cc4cbb191e8481f233d45ed13a46ae5f3f4e82c220bdc7ef997a029b. The scientific configuration SHA256 is a3a468afe61f74ddfb3ad55cf886430b6755e06b4045725893e866de805114e1.

The reference-centered class freezes embeddings, normalization gains, readout, and all unselected weights. Two groups are evaluated separately: the first-layer attention Q/K pair and the final-layer feedforward gate/up pair. For each group, relative perturbation scales are $\eta \in \{10^{-3}, 10^{-2}\}$. The protocol fixes eight paired rank-one directions before outcomes are inspected and evaluates the reference plus both signs of every direction, for a 17-member finite family. The finite-family diagnostic uses 1,024 Rademacher sign draws and 400 paired-sign bootstrap resamples. Candidate positive floors are 0.25, 0.50, and 0.75 times the calibration reference profile. Those checks are calibration-set sufficient tests only; they are not promoted to population radius certificates.

J.2 FP32 NUMERICAL CORRECTION AND REALIZED-BUDGET AUDIT

An initial engineering run used FP16 weights and FP16 captured states. Because the smallest intended perturbations are close enough to FP16 resolution to make quantization a plausible confound, that run is not used as manuscript evidence. The final v2 run preserves the identical scientific configuration, checkpoint revision, text records, groups, scales, directions, and random seeds, but loads the checkpoint in FP32, retains captured residual states in FP32, disables TF32, and uses eager attention.

For every group, scale, direction, sign, and selected matrix, v2 applies the perturbation in FP32 and records the realized effective Frobenius norm, mixed $2,1$ norm, and column cap. The Frobenius norm is used as a valid operator-norm upper bound in the conservative reference-centered budget. Across the complete audit, the largest realized-to-nominal perturbation ratio is 1.00000102, and the largest FP32 rounding residual in Frobenius norm is 1.16×10^{-6}. The theorem calculations use the worst realized budgets over the full frozen direction bank rather than the nominal budgets. Thus the final numerical results do not rely on exact floating-point representation of the intended rank-one update.

J.3 FINITE-FAMILY DISPLACEMENT AND NORMALIZATION DIAGNOSTICS

For the first-layer Q/K family at $\eta = 10^{-3}$, the maximum immediate post-attention displacement is 6.44×10^{-6} and the maximum final-stream displacement is 4.31×10^{-4}. Across signed directions, the ratio of final to immediate displacement ranges from 44.0 to 95.9, with median about 70.1. At $\eta = 10^{-2}$ the corresponding maxima are 6.47×10^{-5} and 4.31×10^{-3}, with essentially the same amplification pattern. For the last feedforward group, the target is already the final residual stream, so target and final maxima coincide: 1.93×10^{-4} at 10^{-3} and 1.93×10^{-3} at 10^{-2}.

Across all target/final and centered/uncentered channels, the largest finite-family diagnostic is 0.671174 and the largest paired sign-bootstrap upper endpoint is 0.673010. The last-feedforward diagnostics are smaller, with maximum 0.132644. As in the controlled study, these are measurements on the fixed 17-member families and do not replace the continuous-class theorem.

J.4 PROPAGATION, REFERENCE-CENTERING, AND THE OBSTRUCTION REGIME

The sharpened thresholds and profile-ceiling calculations added in the current manuscript require no new model evaluation. They are deterministic recomputations from the frozen FP32 archive. The universal constant is $\kappa_\star = 0.1085299426\ldots$, equivalently $\gamma_\star = \kappa_\star/4 = 0.0271324857\ldots$ in the archived Γ_N convention. Relative to the earlier endpoint-pinning condition $\Gamma_N \geq 1$, this enlarges the sample-size obstruction by the factor $(1/\gamma_\star)^2 = 1358.38\ldots$.

At $\eta = 10^{-3}$, the induced-norm channel therefore gives $\log_{10} N_\star^{\mathrm{stage}} > 823.3155$ for first-layer Q/K and > 23.9019 for the last feedforward gate/up group. The corresponding power-estimate values are > 531.6580 and > 21.3217. At $N = 10^6$, holding transport factors fixed, the first-layer Q/K mixed-norm budgets would have to be contracted by an additional factor below $10^{-408.7407}$ in the induced channel (below $10^{-262.9119}$ in the power channel) merely to remove the stage-term obstruction. For the last feedforward group the induced-channel contraction factor is below $10^{-9.0339}$.

The full stagewise reference radius profile is already archived. Because any valid population floor applying to these evaluation sequences must lie below that profile coordinatewise, Proposition 3.6

permits an optimistic ceiling calculation: set the candidate floor equal to the observed reference minima themselves. At $N = 10^6$ and $\eta = 10^{-3}$ this still gives $\log_{10} \Gamma_N = 349.5486$ (induced) and 203.7220 (power) for first-layer Q/K, and 5.9718 (induced) and 4.6817 (power) for the last feedforward group. All are far above $\log_{10} \gamma_\star = -1.5665$. Thus no floor profile compatible with those reference evaluations can remove the stage obstruction under the instantiated ambient caps.

For a linear profile anchored at the observed stage-zero radius 0.005518, the largest slope below every evaluation-stage reference minimum is 1.5681×10^{-3} per residual sublayer; the calibration profile permits 1.8973×10^{-3}. The largest branch Lipschitz certificate is 1.4485×10^5 in the power-estimate channel and 3.0083×10^9 in the induced-norm channel. Hence the empirically allowed reference growth is many orders below the $\zeta \asymp \Lambda$ boundary of Proposition L.3.

Table 3 in the main text reports the power-estimate channel. Its zero-floor propagation product satisfies $\log_{10} \Pi_0 \approx 262.46$ in all four conditions. Replacing the power spectral estimates by the archived induced-norm upper bounds raises this to approximately 408.87 without changing the qualitative conclusion. In the power channel, reference-centering reduces the matched global coefficient by factors 1.32×10^3 and 1.33×10^2 for first-layer Q/K at scales 10^{-3} and 10^{-2}, and by 8.72×10^3 and 8.72×10^2 for the late feedforward group. The selected-stage effective count remains small, approximately 1.63 for Q/K and 1.80 for the feedforward group.

All 12 calibration-set positive-floor sufficient tests fail. The separately held-out evaluation reference nevertheless remains above every calibration-derived candidate profile, with a positive minimum margin in each case. This contrast is not a population certification result; it shows only that failure of the displayed sufficient recursion should not be interpreted as an observed radius collapse on the evaluation records.

For the 32,000-token vocabulary and $c_{\log} = 10$, the capped-loss range is $B = 30.37346$. In every reference-centered/global and power/upper condition, the entropy optimizer is attained at $\alpha = B$ and the final reported bound equals B. Thus the pretrained replication preserves the controlled-study hierarchy: small sampled normalization distortion and large reference-centering gains coexist with enormous worst-case propagation and endpoint saturation.

J.5 Norm-matched trajectory-Jacobian transport audit

A separate frozen audit evaluates the residual-sublayer Jacobian along the identical FP32 reference trajectory and the same 64 evaluation records. No training or weight perturbation is performed. For $\Phi_s(H) = H + F_s(\mathcal{N}(H))$, the primary quantity is the induced norm $\|D\Phi_s(H_{\mathrm{ref}}^{s-1})\|_{\infty,2\to\infty,2}$, where $\|H\|_{\infty,2} = \max_t \|H_t\|_2$. The factor $\sqrt{d}$ relating this norm to $\|\cdot\|_{\infty,\circ}$ cancels in the induced operator norm, so this is the stream norm appearing in the theorem.

The estimator alternates Jacobian-vector and vector-Jacobian products. Given a block-unit input direction, it selects the output token with largest Euclidean norm, places a dual unit vector on that block, applies the transpose Jacobian, and renormalizes each input block to unit Euclidean norm. The full run uses eight iterations and two deterministic restarts for each record–stage pair. Every attained objective is feasible for the primal norm maximization, so the reported value is a lower estimate of the exact induced norm. A six-iteration flattened-ℓ_2 power estimate is archived as a secondary diagnostic and is not identified with the theorem norm.

The full audit contains $64 \times 44 = 2{,}816$ record–sublayer evaluations and passes its internal SHA256 manifest. Three central-difference checks of the JVP implementation have relative errors 4.53×10^{-5}, 6.87×10^{-5}, and 1.01×10^{-4}. The median relative change from the penultimate to final block-power iterate is 1.82×10^{-4} and the 95th percentile is 1.03×10^{-2}; one trace has a larger final increase. This finite-iteration issue is conservative for the conclusion below because additional optimization can only increase the best feasible lower estimate.

Across records, the cumulative realized block-Jacobian lower estimate has $\log_{10}$ range $[38.25, 42.36]$ and median 40.58. The flattened-ℓ_2 diagnostic has median 42.51. By comparison, retaining the ambient branch caps but replacing the zero-floor normalization factor by each record's reference radius gives median power-channel log-product 203.35, while the zero-floor ambient products are 262.46 (power estimate) and 408.87 (induced upper bound). Thus the zero-floor-to-reference-radius substitution accounts for 59.12 median decades, and the numerical gap from the reference-radius

ambient product to the finite-restart realized lower estimate is 162.85 median decades. The total zero-floor-to-realized numerical gap is 221.89 median decades.

The largest median single-stage realized factor is the layer-2 feedforward residual map at 421.54; the first attention stage has median 21.26. Starting immediately after the first attention stage, the product of the remaining realized block estimates has median $\log_{10}$ value 39.23. This is far larger than the roughly 70-fold median downstream amplification of the frozen rank-one Q/K perturbation family. The comparison is diagnostic, not an operator-norm identity, because the rank-one family follows particular parameter-induced directions. Its role is to motivate a future theorem that tracks cross-layer directional compatibility rather than multiplying the induced norm of each stage independently.

Figures 4 and 1 summarize the same archive. The second figure exposes the stagewise factors.

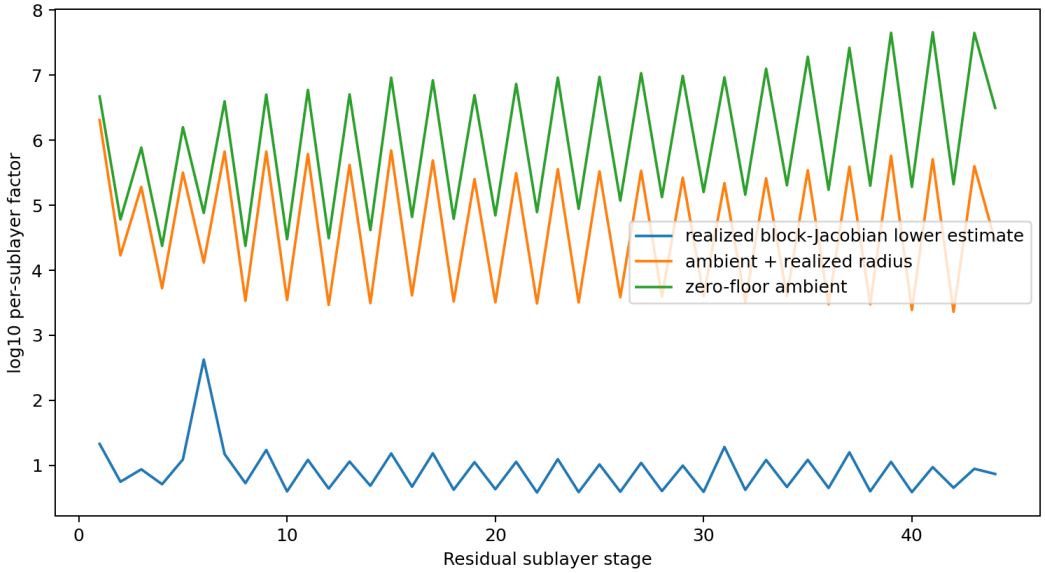


Figure 1: Median per-sublayer transport factors in the frozen trajectory audit, shown on a base-10 logarithmic scale. The realized block-Jacobian quantity is a finite-restart lower estimate; the other two curves retain the ambient branch caps.

J.6 Direct composite-Jacobian audit and logit-cap checks

The separately locked full-composite audit holds the checkpoint and all 64 evaluation records identical to the FP32 pretrained and stagewise transport runs. For each record r, let $J_s^{(r)} = D\Phi_s(H_{\mathrm{ref}}^{s-1,r})$ act on the max-row Euclidean stream space, and define the *full* operator $J_{44:1}^{(r)} = J_{44}^{(r)} \cdots J_1^{(r)}$. Its forward action on a tangent v is computed by successively applying the 44 stage JVPs. The transpose action on a cotangent u applies the 44 stage VJPs in reverse order. After each action, the vector is normalized in the block $\ell_{\infty,2}$ norm and the base-10 logarithmic scaling recorded; because each Jacobian map is linear, this rescaling leaves the operator action exact apart from floating-point effects. The direct induced-norm estimator alternates output-maximizing block selection with the block-unit dual update. It uses 12 iterations and three deterministic restarts for each of the 64 records; an eight-iteration, two-restart flattened-ℓ_2 power estimate is retained for comparison, without identifying the Euclidean and max-block induced norms.

Integrity, convergence, and finite differences. The complete uploaded ZIP passes CRC and contains 85 SHA256-verified files besides its own manifest. The locked protocol hash is `a981a50e5d082dabda2c90f17e9e8bc0189e27a060ebb638a829c4a8dd00ce94`; the notebook records exact SHA256 identities for both prior frozen input archives. The full run contains 64 unique record checkpoints, each with a 13-value primary trace and a 44-stage winning-direction path, and was performed on an NVIDIA A100-SXM4-40GB in FP32 with TF32 disabled.

The median absolute base-10-log change between the penultimate and final primary iterates is 8.24×10^{-8}, with 95th percentile 2.20×10^{-7} and maximum 5.67×10^{-6}. These small terminal changes indicate stable *estimated* values, not rigorous operator-norm upper bounds. The toy dense-matrix chain-rule self-test and first-record continuity check are included in the executed notebook; all 43 consecutive-stage continuity errors on that record are zero at the exported precision. A central-difference test for the whole chain on the first record reports normalized directional errors 4.40×10^{-4}, 1.17×10^{-3}, and 3.54×10^{-3} for relative steps 10^{-2}, 3×10^{-3}, and 10^{-3}, respectively. The best tested step is therefore the largest, consistent with FP32 cancellation at smaller steps.

Main and short-window outputs. The median $\log_{10}$ product of individually estimated stage norms is 40.5751, while the median direct full-composite block estimate is 3.6615; their median *per-record* log-ratio is 36.8438 (range $[34.8538, 38.6478]$). The secondary flattened-ℓ_2 composite estimate has median log-value 3.3608. All are derived from the same 64 frozen records. To inspect when the separation appears, a prespecified 10-row short-window experiment evaluates five stage windows on each of records 0 and 1; Table 10 gives the average *estimated* log-ratios.

Stages of the product	window length	mean log-ratio of lower estimates
1–2	2	0.37
1–4	4	2.03
1–8	8	6.05
17–24	8	5.15
37–44	8	5.43

Table 10: Short-window diagnostic on TinyLlama records 0 and 1: mean $\log_{10}$ ratio of the stagewise product of lower estimates to the directly estimated window composite.

These five selected windows are exploratory finite-set diagnostics.

Remark J.1 (Reading the measured ratio). The 36.84-decade median paired gap compares two finite-search lower estimates, since numerator and denominator are both underestimated; it is an estimate of $\mathfrak{A}_D$ rather than a bound in either direction. The operator that would enter K_{loc} for the first-stage Q/K class is the selected-stage downstream map $J_{44:2}$, taken uniformly over the reference-centered class, and the full-stream composite $J_{44:1}$ at the reference is its trajectory-level proxy. The stage-tail counterfactual (Appendix N) is the quantitative statement about the existing certificate; the composite result shows where a non-product transport proof could improve it.

Figure 2 shows the full-chain comparison for every record.

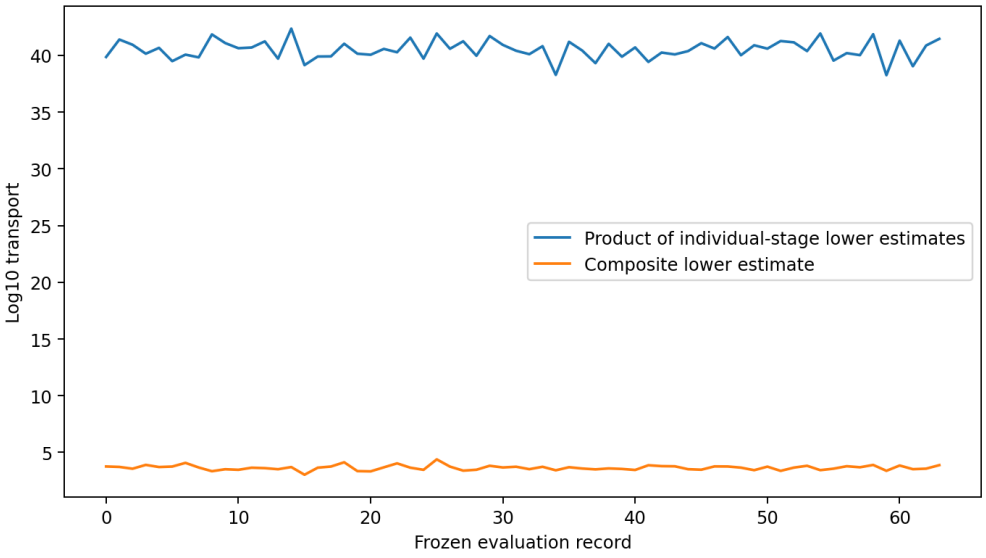


Figure 2: Per-record comparison of the product of 44 separately optimized stage-norm lower estimates and the directly estimated composed Jacobian norm on the TinyLlama 1.1B checkpoint. The median paired separation is 36.84 decades; both curves are finite-search lower estimates.

Limitations and the selected-stage tails. Both the product in the numerator and the directly optimized composite in the denominator of the empirical ratio use finite-restart lower estimates. The exact alignment defect is therefore *not* bounded by this ratio in either direction. The full map $J_{44:1}$ also differs from the downstream map $J_{44:2}$ required after a first-stage Q/K parameter perturbation. Neither the composite estimate nor the short-window measurements can be substituted as a uniform upper transport bound in the population covering argument without an additional theorem. The stagewise-tail coefficient calculation in Appendix H remains the relevant fixed-factor counterfactual for the current certificate.

Loss-cap sensitivity. Over the same 8,192 evaluation tokens, the ordinary unbounded next-token cross-entropy is 2.74619. Applying $\psi_c(z) = c\tanh(z/c)$ to the reference logits yields capped losses 5.95170, 3.81911, 2.81494, and 2.73659 at $c \in \{5, 10, 20, 40\}$. The corresponding conditional induced-upper $\log_{10}\Gamma_{10^6}$ with the reference-centered coefficient held fixed is 407.35, 407.17, 406.95, and 406.70. The strong numerical obstruction persists throughout this grid, but the result is a sensitivity analysis of the bounded-surrogate framework, not a high-probability guarantee for ordinary unbounded cross-entropy. The experiment and all source matrices, deterministic seeds, checkpoints, environment metadata, short-window traces, finite-difference checks, and summary CSVs appear in the frozen composite archive.

J.7 Archive and scope

The final run uses 32 calibration records, 64 evaluation records, all eight frozen directions and both signs, 1,024 sign draws, and 400 bootstrap resamples. It was executed on an NVIDIA A100-SXM4-40GB under PyTorch 2.11.0+cu128; the full loop took approximately 81.5 seconds and performs no training. The compact archive contains an internal SHA256 manifest covering 73 files, including the protocol lock, token archive, model and environment manifests, reference matrix/radius constants, realized perturbation audits, per-condition displacement and sign-maxima records, positive-floor checks, and reference-centered/global bound summaries.

The fixed public checkpoint is independent of the selected text records, but the WikiText records themselves are used here as a frozen empirical corpus. Accordingly, the experiment instantiates the theorem's coefficients and finite-family diagnostics; it is not presented as an empirical coverage test

of the probabilistic generalization statement or as a proof that the corpus records are i.i.d. draws from a specified population.

The completed Jacobian audit shows that simply replacing the ambient branch caps by sharper per-sublayer trajectory norms does not suffice: their product is still enormous. The remaining theoretical direction is therefore a direction-aware or non-product transport inequality whose data use is accounted for explicitly. Any such replacement inside the uniform generalization theorem must retain a formally justified data split, complexity correction, or deterministic structural argument.

K Positioning, symbol guide, and mechanism ledger

Work	Architecture / class	Entropy-to-bound route	Scope of our threshold
Edelman et al. (2022)	bounded-norm attention; sparse-variable analysis	ϵ^{-2}-type covers and a modified Dudley integral	analogous obstruction if normalized objective and range cap are matched; not verbatim
Trauger & Tewari (2024)	norm-bounded single/multi-layer transformers; sequence-length-independent bounds	matrix covers and Dudley integration, including ϵ^{-2} entropy	analogous mechanism; their unspecified constants prevent a direct numerical transfer
Sakai & Imaizumi (2026)	trained deep transformers; post-hoc layerwise Schatten control	Dudley integration with adaptive entropy exponents and indices	applies only to a separately established ϵ^{-2} logarithmic-term specialization, not all indices
Lotfi et al. (2024a;b)	trained large language models	compression / token-based approaches	outside the present continuous-cover obstruction
This work	fixed-reference pre-RMS GQA/SwiGLU decoder	proper covers, explicit ϵ^{-2} stage term and truncated Dudley	theorem applies with stated exact constants

Table 11: Positioning by *proof mechanism*, not just by model size. Generic use of Dudley integration does not automatically transfer our numerical constant or certify that another paper's optimized expression has the exact form of Theorem 3.2. The first two papers exhibit comparable ϵ^{-2} covering mechanisms; the third allows other entropy exponents.

Input	Intermediate quantity	Role in the final certificate
spectral budgets τ_j	q_s, Δ_s	certify state displacement and admissible floor profiles
floors c_s	ω_s, L_s, Π_s	transport perturbations through the residual stack
mixed budgets R_j^{loc}	w_j, K_{loc}	set the proper-cover entropy coefficient
K_{loc}, N	κ_N	decide whether the stage term can affect the optimized objective

Table 12: Mechanism ledger. Reference-centering acts directly on the cover budgets, whereas floor profiles act through transport. The stage-obstruction statistic κ_N is where the two mechanisms meet.

Symbol	Meaning
λ_s, Λ_s	residual coefficient and branch input-Lipschitz certificate at sublayer s
ℓ_j, b_j	sensitivity and input-radius factors for linear stage j
$\tau_j, R_j^{\mathrm{loc}}$	spectral displacement budget and reference-centered mixed-norm cover budget
c_s, ω_s	certified RMS floor and the corresponding normalization factor $(c_s^2+\nu)^{-1/2}$
L_s, Π_s	one-step residual transport factor and downstream product from stage s
w_j, K_{loc}	propagation-weighted stage cost and its $2/3$-power aggregation
$\mathcal{L}_N$	entropy logarithm $\log(2k_*)+\log(2NT+1)$
κ_N	normalized stage-obstruction coefficient $384K_{\mathrm{loc}}\sqrt{\mathcal{L}_N}/(B\sqrt{N})$

Table 13: Symbol guide for the quantities used in the main decomposition. The visually similar λ_s and Λ_s play different roles: the former scales the residual branch, while the latter bounds its input sensitivity.

L Supplementary theory: budget scale, effective stage count, and linear floors

L.1 Critical reference-centered budget scale

Proposition L.1 (Critical reference-centered budget scale)**.** *Hold the ambient caps, floor profile, and hence all propagation factors fixed. Scale the reference-centered embedding, mixed-norm stage, and readout budgets by $q \in [0,1]$. Then $K_{\mathrm{loc}}(q) = qK_{\mathrm{loc}}(1)$. Define*

$$q_\star(N) = \frac{\kappa_\star B\sqrt{N}}{384K_{\mathrm{loc}}(1)\sqrt{\mathcal{L}_N}}. \tag{44}$$

Any $q \geq q_\star(N)$ is still forced to be vacuous by Theorem 3.2; a strictly nonvacuous certificate requires $q < q_\star(N)$. Thus $q_\star$ is a neighborhood-radius form of the same obstruction, conditional on fixed transport factors. If spectral displacement budgets are simultaneously shrunk, the recursion for Δ_s and the set of feasible floor profiles may change; $q_\star$ does not summarize that coupled effect.

Proof of Proposition L.1. Under the stated fixed transport factors, Proposition 2.5 gives $K_{\mathrm{loc}}(q) = qK_{\mathrm{loc}}(1)$, and therefore $\kappa_N(q) = q\kappa_N(1)$. Solving $\kappa_N(q) = \kappa_\star$ gives (44); apply Theorem 3.2 on either side of that threshold. □

For compatibility with the archived endpoint diagnostic, define $\Gamma_N = 96K_{\mathrm{loc}}\sqrt{\mathcal{L}_N}/(B\sqrt{N}) = \kappa_N/4$ and $\gamma_\star = \kappa_\star/4 = 0.0271325\ldots$. If at some N_0, $\Gamma_{N_0} = R \geq \gamma_\star$, then

$$N_\star^{\mathrm{stage}} \geq N_0\left(\frac{R}{\gamma_\star}\right)^2. \tag{45}$$

Proof of (45). Let $\Gamma_{N_0} = R \geq \gamma_\star$ and set $N_1 = N_0(R/\gamma_\star)^2$. Then

$$\Gamma_{N_1}^2 = R^2\frac{\mathcal{L}_{N_1}/N_1}{\mathcal{L}_{N_0}/N_0} = \gamma_\star^2\frac{\mathcal{L}_{N_1}}{\mathcal{L}_{N_0}} \geq \gamma_\star^2,$$

so $\kappa_{N_1} = 4\Gamma_{N_1} \geq \kappa_\star$. Monotonicity gives $N_\star^{\mathrm{stage}} \geq N_1$. □

L.2 Effective stage count

Proposition L.2 (Effective stage-count identity)**.** *For nonnegative stage weights $w_1, \ldots, w_J$, let*

$$\|w\|_{2/3} = \left(\sum_{j=1}^{J} w_j^{2/3}\right)^{3/2}, \qquad J_{\mathrm{eff}} = \left(\frac{\|w\|_{2/3}}{\max_j w_j}\right)^{2/3}. \tag{46}$$

Then $1 \leq J_{\mathrm{eff}} \leq J$ and $\|w\|_{2/3} = J_{\mathrm{eff}}^{3/2}\max_j w_j$. Thus the nominal $J^{3/2}$ stage-accumulation factor occurs only when the weights are balanced; if a few propagation-weighted stages dominate, J_{eff} is small.

Proof of Proposition L.2. Let $M = \max_j w_j$. If $M = 0$, the covering contribution is identically zero and the effective-count diagnostic is immaterial; assume $M > 0$. Since at least one $w_j = M$,

$$M^{2/3} \le \sum_j w_j^{2/3} \le J M^{2/3}.$$

Raising to the $3/2$ power yields $M \le \|w\|_{2/3} \le J^{3/2} M$. The definition (46) therefore gives $1 \le J_{\text{eff}} \le J$, and rearranging gives $\|w\|_{2/3} = J_{\text{eff}}^{3/2} M$. □

For the controlled baseline, the $2/3$-power allocation gives $J_{\text{eff}} = 1.73$; the first-block query/key stages account for 93.8% of the allocation. Raw stage count is therefore not the main source of the observed depth dependence.

L.3 Boundary for a linear growing floor

Proposition L.3 (Boundary for a linear growing floor)**.** *Suppose $\lambda_s = 1$, $\Lambda_s \le \Lambda$, and a nonempty class admits $c_s \ge a + \zeta s$ through depth D. Then*

$$\Pi_s \le e^{\Lambda/a} \left(\frac{a + \zeta D}{a + \zeta s} \right)^{\Lambda/\zeta}. \tag{47}$$

At the same time Proposition F.1 implies the necessary finite-depth condition

$$a + \zeta D \le B_0 + D\Lambda, \qquad \zeta \le \Lambda + \frac{B_0 - a}{D}. \tag{48}$$

Hence, for arbitrarily deep families with fixed a, B_0, Λ, the best possible exponent $p = \Lambda/\zeta$ allowed by these coarse upper and lower growth bounds approaches the boundary $p \ge 1$. The case $\zeta \asymp \Lambda$ is therefore the only potentially benign linear-floor regime in this certificate.

Proof of Proposition L.3. The propagation inequality (47) is exactly Proposition F.2. For the feasibility statement, a nonempty class satisfying the floor must contain at least one stream whose stage-D radius is at least $a + \zeta D$. Proposition F.1 gives the uniform upper bound $B_0 + D\Lambda$ under $\lambda_s = 1$ and $\Lambda_s \le \Lambda$. Hence $a + \zeta D \le B_0 + D\Lambda$, which rearranges to (48). With fixed a, B_0, Λ, letting $D \to \infty$ yields $\limsup \zeta \le \Lambda$ and therefore $\liminf p = \liminf \Lambda/\zeta \ge 1$. □

The necessary feasibility inequality compares a candidate minimum layer radius to a class-wide maximum radius; it is a weak ceiling test, not evidence that a linear floor of slope ζ can be attained. For the frozen pretrained reference, anchoring the line at the observed stage-zero radius 0.005518, the largest slope lying below every evaluation-stage minimum is only 1.57×10^{-3} per residual sublayer, whereas the largest branch Lipschitz bound entering the power-estimate channel is about 1.45×10^5. Thus $\zeta \asymp \Lambda$ is not numerically compatible with the observed reference profile.

L.4 Worked nonvacuous regime for the abstract objective

In the abstract objective (13) with $U = 0$, $B = 1$ and $\kappa = 0.05$, choosing $\alpha/B = \kappa/4$ gives $\Psi \le 0.05[1 + \log(80)] \approx 0.269$; at $N = 1000$ and $\delta = 0.05$, $2\Psi + 3\sqrt{\log(2/\delta)/(2N)} \approx 0.667 < B$. Thus the abstract objective admits a nonvacuous regime; the example is scalar and does not instantiate a trained decoder neighborhood.

M Additional controlled-study material

M.1 Normalization diagnostic and radius recursion

For a finite family $\mathcal{H}_s^{\text{fin}}$ at stage s, let $\widehat{c}_s$ be its minimum observed RMS radius and define

$$\widehat{D}_s = \frac{\widehat{\mathfrak{R}}_m^{\text{vec}}(\mathcal{N} \circ \mathcal{H}_s^{\text{fin}})}{(\widehat{c}_s^2 + \nu)^{-1/2} \widehat{\mathfrak{R}}_m^{\text{vec}}(\mathcal{H}_s^{\text{fin}})}. \tag{49}$$

This is a finite-family diagnostic normalized by the sampled-floor pointwise Lipschitz scale; it is not a theorem about the surrounding continuous neighborhood.

Across all 612 stage rows, $\widehat{D}_s \in [0.3422, 0.9594]$, and every pointwise sign-bootstrap upper endpoint is below one; the largest is 0.9602. For the tested positive candidate floors in $\{0.25, 0.5, 0.75\}$ times the empirical reference profile, the sufficient recursion returns the zero-floor case in every seed–architecture–scale setting. The corresponding final displacement-to-observation ratio ranges from 1.87×10^{18} to 2.81×10^{80}, which quantifies the scale of the propagation term used by the theorem.

Residual scaling at $L = 8$ lowers the product by about 9.6 decades, and raising ν to 10^{-3} lowers the baseline by about 8 decades. These interventions therefore have a large and directly measurable effect on the propagation component even though the optimized objective remains in the same endpoint regime.

M.2 Controlled-study figures

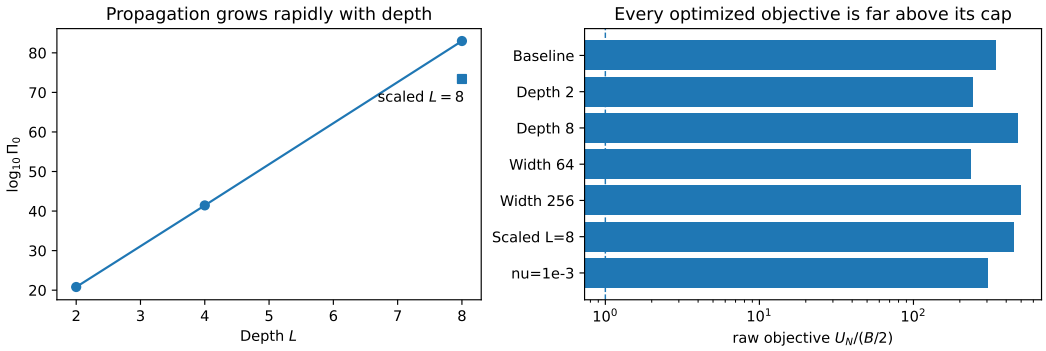


Figure 3: Left: the zero-floor propagation product is approximately exponential in tested depth; residual scaling lowers the propagation product substantially. Right: at $N = 512$ and perturbation scale 10^{-2}, every raw entropy objective is hundreds of times above its $B/2$ cap. Error bars in the left panel are descriptive across-seed standard deviations.

M.3 Additional negative controls

Centering, larger random banks, and the held-out adversarial search all leave the finite-family diagnostic below one (largest value 0.9545, bootstrap upper 0.9557). The frozen 105-run depth–sample-size study is likewise descriptive: within depth the explicit $N^{-1/2}$ factor determines rank ordering, and every final bound remains range-capped. Full tables and five-seed robustness results appear in Appendix I.

M.4 Full Table 2 with empirical gap

The descriptive reference test-minus-fit gap $\widehat{g}$ for each configuration is as follows: Baseline $5.76 \pm 4.84 \times 10^{-3}$; Depth 2: $6.79 \pm 3.54 \times 10^{-3}$; Depth 8: $7.70 \pm 5.05 \times 10^{-3}$; Width 64: $5.99 \pm 4.48 \times 10^{-3}$; Width 256: $6.41 \pm 5.16 \times 10^{-3}$; Scaled $L = 8$: $6.93 \pm 4.46 \times 10^{-3}$; $\nu = 10^{-3}$: $6.63 \pm 3.43 \times 10^{-3}$.

N Additional pretrained-study material

N.1 Loss-cap sensitivity on identical frozen tokens

Over $64 \times 128 = 8{,}192$ next-token outcomes, the unclipped reference loss is 2.746; at $c_{\log} = 10$ the capped loss is 3.819, approximately 39% higher. The following table varies the cap while holding the reference-centered coefficient fixed:

$c_{\log}$	$B(c_{\log})$	reference capped CE	$\log_{10}\Gamma_{10^6}$ (conditional)
5	20.37	5.952	407.3
10	30.37	3.819	407.2
20	50.37	2.815	407.0
40	90.37	2.737	406.7

Changing the loss-range scale moves the conditional obstruction by less than one decade, versus hundreds of decades in ambient transport.

N.2 Transport decomposition figure

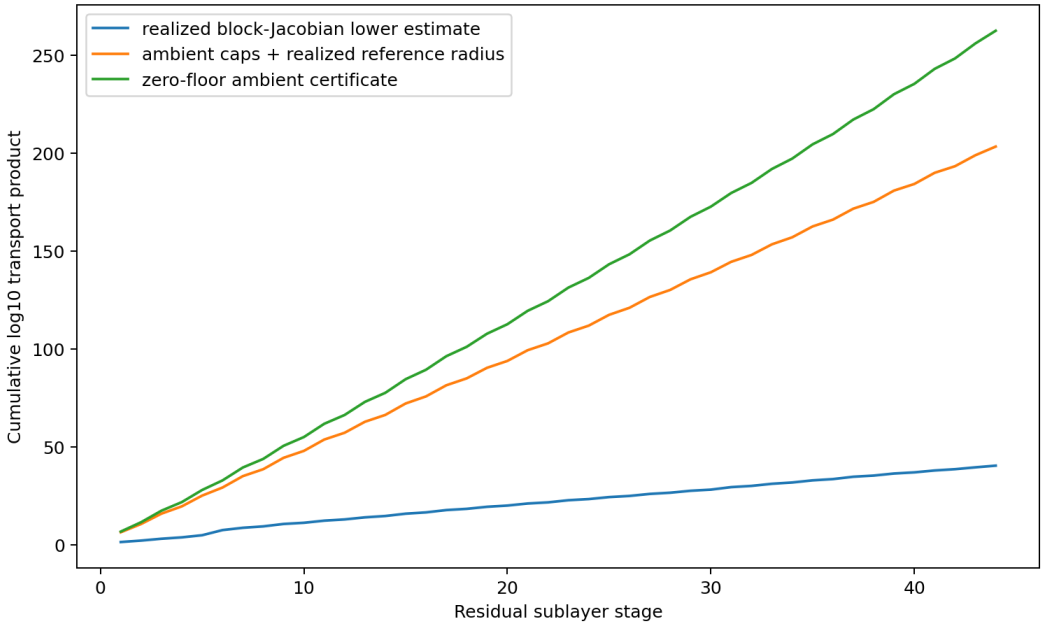


Figure 4: Cumulative transport across the 44 residual sublayers of the frozen 1.1B reference. Curves use stagewise medians over the 64 records. Replacing the zero-floor normalization by realized reference radii removes about 59 decades from the median power-channel log-product. The remaining numerical gap to the realized block-Jacobian lower estimate is about 163 decades, yet the realized product itself still reaches about $10^{40.6}$.

N.3 Stage-tail counterfactual at $N = 10^6$

To put the transport measurement into the units of the sharp obstruction, we replace each selected stage's downstream tail $\Pi_{s(j)}$ by the product of the corresponding finite-search block-Jacobian lower estimates on the same reference record; Π_{out}, the parameter budgets, the loss range and all other factors remain fixed. This is maximally optimistic about per-layer replacements, but remains a finite-trajectory algebraic diagnostic rather than a valid population certificate. Appendix H shows the exact stage-tail recomputation.

Selected group, $\eta = 10^{-3}$	power diagnostic	induced upper	above $\log_{10}\gamma_\star$
First attention Q/K	44.79	47.05	46.35 / 48.62
Last feedforward gate/up	6.18	7.47	7.74 / 9.03

Table 14: Median per-record $\log_{10}\Gamma_{10^6}$ after replacing only selected-stage downstream tails by realized block-Jacobian lower estimates. Even after this optimistic substitution, the stage obstruction remains more than 46 decades above its threshold for the first Q/K group.

N.4 SCOPE OF THE CURRENT EVIDENCE

The controlled experiments use small pre-RMS decoders and a synthetic order-two Markov source, with three independently generated training/data seeds in the main architecture study. Bootstrap intervals are pointwise and conditional on the sampled finite families. The primary F3 table summarizes 21 depth–sample-size cells obtained by averaging the original three seeds, and a prespecified five-seed robustness analysis is reported in Appendix I. The F3 coefficient–gap study is descriptive because its reference is retrained as the sample size changes; by contrast, the pretrained reference is a fixed public checkpoint chosen independently of the text records. The theorem-matched TinyLlama certificate instantiation and parameter-perturbation study concern one checkpoint and one frozen corpus. The separate cross-checkpoint transport audit in Appendix O extends the stagewise-versus-composite diagnostic to three additional pretrained checkpoints, but does not constitute additional population-certificate instantiations. All block-power quantities remain finite-search lower estimates on frozen trajectories and are not inserted into the high-probability theorem.

O CROSS-MODEL COMPOSITE-TRANSPORT ROBUSTNESS

The breadth audit repeats the stagewise-versus-composite experiment on three public SmolLM2 base checkpoints: `HuggingFaceTB/SmolLM2-135M` at revision `28e66ca`, `HuggingFaceTB/SmolLM2-360M` at revision `3ce05f6`, and `HuggingFaceTB/SmolLM2-1.7B` at revision `4fa12ca`. The archived configurations have $(L, D, d) = (30, 60, 576)$, $(32, 64, 960)$, and $(24, 48, 2048)$, with parameter counts 134,515,008; 361,821,120; and 1,711,376,384. The first two use grouped-query attention with $9/3$ and $15/5$ query/key-value heads; the 1.7B checkpoint has $32/32$ heads. All three use RMS-normalization regularizer $\nu = 10^{-5}$.

The protocol freezes 32 deterministic WikiText-103 validation windows of length 128. The three checkpoint tokenizers are verified to agree on the selected text before the common token array is frozen. No training or weight perturbation is performed. Execution is FP32 on an NVIDIA A100-SXM4-40GB with TF32 disabled. For each model–record pair, every residual-sublayer block-$\ell_{\infty,2}$ norm is estimated with five power iterations and two deterministic restarts, while the full composite uses ten iterations and three restarts. The full run therefore contains 96 composite record evaluations and $32(60+64+48) = 5{,}504$ residual-sublayer evaluations. All 117 entries in the uploaded SHA256 manifest verify.

Checkpoint	stagewise product	composite	paired gap	gap min / max
SmolLM2-135M	53.66	4.33	49.21	48.30 / 51.39
SmolLM2-360M	58.75	4.49	54.23	52.47 / 56.85
SmolLM2-1.7B	51.27	5.04	46.17	41.99 / 49.30

Table 15: Per-model $\log_{10}$ summaries for the three SmolLM2 checkpoints. The first three columns are medians of per-record quantities; the paired gap is computed recordwise. The smallest observed gap remains above 41 decades in every model. All operator quantities are finite-search lower estimates.

The composite searches are numerically stable: the median absolute penultimate-to-final change is 7.83×10^{-8}, 1.24×10^{-7}, and 1.10×10^{-7} decades for the 135M, 360M, and 1.7B checkpoints; the largest change among all 96 composite searches is 3.87×10^{-5}.

A prespecified record-0 window diagnostic gives estimated product-minus-composite gaps of $(3.71, 6.76)$ decades after the first $(4, 8)$ stages in the 135M model, $(3.16, 7.15)$ in the 360M model, and $(2.91, 7.33)$ in the 1.7B model. Thus the loss from independent stagewise maximization recurs within the stack and across all three additional pretrained checkpoints. The conclusion remains geometric rather than certificational: both the stagewise product and the direct composite are lower estimates, and none of these trajectory quantities is a uniform upper transport bound over the reference-centered parameter class.