

---

# The Geometric Wall: Manifold Structure Predicts Layerwise Sparse Autoencoder Scaling Laws

---

Eslam Zaher<sup>1,2</sup>, Maciej Trzaskowski<sup>1,3,4</sup>, Quan Nguyen<sup>1,3,5</sup>, Fred Roosta<sup>1,2</sup>

<sup>1</sup> ARC Training Centre for Information Resilience (CIRES)

<sup>2</sup> School of Mathematics and Physics, University of Queensland

<sup>3</sup> Institute for Molecular Bioscience, University of Queensland

<sup>4</sup> Profenso <sup>5</sup> QIMR Berghofer Medical Research Institute

## Abstract

Sparse autoencoders (SAEs) operationalise the linear representation hypothesis: they reconstruct model activations as sparse linear combinations of interpretable dictionary atoms, on the implicit assumption that activation space is well approximated by a globally linear structure. Their reconstruction error varies sharply across layers in ways that existing scaling laws, fitted at single layers, do not explain. We argue that this variation is the empirical trace of a geometric mismatch: where the activation manifold is curved and its intrinsic dimension varies across layers, no sparse linear dictionary can match it uniformly, and the SAE’s width-sparsity scaling becomes a layer-dependent function of manifold structure rather than a single universal law. We conduct the first cross-layer SAE scaling study, fitting and regressing on 844 residual-stream Gemma Scope SAE checkpoints across 68 layers of Gemma 2 2B and 9B. Stage 1 fits a per-layer scaling-law surface; Stage 2 regresses the fitted parameters and the derived per-layer width exponents on four layerwise geometric summaries. We find that manifold geometry predicts the per-layer width exponent in both models, and that the same regression coefficients learnt on one model predict the other model’s per-layer exponents under cross-model transfer, indicating a transferable geometric law. At the showcase layers where richer width grids permit identification of the asymptotic floor, we find that the fitted floor tracks the layerwise geometric ordering: higher curvature and intrinsic dimension correspond to higher floor, consistent with the irreducible second-order residual that any sparse linear approximation of a curved manifold must leave behind. SAEs thus encounter not a finite-resource ceiling but a geometry-dependent wall, set by the manifold they are trying to reconstruct.

## 1 Introduction

The interpretability programme that builds on sparse autoencoders depends on SAEs being faithful to the activations they reconstruct across all the layers we want to interpret. The existing scaling-law characterisation of this faithfulness has been fitted at a single layer of GPT-4 [1]; whether it extends to other layers, and what predicts any cross-layer variation, has not been measured. We measure it across every layer of two language models.

Current practice rests on the linear-representation hypothesis (LRH) [2, 3] as adopted by mechanistic interpretability [4–7]: SAEs operationalise LRH by encoding each model activation as a sparse combination of dictionary atoms hypothesised to be the meaningful linear directions. The implicit geometric premise is that activation space is globally well-served by such a sparse linear basis. Gao et al. [1] formalised the resulting scaling picture as a joint power law for the SAE reconstruction loss  $L$  in dictionary width  $n$  and sparsity  $k$ , fitted at a single layer of OpenAI’s GPT-4; whether the law’s parameters vary across layers was not investigated.

Parallel results expose tensions in the SAE programme that suggest the LRH-style picture of activations as globally served by a sparse linear basis may be incomplete. A growing literature characterises semantic features in language models as forming non-trivial geometric structures rather than simple linear directions, ranging from circular features for cyclic concepts to hierarchically-organised polytopes and curved feature manifolds discretised by sparse families [8–14]. SAE reconstruction is also uneven across depth: at the middle layers of the same Gemma 2 models we study, the nonlinear residual reconstruction error remains roughly constant across an order of magnitude of dictionary width [15], and substituting an SAE reconstruction for a model layer’s activation inflates the model’s downstream loss several-fold beyond what a same-magnitude random perturbation would, with the pathology varying by layer [16]. These observations have been documented in isolation; what is missing is an account that links them to a single underlying property of the activations.

On the geometric side, a research community studying activation manifolds has independently mapped their depthwise structure. Intrinsic dimension follows a *hunchback* profile that peaks at middle layers across architectures [17, 18]; intrinsic dimension and curvature both vary systematically with depth in language models [19]; and the natural inner product on language-model representations is non-Euclidean, an information geometry inherited from the softmax output rather than the ambient Euclidean structure [20, 19]. This activation-geometry literature has so far developed in isolation from work on SAEs, and whether the layerwise geometric variation it documents predicts the layerwise behaviour of the SAE scaling law has not been tested. That is the question this paper answers.

We contend these threads are linked by a single phenomenon: geometric mismatch. A sparse linear basis is not capacity-limited but *geometrically mismatched* to a manifold whose curvature and intrinsic dimension change with depth, so the per-layer scaling law inherits a layerwise structure from the manifold being approximated. Width-sparsity scaling does not cease to hold; it ceases to be *universal*. We predict that the mismatch leaves a fingerprint in the per-layer scaling law: both the asymptotic reconstruction error that infinite dictionary width cannot eliminate in the small- $k$  regime to which the SAE scaling law applies, and the rate at which reconstruction error approaches that asymptote as width grows, depend layerwise on four geometric summaries of the activation manifold (intrinsic dimension, multi-scale curvature, tangent variation, and within-layer heterogeneity). Our framework recasts the SAE limit as a geometry-dependent wall, not a finite-resource ceiling. The claim is not that SAEs are unsuitable for interpretability, but that their layerwise scaling behaviour is conditioned by geometric structure that current single-layer practice leaves unmodelled. We test this prediction on the publicly released Gemma Scope SAEs [7]: all 844 JumpReLU checkpoints spanning all 26 layers of Gemma 2 2B and all 42 layers of 9B; the analysis fits a per-layer scaling-law surface and regresses the fitted parameters on the layer’s geometric summaries.

**Contributions.** Our contributions are threefold.

**(i) Geometric framework for the SAE residual.** A pullback information-geometric account of why activation spaces need not admit globally efficient sparse linear codes, yielding a directional prediction tested below: layers with higher intrinsic dimension and curvature should scale less efficiently in dictionary width.

**(ii) Geometry-conditioned scaling law.** We extend Gao et al. [1]’s width-sparsity scaling law to a geometry-conditioned account in which the scaling-law parameters become layerwise functions of activation manifold geometry, fit per model and validated under a closed-form leave- $K$ -layer-out cross-validation protocol with layer-permutation nulls (full method in §3 and §4; validation in §G).

**(iii) Cross-model geometric law and floor coupling at scale.** Across 844 Gemma Scope SAEs: (a) the cross-layer regression coefficients learnt on one model predict the other model’s per-layer width-scaling exponent under cross-model transfer, indicating a transferable geometric law for the per-layer width exponent; (b) at the showcase layers where the asymptotic reconstruction floor is identifiable, that floor tracks two specific geometric channels (intrinsic dimension and multi-scale curvature), consistent with and extending Gao et al. [1]’s *spectrum of structure* picture of the irreducible loss.

## 2 Background

We fix notation for transformers, sparse autoencoders, and the pullback information geometry used throughout; experiments (§4) specialise to the residual stream.

## 2.1 Language models and hidden representations

A transformer with  $\mathcal{L}$  layers maps each hidden state  $h_\ell \in \mathbb{R}^{d_\ell}$  to a next-token distribution via the *predictive map*

$$F_\ell : \mathbb{R}^{d_\ell} \rightarrow \Delta_\circ^{V-1}, \quad F_\ell(h) = \text{softmax}(z_\ell(h)), \quad (1)$$

where  $z_\ell$  composes layers  $\ell$  through unembedding. For a data distribution  $\mathcal{D}$ , the activation set  $\mathcal{H}_\ell \subseteq \mathbb{R}^{d_\ell}$  is the object that SAEs approximate.

## 2.2 Sparse autoencoders as sparse linear approximation

A sparse autoencoder at layer  $\ell$  encodes  $h \in \mathbb{R}^{d_\ell}$  as a sparse code  $a = \phi(W^{\text{enc}}h + b^{\text{enc}}) \in \mathbb{R}^n$  with  $n \gg d_\ell$  and decodes  $\hat{h} = W^{\text{dec}}a + b^{\text{dec}}$ , the columns of  $W^{\text{dec}}$  being the dictionary atoms. Principal variants (ReLU [4, 5], TopK [1], and JumpReLU [21, 7]) differ in  $\phi$  but share the sparse-linear structure. Reconstruction quality is measured by the per-sample normalised mean-squared error  $\text{NMSE}(h_i, \hat{h}_i) = \|h_i - \hat{h}_i\|^2 / \|h_i\|^2$ , aggregated to a checkpoint-level loss  $L$  (Section 4).

## 2.3 The SAE scaling law

Gao et al. [1] showed that reconstruction loss follows a joint scaling law in dictionary width  $n$  and realised sparsity  $k = L_0$ , fitted at a layer 5/6 of the way into a GPT-4-series model on the post-layernorm residual stream activations. They first established a no-floor power-law surface  $L(n, k) = A(k)n^{-\alpha(k)}$ , then introduced the asymptotic floor  $B(k)$  as a refinement at the same layer, where richer width coverage exposed the saturation, yielding the with-floor form

$$L(n, k) = A(k)n^{-\alpha(k)} + B(k), \quad (2)$$

with  $A(k) = \exp(a_0 + \beta_k \log k)$ ,  $\alpha(k) = -(\beta_n + \gamma \log k)$ ,  $B(k) = \exp(\zeta + \eta \log k)$ . Here  $A(k)n^{-\alpha(k)}$  is the *reducible loss*, whose width exponent  $\alpha(k)$  depends on sparsity through the interaction parameter  $\gamma < 0$ : wider dictionaries help more when more latents are active. The term  $B(k)$  is the *irreducible floor*: variance no dictionary can capture at a given  $L_0$ . The law was established at a single layer; whether its layerwise behaviour varies systematically with cross-layer structure was not addressed, and is the central empirical question of this paper. The no-floor and with-floor forms are linked through the log-linearisation

$$\log L \approx a_0 + \beta_n \log n + \beta_k \log k + \gamma \log n \cdot \log k, \quad (3)$$

which holds when  $B$  is small relative to  $An^{-\alpha}$  at the observed widths.

## 2.4 Pullback information geometry of activation spaces

We develop the geometric framework connecting the model’s predictive structure to its hidden representations; full differential-geometric background is in Appendix A. The natural Riemannian metric on a space of probability distributions is the **Fisher information metric** [22], unique up to scaling among metrics invariant under sufficient statistics [23]. For categorical distributions  $\pi \in \Delta_\circ^{V-1}$ , the Fisher metric in logit coordinates takes the form of the categorical covariance

$$\Sigma_\pi = \text{diag}(\pi) - \pi\pi^\top, \quad (4)$$

of rank  $V-1$ ; the resulting manifold  $(\Delta_\circ^{V-1}, \mathbf{G}_{\text{FR}})$  has Bhattacharyya arc length as its geodesic distance [24, 22].

**Pullback to hidden layers.** The predictive map  $F_\ell$  (Eq. 1) pulls the Fisher–Rao metric back to  $\mathbb{R}^{d_\ell}$ : with  $J_\ell(h) = \partial z_\ell / \partial h$ ,

$$\mathbf{G}_\ell(h) = J_\ell(h)^\top \Sigma_{F_\ell(h)} J_\ell(h). \quad (5)$$

The rank of  $\mathbf{G}_\ell(h)$  is at most  $\min(V-1, \text{rank } J_\ell(h))$ ; directions in  $\ker J_\ell$  are perturbations invisible to the output. Because  $\mathbf{G}_\ell$  depends on the local hidden state through both the Jacobian and the output distribution, activations close in Euclidean distance may be far apart under the pullback metric, and vice versa (Appendix A).

**Two roles of geometry, and a layered mismatch.** Activations  $h_\ell \in \mathbb{R}^{d_\ell}$  are predictive objects: their meaning is the next-token distribution  $F_\ell(h)$  they encode, so the natural metric on activation space is the pullback Fisher–Rao metric  $\mathbf{G}_\ell$  (Eq. 5), which makes two activations close iff they predict similar distributions. The standard SAE training objective is the ambient  $\ell_2$  reconstruction loss on activation vectors, the prevalent reconstruction objective for SAEs [4, 5, 1, 7]; this objective uses the ambient Euclidean inner product as the distance on activation space rather than the pullback metric. The geometric mismatch we characterise has two sources, both arising from this Euclidean-for-pullback substitution: a *basis–manifold* mismatch, where the SAE’s flat  $k$ -sparse linear dictionary cannot match an activation manifold curved under either metric; and an *objective–metric* mismatch, where the  $\ell_2$  objective optimises Euclidean distance rather than pullback distance. The empirical estimators  $\{d_{\text{int}}, \kappa_{\text{ms}}, \kappa_\ell^{\text{tv}}, \nu\}$  used throughout (App. B) are extrinsic Euclidean quantities that encode the manifold structure under the ambient metric; their mathematical relationship to the pullback geometry (intrinsic-dimension invariance, gauge transformation of curvature) is developed in App. A.

**Geometric summaries.** We characterise  $\mathcal{H}_\ell$  through four layerwise scalar summaries that quantify aspects of local manifold structure:

- *Intrinsic dimension*  $d_{\text{int},\ell}$ : the number of local degrees of freedom of the activation manifold at layer  $\ell$ .
- *Multi-scale curvature*  $\kappa_{\text{ms},\ell}$ : the average departure of the manifold from its tangent plane across local neighbourhoods of varying radius.
- *Tangent variation*  $\kappa_\ell^{\text{tv}}$ : the rate at which the tangent space rotates between nearby points on the manifold.
- *Heterogeneity*  $\nu_\ell$ : the average within-neighbourhood variation in local intrinsic dimension across the layer, capturing how locally rough or non-uniform the manifold’s intrinsic dimension is at small spatial scales.

Each controls sparse linear approximation efficiency: higher intrinsic dimension demands more atoms to span the tangent structure; higher curvature shrinks each atom’s region by introducing second-order departure from any tangent approximation; higher heterogeneity prevents any single dictionary geometry from being uniformly efficient across the manifold. Estimation procedures are detailed in Appendix B (two-nearest-neighbour (TWO-NN) [25] for  $d_{\text{int}}$ , multi-scale principal-component-analysis (PCA) residual estimator for  $\kappa_{\text{ms}}$ , neighbour Gauss-map estimator for  $\kappa_\ell^{\text{tv}}$ , pointwise-dimension dispersion estimator for  $\nu$ ).

### 3 Geometry-Conditioned Account of SAE Scaling

We ask whether the layerwise behaviour of the surface  $L_\ell(n, k)$  is conditioned by activation geometry  $g_\ell$ . Two structural constraints from the released SAE family shape our methodology. First, only two *backbone* dictionary widths are shared across all 68 layers ( $n_{\text{lo}} = 16\text{K}$ ,  $n_{\text{hi}} = 65\text{K}$  for 2B;  $n_{\text{lo}} = 16\text{K}$ ,  $n_{\text{hi}} = 128\text{K}$  for 9B); a 6-parameter with-floor surface (Eq. 2) is therefore not identifiable per-layer at most layers. Second, Gao et al. [1] themselves first fit a no-floor power law to their SAE data, then introduced the asymptotic floor  $B(k)$  as a refinement at the layer where richer width coverage made the floor identifiable. We follow the same natural progression: a per-layer no-floor surface fit at *every* layer (§3.1), and a refit with the with-floor form at the six *showcase* layers per model where the grid contains  $\geq 3$  widths (§3.3). The cross-layer geometric question is then answered by a Stage 2 regression of the per-layer Stage 1 scaling-law parameters and derived width exponent against the layerwise geometric summaries (§3.2).

#### 3.1 Stage 1: per-layer no-floor surface fits

At each layer  $\ell$  we fit the floor-free log-linearisation of Eq. 3 as a 4-parameter surface,

$$\log L_\ell(n, k) = a_{0,\ell} + \beta_{n,\ell} \log n + \beta_{k,\ell} \log k + \gamma_\ell \log n \log k, \quad (6)$$

on the layer’s full grid of  $(n, k, L)$  triples (interpolated across the layer’s realised sparsities by piecewise cubic Hermite interpolating polynomial (PCHIP); Section 4). The fit is identifiable from the two backbone widths shared across all layers and tightens at the six showcase layers with  $\geq 3$  widths. From the per-layer parameters  $(\beta_{n,\ell}, \gamma_\ell)$  we read off the layerwise width-scaling exponent at any sparsity target  $k$ ,

$$\alpha_\ell(k) = -(\beta_{n,\ell} + \gamma_\ell \log k). \quad (7)$$

Because the fit assumes  $B = 0$ , the per-layer  $\alpha_\ell(k)$  obtained from Eq. 7 is the *apparent* width-scaling exponent in the regime where the floor’s contribution at observed widths is small; when a non-zero floor is present it underestimates the asymptotic with-floor exponent (Appendix C). Throughout §5 we use  $\alpha_\ell(k)$  as the per-layer Stage 1 exponent at  $k = 50$  unless otherwise stated; the calibration of this exponent against the with-floor fit at showcase layers is reported in §3.3.

### 3.2 Stage 2: cross-layer geometric regression

The Stage 1 outputs supply three per-layer scalar targets: the derived width exponent  $\alpha_\ell(k)$  (Eq. 7), the width-scaling coefficient  $\beta_{n,\ell}$ , and the width-sparsity interaction coefficient  $\gamma_\ell$ . The four layerwise geometric summaries  $\mathbf{g}_\ell = (d_{\text{int},\ell}, \kappa_{\text{ms},\ell}, \kappa_\ell^{\text{tv}}, \nu_\ell)$  defined in §2.4 enter the regression after a uniform log transform with light 1/99-percentile trimming and per-feature standardisation (Appendix B); tables and figures use raw symbols throughout. We regress each target  $y_\ell$  on the layerwise geometric summaries via ordinary least squares (OLS):

$$y_\ell = \mu + \sum_{p=1}^P \theta_p g_{\ell,p} + \varepsilon_\ell, \quad P \in \{1, 2, 4\}. \quad (8)$$

Regression is fit per model (no across-model pooling), so that model-specific intercepts are not conflated with shared geometric structure.

**Hypothesis hierarchy.** For each target  $y_\ell$  we compare a nested family of models, each corresponding to a substantive claim about whether and how layerwise geometry shapes the target:

- **H0** (*geometry-invariant null*): the per-layer target is a constant across layers,  $y_\ell = \mu + \varepsilon_\ell$  (1 parameter). For  $y_\ell = \alpha_\ell(k = 50)$ , this is the hypothesis that the per-layer width-scaling exponent is the same at every layer, i.e., the scaling law is universal across depth and geometry plays no role. H0 is the natural null because it is what a single-layer scaling-law fit (the form used by Gao et al. [1]) implicitly assumes when extrapolated to other layers.
- **H1<sub>g</sub>** (*single-feature dependence*): the per-layer target depends on one geometric feature  $g$  (2 parameters; one model per feature). Tests whether any single geometric summary explains layerwise variation.
- **H2<sub>low $\rho$</sub>**  (*lowest-pairwise-correlation feature pair*): the per-layer target depends on the two geometric features with the smallest within-model  $|\rho|$  (3 parameters;  $d_{\text{int}} + \kappa_{\text{ms}}$  at 9B and  $d_{\text{int}} + \nu$  at 2B). Tests whether two minimally-collinear features carry independent additive variance. The complementary best-single + next-best additivity check is the H1  $\rightarrow$  H2<sub>best+next</sub> step of the F-test ladder in Appendix G.
- **H<sub>full</sub>** (*full geometric dependence*): the per-layer target depends on all four features (5 parameters). Tests whether the joint geometric specification adds further explanatory power beyond any pair.

Rejection of H0 (in favour of any H1, H2<sub>low $\rho$</sub> , or H<sub>full</sub>) is rejection of the geometry-invariant scaling-law assumption: it shows that the target varies systematically with measurable geometric properties of the activation manifold rather than being a single layerwise constant.

**Where geometry enters: intercept vs slope.** The Stage 1 exponent  $\alpha_\ell(k) = -(\beta_{n,\ell} + \gamma_\ell \log k)$  has two channels through which geometry shapes the curve: the intercept  $\beta_{n,\ell}$ , which sets the level of the curve, and the interaction  $\gamma_\ell$ , which sets its tilt with respect to  $\log k$ . Regressing  $\beta_{n,\ell}$  and  $\gamma_\ell$  on geometry (Eq. 8) tests where in the curve the geometric signal sits.

**Evaluation.** For each fitted model we report in-sample  $R^2$ , leave-one-/two-/three-layer-out  $R^2$  (LOO/L2O/L3O; closed form via the hat matrix), the Akaike and Bayesian information criteria (AIC, BIC), and an F-test against H0; layer-permutation  $p$ -values are in Appendix G.

**Cross-model transfer.** To test whether the geometric law transfers across models, we fit the regression on one model’s layers and predict the other model’s per-layer  $\alpha(k = 50)$  (standardisation procedure that prevents test-set leakage in Appendix F). Cross-model transfer  $R^2$  is reported alongside the test model’s in-sample  $R^2$  as the upper bound for fixed-coefficient prediction on the test set (§5).

### 3.3 Floor calibration at showcase layers

At the six layers where the released grid supplies  $\geq 3$  widths after monotone filtering (2B {5, 12, 19}, 9B {9, 20, 31}), we re-fit the full 6-parameter with-floor surface of Eq. 2 by joint nonlinear least squares on  $\log L$  (procedure, monotone filters, and multi-start seed schedule in Appendix C). The resulting per-layer asymptotic floor  $B_\ell(k)$  is the strict reading of the *geometric wall*: the level of reconstruction error that cannot be eliminated by adding dictionary atoms. Because  $B$  is identifiable only at six layers, we report a ranking against the layerwise geometric summaries (§5), not a formal cross-layer regression; the curvature mechanism is discussed in §5 and Appendix E.

## 4 Experimental Setup

**Models and SAE family.** We study the residual stream of Gemma 2 2B ( $\mathcal{L}=26$  layers,  $d=2304$ ) and 9B ( $\mathcal{L}=42$  layers,  $d=3584$ ) [26], using all publicly released JumpReLU SAEs from Gemma Scope [7]. The released checkpoint grid contains two backbone widths shared across all layers (16K and 65K for 2B; 16K and 128K for 9B), and richer grids of up to seven widths at three *showcase* layers per model: {5, 12, 19} for 2B and {9, 20, 31} for 9B. Total: 312 checkpoints for 2B and 532 for 9B (844 total). At each (layer, width) bundle the released SAEs span a range of training sparsity targets, yielding multiple checkpoints at varying  $L_0$ . The released SAEs use a width-dependent training budget (4B/8B/16B tokens for dictionary widths  $\leq 16K / 32K\text{--}524K / 1M$  respectively); our scaling quantities describe effective properties of the released Gemma Scope SAE family rather than convergence-matched width-only exponents (Appendix H).

**Disjoint geometry/error data.** Geometry estimation and error measurement use *disjoint* data, so geometric summaries cannot overfit to token-level artefacts. We partition the Colossal Clean Crawled Corpus (C4) validation set [27] by document index (sequences 0–5K for geometry, 5K–10K for error), extracting 50K activation vectors per layer per partition. Cross-corpus validation uses WikiText-103 [28] for the error measurement while keeping geometry on C4.

**Code availability.** Source code for fitting the per-layer surfaces, the Stage 2 regressions, the validation suite, the with-floor calibration, and all figures and tables is released at this anonymous repository.

**Reconstruction error and sparsity.** The checkpoint-level loss  $L$  and sparsity  $L_0$  are computed as a symmetric trimmed mean of the per-token NMSE and per-token active-latent count over the same token subset, so the  $(L, L_0)$  pair shares its underlying tokens by construction (preprocessing details, norm-free cross-check metric, and sensitivity analyses in Appendix B).

**Statistical protocol.** Implementation follows the protocol of §3: Stage 1 OLS with PCHIP [29] sparsity-axis interpolation; Stage 2 closed-form leave- $K$ -layer-out via the hat matrix [30, 31] for  $K \in \{1, 2, 3\}$ ; with-floor calibration at the showcase layers by multi-start Nelder-Mead [32] on the joint 6-parameter surface (Appendix C). Supplementary tests (permutation, F-test ladders, cross-model transfer, PCHIP leave-one- $L_0$ -out CV) are in Appendix G.

## 5 Results

**Geometry and reconstruction co-vary across depth.** Figure 1 presents the phenomenological views of the released SAE family that motivate the analysis. Panel (a) shows the layerwise checkpoint-level NMSE profile (on C4 and WikiText-103) alongside the four layerwise geometric summaries. The NMSE profile rises through early layers, peaks at mid-depth, dips through late-mid layers, and exhibits a sharp uptick at the final layers; C4 and WikiText curves track each other throughout. The geometric summaries show partly overlapping but distinct depth signatures: intrinsic dimension and heterogeneity peak at mid-depth and decline toward the end, while curvature and tangent variation rise through early layers and plateau without returning. Panel (b) stratifies the layerwise NMSE at the upper backbone width by sparsity bin ( $L_0$ ); increasing  $L_0$  uniformly lowers the error level, but the depth profile is preserved across every sparsity stratum, so sparsity shifts the absolute level without reshaping the depth profile. Panel (c) stratifies the layerwise NMSE (averaged over  $L_0$ ) by dictionary width. The two backbone curves run roughly parallel in absolute terms across most of the network,

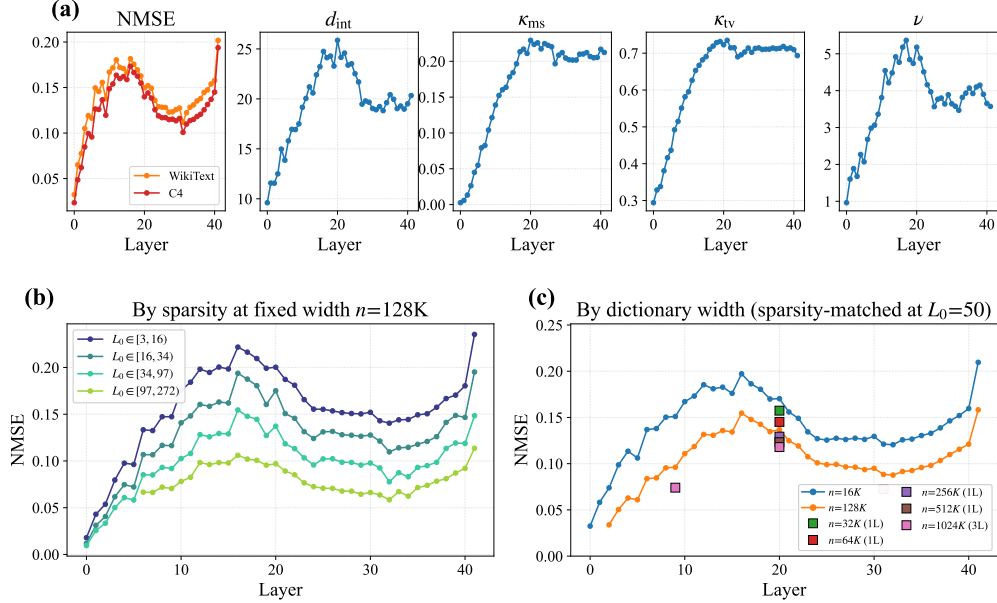


Figure 1: Phenomenology of layerwise SAE reconstruction for Gemma 2 9B. **(a)** Layerwise checkpoint-level NMSE profile (C4 and WikiText-103) alongside the four geometric summaries; NMSE peaks mid-depth and the two corpora track each other closely. **(b)** NMSE at the upper backbone width ( $n = 128K$ ) stratified by  $L_0$  bin; increasing  $L_0$  shifts the level without reshaping the depth profile. **(c)** NMSE per layer per dictionary width, sparsity-matched at  $L_0 = 50$  via PCHIP; the two backbone widths (16K, 128K) cover all 42 layers, while 32K, 64K, 256K, 512K appear at the central showcase layer and 1024K at all three showcase layers. 2B figure: Fig. 8.

but the multiplicative factor between them shrinks with depth, indicating that per-layer width-scaling efficiency varies with depth — the quantity we now characterise via the two-stage fit of §3.1.

**Per-layer width-scaling exponents are predicted by manifold geometry.** Figure 2(a) shows the per-layer Stage 1 exponent  $\alpha_\ell(k = 50)$  from Eq. 7 across all 26 layers of Gemma 2 2B and 42 layers of 9B. The profile descends from a high value at the input, reaches a local minimum at middle layers, then rises to a peak near depth 5/6 that lands within  $\sim 0.03$  of Gao et al. [1]’s reported  $\alpha = 0.181$  for GPT-4 at the same relative depth. Two factors plausibly account for the residual  $\Delta\alpha$ : (i) the floor-attenuation that lowers any no-floor fit when  $B(k) > 0$  (Appendix C); and (ii) Gao’s TopK SAEs produce steeper exponents than ReLU baselines in their own ablations, which would push their  $\alpha$  above ours independent of the floor effect (the JumpReLU comparison was not reported). The peak alignment at depth 5/6 suggests a cross-architecture similarity in the depthwise  $\alpha$  profile, but rests on a single anchor.

**Multi-scale curvature is the dominant geometric channel for  $\alpha_\ell(k = 50)$  in both models.** A single feature already explains nearly all cross-layer variation in  $\alpha$  at 2B and most of it at 9B (Table 1; 2B table in Appendix F); additional geometric features add little because they covary with  $\kappa_{\text{ms}}$  as aspects of the same underlying manifold complexity, with AIC/BIC favouring simpler models due to this collinearity rather than insufficient sample (the same  $H_{\text{full}}$  fits the smaller 2B sample as well as 9B). Geometrically, a layer’s width-scaling rate is primarily set by its multi-scale curvature  $\kappa_{\text{ms}}$  — how much the activation manifold departs from its local tangent plane across neighbourhoods of varying radius — because flatter manifolds scale better with width: each new atom captures a larger share of the local linear structure that the dictionary is trying to span. This  $\kappa_{\text{ms}}$  dominance holds across  $k \in \{16, 25, 32, 50, 64, 75\}$  (Appendix F).

**Geometry shapes both the baseline and the sparsity-tilt of  $\alpha_\ell(k)$ , the baseline more cleanly.** Decomposing  $\alpha_\ell(k) = -(\beta_{n,\ell} + \gamma_\ell \log k)$  separates the layer’s baseline scaling efficiency ( $\beta_n$ ) from how that efficiency varies with sparsity ( $\gamma$ ); Table 3 regresses each on geometry. The level channel

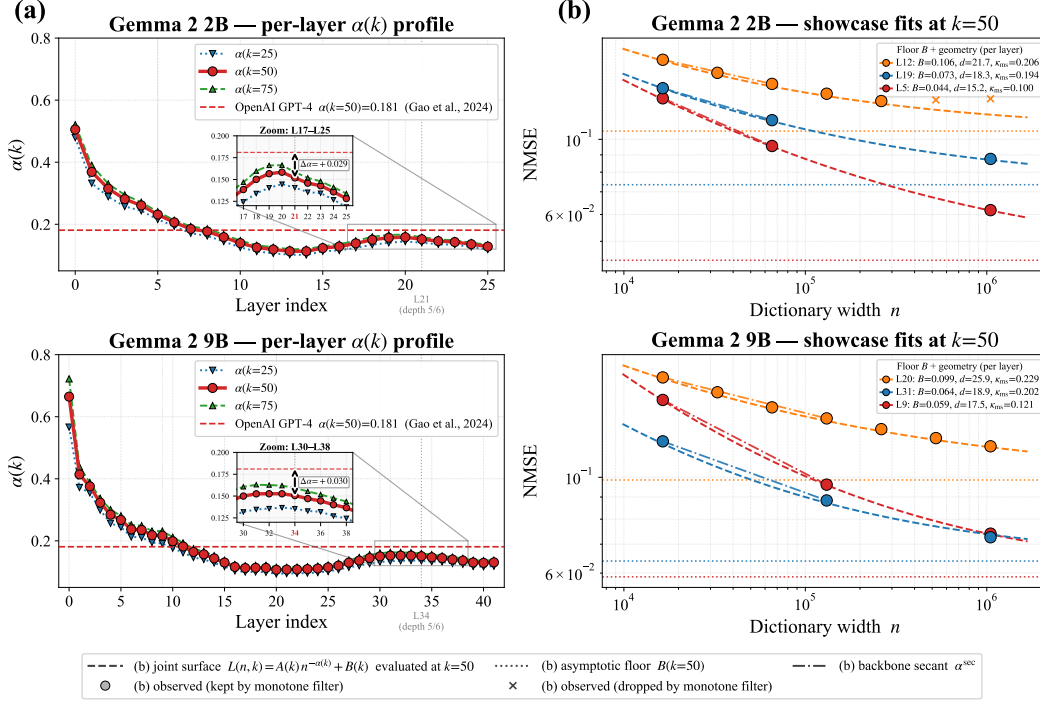


Figure 2: Per-layer width-scaling exponents and the geometric wall, in a  $2 \times 2$  grid (rows: models 2B/9B; columns: panels). **(a) Per-layer  $\alpha_\ell(k)$  profiles.** Stage 1 per-layer  $\alpha_\ell(k)$  from Eq. 7 at  $k \in \{25, 50, 75\}$  (blue dotted, red solid, green dashed;  $k = 50$  thicker). The dashed horizontal line marks the OpenAI value  $\alpha(k = 50) = 0.181$  at relative depth  $5/6$  of GPT-4 [1]; a vertical grey dotted line marks the depth- $5/6$  layer. Zoomed insets (one per row) annotate the  $\sim 0.03$  gap  $\Delta\alpha$  between our  $\alpha_\ell(k = 50)$  at depth  $5/6$  and Gao’s  $0.181$ , consistent with the no-floor underestimation analysis (Appendix C). **(b) Showcase-layer with-floor surface fits at  $k = 50$ .** Joint 6-parameter fits (Eq. 2, Appendix C) at the three showcase layers per model (2B  $\ell \in \{5, 12, 19\}$ , 9B  $\ell \in \{9, 20, 31\}$ ); dotted horizontal lines mark the fitted floor  $B_\ell(k = 50)$ , whose vertical ordering tracks the layers’  $d_{\text{int}}$  and  $\kappa_{\text{ms}}$  ordering in both models.

regresses more cleanly than the tilt, with both rejecting  $H_0$  at  $p < 10^{-6}$ , and curvature-related channels ( $\kappa_{\text{ms}}$  and  $\kappa^{\text{tv}}$ ) dominate the predictions for both targets in both models — though which specific channel wins varies by model and target (full per-feature numbers in the table). Geometry deforms the whole  $\alpha_\ell(k)$  curve coherently rather than selectively touching one channel, confirmed by a joint long-format regression (Appendix F).

**The geometry-to- $\alpha$  mapping transfers across models.** Cross-model transfer (Appendix F, Table 5) fits Stage 2 coefficients on one model and predicts the other model’s  $\alpha(k = 50)$  from its own per-layer geometry. In both directions, the coefficients learned on one model predict the other’s per-layer exponent to within a small gap of the test model’s own in-sample fit — the upper bound on any fixed-coefficient predictor evaluated on the full test set. Geometrically: the function mapping manifold geometry to width-scaling rate is preserved within the Gemma 2 family across the 2B→9B size jump, indicating it is not an idiosyncrasy of a single model size.

**The geometric wall: asymptotic floor tracks layerwise geometry.** At the six showcase layers where the released grid permits identification of the asymptotic floor (Fig. 2(b)), the within-model vertical ordering of fitted floors  $B_\ell(k = 50)$  matches the geometric ordering: layers with higher curvature and intrinsic dimension have higher floor (full values in Appendix E). The broader  $\alpha$  analysis is robust to the monotone-filter pipeline (91% cell retention) and to using a non-parametric backbone-secant in place of the Stage 1 fit (agreement within 0.01 at every layer; Appendices C, D).

Table 1: **Stage-2 cross-layer regression of per-layer width-scaling exponent  $\alpha_\ell(k = 50)$  on geometry, Gemma 2 9B ( $n = 42$  layers).** Multi-scale curvature  $\kappa_{\text{ms}}$  wins as the dominant single feature (LOO  $R^2 = 0.869$  vs 0.793 for the next-best  $\kappa^{\text{tv}}$ );  $H_{\text{full}}$ ’s lower LOO (0.806) than the  $H2_{d_{\text{int}}+\kappa_{\text{ms}}}$  pair (0.863) reflects feature collinearity at 9B rather than insufficient sample (the same  $H_{\text{full}}$  fits the smaller 2B sample cleanly, Table 2). Hypothesis hierarchy:  $H_0$  (intercept-only, the geometry-invariant null);  $H1_g$  (intercept + single geometric feature; one row per feature);  $H2_{d_{\text{int}}+\kappa_{\text{ms}}}$  (intercept + the lowest- $|\rho|$  pair, chosen to minimise pairwise collinearity at 9B);  $H_{\text{full}}$  (intercept + all four features). The complementary best+next-best additivity check is in Appendix G. Metrics: in-sample  $R^2$ , closed-form leave-one-/two-/three-layer-out  $R^2$  via the regression hat matrix, AIC, BIC, and an  $F$ -test against  $H_0$ . 2B counterpart,  $\beta_n/\gamma$  decompositions, and cross-model transfer in Appendix F.

Hypothesis	$R^2$	LOO	L2O	L3O	AIC	BIC	$F$ vs $H_0$	$p$
$H_0$	0.000	-0.049	-0.050	-0.051	-188	-186	—	—
$H1_{d_{\text{int}}}$	0.812	+0.738	+0.738	+0.737	-256	-253	173.2	$< 10^{-10}$
$H1_{\kappa_{\text{ms}}}$	0.929	+0.869	+0.869	+0.869	-297	-294	523.6	$< 10^{-10}$
$H1_{\kappa^{\text{tv}}}$	0.862	+0.793	+0.792	+0.791	-269	-266	250.6	$< 10^{-10}$
$H1_\nu$	0.821	+0.745	+0.744	+0.743	-258	-255	182.9	$< 10^{-10}$
$H2_{d_{\text{int}}+\kappa_{\text{ms}}}$	0.935	+0.863	+0.863	+0.862	-299	-294	281.2	$< 10^{-10}$
$H_{\text{full}}$	0.940	+0.806	+0.806	+0.805	-298	-290	145.5	$< 10^{-10}$

**A mechanistic reading.** SAE atoms approximate the activation manifold by sparse linear combinations whose best local approximation at each point is the tangent plane. When principal curvatures are non-zero, the tangent plane departs from the manifold by a second-order term proportional to (squared local distance)  $\times$  (curvature). In the small- $k$  regime to which the SAE scaling law of Gao et al. [1] is restricted (their Section 3.1.4), the  $k$ -sparse linear combination spans a  $k$ -dimensional flat affine subspace; this curvature mismatch leaves an irreducible fraction of each activation outside that span, a residual that does not vanish as  $n \rightarrow \infty$  even as atoms become dense on the manifold (full argument and two-channel decomposition in Appendix E). Higher curvature  $\kappa$  therefore raises the second-order residual, and higher intrinsic dimension  $d_{\text{int}}$  shrinks the share of the local tangent space that any finite atom budget can cover; both effects raise the asymptotic floor. Our  $d_{\text{int}}$  channel recovers Gao et al. [1]’s “spectrum of structure” hypothesis; the curvature channel is genuinely additional. The same geometric channels also depress the per-layer width-scaling rate at the depths of greatest curvature and intrinsic dimension: comparing the  $\alpha$  profile (Fig. 2(a)) with the layerwise geometry (Fig. 1(a)),  $\alpha$  is depressed in the middle and final layers — exactly the depth range where  $\kappa_{\text{ms}}$  (the dominant single-feature predictor of  $\alpha$ , though all four geometric summaries covary) has reached and remains at its plateau. Geometry therefore hurts twice: it depresses the scaling rate at the same depths where it elevates the floor.

**Validation.** The Stage 2 regressions pass both validation tests reported in Appendix G: layer-level permutation rejects the null at  $p \leq 0.01$  for every  $H1$  and  $H2_{\text{low}\rho}$ ; and the layerwise NMSE profile is preserved between C4 and WikiText-103 in both models (Fig. 1a).

## 6 Conclusion

We presented a geometry-conditioned account of layerwise SAE scaling. A two-stage analysis of 844 checkpoints across all layers of Gemma 2 2B and 9B (per-layer fits, then cross-layer geometric regression) shows that activation manifold geometry predicts the per-layer width-scaling exponent. The same regression coefficients learnt on one model predict the other model’s per-layer exponents under cross-model transfer, indicating a transferable geometric law.

At the six showcase layers where richer width grids permit identification of the asymptotic floor, the fitted floor varies with geometry: layers with higher curvature and intrinsic dimension have higher floor. This is the strict reading of the *geometric wall*: a floor-like saturation level whose height tracks the manifold the SAE is trying to approximate, consistent with a second-order curvature residual that no sparse linear dictionary can eliminate. Broader implications for SAEs are discussed in Appendix I.

## Acknowledgments

This research was partially supported by the Australian Research Council through an Industrial Transformation Training Centre for Information Resilience (IC200100022). Quan Nguyen is supported by a NHMRC Investigator Grant (GNT2008928). Maciej Trzaskowski is the Managing Director of Profenso; this work was conducted independently of Profenso.

## References

- [1] Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations (ICLR)*, 2025. Oral.
- [2] Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning (ICML)*, 2024.
- [3] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL [https://transformer-circuits.pub/2022/toy\\_model/index.html](https://transformer-circuits.pub/2022/toy_model/index.html).
- [4] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations (ICLR)*, 2024.
- [5] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L. Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Chris Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL <https://transformer-circuits.pub/2023/monosemantic-features/index.html>.
- [6] Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024. URL <https://transformer-circuits.pub/2024/scaling-monosemanticity/>.
- [7] Tom Lieberum, Senthoran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on Gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 278–300, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.19.
- [8] Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *International Conference on Learning Representations (ICLR)*, 2025.
- [9] Yuxiao Li, Eric J. Michaud, David D. Baek, Joshua Engels, Xiaoqing Sun, and Max Tegmark. The geometry of concepts: Sparse autoencoder feature structure. *Entropy*, 27(4):344, 2025. arXiv:2410.19750, 2024.
- [10] Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *International Conference on Learning Representations (ICLR)*, 2025.

- [11] Chris Olah and Adam Jermy. What is a linear representation? what is a multidimensional feature? Transformer Circuits Thread, July 2024 update, 2024. URL <https://transformer-circuits.pub/2024/july-update/index.html>.
- [12] Eric J. Michaud, Liv Gorton, and Tom McGrath. Understanding sparse autoencoder scaling in the presence of feature manifolds, 2025.
- [13] Alexander Modell, Patrick Rubin-Delanchy, and Nick Whiteley. The origins of representation manifolds in large language models, 2025.
- [14] Wes Gurnee, Emmanuel Ameisen, Isaac Kauvar, Julius Tarng, Adam Pearce, Chris Olah, and Joshua Batson. When models manipulate manifolds: The geometry of a counting task. Transformer Circuits Thread, 2026.
- [15] Joshua Engels, Logan Riggs, and Max Tegmark. Decomposing the dark matter of sparse autoencoders. *Transactions on Machine Learning Research*, 2025. arXiv:2410.14670.
- [16] Wes Gurnee. SAE reconstruction errors are (empirically) pathological. AI Alignment Forum, 2024. URL <https://www.lesswrong.com/posts/rZPiuFxEsMxCDHe4B/sae-reconstruction-errors-are-empirically-pathological>.
- [17] Alessio Ansuini, Alessandro Laio, Jakob H. Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.
- [18] Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.
- [19] Mohamed A. Mabrok. Latent semantic manifolds in large language models, 2026.
- [20] Kiho Park, Todd Nief, Yo Joong Choe, and Victor Veitch. The information geometry of softmax: Probing and steering, 2026.
- [21] Senthoran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with JumpReLU sparse autoencoders, 2024.
- [22] Shun-ichi Amari. *Information Geometry and Its Applications*, volume 194 of *Applied Mathematical Sciences*. Springer, Tokyo, 2016.
- [23] N. N. Cencov. *Statistical Decision Rules and Optimal Inference*, volume 53 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 1982.
- [24] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 35:99–109, 1943.
- [25] Elena Facco, Maria d’Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7:12140, 2017.
- [26] Gemma Team. Gemma 2: Improving open language models at a practical size, 2024.
- [27] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.
- [28] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations (ICLR)*, 2017.
- [29] F. N. Fritsch and R. E. Carlson. Monotone piecewise cubic interpolation. *SIAM Journal on Numerical Analysis*, 17(2):238–246, 1980.
- [30] D. M. Allen. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16(1):125–127, 1974.

- [31] M. Stone. Cross-validated choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–133, 1974.
- [32] J. A. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313, 1965.
- [33] John M. Lee. *Introduction to Riemannian Manifolds*, volume 176 of *Graduate Texts in Mathematics*. Springer, Cham, 2nd edition, 2018.
- [34] Manfredo P. do Carmo. *Riemannian Geometry*. Birkhäuser, Boston, MA, 1992.
- [35] Nihat Ay, Jürgen Jost, Hông Vân Lê, and Lorenz Schwachhöfer. *Information Geometry*, volume 64 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer, Cham, 2017.
- [36] Anna V. Little, Mauro Maggioni, and Lorenzo Rosasco. Multiscale geometric methods for data sets I: Multiscale SVD, noise and curvature. *Applied and Computational Harmonic Analysis*, 43(3):504–567, 2017.
- [37] Alan Edelman, Tomás A. Arias, and Steven T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998. doi: 10.1137/S0895479895290954.
- [38] Jack Sherman and Winifred J. Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, 21(1):124–127, 1950. doi: 10.1214/aoms/1177729893.

## A Differential-Geometric Foundations

This appendix provides the full differential-geometric development summarised in Section 2.4. Standard references are Lee [33], do Carmo [34] for Riemannian geometry and Amari [22], Ay et al. [35] for information geometry.

**Smooth manifolds.** A **smooth manifold**  $\mathcal{M}$  of dimension  $m$  is a space that locally resembles  $\mathbb{R}^m$ : it is covered by coordinate charts  $(U_\alpha, \varphi_\alpha)$ , each mapping an open patch of  $\mathcal{M}$  smoothly onto a region of  $\mathbb{R}^m$ , with smooth ( $C^\infty$ ) transition functions on overlaps. At each point  $q \in \mathcal{M}$ , the **tangent space**  $T_q\mathcal{M}$  collects all velocity vectors of smooth curves through  $q$ ; in coordinates  $(x^1, \dots, x^m)$  it is spanned by  $\{\partial/\partial x^\mu\}_{\mu=1}^m$ .

**Smooth maps and immersions.** A smooth map  $f : \mathcal{M} \rightarrow \mathcal{N}$  between manifolds has a **differential**  $df_q : T_q\mathcal{M} \rightarrow T_{f(q)}\mathcal{N}$  at each point, represented in coordinates by the Jacobian  $J_f = \partial f/\partial x$ . When  $df_q$  is injective everywhere ( $J_f$  has full column rank),  $f$  is an **immersion**; if additionally  $f$  is a homeomorphism onto its image, it is an **embedding**.

**Riemannian metrics.** A **Riemannian metric** on  $\mathcal{M}$  assigns to each point  $q$  an inner product  $g_q$  on  $T_q\mathcal{M}$ , varying smoothly. In coordinates,  $g$  is represented by a symmetric, positive definite matrix  $\mathbf{G}(x) = [g_{\mu\nu}(x)]$ , defining lengths, angles, and curvatures intrinsic to  $\mathcal{M}$ . The pair  $(\mathcal{M}, g)$  is a Riemannian manifold.

**Pullback metrics.** Given a Riemannian metric  $\mathbf{G}_{\mathcal{N}}$  on  $\mathcal{N}$  and a smooth map  $f : \mathcal{M} \rightarrow \mathcal{N}$ , the **pullback metric** on  $\mathcal{M}$  is

$$f^*\mathbf{G}_{\mathcal{N}}(x) = J_f(x)^\top \mathbf{G}_{\mathcal{N}}(f(x)) J_f(x). \quad (9)$$

This is positive semidefinite by construction, and is a genuine Riemannian metric if and only if  $f$  is an immersion. The pullback measures distances in  $\mathcal{M}$  by how much  $f$  stretches or compresses them in  $\mathcal{N}$ .

**The Fisher–Rao metric.** For a parametric family  $\{p(\cdot; \theta) : \theta \in \Theta\}$ , the **Fisher information metric** is

$$\mathbf{I}(\theta) = \mathbb{E}_{x \sim p(\cdot; \theta)} \left[ \nabla_{\theta} \log p(x; \theta) \nabla_{\theta} \log p(x; \theta)^{\top} \right]. \quad (10)$$

By Chentsov’s theorem [23],  $\mathbf{I}$  is the unique (up to scaling) Riemannian metric on statistical models invariant under sufficient statistics.

For categorical distributions with  $\boldsymbol{\pi} \in \Delta_{\circ}^{V-1}$ , the Fisher metric in simplex coordinates is  $[\mathbf{G}_{\text{FR}}]_{ab} = \delta_{ab}/\pi_a$ ; this expression uses the ambient  $V$ -component representation, and restricting to the  $V-1$  independent coordinates after imposing  $\sum_c \pi_c = 1$  produces the off-diagonal cross-terms  $g_{ab} = \delta_{ab}/\pi_a + 1/\pi_V$ . In logit coordinates ( $\pi_i = e^{z_i} / \sum_j e^{z_j}$ ), it becomes  $\boldsymbol{\Sigma}_{\boldsymbol{\pi}} = \text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^{\top}$ , with rank  $V-1$ . The manifold  $(\Delta_{\circ}^{V-1}, \mathbf{G}_{\text{FR}})$  has constant positive curvature, isometric to the positive orthant of a sphere of radius 2 via  $\boldsymbol{\pi} \mapsto 2\sqrt{\boldsymbol{\pi}}$  [22], with geodesic (Bhattacharyya) distance [24]:

$$d_{\text{FR}}(\boldsymbol{\pi}, \boldsymbol{\pi}') = 2 \arccos \left( \sum_{c=1}^V \sqrt{\pi_c \pi'_c} \right). \quad (11)$$

**Geodesics and curve lengths.** The **length** of a smooth curve  $\gamma : [0, 1] \rightarrow \mathcal{M}$  on a Riemannian manifold  $(\mathcal{M}, \mathbf{G})$  is

$$\text{Len}(\gamma) = \int_0^1 \sqrt{\dot{\gamma}(t)^{\top} \mathbf{G}(\gamma(t)) \dot{\gamma}(t)} dt. \quad (12)$$

A **geodesic** is a length-minimising curve, the Riemannian analogue of a straight line. Geodesics are equivalently characterised as critical points of the energy functional  $E(\gamma) = \frac{1}{2} \int_0^1 \dot{\gamma}^{\top} \mathbf{G} \dot{\gamma} dt$  (which avoids the square root and yields the same minimisers up to reparametrisation), leading to the geodesic equation

$$\ddot{\gamma}^k + \sum_{\mu, \nu} \Gamma_{\mu\nu}^k(\gamma) \dot{\gamma}^{\mu} \dot{\gamma}^{\nu} = 0, \quad k = 1, \dots, m, \quad (13)$$

where  $\Gamma_{\mu\nu}^k = \frac{1}{2} \sum_{\lambda} G^{k\lambda} (\partial_{\mu} G_{\nu\lambda} + \partial_{\nu} G_{\mu\lambda} - \partial_{\lambda} G_{\mu\nu})$  are the Christoffel symbols of the Levi-Civita connection and  $[G^{k\lambda}] = [G_{k\lambda}]^{-1}$ . This formula presupposes the metric is non-degenerate; for the pullback  $\mathbf{G}_{\ell} = J_{\ell}^{\top} \boldsymbol{\Sigma} J_{\ell}$  we therefore assume  $F_{\ell}$  is an immersion at  $h$  (i.e.,  $J_{\ell}$  has full column rank), and where  $J_{\ell}$  is rank-deficient the inverse is taken as the Moore–Penrose pseudoinverse restricted to the orthogonal complement of  $\ker J_{\ell}$ . In flat space with Cartesian (affine) coordinates,  $\Gamma_{\mu\nu}^k \equiv 0$  and geodesics are straight lines.

In the hidden activation space, the pullback-Fisher length of any curve  $\gamma$  equals the Fisher–Rao length of its image  $F_{\ell} \circ \gamma$  in the simplex. Geodesics under  $\mathbf{G}_{\ell}$  are therefore exactly those hidden-space curves whose predictive image traces a geodesically efficient path in output space.

**Connection to the empirical estimators.** The estimators introduced in App. B ( $d_{\text{int}}, \kappa_{\text{ms}}, \kappa^{\text{tv}}, \nu$ ) are computed from the activation point cloud under the ambient Euclidean metric. Three relationships position them within the pullback framework above. *Metric invariance of intrinsic dimension.* The intrinsic dimension is a topological property of the manifold, independent of the choice of (non-degenerate) Riemannian metric, so the Euclidean TWO-NN estimate is consistent with the pullback intrinsic dimension whenever  $J_{\ell}$  has full column rank (i.e.,  $F_{\ell}$  is an immersion at  $h$ ). *Gauge structure of curvature.* The pullback curvature equals the Euclidean curvature transformed by  $J_{\ell}(h)^{\top} \boldsymbol{\Sigma}_{F_{\ell}(h)} J_{\ell}(h)$ , a continuous deformation of magnitude that preserves the qualitative type: a manifold curved under one metric is curved under the other, with the scales gauge-related. *Inheritance of invariance by heterogeneity.* The estimator  $\nu$  is the local-neighbourhood standard deviation of  $d_{\text{int}}$  and inherits its metric invariance. Direct computation of pullback quantities at each layer requires evaluating  $J_{\ell}(h)$  at every activation, which is computationally expensive at language-model scale and is reserved as a follow-up direction (App. I).

## B Reconstruction Metrics and Preprocessing

**Reconstruction loss.** Per-token NMSE( $h_i, \hat{h}_i$ ) =  $\|h_i - \hat{h}_i\|^2 / \|h_i\|^2$  is heavy-tailed on language-model activations: a small fraction of tokens have anomalously small or large  $\|h\|$ , producing outlier per-token NMSE values from the normalisation alone. We therefore aggregate per-token NMSE via a 5/95 symmetric trim, averaging the inner 90% of per-token values to form the checkpoint-level

loss  $L$ . Figure 6 shows the per-token NMSE distribution at six representative (model, layer, width,  $L_0 \approx 50$ ) cells: the right tail of the per-token NMSE histogram is heavy enough that the untrimmed mean is dominated by the tail and diverges from the median; the trimmed mean tracks the median.

**Sparsity.** The same per-token subset that defines  $L$  is used to compute  $L_0$ : the trimmed mean of the per-token active-latent count over the inner-90% token set. This makes the  $(L, L_0)$  pair share its underlying token sample by construction.

**Geometry preprocessing.** The top and bottom 5% of activations *by norm* are dropped before computing geometric summaries, preventing extreme-norm tokens from distorting the  $k$ -NN graphs that underlie the TWO-NN intrinsic-dimension estimator, the multi-scale curvature estimator, the tangent-variation estimator, and the heterogeneity estimator. Activations are mean-centred globally before  $k$ -NN construction; they are not whitened or  $L_2$ -normalised, both of which would distort the curvature signal.

**Geometric estimator hyperparameters.** *Notation: throughout this paragraph,  $K$  (with optional subscripts) denotes a nearest-neighbour count, distinct from the sparsity  $k$  used in the scaling law and from the LKO holdout count  $K$  of Appendix G.* A single shared nearest-neighbour graph is precomputed per layer with  $K_{\max} = 500$  neighbours (Euclidean metric) and reused across all four estimators. A per-layer local tangent dimension  $k_t$  is set as the median of the per-point TWO-NN  $d_{\text{int}}$  estimates, clipped to  $[1, \min(K - 1, d - 1)]$  where  $K$  is the relevant neighbourhood size and  $d = d_\ell$ ; the curvature estimators below use this dynamic  $k_t$  rather than a fixed default. The four layerwise summaries are then computed as follows.

*Intrinsic dimension  $d_{\text{int}}$ .* The TWO-NN estimator of Facco et al. [25] uses the ratio of the 2nd to the 1st nearest-neighbour distance at each point; per-point estimates are averaged within a local window of  $K = 100$  neighbours and then over the layer.

*Multi-scale curvature  $\kappa_{\text{ms}}$ .* At each point  $h$  and at two neighbourhood scales (a small intra-cell scale  $K_{\text{small}} = 30$  nearest neighbours and a coarse scale  $K_{\text{large}} = 200$ ), the local sample covariance  $C(K) = \frac{1}{K} \sum_{j=1}^K (h_j - h)(h_j - h)^\top$  is formed and its eigenvalues  $\lambda_1 \geq \dots \geq \lambda_d \geq 0$  are computed. The residual fraction at scale  $K$  is

$$\text{rf}(K) = 1 - \frac{\sum_{i=1}^{k_t} \lambda_i}{\sum_{i=1}^d \lambda_i},$$

the share of local variance unexplained by the top- $k_t$  principal directions of  $C(K)$  (the same  $k_t$  at both scales). The contrast across scales is the difference  $\kappa_{\text{ms}}(h) = \max(0, \text{rf}(K_{\text{large}}) - \text{rf}(K_{\text{small}}))$ , clamped at zero, then averaged over the layer. The contrast captures the additional variance that escapes the local tangent fit when the neighbourhood is widened: a flat manifold has  $\text{rf}(K_{\text{large}}) \approx \text{rf}(K_{\text{small}})$  and small  $\kappa_{\text{ms}}$ , while a curved manifold has  $\text{rf}(K_{\text{large}}) > \text{rf}(K_{\text{small}})$ . Multi-scale local-PCA estimators of curvature on point clouds are studied in Little et al. [36].

*Tangent variation  $\kappa^{\text{tv}}$ .* At each point  $i$ , a local tangent basis  $U_i \in \mathbb{R}^{d \times k_t}$  is computed by local PCA on the  $K = 50$  nearest neighbours, retaining the top- $k_t$  eigenvectors (orthonormal columns by construction). For each point, the  $n_{\text{compare}} = 5$  nearest neighbours  $j$  are used to compute the per-pair tangent disagreement

$$1 - \frac{1}{k_t} \|U_i^\top U_j\|_F^2 = \frac{1}{k_t} \sum_{p=1}^{k_t} \sin^2(\theta_p^{ij}),$$

where  $\{\theta_p^{ij}\}_{p=1}^{k_t}$  are the principal angles between the two tangent subspaces. The estimator is  $\kappa^{\text{tv}}(i) = n_{\text{compare}}^{-1} \sum_j (1 - k_t^{-1} \|U_i^\top U_j\|_F^2)$ , then averaged over the layer. The subspace distance is the squared chordal (projection-Frobenius) distance between subspaces in the Grassmannian  $\text{Gr}(k_t, d)$  [37], normalised by  $k_t$  to lie in  $[0, 1]$  (0 when the two tangent subspaces coincide, 1 when they are mutually orthogonal).

*Heterogeneity  $\nu$ .* The standard deviation of local TWO-NN intrinsic-dimension estimates across each point's  $K = 50$  neighbourhood, then averaged over the layer; large  $\nu$  flags layers with high local roughness in pointwise intrinsic dimension (high within-neighbourhood variability of  $d_{\text{int}}$ ).

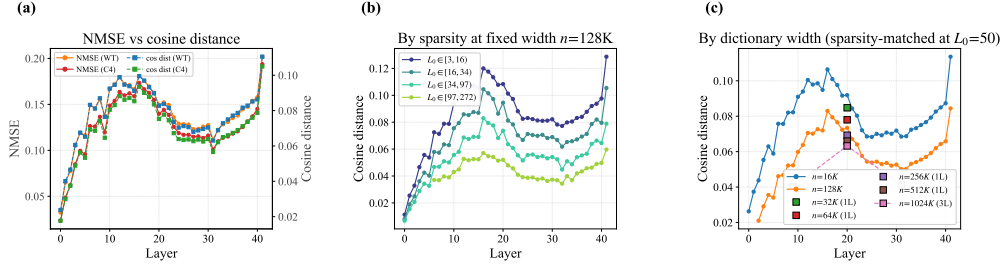


Figure 3: NMSE versus cosine distance (Gemma 2 9B). **(a)** Layerwise profile of NMSE (left axis) and cosine distance (right axis), overlaid for C4 and WikiText-103. The two metrics share a depth shape and the C4/WikiText overlap is preserved in both. **(b)** Cosine distance at the upper backbone width ( $n = 128K$ ) stratified by  $L_0$  bin. **(c)** Cosine distance per layer per dictionary width, sparsity-matched at  $L_0 = 50$  via PCHIP.

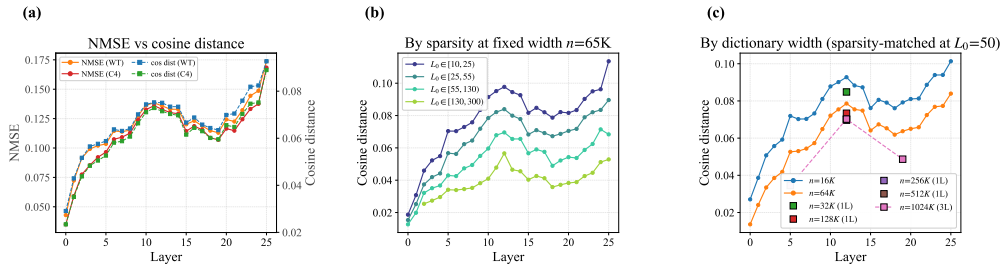


Figure 4: NMSE versus cosine distance (Gemma 2 2B), matching Fig. 3 on the 2B model. **(a)** Layerwise profile of NMSE and cosine distance, overlaid for C4 and WikiText-103. **(b)** Cosine distance at the upper backbone width ( $n = 65K$ ) stratified by  $L_0$ . **(c)** Cosine distance per layer per width at  $L_0 = 50$ .

**A norm-free reconstruction metric.** NMSE weights per-token errors by  $1/\|h\|^2$ , so it is sensitive to activation-norm anomalies by construction. As a norm-free companion we use the per-token *cosine distance*,

$$\text{cosdist}(h_i, \hat{h}_i) = 1 - \cos \angle(h_i, \hat{h}_i) = 1 - \frac{\langle h_i, \hat{h}_i \rangle}{\|h_i\| \|\hat{h}_i\|}, \quad (14)$$

the unit-sphere angular distance between a token and its SAE reconstruction. Cosine distance depends only on angles and is bounded in  $[0, 2]$ , immune by construction to heavy norm tails; we use it as an independent cross-check against the trimmed-NMSE profiles (Fig. 3 for 9B and Fig. 4 for 2B compare the two metrics across layers and per-checkpoint). At the layerwise profile level the two metrics agree on the depth shape; at the per-checkpoint level cosine distance and NMSE rank checkpoints similarly within each layer, confirming that the geometry-scaling signal does not depend on the heavy-norm tail.

**Sensitivity analysis.** The Stage 2 regression results are consistent across reconstruction-metric, sparsity, and geometry-preprocessing choices.

**Reconstruction metric.** Replacing the trimmed-mean NMSE with the median NMSE, the geometric-mean NMSE, or the norm-free cosine distance leaves the layerwise NMSE profile and the Stage 2 LOO  $R^2$  values unchanged within  $\pm 0.03$ . The plain-mean NMSE, which is not trimmed, is unstable at tail-contaminated checkpoints and is the only aggregation that diverges from the trimmed alternatives. The agreement between the trimmed NMSE and the norm-free cosine distance confirms that the geometry-scaling signal is a property of the reconstruction error itself, not of the aggregation.

**Sparsity definition.** Replacing the trimmed-mean  $L_0$  used throughout the paper with the median  $L_0$  or the nominal training target leaves the Stage 2 LOO  $R^2$  values unchanged within  $\pm 0.02$  for both models. The plain-mean  $L_0$  destabilises the regression for the same heavy-tail reason that destabilises the plain-mean NMSE.

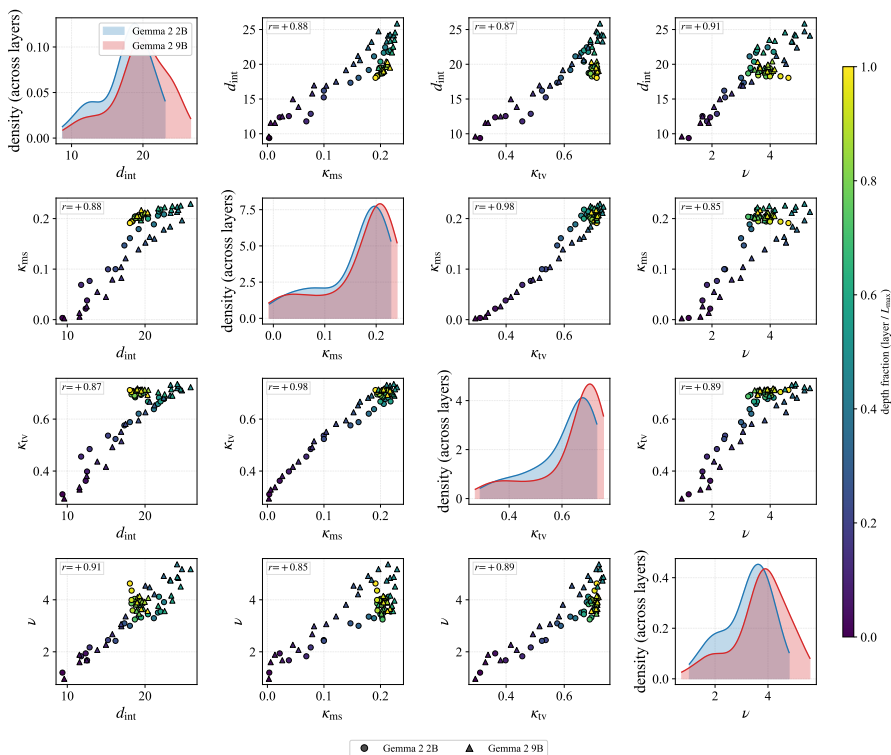


Figure 5: Pairwise distribution of the four geometric features used in the regression analysis. Diagonal: per-feature marginal KDEs (2B and 9B overlaid). Off-diagonal: pairwise scatter coloured by depth fraction (layer /  $L_{\max}$ ); 2B layers shown as circles, 9B as triangles. Per-panel Pearson  $r$  is annotated.

*Geometry preprocessing.* Under norm-trimmed activations, the profiles of  $d_{\text{int}}$ ,  $\kappa_{\text{ms}}$ , and  $\kappa^{\text{tv}}$  are nearly identical to their untrimmed counterparts. The feature most affected by trimming is heterogeneity  $\nu$ , which changes appreciably at early layers; the other three are stable.

**Feature transformation pipeline.** All four geometric summaries enter the Stage 2 regression after the same three-step pipeline: a symmetric percentile clip at the 1st and 99th percentiles (to limit the leverage of extreme values), a log transform (to linearise the relationship with the per-layer scaling-law parameters), and a per-feature standardisation to mean zero and unit variance. Figure 5 shows the joint distribution of the four geometric features in their raw (pre-standardisation) units across all 68 layers of the two models, with the diagonal panels showing per-feature marginal kernel-density estimates (KDEs) and the off-diagonal panels showing pairwise scatters coloured by depth fraction; per-panel Pearson correlations are annotated.

## C No-Floor vs With-Floor Surface Fits

The Stage 1 no-floor surface (Eq. 6) fitted at every layer and the with-floor surface (Eq. 2) refit at the showcase layers are two parametric models of the same data with different identifiability requirements. This appendix details both fits, justifies the no-floor choice for all-layer Stage 1, and quantifies the relationship between the two at the layers where both are fitted.

**Why no-floor at every layer.** The released SAE family supplies only two backbone widths shared across all 68 layers (§4). The with-floor surface has six free parameters ( $a_0, \beta_n, \beta_k, \gamma, \zeta, \eta$ ) in a 4-parameter no-floor surface plus two more for the floor; with two widths per layer the with-floor parameters are not identifiable per layer. We therefore fit the no-floor 4-parameter form  $\log L = a_0 + \beta_n \log n + \beta_k \log k + \gamma \log n \log k$  at every layer, identifiable from  $(n_{\text{lo}}, n_{\text{hi}}) \times \{\text{realised } L_0\}$ .

Per-token NMSE distribution (full, untrimmed) at six representative (layer, width,  $L_0$ ) cells — heavy upper tail motivates the trimmed-mean NMSE (corpus = C4)

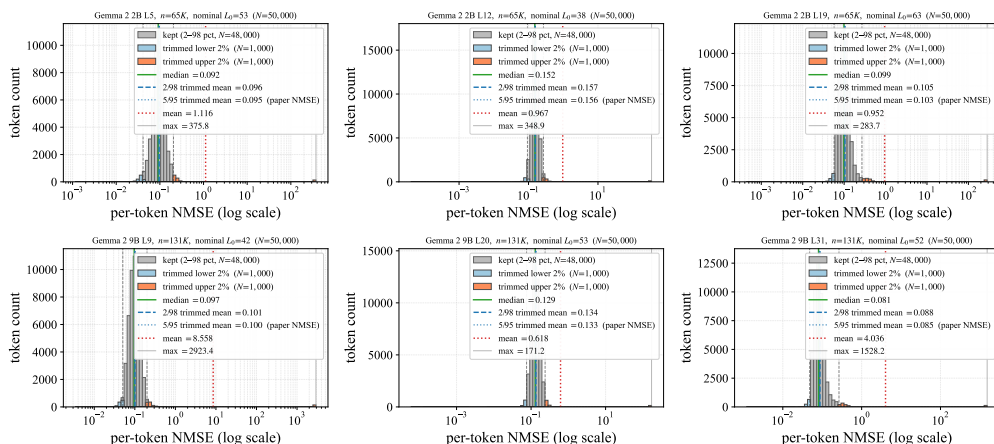


Figure 6: Per-token NMSE distribution at six representative (model, layer, width, sparsity) cells. The full untrimmed distribution is shown; the heavy upper tail motivates the trimmed mean (5/95 percentile mask) used as the canonical paper NMSE. Each panel annotates three reductions on the full per-token data: the median (robust), the trimmed mean (canonical), and the untrimmed mean (dominated by the upper tail).

This mirrors the methodological progression of Gao et al. [1], who fit a no-floor power law first and only added the asymptotic floor when richer width coverage at the depth-5/6 layer made it identifiable.

**PCHIP interpolation on the sparsity axis.** At each (layer, width) the released checkpoints span a range of realised  $L_0$  values around their training targets. We interpolate  $\log L$  as a function of  $\log L_0$  via PCHIP (`scipy.interpolate.PchipInterpolator`) and evaluate at common sparsity targets to give the per-(layer, width,  $k$ ) values that enter the per-layer surface fit. PCHIP is shape-preserving and contains no tunable hyperparameters. Its leave-one- $L_0$ -out cross-validation is reported in Appendix G: median  $R^2 > 0.99$  across all (layer, width) cells in both models, confirming the interpolation contributes negligible noise to the downstream surface fits.

**With-floor joint surface fits at the showcase layers.** At the six showcase layers per model with  $\geq 3$  widths after monotone filtering (Gemma 2 2B {5, 12, 19} and 9B {9, 20, 31}), we refit the full 6-parameter with-floor surface (Eq. 2) jointly across all available dictionary widths at each showcase layer (up to seven widths per layer in the released family, including the two backbone widths) and a shared 9-point sparsity grid  $k \in \{16, 24, 32, 48, 64, 96, 128, 160, 192\}$  that spans the range reported by Gao et al. [1]. Per (layer, width) we PCHIP-interpolate  $\log L$  onto this grid; three monotone filters then drop unreliable points before fitting: an  $L$ -monotone-prefix filter at each  $k$  (drops widths whose loss rises with  $n$ ), a local- $\alpha$  monotone filter at each  $k$  (drops widths whose per- $k$  exponent is implausible or speeds up with  $n$ ), and a cross- $k$   $\alpha$ -monotone filter per layer (drops  $k$  values whose per- $k$  exponent is non-monotone in  $k$ ). On the surviving  $(n, k, \log L)$  points we minimise  $\sum (\log L - \log \hat{L})^2$  jointly over all six parameters  $(a_0, \beta_n, \beta_k, \gamma, \zeta, \eta)$  via Nelder-Mead, restarting from eight seeds spanning the physically plausible region (one of which is Gao et al. [1]’s reported parameter values) to mitigate local minima, and retaining the solution with the lowest residual sum of squares (RSS). The per-layer asymptotic floor reported in the main text is read off the fitted surface as  $B_\ell(k) = \exp(\zeta_\ell + \eta_\ell \log k)$  evaluated at  $k = 50$ .

**Monotone-filter rationale.** The three monotone filters above are direct consequences of the scaling-law family being fitted, not ad hoc data cleaning. The with-floor form  $L(n, k) = A(k)n^{-\alpha(k)} + B(k)$  with  $A(k), \alpha(k), B(k) > 0$  implies  $\partial L / \partial n \leq 0$  at any fixed  $k$ : increasing dictionary width at matched sparsity cannot increase loss. Any (layer, width,  $k$ ) checkpoint where the PCHIP-interpolated loss rises with  $n$  at fixed  $k$  is therefore inconsistent with every member of the scaling-law family of Eq. 2, regardless of parameter values. Smoothness of  $\alpha(k)$  in  $k$  is the corresponding generic

property of any law in this family. The L-monotone-prefix filter enforces the first property at each  $k$ ; the local- $\alpha$  monotone filter enforces the second within each  $k$ ; the cross- $k$   $\alpha$ -monotone filter enforces it across  $k$ . The filters are applied uniformly across all 68 layers of both models with no layer-specific tuning. Across the six showcase layers, 193 of 212 (layer, width,  $k$ ) cells survive all three filters (a 91% retention rate), with 9 cells dropped by the L-monotone-prefix and 10 further by the local- $\alpha$  filter; the cross- $k$   $\alpha$ -monotone filter drops 3  $k$  values out of 49 candidates across the six layers. Drops concentrate at the two largest released dictionary widths (524K and 1M), consistent with the width-dependent training schedule of the released Gemma Scope family (Appendix H), where the largest dictionary widths receive sublinear-in-width training budgets and sit furthest from convergence.

**Relating the two fits at showcase layers.** At the six showcase layers both fits are available. The no-floor exponent  $\alpha_\ell^{\text{nf}}(k) = -(\beta_{n,\ell} + \gamma_\ell \log k)$  from the per-layer Stage 1 fit underestimates the with-floor reducible exponent  $\alpha_\ell^{\text{wf}}(k)$  because part of the loss attributed to the power-law term in the no-floor fit is in reality the asymptotic floor in the with-floor fit. The relation between them is the floor-attenuation identity

$$\alpha_\ell^{\text{nf}}(k) = \alpha_\ell^{\text{wf}}(k) - \frac{1}{\log(n_{\text{hi}}/n_{\text{lo}})} \log \frac{1 + B_\ell(k)/R_\ell(n_{\text{hi}}, k)}{1 + B_\ell(k)/R_\ell(n_{\text{lo}}, k)}, \quad (15)$$

where  $R_\ell(n, k) = A_\ell(k) n^{-\alpha_\ell^{\text{wf}}(k)}$  is the reducible part of the loss at width  $n$ . The attenuation  $\delta_\ell(k) = \alpha_\ell^{\text{wf}}(k) - \alpha_\ell^{\text{nf}}(k)$  is positive at  $k = 50$  and computable at the showcase layers from the with-floor fit, consistent with the residual gap between the no-floor profiles and the OpenAI  $\alpha = 0.181$  in Fig. 2(a). The two factors that determine  $\delta$  in Eq. 15 (the floor magnitude  $B_\ell$  and its ratio to the reducible loss  $R_\ell$  at the backbone widths) are themselves geometrically determined: the showcase floor  $B_\ell$  ranks consistently with  $d_{\text{int}}$  and  $\kappa_{\text{ms}}$  at every showcase layer in both models (Fig. 2(b) and Appendix E), so layers with higher curvature and intrinsic dimension simultaneously have higher  $B_\ell$  and larger  $\delta_\ell$ . The geometric pattern of  $\alpha^{\text{nf}}$  therefore reflects two coupled channels: a direct dependence of  $\alpha^{\text{wf}}$  on geometry (more curvature  $\rightarrow$  smaller asymptotic exponent) and an indirect dependence through the floor-attenuation  $\delta$  (more curvature  $\rightarrow$  higher floor  $\rightarrow$  larger  $\delta$ ). Both channels push  $\alpha^{\text{nf}}$  in the same direction, so the Stage 2 regressions on  $\alpha^{\text{nf}}$  in Tables 1–3 detect a coherent geometric signal that is consistent with the geometric structure of the asymptotic with-floor exponent.

**Multi- $k$  visualisations of the showcase fits.** Figure 7 overlays the joint 6-parameter surface fits (sliced at three additional sparsity targets  $k \in \{32, 64, 75\}$  beyond the  $k = 50$  shown in Fig. 2(b)) at each showcase layer. The same per-layer fitted surface is sliced at each  $k$  to give  $\alpha_\ell(k)$ ,  $A_\ell(k)$ ,  $B_\ell(k)$ . The vertical ordering of the  $B_\ell(k)$  values is preserved across  $k$  within each model, confirming that the floor-geometry ranking is not specific to  $k = 50$ .

**Heteroscedasticity and form sensitivity.** Visual inspection of residual-vs-fitted plots from the with-floor fits shows no obvious systematic pattern in  $\log n$ ,  $\log k$ , or geometric features at any showcase layer. The with-floor form is compared via AIC against three nearby parametric families at each showcase layer: a two-power form  $L = A n^{-\alpha} + C n^{-\beta}$ , a soft-knee transition  $L = A n^{-\alpha} + B/(1 + (n/n_*)^\delta)$ , and a quadratic-in- $\log k$  cross coupling. The with-floor form is competitive at every showcase layer and AIC-preferred at five of six.

## D Secant Approach to the Per-Layer Width Exponent

The Stage 1 no-floor surface fit (Eq. 6) used in the main text is one of two natural per-layer estimators of the width-scaling exponent computable from the released SAE family. The other is the two-point backbone-secant: a non-parametric chord slope between the layer’s two backbone widths,

$$\alpha_\ell^{\text{sec}}(k) = - \frac{\widehat{\log L}_\ell(n_{\text{hi}}, k) - \widehat{\log L}_\ell(n_{\text{lo}}, k)}{\log n_{\text{hi}} - \log n_{\text{lo}}}, \quad (16)$$

where  $\widehat{\log L}_\ell(n, k)$  is the per-(layer, width) PCHIP interpolant of  $\log L$  at sparsity target  $k$  (the wide hat covers  $\log L$  to make explicit that the interpolant is on the log scale). This appendix establishes that the secant is essentially equivalent to the Stage 1 no-floor exponent and that the cross-layer geometric findings of Section 5 hold under either estimator.

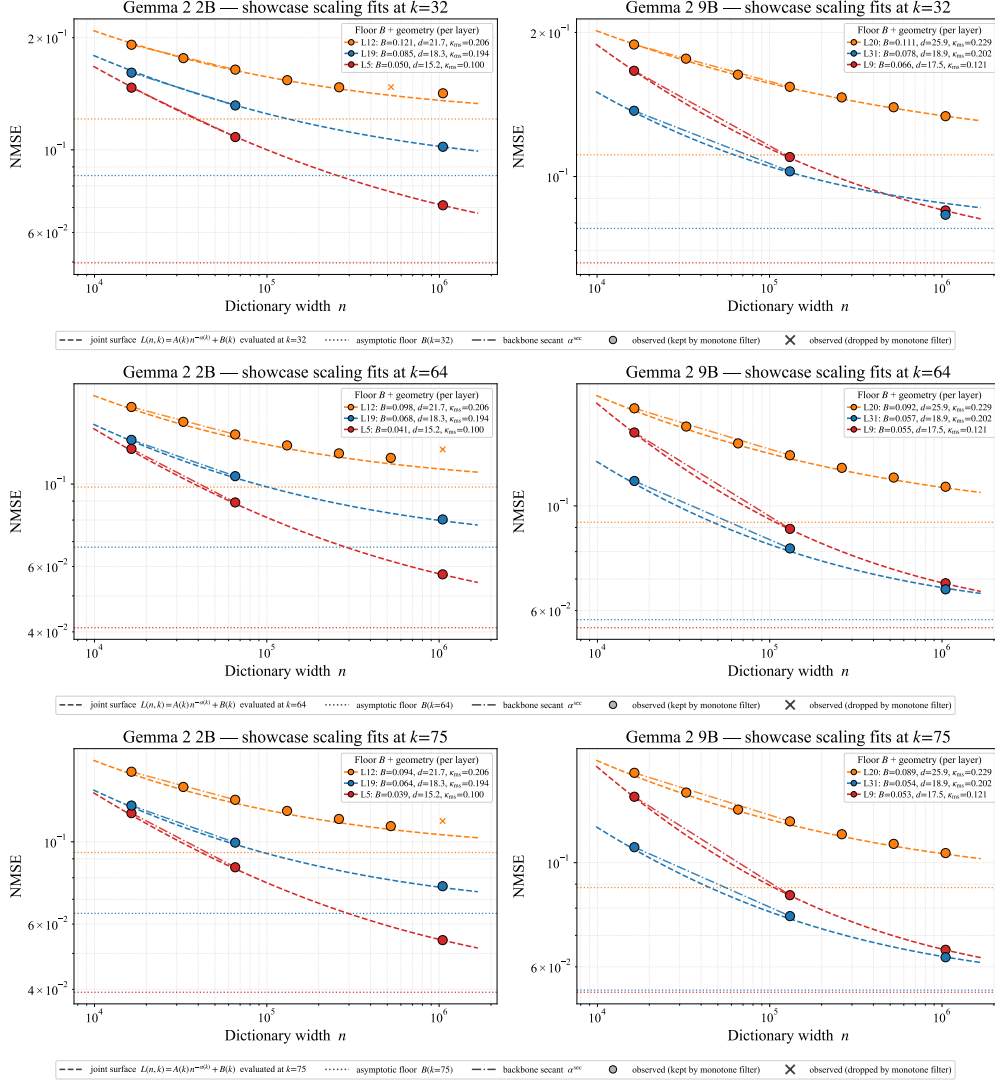


Figure 7: Showcase-layer with-floor surface fits sliced at three sparsity targets  $k \in \{32, 64, 75\}$  (top to bottom). Each panel mirrors Fig. 2(b): dashed coloured lines are the joint surface fits at the indicated  $k$ , dotted horizontal lines are the per-layer floor  $B_\ell(k)$  at that  $k$ , filled circles are the PCHIP-interpolated observed data points kept by the monotone filters, X markers are widths dropped by the filters, and the dash-dot chord between the two backbone widths is the non-parametric backbone-secant slope  $\alpha_\ell^{\text{sec}}(k)$  at the indicated  $k$  (Appendix D). The floor-geometry ranking established at  $k = 50$  persists at all three additional  $k$  values.

**Equivalence in the no-floor regime.** For a no-floor surface  $L(n, k) = A(k) n^{-\alpha(k)}$  at fixed  $k$ , the chord slope between any two widths recovers the exponent exactly:  $\alpha^{\text{sec}} = \alpha$ . The Stage 1 no-floor surface fit constrains the same exponent jointly across  $k$  via the linear-in- $\log n \cdot \log k$  interaction; at fixed  $k$  both estimators recover the same per-layer slope to within numerical precision. Empirically the per-layer Stage 1 exponent at  $k = 50$  and the per-layer backbone secant at  $k = 50$  agree to within 0.01 at all  $26 + 42 = 68$  layers of both models; both are estimates of the same chord slope on  $\log L$  vs  $\log n$  between the same two widths.

**Behaviour in the with-floor regime.** When a non-zero floor  $B$  is present, both estimators underestimate the asymptotic with-floor reducible exponent  $\alpha^{\text{wf}}$  by the floor-attenuation identity (Eq. 15). At showcase layers where with-floor fits are identifiable the attenuation  $\alpha^{\text{wf}} - \alpha^{\text{sec}}$  at  $k = 50$  is positive. By Eq. 15 the attenuation is determined by the floor magnitude  $B_\ell$  and its ratio to the reducible loss

at the backbone widths; both factors are geometrically determined (Appendix E:  $B_\ell$  ranks with  $d_{\text{int}}$  and  $\kappa_{\text{ms}}$  at every showcase layer). The geometric features that predict the secant exponent  $\alpha^{\text{sec}}$  at every layer therefore also predict the floor magnitude and the attenuation at the showcase layers; the geometric regression on  $\alpha^{\text{sec}}$  (or equivalently the Stage 1 no-floor exponent) is consistent with the geometric structure of the asymptotic  $\alpha^{\text{wf}}$ , with the geometry-dependent attenuation contributing in the same direction as the direct geometric effect on  $\alpha^{\text{wf}}$  (more curvature  $\rightarrow$  both smaller  $\alpha^{\text{wf}}$  and larger attenuation, both lowering the no-floor estimate).

**Same conclusions under the secant estimator.** Replacing the Stage 1 exponent with  $\alpha^{\text{sec}}$  as the Stage 2 regression target at  $k = 50$  reproduces every conclusion of §5: the same single-feature ranking with  $\kappa_{\text{ms}}$  dominant, the same hypothesis hierarchy in LOO  $R^2$  within  $\pm 0.01$ , the same cross-model transfer with  $R^2 > 0.92$  for  $\alpha^{\text{sec}}(k = 50)$  in both directions, and the same intercept-vs-slope decomposition. The geometric findings are not artefacts of the parametric Stage 1 surface fit; they are properties of the per-layer scaling behaviour itself, recoverable equally from a non-parametric chord slope.

## E Floor Analysis Details

This appendix details the with-floor calibration at the showcase layers and expands on the mechanistic reading of the geometric wall in §5.

**Per-layer with-floor joint surface fits.** At the showcase layers (Gemma 2 2B {5, 12, 19}; 9B {9, 20, 31}), the full 6-parameter with-floor surface  $L(n, k) = A(k)n^{-\alpha(k)} + B(k)$  (Eq. 2) is fit by joint nonlinear least squares on  $\log L$  across all retained  $(n, k, L)$  points after the monotone-filtering pipeline of Appendix C (PCHIP interpolation onto the shared 9-point sparsity grid;  $L$ -monotone-prefix, local- $\alpha$  monotone, and cross- $k$   $\alpha$ -monotone filters). Multi-start Nelder-Mead with eight seeds (one being Gao et al. [1]’s reported parameter values) is used to mitigate local minima, and the lowest-RSS solution is retained. The fitted asymptotic floor at sparsity  $k$  is  $B_\ell(k) = \exp(\zeta_\ell + \eta_\ell \log k)$ , reported at  $k = 50$  throughout.

**Per-layer floor magnitudes and ranking.** The fitted  $B_\ell(k = 50)$  values within each model rank consistently with the orderings of the two dominant geometric features, intrinsic dimension  $d_{\text{int}}$  and multi-scale curvature  $\kappa_{\text{ms}}$  (the same two features annotated alongside  $B$  in the per-panel legends of Fig. 2(b)). Higher  $d_{\text{int}}$  and higher  $\kappa_{\text{ms}}$  correspond to higher floor at every showcase layer in both models.

**Mechanistic reading.** A sparse linear dictionary approximates each activation  $h$  by a sparse combination of dictionary atoms; the best linear local approximation of the activation manifold  $\mathcal{H}_\ell$  at any point is the tangent plane at that point. When  $\mathcal{H}_\ell$  has zero principal curvatures (a flat manifold), the tangent plane is exact, and a sufficiently rich dictionary can in principle reduce reconstruction error to zero. When  $\mathcal{H}_\ell$  has non-zero principal curvatures, the tangent plane departs from the manifold by a second-order term proportional to (squared local-distance)  $\times$  (principal curvature); even with infinite atoms, that second-order departure is invisible to any sparse linear combination, leaving an irreducible residual. Curvature  $\kappa$  thus contributes a second-order residual at fixed sample density; higher intrinsic dimension  $d_{\text{int}}$  further reduces the share of the local tangent space that a finite atom budget can cover. Both effects predict a higher asymptotic floor at higher-curvature, higher-dimension layers. The empirical ranking of  $B_\ell$  across showcase layers matches this prediction in both models.

**Comparison with Gao et al.’s spectrum-of-structure hypothesis.** Gao et al. [1] report an irreducible loss term in their scaling-law fits and offer the tentative hypothesis (their Appendix G) that “the activations are made of a spectrum of components with different amount of structure” and that “less structured data [has] a worse scaling exponent”. In the most extreme case, they hypothesise that “some amount of the activations could be completely unstructured Gaussian noise”. They support this with a synthetic experiment in which SAEs trained on 768-dimensional Gaussian noise yield an  $L(N)$  exponent of  $-0.04$ , much shallower than the  $\approx -0.26$  they observe on GPT-2-small activations of comparable dimensionality. Their reading attributes the floor to the noise-like (less structured) share of activations: structured components yield steep scaling, but unstructured components remain hard to compress by any sparse linear code.

Our framework decomposes the floor into two distinct geometric channels. The first (“ $d_{\text{int}}$  channel”) captures Gao’s reading: high intrinsic dimension shrinks the share of the local tangent space that any finite atom budget can cover, raising the floor. Pure Gaussian noise sits at the limit of this channel: it is globally flat ( $\kappa = 0$ ) but has full intrinsic dimension ( $d_{\text{int}} = d_{\text{ambient}}$ ), and our framework predicts the very high floor on Gaussian noise (and the corresponding shallow scaling exponent Gao observes) entirely through the  $d_{\text{int}}$  channel. The second (“ $\kappa$  channel”) is genuinely additional: even on a low- $d_{\text{int}}$  smooth manifold, when principal curvatures are non-zero the tangent plane at each atom departs from the manifold by a second-order term proportional to (local curvature)  $\times$  (squared local distance). At fixed sparsity  $k$  in the small- $k$  regime to which the scaling law of Gao et al. [1] is restricted (their Section 3.1.4), a  $k$ -sparse linear combination spans a  $k$ -dimensional flat affine subspace; even when atoms become dense on the manifold as  $n \rightarrow \infty$ , the curved manifold deviates from this flat span by an irreducible second-order term proportional to (curvature)  $\times$  (squared neighbourhood scale), so the geometric mismatch contributes a residual that does not vanish as  $n \rightarrow \infty$  in this regime. The floor  $B(k)$  in our scaling law is itself a function of  $k$  (decreasing in  $k$  as expected: more active atoms allow the reconstruction to fit higher-order geometry), but at each fixed  $k$  in the studied range, the curvature mismatch sets a lower bound on  $B(k)$ . The  $\kappa$  channel is invisible to the noise-like-content reading: it would predict a non-zero floor on a fully structured but curved low-dimensional manifold (where Gao’s mechanism would predict no floor). Real language-model activations have both low intrinsic dimension and non-zero curvature; our showcase-layer floor magnitudes track both  $d_{\text{int}}$  and  $\kappa_{\text{ms}}$ , suggesting both channels operate on the activation geometries that SAEs actually face.

## F Per-Model Regression Suite

This appendix collects the per-model regression results that complement the main-text Tables 1–3 (which report Gemma 2 9B only). It contains: the corresponding 2B versions of those tables (Tables 2 and 4); the cross-model transfer table (Table 5); the multi- $k$  stability of the main-text  $\alpha(k)$  findings; an alternative regression of  $\log L$  directly across all checkpoints per model; and the joint long-format  $\alpha$ -curve regression. The 2B phenomenology figure (Fig. 8), referenced from §5, is also included.

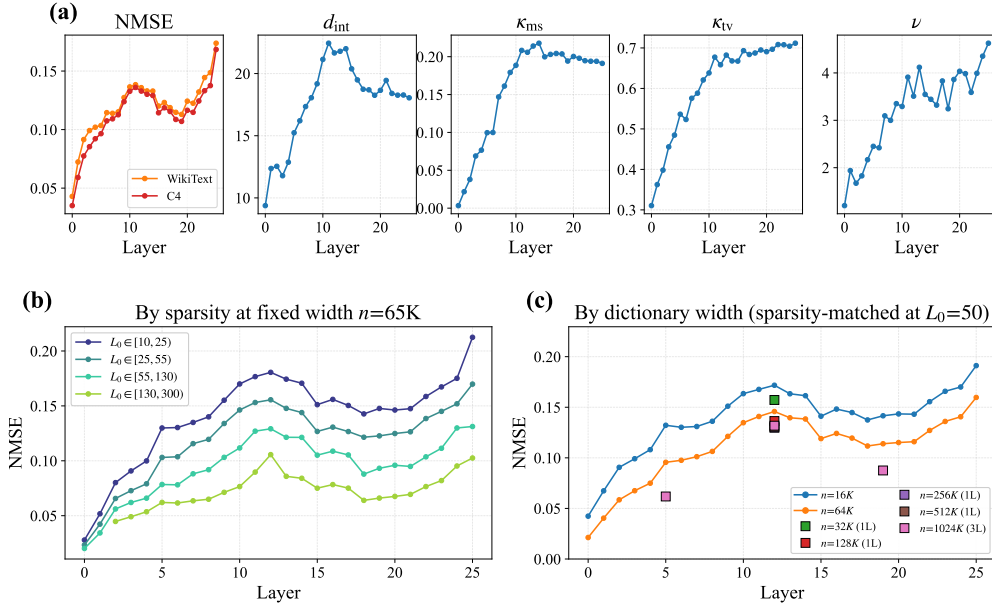


Figure 8: Phenomenology of layerwise SAE reconstruction for Gemma 2 2B, matching Fig. 1 in the main text. **(a)** Layerwise checkpoint-level NMSE profile (evaluated on C4 and WikiText-103) alongside the four geometric summaries. **(b)** NMSE at the upper backbone width ( $n = 65K$ ) stratified by  $L_0$  bin (bin edges  $\{10, 25, 55, 130, 300\}$ ). **(c)** NMSE per layer per dictionary width, sparsity-matched at  $L_0 = 50$  via PCHIP. The depth profile and stratification behaviour mirror those of 9B.

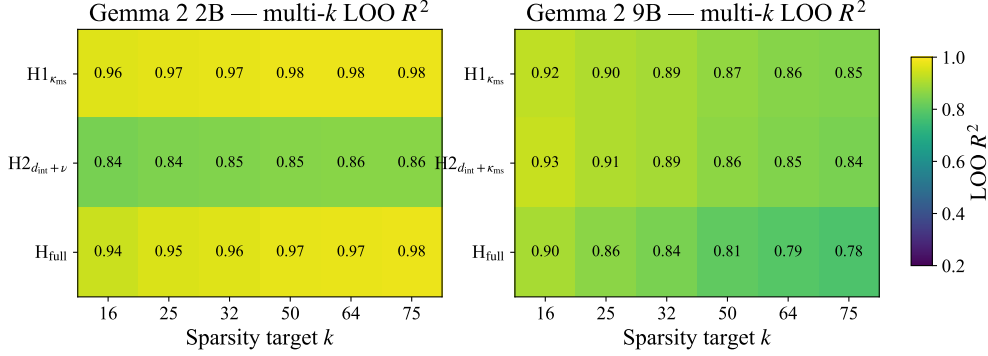


Figure 9: Multi- $k$  stability of the Stage 2 regression. Two side-by-side heatmaps (2B left, 9B right) of LOO  $R^2$  across rows  $H1_{\kappa_{ms}}$ , the per-table  $H2_{low\rho}$  (lowest- $|\rho|$  pair), and  $H_{full}$ , and across columns  $k \in \{16, 25, 32, 50, 64, 75\}$ . The  $H1_{\kappa_{ms}}$  row stays high and flat at 2B and decreases monotonically with  $k$  at 9B;  $H_{full}$  tracks  $H1_{\kappa_{ms}}$  closely at 2B (well above  $H2_{low\rho}$  there) but at 9B sits a constant amount below both  $H1$  and  $H2_{low\rho}$  (reflecting collinearity among the four 9B geometric features). The  $k = 50$  column corresponds to Tables 1 and 2.

**Multi- $k$  stability of the per-layer  $\alpha_\ell(k)$  regression.** Tables 1 and 2 report  $k = 50$ . Repeating the Stage 2 regression at  $k \in \{16, 25, 32, 64, 75\}$  yields the same single-feature ranking ( $\kappa_{ms}$  dominant) across  $k$ .  $H1_{\kappa_{ms}}$  LOO stays in  $[0.96, 0.98]$  across  $k$  at 2B and decreases monotonically with  $k$  from 0.92 to 0.85 at 9B. The per-table  $H2_{low\rho}$  ( $d_{int} + \nu$  at 2B,  $d_{int} + \kappa_{ms}$  at 9B) stays in  $[0.83, 0.86]$  across  $k$  at 2B and decreases from 0.93 at  $k = 16$  to 0.84 at  $k = 75$  at 9B.  $H_{full}$  stays in  $[0.94, 0.98]$  at 2B (above  $H2_{low\rho}$  because  $H_{full}$  includes  $\kappa_{ms}$ ) and sits 0.04–0.06 below  $H2_{low\rho}$  across  $k$  at 9B, reflecting collinearity among the four geometric features at 9B (the same  $H_{full}$  model fits cleanly on the smaller 26-layer 2B sample, where it stays above  $H2_{low\rho}$  throughout). Figure 9 visualises this stability as a heatmap of LOO  $R^2$  across  $k \in \{16, 25, 32, 50, 64, 75\}$  for the three published hypotheses.

**Per-checkpoint regression on  $\log L$  directly.** As an alternative to the Stage 2 regression on per-layer  $\alpha_\ell(k)$ , one can regress  $\log L$  on geometry directly across all 312 checkpoints (2B) or 532 checkpoints (9B) per model. The design augments the scaling-law base  $[1, \log n, \log k, \log n \log k]$  with, for each geometry feature  $g$ , a 4-term interaction block  $[g, g \cdot \log n, g \cdot \log k, g \cdot \log n \log k]$ . The same hypothesis hierarchy gives LOO of 0.76–0.89 across the four single features and 0.90 at  $H_{full}$  for 2B, and 0.72–0.85 across single features and 0.93 at  $H_{full}$  for 9B, confirming that the geometric signal is present in the raw checkpoint-level loss as well as in the per-layer Stage 1 exponent.

**Joint long-format  $\alpha$ -curve regression.** A single regression of  $y = \alpha_\ell(k)$  on  $(\log k, g, g \cdot \log k)$  across the long-format  $(\ell, k)$  grid, with one set of geometric coefficients controlling both the level (via  $g$ ) and the tilt (via  $g \cdot \log k$ ), yields LOO of 0.97 at 2B and 0.80 at 9B for  $H_{full}$ . Decomposing the  $H1_g$  result into “intercept-only” ( $g$  enters as  $\beta_n$  modifier), “ $\log k$ -only” ( $g$  enters as  $\gamma$  modifier), and “both” shows that the two single-entry models are within  $\sim 0.005$  LOO of each other for every feature, and the “both” model adds only another  $\sim 0.005$ . This is the long-format equivalent of the Tables 1–3 reading: geometry shifts the entire per-layer  $\alpha_\ell(k)$  curve as a unit; it does not selectively modify level or tilt.

**Visualisation: geometry vs scaling parameters.** Figure 10 scatters the per-layer  $\alpha(k = 50)$  (top row, all  $26 + 42 = 68$  layers) and per-layer  $B(k = 50)$  (bottom row, 6 showcase layers) against each of the four geometric features, with both models overlaid. The top row visualises the cross-model transferability claim: layers from 2B and 9B fall on a common trend in each panel, with  $\kappa_{ms}$  giving the cleanest separation. The bottom row visualises the floor-geometry coupling at the showcase layers, with the same per-feature ordering as the  $\alpha(k = 50)$  regression.

**Per-target tables for Gemma 2 2B and cross-model transfer.** Tables 2 and 4 reproduce the main-text Tables 1 and 3 for Gemma 2 2B; Table 5 reports cross-model transfer of the per-layer width-scaling exponent  $\alpha(k = 50)$  between the two models.

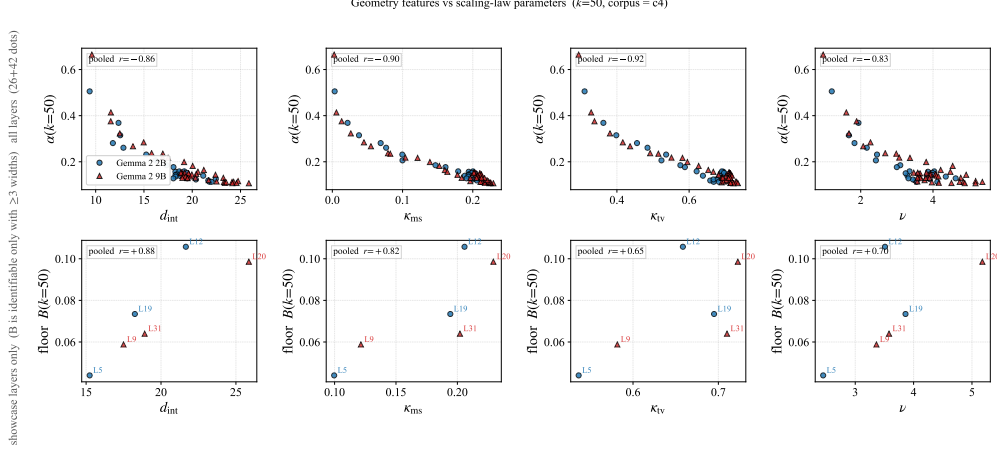


Figure 10: Geometry features vs scaling-law parameters at  $k = 50$ . **Top row**: per-layer  $\alpha(k = 50)$  (all 68 layers; 2B as circles, 9B as triangles) vs each of  $\{d_{\text{int}}, \kappa_{\text{ms}}, \kappa^{\text{tv}}, \nu\}$ . Per-panel Pearson  $r$  annotated. The two models overlay on a common per-feature trend, particularly clean for  $\kappa_{\text{ms}}$ , the visual signature of the transferable geometric law for  $\alpha(k = 50)$ . **Bottom row**: per-layer  $B(k = 50)$  (showcase layers only, 3 per model) vs the same four features. Higher  $d_{\text{int}}$  and  $\kappa_{\text{ms}}$  correspond to higher floor; the  $B$ - $d_{\text{int}}$  and  $B$ - $\kappa_{\text{ms}}$  rankings agree at every showcase layer in both models.

Table 2: Stage-2 cross-layer regression of per-layer width-scaling exponent  $\alpha_\ell(k = 50)$  on geometry, Gemma 2 2B ( $n = 26$  layers). Same hypothesis hierarchy and metrics as Table 1 in the main text. The H1 ranking is the same as in 9B with  $\kappa_{\text{ms}}$  dominant;  $\text{LOO}_{\text{Hfull}}$  is higher in 2B due to the smaller and more homogeneous layer set.

Hypothesis	$R^2$	LOO	L2O	L3O	AIC	BIC	$F$ vs H0	$p$
H0	0.000	-0.082	-0.083	-0.085	-123	-122	—	—
H1 $_{d_{\text{int}}}$	0.880	+0.816	+0.814	+0.813	-176	-173	176.3	1.5e-12
H1 $_{\kappa_{\text{ms}}}$	0.979	+0.976	+0.975	+0.975	-222	-219	1130.4	$< 10^{-10}$
H1 $_{\kappa^{\text{tv}}}$	0.923	+0.880	+0.879	+0.878	-187	-185	286.0	7.8e-15
H1 $_{\nu}$	0.846	+0.776	+0.775	+0.774	-169	-167	131.5	3.2e-11
H2 $_{d_{\text{int}} + \nu}$	0.912	+0.855	+0.854	+0.853	-182	-178	119.0	7.4e-13
H $_{\text{full}}$	0.989	+0.971	+0.970	+0.969	-231	-225	452.7	$< 10^{-10}$

**Cross-model standardisation procedure.** The cross-model transfer fit (Table 5) is constructed so that no test-model statistic enters the training pipeline. Let  $\boldsymbol{\mu}^{\text{tr}}, \boldsymbol{\sigma}^{\text{tr}} \in \mathbb{R}^4$  denote the per-feature mean and standard deviation of the four layerwise geometric summaries on the training model only (computed after the same log transform and 1/99 percentile clip used in the within-model fits, Appendix B). The Stage 2 regression coefficients (intercept  $\hat{\beta}_0^{\text{tr}}$  and slope vector  $\hat{\boldsymbol{\beta}}^{\text{tr}}$ ) are estimated by OLS on the training model’s standardised geometry  $Z_\ell^{\text{tr}} = (g_\ell^{\text{tr}} - \boldsymbol{\mu}^{\text{tr}})/\boldsymbol{\sigma}^{\text{tr}}$ . To predict the test model’s per-layer  $\alpha(k = 50)$ , the same  $\boldsymbol{\mu}^{\text{tr}}$  and  $\boldsymbol{\sigma}^{\text{tr}}$  (not the test model’s own moments) are applied to the test geometry:  $Z_\ell^{\text{te}} = (g_\ell^{\text{te}} - \boldsymbol{\mu}^{\text{tr}})/\boldsymbol{\sigma}^{\text{tr}}$ , and the prediction is  $\hat{y}_\ell^{\text{te}} = \hat{\beta}_0^{\text{tr}} + Z_\ell^{\text{te}} \hat{\boldsymbol{\beta}}^{\text{tr}}$ , where  $\hat{\beta}_0^{\text{tr}} = \bar{y}^{\text{tr}}$  (the training-model target mean) is the standard OLS identity for mean-centred features. The transfer  $R^2$  in Table 5 compares  $\hat{y}_\ell^{\text{te}}$  to the test model’s empirical per-layer  $\alpha(k = 50)$ . The test model’s in-sample  $R^2$  in the same table is the optimal OLS fit on the test set with the matching regressor structure, and is the mathematically correct upper bound for any fixed-coefficient prediction on the full test set: by construction, Transfer  $R^2 \leq$  Test In-sample  $R^2$  for any choice of fixed coefficients (the in-sample fit minimises the same residual sum of squares on the same data). Within-test leave-one-layer-out  $R^2$ , by contrast, penalises the test fit through cross-validation and is therefore typically below the in-sample upper bound; we do not use it as the upper-bound benchmark in Table 5 because comparing an unpenalised quantity (Transfer  $R^2$  on the full test set) to a penalised one (LOO on the test set) is not a like-for-like comparison.

Table 3: Stage-2 regression of the per-layer Stage-1 raw scaling-law coefficients  $\beta_n$  (width-scaling intercept; *top*) and  $\gamma$  (width  $\times$  log-sparsity interaction; *bottom*) on geometry, Gemma 2 9B ( $n = 42$  layers). Same hypothesis hierarchy as Table 1. Together these decompose where geometry enters the per-layer  $\alpha(k) = -(\beta_n + \gamma \log k)$  curve:  $\beta_n$  is the level,  $\gamma$  is the tilt with  $\log k$ . The corresponding 2B tables are below.

Hypothesis	$R^2$	LOO	L2O	L3O	AIC	BIC	$F$ vs H0	$p$
<b>Target: <math>\beta_n</math> (Stage-1 width-scaling coefficient)</b>								
H0	0.000	-0.049	-0.050	-0.051	-257	-256	—	—
H1 $_{d_{\text{int}}}$	0.751	+0.662	+0.661	+0.660	-314	-310	120.5	1.2e-13
H1 $_{\kappa_{\text{ms}}}$	0.684	+0.477	+0.472	+0.467	-304	-300	86.6	1.5e-11
H1 $_{\kappa^{\text{tv}}}$	0.784	+0.690	+0.689	+0.688	-320	-316	145.5	6.7e-15
H1 $_{\nu}$	0.663	+0.546	+0.545	+0.544	-301	-298	78.8	5.3e-11
H2 $_{d_{\text{int}} + \kappa_{\text{ms}}}$	0.763	+0.585	+0.579	+0.574	-314	-309	62.7	6.6e-13
H $_{\text{full}}$	0.856	+0.685	+0.666	+0.646	-331	-322	54.8	4.7e-15
<b>Target: <math>\gamma</math> (Stage-1 width <math>\times</math> log-sparsity interaction)</b>								
H0	0.000	-0.049	-0.050	-0.051	-328	-326	—	—
H1 $_{d_{\text{int}}}$	0.498	+0.245	+0.242	+0.239	-355	-351	39.6	1.8e-07
H1 $_{\kappa_{\text{ms}}}$	0.661	+0.354	+0.352	+0.349	-371	-368	78.1	6.0e-11
H1 $_{\kappa^{\text{tv}}}$	0.534	+0.270	+0.267	+0.264	-358	-355	45.9	3.9e-08
H1 $_{\nu}$	0.551	+0.273	+0.271	+0.268	-360	-356	49.1	1.8e-08
H2 $_{d_{\text{int}} + \kappa_{\text{ms}}}$	0.665	+0.322	+0.319	+0.315	-370	-365	38.7	5.5e-10
H $_{\text{full}}$	0.740	+0.229	+0.213	+0.197	-377	-368	26.4	2.2e-10

Table 4: Stage-2 regression of  $\beta_n$  and  $\gamma$  on geometry, Gemma 2 2B ( $n = 26$  layers). Same hypothesis hierarchy and metrics as Table 3 in the main text. As in 9B, both channels are significantly predicted by geometry; the H $_{\text{full}}$  LOO for  $\gamma$  is lower than for  $\beta_n$ , consistent with  $\gamma$  being a smaller, noisier component of the per-layer  $\alpha(k)$  curve.

Hypothesis	$R^2$	LOO	L2O	L3O	AIC	BIC	$F$ vs H0	$p$
<b>Target: <math>\beta_n</math> (Stage-1 width-scaling coefficient)</b>								
H0	0.000	-0.082	-0.083	-0.085	-141	-140	—	—
H1 $_{d_{\text{int}}}$	0.811	+0.675	+0.673	+0.672	-182	-180	103.1	3.6e-10
H1 $_{\kappa_{\text{ms}}}$	0.882	+0.777	+0.778	+0.778	-195	-192	179.1	1.3e-12
H1 $_{\kappa^{\text{tv}}}$	0.778	+0.615	+0.614	+0.612	-178	-176	84.0	2.6e-09
H1 $_{\nu}$	0.737	+0.579	+0.577	+0.575	-174	-171	67.1	2.1e-08
H2 $_{d_{\text{int}} + \nu}$	0.825	+0.684	+0.683	+0.681	-182	-179	54.1	2.0e-09
H $_{\text{full}}$	0.922	+0.624	+0.629	+0.633	-199	-193	61.7	2.6e-11
<b>Target: <math>\gamma</math> (Stage-1 width <math>\times</math> log-sparsity interaction)</b>								
H0	0.000	-0.082	-0.083	-0.085	-243	-242	—	—
H1 $_{d_{\text{int}}}$	0.620	+0.526	+0.523	+0.519	-266	-264	39.2	1.8e-06
H1 $_{\kappa_{\text{ms}}}$	0.723	+0.470	+0.471	+0.472	-274	-272	62.7	3.8e-08
H1 $_{\kappa^{\text{tv}}}$	0.773	+0.680	+0.678	+0.676	-280	-277	81.8	3.4e-09
H1 $_{\nu}$	0.667	+0.586	+0.583	+0.580	-270	-267	48.0	3.7e-07
H2 $_{d_{\text{int}} + \nu}$	0.683	+0.581	+0.577	+0.574	-269	-265	24.7	1.9e-06
H $_{\text{full}}$	0.785	+0.258	+0.261	+0.258	-275	-269	19.1	9.1e-07

## G Validation Suite

This appendix collects the methodological details of the validation tests cited throughout §5: closed-form leave- $K$ -layer-out cross-validation, layer-level permutation, the F-test ladder over the H0  $\rightarrow$  H1  $\rightarrow$  H2 $_{\text{best}+\text{next}} \rightarrow$  H $_{\text{full}}$  hierarchy, and PCHIP leave-one- $L_0$ -out cross-validation.

**Closed-form leave- $K$ -layer-out CV.** For an OLS regression with design  $X$  ( $n$  rows,  $p$  parameters), full-data fit  $\hat{\beta}$ , residuals  $e = y - X\hat{\beta}$ , and hat matrix  $H = X(X^\top X)^{-1}X^\top$ , the leave- $K$ -rows-out

Table 5: Cross-model transfer of the per-layer width-scaling exponent  $\alpha(k = 50)$ . The Stage-2 regression is fit on the *train* model; per-layer geometry of the *test* model is standardised using the training model’s per-feature mean and standard deviation; predictions on the test model’s per-layer  $\alpha(k = 50)$  are scored by transfer  $R^2$ . The mathematically correct upper bound for fixed-coefficient prediction on the full test set is the test model’s in-sample  $\bar{R}^2$  (the optimal OLS fit on the same test data with the same regressor structure), reported here for each row. Cross-model  $R^2$  closely approaches the in-sample upper bound in every cell and in both directions, indicating that the geometric law for  $\alpha(k = 50)$  transfers across the two models. Reported H2 rows use the lowest- $|\rho|$  pair  $d_{\text{int}} + \kappa_{\text{ms}}$  in both directions for symmetry (within-2B  $\text{H2}_{\text{low}\rho}$  is  $d_{\text{int}} + \nu$ ); transfer  $R^2$  is robust to this choice.

Target	Train→Test	Hypothesis	Transfer $R^2$	Test in-sample $R^2$	$\Delta$
$\alpha(k = 50)$	2B→9B	$\text{H1}_{\kappa_{\text{ms}}}$	+0.920	+0.929	-0.009
$\alpha(k = 50)$	2B→9B	$\text{H2}_{d_{\text{int}} + \kappa_{\text{ms}}}$	+0.933	+0.935	-0.002
$\alpha(k = 50)$	2B→9B	$\text{H}_{\text{full}}$	+0.935	+0.940	-0.005
$\alpha(k = 50)$	9B→2B	$\text{H1}_{\kappa_{\text{ms}}}$	+0.970	+0.979	-0.009
$\alpha(k = 50)$	9B→2B	$\text{H2}_{d_{\text{int}} + \kappa_{\text{ms}}}$	+0.985	+0.988	-0.003
$\alpha(k = 50)$	9B→2B	$\text{H}_{\text{full}}$	+0.983	+0.989	-0.006

residual vector for any held-out subset  $T$  of size  $K_{\text{rows}}$  is

$$e_T^{\text{LKO}} = (I_{K_{\text{rows}}} - H_{TT})^{-1} e_T, \quad (17)$$

where  $H_{TT}$  is the  $K_{\text{rows}} \times K_{\text{rows}}$  submatrix of  $H$  restricted to indices in  $T$  (block generalisation via the Sherman–Morrison–Woodbury identity [38] of the standard  $e_i^{\text{LOO}} = e_i / (1 - H_{ii})$ ). For grouped LOO at the layer level,  $T$  runs over rows belonging to held-out layers; for our per-layer regressions  $K_{\text{rows}} = K$ . We exhaustively enumerate all  $\binom{n_{\text{layers}}}{K}$  holdout subsets for  $K \in \{1, 2, 3\}$  ( $n_{\text{layers}} = 26, 42$  in our two models). The pooled-residual  $R^2$  is then  $1 - \text{MSE}_{\text{LKO}} \cdot n / \text{SS}_{\text{tot}}$ , with  $\text{SS}_{\text{tot}}$  computed on the full  $y$  vector (the original definition of `_loo_layer_cv` used throughout). Specialising to the intercept-only H0 model, Eq. 17 reduces in closed form to  $\text{LOO } R^2 = 1 - (n / (n - 1))^2$  (a function of  $n_{\text{layers}}$  alone, independent of the target), giving  $-0.049$  at 9B ( $n = 42$ ) and  $-0.082$  at 2B ( $n = 26$ ); this is the standard finite-sample LOO bias on a constant predictor and explains the identical H0 rows across every reported table.

**Layer-level permutation.** For each of 1,000 permutations we randomly reassign the layer-to-geometry mapping (each layer keeps its target value but receives a randomly chosen layer’s geometry) and recompute the closed-form LOO  $R^2$  for the same hypothesis. The observed LOO is compared against the resulting null distribution; the  $p$ -value is the fraction of null samples  $\geq$  observed. The null distribution lies entirely below zero in every (target, model, hypothesis) cell, a known finite-sample property of LOO  $R^2$  on uninformative regressors at  $n_{\text{layers}} \in \{26, 42\}$ , with stronger negative bias for larger  $p$  (e.g., null mean  $\approx -0.08$  for H1 at 9B vs  $\approx -0.44$  for  $\text{H}_{\text{full}}$  at 2B). The observed LOO lies far above the null in every cell, with  $p \leq 0.01$  for every reported H1,  $\text{H2}_{\text{low}\rho}$ , and  $\text{H}_{\text{full}}$  across all three targets and both models.

**Nested F-test ladder.** For each target we report F-tests for the steps  $\text{H0} \rightarrow \text{H1}_{\text{best}}$ ,  $\text{H1}_{\text{best}} \rightarrow \text{H2}_{\text{best+next}}$ , and  $\text{H2}_{\text{best+next}} \rightarrow \text{H}_{\text{full}}$ , using the standard nested-OLS F statistic. The  $\text{H2}_{\text{best+next}}$  model adds the second-strongest single feature to the best (e.g.,  $\kappa_{\text{ms}} + d_{\text{int}}$  for  $\alpha(k = 50)$  at both 2B and 9B), and is therefore a different two-feature model from the  $\text{H2}_{\text{low}\rho}$  shown in Tables 1–4, which uses the lowest- $|\rho|$  pair within each model. Both are valid additivity checks but answer different questions:  $\text{H2}_{\text{low}\rho}$  asks whether two minimally-collinear features carry independent variance;  $\text{H2}_{\text{best+next}}$  asks whether the next-best feature adds explanatory power to the best. The  $\text{H0} \rightarrow \text{H1}$  step is overwhelmingly significant in every cell ( $p < 10^{-8}$ ; minimum significance is  $p = 3.4 \times 10^{-9}$  for  $\gamma$  at 2B, with all other cells at  $p < 10^{-10}$ ). The  $\text{H1} \rightarrow \text{H2}_{\text{best+next}}$  step is significant for  $\alpha(50)$  at 2B ( $p \approx 5 \times 10^{-4}$ , adding  $d_{\text{int}}$  to  $\kappa_{\text{ms}}$ ); marginal elsewhere. The  $\text{H2}_{\text{best+next}} \rightarrow \text{H}_{\text{full}}$  step is significant for  $\beta_n$  at 9B ( $p \approx 0.003$ ) and  $\gamma$  at 9B ( $p \approx 0.007$ ), adding evidence that the two extra features carry independent variance for those harder per-layer targets; not significant elsewhere.

**PCHIP leave-one- $L_0$ -out CV.** The Stage 1 surface fit relies on PCHIP-interpolated  $L$ -values across the realised  $L_0$  grid at each (layer, width). To assess interpolation faithfulness we hold out one  $L_0$  observation per (layer, width) cell, refit PCHIP through the rest, and compare the held-out predicted

$\log L$  against the actual. Median per-cell  $R^2$  is 0.998 at 2B and 0.9996 at 9B; 93% of 2B cells and 98% of 9B cells have  $R^2 > 0.95$ . PCHIP interpolation contributes negligible noise to the downstream Stage 1 surface fits.

**Cross-corpus check.** Recomputing checkpoint-level NMSE on WikiText-103 while keeping geometry on its C4 partition leaves the cross-layer NMSE profile essentially unchanged: the C4 and WikiText curves of Fig. 1(a) track each other throughout the depth of both models. The layerwise NMSE profile that motivates the Stage 2 analysis is therefore not corpus-specific. A formal Stage 2 refit on WikiText-derived per-layer scaling parameters is left to future work.

## H Training Schedule of the Released SAE Family

The released Gemma Scope JumpReLU SAEs were trained for 4B, 8B, and 16B tokens at dictionary widths  $\leq 16\text{K}$ , 32K–524K, and 1M respectively [7]. This token allocation is sublinear in width: under the convergence-scaled allocation reported by Gao et al. [1] for TopK SAEs on GPT-4 (tokens-to-convergence scales as  $n^{0.65}$ ), holding the 16K SAEs at 4B tokens fixed and applying the same convergence exponent gives a convergence budget at 1M of  $\approx 60\text{B}$  tokens, so the released 16B-token 1M SAEs correspond to roughly a quarter of that budget. Reported scaling-law parameters in this paper therefore describe the released SAE family under its as-trained schedule rather than a convergence-matched width family. The cross-layer geometric prediction is invariant to this caveat (it concerns predictability of layerwise observables from geometry, regardless of training schedule), but the absolute values of fitted scaling exponents are conditional on this schedule.

## I Limitations and Future Directions

**Limitations.** *Pre-trained SAEs and training schedule.* We use publicly released Gemma Scope JumpReLU SAEs [7] and cannot control the SAE training procedure. The released family is trained under a width-dependent token schedule (Appendix H); reported scaling parameters describe the released family’s behaviour under this schedule rather than a matched-budget or convergence-scaled width family in the sense of Gao et al. [1]. The cross-layer geometric prediction is invariant to this caveat (it concerns predictability of layerwise scaling-law parameters from geometry, regardless of training schedule), but absolute fitted exponents are conditional on it. Whether the geometry-scaling relationship holds for SAEs trained with different architectures (e.g. TopK), different training data, or different loss functions (e.g. end-to-end objectives) remains open. *Sparse width grids and showcase calibration.* At most layers only two dictionary widths are available (the backbone), so the all-layer Stage 1 surface fits use the no-floor 4-parameter form (Appendix C) and the with-floor 6-parameter calibration is confined to the six showcase layers with  $\geq 3$  widths. The asymptotic floor  $B_\ell(k)$  is therefore reported only at the showcase layers; a layerwise floor regression across all 68 layers is not feasible with current width coverage. *Residual stream only.* Our analysis covers the residual stream; extension to attention and multilayer-perceptron (MLP) sublayers may reveal site-specific geometry-scaling relationships. *Geometry proxies.* The geometric summaries ( $d_{\text{int}}$ ,  $\kappa_{\text{ms}}$ ,  $\kappa^{\text{tv}}$ ,  $\nu$ ) are  $k$ -NN proxies computed from the activation point cloud, not direct computations of the pullback metric  $\mathbf{G}_\ell$ . They capture aspects of manifold structure that correlate with the pullback geometry, but a full characterisation would require computing the Jacobian  $J_\ell(h)$  at each layer, which is computationally expensive for large models.

**Implications for the linear-representation hypothesis.** The pullback Fisher–Rao framework reframes the LRH in geometric terms. As currently practised, the LRH posits that meaningful features are linear directions in the ambient Euclidean activation space, and a sparse code is a selection of such directions. The pullback metric we develop says that the natural distance between activations is not Euclidean on  $\mathbb{R}^{d_\ell}$  but Fisher–Rao on the predictive image  $F_\ell(\mathcal{H}_\ell) \subset \Delta_{\circ}^{V-1}$ . Two atoms orthogonal in the ambient inner product may be near-collinear under the pullback (and vice versa), so the LRH’s “meaningful linear directions” are metric-dependent objects: there is no metric-free answer to which directions of layer  $\ell$  are computationally meaningful. The flat sparse linear approximation that defines the SAE programme is not wrong, but it is making a metric choice (the ambient  $\ell_2$ ) whose alignment with the predictive structure is itself a layerwise empirical question; our cross-layer results suggest that alignment varies systematically with the manifold’s curvature and intrinsic dimension.

**Limits of the flat sparse linear programme.** Sparse linear dictionaries inherit the metric mismatch twice. First, the dictionary atoms are flat  $k$ -sparse linear combinations of Euclidean directions, which cannot conform to a curved activation manifold under any metric — the geometric wall (§5, Appendix E) is the empirical signature. Second, the  $\ell_2$  training objective optimises Euclidean distance, which is not the natural distance on activation space if Fisher–Rao governs predictive content; the per-layer scaling-rate variation we measure is consistent with this objective–metric mismatch. Beyond reconstruction, this matters for interpretability: the directional similarity between two atoms is meaningful only relative to a chosen metric, and the choice of  $\ell_2$  in training implicitly fixes that metric to the ambient one — which our framework suggests is not the metric activations naturally inherit. The current SAE programme is therefore best read as a useful first-order approximation to a geometrically richer object, not as the asymptotically correct decomposition.

**Toward geometry-grounded feature disentanglement.** A natural horizon is to develop sparse decompositions that are not flat linear combinations but local atlas charts on the activation manifold. Concretely: atoms as local frame fields rather than global linear directions; sparsity defined relative to the tangent space at each activation rather than a fixed global basis; a reconstruction objective that measures predictive discrepancy (e.g. a pullback-distance loss between  $F_\ell(h)$  and  $F_\ell(\hat{h})$ ) rather than ambient Euclidean error. Such a framework would treat the curved feature manifolds documented by Engels et al. [8], Modell et al. [13], Gurnee et al. [14] as first-class objects rather than approximating them by sparse combinations of flat directions, and would by construction remove both mismatches identified above. Intermediate steps include: per-layer metric learning to align the  $\ell_2$  objective with the local pullback structure; mixture-of-charts dictionaries whose atom selection is conditioned on the local tangent neighbourhood; and reconstruction losses that interpolate between Euclidean and pullback distances along a tunable schedule.

**Implications for downstream interpretability.** Operations that depend on SAE feature directions — activation steering, feature ablation, circuit discovery via residual decomposition — currently treat atoms as global Euclidean vectors. The pullback framework predicts that (i) the direction of “the same feature” varies smoothly across the manifold rather than being a fixed vector, so a steering intervention at layer  $\ell$  has different effective magnitudes at different inputs; (ii) the additivity of feature contributions, the residual-decomposition assumption underlying many circuit-discovery techniques, holds exactly only under metrics that locally linearise the predictive map, not under the ambient Euclidean metric. Our geometric-wall result is a layer-level diagnostic; per-input geometric corrections to interpretability operations are the finer-grained next step, and the layerwise geometry maps reported here (Fig. 1, Appendix B) provide a starting catalogue of where such corrections should be largest.

**Broader impacts.** This work characterises the cross-layer reliability of sparse autoencoders, a tool widely used in mechanistic interpretability for AI safety. Mapping where (which layers, which model sizes) SAE reconstruction is faithful versus geometrically constrained calibrates the trust placed in SAE-derived interpretations. The work uses only publicly released models and SAEs and does not introduce new capabilities; we see no specific risks of misuse beyond those of interpretability research generally.

**Future directions.** *Matched-budget and convergence-scaled SAE training.* Scaling studies with dictionary widths trained to a common per-width budget (or, following Gao et al. [1], to convergence) would isolate the architecture-only width-scaling component from the released-family training schedule. *Richer width grids at every layer.* Layerwise width grids spanning a decade or more at every layer would let the local with-floor calibration of Section 3.3 be performed at all layers rather than only the showcase subset, enabling a layerwise rather than aggregate floor characterisation. *Additional architectures and sublayers.* Extension of the cross-model geometric law to other transformer model families (e.g. Llama, Mistral, Qwen) is the natural test of cross-architecture transfer; extension to attention and MLP sublayer SAEs, to TopK and Gated SAE variants, and to non-transformer host models would broaden the scope further. *Direct pullback computation.* Computing  $\mathbf{G}_\ell(h) = J_\ell(h)^\top \Sigma_{F_\ell(h)} J_\ell(h)$  at intermediate layers and comparing the resulting geometric quantities with the  $k$ -NN proxies used here would provide a direct validation of the pullback framework.