REACHABILITY FLOORS OF LOW-RANK LANGUAGE-MODEL HEADS

Anonymous authors
Paper under double-blind review

ABSTRACT

Final normalization bounds the hidden states presented to a language model's output head, so a fixed head can realize only a restricted set of token distributions. With this finite state budget, reachability depends on the head's scale as well as its rank. We define the reachability floor as the infimum target loss over normalization-feasible states and compute two-sided bounds at full vocabulary. For one-hot targets the average floor gives an exact decomposition of validation loss into a head-conditioned floor and a realization gap. Across a controlled intervention and two large-model heads, these floors separate into two regimes.

In a paired multi-seed experiment, the factorized low-rank head has a much higher one-hot floor than the full-rank-forward control, while their validation losses differ far less. Bias-fixed matching of the learned output-map scale removes most of this contrast. Typical realized losses nevertheless remain well above their floors. Context-conditioned teacher targets reveal the second regime: their floors are smaller, but they make up a substantial share of realized teacher–student divergence, and their absolute floors are comparatively robust to uniform scale matching. A factorization-aware initialization and decay scheme improves both kinds of floor and validation loss relative to same-variance factorization, and an independent twenty-seed replication reproduces the improvement, with mean validation loss close to the full-rank-forward control's. A separate observational comparison of two 2B checkpoints also finds a large separation in teacher-distribution floors.

The theory separates the roles of rank and scale, and the experiments distinguish uniform-scale sensitivity from residual learned-head geometry in output reachability. Reachability floors complement validation loss by locating restrictions in the learned output head.

1 INTRODUCTION

In many modern language models the last operation is a linear map $W \in \mathbb{R}^{V \times d}$ applied to a hidden state, followed by a softmax. With V far larger than d, the head restricts log probabilities to a low-dimensional family, the softmax bottleneck (Yang et al., 2018). In the models studied here, final normalization imposes a second restriction. It bounds the hidden-state norm, so reachability depends on the head's scale as well as its rank. The centered map determines the available logit subspace, rank counts its dimensions, and scale sets how far the model can move within it. With the learned bias fixed, uniformly scaling the map is exactly equivalent to enlarging the state budget (Theorem 1). Uniform head scaling is irrelevant in the unbounded formulation and consequential under a finite state budget.

Prior work characterizes the restriction as representability (Yang et al., 2018), argmaxability (Grivas et al., 2022), hull-interior probability limits (Demeter et al., 2020), and bounds on the most likely tokens (Basri & Jacobs, 2026). Closest to our setting, Ganea et al. (2019) characterize the infimum cross-entropy of a fixed, bias-free head over unconstrained hidden states. We study the finite-radius, bias-aware counterpart induced by final normalization, bound it for learned heads at vocabulary scale, and use it to decompose validation loss.

The decomposition. For a frozen head, the *reachability floor* of a token target is the infimum of its NLL over normalization-feasible states. The *average one-hot reachability floor* F is the mean

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.

of these floors over the evaluation targets. Any model using that head and state interface satisfies $L \geq F$, so

$$L = F + G, \qquad G = L - F \geq 0, \tag{1}$$

where G, the *realization gap*, is the residual between the realized loss and the head-conditioned floor. Both terms depend on the learned head because G is defined relative to F, and they need not move together: the sum can change little while each changes substantially.

What we find. Theorem 1 equates map scale with state budget, and for the tested parameterization Propositions 4 and 5 predict the $\sigma\sqrt{k}$ initial scale ratio (initialization standard deviation σ, inner rank k) and one additional decay contraction, predictions we verify directly. A paired controlled comparison provides the setting: a rank-48 factorized-forward head against a full-rank forward head whose backward signal is projected into 48 dimensions, matched within each seed. The factorized-forward arm ends with an average floor 1.62 nats higher and a validation NLL 0.101 nats higher, raising two questions the theory makes precise: *how much of the contrast is scale-recoverable*, and *whether training can reduce it*. With the learned bias fixed, amplifying the trained map to the control's centered scale is exactly a budget relaxation, and it removes most of the floor contrast in every seed. A larger gain reverses the floor ordering. Applying the same gains at inference sharply raises validation loss (§5). The factorization-aware training scheme the propositions motivate lowers the floor contrast and improves validation loss, and an independent twenty-seed replication reproduces the improvement. Context-conditioned teacher targets show a second regime: those floors are smaller in absolute terms, yet their between-arm differences remain substantial under the tested matching conventions. The contrasts reproduce on twenty fresh seeds, and a separate observational comparison of two 2B checkpoints finds a large teacher-distribution floor difference. The rescaling analyses characterize sensitivity at the trained endpoints. Whether scale mediates the contrast during training is a separate question.

This paper makes three contributions.

- **Finite-budget reachability theory.** With the bias fixed, scaling the centered effective map is exactly enlarging the feasible-state radius, the floor's scale response is convex with a shadow-price characterization, row margins give exponentially decaying upper bounds under amplification, and tokens whose centered head row lies in the convex hull of the competing rows keep a scale-independent positive lower bound. For the tested two-factor parameterization, reconstructed initial models agree with the predicted $\sigma\sqrt{k}$ RMS-Frobenius scale ratio, and under the isolated decay operator factorization adds one relative contraction.
- **Empirical separation of scale-sensitive and persistent restrictions.** In the controlled five-seed comparison the one-hot floor contrast is $16\times$ the validation-NLL contrast by point estimate and is mostly scale-recoverable: bias-fixed matching shrinks it in every seed while leaving it positive, further amplification reverses the ordering, and applying the matching gains at inference raises realized loss. Teacher-distribution floors are smaller but persist under the tested matching conventions, their contrasts replicate on twenty fresh seeds, and an observational 2B comparison also shows a large teacher-distribution floor separation.
- **A theory-guided training intervention with independent replication.** The factorization-aware scheme motivated by the two predictions matches the predicted initial scale and corrects the extra decay contraction. It lowers both kinds of floor and improves validation loss over the same-variance factorization, and across twenty fresh seeds the one-sided 95% upper bound on the mean validation-loss difference from the full-rank-forward arm is below 0.05 nats (§5).

2 THE FLOOR AND THE DECOMPOSITION

Output layer and feasible states. Let $U \in \mathbb{R}^{V\times d}$ and $b_0 \in \mathbb{R}^V$ be the output head as it appears in the network, and let $\hat{h}$ be the normalized pre-affine state of the final normalization layer (Ba et al., 2016; Zhang & Sennrich, 2019; Xiong et al., 2020). Folding that layer's learned gain γ and shift β into the head gives

$$z = U(\gamma \odot \hat{h} + \beta) + b_0 = W\hat{h} + b, \qquad W = U\mathrm{diag}(\gamma), \quad b = b_0 + U\beta. \tag{2}$$

For LayerNorm, write $\hat{h} = Q_\perp h$ with $Q_\perp$ an orthonormal basis of $\mathbf{1}^\perp$, and replace W by $WQ_\perp$. For RMSNorm, set $h = \hat{h}$. Below, h and W denote the resulting coordinate and effective map, so

that $z = Wh + b$ holds exactly. All budget comparisons below refer to the normalized coordinate $\hat{h}$. The normalization fixes the feasible set at $\hat{h}$. Under **LayerNorm**, $\hat{h}$ is centered with $\|\hat{h}\|^2 = d\,\sigma^2/(\sigma^2 + \epsilon) < d$, where σ^2 is the coordinate variance, so it lies strictly inside the ball of radius $R = \sqrt{d}$ and approaches it only as $\epsilon/\sigma^2 \to 0$. The state also lies in $\mathbf{1}^\perp$, so the reduced coordinate h has $d' = d - 1$ entries (Brody et al., 2023). This is the geometry of the controlled model ($d = 96$, $d' = 95$, $R = \sqrt{96}$). Under **RMSNorm** there is no centering and no normalization shift, so $\hat{h}$ occupies the uncentered d-dimensional ball $\|\hat{h}\| < \sqrt{d}$ with no $\mathbf{1}^\perp$ reduction. The evaluated checkpoints use a bias-free output head, so $b = 0$. This is the geometry of the 2B models ($d = 4096$, $R = 64$). In both cases normalized states satisfy $\|\hat{h}\| < R = \sqrt{d}$.

The target-distribution floor. Norms are Euclidean throughout. For a target distribution q define

$$T_R(q) \;=\; \min_{\|h\| \le R} D_{\mathrm{KL}}\big(q \,\|\, \mathrm{softmax}(Wh + b)\big). \tag{3}$$

Normalized states range over the open ball. By continuity, the infimum over that open set equals the minimum over its closed-ball closure B_R, which we compute. A boundary minimizer may not be attained pre-normalization, but its value is the exact infimum. This is a convex program: $D_{\mathrm{KL}}(q \,\|\, \mathrm{softmax}(\cdot))$ is convex in the logits and the logits are affine in h, over a convex feasible set. At a one-hot target $q = e_v$ it reduces to the infimum target NLL over normalized states, equivalently the minimum over B_R,

$$T_R(e_v) \;=\; \min_{\|h\| \le R} \; -\log \mathrm{softmax}(Wh + b)_v \;=\; -M_R(v) \;\ge\; 0, \tag{4}$$

where $M_R(v)$ is the maximum log-probability of token v over B_R, equivalently its supremum over the open feasible set induced by normalization. For finite R, finite parameters and $V \ge 2$, $T_R(e_v)$ is strictly positive. A small value means the frozen head can assign v probability close to one within the budget, and its unbounded-radius limit may vanish when v can be made arbitrarily likely. Appendix A proves this and derives the dual used in §4.

Average floor and realization gap. Let y_i denote the target at evaluation position $i = 1, \ldots, n$, and let L be the mean negative log-likelihood over these positions. Define

$$F \;=\; \frac{1}{n} \sum_{i=1}^{n} T_R(e_{y_i}), \qquad G \;=\; L - F. \tag{5}$$

F is the *average one-hot reachability floor*, frequency-weighted and target-aware: at each position it takes the infimum loss the learned head can assign to that position's target, with a separately optimal feasible state allowed at every position. Two heads with equal F may still differ in which distributions they can emit. The model's realized state at position i is feasible, so its loss there upper-bounds $T_R(e_{y_i})$. Averaging over positions gives $F \le L$ and $G \ge 0$ in exact arithmetic. §4 records the corresponding numerical comparison as a diagnostic.

Among lower bounds that use only per-position feasibility, F is the sharpest: it is the minimum of the mean target NLL over independently chosen feasible states, and any uniform feasible-state bound lies below it (Appendix A).

G measures the distance from realized loss to the per-target floor. Because the floor allows a different target-aware state at every position while the model's state comes from shared parameters, G can reflect conditional uncertainty and one shared trunk serving every context. Both terms may remain positive at convergence.

The floor F is specific to the learned head and its feasible-state set.

3 Rank and scale under final normalization

Adding a constant to every logit leaves the softmax unchanged, so only the centered map matters. Write $P = I - \frac{1}{V}\mathbf{1}\mathbf{1}^\top$ for that centering. Every head scale in this paper is measured on PW, since $\|W\|$ itself is gauge dependent. The closure of the reachable centered-logit set is

$$\mathcal{Z}_R \;=\; Pb + PW\,B_R, \qquad B_R = \{h : \|h\| \le R\}, \tag{6}$$

a possibly degenerate ellipsoid centered at Pb with semi-axes $R\,\sigma_i(PW)$: the rank of PW counts the available directions, the singular vectors orient them, the singular values and R set their extent, and the bias places the center.

Theorem 1 (Scale is radius)**.** *For every $c > 0$, $T_R(q; cW, b) = T_{cR}(q; W, b)$, and the reachable sets agree, $\mathcal{Z}_R(cW, b) = \mathcal{Z}_{cR}(W, b)$.*

Proof. Substitute $u = ch$. As h ranges over B_R, u ranges exactly over B_{cR}, and $cWh + b = Wu + b$. □

Multiplying W by c preserves the rank of PW, its range, and the unbounded affine family $Pb + \mathrm{range}(PW)$, while the finite-radius ellipsoid dilates. Final normalization fixes the state budget, so output-head scale controls how far the model can move along its available logit directions.

Proposition 2 (Convex scale response)**.** *Fix (W, b, q). The map $\phi(r) = T_r(q; W, b)$ is finite, continuous, nonincreasing and convex on $r \geq 0$. For $r > 0$, $\phi'(r) = -\lambda_r$ wherever the derivative exists, where $\lambda_r \geq 0$ is the multiplier of the radius constraint, and at a boundary solution $\lambda_r = \|W^\top(p_r - q)\|$ with p_r the optimal softmax. (Proof in Appendix B.)*

Through Theorem 1, $\frac{d}{dc}\,T_R(q; cW, b) = -R\,\lambda_{cR}$ where the derivative exists, and convexity makes λ_r nonincreasing: amplification has diminishing returns. A strictly interior optimizer implies $\lambda_r = 0$, and a positive multiplier implies the budget binds. A zero multiplier means additional scale has zero local value, though the selected optimizer need not be interior.

Corollary 3 (Margins give exponential suppression)**.** *For a token v, with w_j the rows of the folded, reduced W, define $\gamma_v = \max_{\|u\|\leq 1} \min_{j\neq v}(w_v - w_j)^\top u$. If $\gamma_v > 0$, then for every $c > 0$,*

$$T_R(e_v;\, cW, b) \;\leq\; \log\Big[1 + e^{-cR\gamma_v} \textstyle\sum_{j\neq v} e^{\,b_j - b_v}\Big]. \tag{7}$$

Since $u = 0$ is feasible, $\gamma_v \geq 0$ always, with $\gamma_v = 0$ exactly when w_v lies in the convex hull of the other rows, the hull condition of Demeter et al. (2020), and then amplification cannot eliminate the floor: for any convex representation $w_v = \sum_{j\neq v} \mu_j w_j$, for every c and R,

$$T_R(e_v;\, cW, b) \;\geq\; \log\Big[1 + e^{\,\sum_{j\neq v} \mu_j b_j \,-\, b_v}\Big] \;>\; 0, \tag{8}$$

a positive lower bound independent of scale and radius (proof in Appendix B). Amplification can reduce the finite-radius loss while this positive lower bound remains in force. It cannot create a separating direction or drive the loss below the bound. For tokens with positive margin, scale suppresses the competing logits exponentially in $cR\gamma_v$ without changing the reachable subspace.

Why the factorized head starts smaller. We analyze initialization and factorwise decay for the two-factor parameterization used in the controlled experiment.

Proposition 4 (Same-variance factor initialization)**.** *Let every entry of the raw direct head $U \in \mathbb{R}^{V\times d}$ and of the factors $A \in \mathbb{R}^{V\times k}$, $B \in \mathbb{R}^{k\times d}$ be independent $\mathcal{N}(0, \sigma^2)$, and let $Q_\perp$ span $\mathbf{1}^\perp \subset \mathbb{R}^d$. Then $\mathbb{E}\,\|PUQ_\perp\|_F^2 = (V-1)(d-1)\,\sigma^2$ and $\mathbb{E}\,\|PABQ_\perp\|_F^2 = (V-1)(d-1)\,k\,\sigma^4$, so the ratio of RMS Frobenius scales is $\sigma\sqrt{k}$. (Proof in Appendix B.)*

For the configuration used here ($\sigma = 0.02$, $k = 48$) the predicted RMS-Frobenius ratio is 0.139, and deterministic reconstructions of the five initial models give 0.136 to 0.141. The deficit is a property of the same-variance initialization itself, and scale-aware factor initialization and product-aware regularization address it (Khodak et al., 2021).

Proposition 5 (Factorwise decay)**.** *Apply decoupled weight decay to every parameter with per-step multiplier $a_t = 1 - \eta_t\lambda$ (learning rate η_t, decay coefficient λ), and suppress the gradient updates. With $\alpha = \prod_t a_t$, a raw direct head U maps to αU and a two-factor product AB to $\alpha^2 AB$. Since the normalization gain decays as well, the folded direct and factorized maps receive α^2 and α^3.*

With gradient updates suppressed, the factorized map receives one additional contraction factor α relative to the direct map, equal to 0.740 over this schedule. During training, gradient updates and factor cross-terms also affect the learned scale. LayerNorm's shift initializes at zero, so both folded biases start at zero and the initial deficit is a property of the effective map alone.

4 COMPUTING TWO-SIDED BOUNDS ON THE FLOOR

We compute lower and upper bounds on each floor under the arithmetic model specified in Appendix C. Write $H(p) = -\sum_i p_i \log p_i$ with $0 \log 0 = 0$. The dual of equation 3 is the problem $\max_{\mu \in \Delta_V} D_R(\mu)$, where $\Delta_V \subset \mathbb{R}^V$ is the probability simplex and

$$D_R(\mu) \;=\; H(\mu) - H(q) + (\mu - q)^\top b \;-\; R \left\| W^\top (\mu - q) \right\|. \tag{9}$$

Appendix A derives this expression and establishes strong duality. We use only the weak inequality $D_R(\mu) \le T_R(q) \le D_{\mathrm{KL}}(q \,\|\, \mathrm{softmax}(Wh + b))$, which holds for every $\mu \in \Delta_V$ and every feasible h, so a feasible state and a simplex witness yield a two-sided bracket.

For a valid enclosure, the computed softmax is normalized exactly onto the simplex in rational arithmetic and every endpoint is rounded outward, with the error model and the full-vocabulary reformulation in Appendix C.

We aggregate the per-target bounds using exact integer counts and rational arithmetic with directed rounding. Write F_{hi} for the upper endpoint of the aggregated floor enclosure. In the one-hot runs every rigorous upper endpoint is at most the best observed fp64 NLL for that target, and every run satisfies $F_{\mathrm{hi}} \le L$. Neither comparison enters the reported bounds.

Validity. A bracket is valid when it is finite, ordered, and attached to a feasible state. The evaluator requires a valid bracket for every target of positive empirical frequency. Every run included in the analysis satisfies this requirement. The widest per-target bracket is 9.47×10^{-11} nats in the original one-hot runs at the architectural radius, 1.49×10^{-10} in the replication's one-hot runs, 6.50×10^{-6} at $8R$, and for dense teacher targets 1.02×10^{-3} per context (5.59×10^{-5} for aggregate N), rising to 3.24×10^{-2} (1.51×10^{-3}) in the rescaled cells. The 2B aggregate floor intervals are 2.5–3.2×10^{-5} nats wide, both negligible relative to the contrasts they enter. All displayed interval and one-sided-bound endpoints are rounded outward at the displayed precision. Point estimates and descriptive ranges are rounded to nearest.

Per-run bounds and across-seed intervals. Each per-run floor lies in a two-sided numerical enclosure, so each seed contributes an interval rather than a point. Across-seed intervals propagate those bounds through the seed-level Student-t calculation, taking the smallest interval that encloses the Student-t intervals of all 2^n vertices of the box (32 at five seeds, 1,048,576 at twenty), since the endpoints are convex and concave in the sample vector (Appendix C). We call the result a *bounds-propagated* interval. The numerical enclosures for F, N, ΔF, and ΔN do not depend on the stored-loss values. L is a stored fp64 value, so ΔL and ΔG intervals are conditional on those stored losses, with the F bounds propagated through $G = L - F$. The enclosures cover the measured seed-level effects rigorously, while population coverage additionally assumes the seed-level effects are independent draws from a common, approximately Gaussian distribution.

5 EXPERIMENTS: SCALE SENSITIVITY AND FACTORIZATION-AWARE TRAINING

Design. We use the controlled subword setting of Murugan (2026): a 6-layer transformer with $d = 96$, $V = 8{,}192$, and sequence length 128, trained on WikiText-2 for 1,200 steps under a common learning-rate schedule across arms (configuration in Appendix D). The setting includes an unrestricted baseline and two nominal-rank-48 interventions that place the restriction at different points in the computation. The three arms are:

A plain full-rank forward head, ordinary backpropagation.

B full-rank forward head with a *backward-only* rank-48 restriction. The forward pass is untouched, but the gradient sent back to the trunk is projected onto the 48-dimensional subspace spanned by the head's leading right singular directions, so the signal reaching the trunk is 48-dimensional while the forward head stays full rank. The projected gradient is rescaled to match the root-mean-square magnitude of the ordinary backward gradient, and the subspace is recomputed every 20 steps.

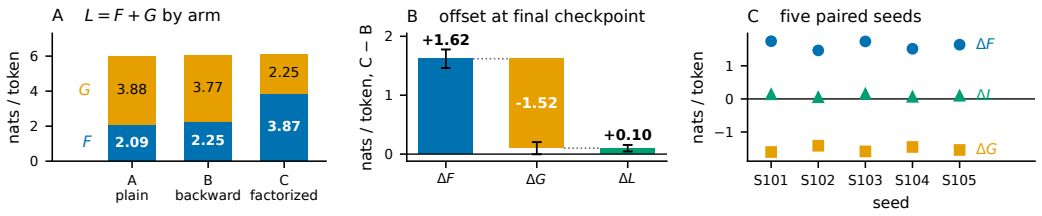


Figure 1: **Floor and validation contrasts.** **A**: the decomposition $L = F + G$ per arm, averaged over five seeds. **B**: the paired $C - B$ contrast. **C**: the same contrast per seed. Error bars are the 95% intervals of §4.

C rank-48 *factorized-forward* head, with ordinary backpropagation through it. The output map is written as a product $W_{\text{lm}}W_{\text{proj}}$ with $W_{\text{proj}} \in \mathbb{R}^{48\times 96}$ and $W_{\text{lm}} \in \mathbb{R}^{V\times 48}$, so the state is projected to 48 dimensions before reaching the vocabulary and the effective head has rank at most 48.

Restricting the gradient alone tests whether compressing the signal reaching the trunk to 48 dimensions is harmful, the mechanism proposed by Godey & Artzi (2026), while factorizing the forward map also limits the logits reachable from feasible states. Arm B is the nominal-rank-48 comparator. The backward signal is restricted in both arms: explicitly in B, and in C because ordinary backpropagation through the factorization leaves the gradient reaching the trunk rank restricted as a consequence.

Within each seed the arms share the data order and initial backbone parameters. A 100-step full-rank diagnostic confirms agreement between the custom and ordinary backward passes at relative tolerance 10^{-5}. Seeds are the units of inference, five in the original experiment. Each run is evaluated on all 286,727 positions of the prepared validation stream, with 7,366 distinct targets evaluated per run.

Result. The $C - B$ comparison gives a one-hot-floor difference of $+1.620$ nats and a validation-NLL difference of $+0.101$ nats (Table 1). The 0.101-nat increase is statistically resolved and corresponds to a perplexity multiplier of 1.106. The corresponding G difference is -1.519 nats: the factorized arm sits closer to a higher floor, and this opposing change offsets 93.8% of the F difference by point estimate.

Positionwise loss headroom. For position i with realized NLL ℓ_i, define $s_i = \ell_i - T_R(e_{y_i})$. This is the upper limit on its NLL reduction if that position could select any feasible hidden state independently while the head, bias, and state interface remain fixed. Its mean is G. At the first percentile, the ordering varies by seed. At the 5th, 10th, 25th, and 50th percentiles, arm C has lower headroom than arm B in every seed. Per-seed medians are 1.98–2.03 nats for C and 3.41–3.63 for B. Fifth percentiles are 0.55–0.60 and 1.17–1.26. Arm C's median headroom is about two nats, leaving substantial scope between realized loss and the per-target floor. Appendix F reports the per-seed empirical cumulative distribution functions of s_i, its quantiles, and the secondary floor-tercile analysis.

Learned scales and sensitivity. At the final checkpoint the factorized head's centered folded scales are about a third of arm B's, a gap the initialization and decay mechanisms of §3 anticipate (measurements in Appendix D). The centered bias norm $\|Pb\|$ is also far smaller. Subsequent training also shapes the final ratio. Within each seed we amplify arm C's folded map to arm B's centered scale, bias fixed, an exact radius change (Theorem 1) evaluated directly by the bounding procedure. Spectral matching equates the largest semiaxis of the two centered logit ellipsoids, and Frobenius matching the total squared semiaxis extent. Both are frozen-head sensitivity conventions, and their residual contrasts retain the bias difference, spectrum shape, row arrangement, and training history. Under the primary spectral convention per-seed gains span 2.56–3.45 with matched contrast $[0.111, 0.252]$, and under the Frobenius convention gains span 2.25–2.75 with matched contrast $[0.336, 0.450]$. Both matched contrasts remain positive in every seed. Per seed the identity $F_C(R) - F_B(R) = [F_C(R) - F_C(c_sR)] + [F_C(c_sR) - F_B(R)]$ is exact: at the point-estimate level

Table 1: Main contrasts. Panel A: the raw $C - B$ comparison on the original five seeds. Italics are point estimates, brackets 95% intervals (ΔF bounds-propagated, ΔL and ΔG conditional on stored fp64 losses with the F bounds propagated, Student-t across seeds). The five-seed C' aggregates are in Appendix D. Panel B: the independent twenty-seed replication. The validation-NLL rows report each mean and its one-sided 95% upper bound across the twenty seeds, and the remaining rows are bounds-propagated two-sided 95% intervals. C' is the factorization-aware variant of arm C and N the teacher-distribution floor, both introduced in this section.

A. original five seeds

	ΔF (one-hot floor)	ΔL (validation NLL)	ΔG (gap)
raw $C - B$	$+1.620$ [1.462, 1.777]	$+0.101$ [0.045, 0.156]	-1.519 [−1.623, −1.415]

B. independent twenty-seed replication

quantity	contrast	result
validation NLL	$L_{C'} - L_C$	mean -0.086, $U_{95} = -0.072$
validation NLL	$L_{C'} - L_B$	mean 0.013, $U_{95} = 0.027$
one-hot floor	$F_{C'} - F_C$	$[-0.745, -0.639]$
one-hot floor	$F_{C'} - F_B$	$[0.886, 0.977]$
teacher-distribution floor	$N_C - N_B$	$[0.071, 0.085]$
teacher-distribution floor	$N_{C'} - N_C$	$[-0.056, -0.042]$
teacher-distribution floor	$N_{C'} - N_B$	$[0.024, 0.035]$

the spectral rescaling removes 1.438 of the 1.620-nat contrast. The rescaling holds arm C's learned bias fixed, so the between-arm bias difference remains part of the matched contrast.

Crossover under amplification. $D(c) = F_C(cR) - F_B(R)$ is continuous, convex and nonincreasing in c (Proposition 2), so $c^\star = \inf\{c \geq 1 : D(c) \leq 0\}$ is well defined. In every seed the lower endpoint of $D(2)$ is positive and the upper endpoint of $D(8)$ negative, and exact-rational secant and chord constructions bracket each seed's crossover within 2.93 to 4.18 (per-seed brackets in Appendix D): at a finite gain the contrast crosses zero in every seed, under an amplification that preserves rank and every logit direction. At a common eightfold budget, B's mean floor falls to 1.0×10^{-5} nats while C retains 1.227, with $C - B \in [1.155, 1.300]$.

Against arm A, the $C - A$ floor difference is resolved at five seeds, while the $B - A$ interval includes zero there. The twenty-seed replication below resolves it (Appendix D).

Inference-time rescaling. With each arm-C checkpoint's states and bias fixed, validation NLL rises by [3.519, 6.153] nats at the spectral gain and [2.618, 4.014] at the Frobenius gain (paired intervals across five seeds from stored fp64 point values). The same gains lower the average floor but sharply raise validation loss at the frozen checkpoint. C' tests a different intervention: jointly changing factor initialization and product decay during training.

Factorization-aware training. A fourth arm C' repeats arm C with a factorization-aware initialization and product-decay correction: factor draws rescaled to the direct head's expected squared centered folded Frobenius scale, and factor decay corrected to the direct head's product multiplier under the isolated decay operator. C' estimates the combined effect of the initialization and decay corrections. Across the original five seeds, C' has a lower floor than C and a higher floor than B in every seed. The per-seed enclosures of the retained fraction $(F_{C'} - F_B)/(F_C - F_B)$ span [0.523, 0.642] (Appendix D). Relative to C the floor falls and the realization gap rises: the reduction in the floor is partly offset by a larger realization gap, so validation loss improves by much less. Final centered spectral scales remain well below arm B's (34.2–41.3 against 98.8–104.4). Appendix D reports the per-seed outcomes, exact recipe, and head-geometry summaries.

Twenty-seed replication. Across twenty new seeds, mean $L_{C'} - L_C = -0.086$ with one-sided 95% upper bound -0.072, and mean $L_{C'} - L_B = 0.013$ with upper bound 0.027, below the 0.05-nat margin. That margin is about half the original 0.101-nat difference and corresponds to a 5.1% perplexity increase. The one-hot and teacher-distribution contrasts keep their directions (Table 1,

panel B). Against the unrestricted arm, mean $L_{C'} - L_A = 0.065$ with two-sided 95% interval $[0.046, 0.083]$. Appendix D reports the seedwise signs and the arm-level summaries.

Teacher-distribution targets. With a teacher distribution q, $N = \mathbb{E}\, T_R(q)$ is the floor and K the realized divergence, each averaged over the evaluated contexts, and $Q = N/K$ is the unremovable share, the ratio of those averages, reported only when the lower endpoint of K is at least 10^{-3} nats. The same question for context-conditioned targets uses arm A as within-seed teacher on a fixed list of $4{,}096$ contexts. The restriction is pronounced when asking how strongly the head can favor the realized label, but weaker for context-conditioned target distributions. On the original five seeds, per-seed N_B spans 0.035–0.053 nats and N_C 0.115–0.125, far below the one-hot floors. The ordering persists: $N_C - N_B \in [0.072, 0.086]$, $N_{C'} - N_C \in [-0.061, -0.043]$, and the unremovable share is 0.50–0.53 for C against 0.24–0.32 for B. Under spectral and Frobenius matching, the changes in N_C relative to the unmatched cell are $[-0.020, -0.013]$ and $[-0.017, -0.011]$. In those rescaled arms the realized divergence K rises sharply and Q falls (Appendix D). The teacher is itself a trained model, so the diagnostic is teacher-relative (per-seed values in Appendix D).

The 2B comparison. At larger scale we compare two observational cells of a 2B suite (Godey & Artzi, 2026): a rank-32 and a rank-512 head evaluated against the same $d = 4096$ teacher. The rank-32 floor's lower endpoint is at least 13.8 times the rank-512 upper endpoint, with unremovable shares about 60.6% and 12.6%. The centered spectral scales differ by 0.261%: matching them requires amplifying the rank-512 map by the factor $c = 1.0026$. Write $\underline{N}$ and $\overline{N}$ for enclosure endpoints. Theorem 1 equates this amplification with radius cR, and since $1 < c < 2$, monotonicity gives $0 \le N_{512}(R) - N_{512}(cR) \le \overline{N}_{512}(R) - \underline{N}_{512}(2R) \le 2.1 \times 10^{-6}$ from the stored enclosures: the contrast persists beyond uniform centered spectral scale. The match is spectral, and the centered Frobenius scales remain far apart (rank-32 over rank-512: 0.433). Under the rank-512-referenced Frobenius convention, amplifying the rank-32 map by 2.31 stays within the evaluated $8R$ range, and monotonicity with the radius ladder gives a lower bound of 0.40697 nats on the matched rank-32 floor, still at least 13.5 times the rank-512 upper endpoint. The residual summarizes the remaining differences between the two learned heads.

The endpoint decomposition measures how the final-checkpoint difference divides between F and G. How training produces these opposing between-arm differences remains open.

6 Related work

Attainable loss. Ganea et al. (2019) provide the direct predecessor: their Theorem 2 characterizes the unconstrained-state minimum cross-entropy of a fixed, bias-free head. T_R adds the architecture's finite budget and the learned bias, bounded at vocabulary scale and aggregated into $L = F + G$.

Token-level reachability. Demeter et al. (2020) identify stolen probability for tokens interior to the convex hull of a bias-free dot-product head. Grivas et al. (2022) give the bias-aware criterion and an exact detector over a specified bounded input region. Basri & Jacobs (2026) bound top-m specifiability with unbounded norm. These results concern which tokens or top-m sets a head can favor, and how hull geometry bounds an interior token's probability. For a fixed head, Equation 3 defines the minimum target loss over the closed feasible ball B_R.

Factorized layers. Khodak et al. (2021) observe that standard factor initialization misscales the product, and propose spectral initialization, which starts the factors so their product approximates a target unfactorized initialization, and Frobenius decay, which regularizes the product rather than the individual factors. Kamalakara et al. (2022) train GPT-style models with factorized layers and caution that singular-value scale alone does not track predictive performance. Our propositions quantify initialization and isolated-decay effects on output-head scale under final normalization, where Theorem 1 converts scale into state budget.

Causal tests of the head. Murugan (2026) separates geometry from causation with paired seeds and finds that in the half-rank subword experiment a factorized forward head costs roughly three times an equally ranked backward-only restriction, concluding that strong geometric compression is confirmed but not established as harmful. We adopt that design and add the target-wise reachability floor the learned head imposes.

Geometry of hidden states. Trained representations are typically anisotropic in angle, so a tighter measured state set can only raise these floors. Appendix E adds the bottleneck-rank literature, output-layer conventions, and the convex-analysis toolkit.

7 DISCUSSION

Two regimes. The one-hot and teacher-distribution floors behave differently. The one-hot contrast is strongly scale-sensitive, and most realized losses sit well above their floors, so it chiefly reveals latent restrictions in the state-to-output interface. Teacher-distribution floors are far smaller in absolute terms yet carry a substantial share of realized teacher divergence and respond much less to uniform rescaling, in the controlled arms and in the observational 2B comparison alike. The factorization-aware training scheme improves both kinds of floor and validation loss, retains a positive floor difference from B, and its improvement estimates the joint effect of a changed training trajectory, spanning head, trunk, bias, and optimization path.

What the decomposition adds. Validation NLL reports the sum $F+G$. The decomposition shows how a between-arm difference divides between the head-conditioned floor and the realization gap, and here the between-arm changes in F and G are much larger in magnitude than their sum, ΔL. A desired average loss below the fixed-head floor requires changing the head or its state interface. In the tested parameterization, the joint correction to factor initialization and product decay improves validation loss. Head comparisons should report effective-map scale alongside rank, which alone leaves the reachable geometry unspecified. Intervention comparisons benefit from reporting both the floor and realized loss: inference-time amplification lowers F while raising L. For teacher targets, N bounds the achievable teacher–student divergence from below, and $Q = N/K$ measures its unremovable fraction. Whether these diagnostics select better architectures more broadly remains open.

Limitations. The primary experiment is one architecture ($d = 96$, $V = 8{,}192$, untied output head), one corpus, one rank pair, and one training horizon, with five original and twenty replication seeds. The $C - B$ contrast compares two complete interventions with different forward maps and rank-48 backward operators and does not isolate a forward-only causal effect. C' changes initialization and decay jointly and with them the full training trajectory, so the experiment estimates that bundled intervention, and mediation through F is unidentified. The 2B comparison is observational. F is a property of the entire learned head, and two heads of equal rank can have very different floors, so our results characterize the full learned geometry of these specific heads.

8 CONCLUSION

Final normalization makes head scale part of the constraint on output distributions: rank counts the available directions, and scale sets how far the model can move along them. With the bias fixed, the scale–radius identity gives an exact probe of that constraint, preserving rank and every logit direction. The target matters as well. Much of the one-hot floor contrast is scale-recoverable, while substantial teacher-distribution floor differences persist under the tested matching conventions, in the controlled arms and in an observational 2B comparison alike. Rank and uniform scale therefore leave important aspects of learned-head geometry unspecified.

The same theory motivates a training intervention. For the tested parameterization, jointly correcting factor initialization and product decay improves both kinds of floor and validation loss relative to same-variance factorization, and twenty fresh seeds reproduce the improvement. At the tested gains, amplifying the frozen arm-C checkpoints lowers the floor while worsening validation loss. Reachability and predictive performance therefore carry distinct information when a head intervention is evaluated.

Head comparisons should include effective-map scale alongside rank, and reachability floors alongside realized loss. The floor states a fixed-head feasibility limit, while validation loss measures the trained model's performance. A desired average loss below F requires changing the head or its state interface.

AI USE STATEMENT

Generative AI tools assisted with coding and prose revision, including trimming, grammar checking, and stylistic polishing. The authors conceived the research ideas, developed the theory, designed the experiments, reviewed the code, and analyzed and interpreted the results. The authors take full responsibility for the work.

REPRODUCIBILITY STATEMENT

Validation code and the stored analysis records are available at `https://anonymous.4open.science/r/repo-B99C/`. One script, `reproduce.sh`, runs the numerical regression suites for the solver, evaluator, and directed-interval machinery, verifies the full-rank custom backward pass against ordinary backpropagation at relative tolerance of 10^{-5}, and recomputes selected aggregates and contrasts from the records, checking those results and the artifact hashes against what the records hold. The headroom distribution arrays are rebuilt from the per-seed arrays the export carries and checked against their recorded hash. The README lists the dependencies and hardware requirements and identifies the larger arrays the export does not carry, each with the hash recorded for it. The script neither regenerates nor independently verifies those. Training, checkpoint evaluation, display and figure generation, and the manuscript build are outside its scope: those stages run at pinned code and dataset revisions and write versioned, checksummed artifacts, and the validation runs against the stored versions. Section 2 states the assumptions behind the decomposition, Appendix A proves the duality results, Appendix B proves the scale results, Appendix C gives the complete bounding procedure behind the reported enclosures, and Appendix D reports per-seed outcomes, training specifications, and sensitivity conventions.

ETHICS STATEMENT

This work trains small language models on WikiText-2 and analyzes released research checkpoints. It introduces no new data collection, no human subjects, and no deployed system. We use the floor to diagnose output-head constraints and study changes to initialization and decay in a small controlled setting. The work inherits the concerns attached to the resources it builds on and to language-model research generally, including corpus provenance and licensing, representation in the training text, and the computational cost of training and evaluation.

REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv:1607.06450*, 2016.

Ronen Basri and David W. Jacobs. The softmax bottleneck does not limit the probabilities of the most likely tokens. In *ICLR*, 2026.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Shaked Brody, Uri Alon, and Eran Yahav. On the expressivity role of LayerNorm in transformers' attention. In *Findings of ACL*, 2023.

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. Isotropy in the contextual embedding space: Clusters and manifolds. In *ICLR*, 2021.

Nicholas Carlini, Daniel Paleka, Krishnamurthy Dj Dvijotham, Thomas Steinke, Jonathan Hayase, A. Feder Cooper, Katherine Lee, Matthew Jagielski, Milad Nasr, Arthur Conmy, Eric Wallace, David Rolnick, and Florian Tramèr. Stealing part of a production language model. In *ICML*, 2024.

Haw-Shiuan Chang and Andrew McCallum. Softmax bottleneck makes language models unable to represent multi-mode word distributions. In *ACL*, 2022.

David Demeter, Gregory Kimmel, and Doug Downey. Stolen probability: A structural weakness of neural language models. In *ACL*, 2020.

Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *EMNLP*, 2019.

Matthew Finlayson, John Hewitt, Alexander Koller, Swabha Swayamdipta, and Ashish Sabharwal. Closing the curious case of neural text degeneration. In *ICLR*, 2024a.

Matthew Finlayson, Xiang Ren, and Swabha Swayamdipta. Logits of API-protected LLMs leak proprietary information. In *COLM*, 2024b.

Octavian-Eugen Ganea, Sylvain Gelly, Gary Bécigneul, and Aliaksei Severyn. Breaking the softmax bottleneck via learnable monotonic pointwise non-linearities. In *ICML*, 2019.

Nathan Godey and Yoav Artzi. Lost in backpropagation: The LM head is a gradient bottleneck. In *COLM*, 2026. arXiv:2603.10145.

Nathan Godey, Éric de la Clergerie, and Benoît Sagot. Anisotropy is inherent to self-attention in transformers. In *EACL*, 2024a.

Nathan Godey, Éric de la Clergerie, and Benoît Sagot. Why do small language models underperform? Studying language model saturation via the softmax bottleneck. In *COLM*, 2024b.

Andreas Grivas, Nikolay Bogoychev, and Adam Lopez. Low-rank softmax can have unargmaxable classes in theory but rarely in practice. In *ACL*, 2022.

Andreas Grivas, Antonio Vergari, and Adam Lopez. Taming the sigmoid bottleneck: Provably argmaxable sparse multi-label classification. In *AAAI*, 2024.

Hakan Inan, Khashayar Khosravi, and Richard Socher. Tying word vectors and word classifiers: A loss framework for language modeling. In *ICLR*, 2017.

Siddhartha Rao Kamalakara, Acyr Locatelli, Bharat Venkitesh, Jimmy Ba, Yarin Gal, and Aidan N. Gomez. Exploring low rank training of deep neural networks. *arXiv preprint arXiv:2209.13569*, 2022.

Sekitoshi Kanai, Yasuhiro Fujiwara, Yuki Yamanaka, and Shuichi Adachi. Sigsoftmax: Reanalysis of the softmax bottleneck. In *NeurIPS*, 2018.

Mikhail Khodak, Neil Tenenholtz, Lester Mackey, and Nicolò Fusi. Initialization and regularization of factorized neural layers. In *ICLR*, 2021.

Anand Murugan. Does the LM head create a harmful gradient bottleneck? A causal test. *arXiv preprint arXiv:2608.16671*, 2026.

Ofir Press and Lior Wolf. Using the output embedding to improve language models. In *EACL*, 2017.

R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *COLM*, 2024.

William Timkey and Marten van Schijndel. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *EMNLP*, 2021.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer architecture. In *ICML*, 2020.

Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W. Cohen. Breaking the softmax bottleneck: A high-rank RNN language model. In *ICLR*, 2018.

Zhilin Yang, Thang Luong, Ruslan Salakhutdinov, and Quoc V. Le. Mixtape: Breaking the softmax bottleneck efficiently. In *NeurIPS*, 2019.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *NeurIPS*, 2019.

A THE DUAL, AND STRICT POSITIVITY AT FINITE RADIUS

Canonicality. F is itself the minimum of the mean target NLL over independently chosen feasible states, $F = \min_{h_1,\dots,h_n \in B_R} \frac{1}{n} \sum_i -\log p_{y_i}(h_i)$ with $p(h) = \mathrm{softmax}(Wh + b)$, because the objective separates and each per-position minimum is $T_R(e_{y_i})$ by equation 4. Consequently any target-wise constant that holds uniformly over feasible states lies below $T_R(e_v)$ tokenwise, and its induced average lies below F. Any tighter lower bound must use information beyond $h_i \in B_R$.

Throughout, $W \in \mathbb{R}^{V \times d}$ and $b \in \mathbb{R}^V$ are finite, $q \in \Delta_V$, and $R \geq 0$ is finite, with $H(p) = -\sum_i p_i \log p_i$ and the convention $0 \log 0 = 0$, which makes H continuous on the closed simplex. The feasible set $B_R = \{h : \|h\| \leq R\}$ and the simplex Δ_V are compact and convex.

Normalized states range over the open ball, whereas the minimum in equation 3 is taken over its closure B_R. The objective is continuous and, for $R > 0$, the open ball is dense in B_R, so $T_R(q)$ equals the infimum over normalization-feasible states.

Derivation of equation 9. Writing $z = Wh + b$ and using $D_{\mathrm{KL}}(q \,\|\, \mathrm{softmax}(z)) = \mathrm{LSE}(z) - q^\top z - H(q)$,

$$T_R(q) \;=\; \min_{h \in B_R} \big[\mathrm{LSE}(Wh + b) - q^\top (Wh + b) - H(q)\big].$$

The log-sum-exp is the convex conjugate of the negative entropy on the simplex,

$$\mathrm{LSE}(z) \;=\; \max_{\mu \in \Delta_V} \big\{\mu^\top z + H(\mu)\big\}, \tag{10}$$

the maximum being attained at $\mu = \mathrm{softmax}(z)$. Substituting,

$$T_R(q) \;=\; \min_{h \in B_R} \max_{\mu \in \Delta_V} \big[(\mu - q)^\top (Wh + b) + H(\mu) - H(q)\big].$$

The bracketed function is affine, hence convex and continuous, in h for each fixed μ, and concave and continuous in μ for each fixed h, since H is concave on Δ_V and the remaining terms are affine. With B_R and Δ_V compact and convex, Sion's minimax theorem (Sion, 1958) permits the exchange, and continuity on compact sets means both extrema are attained. Exchanging and evaluating the inner minimum with

$$\min_{h \in B_R} h^\top W^\top (\mu - q) \;=\; -R \left\| W^\top (\mu - q) \right\|_2$$

gives $T_R(q) = \max_{\mu \in \Delta_V} D_R(\mu)$ with D_R as in equation 9. Strong duality therefore holds, and weak duality is the specialization to a single μ and a single h. We compute the lower bound from a simplex witness and the upper bound from a feasible state.

Strict positivity at finite radius. Fix v and $h \in B_R$. Since W and b are finite, every logit is finite, and for $V \geq 2$,

$$\mathrm{softmax}(Wh + b)_v \;=\; \frac{e^{z_v}}{\sum_j e^{z_j}} \;<\; 1,$$

because $\sum_{j \neq v} e^{z_j} > 0$. Hence $-\log \mathrm{softmax}(Wh + b)_v > 0$ pointwise on B_R. The map $h \mapsto -\log \mathrm{softmax}(Wh + b)_v$ is continuous and B_R is compact, so the minimum is attained at some $h^\star \in B_R$ and

$$T_R(e_v) \;=\; -\log \mathrm{softmax}(Wh^\star + b)_v \;>\; 0.$$

The floors are therefore strictly positive at any finite budget. For a *fixed* finite head (W, b), $T_R(e_v)$ can approach zero only as $R \to \infty$. With b fixed, scaling W by $c > 1$ at radius R is equivalent to increasing the radius to cR.

B PROOFS FOR SECTION 3

Proof of Proposition 2. Finiteness holds because each minimum is attained on the compact ball, and monotonicity because $B_{r_1} \subseteq B_{r_2}$ for $r_1 \leq r_2$. For continuity, the objective is globally Lipschitz in h: its gradient is $W^\top (p - q)$ with $\|p - q\|_1 \leq 2$, so $L = 2\,\sigma_{\max}(W)$ is a Lipschitz constant. For $r < s$, an optimizer h_s at radius s shrinks radially to $(r/s)\, h_s \in B_r$ at distance at most $s - r$, so $\phi(r) \leq \phi(s) + L\,(s - r)$, while $\phi(s) \leq \phi(r)$ by monotonicity. Hence $|\phi(r) - \phi(s)| \leq L\,|r - s|$,

including at $r = 0$. For convexity, take radii r_1, r_2 with minimizers h_1, h_2 and $\theta \in [0, 1]$. The point $\theta h_1 + (1-\theta)h_2$ is feasible at radius $\theta r_1 + (1-\theta)r_2$, and the objective $h \mapsto \mathrm{LSE}(Wh + b) - q^\top(Wh + b) - H(q)$ is convex in h, so

$$\phi\big(\theta r_1 + (1-\theta)r_2\big) \;\le\; \theta\,\phi(r_1) + (1-\theta)\,\phi(r_2).$$

For $r > 0$ the point $h = 0$ is strictly feasible, so strong Lagrange duality holds for the constraint $\|h\| \le r$, and the standard perturbation argument gives $-\lambda_r \in \partial\phi(r)$: if $h_r^\star$ is optimal at radius r with multiplier λ_r, then for any r' the Lagrangian bound yields $\phi(r') \ge \phi(r) - \lambda_r(r' - r)$. A convex function is differentiable almost everywhere, and where it is, $\phi'(r) = -\lambda_r$. At a boundary solution the stationarity condition $W^\top(p_r - q) + \lambda_r h_r^\star/\|h_r^\star\| = 0$ gives $\lambda_r = \|W^\top(p_r - q)\|$. Convexity makes the subgradient $-\lambda_r$ nondecreasing in r, so λ_r is nonincreasing. □

Proof of Corollary 3. Let u_v attain γ_v and take the feasible state $h = Ru_v$. For every $j \ne v$, $(w_j - w_v)^\top u_v \le -\gamma_v$, so the logit gaps under cW satisfy $z_j - z_v = cR\,(w_j - w_v)^\top u_v + b_j - b_v \le -cR\gamma_v + b_j - b_v$. Hence

$$\begin{aligned} T_R(e_v;\, cW, b) \;\le\; -\log \mathrm{softmax}(cWh + b)_v \;&=\; \log\Big[1 + \textstyle\sum_{j\ne v} e^{\,z_j - z_v}\Big] \\ &\le\; \log\Big[1 + e^{-cR\gamma_v} \textstyle\sum_{j\ne v} e^{\,b_j - b_v}\Big]. \end{aligned}$$ □

Proof of Proposition 4. P is a symmetric idempotent with $\mathrm{tr}(P) = V - 1$, and $Q_\perp^\top Q_\perp = I_{d-1}$. For the raw direct head, $\mathbb{E}[U^\top PU] = \sigma^2\,\mathrm{tr}(P)\,I_d$, so

$$\mathbb{E}\,\|PUQ_\perp\|_F^2 \;=\; \mathrm{tr}\big(Q_\perp^\top\,\mathbb{E}[U^\top PU]\,Q_\perp\big) \;=\; (V-1)(d-1)\,\sigma^2.$$

For the product, condition on B: $\mathbb{E}_A[A^\top PA] = \sigma^2(V-1)I_k$, so $\mathbb{E}\,\|PABQ_\perp\|_F^2 = \sigma^2(V-1)\,\mathbb{E}\,\|BQ_\perp\|_F^2 = \sigma^2(V-1)\,\sigma^2 k(d-1)$. The RMS ratio is $\sqrt{k\sigma^4/\sigma^2} = \sigma\sqrt{k}$. □

Proof of the hull floor equation 8. Let $w_v = \sum_{j\ne v}\mu_j w_j$ with μ on the simplex. For any feasible h and any $c > 0$, the logits under cW satisfy $\max_{j\ne v} z_j \ge \sum_{j\ne v}\mu_j z_j = c\,w_v^\top h + \sum_{j\ne v}\mu_j b_j$, so $z_v - \max_{j\ne v} z_j \le b_v - \sum_{j\ne v}\mu_j b_j$, a constant free of h, c and R. Hence $\mathrm{softmax}(cWh + b)_v \le \big(1 + e^{\sum_j \mu_j b_j - b_v}\big)^{-1}$ and the stated lower bound on $T_R(e_v; cW, b)$ follows. For the equivalence: if w_v lies outside the convex hull of the other rows, a strictly separating direction gives $\gamma_v > 0$. If it lies inside, the displayed identity forces $\min_{j\ne v}(w_v - w_j)^\top u \le 0$ for every u, so $\gamma_v = 0$. □

Proposition 5. With gradients suppressed, the decoupled step multiplies every parameter tensor by a_t, so a two-factor product carries a_t^2 per step, and folding the decayed normalization gain contributes one further common factor to each arm. □

C Numerical bounding details

Error model. The affine product Wh is not exact in fp64, so we enclose it with an underflow-aware bound: for a length-n dot product, $|\mathrm{fl}(x^\top y) - x^\top y| \le \gamma_n \sum_i |x_i||y_i| + 2n\eta$ with $\gamma_n = nu/(1-nu)$, $u = 2^{-53}$ and $\eta = 2^{-1074}$ the smallest subnormal. The magnitude term is itself computed in floating point, so it is inflated mechanically as $M \le (\hat{M} + 2n\eta)/(1-\gamma_n)$ with the denominator rounded down. The γ_n bound covers arithmetic reductions only. Entropy terms involving $\log$ are evaluated in interval arithmetic.

Directed rounding. Casting an interval endpoint with the default round-to-nearest can pull it *inside* the true enclosure and silently invalidate the bound. Every endpoint is therefore moved outward by one unit in the last place.

Environment guards. The error model assumes round-to-nearest, gradual underflow, and a single deterministic reduction order. We query the rounding mode directly because the common test $1 + 2^{-53} = 1$ also passes under round-down and round-toward-zero. We confirm that subnormals survive the scalar, dot and matrix–vector paths, and assert single-threaded BLAS, since a multi-threaded reduction changes summation order.

Scalable evaluation. Evaluating equation 9 entirely in interval arithmetic costs $O(Vd)$ interval operations per token. Writing

$$W^\top(\mu - q) \;=\; \frac{W^\top \hat{\mu}}{S_\mu} \;-\; \frac{W^\top \hat{q}}{S_q}$$

reduces this to four rigorously enclosed fp64 matrix–vector products plus d interval operations, enabling evaluation at $V = 49{,}152$.

Boundary candidates. An outward norm bound on a candidate at $\|h\| = R$ necessarily exceeds R. When a candidate lies on the boundary, we shrink it strictly inside. Feasibility preserves primal validity, and the resulting candidate is reevaluated for both endpoints.

Parallelism. The evaluation distributes independent contexts across worker processes by copy-on-write forking and writes results by index.

Exact normalization and mixed precision. Stored binary64 values are treated as exact dyadic rationals. The target and dual witness are normalized exactly in rational arithmetic. The witness is constructed from the candidate logits and may differ from the exact softmax at h. Its simplex membership is sufficient for weak duality. The transcendental terms (interval $\log$, $\exp$ and division, and the log-sum-exp built from them) are evaluated with `mpmath` 1.3.0's interval context (`mpmath.iv`) at 60-digit working precision. The documentation labels `mpmath.iv` experimental. The bounding is therefore *mixed precision*: outward-rounded fp64 with explicit error bounds for the arithmetic reductions, and pinned interval arithmetic for the transcendentals. Its validity is conditional on the documented enclosure semantics of those `mpmath.iv` operations at that pinned version. Matrix products carry the explicit underflow-aware bound above. The implementation rejects any run with overflow, non-finite values, an infeasible state, an incorrect rounding mode, multithreaded BLAS, or a failed bracket.

Independent cross-check. We test that dependence on `mpmath.iv` with a second implementation that re-evaluates released witnesses using Arb ball arithmetic through `python-flint` at 300-bit precision, sharing no code with the primary implementation. For each sampled target, it verifies feasibility exactly by checking $\sum_i h_i^2 \le R^2$ (with $R^2 = d$ at the architectural radius) with zero-radius arithmetic, then recomputes the primal at the released state and the dual at the softmax of its logits. The dual lower-bounds T_R and the primal upper-bounds it, so Arb sandwiches the true value.

The check requires the production interval to contain the independently computed Arb bracket. This is a sufficient validation condition: a valid production lower bound can exceed the Arb lower bound when the two implementations use different dual witnesses. Across 300 one-hot targets drawn from all three arms, every witness is exactly feasible and every production interval contains the sandwich. Two negative controls test endpoint containment and witness feasibility: raising one lower endpoint by 10^{-9} yields exactly one containment failure, and scaling one witness just outside the architectural ball yields exactly one feasibility failure. A parallel dense-target check covers 550 frozen contexts: the four controlled arms and the replication's C' at the architectural radius and at $8R$, and both 2B cells at $R = 64$ and $R = 512$. Wherever stored per-context endpoints exist, they are reproduced exactly through the production path and then tested directly, exercising the entropy sums, $W^\top$ applied to dense differences, and the exact rational normalization of the teacher vectors. All contexts pass with zero containment or feasibility failures, and each negative control produces exactly one failure. The sampled ids, manifests, and generators are released with the code.

Where the enclosure begins. The evaluator treats the materialized binary64 folded and reduced W and b as exact dyadic inputs, and the enclosure covers the subsequent optimization and evaluation. It does not enclose the rounding introduced while folding the raw parameters or constructing the reduction basis, and the independent check validates the materialized-map computation on the same inputs.

Split radius convention. The architectural feasible set is the ball of radius $\sqrt{d}$, and $\mathrm{fl}(\sqrt{d})$ can fall strictly below it, as it does for $d = 96$. The two endpoints therefore use different radii. The dual

is evaluated at $R_{\mathrm{up}} \geq \sqrt{d}$, which is sound because D_R is decreasing in R, so $D_{R_{\mathrm{up}}} \leq T_{R_{\mathrm{up}}} \leq T_{\sqrt{d}}$ and the lower endpoint transfers. The primal witness, by contrast, bounds $T_{\sqrt{d}}$ from above only if it is feasible for the exact ball, so each candidate is shrunk until $\sum_i h_i^2 \leq R^2$ holds exactly, decided in integer arithmetic: every binary64 value is a dyadic rational, so the squares share a common power-of-two denominator and the comparison reduces to one between integers. Using R_{up} for both endpoints would enclose $T_{R_{\mathrm{up}}}$. Exact-ball feasibility of the primal witness keeps the upper bound valid for $T_{\sqrt{d}}$.

Square roots. The norm routines require an upper bound on $\sqrt{x}$. IEEE-754 mandates a correctly rounded square root. The returned r is checked in exact rational arithmetic and nudged outward until $r^2 \geq x$ holds, establishing the upper bound independently of platform square-root behavior.

Candidate generation. Candidates come from fp64 projected gradient descent with backtracking line search on equation 3, started at $h = 0$ (or at a supplied warm start, projected into B_R), run for at most 4000 iterations and stopped when the primal–dual gap falls below 10^{-10}. Candidate quality affects only how tight the bracket is, never its validity: any feasible h and any simplex μ bracket T_R. Evaluating one arm and seed of §5 takes about eight seconds for the flat-stream forward pass and about four minutes to bound all $7{,}366$ observed targets across twelve worker processes.

Across-seed vertex enumeration. Each seed contributes an interval rather than a point, so the vector of seed-level effects lies in a box. With n seeds, sample mean $\bar{x}$, sample standard deviation s, and $t = t_{0.975,n-1}$, the usual 95% Student-t interval has endpoints $\bar{x} \pm ts/\sqrt{n}$. The upper endpoint is a convex function of that vector, since the mean is linear and the sample standard deviation is a norm of the centered vector, so its maximum over a box is attained at a vertex. The lower endpoint is concave and its minimum is likewise attained at a vertex. We therefore enumerate all 2^n vertices, $2^5 = 32$ for the five-seed experiment and $2^{20} = 1{,}048{,}576$ for the twenty-seed replication, and report the smallest interval enclosing all the resulting Student-t intervals. The box endpoints are interval differences computed with outward-directed subtraction and audited against exact rational values, and the same construction covers the one-hot contrasts ΔF and the teacher-distribution contrasts ΔN. Taking only the all-lower and all-upper vertices would not be conservative, because mixed vertices can have larger sample variance.

D Additional analyses and implementation details

Backbone matching. The primary analysis requires identical data order and identical initial backbone parameters across arms within a seed. In the reference implementation the factorized arm constructs a projection and a smaller output matrix before the model-wide reinitialization, advancing the random generator differently from the full-rank head, so its initial backbone differs. The primary analysis implements the pairing by copying arm A's initial backbone into every arm. Arms A and B consume the same random draws before reinitialization and already match, so the copy changes only arm C.

Table 2 compares the matched convention with the reference implementation's unmatched one across all five seeds. Both give positive 95% intervals for ΔF and ΔL.

Table 2: Primary $C - B$ contrast under both initialization conventions. ΔF brackets are bounds-propagated 95% intervals, and ΔL brackets are paired Student-t 95% intervals. All values are in nats.

variant	ΔF	ΔL
matched (primary)	$[1.462,\ 1.777]$	$[0.045,\ 0.156]$
unmatched (reference implementation)	$[1.567,\ 1.762]$	$[0.077,\ 0.155]$

Budget relaxation. At twice and eight times the architectural radius, the backward-only floor falls to about 10^{-5} nats at $8R$ while the factorized arm retains 1.227, with paired contrast $[1.155,\ 1.300]$. The contrast is nonmonotonic because the two convex responses fall at different rates. An eightfold relaxation changes each 2B cell's mean floor by less than 0.01 nats. These ladders relax both arms

Table 3: Floors along the radius ladder (8,192 contexts, all valid at every radius). The four radius columns report upper endpoints of the mean floor. $\bar{D}_8$ is a two-sided interval.

	mean floor, upper endpoint				
cell	R=64	R=128	R=256	R=512	$\bar{D}_8 = \mathbb{E}[T_{64} - T_{512}]$
rank 32	0.41522	0.40731	0.40705	0.40701	[0.008195, 0.008208]
rank 512	0.030065	0.030063	0.030063	0.030063	$[6.8\times10^{-7}, 2.7\times10^{-6}]$

Table 4: Controlled-model radius analysis over five seeds. Arm rows report mean floor point estimates in nats. The ΔF row reports bounds-propagated 95% intervals. The architectural radius is $R = \sqrt{96}$.

	R	$2R$	$8R$
arm B, full-rank forward	2.252	0.448	1.0×10^{-5}
arm C, factorized forward	3.871	2.911	1.227
ΔF $(C - B)$	[1.462, 1.777]	[2.355, 2.572]	[1.155, 1.300]

together and bound $T_R - T_{8R}$, and relaxing arm C alone is the amplification analysis above. Without T_∞, the floor at $8R$ remains the unbounded reachable-family restriction plus a remaining finite-budget effect.

Learned mean scales. At the final checkpoint the mean centered folded scales are: spectral 34.3 (C) against 101.2 (B), Frobenius 46.6 against 115.8, centered bias norm $\|Pb\|$ 5.3 against 33.4.

Initialization and decay measurements. The same-variance initialization predicts an initial RMS-Frobenius ratio of 0.139 for the factorized effective map, and deterministic reconstructions of the five initial models give 0.136 to 0.141. Under the decay operator alone the direct head contracts by $\alpha = 0.740$ over the schedule and the factor product by $\alpha^2 = 0.548$, the folded maps receiving α^2 and α^3.

Per-seed validation losses. Table 5 reports the stored per-seed validation NLL of the three original arms.

Table 5: Per-seed validation NLL (nats), original arms.

seed	L_A	L_B	L_C
S101	5.969	5.998	6.138
S102	5.953	6.068	6.119
S103	5.982	5.986	6.137
S104	5.971	6.061	6.125
S105	5.992	6.010	6.107

Training configuration. All arms train for 1,200 steps with AdamW (learning rate 0.005, batch size 8, weight decay 0.1, betas 0.9 and 0.95, epsilon 10^{-8}, cosine decay after 120 linear-warmup steps, gradient clipping at 1.0), evaluated at the final checkpoint on the full prepared validation stream. The corpus is the WikiText-2 raw split at pinned revision `b08601e`, tokenized at BPE-8,192. The 2B comparison evaluates the released `d32_170k` and `d512_170k` checkpoints on the frozen 8,192-context list.

The C' training specification. Within each seed, C' starts from arm C's factor draws before any rescaling. Both factors are rescaled symmetrically to standard deviation $\tau = (\sigma^2/k)^{1/4} \approx 0.0537$, which equates the expected squared centered folded Frobenius scale of the product with the direct head's at initialization. In the optimizer, the two factor tensors form a parameter group with weight decay disabled, and at every step both receive the multiplier $\sqrt{1 - \eta_t\lambda}$, with η_t the

scheduled learning rate and $\lambda = 0.1$ the configured decay, so their product receives the same $1-\eta_t\lambda$ multiplier as a directly parameterized head. The correction matches the isolated decay operators. Adaptive gradient updates can produce different training trajectories. The output head carries no bias parameter, and the effective bias is the folded final-normalization shift. The final normalization gain and shift remain in the ordinary parameter group and receive its usual decay. Initial backbones and data order are paired within each seed. C' jointly modifies initialization and product decay, so the comparison estimates their combined effect.

C' results. At five seeds the aggregate contrasts are $\Delta F \in [0.745,\ 1.094]$, $\Delta L \in [-0.050,\ 0.066]$, and $\Delta G \in [-1.029,\ -0.793]$ against B, and $\Delta F \in [-0.762,\ -0.639]$, $\Delta L \in [-0.117,\ -0.068]$, and $\Delta G \in [0.562,\ 0.654]$ against C, with retained contrast $(F_{C'}-F_B)/(F_C-F_B) \in [0.523,\ 0.642]$. Table 6 reports the C' outcomes per seed: floor and validation loss, the paired contrasts, the retained-fraction box, and the headroom summaries. The aggregate contrasts above are bounds-propagated Student-t 95% intervals across seeds, the retained contrast $[0.523,\ 0.642]$ is the descriptive span of the five per-seed enclosure boxes, and in Table 6 the floor and floor-contrast cells are numerical enclosures, printed as a single value only when both endpoints round identically at three decimals, while $L_{C'}$ is the stored fp64 measurement and the ΔG and headroom columns combine the enclosures with those stored losses. Across seeds the headroom medians are 2.57–2.71 nats and the fifth percentiles 0.91–0.97. Final centered folded scales: spectral 34.2–41.3 (C') against 30.3–38.6 (C) and 98.8–104.4 (B), Frobenius 50.6–56.5 against 43.2–50.3 and 113.2–118.8, centered bias norm $\|Pb\|$ 2.7–5.9 against 4.0–7.4 and 29.8–39.5.

Table 6: Per-seed C' outcomes: floor, validation loss, paired contrasts, retained fraction, and headroom summaries. Loss quantities and headroom are in nats, and the retained fraction is dimensionless.

seed	$F_{C'}$	$L_{C'}$	ΔF vs B	ΔF vs C	ΔG vs C	retained	med s	s at 5%
S101	3.294	6.066	1.119	-0.624	0.552	0.642	2.572	0.910
S102	3.163	6.027	0.783	-0.682	0.590	0.535	2.669	0.952
S103	3.110	6.016	0.985	-0.752	0.631	0.567	2.706	0.931
S104	3.136	6.025	0.793	-0.722	0.622	0.523	2.689	0.954
S105	3.152	6.030	0.918	-0.721	0.643	0.560	2.674	0.967

Descriptive head geometry. Final centered folded stable ranks $\|PW\|_F^2/\sigma_1(PW)^2$: B 1.3–1.3, C 1.7–2.0, C' 1.9–2.2. Figure 2 shows the normalized singular-value curves averaged over seeds. Every head is dominated by one direction, and C' carries slightly more relative weight beyond the leading direction than C.

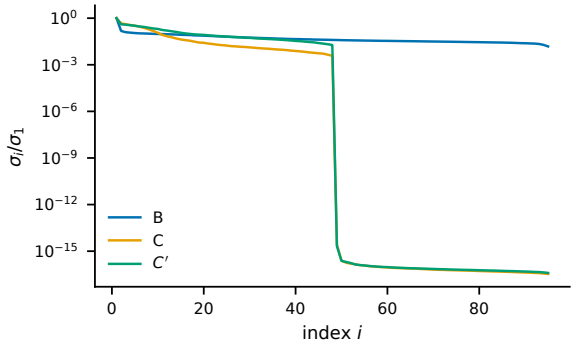


Figure 2: Normalized singular-value curves σ_i/σ_1 of the final centered folded maps, averaged over seeds (log scale).

Per-seed crossover brackets. Secant and chord constructions in exact rational arithmetic bracket each seed's floor crossover: $[3.70, 3.97]$, $[2.93, 3.37]$, $[3.62, 4.18]$, $[3.10, 3.54]$ and $[3.33, 3.63]$ for seeds S101 through S105, displayed endpoints rounded outward.

Table 7: Teacher-distribution floors (nats) and unremovable shares per seed. Teacher: arm A of the same seed.

seed	N_B	N_C	$N_{C'}$	$N_{C^{\mathrm{spec}}}$	$N_{C^{\mathrm{frob}}}$	Q_B	Q_C	$Q_{C'}$
S101	0.035	0.117	0.075	[0.098, 0.101]	0.102	[0.251, 0.252]	0.519	0.429
S102	0.053	0.125	0.065	[0.109, 0.111]	[0.111, 0.112]	0.322	0.526	0.396
S103	0.036	0.122	0.064	[0.104, 0.106]	[0.106, 0.108]	0.261	0.529	0.375
S104	0.041	0.119	0.068	0.106	[0.107, 0.108]	0.306	0.521	0.400
S105	0.036	0.115	0.066	[0.096, 0.098]	[0.098, 0.100]	[0.243, 0.244]	0.505	[0.357, 0.358]

Teacher-distribution targets in the controlled model. With arm A of the same seed as teacher, 4,096 evenly spaced full-window contexts, and identical contexts and ordering for every arm, Table 7 reports the per-seed N enclosures and unremovable shares $Q = N/K$. All 4,096 contexts are valid in every cell, and N for arm A evaluated against itself is zero to displayed precision. Q is shown for the training arms, with the C' share spanning 0.36–0.43. The remaining contrasts are $N_{C'} - N_B \in [0.014,\ 0.041]$ and $N_B - N_A \in [0.030,\ 0.050]$, both resolved. For the two rescaled-C cells the realized divergence from the teacher rises sharply, echoing the inference-time result, and Q falls below 0.04.

Secondary replication outcomes. Every seedwise $L_{C'} - L_C$ difference is negative, and the two-sided 90% interval for $L_{C'} - L_B$ is $[-0.001,\ 0.027]$. Under exact 2^{20}-vertex bounds propagation, $F_B - F_A \in [0.146,\ 0.254]$: the twenty seeds resolve the $B - A$ floor difference that five left unresolved, and the arm and unremovable-share orderings are unchanged. Across the twenty seeds the arm-level enclosure ranges are $N \in [0.028,\ 0.067]$, $K \in [0.126,\ 0.186]$, and $Q \in [0.203,\ 0.358]$ for B, $N \in [0.094,\ 0.152]$, $K \in [0.210,\ 0.260]$, and $Q \in [0.410,\ 0.588]$ for C, and $N \in [0.056,\ 0.098]$, $K \in [0.157,\ 0.203]$, and $Q \in [0.315,\ 0.483]$ for C'.

The 2B teacher construction. Each teacher distribution q is the stored fp64 softmax of the full-rank ($d = 4096$) teacher at the frozen context. The bounding layer treats the stored vector as exact dyadic rationals and normalizes it exactly, $q_j = \hat{q}_j / S$ with S the exact rational sum, validated to lie within 10^{-6} of one. N is the mean of the per-context two-sided T_R enclosures, K the mean enclosure of the realized-state divergence $D_{\mathrm{KL}}(q \,\|\, p_X)$ with no projection of the realized state, and $Q = N/K$ under the eligibility rule that the lower endpoint of K is at least 10^{-3} nats. The two-cell intervals at R: $N_{32} \in [0.4151,\ 0.4153]$, $N_{512} \in [0.0300,\ 0.0301]$, $K_{32} \in [0.6850,\ 0.6851]$, $K_{512} \in [0.2377,\ 0.2378]$. The scale match behind the main-text derivation is spectral only, the centered Frobenius scales remain substantially different (ratio 0.433), and both heads are bias-free under RMSNorm. The drop bound subtracts the stored endpoints $\overline{N}_{512}(R) = 0.030064542264737645$ and $\underline{N}_{512}(2R) = 0.030062480587069836$ with outward rounding.

The backward-only arm. Relative to arm A, arm B has a floor difference of $+0.163$ nats, with an interval of $[-0.010,\ 0.336]$ that contains zero, so the difference is unresolved at five seeds. The forward comparison from the same baseline, $C - A$, gives $[1.733,\ 1.833]$. Both restrictions carry the same nominal rank. The twenty-seed replication resolves the $B - A$ difference.

Table 8: Secondary contrasts over five seeds. ΔF brackets are bounds-propagated 95% intervals, ΔL brackets are Student-t 95% intervals, and the ΔG entries are point estimates.

contrast	ΔF	ΔL	ΔG
$C - A$ factorized	$+1.783$ [1.733, 1.833]	$+0.152$ [0.125, 0.179]	-1.631
$B - A$ backward-only	$+0.163$ [-0.010, 0.336]	$+0.051$ [-0.009, 0.112]	-0.112

E Extended related work

The softmax bottleneck as a rank constraint. Yang et al. (2018) identified the bottleneck as a rank constraint on attainable log-probability matrices. Kanai et al. (2018) and Yang et al. (2019) raise that

rank, and Chang & McCallum (2022) exhibit multi-mode distributions as the casualty. Godey et al. (2024b) connect it to small-model saturation and Godey & Artzi (2026) to gradient compression through the rank-d head. The same low-rank image lets Finlayson et al. (2024b) and Carlini et al. (2024) estimate the dimension of the softmax-identifiable image from API-exposed output scores, recovering d when the centered head and the queried states both span d dimensions, and, for a cross-entropy-optimal model, it induces the moment-matching fiber $W^\top p = W^\top \hat{p}$ that Finlayson et al. (2024a) exploit.

Geometry of hidden states. Trained representations are typically anisotropic, concentrated in angle (Ethayarajh, 2019; Godey et al., 2024a), with rogue dimensions (Timkey & van Schijndel, 2021), massive activations (Sun et al., 2024), and local cluster structure (Cai et al., 2021). Our budget is a norm bound on the normalized coordinate and assumes no isotropy. Restricting the feasible set to a tighter measured state set can only increase the floors.

Grivas et al. (2024) carry argmaxability to sigmoid outputs.

We measure scale on the centered folded map and analyze both learned heads and bias-fixed rescalings of frozen checkpoints. The scale–radius identity connects these rescalings to the finite-radius floor.

Output-layer conventions. We state the budget at the normalized coordinate of the final normalization layer with the affine parameters folded into W and b (Ba et al., 2016; Zhang & Sennrich, 2019), and the formulation applies unchanged to tied heads (Press & Wolf, 2017; Inan et al., 2017). Brody et al. (2023) identified the same two LayerNorm facts for attention.

Convex analysis. The analysis uses convex duality (Rockafellar, 1970; Boyd & Vandenberghe, 2004), with the minimax exchange of Appendix A due to Sion (1958).

F POSITIONWISE LOSS HEADROOM

For each validation position i, the headroom $s_i = \ell_i - T_R(e_{y_i})$ is computed from the stored fp64 NLL ℓ_i and the rigorous per-token floor bracket, with outward subtraction at both endpoints. Every summary here is a floor-bound-propagated numerical band conditional on the stored fp64 NLL. Figure 3 and Table 9 show every seed separately, and comparisons across runs rely on the five seeds. Table 10 is a descriptive secondary analysis: within each seed it forms terciles over positions from the positionwise mean of the two arms' floor-band midpoints, applies the identical positional masks to both arms, and only then pools the masked positions across seeds. In the tables, a cell showing a single value is a descriptive nearest-rounded number whose two underlying band endpoints round identically at three decimals, and a cell showing a band is rounded outward.

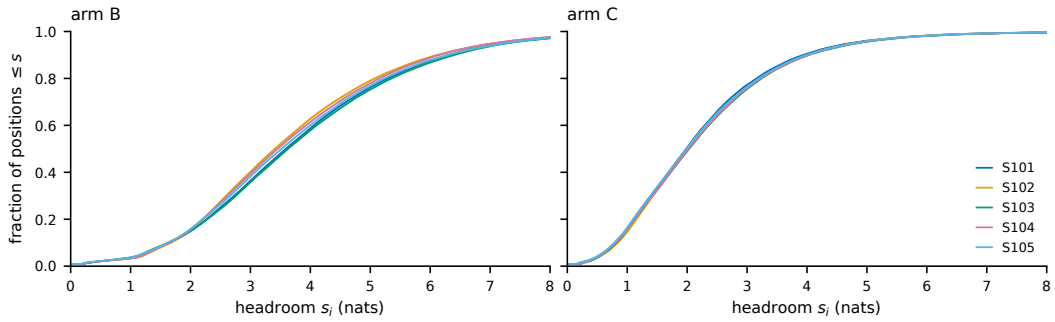


Figure 3: Empirical cumulative distribution of the positionwise headroom per seed, arms B and C. The vertical axis is the fraction of validation positions with headroom at most s. In each arm panel every seed contributes two curves, the band's lower and upper endpoints, visually coincident at this scale. The axis stops at 8 nats for readability, while the released arrays retain the full tails, which reach roughly 18 nats.

Table 9: Headroom quantiles per seed (nats).

arm	seed	1%	5%	10%	25%	50%
B	S101	0.199	1.199	1.645	2.498	3.591
B	S102	0.199	1.261	1.677	2.404	3.407
B	S103	0.196	1.180	1.629	2.531	3.626
B	S104	0.206	1.226	1.650	2.409	3.452
B	S105	0.189	1.171	1.628	2.436	3.516
C	S101	0.206	0.559	0.794	1.245	1.982
C	S102	0.170	0.599	0.840	1.286	2.029
C	S103	0.225	0.586	0.818	1.271	2.024
C	S104	0.217	0.572	0.806	1.266	2.020
C	S105	0.151	0.555	0.791	1.241	1.985

Table 10: Descriptive secondary analysis: headroom quantiles by shared floor tercile (nats), masked positions pooled across seeds after the within-seed construction.

arm	tercile	1%	5%	10%	25%	50%
B	low	0.003	0.337	0.975	1.578	2.481
B	mid	1.372	1.791	2.063	2.620	3.490
B	high	1.916	2.397	2.756	3.495	4.536
C	low	0.002	0.303	0.669	1.165	1.961
C	mid	0.374	0.627	0.816	1.275	2.040
C	high	0.450	0.708	0.902	1.349	2.021