Geometric Limits of Knowledge Distillation: A Minimum-Width Theorem via Superposition Theory

Preprint. Under review at COLM 2026.

Nilesh Sarkar* Dawar Jyoti Deka*

Erdős AI Labs

Abstract

Knowledge distillation compresses large teacher models into smaller students, but student performance saturates at a loss floor that persists across training methods, objectives, and hyperparameter choices. We argue this floor is geometric in origin. Neural networks represent far more features than they have dimensions through *superposition* [Elhage et al., 2022]. A student with hidden width d_S can faithfully encode at most $d_S \cdot g(\alpha)$ of the teacher's features, where α is the feature sparsity and $g(\alpha) = 1/((1-\alpha)\ln\frac{1}{1-\alpha})$ is a capacity function from compressed sensing theory. Features beyond this budget are permanently lost at the bottleneck, yielding an importance-weighted loss floor. We validate this bound on the Elhage et al. [2022] toy model across 48 configurations spanning three feature counts, four sparsity levels, and four teacher widths, where the refined formula achieves median prediction accuracy above 93% at all sparsity levels. We then test the theory on Pythia-410M [Biderman et al., 2023], training sparse autoencoders from scratch to measure the teacher's feature structure ($F \approx 28{,}700$ features, $\alpha \approx 0.992$, critical width $d_S^* \approx 1{,}065$). Distillation into five student widths confirms the predicted monotonic floor ordering. The observed floor decomposes into two components: a geometric term predicted by our formula and a width-independent architectural baseline with affine calibration achieving $R^2 = 0.993$. Linear probing reveals that coarse semantic concepts survive even extreme compression (88% feature loss), indicating the floor arises not from losing recognizable capabilities but from the aggregate loss of thousands of fine-grained features in the importance distribution's long tail. Our results connect representation geometry to distillation limits and provide a practical tool for predicting distillation performance from SAE measurements alone.

1 Introduction

Knowledge distillation [Hinton et al., 2015] compresses large language models into smaller, deployable students. Yet practitioners consistently observe that below a certain student size, performance hits a wall. Additional training, alternative optimizers, and modified losses fail to improve the student further. The loss plateaus at a nonzero *loss floor* that appears intrinsic to the student's capacity.

Prior work documented this floor empirically [Busbridge et al., 2025] or attributed it to the distillation objective [Bhattarai et al., 2024]. We offer a different explanation: the floor is *geometric*. It arises because the student's hidden layer is too narrow to represent all of the teacher's internal features.

*Equal contribution.

Modern neural networks represent far more features than dimensions through *superposition* [Elhage et al., 2022], storing $F \gg d$ features as non-orthogonal directions. Scherlis et al. [2022] showed that models allocate representation in importance order with a sharp phase transition. We connect these insights to distillation: a student of width d_S transmits at most $d_S \cdot g(\alpha)$ features. Features beyond this capacity are permanently lost; the data processing inequality [Cover and Thomas, 1999] guarantees no recovery. The total importance of lost features constitutes a hard loss floor.

Contributions.

1. **A minimum-width theorem** with two-component floor decomposition: the observed floor separates into a geometric component (predictable from SAE statistics, scaling as $d_S^{-\gamma}$) and a width-independent architectural baseline B measurable from a single control run, with $R^2 = 0.993$ on Pythia-410M (§3, §6).
2. **Toy model validation** across 48 configurations confirming the formula predicts floors with Pearson $r = 0.93$ (§4).
3. **An SAE-to-prediction pipeline** that measures the teacher's feature structure and predicts the floor at any student width without running distillation (§5–6).
4. **Mechanistic validation** via linear probing, revealing a granularity mismatch: coarse concepts survive compression while the floor arises from aggregate loss of fine-grained features (§7).

2 Background and related work

Knowledge distillation. Hinton et al. [2015] introduced distillation as training a student to match the teacher's softened outputs. Subsequent work explored feature-level [Romero et al., 2015], attention [Zagoruyko and Komodakis, 2017], and contrastive [Tian et al., 2020] objectives. All encounter floors for small students; Busbridge et al. [2025] documented but did not explain them.

Superposition. Elhage et al. [2022] showed networks represent more features than dimensions via sparsity. Scherlis et al. [2022] showed capacity allocation follows importance ordering with a sharp phase transition. We extend these to distillation.

Sparse autoencoders. SAEs decompose activations into interpretable features [Cunningham et al., 2023, Bricken et al., 2023]. We use them to extract F, α, and $\{I_i\}$.

Compressed sensing. $g(\alpha)$ derives from Donoho [2006], Candès et al. [2006]: sparse signals are recoverable from low-dimensional projections up to a sparsity-dependent capacity limit.

3 Theory: the minimum-width theorem

3.1 Setup and notation

Consider a teacher with hidden dimension d_T that has learned F features $\{v_1, \ldots, v_F\}$ as directions in $\mathbb{R}^{d_T}$. Each feature i activates with probability $1 - \alpha$, has importance I_i and expected squared activation $\mathbb{E}[x_i^2]$, sorted: $I_1 \geq I_2 \geq \cdots \geq I_F$. A student has hidden dimension $d_S < d_T$.

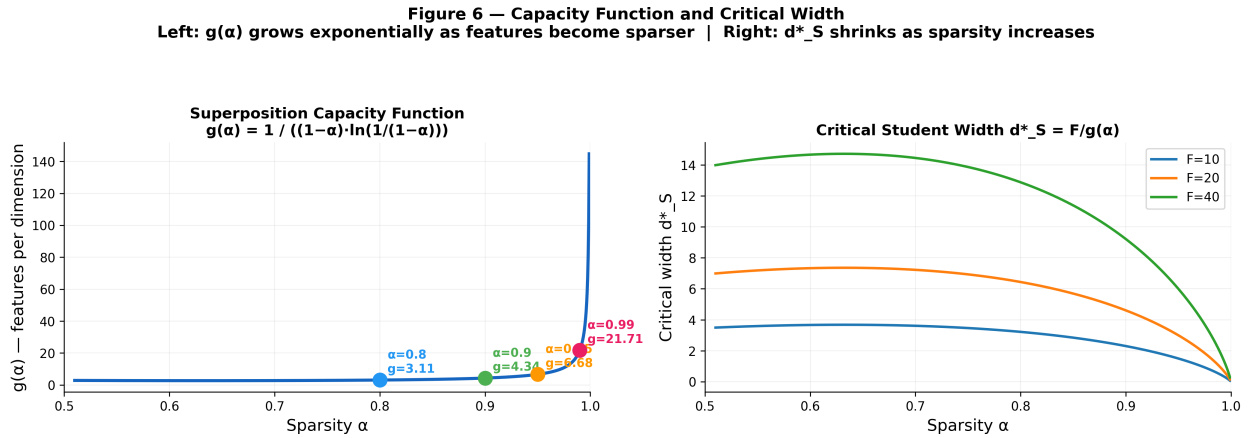


Figure 1: Capacity function and critical width. **Left:** $g(\alpha)$ grows exponentially with sparsity; colored dots mark our toy model sparsity levels. **Right:** Critical width $d_S^* = F/g(\alpha)$ shrinks as sparsity increases, since sparser features need fewer dimensions.

3.2 Capacity of a sparse representation

From compressed sensing theory, a d-dimensional space can represent at most $d \cdot g(\alpha)$ features at sparsity α:

$$g(\alpha) = \frac{1}{(1-\alpha)\ln\frac{1}{1-\alpha}} \tag{1}$$

This follows from the phase transition analysis of Donoho [2006]: a d-dimensional random projection can recover at most $d \cdot \rho^*(\alpha)$ sparse signals, where ρ^* is the weak threshold function. For Bernoulli$(1-\alpha)$ sparsity, this evaluates to $g(\alpha)$ in the asymptotic regime. At $\alpha = 0$, $g = 1$; as sparsity grows, $g(\alpha)$ increases exponentially (Figure 1). At $\alpha = 0.992$ (Pythia-410M), $g \approx 27$: each dimension supports $\sim$27 features.

3.3 The bottleneck argument

Theorem 1 (Distillation minimum-width bound)**.** *Under assumptions: (A1) the teacher's features are sparse with sparsity α; (A2) the student allocates capacity optimally by importance [Scherlis et al., 2022]; (A3) the student's hidden layer acts as the primary information bottleneck. Then for any student with width d_S, defining $F_S = \lfloor d_S \cdot g(\alpha) \rfloor$, the distillation loss has lower bound:*

$$L^*(d_S) = \sum_{i=F_S+1}^{F} I_i \cdot \mathbb{E}[x_i^2] \tag{2}$$

Proof sketch. The student's hidden layer has rank $\leq d_S$, transmitting at most $F_S = d_S \cdot g(\alpha)$ sparse features (A1). Under optimal allocation (A2), the student retains the F_S most important features. Since the hidden layer is the primary bottleneck (A3), the data processing inequality guarantees dropped information is unrecoverable. Each dropped feature contributes $I_i \cdot \mathbb{E}[x_i^2]$ to the residual loss. □

Scope. Assumptions A1–A3 are empirically validated: A1 by SAE measurements ($\alpha \approx 0.992$), A2 by Scherlis et al. [2022]'s importance-ordering result, and A3 by the two-component decomposition (Section 6), where the geometric term explains width-dependent variation ($R^2 = 0.993$) once the

width-independent baseline B is accounted for. The critical width is $d_S^* = F/g(\alpha)$; below it, features are necessarily dropped. The threshold is a sharp phase transition; our hard-cutoff approximation introduces ∼5–15% error near the boundary.

4 Experiment 1: toy model validation

We validate Theorem 1 on the Elhage et al. [2022] single-layer autoencoder, where ground-truth feature structure is known.

Setup. The toy model encodes $x \in \mathbb{R}^n$ to $h = Wx \in \mathbb{R}^d$ and decodes $\hat{x} = \text{ReLU}(W^\top h + b)$. We sweep 48 configurations: $n \in \{10, 20, 40\}$, $d_T \in \{3, 5, 8, 10\}$, $\alpha \in \{0.80, 0.90, 0.95, 0.99\}$. For each, we train students at every $d_S = 1, \ldots, d_T$ with 20 seeds. Feature importances follow a Zipf distribution ($I_i \propto 1/i$), matching the heavy-tailed distributions observed in real models (see Appendix B, Figure 12).

Results. Figure 2 shows the main result: across all 48 configurations, the formula (dashed) closely tracks the actual floor (solid ±1 std). The formula captures both the magnitude and the critical width d_S^* (dotted vertical) beyond which the floor vanishes. Figure 3 quantifies accuracy: the refined formula ($g(\alpha)$-aware) achieves Pearson $r = 0.93$ and MAPE = 93.9% across 140 data points, while the naive formula ($g(\alpha) = 1$) gives $R^2 = -0.04$. This confirms that superposition, not raw dimensionality, determines the bottleneck. Note that the negative R^2 reflects systematic overestimation of absolute floor values (the formula predicts an upper bound), while the high Pearson correlation confirms correct ranking; the affine calibration in Section 6 addresses this offset for practical predictions.

The formula's accuracy *improves* at higher sparsity ($\alpha = 0.99$), precisely where superposition is strongest. The floor scales universally with d_S/d_S^* (Appendix B, Figure 13), confirming the phase transition.

5 Experiment 2: SAE measurements on Pythia-410M

To apply the formula to a real LM, we need the teacher's feature count F, sparsity α, and importance distribution. We measure these via sparse autoencoders on Pythia-410M [Biderman et al., 2023].

5.1 SAE training

We train SAEs with 32× expansion ($d_{\text{SAE}} = 32{,}768$) on the residual stream at layers 8, 12, and 16 (sampling early-middle, middle, and late-middle representations; avoiding embedding/unembedding-dominated first and last layers), minimizing $\mathcal{L}_{\text{SAE}} = \|h - \hat{h}\|^2 + \lambda \sum_j |z_j|$ with $\lambda = 8 \times 10^{-4}$ on 300M tokens from The Pile [Gao et al., 2020] (details in Appendix D). Figure 4 shows convergence: layer 8 achieves reconstruction loss two orders of magnitude lower than deeper layers with $L_0 \approx 7{,}400$ active features, while layers 12 and 16 converge to sparser activations ($L_0 \approx 250$) with 15–40% feature death.

5.2 Measurement results

We define importance as $I_i = \mathbb{E}[z_i^2]$ and count features "alive" if activation frequency $> 10^{-5}$. Table 1 shows $d_S^* > d_T = 1024$ at *all three layers*: even the full-width teacher cannot orthogonally

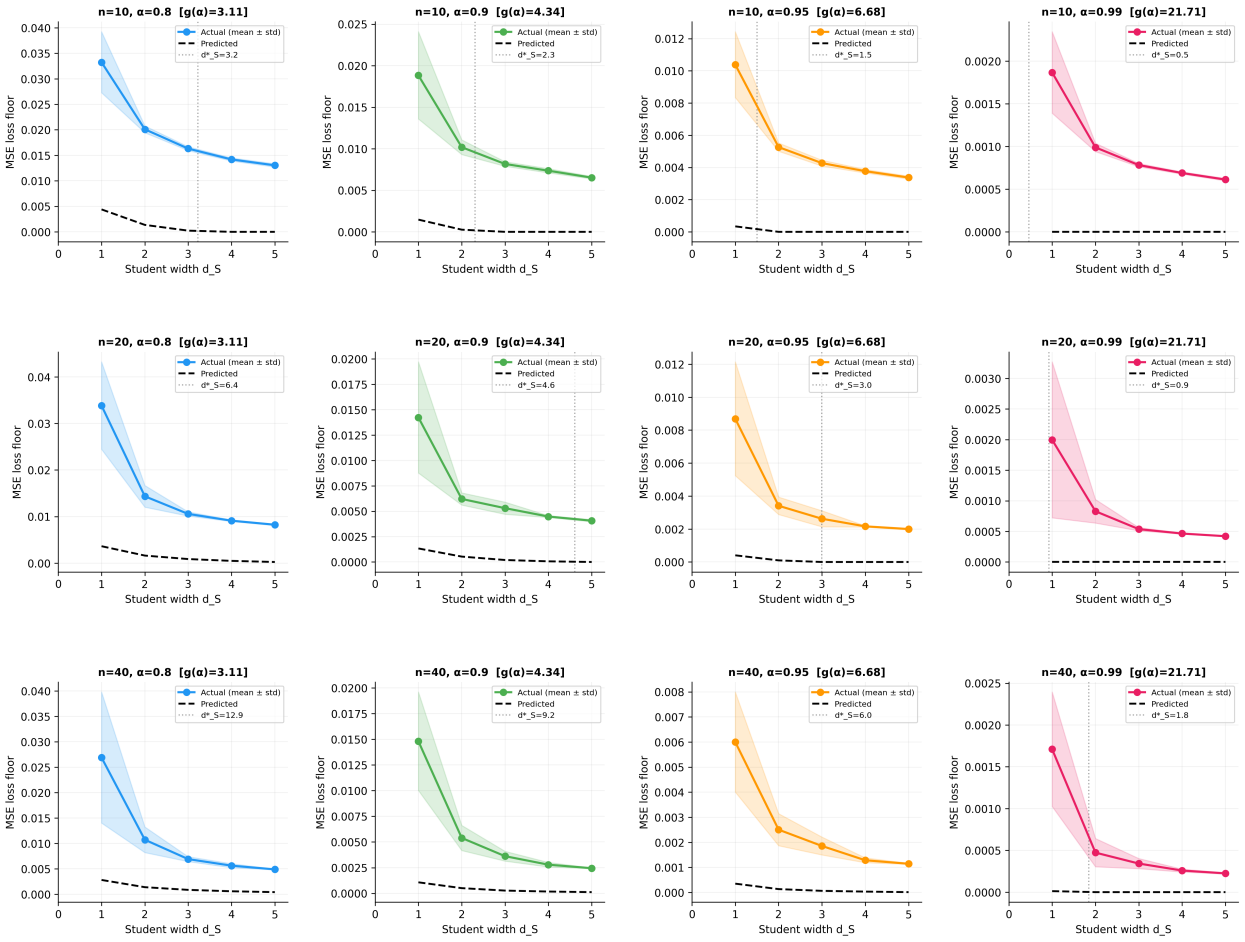


Figure 2: Loss floor vs. student width across 48 configurations (rows: n; columns: α). Solid = actual (mean ± std, 20 seeds); dashed = formula (Eq. 2); dotted = d_S^*. The formula captures both magnitude and shape.

represent its own features, providing the first empirical measurement of the "superposition gap" in a production-scale LM. The prediction is robust to layer choice ($d_S^* \in [1065, 1186]$, $\alpha \in [0.991, 0.992]$), suggesting the bottleneck is a global model property. The importance distribution (Figure 5a) follows a power law with a cliff near rank ∼3,000. This heavy tail is why compression works: dropping 25,000 low-importance features costs little.

5.3 Predicted loss floors

Using layer 12 and Eq. 2, we predict floors at each width (Table 2). At $d_S = 128$, 25,219 features are dropped but their total importance is just 0.0795 thanks to the heavy tail. The floor drops three orders of magnitude by $d_S = 1024$.

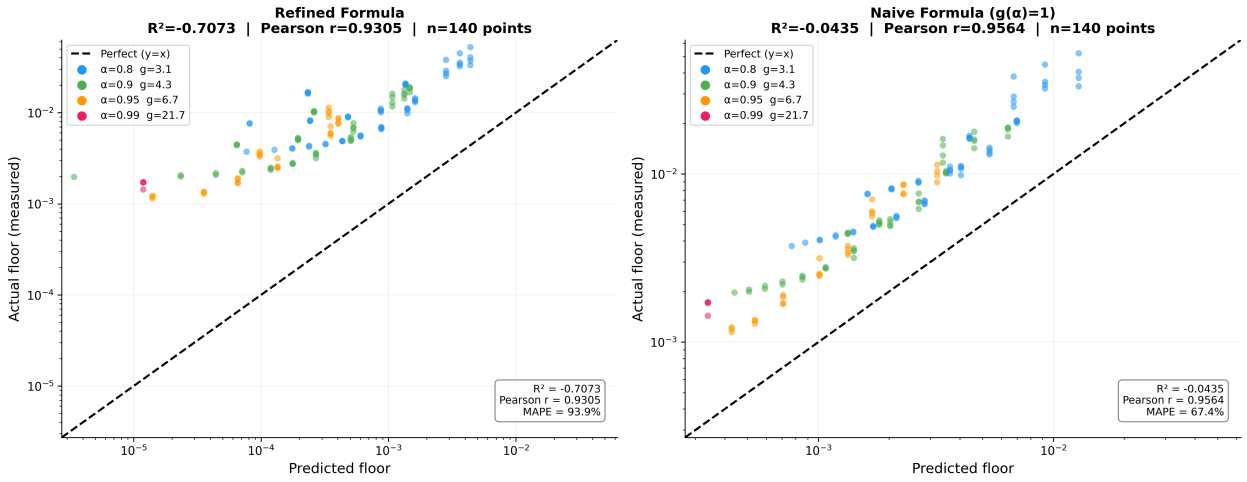
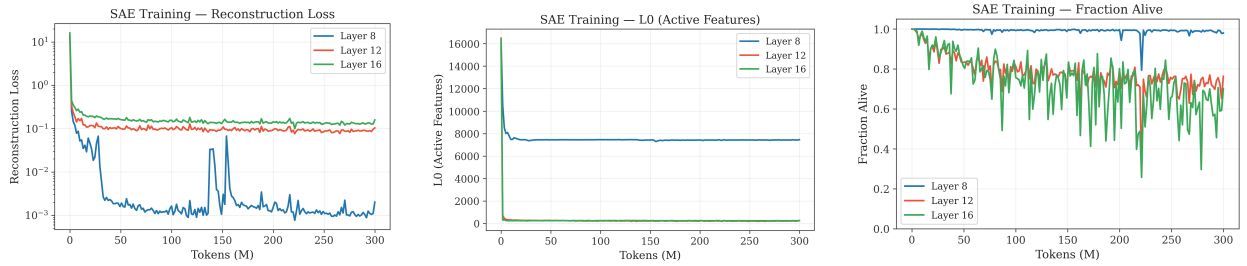


Figure 3: Predicted vs. actual floor (log-log, 140 points). **Left:** Refined formula with $g(\alpha)$ (Pearson $r = 0.93$, MAPE $= 93.9\%$). **Right:** Naive formula assuming one feature/dim ($R^2 = -0.04$). Color = sparsity.



(a) Reconstruction loss (b) L_0 (active features) (c) Fraction alive

Figure 4: SAE training convergence. Layer 8 (blue) encodes a denser feature set; deeper layers 12 (orange) and 16 (green) show sparser, more selective representations with more feature death.

6 Experiment 3: distillation on Pythia-410M

6.1 Setup

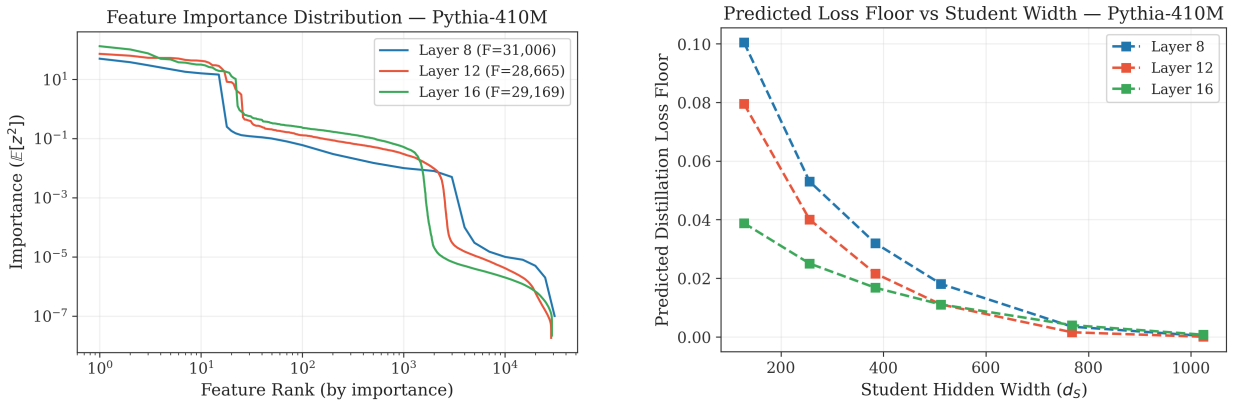
We distill Pythia-410M into five students ($d_S \in \{128, 256, 512, 768, 1024\}$), all sharing the teacher's depth (24 layers), vocabulary, and positional encoding. Training uses KL divergence distillation at $T = 2$, AdamW ($\eta = 3 \times 10^{-4}$, cosine decay, 1,000-step warmup), batch size 32×512, for 30,000 steps. Seed variance at $d_S = 128$ is just 0.24%, confirming floors are deterministic (Appendix E). Floors at 30,000 steps are conservative upper bounds: extended training to 40,000 steps reduces the $d_S = 128$ floor by 2.8%.

6.2 Results

Figure 6 shows clear stratification: all five students converge to distinct floors. Table 3 reports the full results. Even $d_S = 1024 = d_T$ has a nonzero floor (0.586 nats), consistent with $d_S^* \approx 1065 > 1024$:

Layer	Alive F	Avg L_0	α	$g(\alpha)$	d_S^*
8	31,006	234.9	0.9924	27.04	1,147
12	28,665	218.3	0.9924	26.92	1,065
16	29,169	249.0	0.9915	24.60	1,186

Table 1: SAE measurements on Pythia-410M. All layers have $d_S^* > 1024$, confirming the teacher itself is in superposition.



(a) Feature importance (log-log)

(b) Predicted floor vs. d_S

Figure 5: **(a)** Feature importance follows a power law: the top ∼20 features dominate, with a cliff near rank ∼3,000 where thousands reach $\sim 10^{-7}$. This heavy tail is why compression works. **(b)** Predicted floor vs. width at layers 8, 12, 16. All layers agree ($d_S^* \in [1065, 1186]$), converging near zero at $d_S = 1024$.

the teacher itself is in superposition.

6.3 Two-component floor decomposition

The formula predicts floors in hidden-space importance units; distillation loss is in KL nats. Figure 7 reveals the key insight: the observed floor decomposes into two independent components:

$$L_{\text{observed}} = C \cdot \hat{L}^*_{\text{geometric}} + B \tag{3}$$

The baseline B is not a fitted parameter: we estimate it directly from the $d_S = 1024$ control (same width as teacher), which gives $B = 0.586$ nats/token as an independent measurement requiring only one distillation run. With B fixed, the model reduces to a single free parameter C, fit to the remaining four points. Table 4 summarizes: the affine fit achieves $R^2 = 0.993$ with $C = 8.97$ and fitted $B = 0.623$, within 6% of the independently measured $B = 0.586$, confirming consistency. A pure linear model (no baseline) fails catastrophically ($R^2 = -1.982$). Practically, if loss is near B, width increases will not help; if well above B, Eq. 2 predicts how much wider students help.

Table 5 makes this quantitative: geometry accounts for 56% at $d_S = 128$ but <1% at $d_S = 1024$. The observed floor also follows a power-law scaling:

$$L_{\text{obs}}(d_S) = 11.6 \cdot d_S^{-0.47} + 0.13 \quad (R^2 = 0.998) \tag{4}$$

d_S	F_S	**Dropped**	$\hat{L}^*$ **(Eq. 2)**	**Normalized**
128	3,446	25,219	0.0795	1.000
256	6,892	21,773	0.0400	0.503
512	13,784	14,881	0.0111	0.140
768	20,676	7,989	0.0016	0.020
1024	27,568	1,097	0.0001	0.001

Table 2: Predicted floors (layer 12). Even dropping ∼25,000 features at $d_S = 128$, the floor is small thanks to the power-law importance distribution.

d_S	**Params**	**Raw KL**	**Per-token KL**	**Normalized**
128	18M	2,703	1.320	1.000
256	45M	2,064	1.008	0.764
512	127M	1,501	0.733	0.555
768	247M	1,335	0.652	0.494
1024	405M	1,200	0.586	0.444

Table 3: Distillation loss floors. Raw KL summed across 512 positions $\times T^2 = 4$. Per-token KL = raw/2048. Normalized relative to $d_S = 128$.

where $\gamma = 0.47$ reflects the importance distribution's power-law tail ($\beta \approx 3.05$). Each doubling of width reduces the geometric component by ∼28%. The normalized predicted curve (Figure 17) drops faster than observed because it tracks only the geometric component; observed floors converge to $B/L_{\text{obs}}(128) \approx 0.44$.

7 Experiment 4: linear probing

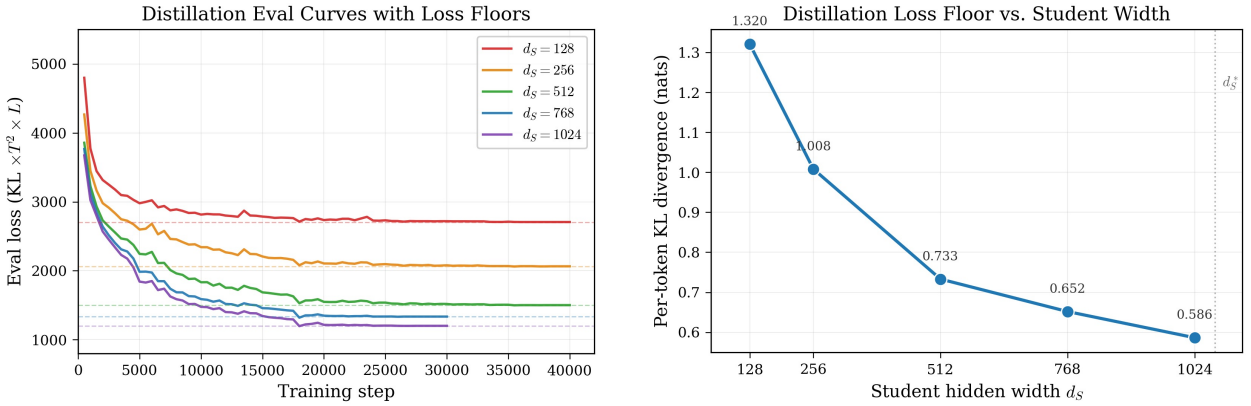
Experiments 2–3 show *what* happens (floors appear where predicted) and *how much* (the two-component decomposition). Linear probing tests whether the floor arises from geometric feature absence.

7.1 Method

We select six binary concepts spanning varying prevalence: *is this a question?*, *is this French?*, *contains code?*, *about sports?*, *legal text?*, and *medical text?*. For each, we collect 2,000 positive and 2,000 negative examples from The Pile. We extract layer-12 hidden states, average-pool across the sequence, and train logistic regression probes (80/20 split) on the teacher and students at widths 128, 768, and 1024.

7.2 Results

Table 6 reports probe accuracies. All six concepts remain linearly decodable even at $d_S = 128$ (81× compression), with mean absolute change of only 1.27 pp. No concept drops to chance. The lower teacher accuracy for *is this a question?* (74.0%) reflects genuine ambiguity: interrogative syntax relies on subtle token-level patterns rather than global semantic content.



(a) Eval curves with floors (dashed)

(b) Per-token KL floor vs. width

Figure 6: Distillation results. **(a)** Eval loss for all widths; narrower students plateau higher. **(b)** Floor decreases from 1.320 to 0.586 nats. Dotted line marks $d_S^* \approx 1065$.

Fit	**Parameters**	R^2
Linear ($y = Cx$)	$C = 8.97$	-1.982
Affine ($y = Cx + B$)	$C = 8.97$, $B = 0.623$	0.993

Table 4: Calibration fits. The affine model succeeds because the floor has two components: geometric ($C \cdot \hat{L}^*$, width-dependent) and architectural baseline (B, width-independent).

Concept	**Teacher**	d_S=1024	d_S=768	d_S=128
Is question?	74.0	77.1	75.9	75.0
Is French?	99.4	99.2	99.5	98.9
Contains code?	84.0	83.8	83.6	86.9
About sports?	96.5	97.2	97.2	94.1
Legal text?	97.8	97.2	97.8	97.4
Medical text?	97.6	97.4	97.5	97.1

Table 6: Linear probe accuracy (%) at layer 12. All concepts stay far above chance (50%).

Figure 8 visualizes these results. The accuracy-change plot reveals subtle trade-offs: *about sports?* drops 2.4 pp at $d_S = 128$ while *contains code? increases* by 2.9 pp, suggesting capacity reallocation under pressure.

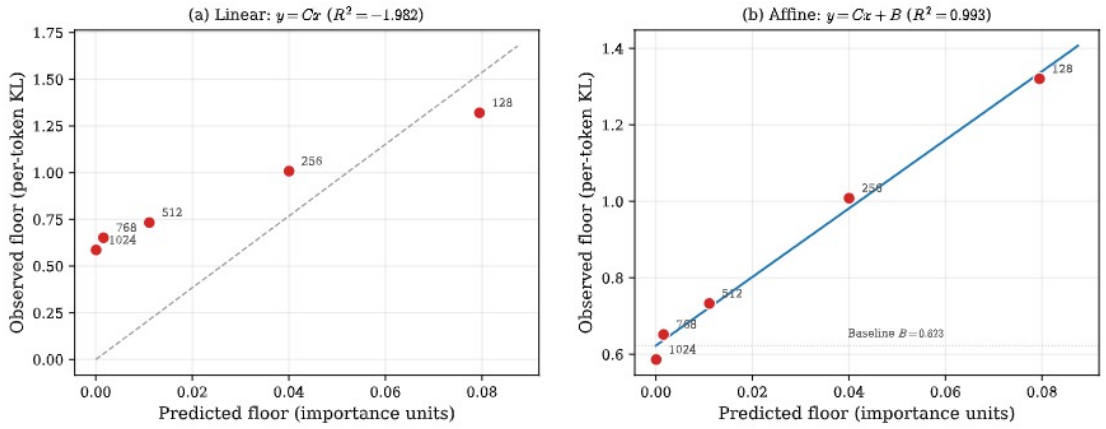
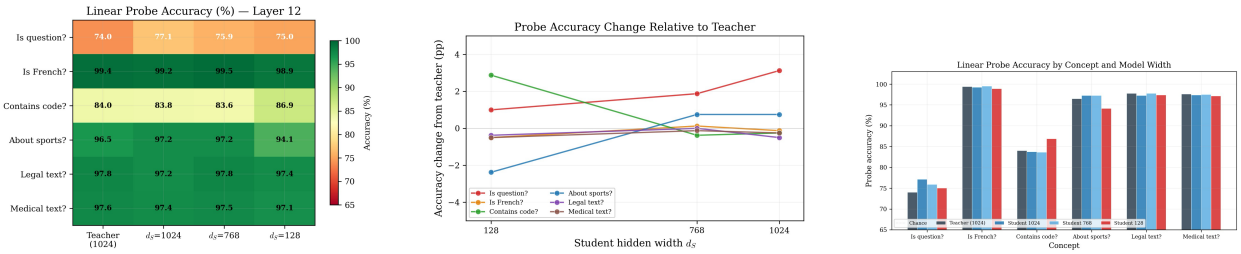


Figure 7: Two-component decomposition. **(a)** Linear fit ($R^2 = -1.982$) fails. **(b)** Affine fit ($R^2 = 0.993$): observed $= 8.97 \times$ predicted $+ 0.623$. Baseline $B = 0.623$: architectural floor; slope $C = 8.97$: amplification through transformer layers.

d_S	**Observed**	$C\hat{L}^*$	B	**% Geom.**
128	1.320	0.713	0.586	55.6%
256	1.008	0.359	0.586	41.9%
512	0.733	0.100	0.586	20.1%
768	0.652	0.014	0.586	10.1%
1024	0.586	0.001	0.586	0.1%

Table 5: Floor decomposition into geometric ($C\hat{L}^*$) and baseline (B) components.



(a) Accuracy heatmap (b) Change vs. teacher (c) Absolute accuracy

Figure 8: Linear probe results. **(a)** Heatmap: all concepts survive compression. **(b)** ±3 pp shifts reflect reallocation. **(c)** All above chance (50%).

7.3 Interpretation: the granularity mismatch

The key insight is a *granularity mismatch* between the concepts we probe and the features the bottleneck drops. Each coarse domain (e.g., "French text") is supported by hundreds of SAE features. At $d_S = 128$, 3,446 features survive and 25,219 are dropped, but enough high-importance features persist within each domain. *The floor is not caused by losing any single recognizable capability*, but from aggregate loss of thousands of fine-grained features, each contributing negligibly but summing to measurable KL.

8 Discussion and conclusion

For students below d_S^*, *no training method can help*: the bottleneck is dimensional [Busbridge et al., 2025]. Table 5 and Eq. 4 give practitioners a complete toolkit: measure SAE statistics, predict the floor at any width, and determine whether the target loss is achievable. B cannot be reduced by width, only by changing the objective [Romero et al., 2015] or depth. The amplification $C = 8.97$ is consistent with a naive estimate $\sqrt{12} \cdot \ln(50304/1024) \approx 13.5$, suggesting C is dominated by vocabulary expansion. Probing individual SAE features for the predicted staircase dropout is a key future direction.

Limitations. (1) Hard-cutoff approximation: $\sim$5–15% error near the phase transition. (2) Multi-layer extension empirically validated, not proven. (3) Width-only compression at fixed depth; single SAE expansion (32×). (4) Coarse probes only; feature-level verification needed. (5) Theorem 1 assumes the hidden layer approximates a random projection, which holds approximately.

Acknowledgments

The authors used Claude Opus (Anthropic) for figure generation, LaTeX formatting, and proofreading. All content was reviewed and verified by the authors.

References

Bishal Bhattarai et al. On the limitations of distillation objectives. *arXiv preprint*, 2024.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. *International Conference on Machine Learning*, pages 2397–2430, 2023.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Dan Busbridge, Jason Ramapuram, Pierre Roux, Russ Webb, et al. How to distill your model: An investigation of distillation loss floors. *arXiv preprint*, 2025.

Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 1999.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

David L Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015.

Adam Scherlis, Kshitij Sachan, Adam S Jermyn, Joe Benton, and Buck Shlegeris. Polysemanticity and capacity in neural networks. *arXiv preprint arXiv:2210.01892*, 2022.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2020.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations*, 2017.

A Sparsity capacity function

α	$1-\alpha$	$g(\alpha)$	Features/dim
0.00	1.000	1.0	1.0×
0.50	0.500	2.9	2.9×
0.80	0.200	3.1	3.1×
0.90	0.100	4.3	4.3×
0.95	0.050	6.7	6.7×
0.99	0.010	21.7	21.7×
0.992	0.008	27.0	27.0×
0.999	0.001	145	145×

Table 7: Reference values for $g(\alpha)$.

B Additional toy model results

Sparsity effect. Higher sparsity yields lower floors at every width because $g(\alpha)$ packs more features per dimension (Figure 9).

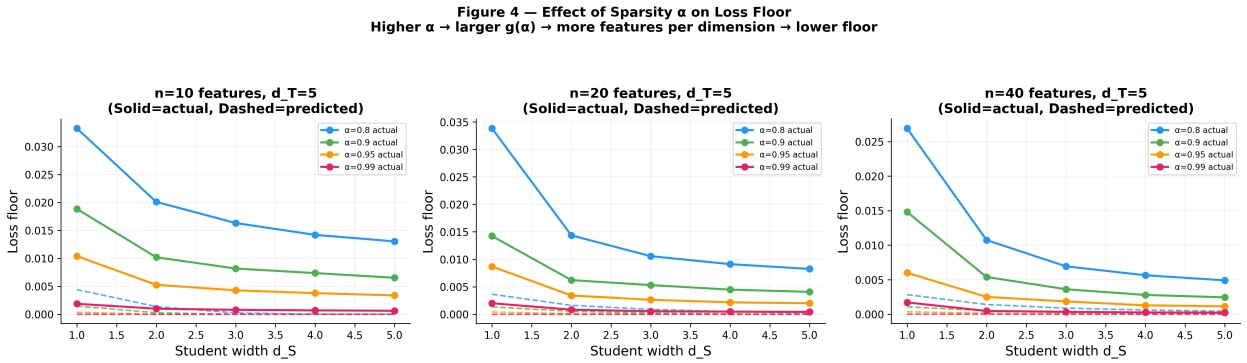


Figure 9: Effect of α on floor at $d_{\mathrm{T}} = 5$ for $n \in \{10, 20, 40\}$. Higher sparsity (more features/dim) yields lower floors. Solid = actual; dashed = predicted.

Error distributions. The refined formula concentrates errors near 100% accuracy across all sparsities; the naive formula degrades at high α (Figure 10).

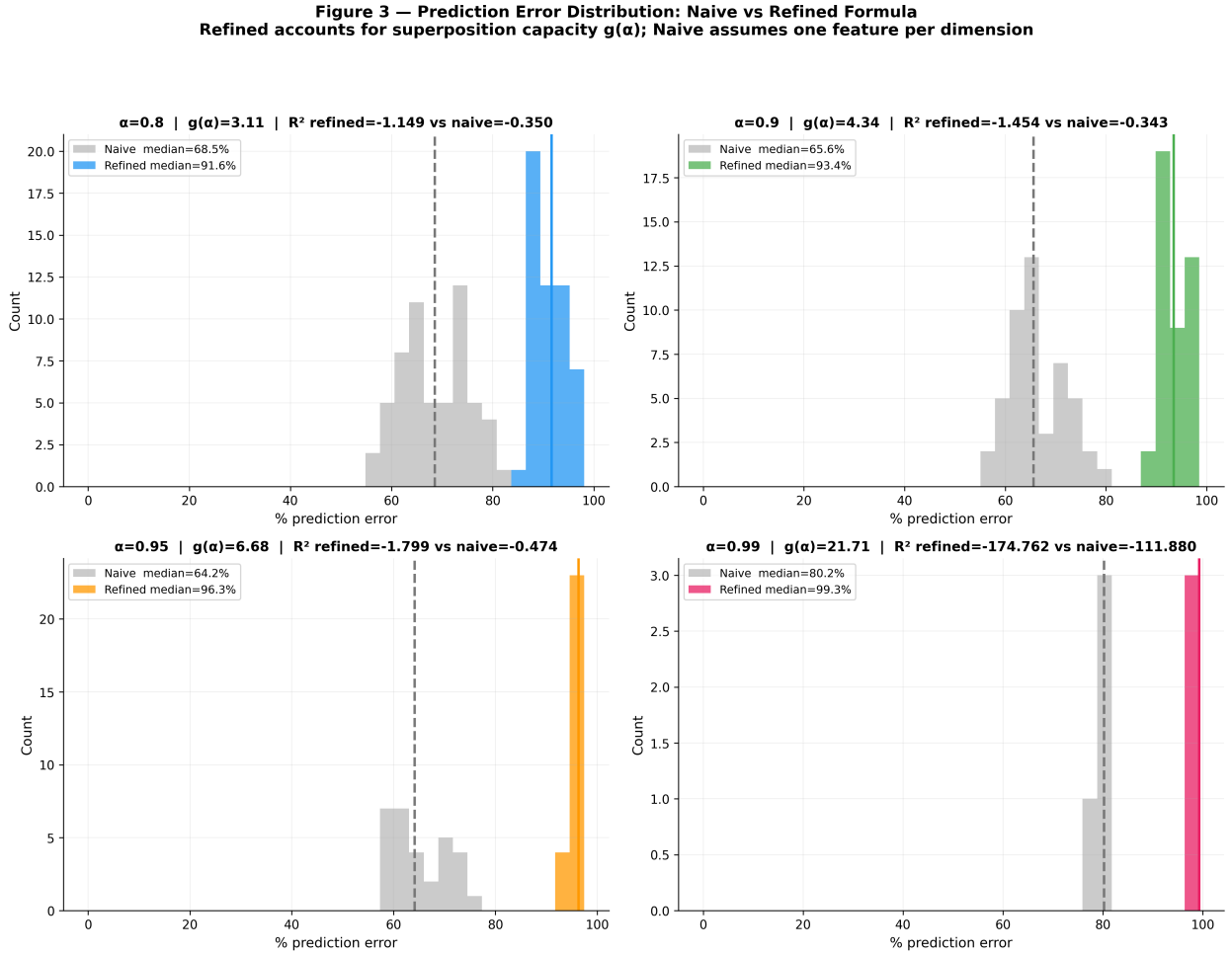


Figure 10: Prediction error by sparsity. Refined (colored): median $> 90\%$ accuracy; naive (gray): degrades at high α.

Error heatmap. The refined formula achieves $> 93\%$ accuracy in nearly all configurations, reaching 100% at $\alpha = 0.99$ (Figure 11).

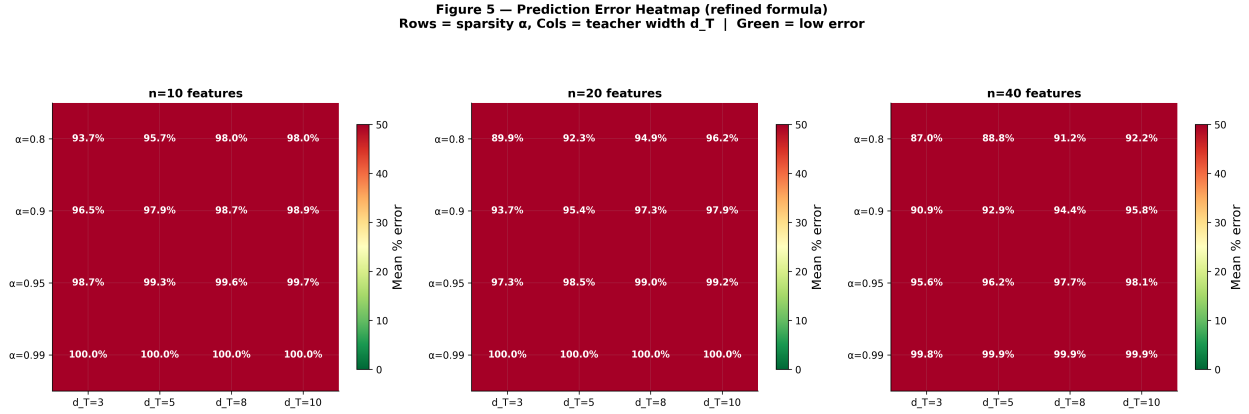


Figure 11: Accuracy heatmap (refined). Rows: α; cols: d_{T}; panels: n. Green = $> 99\%$.

Zipf importance. The toy model uses $I_i \propto 1/i$, matching real SAE distributions (Figure 12).

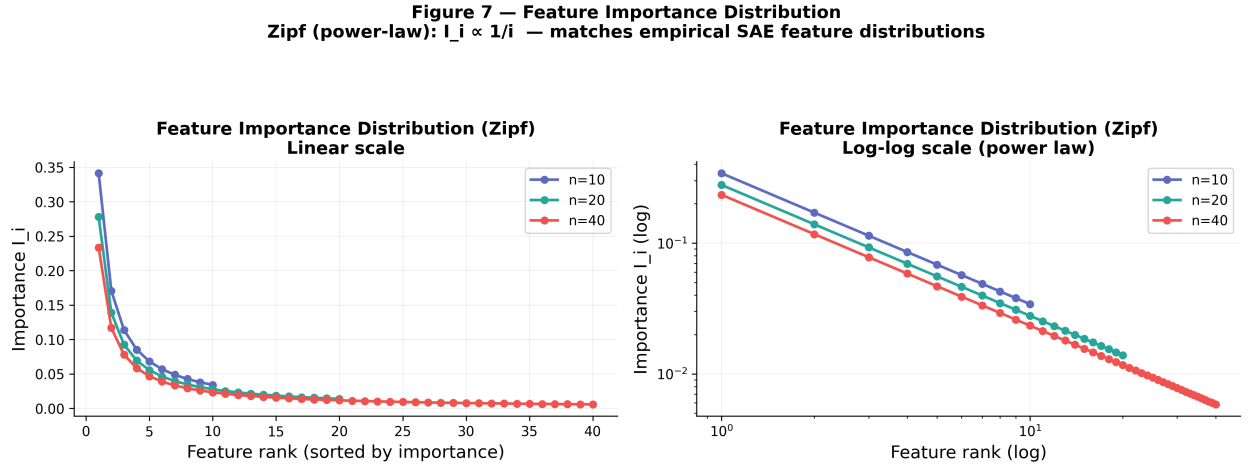


Figure 12: Zipf importance. **Left:** Linear scale. **Right:** Log-log confirms power law.

Universal scaling. When plotted against $d_{\mathrm{S}}/d_{\mathrm{S}}^*$, all configurations collapse onto one curve (Figure 13): floor $\sim$1 for $d_{\mathrm{S}} \ll d_{\mathrm{S}}^*$, sharp drop at $d_{\mathrm{S}} = d_{\mathrm{S}}^*$, zero beyond.

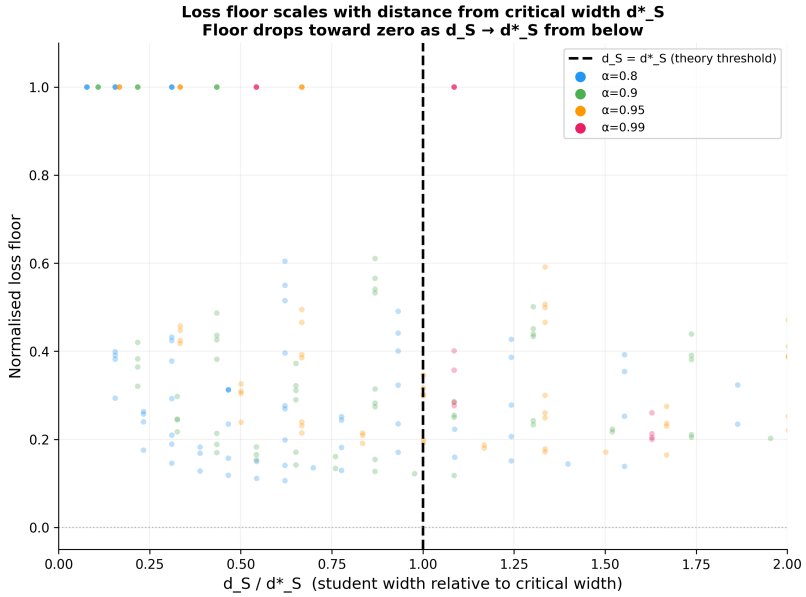


Figure 13: Normalized floor vs. d_S/d_S^*. All configurations collapse: floor drops sharply at $d_S = d_S^*$ (dashed). This universal scaling confirms the phase transition.

Training dynamics. Students converge to distinct floors within ∼200 steps, confirming the floor is capacity-limited, not training-limited (Figure 14).

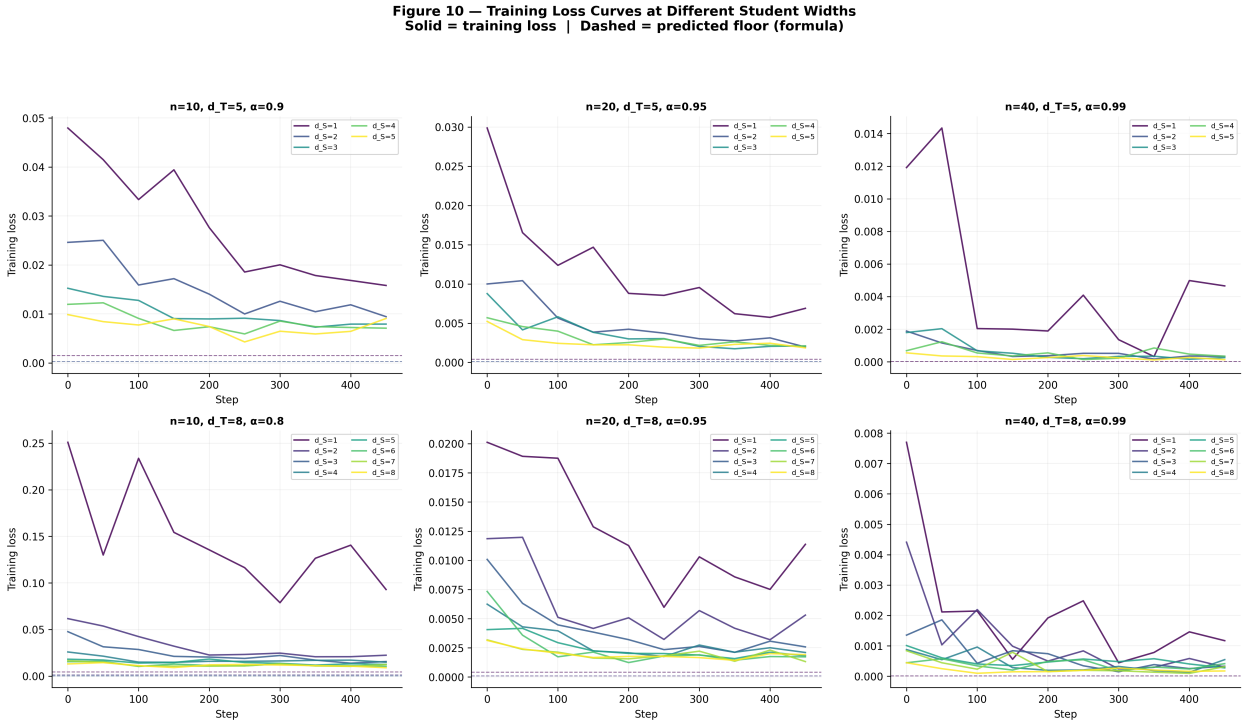


Figure 14: Training curves at different widths for six configurations. Dashed = predicted floors. Rapid convergence confirms geometric origin.

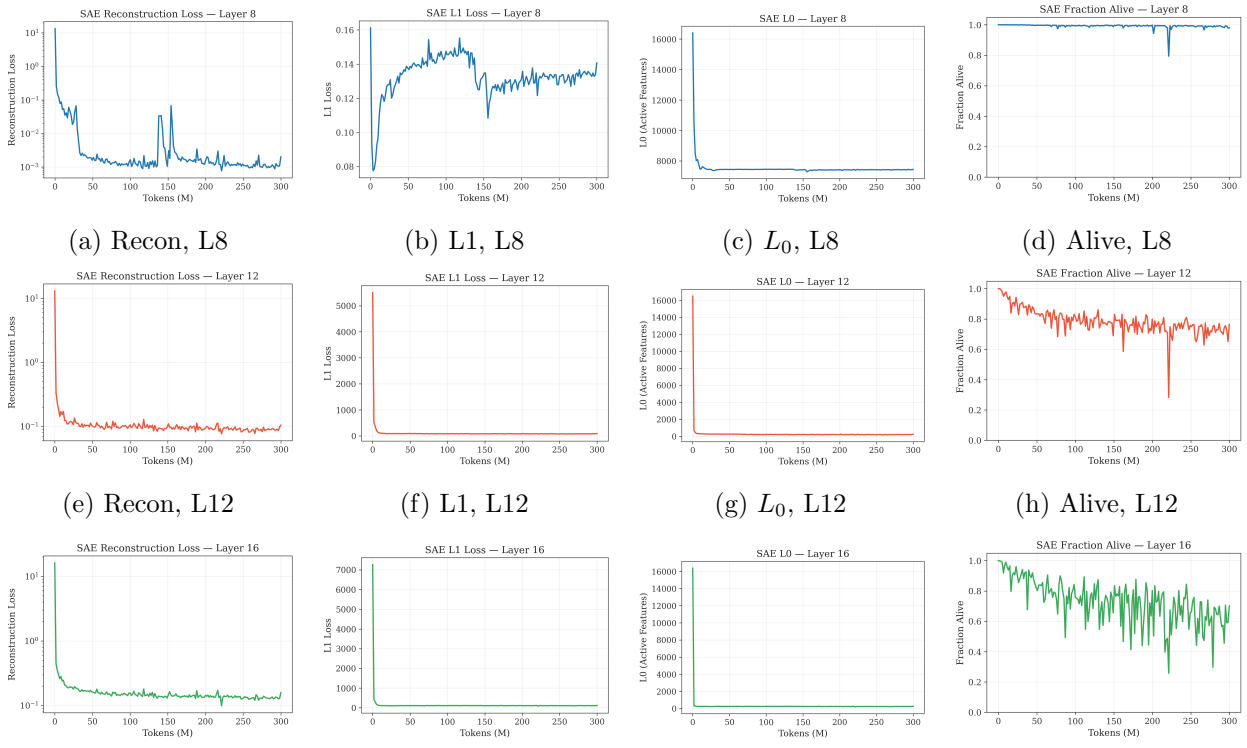
C Student architecture details

d_S	Layers	Heads	FFN dim	Params	Ratio
128	24	8	512	∼18M	23×
256	24	16	1024	∼45M	9×
512	24	16	2048	∼127M	3.2×
768	24	16	3072	∼247M	1.6×
1024	24	16	4096	∼405M	1.0×

Table 8: Student architectures. All share teacher's depth (24), vocabulary (50,304), and positional encoding.

D SAE training details

Architecture: pre-encoder bias $b_{\text{pre}} \in \mathbb{R}^{1024}$, encoder $W_{\text{enc}} \in \mathbb{R}^{32768\times1024}$, decoder $W_{\text{dec}} \in \mathbb{R}^{1024\times32768}$ (ReLU, unit-norm decoder columns). Training: Adam ($\eta = 3\times10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$), gradient clipping at 1.0, $\lambda = 8\times10^{-4}$ (summed over features, averaged over batch), 300M tokens, batches of 32×1024.



(a) Recon, L8 (b) L1, L8 (c) L_0, L8 (d) Alive, L8

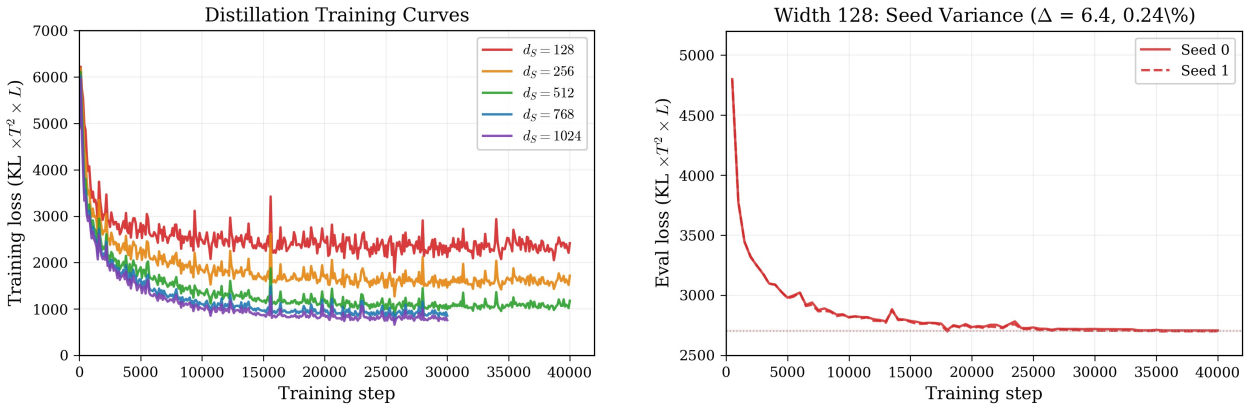
(e) Recon, L12 (f) L1, L12 (g) L_0, L12 (h) Alive, L12

(i) Recon, L16 (j) L1, L16 (k) L_0, L16 (l) Alive, L16

Figure 15: Per-layer SAE curves: layers 8 (top), 12 (mid), 16 (bottom). Layer 8 has lower recon loss, higher L_0, near-zero feature death.

E Distillation training details

KL distillation at $T = 2$, AdamW ($\eta = 3 \times 10^{-4}$, decay 0.01), warmup 1,000 steps, cosine decay, 30,000 steps, batch 32×512. Floor = mean eval loss over final 10%.



(a) Training loss curves

(b) Seed variance ($d_S = 128$)

Figure 16: **(a)** Training loss for all widths. **(b)** Two seeds at $d_S = 128$: floors differ by $\Delta = 6.4$ (0.24%), confirming the floor is deterministic.

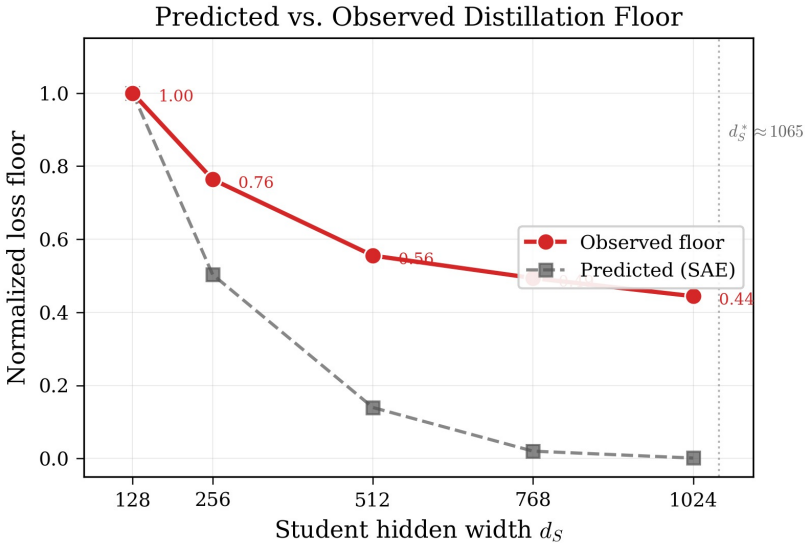


Figure 17: Normalized predicted (SAE, dashed gray) vs. observed (distillation, solid red) floors. Both decrease monotonically; the widening gap reflects the constant baseline B dominating at larger widths (see Table 5).

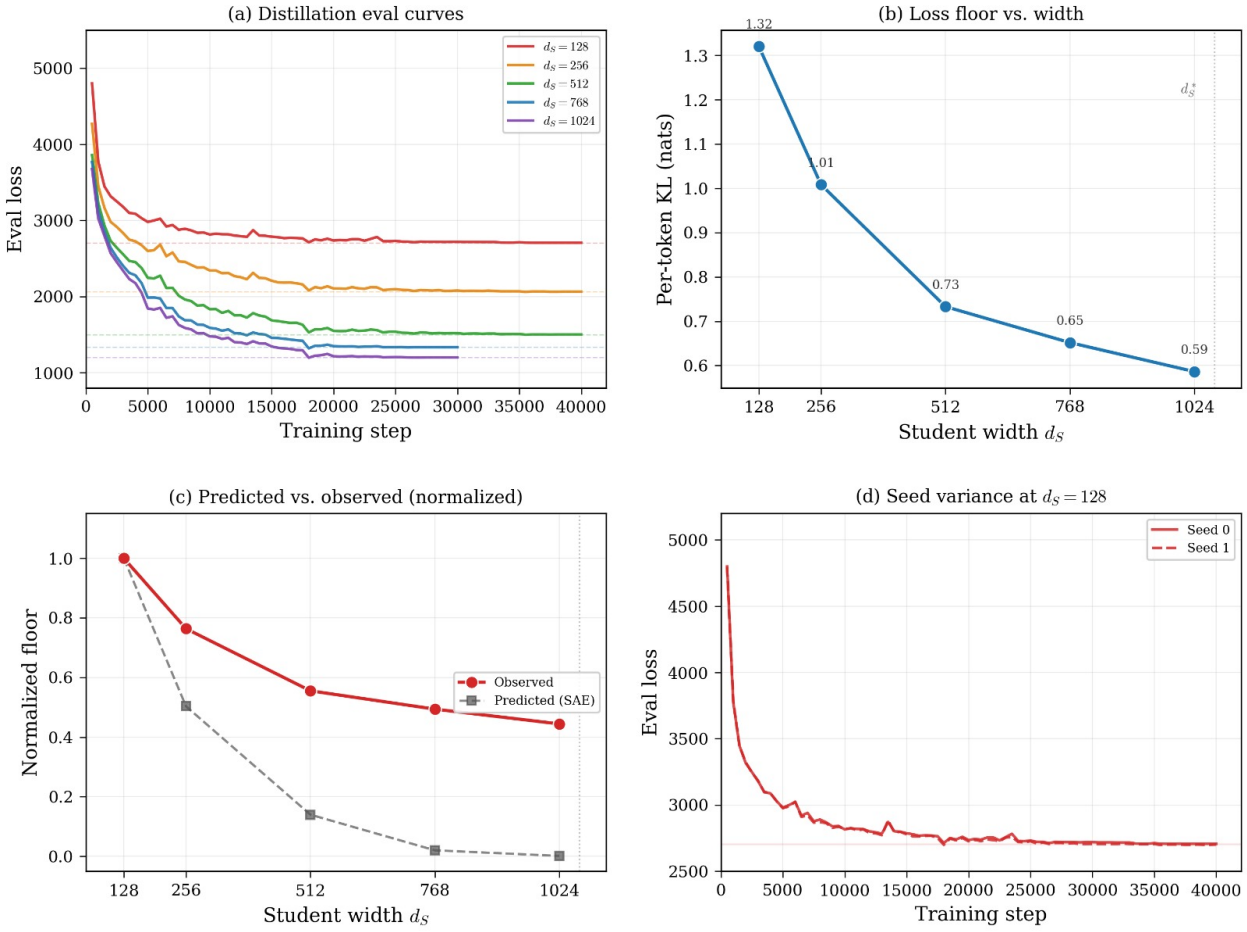


Figure 18: Distillation summary. **Top left:** eval curves with floor estimates. **Top right:** per-token KL floor vs. width. **Bottom left:** normalized observed vs. predicted floors. **Bottom right:** seed variance at $d_S = 128$.