Width-to-Depth Ratio Tracks Readout Misalignment in Language Models

Weilun Xu [1]

Abstract

Interpretability tools like the logit lens read a model's *intermediate* layers through the same output map that turns its final layer into a next-token prediction, assuming those layers are already oriented toward that map. Whether, and when, that assumption holds has not been measured. We introduce **Predictive-Geometric Alignment (PGA)**, a null-calibrated score of how much of a layer's geometry lies in the output map's own directions, and find a late-layer collapse of it in narrow models: abrupt, invisible to loss and to spectral measures, largely absent in wide ones. Width relative to depth (d/L), not parameter count, orders its severity across a seven-model Pythia suite and near-perfectly by rank on a held-out corpus. Under a pre-registered change of training data, the ordering's extremes survive but its sharp boundary does not. The collapse masks rather than erases predictive structure, which is recovered by removing a few dominant directions or by a small trained per-layer reader. It switches on late in training, long after the copying mechanism forms, and falls selectively on predictions copyable from the context. A failed reading at such a layer reflects the readout, not missing knowledge, so tools that read through the output map should treat these layers as readout-limited. How far a model's geometry stays readable this way is a property that emerges during training, tracked by width relative to depth.

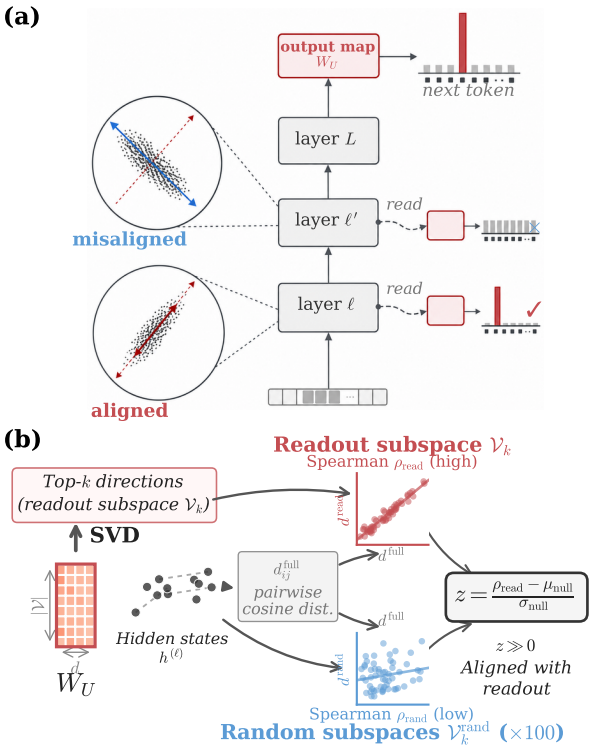

Figure 1. **The premise and the measurement** **(a)** Only the final layer's state becomes a prediction, yet interpretability methods read *intermediate* layers through the same output map; the reading is informative only if the layer's geometry is oriented toward that map (aligned) rather than elsewhere (misaligned). **(b)** PGA scores this orientation: how well the readout's preferred directions preserve a layer's distance structure, against equal-size random directions. Positive: the geometry concentrates in the readout's directions; below zero: even random directions preserve it better.

1. Introduction

A language model computes through many layers, but only the final one turns its internal state into a prediction: the output layer scores every candidate next token, and training constrains nothing before it. A growing family of interpretability methods nevertheless reads *intermediate* layers through that same output map: the logit lens (nostalgebraist, 2020) asks what token an intermediate state would predict, the tuned lens (Belrose et al., 2023) learns a small correction before reading, and vocabulary-space analyses of internal components build on the same operation (Geva et al., 2022; Ghandeharioun et al., 2024).

These readings can only report what falls within the directions the output map responds to, so they carry an unstated

[1]School of Computer and Communication Sciences, EPFL, Lausanne, Switzerland. Correspondence to: Weilun Xu <weilun.xu@epfl.ch>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

premise: the layer being read must already be oriented toward that map. Existing analyses describe the *shape* of a layer's geometry (how many directions it occupies, how quickly they decay (Li et al., 2025; Cancedda, 2024)), but not whether that shape is turned toward the model's own readout:

Does a layer's geometry align with the model's predictive readout, and when does it not?

Training leaves this open (related work in Appendix A). We introduce **PGA** (predictive–geometric alignment; §2) to measure it. The output layer responds strongly to some directions of the model's internal space and barely to others; PGA asks how much of a layer's distance structure (which contexts sit near which) survives when viewed through only the readout's preferred directions, compared with equally many random ones, and reports the excess as a calibrated score (Figure 1).

Contributions (1) We introduce PGA, a null-calibrated score of how much of a layer's distance structure lives in the readout's directions (§2). With it we identify a late, abrupt loss of readout alignment in narrow models, invisible to the training loss and to spectral measures. A single ratio, width to depth (d/L), tracks its severity, and the ordering holds on a held-out corpus and, at its extremes, under a change of training data (§3, §4). (2) We characterize the collapse: it hides predictive structure rather than destroying it, switches on late and abruptly in training, and falls hardest on predictions copyable from the context (§5). (3) We draw the consequence for readout-based interpretability: at misaligned layers a failed logit-lens reading reflects the readout, not missing knowledge, and a small trained reader restores the decode (§6).

2. Method

PGA asks of each layer: if only the output layer's preferred directions were visible, would the layer's arrangement of contexts survive better than chance?

2.1. Predictive-Geometric Alignment (PGA)

Readout subspace The output map amplifies some directions of hidden space strongly and others barely at all. The *readout subspace* $\mathcal{V}_k$ is the span of the k it amplifies most. Formally, let $W_U = U\Sigma V^\top$ be the singular value decomposition (SVD) of the unembedding matrix: the right singular vectors V live in hidden space, the singular values σ_i give the gain along each of them, and $\mathcal{V}_k$ is the span of the top-k right singular vectors. A vector orthogonal to $\mathcal{V}_k$ still reaches the logits (the pre-softmax scores $W_U h$ assigns to each candidate next token), but only through the weaker remaining directions, with gain bounded by σ_{k+1}. So $\mathcal{V}_k$ is W_U's *highest-leverage* subspace, not its exclusive channel: projecting states onto $\mathcal{V}_k$ changes predictions measurably (Appendix E.5). The cutoff k is a measurement choice, and we control for it directly (§2.3, Appendices D and D.3).

One subspace serves every layer. Each layer reads from and adds to a residual stream of the same width d, so a subspace defined once from W_U can be applied at any depth. Computing the metric at layer ℓ then asks whether that layer's geometry is already oriented toward the readout, much as the logit lens applies W_U to intermediate states.

Two clarifications. First, the readout is itself geometric: each logit is a dot product $\mathrm{logit}_i = W_U^{(i)} \cdot h$, so decoding one token is already a distance-like (margin-to-hyperplane) operation. PGA does not re-ask that per-point decodability question (which the logit lens covers); it asks the *structural* one: whether the *pairwise* distance geometry of many hidden states survives when only the readout directions $\mathcal{V}_k$ are visible. Second, $\mathcal{V}_k$ is a *functional* subspace, not a variance one: it is fixed by W_U and need not align with the directions of greatest variance, unlike PCA of the residual stream, whose leading directions are the high-variance "background" we *remove* as anisotropy (§2.2).

Distance preservation We compare hidden states by *cosine* rather than Euclidean distance: for transformer representations the meaningful axis of comparison is direction, not magnitude: residual-stream norms are dominated by a few outlier dimensions, and the final LayerNorm normalizes magnitude before the readout acts (Ethayarajh, 2019). For a layer ℓ and n contexts, we take the anisotropy-corrected last-token hidden states $\{\tilde{h}_i\}_{i=1}^n$ (§2.2) and compute pairwise cosine distances in the full hidden space and in the readout subspace, over all context pairs $1 \le i < j \le n$:

$$d_{ij}^{\text{full}} = 1 - \frac{\tilde{h}_i \cdot \tilde{h}_j}{\|\tilde{h}_i\|\|\tilde{h}_j\|}, \quad d_{ij}^{\text{read}} = 1 - \frac{(P_k\tilde{h}_i)\cdot(P_k\tilde{h}_j)}{\|P_k\tilde{h}_i\|\|P_k\tilde{h}_j\|}, \tag{1}$$

where P_k projects onto $\mathcal{V}_k$. The full distance is how far apart hidden states are in the original geometry; the projected distance is how far apart they are when only the readout directions are visible. The *readout correlation* is the Spearman rank correlation between the two,

$$\rho_{\text{read}}^{(\ell)} = \rho_{\text{Spearman}}(\mathbf{d}^{\text{full}}, \mathbf{d}^{\text{read}}), \tag{2}$$

and measures how much of the layer's distance structure survives when everything outside $\mathcal{V}_k$ is discarded. We use rank correlation to stay insensitive to the anisotropic scale of contextual distances; the conclusions are unchanged if Pearson correlation is used instead (Appendix D.4).

Random null and the z-score ρ_{read} alone is hard to interpret: any sufficiently high-dimensional projection preserves *some* structure. To isolate the readout's contribution we com-

pare it against *random lenses*: subspaces of the same dimension as $\mathcal{V}_k$ but with no special orientation. Concretely we draw 100 *uniformly random* orthonormal k-dimensional subspaces (i.i.d. Gaussian $d \times k$ matrices, QR-orthonormalized). Each such lens is the same size as the readout subspace but points nowhere in particular; projecting the hidden states through it gives a correlation ρ_{rand}, and the 100 draws form a null with mean μ_{null} and standard deviation σ_{null}. PGA is the resulting z-score:

$$\text{PGA}^{(\ell)} = z^{(\ell)} = \frac{\rho_{\text{read}}^{(\ell)} - \mu_{\text{null}}^{(\ell)}}{\sigma_{\text{null}}^{(\ell)}}. \tag{3}$$

z measures how privileged the readout directions are over arbitrary k-dimensional subspaces: $z \gg 0$, the layer's distance structure concentrates in the readout directions; $z \approx 0$, the readout is no better than chance; $z < 0$, random subspaces preserve more structure than the readout, which has lost its privileged status. The construction parallels representational similarity analysis (RSA; Kriegeskorte et al., 2008), with one compared representation replaced by a functionally defined subspace and the other by a dimensionality-matched random null. z is a standard score ($z = 10$: the readout is ten standard deviations more distance-preserving than the null), but because the null is estimated from 100 Monte-Carlo draws of a bounded statistic and its variance can be small, we read large $|z|$ as "decisively more (or less) structure-preserving than random" rather than as a calibrated tail probability; absolute ρ_{read} values are in Appendix D.10.

2.2. Anisotropy Correction

Transformer hidden states are anisotropic (Ethayarajh, 2019): a single dominant direction in the residual stream carries most of the variance and inflates the cosine similarity between otherwise unrelated states. This direction biases z, because the readout subspace and a random subspace generally capture *different amounts* of it, and each subspace's share of it then dominates the pairwise-distance correlations: measured on raw states, z would conflate genuine readout alignment with overlap with the anisotropy direction. The correction is complementary to, not redundant with, LayerNorm: LayerNorm rescales each token's vector individually but leaves in place the direction *shared* across tokens, exactly the component that inflates pairwise cosine similarity.

We remove this confound with *Common Component Removal* (CCR-k) (Mu et al., 2018), defined generally as: mean-center the hidden states, then project out the top k principal components of the centered hidden-state matrix. Writing μ for the mean hidden state and $v_1, \ldots, v_k$ for those top-k directions (the leading right singular vectors of the centered matrix),

$$\bar{h} = h - \mu, \qquad \tilde{h} = \bar{h} - \sum_{i=1}^{k} (\bar{h} \cdot v_i)\, v_i.$$

CCR-k strips these shared background directions so the metric is computed on the residual, content-bearing geometry. The variant is named by k: *CCR-1* (our default) removes only the single dominant direction; *CCR-5* and *CCR-10* remove the top 5 and 10, which we use to test whether a collapsed layer's predictive structure is merely *masked* (§5.1). After CCR-1, pairwise isotropy rises to near-ceiling values in the mid-to-late layers (values in Appendix D.1).

CCR-1 itself could in principle remove readout-aligned variance and bias z in the opposite direction. Three observations bound this concern (full values in Appendix D.2): the CCR-1 direction lies mostly outside the readout subspace at interior layers; the overlap *inverts* across depth (smallest exactly where z goes negative, largest at the final layer where z is most positive, the opposite of what a stripping artifact predicts); and, with mean-centering only (CCR-0, no PC removal), the collapse is *deeper*, not absent, so CCR-1 attenuates rather than manufactures the effect.

2.3. Experimental Setup

Models Our primary suite is the seven Pythia models from 70M to 6.9B (Biderman et al., 2023), all trained on the same data (The Pile) with the same tokenizer (GPT-NeoX), so the models differ principally in capacity (full d/L in Table 2). Pythia uses *untied* embeddings, which matters here: tying makes layer-0 states identical to the input embeddings, trivially inflating early-layer z. The 1B / 1.4B pair shares width but differs in depth, isolating depth's contribution. For cross-family validation we use OLMo-1B (Groeneveld et al., 2024), Phi-1.5 (Li et al., 2023b), and Gemma-2-2B (Gemma Team, 2024), spanning both tied and untied readouts.

Data We sample 1,000 contexts from OpenWebText with at least 64 words, truncated to 512 tokens; the same texts are used for every model, so hidden-state samples differ only in the model that produced them. Each context contributes one last-token state, so the uncertainty in z is governed by the number of contexts, not by the $\sim 5 \times 10^5$ pairwise distances derived from them; z is stable across sample sizes varying by more than an order of magnitude, with tight bootstrap confidence intervals across models (Appendices D.12, D.11).

Computation For each model and layer, we extract the last-token hidden state of each context, apply mean + CCR-1 correction, and compute PGA with $k = 100$ readout dimensions and 100 random subspaces. For five of these models (Pythia-160M/410M/1B/2.8B and OLMo-1B) we

also compute six spectral metrics (RankMe, stable rank, participation ratio, α-ReQ, isotropy, condition number) for Appendix E.4. Final-layer states are taken post-LayerNorm, matching the geometry W_U operates on.

Readout coverage A fixed k=100 means different things at different widths: the same hundred directions are a larger share of a narrow model's readout than of a wide one's. That share sets the null's variance and hence the *size* of z (Appendix D.9), so z magnitudes from models of different width do not sit on a common scale, and our claims are built to avoid comparing them. What we read off z within a model is its *sign* along the depth trajectory: the narrow-model late-layer collapse is the readout becoming worse than random, and that sign is unchanged at every k tested in the three models swept (Appendix D). The one cross-model statement we make is ordinal, the d/L ordering of interior minima (§4). Both survive a constant-coverage control that gives every model the same k/d (k=$d/10$), so neither is an artifact of unequal coverage. Wide models are the case where coverage genuinely matters: the sign of their *interior minimum* (the lowest z over a model's non-embedding, non-final layers) can flip as coverage changes, and at matched coverage their behavior is modulated by depth instead. For those models we therefore report the coverage alongside the result rather than calling a collapse present or absent outright (Appendix D.3).

3. Narrow Models Lose Readout Alignment at Late Layers

The first result is positive: every model tested has layers arranged along the directions the output layer reads, far more so than chance (Figure 2a; Table 2; per-layer values in Appendix C.2), consistent with theory relating next-token training to readout-aligned representations (Crutchfield & Young, 1989; Zhao et al., 2024). How far that alignment extends varies with capacity: in the wider models most layers sit above chance; in the narrowest only a handful do, at scattered early layers and at the output.

Late-layer collapse In Pythia-160M and 410M (and, more shallowly, in 70M; Table 2), predictive alignment collapses at late-but-not-final layers (Figure 2a, shaded). What is striking is the regularity: within the Pythia suite every model that collapses recovers at the very last layer, and the collapse always sits in the second half of the network. Gemma-2-2B is the exception in the other direction, collapsing *at* the final layer rather than before it; that is a separate, family-specific effect (§7). Low z at late intermediate layers, positive again at the output: we call this trajectory a *detour*.

Invisible to loss and to spectral metrics The collapse leaves no fingerprint on the training objective (§5.2). Nor do spectral metrics surface it: in the deepest collapse zone the spectral profile is indistinguishable from surrounding layers while z swings by tens of standard deviations, and no spectral metric we tested consistently tracks z across models (Figure 11; Appendix E.4).

Robustness Beyond the sampling and coverage controls of §2.3, the collapse replicates on both held-out corpora for Pythia-160M and deepens on code for both narrow models, though Pythia-410M flattens to chance on WikiText; it recurs in every *independent training re-init* we ran; orthogonal PGA finds no above-chance alignment outside $\mathcal{V}_k$; and replacing W_U with the input embedding W_E yields chance-level z, so the alignment is specific to the predictive interface. Full values for all of these controls are in Appendix D.

4. Collapse Severity Follows the Width-to-Depth Ratio

How severe the collapse becomes follows not a model's size but its shape: how much width it has per layer of depth. A single variable, the width-to-depth ratio d/L, orders collapse severity across the suite (Figure 2b).

Width, depth, and parameter count all rise together across the Pythia suite, so no single-model comparison can separate them; matched pairs can. At *fixed depth*, the narrower Pythia-410M collapses deeply while the wider Pythia-1.4B only grazes zero, so width helps. At *fixed width*, the shallower Pythia-1B stays positive while the deeper Pythia-1.4B crosses to negative interior z, so depth hurts. Only one matched pair is available in each direction, but both point the same way, and the from-scratch sweep below addresses this directly. The operative variable is therefore width *relative to* depth.

We read this ratio as *per-layer capacity*: the room a layer has left to keep a separate readout-aligned set of directions after hosting the dominant variance direction it shares with every token (§5.1). Because the ratio is set within one training recipe, it is a controlled variable, not a confound; and the apparent width boundary at default coverage dissolves into a continuum once coverage is matched.

Two out-of-sample tests probe the ordering. On Python code (a held-out corpus: same models, new domain), the interior minima of all six Pythia models measured are near-perfectly rank-ordered by d/L (tie-corrected r_s=0.99, n=6; ties and margins in Appendix D.5), and the dip/no-dip directions for the two models added to that run were stated before it ran and landed as predicted. A fully pre-registered test on the deduped-Pythia suite (same shapes, different training data) confirms the ordering only at its extremes. Five of the seven directional calls landed: the deep-collapse and

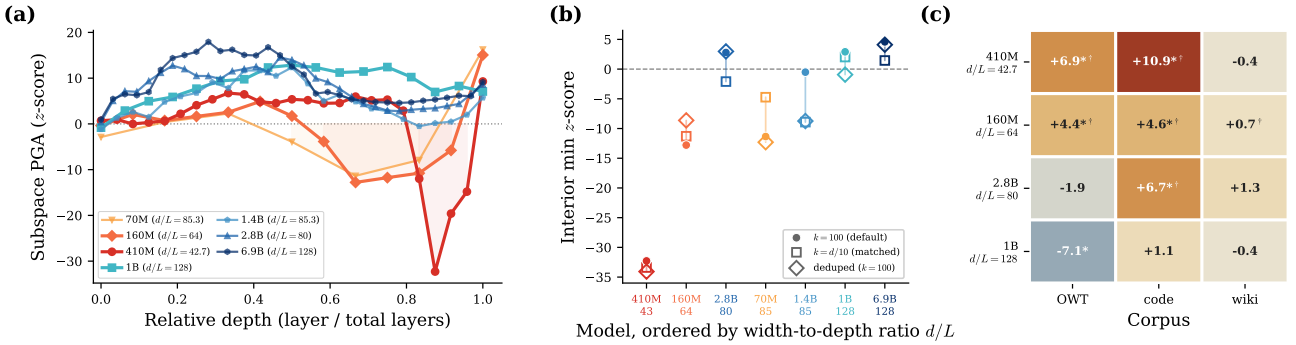

Figure 2. **The phenomenon: a late-layer collapse of readout alignment, indexed by d/L.** The panels read left to right as claim, index, and content. **(a)** PGA z against relative depth, one curve per Pythia model. The narrow models (warm) fall below zero in the last third of the network and climb back at the output layer. The wide ones (cool) show no such late collapse; 1.4B comes closest, grazing zero at 83% depth, and only the embedding layer of 1B and 1.4B sits marginally below. **(b)** Each model's worst interior z, models laid out in order of width-to-depth ratio, with three readings joined per model: default readout coverage (filled circle), matched coverage (open square), and the pre-registered deduped suite (open diamond). Severity broadly rises as d/L falls, though 2.8B and 70M are inverted, which is why the rank agreement is r_s=0.87 rather than 1. **(c)** How much deeper the collapse falls on contexts whose next token is copyable, per (model, corpus) cell, rows ordered by d/L. A significantly positive gap appears only in cells that themselves collapse (†), never in one that does not, while the converse fails: collapsing is necessary for the selectivity rather than sufficient.

stable models reproduced their standard twins almost point-for-point, while both misses were borderline models that shifted toward collapse, and the pre-registered rank falsifier fired (r_s=0.67, below the 0.8 threshold we set in advance). The extremes are thus training-data-invariant; the binary boundary and the middle ranks are not (Figure 2b, open diamonds; Appendix D.6).

A from-scratch sweep varies width and depth on purpose, isolating d/L from the confounds of the pretrained suite. At the scale we could afford it does *not* reproduce the full detour, so it is a directional datum rather than a causal demonstration (§7.3). What it does show is a sign that follows d/L: among the cells trained to at least 275k steps the final interior minimum rises monotonically with d/L, the two cells at d/L=32 ending clearly negative despite their different widths and depths, the d/L=64 cell grazing zero, and the d/L=128 cell ending positive (Appendix D.8).

5. Masking, Late Onset, and Content Selectivity

Three properties characterize the collapse: the structure it hides survives underneath it, it arrives late and abruptly in training, and it does not affect all predictions alike.

5.1. Masked, Not Erased

When a layer's structure no longer registers through the readout, the structure itself is not gone: it is hidden. In the affected layers, the readout correlation ρ_{read} stays moderate, so distance information has not vanished from $\mathcal{V}_k$. The random-subspace correlation μ_{null}, however, rises toward ceiling: an arbitrary k-dimensional subspace now captures the layer's distance structure as well as the readout does (per-layer values in Table 11); the readout's privileged status disappears, and the predictive structure is not destroyed but *overwhelmed* (Figure 4a; per-layer values in Appendix D.10).

To see why, consider the dominant variance directions. Following Cancedda (2024), we partition hidden space into the readout subspace $\mathcal{V}_k$ (the *bright* directions, where W_U's leverage concentrates) and its orthogonal complement (the *dark* directions, which W_U amplifies only weakly, through singular values below σ_k). Across training, the leading principal components of the (raw) residual stream rotate from bright into dark at late layers; Figure 5a,b (Appendix C.2) tracks PC1, measured on raw states. Since z is computed on CCR-1-corrected states, PC1 itself cannot drive the metric; what drives it is that the migration involves a *group* of dominant directions: in the collapse zone the surviving second component is even darker than the first (per-component dark-fractions in the caption of Figure 5), so the corrected stream is still dominated by dark directions. This is what turns z negative (Figure 5c). Once a dark direction dominates the layer's full-space distances, a random subspace picks up a share of it (about k/d of its variance) and so tracks the full-space structure, while the readout subspace, built to exclude the dark directions, does not. The random lens preserves more of the structure than the readout, and z falls below zero. This is also why removing one dominant direction (CCR-1) does not undo the collapse but removing the surviving dark group (CCR-5/10) does.

This accounts for how collapse looks but not why it happens only in small models. Pythia-1B undergoes comparable PC1 migration yet z stays positive (§5.2). The simplest explana-

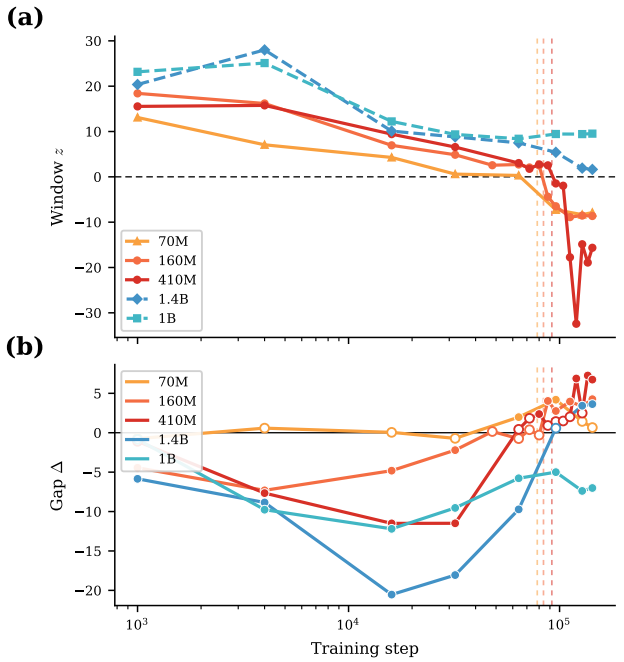

Figure 3. **The collapse switches on late in training, and only in the narrow models.** The two panels share the training-step axis; dashed guides mark where each narrow model's alignment crosses zero. **(a)** Readout alignment over training. The narrow models (warm) start well aligned and lose it within one late stretch of training; the wide models (cool) never do. **(b)** Over the same axis, how much deeper the collapse falls on contexts whose next token is copyable from the context. Filled markers are significant ($p<0.05$ against the $B=1000$ balanced-partition null), open markers are not. The three collapsing models (warm) turn significantly positive around each one's crossing. Pythia-1.4B (the borderline case) reverses too, though its window z only grazes zero rather than crossing, so the selectivity can flip on a marginal approach. The stable control 1B (cool) never reverses, staying negative throughout, so it is the reversal, not the negative early phase, that marks a collapse.

tion is capacity: small models lack enough dimensions to host a dominant variance direction for intermediate computation *and* a separate set of readout-aligned directions, so the two collide; large models have room for both. Removing the top few principal components turns the collapsing models' interior troughs positive, decisively so once the surviving dark group is removed (CCR-10; Figure 4b).

Masking has a direct functional signature: a *learned* per-layer affine readout (tuned lens; Belrose et al., 2023) repairs decoding straight through the collapse zones, cutting the zero-shot lens's error by roughly an order of magnitude in Pythia-160M's zone, while in Pythia-410M's zone the tuned-lens error sits on the same depth trend as the never-collapsing Pythia-1B (full numbers in Appendix D.15). Several independent lines converge on the same conclusion, *masking, not erasure*: the correlation decomposition above, restoration under CCR-5/10, tuned-lens repair, and two further controls in the appendix (linear probing stays flat; the readout subspace keeps its predictive-content correlation). The collapse degrades specifically the *zero-shot* W_U readout, not the predictive structure beneath: linear-probe accuracy stays high and flat ($\approx$0.85) straight through Pythia-410M's collapse zone (Appendix E.2), so z tracks W_U-alignment, not raw information content (§6, Appendix E.3).

5.2. Alignment Forms Early, Masking Emerges Late

The collapse is not built into the architecture: tracking PGA across the full range of public training checkpoints shows it emerges during training, and whether it emerges at all depends on the width-to-depth ratio (Figure 6 and Tables 3, 4, Appendix C.2); as with other internal transitions localized by progress measures (Nanda et al., 2023), the onset is abrupt rather than a smooth drift.

Default alignment All three checkpointed models (160M, 410M, 1B) align geometry with predictive directions within the first thousand steps (Tables 3, 4); readout-aligned geometry is what training establishes first.

Progressive masking In the small models ($d \leq 1024$), the dominant variance direction at late layers begins to drift away from the readout in the second half of training (Figure 6a,b), and z at those layers drops below zero. The timing coincides with the "compression-seeking" phase of Li et al. (2025): effective rank declines and a few directions dominate the residual stream; if they point away from $\mathcal{V}_k$, masking follows. The final layer, held in place by the cross-entropy loss, does not participate. Pythia-1B undergoes comparable PC migration, but develops no deep late-layer collapse (Figure 6c, Table 4), consistent with the capacity reading.

Loss does not track the collapse Validation loss descends smoothly and similarly across all three models; the deepest-collapsing model in fact ends training at slightly *lower* validation loss than the stable control (Figure 6d).

The content selectivity switches on with the collapse Tracking the content contrast of §5.3 across training localizes *when* it appears, in all three collapsing models, shown in the main figure (Figure 3b; full three-band version in Figure 7, Appendix C.2). What the three collapsing models share is the onset itself: in each, the copyable-deeper gap becomes significantly positive in the same late checkpoint window in which its window z crosses zero (80k→88k for 160M, 88k→96k for 410M, 64k→96k for 70M). What they do not share is the phase on either side: in 160M and 410M the gap is significantly *negative* early and stays positive to the end of training; in 70M it is indistinguishable from zero on either side of the crossing. Two further trajectories sharpen the picture. Pythia-1B, which never collapses, has a gap that is negative at every checkpoint and never reverses,

so a negative gap tracks nothing by itself and only the reversal co-times with collapse. Pythia-1.4B, the borderline model, reverses like the collapsing three even though its window z only grazes zero, so the onset strengthens continuously with proximity to collapse rather than switching on at a threshold. Its interior still dips, so this is the continuum behind the per-cell rule of §5.3, not a counterexample to it. The copy machinery does not change at the flip: the prefix-matching / in-context-learning score is already at its mature value at the earliest checkpoint measured, so the induction heads that serve copyable predictions (Olsson et al., 2022) are in place nearly two orders of magnitude earlier in training steps than the reorganization (formation itself predates our first checkpoint).

5.3. The Collapse Is Content-Selective

If the detour were a uniform side-effect of anisotropy, it should affect all predictions alike. It does not. We split the shared contexts by whether the model's own top-1 next token is *copyable from the context* (it already appears in the prefix, the population induction heads serve; Olsson et al., 2022) or *parametric* (it does not), subsample to equal size, and recompute the per-layer z for each: a within-model, within-layer contrast on matched sample sizes, so differences in model quality or in-context-learning ability cannot produce it.

Two controls discipline the comparison. First, the layer windows are *pre-registered* from independent data: for collapsing models, the interior layers where the flagship run has $z < 0$; for non-collapsing models, a fixed relative-depth band $[0.65, 0.95]$. The window-mean gap $\Delta = \bar{z}_{\text{parametric}} - \bar{z}_{\text{copyable}}$ is calibrated against $B{=}1000$ balanced random partitions of the same contexts. Second, because the copyable and parametric populations differ sharply in token frequency (Table 5), and token frequency shapes exactly the anisotropy structure z is sensitive to, we recompute Δ on *frequency-matched* subsets (quintile-binned equal-count subsampling; within-bin partition null of the same size). We run this on three corpora: OpenWebText, Python code, and a shuffled WikiText-103 sample (shuffled to avoid the topical homogeneity of first-N sampling).

Figure 2c reports the frequency-matched gap on all three corpora (exact values and p-values in Table 5, Appendix C.2). Because collapse itself is corpus-dependent (Table 9), the unit of analysis is the (model, corpus) *cell*, classified by whether that model's interior z dips below -2 at a late layer (relative depth ≥ 0.5) *on that corpus*. Under this per-cell reading the pattern is clean: **the frequency-matched gap is significantly positive only where the model collapses on that corpus**. On OpenWebText both collapsing models show the gap and neither stable model does. On code the same rule holds, including for Pythia-2.8B, which dips mildly on code and shows the gap there while showing none on OpenWebText, where it does not dip. The WikiText cells sharpen it into a strictly one-directional rule: three models do not collapse on shuffled WikiText and show no gap, but Pythia-160M *does* collapse there, inside its pre-registered window, with a null gap: collapse without selectivity. Across all twelve classifiable cells the rule is one-directional. **A significant copyable-deeper gap appears only where the model collapses on that corpus**, with no counterexample. The converse does not hold: collapse alone does not produce the gap, as the 160M–WikiText cell shows. Pythia-1B is not a counterexample to that rule, which asks only that a non-collapsing cell show no *positive* gap: 1B does not collapse and its gap is negative. That its OpenWebText gap is significantly negative rather than null, copyable-target geometry being *more* readout-aligned there, is a separate effect the rule does not address, and we leave it open. The selectivity is geometric, not behavioral: slicing each model's own token-level cross-entropy by the same copyability criterion and by position (pre-registered either-way framing) shows only a smooth scale trend, with no regime-specific anomaly (committed analysis in the repository), on both OpenWebText and code (extending the invisibility result of §5.2).

6. Consequences for Reading Intermediate Layers

The findings change what a failed reading through the output map means: sometimes the fault lies with the readout, not the layer.

What z licenses: interpreting readout-based *failures* z is a task-free, label-free diagnostic of whether the zero-shot readout retains privileged geometric status at a layer, and its practical content is about interpreting failure rather than licensing success. At a $z{<}0$ layer, a poor logit-lens or direct-logit-attribution reading (both readouts of intermediate states through W_U) is a *readout artifact*: the distance structure is present (probing stays high; Appendix E.2) and a learned per-layer reader recovers most of the decode (Appendix D.15). At a $z{>}0$ layer the same failure cannot be blamed on readout misalignment.

The diagnostic is deliberately narrow, and three boundaries keep it so. z is not a lens-usability score: logit-lens failure is universal at intermediate depths across scales, so positive z does not make the zero-shot lens accurate. It does not track the *magnitude* of the tuned-versus-zero-shot repair gap, with which its rank correlation is approximately zero. And it is not a general probe-layer-selection score: across the suite and three tasks, simple depth-monotonic signals predict linear-probe accuracy as well or better (Appendices E.1, D.15, E.3). Its value is this specificity to W_U-alignment rather than to generic layer quality.

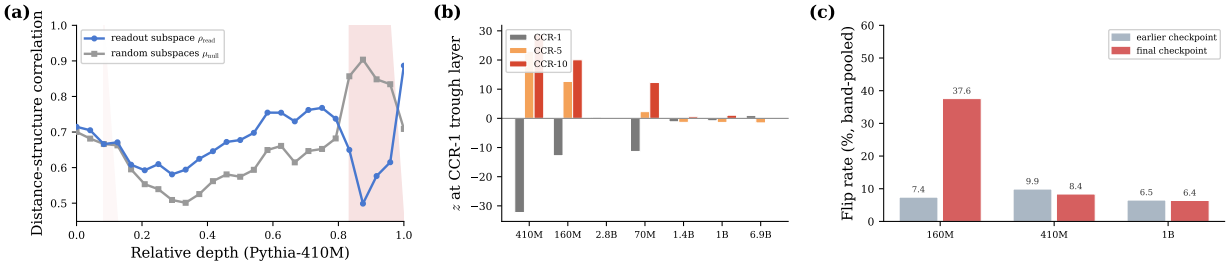

Figure 4. **The structure is masked rather than erased, and where the masking becomes functional.** **(a)** What a negative z is made of, for Pythia-410M. The strongly shaded late band is the collapse zone (deep negative z); the faint early mark is a single layer that dips just below zero. Across the collapse zone the readout-subspace correlation stays moderate while the random-subspace mean rises past it: the readout loses its privileged status, but the distance structure itself does not vanish. **(b)** z at each model's trough layer as more dominant directions are removed (CCR-1 baseline in grey, CCR-5/10 in warm tones), models ordered by d/L. The collapsing models' troughs (large grey bars) turn positive once a few directions are stripped, so the predictive structure was present but overwhelmed; the wide models' troughs sit at the near-zero embedding layer and barely move. **(c)** The functional consequence, pre-registered: how often the final top-1 token is missing from the zero-shot lens but recovered by a trained one, at the pre-onset and final checkpoints. Pythia-160M rises sharply; the stable control 1B and the deepest collapser 410M stay flat.

At the narrow end, the failure is functional and timed to the onset A pre-registered companion analysis (Appendix G) measures what the collapse does to W_U-based reading in practice via the *flip rate*: the fraction of contexts whose final top-1 token is absent from the zero-shot lens top-5 but present in the tuned-lens top-5, pooled over each model's pre-registered late-layer band. The flip rate rises sharply from the pre-onset checkpoint to the end of training in Pythia-160M, while the stable control stays flat (Figure 4c); the same onset-timed rise holds for 70M as a pre-registered out-of-sample call (Table 17). The deepest-collapsing model, Pythia-410M, is functionally *silent*. The dissociation tracks W_U's spectrum: functional failure appears only in the models whose unembedding is nearly rank-1 (70M/160M), while 410M's much flatter spectrum leaves directions below the readout cutoff carrying the decode (Appendix D.9). Spectrum and size are confounded in public models, however, so the interaction is a measured observation, not an isolated cause.

Layer selection in small models Collapse-zone layers remain usable for probing, but the zero-shot readout there has lost its geometric privilege: W_U-based readings should be treated as readout-limited rather than content-limited, and layer selection should account for geometric orientation, not just decodability.

7. Discussion

7.1. Two Geometric Regimes on a Capacity Continuum

Taken together, the results (§3–§5) sort language models into two regimes ordered by representational *capacity* in the sense of §4. In the **direct** regime (high d/L), geometry stays readout-aligned across essentially all intermediate layers; in the **detour** regime (low d/L), late layers develop a dominant variance direction pointing away from the readout subspace, and the final layer is realigned with it. The labels are operational (the sign of the interior minimum at a stated coverage), and the boundary is a d/L continuum, not a width threshold (§4).

Three observations point to capacity, not architecture, as the axis that orders the collapse. (i) PC1 migrates comparably across scales (§5.2, Figure 6c): what differs is whether the model has enough remaining dimensions to absorb it. (ii) Depth contributes alongside width (§4): more layers give anisotropy more room to accumulate, and enough depth can tip even a wide model into the detour. "Capacity" thus admits two non-exclusive readings: an instantaneous *per-layer budget*, under which severity should follow d/L regardless of how width and depth combine, and *depth-wise accumulation*, under which deeper models should collapse more at fixed width. The two d/L=32 cells of the from-scratch sweep, which reach a similar collapse depth from different widths and layer counts, favor the per-layer-budget reading, but our observational data do not fully separate the two. (iii) KL projection of late-layer hidden states onto $\mathcal{V}_k$ shows the collapse zone preserves task-relevant information (Appendix E.5), so the detour reorganizes geometry without changing what is computed. Why capacity acts as sharply as it does at default coverage remains open; a theory of how a model allocates residual dimensions between a dominant variance direction and a readout-aligned set is a natural next step, and might also explain the family-specific final-layer collapse of Appendix F.

7.2. Cross-Architecture Validation

Three cross-family models, plus an extended nine-model sweep (Appendix F), test whether the regime structure depends on Pythia-specific choices. Phi-1.5 (untied) shows no late-layer loss, consistent with wide-model stability generalizing beyond Pythia; its z-scores are on average some-

what stronger than Pythia-1.4B's at matching width (Appendix C.1), suggesting factors beyond capacity may also matter. OLMo-1B (tied) shows no loss at any layer. Gemma-2-2B shows no *intermediate*-layer loss, but its *final* layer collapses: a distinct, family-specific phenomenon (also in Qwen-2.5; absent in Llama-2/3.2, OLMo, TinyLlama) that is *not* accounted for by embedding tying, width, or vocabulary size (counterexamples in Appendix F); we leave its mechanism to future work.

7.3. Limitations

Metric scope PGA assumes a linear readout. It is measured in the native residual basis, which is not an arbitrary gauge choice: z is provably invariant to the orthogonal reparameterizations that preserve both the logits and the cosine metric (Appendix D.7). Features routed nonlinearly before W_U may not be captured; a per-layer affine does eliminate the collapse (Appendix D.15), scoping the phenomenon precisely to the zero-shot readout. Tied embeddings inflate early-layer alignment trivially. SSMs and MoEs use W_U differently and would need a separate analysis.

CCR sensitivity Negative z-scores disappear under CCR-5. We read this as informative rather than as an artifact: the collapse is driven by a few dominant masking dimensions, not an absence of predictive structure; the choice of correction strength determines the object of study.

Generalization The width-driven *intermediate*-layer detour is established within the controlled Pythia suite, and the public model pool holds no untied non-Pythia model narrow enough ($d < 2048$) to re-test the narrow regime across families. What does generalize across families is the wide-model side: intermediate geometry stays readout-aligned and wide models avoid the detour (§7.2). We test up to 6.9B parameters (frontier-scale untested); probing uses one task (AG News), and measurements come from final-token hidden states only (Valeriani et al., 2023).

Causality The evidence is correlational but robust: d/L reliably orders readout misalignment within the Pythia suite and on a held-out corpus (§4), and a learned per-layer reader repairs the decode through the collapse zone (Appendix D.15). We do not establish a causal mechanism. Three stronger interventional tests returned null (KL-asymmetry under projection onto/off $\mathcal{V}_k$, linear recoverability of the model's own prediction from $\mathcal{V}_k$, and a freeze of the readout-aligned bright components through the collapse window; committed analyses in the repository), and the from-scratch sweep does not reproduce the full detour signature at the scale we could train (Appendix D.8); each bounds a causal reading of z, none the observational ordering. The capacity account therefore rests on the pretrained suite together with its held-out-corpus and deduped-suite predictions, with the from-scratch sweep supplying a directional sign pattern rather than a demonstration. Because z measures W_U-alignment, we make no claim that the collapse degrades the model's computation.

8. Conclusion

PGA asks whether a layer's geometry is aligned with the model's own readout, a question the shape-only spectral metrics do not address. The answer is yes, but by an amount that follows per-layer capacity (d/L) rather than parameter count. In the narrow models this alignment collapses at late layers, late in training, even as the loss keeps improving. The geometry is masked, not erased: the reorganization is late, abrupt, invisible to the training objective, and selective in what it affects. A model's width-to-depth ratio therefore tracks not only its loss but how far its intermediate geometry stays readable through W_U, which matters for every interpretability tool that reads intermediate layers that way.

Acknowledgements

The author gratefully acknowledges the EPFL School of Computer and Communication Sciences (IC) for financial support enabling attendance at ICML 2026 and the presentation of this work.

References

Arefin, M. R., Subbaraj, G., Gontier, N., LeCun, Y., Rish, I., Shwartz-Ziv, R., and Pal, C. Seq-VCR: Preventing collapse in intermediate transformer representations for enhanced reasoning. In *International Conference on Learning Representations*, 2025. arXiv:2411.02344.

Bansal, Y., Nakkiran, P., and Barak, B. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Belrose, N., Ostrovsky, I., McKinney, L., Furman, Z., Smith, L., Halawi, D., Biderman, S., and Steinhardt, J. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Biderman, S., Schoelkopf, H., Anthony, Q., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430, 2023.

Cancedda, N. Spectral filters, dark signals, and attention sinks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

Crutchfield, J. P. and Young, K. Inferring statistical complexity. *Physical Review Letters*, 63(2):105–108, 1989.

Dar, G., Geva, M., Gupta, A., and Berant, J. Analyzing transformers in embedding space. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023.

Delétang, G., Ruoss, A., Duquenne, P.-A., Catt, E., Genewein, T., Mattern, C., Grau-Moya, J., Li Kevin Wenliang, Aitchison, M., Orseau, L., Hutter, M., and Veness, J. Language modeling is compression. In *International Conference on Learning Representations*, 2024.

Dong, Y., Cordonnier, J.-B., and Loukas, A. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning*, pp. 2793–2803, 2021.

Ethayarajh, K. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Empirical Methods in Natural Language Processing*, pp. 55–65, 2019.

Fahim, M. M. M. and Karim, M. R. The depth delusion: Why transformers should be wider, not deeper. *arXiv preprint arXiv:2601.20994*, 2026.

Gao, J., He, D., Tan, X., Qin, T., Wang, L., and Liu, T.-Y. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations*, 2019.

Gärdenfors, P. *Conceptual Spaces: The Geometry of Thought*. MIT Press, 2000.

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Geva, M., Caciularu, A., Wang, K. R., and Goldberg, Y. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

Ghandeharioun, A., Caciularu, A., Pearce, A., Dixon, L., and Geva, M. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *International Conference on Machine Learning (ICML)*, 2024.

Groeneveld, D., Beltagy, I., Walsh, P., Bhagia, A., Kinney, R., Tafjord, O., et al. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024.

Gröger, F., Wen, S., and Brbić, M. Revisiting the platonic representation hypothesis: An aristotelian view. *arXiv preprint arXiv:2602.14486*, 2026.

Gurnee, W. and Tegmark, M. Language models represent space and time. In *International Conference on Learning Representations*, 2024.

Huh, M., Cheung, B., Wang, T., and Isola, P. Position: The platonic representation hypothesis. In *International Conference on Machine Learning*, volume 235, pp. 20617–20642, 2024.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Kriegeskorte, N., Mur, M., and Bandettini, P. A. Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008.

Kulkarni, A., Springer, J. M., Subramonian, A., and Swayamdipta, S. Disentangling geometry, performance, and training in language models. *arXiv preprint arXiv:2602.20433*, 2026.

Lee, A., Weber, M., Viégas, F., and Wattenberg, M. Shared global and local geometry of language model embeddings. In *Conference on Language Modeling*, 2025. Outstanding Paper Award.

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., and Wattenberg, M. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *International Conference on Learning Representations*, 2023a.

Li, M. Z., Agrawal, K. K., Ghosh, A., Teru, K. K., Santoro, A., Lajoie, G., and Richards, B. A. Tracing the representation geometry of language models from pretraining to post-training. In *Advances in Neural Information Processing Systems*, 2025. arXiv:2509.23024.

Li, Y., Bubeck, S., Eldan, R., Del Giorno, A., Gunasekar, S., and Lee, Y. T. Textbooks are all you need II: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*, 2023b.

Marks, S. and Tegmark, M. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Conference on Language Modeling*, 2024.

Masarczyk, W., Ostaszewski, M., Imani, E., Pascanu, R., Miłoś, P., and Trzciński, T. The tunnel effect: Building data representations in deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Miller, J., Chughtai, B., and Saunders, W. Transformer circuit faithfulness metrics are not robust. In *Conference on Language Modeling (COLM)*, 2024.

Mu, J., Bhat, S., and Viswanath, P. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*, 2018.

Nanda, N., Chan, L., Lieberum, T., Smith, J., and Steinhardt, J. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations (ICLR)*, 2023.

nostalgebraist. interpreting GPT: the logit lens. LessWrong, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens. Accessed: 2026-03-30.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning*, volume 235, pp. 39643–39666, 2024.

Park, K., Choe, Y. J., Jiang, Y., and Veitch, V. The geometry of categorical and hierarchical concepts in large language models. In *ICLR*, 2025. Oral; arXiv:2406.01506.

Saurez, A., Lee, Y., and Har, D. Why linear interpretability works: Invariant subspaces as a result of architectural constraints. *arXiv preprint arXiv:2602.09783*, 2026.

Shai, A., Teixeira, L., Gietelink Oldenziel, A., Marzen, S. E., and Riechers, P. M. Transformers represent belief state geometry in their residual stream. In *Advances in Neural Information Processing Systems*, 2024.

Shepard, R. N. Toward a universal law of generalization for psychological science. *Science*, 237(4820):1317–1323, 1987.

Shwartz-Ziv, R. and Tishby, N. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Skean, O., Arefin, M. R., Zhao, D., Patel, N., Naghiyev, J., LeCun, Y., and Shwartz-Ziv, R. Layer by layer: Uncovering hidden representations in language models. In *International Conference on Machine Learning*, 2025. Oral; arXiv:2502.02013.

Tay, Y., Dehghani, M., Rao, J., Fedus, W., Abnar, S., Chung, H. W., Narang, S., Yogatama, D., Vaswani, A., and Metzler, D. Scale efficiently: Insights from pre-training and fine-tuning transformers. In *International Conference on Learning Representations (ICLR)*, 2022.

Tishby, N., Pereira, F. C., and Bialek, W. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, pp. 368–377, 1999.

Valeriani, L., Doimo, D., Cuturello, F., Laio, A., Ansuini, A., and Cazzaniga, A. The geometry of hidden representations of large transformer models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Zhao, Y., Behnia, T., Vakilian, V., and Thrampoulidis, C. Implicit geometry of next-token prediction: From language sparsity patterns to model representations. In *Conference on Language Modeling*, 2024. arXiv:2408.15417.

Reproducibility Statement

All models are publicly available on HuggingFace (Pythia suite from EleutherAI, OLMo-1B from AI2, Phi-1.5 from Microsoft, Gemma-2-2B from Google). Evaluation uses N=1,000 OpenWebText contexts ($\geq$ 64 tokens, truncated to 512), readout subspace dimension k=100, and 100 random-subspace draws per layer. Anisotropy correction is mean-centering followed by CCR-1. Statistical calibration uses the 100-draw random-subspace null behind every z (with bootstrap confidence intervals, Appendix D.11) and empirical balanced-partition permutation nulls (B=1000) for the subset contrasts. Results are stable across random seeds and sample sizes (Appendix D.12). Total compute: $\sim$60–80 A100 GPU-hours. Code will be released upon acceptance.

A. Related Work

The premise that geometry carries content is well grounded: what a representation encodes is shaped by where it sits relative to others (Shepard, 1987; Gärdenfors, 2000); semantic properties appear as linear directions (Marks & Tegmark, 2024; Gurnee & Tegmark, 2024; Li et al., 2023a); and concepts form systematic structures (Park et al., 2025) that converge across architectures (Huh et al., 2024).

Our question (whether each layer's geometry is oriented toward the predictive readout, and how this changes with scale) intersects several lines of work, each of which characterizes representation geometry in a way adjacent to, but distinct from, predictive alignment. For each, we state what it establishes and the specific gap that PGA fills.

Predictive Geometry Computational mechanics (Crutchfield & Young, 1989) and ϵ-machines predict grouping by conditional futures. Shai et al. (2024) confirmed this in HMM-trained transformers; Zhao et al. (2024) proved a formal analogue for NTP. The information bottleneck (Tishby et al., 1999; Shwartz-Ziv & Tishby, 2017) and compression–prediction equivalence (Delétang et al., 2024) suggest compression toward predictive-relevant directions. Saurez et al. (2026) prove that any feature decoded through a linear interface must occupy an invariant linear subspace of that interface, grounding why readout alignment should exist. These works imply predictive alignment but provide no per-layer metric for its degree.

Spectral Geometry Li et al. (2025) tracked RankMe and α-ReQ across training, identifying geometric phases (warmup, entropy-seeking, compression-seeking). The Platonic Representation Hypothesis (Huh et al., 2024) argues that representations converge across models and modalities; Gröger et al. (2026) caution that raw representational similarity is confounded by depth and width and shrinks under a null-calibrated comparison, the same correction our dimensionality-matched random null embodies. Lee et al. (2025) showed that embedding geometry has shared global and local structure. These works describe the shape of representation geometry; PGA asks whether that shape is oriented toward the model's predictive task.

Intermediate Layer Analysis The logit lens (nostalgebraist, 2020), tuned lens (Belrose et al., 2023), and Patchscopes (Ghandeharioun et al., 2024) decode intermediate hidden states by reading them through the vocabulary space (Geva et al., 2022); the tuned lens requires learned affine corrections, demonstrating "representational drift." Skean et al. (2025) showed that intermediate layers outperform the final layer on downstream tasks by up to 16%. The logit lens measures absolute decodability (does W_U produce the right token?); PGA measures geometric alignment (is the distance structure aligned with W_U?). Logit lens failure occurs at intermediate layers in *all* models, while loss of predictive alignment is specific to small models: these are distinct phenomena (Appendix E.1). That W_U-based readouts carry a model-dependent validity condition is complementary to broader evidence that interpretability measurements are sensitive to methodological choices (Miller et al., 2024).

Representation Collapse and Degeneration Dong et al. (2021) showed self-attention loses rank doubly exponentially with depth; Gao et al. (2019) identified representation degeneration; Arefin et al. (2025) showed that preventing intermediate-layer collapse improves reasoning. The loss of predictive alignment we observe is distinct (§5.1): representations retain substantial effective-rank-based diversity, and the phenomenon tracks model shape rather than size, and is not a pathology requiring regularization. It also differs from the depth-induced *tunnel effect* (Masarczyk et al., 2023), in which later layers compress representations; our detour is late-*training* and readout-specific: information is preserved and only its alignment with W_U is lost.

Unembedding Subspace Analysis Cancedda (2024) used SVD of the unembedding matrix to partition logit outputs into spectral bands, revealing a "dark subspace" of tail singular vectors that drives attention sink behavior. Dar et al. (2023) treat

the embedding space as a universal reference frame for interpreting transformer computations. PGA builds on this intuition but asks a different question: whether the *distance structure* of hidden states is aligned with readout directions, not whether specific computations occur in particular subspaces. The dimensionality-matched random-subspace null is what reveals the capacity-dependent loss of alignment. Kulkarni et al. (2026) study the effective rank of W_U across 108 models, finding spectral properties alone insufficient to predict performance, consistent with our finding that spectral metrics do not capture predictive alignment.

Width, Depth, and Aspect Ratio The width-to-depth ratio is Kaplan's *aspect ratio* (Kaplan et al., 2020), over which test loss is nearly invariant across a broad range; Tay et al. (2022) find that architectural shape, not parameter count, governs downstream transfer, and Fahim & Karim (2026) argue transformers should be wider relative to depth. These results concern loss and performance; we show the same aspect-ratio axis governs an *internal* geometric property (whether a layer's distance geometry stays aligned with its own readout) inside the regime scaling laws treat as shape-invariant.

B. Predictive Structure Concentrates in Readout Directions

Predictive structure concentrates in readout-aligned directions. Orthogonal PGA, computed in the complement of the readout subspace, never exceeds the 95th percentile of dimensionality-matched random subspaces at any model×layer combination (0/12; Table 1). In Pythia-410M, the orthogonal complement holds most of the representation variance (72–89% across depths; 72% at the final layer) yet carries no above-chance predictive alignment. The readout subspace also has the lower intrinsic dimensionality ($\mathrm{ID}_{\mathrm{readout}} = 28.5$ vs. $\mathrm{ID}_{\mathrm{ortho}} = 51.7$ at the final layer, TwoNN), consistent with compression into a lower-dimensional predictive manifold.

Intermediate-layer predictions match final predictions only in the last layers: intermediate-to-final prediction correlation $\rho < 0.5$ through ≈80% depth, then rises sharply to $\rho \approx 1.0$. The timing of this convergence overlaps with the collapse zone in small models.

Table 1. **Orthogonal PGA.** Computed within the complement of the readout subspace, predictive alignment never exceeds the 95th percentile of random draws (0/12). The geometry in the complementary space carries no above-chance alignment with the readout. All values: k=100, N=1000 contexts, 300 random subspaces (a larger null than the main protocol's 100, for a tighter 95th-percentile estimate), CCR-1 applied (`exp16`).

MODEL	DEPTH	ORTHO VAR%	ORTHO ρ	NULL P95	>?
P-160M	25%	85.0	0.350	0.365	✗
	50%	83.3	0.311	0.338	✗
	75%	82.4	0.219	0.255	✗
	100%	69.9	0.109	0.242	✗
P-410M	25%	88.7	0.058	0.064	✗
	50%	87.6	0.122	0.127	✗
	75%	82.0	0.281	0.312	✗
	100%	72.4	0.164	0.193	✗
P-1B	25%	94.0	0.131	0.142	✗
	50%	92.9	0.093	0.106	✗
	75%	86.4	0.186	0.215	✗
	100%	78.2	0.236	0.241	✗

C. Extended Quantitative Results

C.1. Model-Specific Geometric Profiles

Pythia-2.8B d=2560, L=32. Predictive alignment is maintained across all layers ($z > 2$ at 32/33 layers). Peak $z = +14.5$ at L15; minimum $z = +0.4$ at L0 (embedding layer). No collapse observed. Late layers show gradual decline from peak but remain positive throughout, with recovery at the final layer ($z = +8.6$).

OLMo-1B d=2048, L=16 (trained on Dolma with tied embeddings). Strong predictive alignment throughout: $z > 2$ at all 17 layers, $z > 5$ at 13/17 layers. Peak $z = +24.0$ at L15. The broadly increasing profile in late layers contrasts with Pythia-1B's plateau, possibly reflecting architectural differences (tied vs. untied embeddings). No loss of alignment

observed, consistent with the capacity interpretation given d=2048.

Pythia-6.9B d=4096, L=32. Peak $z = +17.93$ at L9. No mid- or late-layer masking, consistent with the other large Pythia models.

Phi-1.5 d=2048, L=24 (trained on textbook-quality data). Peak $z = +12.39$ at L24. No intermediate-layer collapse. Its alignment is on average stronger than Pythia-1.4B's at matching width (mean z 5.99 vs. 5.26; higher at 14 of 24 layers), the comparison behind the "factors beyond capacity" remark in §7.

Gemma-2-2B d=2304, L=26 (tied embeddings). Strong alignment at intermediate layers (peak $z = +16.36$ at L22) but collapses at the final layer ($z = -16.5$ at L26). This family-specific final-layer collapse is examined in Appendix F: it also appears in the Qwen-2.5 family but is *not* explained by embedding tying (Llama-3.2 is tied yet stable; Qwen-7B is untied yet collapses), width, or vocabulary size, and is distinct from the width-driven intermediate-layer detour that is our focus.

C.2. Comprehensive Numeric Tables

This section provides exact numeric values corresponding to the visualizations in the main text. Table 2 provides peak and trough z-scores across the complete model suite (corresponding to Figure 2a). Table 3 and Table 4 provide z-scores throughout training for Pythia-410M, 160M, and 1B (corresponding to Figure 6). Figure 5 shows the bright→dark PC1 migration behind §5.1; Figure 6 the training-dynamics trajectories and loss curves behind §5.2; and Figure 7 the three co-timed bands (window z, frequency-matched content gap, and copy-machinery maturity) behind the onset analysis of §5.2.

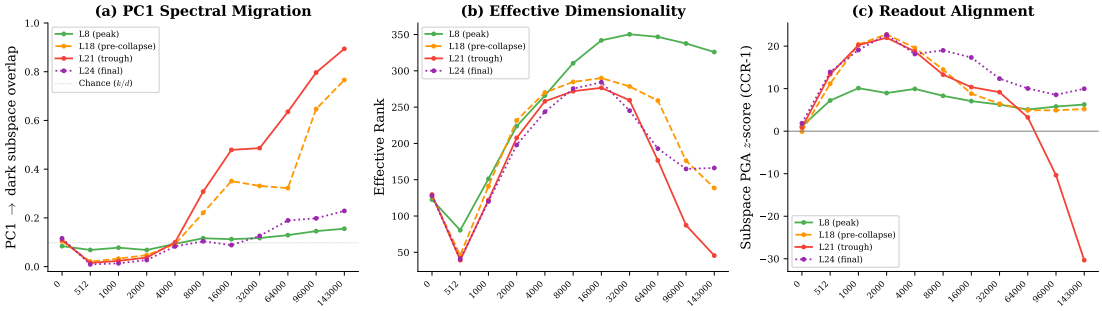

Figure 5. **PC1 Migrates Bright → Dark as Collapse Emerges (Pythia-410M). (a)** PC1's dark-subspace overlap increases at late layers during training; the final layer stays anchored. **(b)** Effective rank drops as anisotropy concentrates on dark directions. **(c)** z-score goes negative when PC1→dark > 0.80 and effective rank < 100. The migration involves a *group* of dominant directions, not one: in the collapse zone the surviving second component is darker than the first.

Table 2. **Predictive Alignment Across Scales.** Mid-layer geometry is significantly aligned with the readout in all models with sufficient depth. Small models lose this alignment at late layers; large models maintain it throughout. P = Pythia, G = Gemma-2, O = OLMo. All results: k=100, 100 random subspaces, 1000 contexts.

MODEL	d	L	PEAK z	@ L	z>5	z>2	LATE-LAYER LOSS?
P-70M	512	6	+16.26	L6	1/7	2/7	YES (z→−11, L3–5)
P-160M	768	12	+15.07	L12	1/13	4/13	YES (z→−13, L7–11)
P-410M	1024	24	+9.27	L24	9/25	16/25	YES (z→−32, L20–23)
P-1B	2048	16	+12.85	L8	14/17	16/17	NO
P-1.4B	2048	24	+12.34	L12	12/25	17/25	BORDERLINE (z=−0.5, L20)
P-2.8B	2560	32	+14.53	L15	23/33	32/33	NO
P-6.9B	4096	32	+17.93	L9	28/33	32/33	NO
PHI-1.5	2048	24	+12.39	L24	15/25	24/25	NO
O-1B	2048	16	+24.00	L15	13/17	17/17	NO
G-2-2B	2304	26	+16.36	L22	26/27	26/27	FINAL ONLY (z=−16.5, L26)

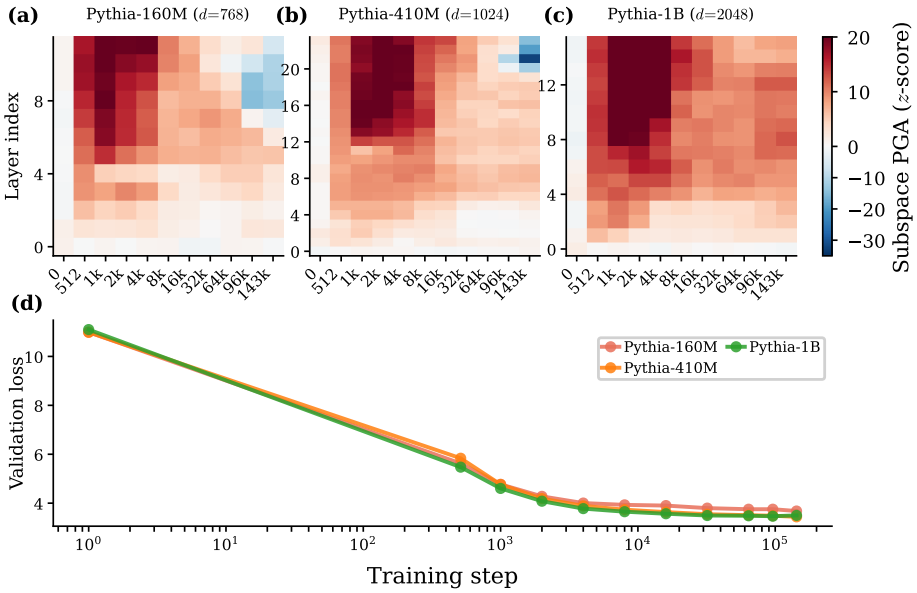

Figure 6. **Masking is learned and capacity-dependent.** z-score across training steps and layers for **(a)** 160M, **(b)** 410M, **(c)** 1B. Small models develop late-layer masking by step ∼96k; 1B never does. **(d)** Validation loss converges comparably: masking does not affect the training objective.

Table 3. **Training Dynamics (Pythia-410M).** Predictive alignment is established by step 512 (24/25 layers with $z>2$). Loss of alignment appears after step 64,000 and deepens through the end of training, even as loss continues to improve.

Step	$z>2$	Peak z	Min z	Masked?
0	1/25	+2.1	−1.0	No
512	24/25	+10.8	+0.1	No
1,000	24/25	+20.4	+0.6	No
4,000	22/25	+20.5	+0.3	No
16,000	23/25	+18.4	−0.7	No
64,000	19/25	+10.1	−0.5	No
96,000	16/25	+8.1	−10.2	Yes
143,000	16/25	+9.3	−32.3	Yes

Table 4. **Multi-Model Training Dynamics.** 160M develops masking at the same time as 410M (step ∼96k). 1B never develops masking despite comparable PC1 migration, supporting the capacity interpretation. 1B's negative **Min** z entries fall at the embedding layer (L0, never below −0.8 after step 4k); its late layers stay positive throughout (late-layer minimum ≈+7), unlike the late-layer collapse behind the −14.8/−32.3 minima of 160M/410M.

	160M (d=768)		**410M** (d=1024)		**1B** (d=2048)	
Step	$z>2$	Min z	$z>2$	Min z	$z>2$	Min z
0	0/13	−1.1	1/25	−1.0	0/17	−2.0
512	12/13	−0.2	24/25	+0.1	16/17	+0.2
1,000	12/13	+0.5	24/25	+0.6	16/17	+0.8
4,000	12/13	+0.5	22/25	+0.3	15/17	−1.0
16,000	8/13	−1.4	23/25	−0.7	15/17	+0.1
64,000	8/13	−1.4	19/25	−0.5	16/17	+0.1
96,000	7/13	−14.8	16/25	−10.2	16/17	−0.5
143,000	4/13	−13.0	16/25	−32.3	16/17	−0.8

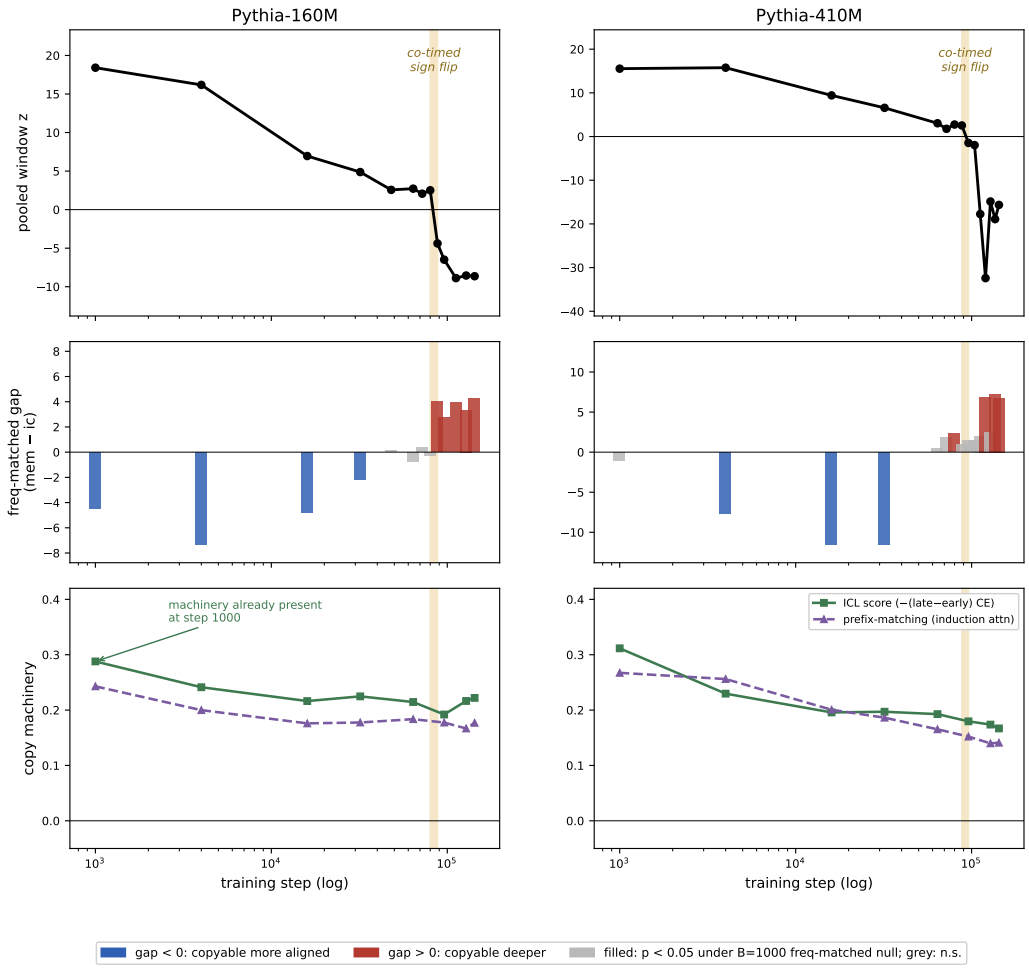

Figure 7. **The content selectivity switches on with the collapse, long after the copy machinery forms** (Pythia-160M and 410M, three x-aligned bands per model; the shaded band marks the window-z zero crossing). **Top:** the masking trajectory, pooled z over each model's pre-registered window, strongly positive at the first checkpoint. **Middle:** the frequency-matched gap between parametric and copyable contexts, positive when copyable contexts collapse deeper; bars carry full colour where the B=1000 within-frequency-bin partition null rejects at 0.05. Early in training the gap is *negative*, so copyable-target geometry is the better aligned of the two while $z > 0$; it turns positive as z crosses below zero, within one late window per model. **Bottom:** the copy machinery itself, measured behaviourally and by prefix-matching attention, is already at its mature value at the first checkpoint and roughly flat thereafter, nearly two orders of magnitude in training steps before the flip.

Table 5. **Content-selectivity follows collapse per (model, corpus) cell** (frequency-matched; $\Delta = \bar{z}_{\text{parametric}} - \bar{z}_{\text{copyable}}$ over pre-registered windows, positive = copyable collapses deeper; p against B=1000 within-frequency-bin balanced partitions; subsets contain 138–474 contexts each, smallest on code and largest on WikiText). † marks cells where the model's interior z dips below -2 at relative depth $\geq$0.5 *on that corpus* (Table 9); in all eight cells where both quantities are measured, Δ is significantly positive exactly when the cell collapses (8/8). A significant positive Δ occurs *only* in † cells (5/5; no counterexample in twelve cells); the 160M–WikiText cell (†: 160M collapses on the shuffled sample, $z_{\min}$=-12.4 at L9, inside its pre-registered window; null gap) shows the converse does not hold: collapse is necessary for the selectivity, not sufficient. The frequency confound is large and corpus-dependent: median target-token frequency differs by 21–43× on OpenWebText and by 12× on code, and inverts on WikiText, where copyable targets are the rarer ones. It is also consequential: Pythia-2.8B's raw OpenWebText gap of $+3.2$ ($p < .001$) reverses to the -1.9 (p=.22) shown here once targets are frequency-matched.

	frequency-matched Δ (p)		
Model	OpenWebText	code	WikiText
Pythia-160M	$+4.4^{\dagger}$ ($<.001$)	$+4.6^{\dagger}$ (.001)	$+0.7^{\dagger}$ (.60)
Pythia-410M	$+6.9^{\dagger}$ (.001)	$+10.9^{\dagger}$ ($<.001$)	-0.4 (.81)
Pythia-1B	-7.1 ($<.001$)	$+1.1$ (.27)	-0.4 (.64)
Pythia-2.8B	-1.9 (.22)	$+6.7^{\dagger}$ ($<.001$)	$+1.3$ (.14)

D. Robustness and Methodological Controls

Summary (expanded from §3): results are stable under bootstrap resampling (Appendix D.11), across sample sizes (Appendix D.12), and across the readout dimension $k \in \{50, 100, 200\}$ (the sign pattern is unchanged: 160M/410M interior minima $-7.2/-13.0$ at k=50 and $-33.3/-76.0$ at k=200, 1B positive at both); the late-layer collapse further persists at matched coverage k=$d/10$ (Appendix D.3) and on two held-out corpora (WikiText-103 and Python code; Appendix D.5), where the narrow-model collapse recurs at the same layers (deeper on code) and every Pythia model still recovers at the output layer. The detour also reproduces across *independent training seeds* ($4/4$ re-inits collapse and recover): Pythia 70M, 410M, and two 160M seed re-inits all reach negative interior z (interior minima -10 to -23) and recover at the output ($4/4$). *Orthogonal* PGA (the same metric computed in $\mathcal{V}_k^{\perp}$) never exceeds the random null at any model-layer combination tested (0/12; Appendix B), so the alignment really is concentrated in $\mathcal{V}_k$ rather than diffused across hidden space. Replacing W_U with the input embedding W_E yields z-scores indistinguishable from random (Appendix D.13), so the alignment is specific to the predictive interface and not an artifact of any matrix the model contains.

D.1. Anisotropy Correction

After mean+CCR-1, pairwise isotropy improves substantially at every layer but does not reach 1. Isotropy is computed as $1 - \lambda_1 / \sum_i \lambda_i$ where λ_i are eigenvalues of the pairwise cosine similarity matrix. Pre-correction, isotropy ranges from 0.70 to 0.96 across layers; post-correction it rises to 0.95–0.98 in the mid-to-late layers (global maximum 0.978, OLMo-1B), with layer 0 remaining lower (≈ 0.72–0.74) where the dominant direction carries little of the variance. This substantially reduces the anisotropy confound.

D.2. CCR-1 / Readout Overlap

The CCR-1 correction removes the direction v_1 of largest variance in the centered hidden states. To check that this direction is largely outside the readout subspace (so that CCR-1 is not silently removing readout-aligned variance and biasing z), we report two complementary measures.

The primary measure (used in §2.2) is $\|P_k v_1\|^2$, the fraction of the CCR-1 direction v_1's variance that lies in the readout subspace $\mathcal{V}_k$ (k=100). At the layers where collapse occurs this fraction is small, ≈ 0.13 for Pythia-410M (L6–L23) and ≈ 0.29 for Pythia-1B, so v_1 lies mostly (70–87%) outside the readout. At the final layer of both models it rises to ≈ 0.60–0.63 (where $z > 0$). CCR-1 therefore removes a direction that is largely outside the readout at the layers where collapse happens, so any change in z after correction reflects geometric reorganization rather than removal of readout-aligned variance. (Relative to the k/d chance overlap of a random direction, 0.10 and 0.05, the absolute overlaps $0.13/0.29$ are a mild *enrichment*, so absolute smallness alone does not settle the question; the depth inversion above is the diagnostic pattern.) The decisive control removes the correction entirely: with mean-centering only (CCR-0, no PC removal), the collapse is *deeper*, not absent (interior minima -20.6 for 160M, -59.0 for 410M; 1B stays positive), so CCR-1 attenuates rather than manufactures the effect.

The second measure is the single-vector cosine $|\cos(v_1, u_1^{W_U})|$: how much v_1 aligns with W_U's top right singular vector alone. This is a strict lower bound on $\|P_k v_1\|$ but easier to summarize per model. Table 6 reports $\max_\ell |\cos(v_1^{(\ell)}, u_1^{W_U})|$ across all layers for the Pythia suite. All values are < 0.18, consistent with the projection-norm argument above.

Table 6. **CCR-1 / readout single-vector overlap.** Maximum single-vector cosine $|\cos(v_1, u_1^{W_U})|$ across layers, a lower bound on $\|P_k v_1\|$ that summarizes CCR-1 / readout overlap per model.

MODEL	MAX $\lvert\cos(v_1, u_1^{W_U})\rvert$	LAYER
P-160M	0.060	L0
P-410M	0.134	L19
P-1B	0.122	L14
P-2.8B	0.173	L29

D.3. Constant-Coverage Control

Fixing k=100 gives the readout subspace a different coverage k/d at each scale, which sets the random-null mean (§D.9); a reviewer raised that this could make the cross-scale collapse a coverage artifact. We therefore recompute PGA at *matched*

coverage $k = \mathrm{round}(d\,c)$, the same k/d for every model, for $c \in \{0.05, 0.10\}$ (`exp21`). Holding k/d fixed pins the coverage-dependent part of the null across scales by construction, so any residual scale-dependence reflects geometry, not coverage. Table 7 reports the minimum interior (non-embedding, non-final) z per model.

Table 7. **Constant-coverage control.** Minimum interior PGA z at matched coverage k=round$(d\,c)$. Narrow models ($d \leq 1024$) still collapse (negative) at matched coverage, just as at k=100; among the wide ($d \geq 2048$) models the deeper 1.4B/2.8B collapse at c=0.10 while the shallow 1B stays positive. `exp21`.

MODEL	d	L	MIN z (c=0.05)	MIN z (c=0.10)
P-70M	512	6	−1.6	−4.7
P-160M	768	12	−4.7	−11.3
P-410M	1024	24	−12.8	−33.4
P-1B	2048	16	+3.3	+2.0
P-1.4B	2048	24	−0.3	−8.9
P-2.8B	2560	32	+2.0	−2.1
P-6.9B	4096	32	+3.5	+1.4

The narrow models ($d \leq 1024$: 70M, 160M, 410M) still collapse to negative interior z at matched coverage, just as at k=100, so the late-layer collapse is *not* an artifact of the k/d ratio. Among the wide ($d \geq 2048$) models the picture is depth-modulated: the shallow Pythia-1B (L=16) stays positive at both coverages, whereas the deeper Pythia-1.4B (L=24) and Pythia-2.8B (L=32) are stable at c=0.05 but collapse at c=0.10, consistent with depth giving anisotropy more layers to accumulate (§7). Width alone therefore does not guarantee coverage-robust stability; we accordingly rest claims on the *sign* of the interior minimum and the within-model depth trajectory, not on z magnitudes compared across coverages, which lie on different scales. Consistent with a capacity reading, a single width-per-layer variable d/L orders the interior collapse severity across the Pythia suite (Spearman r_s=0.87, p=0.01, n=7, at both c=0.05 and c=0.10), more stably than width alone (r_s=0.79 at c=0.05, 0.58 at c=0.10) or depth alone (0.49 / 0.24); with n=7 the margin over width is not itself significant (leave-one-out p up to 0.059), so we treat d/L as an ordering, not a law. On a held-out corpus (same models, new domain), however, the ordering holds: on code the interior minima of all six Pythia models measured are rank-ordered by d/L (tie-corrected Spearman r_s=0.99, n=6; 1B and 6.9B tie at d/L=128, and the adjacent 1.4B/2.8B minima differ by 0.2 z-units, within noise), and the two models added to that run after the prediction was stated, Pythia-1.4B (d/L=85, predicted to dip) and Pythia-6.9B (d/L=128, predicted not to), landed on the predicted side (−4.1 and +0.4; Appendix D.5). This ordering is *within-recipe*: across model families the same coverage puts wide non-Pythia models into a large-k regime whose interior z is not comparable (e.g. Llama-2-7B and Pythia-6.9B share d=4096 and k but diverge), so we do not extend the d/L ordering across families; the family-specific cross-family behaviour is characterized separately at k=100 in Appendix F.

D.4. Robustness to the RSA Correlation: Spearman vs. Pearson

The readout correlation ρ_{read} (§2) and the null distribution both use Spearman rank correlation between pairwise-distance vectors. To check that the conclusions do not hinge on this choice, we recompute the entire per-layer z trajectory with Pearson correlation instead, on the *same* cosine distances and the *same* 100 random null subspaces, so that only the coefficient changes (`compute_subspace_pga(corr=...)`; `exp26`). Magnitudes are not comparable across coefficients (Pearson z runs ∼20% smaller at the peaks and larger at the troughs), so we compare the *sign* of the interior alignment, the peak layer, and the rank ordering of the trajectory: the same invariants the main claims rest on.

Across all ten models every qualitative conclusion is preserved under Pearson (Table 8): interior-minimum signs and final-layer signs are concordant in all ten (including Gemma-2-2B's final-layer collapse, −17.6 Spearman vs. −22.4 Pearson), the interior peak layer coincides within one layer wherever the trajectory is non-flat, and the interior rank correlations span 0.885–1.000. ([†]Gemma's interior is a flat plateau ($z \approx +7$–$+17$ everywhere), on which rank correlation is uninformative; its invariant, sign concordance at 25/25 interior layers and at the final layer, holds.) The eight interior sign disagreements across the ten models all sit at transitional layers where at least one coefficient's $|z| \leq 0.48$ (and both ≤ 1.75); no layer that both coefficients place at $|z| \geq 2$ disagrees, so no positive claim flips. Six of the ten models agree at every interior layer. (At the embedding layer ℓ=0, which the interior claims exclude, the coefficients can disagree, e.g. OLMo-1B has z_{Spear}=+7.5 vs. z_{Pear}=−18.3, because pre-contextual token geometry is dominated by frequency-driven norm structure that the rank transform and the linear correlation treat differently.)

Table 8. **Spearman vs. Pearson RSA is qualitatively invariant.** Per model (ten models): rank correlation between the two per-layer z trajectories over the interior ($1 \leq \ell \leq n_L - 1$); the interior peak layer under each coefficient; the interior minimum z under each; and the number of interior layers whose z-sign agrees. The narrow models ($d \leq 1024$) collapse to negative interior z under *both* coefficients. The eight interior sign disagreements sit in four models and only at transitional layers where at least one coefficient's $|z| \leq 0.48$ (both ≤ 1.75); six of the ten models agree at every interior layer. `exp26`, n=1000 shared contexts, k=100.

Model	d	$\rho_{\mathrm{Spear-Pear}}$	peak ℓ (S/P)	interior min z (S/P)	sign agree
Pythia-70M	512	1.000	$2/2$	$-11.3/-12.4$	$4/5$
Pythia-160M	768	0.973	$5/5$	$-12.9/-15.5$	$9/11$
Pythia-410M	1024	0.956	$8/8$	$-32.2/-47.0$	$21/23$
Pythia-1B	2048	0.979	$8/8$	$+2.9/+1.6$	$15/15$
Pythia-1.4B	2048	0.983	$12/12$	$-0.5/-1.5$	$20/23$
Pythia-2.8B	2560	0.981	$15/16$	$+2.8/+2.0$	$31/31$
Pythia-6.9B	4096	0.885	$9/9$	$+4.6/+2.7$	$31/31$
OLMo-1B	2048	0.968	$15/15$	$+2.1/+0.8$	$15/15$
Phi-1.5	2048	0.950	$22/22$	$+2.7/+0.2$	$23/23$
Gemma-2-2B	2304	$0.054^{\dagger}$	$22/9$	$+6.6/+2.9$	$25/25$

D.5. Corpus Robustness

All headline measurements use one fixed OpenWebText context set, and z is context-set-sensitive, so we recompute full per-layer trajectories on two held-out corpora, WikiText-103 and Python code (1,000 contexts each, same recipe), and compare *signs and trajectories* only (magnitudes are not comparable across context sets). Table 9 summarizes.

Table 9. **Corpus robustness.** Interior minimum z (and final-layer z) per corpus. Invariants that replicate on all corpora: the narrow models' interior collapse (160M on all three, at the same layers; 410M on OWT and code, at the same layer L21, deeper on code), the narrow$\ll$wide ordering, and final-layer recovery in every Pythia model. Corpus-dependent: collapse *depth* (410M flattens to chance on WikiText, which the shuffled control reproduces); on code the two mid-capacity wide models dip mildly, in line with the d/L ordering (see below). Gemma-2-2B's family-specific *final*-layer collapse replicates on all three corpora.

Model	OWT	code	WikiText
Pythia-160M	-12.8 ($+15.1$)	-14.6 ($+4.5$)	-11.7 ($+12.3$)
Pythia-410M	-32.3 ($+9.3$)	-43.2 ($+0.9$)	-1.1 ($+8.9$)
Pythia-1B	$+2.9$ ($+7.1$)	$+0.2$ ($+2.6$)	-2.7 ($+7.2$)
Pythia-1.4B	-0.5 ($+5.7$)	-4.1 ($+0.2$)	-0.7 ($+8.3$)
Pythia-2.8B	$+2.8$ ($+8.6$)	-4.3 ($+1.4$)	$+0.1$ ($+8.7$)
Pythia-6.9B	$+4.6$ ($+9.1$)	$+0.5$ ($+2.3$)	-1.3 ($+9.6$)
Gemma-2-2B	$+6.6$ (-17.6)	$+4.2$ (-9.0)	-6.6 (-19.2)

Beyond replication, the code corpus provides a held-out-corpus test of the capacity ordering (same models, new domain): interior minima of all six Pythia models are rank-ordered by d/L (tie-corrected Spearman r_s=0.99, n=6: -43.2 at d/L=43 up to $+0.5$ at 128; 1B/6.9B tie on the predictor, and the adjacent 1.4B/2.8B minima differ by 0.2 z-units), and the two models added to the run *after* the directional prediction was stated, 1.4B (predicted dip) and 6.9B (predicted no dip), both landed as predicted. We report the WikiText flattening of 410M honestly, and a shuffled-sample control settles its cause: the flattening *replicates* on a shuffled WikiText-103 sample ($z_{\min}$=-1.8 vs. -1.1 first-N), so it is a genuine corpus effect, not an artifact of the original sample's topical homogeneity. The surviving reading is content demand: the collapse is content-selective (§5.3), and a corpus with less in-context-copy demand shows a shallower detour, although the 160M–WikiText cell (collapse without selectivity, §5.3) shows corpus demand cannot be the whole story.

D.6. Pre-Registered Test: the Deduped-Pythia Suite

The deduped-Pythia suite (identical architectures, tokenizer, and training recipe; deduplicated training data) provides seven same-shape models never used in forming the d/L hypothesis. Before any deduped-model extraction, we froze directional predictions, a rank prediction, and explicit falsifiers in a version-controlled pre-registration file (`results/phenomenon/exp39_deduped_prereg/PREREGISTRATION.md`; the measurement script refuses to run in its absence, so every measurement provably postdates the predictions).

The outcome splits exactly along the continuum framing of the main text (Table 10). At the extremes the geometry is training-data-invariant to a degree we did not anticipate: the three deep-collapse models and the two clearly-stable models

Table 10. **Pre-registered deduped-suite outcomes** (OWT shared contexts, k=100, CCR-1; "twin" = the standard-Pythia value). Five of seven directional calls landed; both misses are borderline models shifting *toward* collapse under deduplicated data.

Model (deduped)	d/L	predicted	observed $z_{\min}$ (twin)	hit
410M	42.7	dips, deepest	-34.1 (-32.3)	✓
160M	64	dips	-8.6 (-12.8)	✓
70M	85.3	dips	-12.3 (-11.4)	✓
2.8B	80	no dip	$+3.0$ ($+2.8$)	✓
6.9B	128	no dip	$+4.1$ ($+4.6$)	✓
1.4B	85.3	shallow dip > -5	-8.7 (-0.5)	✗
1B	128	no dip	-0.9 ($+2.9$)	✗

all landed as predicted, with interior minima tracking their standard twins almost point-for-point (-34.1 vs. -32.3; $+4.1$ vs. $+4.6$). The two *borderline* models missed their calls, both in the same direction: deduped 1.4B dips to -8.7 (twin: grazes zero) and deduped 1B slips just below zero (-0.9; twin $+2.9$). Our pre-registered rank falsifier accordingly fired (tie-corrected r_s=0.67 over the six-model set, below the frozen 0.8 threshold): the miss is concentrated in the middle of the ordering, not at its ends. We read this as convergent evidence for the paper's central scoping: the *ordering* of collapse severity by d/L and the separation of its extremes are robust, across corpora (Appendix D.5) and now across training data, while any *binary* dip/no-dip boundary is condition-dependent (coverage, corpus, and now data), which is why the paper claims a continuum rather than a threshold. Why deduplication shifts borderline models toward collapse is an open question; one candidate, consistent with the content-selectivity of §5.3 (the collapse is deeper for context-copyable predictions), is that deduplicated data reduces memorizable repetition and increases reliance on in-context copying; we flag this as a post-hoc hypothesis, not a finding.

D.7. Reparameterization (Gauge) Invariance

A reviewer may worry that z depends on an arbitrary choice of residual-stream basis. Two reparameterizations must be distinguished, both of which leave the model's logits, hence its function, exactly unchanged. **Orthogonal:** $h \mapsto hQ$, $W_U \mapsto W_U Q$ with Q orthogonal preserves the cosine metric as well as the logits, and z is invariant *by construction*: cosine distances are unchanged and the readout subspace co-rotates, so the projected coordinates $hV_k^\top$ are identical. We confirm this numerically on real hidden states (Pythia-410M and 1B, three depths each, random orthogonal Q): ρ_{read} is invariant to machine precision ($|\Delta\rho_{\text{read}}| \leq 5\times10^{-7}$), and the residual wobble in z ($\leq$0.5 at B=2000 null draws) is Monte-Carlo noise in the random-subspace null, not basis dependence (`exp27`). **General linear:** $h \mapsto A^{-1}h$, $W_U \mapsto W_U A$ with A invertible also preserves the logits but *not* the cosine metric, and ρ_{read} genuinely moves (by up to 3×10^{-3}, well above machine precision, in the same experiment). This dependence is not a defect: $\mathrm{GL}(d)$ is not a symmetry of a trained transformer (LayerNorm acts coordinate-wise and training fixes the residual basis), so the native basis in which we measure is the one the model's own normalization and readout operate in, not an arbitrary convention. z is invariant to exactly the transformations that preserve the geometry the model uses.

D.8. From-Scratch Sweep: Design and Final-Layer Regime

The powered from-scratch sweep of §7.3 has two arms: fixed d=512 with $L \in \{8, 16\}$ complete to 300k steps and L=32 truncated at 75k by infrastructure failure; and fixed L=8 with $d \in \{256, 1024\}$; each cell gets ∼5B tokens, 5× the length of our earlier attempt, but still a small fraction of the scale at which the pretrained models reorganize (Pythia's onset occurs after ∼190B tokens). No cell develops a stable late-layer collapse zone with final-layer recovery; indeed these briefly-trained models' *final* layers are themselves readout-misaligned ($z \approx -8$ to -21), placing them in the family-specific final-collapse regime of Appendix F rather than the Pythia regime. The surviving directional observations, that both d/L=32 cells drift to negative interior z in lockstep (trajectory minima -3.5 and -6.8, final -3.0 and -5.2), while the d/L=64 cell only grazes zero (final -0.7) and the d/L=128 cell ends positive ($+2.4$)..

D.9. Unembedding Spectral Concentration

Figure 8 plots cumulative variance explained by the top-k right singular vectors of W_U across the Pythia suite. Small models (70M, 160M) have near-rank-1 unembeddings (the top singular value alone explains ∼92% of variance and k=100 captures effectively all informative directions), while larger models have flatter spectra (k=100 captures ∼30% for 410M, ∼20%

for 1B). The steep small-model decay reflects W_U's limited capacity when d is small relative to $|\mathcal{V}|$. Because coverage varies with scale, we do not rely on cross-scale z magnitudes; the constant-coverage control (Appendix D.3) confirms the narrow-model collapse is not an artifact of the k/d ratio.

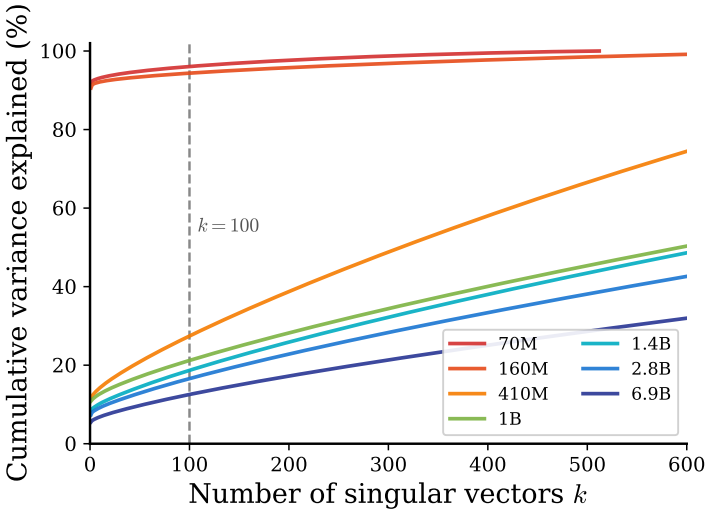

Figure 8. **Cumulative variance explained by top-k right singular vectors of W_U.** Small models (warm colors) have steeply decaying spectra; large models (cool colors) have flatter spectra. Dashed line marks k=100, the default readout subspace dimension.

D.10. Absolute Correlation Values Per Layer

Table 11 reports the absolute readout correlation ρ_{read}, null distribution mean μ_{null} and standard deviation σ_{null}, and z-score for every layer of four representative Pythia models. This contextualizes large-magnitude z-scores: at Pythia-410M L21 ($z = -32.3$), $\sigma_{\text{null}} = 0.0125$ is small but the absolute gap ($\rho_{\text{read}} = 0.499$ vs. $\mu_{\text{null}} = 0.903$) is large: the effect is substantive, not an artifact of vanishing σ_{null}. Across the collapse zones (bolded interior rows), ρ_{read} stays moderate ($\approx$0.50–0.74) while μ_{null} rises to $\approx$0.80–0.90, the correlation-level decomposition behind §5.1.

D.11. Bootstrap Confidence Intervals

To assess the statistical reliability of our z-score estimates, we computed bootstrap confidence intervals by resampling the 1,000-text evaluation set with replacement (1,000 resamples) and recomputing PGA at representative layers for all five models. Table 12 reports point estimates and 95% bootstrap CIs for peak-alignment, trough (collapse), and final layers.

D.12. Sample Stability

To verify that results do not depend on the specific sample size used (N=1000), we evaluated PGA on Pythia-410M at $N \in \{100, 200, 500, 1000\}$, repeating each sample size 5 times with different random text subsets. Table 13 reports mean and standard deviation of z-scores across repeats at the peak (= final, L24) and trough (L21) layers. Since peak and final coincide for Pythia-410M, they share one row.

To confirm the headline numbers do not depend on N=1000, we extended to $N \in \{1000, 2000, 5000\}$ for a small (160M), a collapse (410M), and a large (1B) model, drawing from a larger pool (1,200 shared contexts plus OpenWebText; 3 repeats). Values are flat from N=1000 to 5000 (the 410M collapse holds at $z \approx -32$, the narrow/wide separation is unchanged), while the across-repeat standard deviation shrinks with N (Table 14). Absolute z differs slightly from Table 2 (different pool); the peak/trough pattern is identical. Five times the data does not move the conclusions.

D.13. Specificity to the Predictive Interface (W_E Control)

Would *any* embedding matrix produce comparable alignment, or specifically W_U? We recompute PGA with the input embedding W_E (used in tokenization, not prediction) for Pythia-410M, 1B, and 160M (all untied, $W_E \neq W_U$). The top-100 W_U and W_E subspaces are nearly orthogonal (Grassmann overlap 0.10 for 410M, 0.05 for 1B). W_U gives mean $|z| = 6.5$

Table 11. **Per-layer absolute correlation values.** ρ_{read}: readout subspace Spearman correlation; μ_{null}, σ_{null}: null distribution statistics from 100 random subspaces. Layers with $z < 0$ are **bolded**.

Model	Layer	ρ_{read}	μ_{null}	σ_{null}	z
P-160M	0	**0.746**	**0.753**	**0.0197**	**-0.4**
	5	0.743	0.655	0.0179	4.9
	8	**0.736**	**0.887**	**0.0118**	**-12.8**
	10	**0.637**	**0.804**	**0.0155**	**-10.8**
	12	0.850	0.599	0.0167	15.1
P-410M	7	0.581	0.509	0.0134	5.4
	14	0.755	0.649	0.0234	4.5
	18	0.768	0.652	0.0218	5.3
	20	**0.650**	**0.857**	**0.0173**	**-12.0**
	21	**0.499**	**0.903**	**0.0125**	**-32.3**
	23	**0.615**	**0.834**	**0.0148**	**-14.8**
	24	0.887	0.709	0.0192	9.3
P-1B	0	**0.704**	**0.719**	**0.0191**	**-0.8**
	5	0.600	0.448	0.0163	9.4
	8	0.635	0.445	0.0148	12.9
	12	0.831	0.573	0.0207	12.4
	16	0.872	0.711	0.0227	7.1
P-2.8B	0	0.711	0.703	0.0212	0.4
	6	0.519	0.392	0.0100	12.7
	15	0.625	0.406	0.0151	14.5
	24	0.810	0.712	0.0336	2.9
	32	0.860	0.661	0.0232	8.6

(410M) and 8.6 (1B), reaching $+9.3$ and -32.3 at individual 410M layers; W_E gives mean $|z| \approx 1.0$ for all three models, with $|z| < 3$ at every non-embedding layer (indistinguishable from random), except L0, where z_{W_E} is large and positive ($+6.2$ for 410M) since layer-0 representations *are* the embeddings. Alignment is specific to the predictive readout, not to any matrix the model contains.

D.14. The Readout Subspace Carries Predictive, Not Surface, Structure

A complementary specificity check asks what the readout-subspace geometry correlates with *in content*. For each layer we correlate pairwise representation distances (restricted to a subspace) with (a) pairwise distances between the model's next-token *distributions* (predictive structure) and (b) pairwise lexical overlap of the contexts (surface structure), using a partial-correlation abstraction index d_A that credits a subspace for tracking prediction beyond what surface overlap explains. Across Pythia-160M/410M/1B (55 layers, 100-draw nulls): the readout subspace's correlation with predictive structure

Table 12. **Bootstrap 95% CIs for PGA z-scores.** 1,000 bootstrap resamples per model. Peak = layer with highest z; trough = layer with lowest z (collapse zone for small models); final = last layer. CIs confirm that collapse in small models and alignment in large models are not sampling artifacts.

MODEL	LAYER	z-SCORE	95% CI
P-160M	PEAK/FINAL (L12)	$+15.1$	$[+13.7, +16.1]$
	TROUGH (L8)	-12.8	$[-13.6, -11.7]$
P-410M	PEAK/FINAL (L24)	$+9.3$	$[+8.6, +9.8]$
	TROUGH (L21)	-32.3	$[-35.0, -29.1]$
P-1B	PEAK (L8)	$+12.9$	$[+11.6, +13.4]$
	FINAL (L16)	$+7.1$	$[+6.5, +7.6]$
P-2.8B	PEAK (L15)	$+14.5$	$[+13.1, +15.2]$
	FINAL (L32)	$+8.6$	$[+7.9, +9.1]$
O-1B	PEAK (L15)	$+24.0$	$[+21.1, +25.2]$
	FINAL (L16)	$+14.4$	$[+13.0, +15.2]$

Table 13. **Sample Stability (Pythia-410M).** z-scores (mean $\pm$ std, 5 repeats) are stable across sample sizes. Collapse is detectable even at N=100.

LAYER	N=100	N=200	N=500	N=1000
PEAK/FINAL(L24)	$+8.5\pm0.8$	$+9.1\pm0.7$	$+9.6\pm0.8$	$+9.8\pm0.9$
TROUGH(L21)	-30.2 ± 4.7	-35.2 ± 4.8	-35.0 ± 3.7	-33.3 ± 1.6

Table 14. **Sample Stability at Large N (Pythia-160M / 410M / 1B).** PGA z (mean $\pm$ std, 3 repeats) at representative layers for N up to 5,000 contexts. z is unchanged from N=1000 to N=5000; the spread narrows with N.

MODEL	LAYER	N=1000	N=2000	N=5000
P-160M	PEAK/FINAL (L12)	$+17.1\pm1.8$	$+16.8\pm0.6$	$+17.2\pm0.7$
	TROUGH (L8)	-11.9 ± 1.0	-12.1 ± 1.0	-12.1 ± 0.9
P-410M	PEAK/FINAL (L24)	$+8.9\pm0.6$	$+9.2\pm0.2$	$+9.1\pm0.4$
	TROUGH (L21)	-31.1 ± 1.1	-30.8 ± 1.7	-32.0 ± 0.8
P-1B	PEAK (L8)	$+12.2\pm0.6$	$+12.2\pm0.8$	$+12.5\pm0.3$
	FINAL (L16)	$+7.5\pm0.6$	$+7.4\pm0.5$	$+7.6\pm0.7$

ranges r=0.06–0.46 while its correlation with surface structure is negligible (r=+0.05 on average over the 55 layers measured, range -0.02 to $+0.09$), and d_A in the readout subspace exceeds the random-subspace mean at $49/55$ layers. Notably, in Pythia-410M's collapse zone (L20–L23, z=-12 to -32) the readout subspace's predictive-content correlation is at its *highest* (r=0.26–0.35) and d_A still exceeds random, a further, content-level line of evidence that the collapse masks the readout's privileged *distance-preservation* status without erasing the predictive structure the subspace carries. (d_A is a content measure and accordingly anti-correlates with z across depth; it complements, and does not proxy for, alignment.)

D.15. A Learned Affine Repairs Decoding Through the Collapse Zone

The anticipated objection to masking-not-erasure is that CCR-5's restoration only shows the misalignment is low-rank: a learned per-layer readout might, or might not, recover actual *decoding*. We test this directly with a tuned-lens construction (Belrose et al., 2023): per layer, a residual affine translator $h' = h + Ah + b$ trained to minimize KL(final logits $\|$ lens logits) on $\sim 10^5$ token positions of the shared contexts (700 train / 300 held-out contexts), evaluated against the zero-shot logit lens (A=0, b=0).

The learned affine repairs decoding smoothly *through* the collapse zones. At Pythia-160M's collapse-zone layers L8–L11, the zero-shot lens is uniformly poor (held-out KL $= 11.7$–13.0, top-1 agreement 8–28%) while the tuned lens cuts through it (KL $= 0.57$–2.14, top-1 50–73%), a 5–20$\times$ reduction; pooled over collapse-zone layers, top-1 agreement rises from 18% (zero-shot) to 60% (tuned). In Pythia-410M's zone (L20–L23, $z = -12$ to -32), tuned-lens KL ($0.74 \rightarrow 0.15$) sits on the same depth trend as the never-collapsing Pythia-1B at matched relative depths ($0.59 \rightarrow 0.20$): *under a learned readout the collapse zone is not anomalous at all*. The *magnitude* of the tuned-vs.-zero-shot repair gap, however, does not track z: across the 19 layer-points where both are measured (160M/410M/1B), the pooled rank correlation between z and the repair gap is $\approx$0 (Spearman r_s=-0.02; within-model directions conflict; committed analysis in the repository); the gap's size is governed by depth and model as much as by alignment (§6). Together with CCR-5 restoration, flat probing, and the content-level correlation of Appendix D.14, this closes the functional side of masking-not-erasure: what the collapse degrades is specifically the *zero-shot* W_U readout (the channel the logit-lens family uses), while an inexpensive learned correction recovers decoding, consistent with a low-rank rotation of predictive structure away from the readout basis rather than its destruction.

E. Theoretical Dissociations and Mechanisms

E.1. Logit Lens Failure vs. Loss of Predictive Alignment

The logit lens evaluates absolute decodability (whether W_U maps a hidden state to the correct token); PGA evaluates distance-structure alignment. Figure 9 shows these dissociate: all models exhibit logit lens failure at intermediate depths, but large models maintain positive PGA through this zone: geometric structure remains aligned with the readout even when

absolute coordinates have drifted.

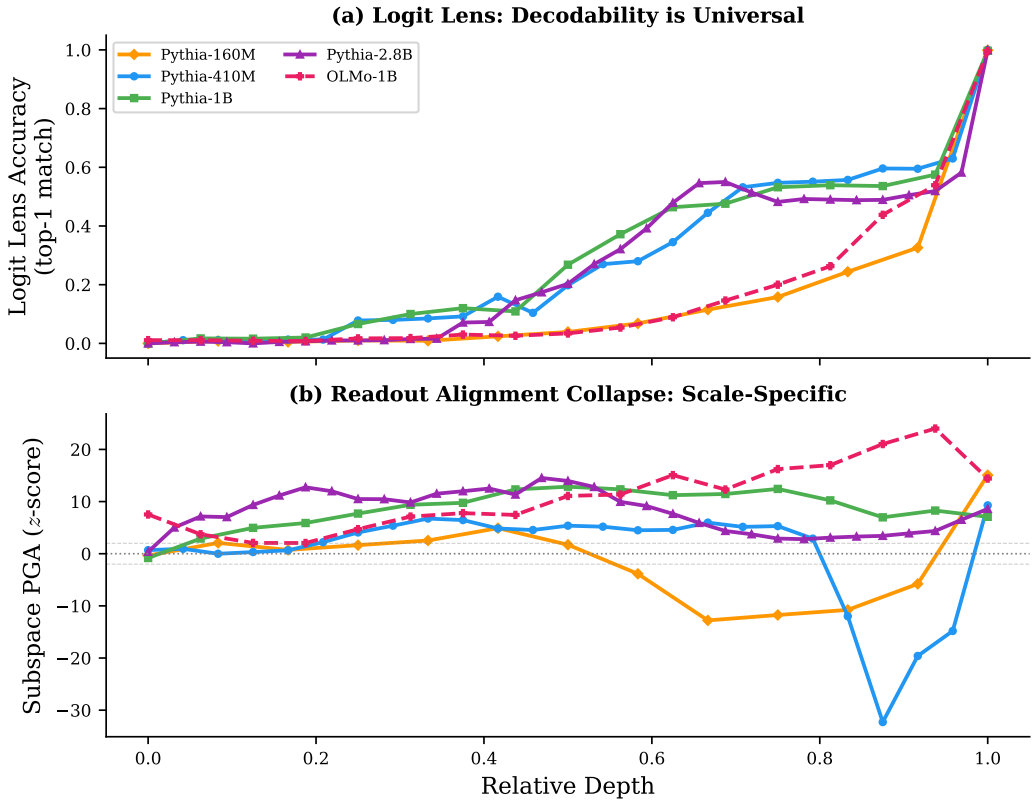

Figure 9. **Logit Lens vs PGA. (a)** Absolute decodability fails at intermediate depths in every model shown. **(b)** Intrinsic geometric alignment collapses only in small-capacity models. The dissociation confirms that geometric alignment with the readout and absolute decodability are distinct properties.

E.2. Probing: z versus Linear-Probe Accuracy

To check whether z tracks something that matters beyond geometry, we trained linear probes for AG News topic classification on per-layer hidden states of Pythia-410M and 1B, and compared probe accuracy to PGA z-scores (Figure 10).

In the geometrically stable Pythia-1B, z and probing accuracy track each other layer by layer (Spearman $r_s = 0.69$, $p = 0.002$): the same layers whose distance structure is most aligned with prediction also yield more linearly separable features for an unrelated downstream task. This is *not*, however, a generic layer-quality score, and we do not claim z as a practical layer selector. In a controlled comparison across the suite and three tasks (Appendix E.3), the task-free signals that best predict probe-ability (logit-lens accuracy, effective rank, isotropy) are all monotonic in depth, whereas z is not; z's agreement with probe-ability is task- and scale-dependent. That dissociation is exactly the point: z tracks the specific W_U-alignment geometry, not the depth-driven maturity that probe-ability follows, so it captures something the monotonic baselines do not.

In Pythia-410M, the collapse zone tells the complementary story. Probe accuracy stays around 0.85 even at layers where z falls well below zero. The information needed for the probe is still present in those layers; what changed is its alignment with W_U. This is direct evidence for masking, not erasure (§5.1). PGA thus measures alignment with W_U, not raw information content: a layer can be highly probe-able yet W_U-misaligned. Its functional relevance is correspondingly specific: to W_U-based readouts (the logit lens above all) and to the cross-model convergence observed in $\mathcal{V}_k$ (Appendix E.7), not to a layer's raw probe-ability.

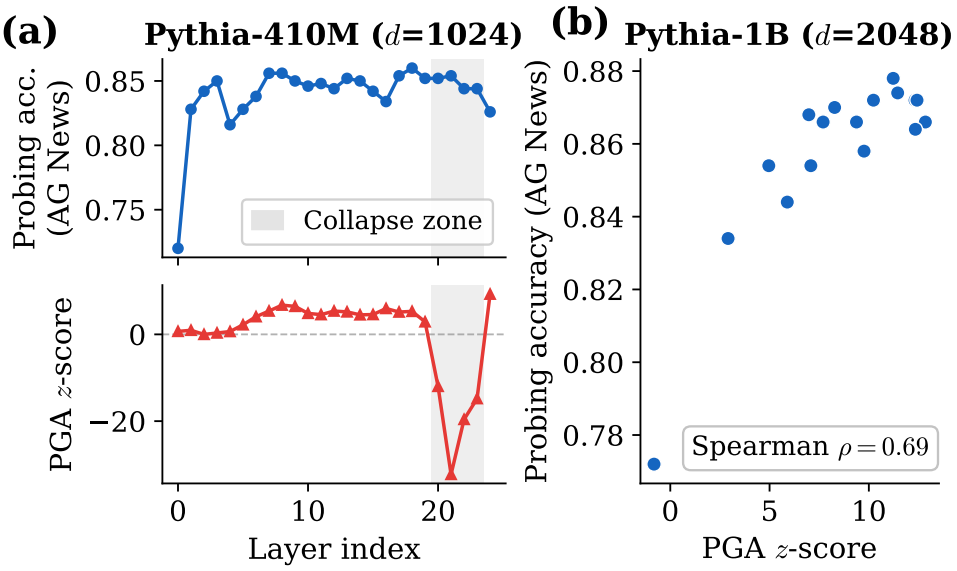

Figure 10. **Probing vs. predictive alignment. (a)** Pythia-410M: probing accuracy stays high through the collapse zone (masking, not erasure). **(b)** Pythia-1B: probing accuracy and z correlate (Spearman $r_s = 0.69$, $p = 0.002$; one model and one task).

E.3. Layer Selection: z vs. Task-Free Baselines

A control for the probing analysis (Appendix E.2): if z were a practical task-free layer-selection signal, it should predict which layer probes best, and do so better than simpler signals. We test this directly. For Pythia-410M/1B/2.8B/6.9B and three tasks (AG News, SST-2, DBpedia-14) we train per-layer linear probes (mean-pooled, CCR-1 features) at $n_{\text{train}} \in \{50, 100, 200, 500, 2000\}$, and for each task-free signal compute the per-layer Spearman correlation with probe accuracy across non-embedding layers. Table 15 reports the mean over tasks at n_{train}=2000; the low-resource sweep gives the same ordering.

Table 15. **Per-layer correlation of task-free signals with linear-probe accuracy** (Spearman r_s, mean over AG News / SST-2 / DBpedia-14, n_{train}=2000). z is *not* a superior predictor of probe-ability: the depth-monotonic signals (logit-lens accuracy, RankMe, isotropy) track it as well or better, and z's agreement is task/scale-dependent, weakening to slightly negative at 2.8B/6.9B. `exp22`.

MODEL	z (OURS)	LOGIT-LENS	RANKME	ISOTROPY
P-410M	+0.27	+0.53	+0.39	+0.09
P-1B	+0.52	+0.61	+0.71	+0.70
P-2.8B	−0.13	+0.62	+0.23	+0.12
P-6.9B	−0.13	+0.56	+0.22	−0.06

The signals that best track probe accuracy are all monotonic in depth, and probe accuracy is itself roughly depth-monotonic, so they agree largely by construction. z is non-monotonic (it peaks at intermediate depth and can collapse late), and its correlation with probe accuracy is task- and scale-dependent. We read this as confirming, not undercutting, the metric's interpretation: z measures the specific W_U-alignment geometry rather than the depth-driven maturity that probe-ability follows, which is exactly why a layer can be highly probe-able yet W_U-misaligned (the masking dissociation, Appendix E.2). We therefore position z as a signal of W_U-readout grounding, not as a general probe-layer selector.

E.4. Spectral Metrics Do Not Track Predictive Alignment

Spectral metrics characterize the *intrinsic* shape of a layer's geometry (its rank, anisotropy, and spectral decay) without reference to W_U. PGA instead measures an *extrinsic* property: how that geometry is oriented relative to the readout subspace. The two need not move together. Across the models tested, no spectral metric consistently tracks z (Figure 11). RankMe comes closest: positively correlated with z in all five models with committed correlations (r_s=0.31–0.71), but significant in only three, non-significant in Pythia-160M and OLMo-1B, and blind to *where* the collapse occurs; the remaining metrics

(stable rank, participation ratio, α-ReQ, isotropy, condition number) are model-dependent in both sign and size. The cleanest illustration is Pythia-410M's collapse zone (L20–L23), where α-ReQ is indistinguishable from surrounding layers while z swings by more than 40 standard deviations: nearly the same intrinsic shape, a very different orientation relative to W_U. The claim is not that function is independent of geometry (it is downstream of it) but that the spectral descriptors in common use do not capture the specific relationship between a layer's geometry and the unembedding matrix, which is what determines whether W_U-based tools apply.

The dissociation occurs because spectral metrics measure the *intrinsic shape* of the coordinate space (variance decay across principal components), while PGA measures its *extrinsic orientation* relative to the readout (W_U). When a small model takes the geometric detour, its manifold rotates away from the readout as a near-rigid transformation, so the eigenvalue density (α-ReQ) stays flat even as functional alignment collapses.

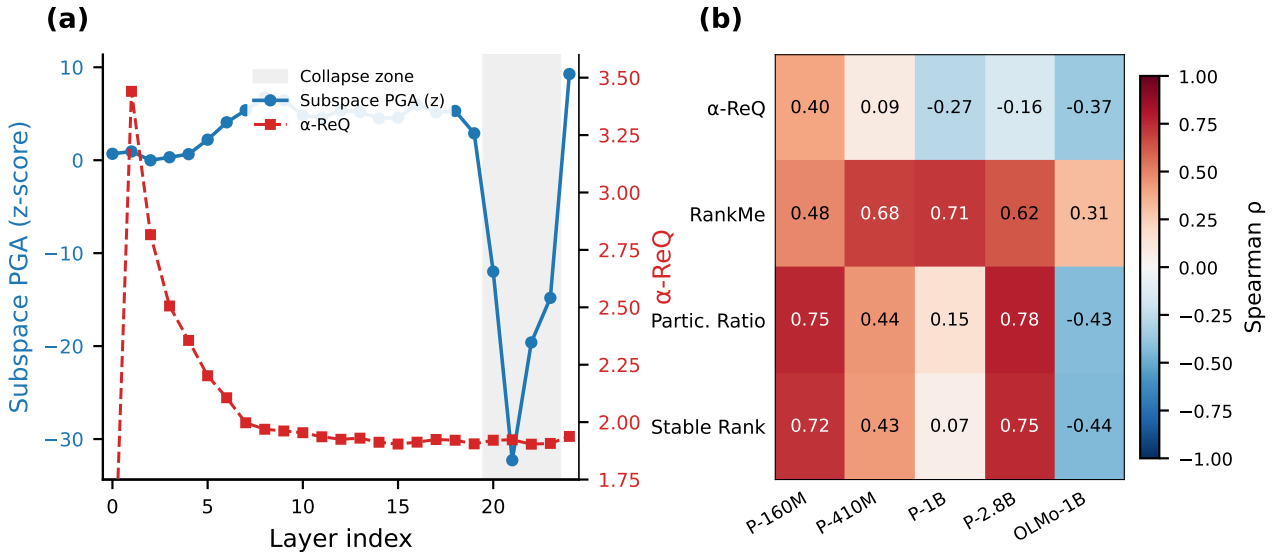

Figure 11. **Spectral Metrics Do Not Track Predictive Alignment. (a)** PGA z-score and α-ReQ across layers (Pythia-410M). α-ReQ is flat through the zone where predictive alignment collapses (shaded) despite a >40-standard-deviation PGA swing. **(b)** Spearman correlation between each spectral metric and PGA z-scores across layers. Correlations are model-dependent; no metric is a consistent predictor across the suite.

E.5. KL Projection Analysis

To test whether prediction-irrelevant directions carry functional information, we projected hidden states onto the readout subspace and measured the effect on next-token predictions. Late layers of small models are *least* sensitive (KL $= 3.6$, 32.5% top-1 preserved at Pythia-410M L21) and mid-layers most sensitive (KL $= 8.8$ at L8): the affected layers reorganize geometry rather than compute new predictions. The same pattern replicates in the second collapse model: Pythia-160M's collapse zone (L9–L11) shows KL $= 1.9$–2.9 with top-1 agreement 23–42%, versus KL $= 4.3$–9.8 and top-1 $\leq 6\%$ at its mid layers. Pythia-1B at a comparable late layer (L12, ≈75% depth) shows KL $= 4.2$ but only 12.5% top-1 agreement, so large models' non-predictive computation there carries more functional weight.

E.6. Theoretical Connections

Computational mechanics (Crutchfield & Young, 1989) predicts that optimal prediction groups histories with identical conditional futures; positive PGA is consistent with this at mid-layers, while the late-layer collapse in small models marks an empirical boundary the theory does not specify. Zhao et al. (2024)'s implicit-alignment theorem predicts collinear representations for contexts sharing next-token support: true at most layers, but broken where anisotropy overwhelms the predictive signal. Under Saurez et al.'s invariant-subspace theorem, the architectural necessity of W_U-aligned features appears fully satisfied only at the final layer.

The information bottleneck (Tishby et al., 1999; Shwartz-Ziv & Tishby, 2017) seeks maximally compressed representations

that preserve task-relevant information. Our dynamics show *selective* compression: continued training compresses small models' late-layer geometry into prediction-irrelevant directions, reducing functional alignment even as loss improves: compression, but not toward prediction, a distinction the bottleneck framework does not draw.

The linear representation hypothesis (Park et al., 2024) assumes that geometric proximity reflects functional similarity. Our results qualify this: it holds at most layers of all models, but at late layers of small models, the dominant geometric directions are orthogonal to prediction. Geometric proximity there reflects non-predictive similarity.

E.7. Cross-Model Convergence in the Readout Subspace

If representation geometry encodes structure that converges across architectures (Huh et al., 2024; Lee et al., 2025), and the readout subspace is the part of that geometry tied to a specific function, then convergence should be most pronounced there. We test this with cross-model *representational similarity analysis* (RSA; Kriegeskorte et al., 2008), which compares two representations by correlating their matrices of pairwise distances: for each pair from {Pythia-1B, OLMo-1B, Phi-1.5}, we compare the Spearman correlation of pairwise distances in the full $2{,}048$-dimensional space against the same correlation in the 100-dimensional readout subspace.

The 100-dimensional readout-RSA exceeds the random-subspace null in all three pairs (Figure 12), so the predictive readout directions carry more cross-model agreement than arbitrary directions do. Whether the readout *concentrates* that agreement beyond the full space is a weaker, inconsistent signal: the readout-RSA exceeds the full $2{,}048$-dimensional RSA only at some depths, most clearly for the most similar pair (Pythia-1B vs. OLMo-1B at the final layer, readout-RSA 0.76 vs. full 0.64); for the pairs involving Phi-1.5 the margin is small and the absolute cross-model agreement is low. We therefore report the concentration effect as a preliminary, single-scale observation rather than a general pattern. Even the robust part, that what independently trained models agree on geometrically is captured above chance by predictive directions, is tested at one scale with similar-capacity models; a stronger causal test (e.g., model stitching; Bansal et al., 2021) and a wider model set are left to future work.

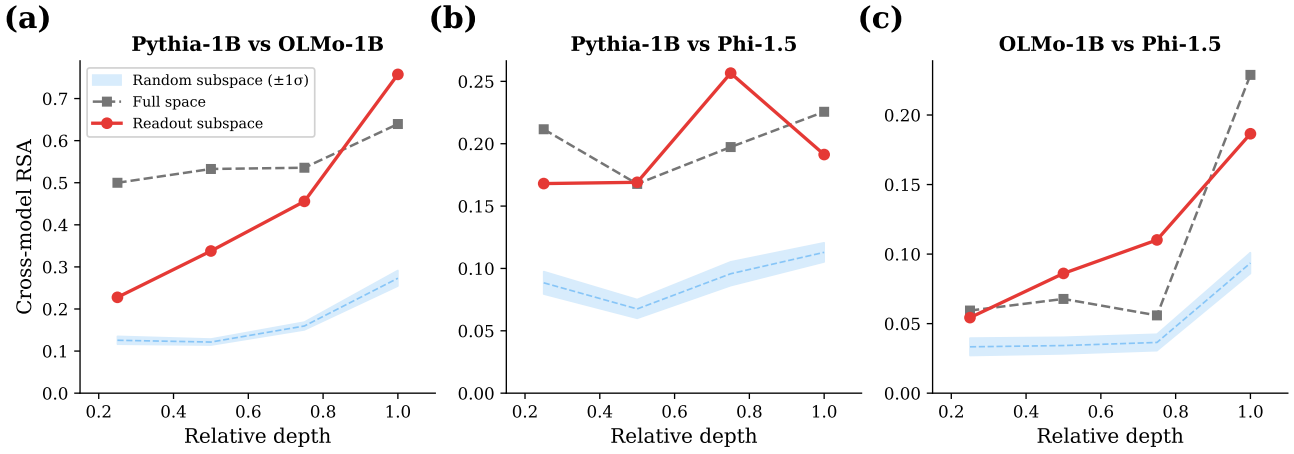

Figure 12. **Cross-model RSA** for each model pair: readout-subspace RSA (red) vs. full-space RSA (grey) vs. dimension-matched random-subspace null (blue band).

F. Cross-Family Extension

To probe how far the regime structure generalizes, we ran PGA on nine models beyond the three in Appendix E.7, spanning the Qwen-2.5, Llama-2/3.2, OLMo, and TinyLlama families (same protocol: k=100, 100-subspace null, CCR-1, final-token states, shared contexts; Llama weights from ungated re-uploads identical to the official releases).

What generalizes Two findings hold across families (Table 16): (i) *intermediate* geometry concentrates in the readout subspace: every model tested has an aligned interior at its peak (positive z at mid-depth), with one narrow-model qualification below; and (ii) *wide* models do not develop the intermediate-layer detour. No untied non-Pythia model with $d < 2048$ exists (small models tie embeddings to save parameters), so the narrow-model *intermediate* collapse is established within the controlled Pythia suite rather than re-tested cross-family. Directionally consistent with it, however, the one narrow

cross-family model available, Qwen2.5-0.5B (d=896, tied), dips to negative interior z around 60% depth (trough -3.8) before its family-specific final-layer collapse; we keep this as supporting rather than primary evidence because tying confounds the comparison.

A distinct final-layer collapse in some families Separately, the Qwen-2.5 family (all sizes) and Gemma-2 show a *final*-layer collapse: an aligned interior followed by a strongly negative z at the last layer; Llama-2/3.2, OLMo, TinyLlama, Phi, and Pythia do not. This effect is **not** explained by the obvious candidates:

- **Not embedding tying.** Llama-3.2 (tied) is stable (z_{final}=+13 to +17); Qwen-7B (untied) collapses (z_{final}=-6.9).
- **Not a measurement artifact.** For all nine cross-family models tested (including Gemma-2-2B recomputed post-norm) and the two Pythia anchors, the final-token `output_hidden_states[-1]` is bit-identical to the captured `lm_head` input (post-final-norm), so the metric is computed on exactly the state W_U reads; the collapse is in the readout-relevant geometry, not a pre-norm convention.
- **Not the subspace dimension k.** Qwen's final-layer z stays negative across $k \in \{50, 100, 200\}$ and grows *more* negative as k increases, the opposite of an under-resolved readout, which would recover as k grows.
- **Not vocabulary size or W_U rank.** Llama-3.2 (vocab 128k, stable) and Qwen (vocab 152k, collapse) are too close for vocabulary to separate them, and W_U spectral concentration does not order the groups (Llama-2 has the lowest top-100 cumulative variance, 0.12, yet is stable).

The final-layer collapse is thus a genuine, family-specific phenomenon, distinct from the width-driven *intermediate*-layer detour that is this paper's focus. We surface it as an open observation; its mechanism (plausibly tied to the output distribution or training of these specific families) is left to future work.

Table 16. **Cross-family PGA.** All models have an aligned interior; z_{final} is the final-layer z (k=100). Wide models are stable, while Qwen-2.5 (all sizes) and Gemma-2 show a distinct final-layer collapse absent in Llama/OLMo/TinyLlama, not explained by tying, width, vocabulary, or the metric. Values come from this run; independent runs of the same protocol place Gemma-2-2B's final-layer z between -13.6 and -17.6, the spread expected when a 100-draw Monte-Carlo null sits in the denominator, so the sign and not the magnitude is the reportable quantity.

MODEL	FAMILY	d	VOCAB	TIED	z_{final}	PATTERN
QWEN2.5-0.5B	QWEN	896	152K	Y	-10.8	FINAL-COLLAPSE
QWEN2.5-1.5B	QWEN	1536	152K	Y	-11.1	FINAL-COLLAPSE
QWEN2.5-3B	QWEN	2048	152K	Y	-8.4	FINAL-COLLAPSE
QWEN2.5-7B	QWEN	3584	152K	N	-6.9	FINAL-COLLAPSE
GEMMA-2-2B	GEMMA	2304	256K	Y	-13.6	FINAL-COLLAPSE
LLAMA-3.2-1B	LLAMA	2048	128K	Y	$+13.3$	STABLE
LLAMA-3.2-3B	LLAMA	3072	128K	Y	$+16.9$	STABLE
LLAMA-2-7B	LLAMA	4096	32K	N	$+5.1$	STABLE
TINYLLAMA-1.1B	LLAMA	2048	32K	N	$+10.9$	STABLE
OLMO-7B	OLMO	4096	50K	N	$+30.7$	STABLE

G. Pre-Registered Readout-Reliability Companion Analysis

Provenance note. Raw per-cell JSONs are committed under `results/readout_tools/exp41_readout_reliability/`, including the post-hoc-rescued 70M final-checkpoint and OLMo mid-checkpoint files; three cells (70M pre-onset, OLMo final, 2.8B) are recorded from run-time logs rather than committed raw files, and are marked accordingly.

This appendix documents the pilot cells of a pre-registered companion analysis of *functional* readout reliability, summarized in §6. Predictions were frozen in a version-controlled pre-registration file before any measurement (the measurement script refuses to run in its absence, as in Appendix D.6); the full reliability treatment (when intermediate-layer W_U readings can be trusted, what predicts failure, and how to repair it) is companion work.

Design Zero-shot logit lens vs. per-layer tuned lens (the recipe of Appendix D.15) on the shared OpenWebText contexts; errors are KL(lens || final output) and top-1 agreement with the model's own final prediction, pooled over each model's

pre-registered late-layer band (relative depth 0.65–0.95, or the frozen collapse windows). The headline statistic is the qualitative *flip rate*: the fraction of contexts whose final top-1 token is absent from the zero-shot lens top-5 but present in the tuned-lens top-5. Frozen predictions: **P1** (regime excess: zero-shot error of collapsing models exceeds stable models at matched relative depth, with no comparable tuned-lens excess); **P2** (within-model onset contrast, the primary test, immune to cross-model quality confounds: the excess at step 143k is absent at step 64k, pre-onset); **P3** (content: larger zero-shot error increment for copyable than for frequency-matched parametric targets in collapse bands); plus an elevated-flip-rate prediction, each with an explicit falsifier, reported below as fired or not.

Table 17. **Readout-reliability pilot cells** (band-pooled over each model's pre-registered late-layer band; shared OpenWebText contexts; step 64k is pre-onset, 143k the final checkpoint). Flip rate = final top-1 absent from the zero-shot lens top-5 but present in the tuned-lens top-5. `exp41`.

MODEL	STEP	KL (ZERO-SHOT)	KL (TUNED)	TOP-1 (ZS/TUNED)	FLIP RATE
P-160M	64K	1.25	0.52	0.63/0.74	7.4%
P-160M	143K	3.98	1.09	0.30/0.70	37.6%
P-410M	64K	1.30	0.58	0.57/0.74	9.9%
P-410M	143K	1.69	1.12	0.59/0.72	8.4%
P-1B	64K	1.23	0.59	0.60/0.74	6.5%
P-1B	143K	1.39	0.70	0.59/0.72	6.4%
P-1.4B	143K	1.83	0.75	0.61/0.77	11.1%

Outcomes, reported per the frozen falsifiers Table 17 gives the band-pooled cells. **P2 is confirmed for Pythia-160M**: the flip rate rises $7.4\% \rightarrow 37.6\%$ from pre-onset to final checkpoint and band zero-shot KL rises $1.25 \rightarrow 3.98$, while the stable control (Pythia-1B) is flat ($6.5\% \rightarrow 6.4\%$; KL $1.23 \rightarrow 1.39$). **P2 is refuted for Pythia-410M**: the deepest-collapsing model is functionally silent; its band metrics are indistinguishable from the stable control ($9.9\% \rightarrow 8.4\%$; KL $1.30 \rightarrow 1.69$). **P3's falsifier fired**: in the collapse bands, copyable targets decode *better*, not worse, than frequency-matched parametric targets under the zero-shot lens (e.g. Pythia-160M at 143k: zero-shot KL 3.93 copyable vs. 5.21 parametric), so the *geometric* content-selectivity of §5.3 does not transfer to zero-shot decoding error.

The interaction hypothesis and its out-of-sample cell A post-hoc hypothesis reconciles the split P2 outcome: functional zero-shot failure requires *both* geometric collapse *and* a concentrated W_U spectrum (Pythia-160M: top singular direction $\approx$91% of variance; Pythia-410M: k=100 covers only $\approx$30%; Appendix D.9), because a flat spectrum leaves sub-σ_k directions carrying the decode (cf. the leverage-not-exclusive reading of §2). This hypothesis was frozen in a pre-registered addendum and then tested out-of-sample on Pythia-70M (collapsing, near-rank-1 W_U): the predicted onset-timed flip-rate rise appeared ($\approx$+31 percentage points from pre-onset to final checkpoint), and a cross-recipe stable-arm control (OLMo-1B, the only non-Pythia family with public intermediate checkpoints) was pre-registered in a second addendum. Because W_U spectral concentration and model size are confounded in the public-model pool (the concentrated regime is occupied only by the two smallest Pythia models), isolating the collapse $\times$ spectrum interaction requires controlled training; this, and the repair analysis, belong to the companion work.

H. Experimental Details

Context Sampling We sample 1,000 texts from OpenWebText, filtering for ≥ 64 tokens and truncating to 512 tokens. To avoid autocorrelation between tokens within a sequence, we use only the *final-token* hidden state from each of the 1,000 independent contexts.

CCR-1 Estimation Requirements CCR-1 requires estimating the top eigenvector of the covariance matrix. With N=1,000 independent samples, this estimate is stable even at d=4096 (Pythia-6.9B), since only the leading eigenvector (not the full covariance) is needed.

Statistical Methodology Pairwise distance matrices violate independence assumptions required by parametric tests, so all calibration is empirical. Every z is a standard score against its own 100-draw random-subspace null (§2); uncertainty over contexts is quantified by bootstrap over contexts (Appendix D.11) and stability sweeps over N and seeds (Appendix D.12);

subset contrasts (the context split) use balanced-partition permutation nulls with pre-registered layer windows and frequency matching.

Compute and Hardware All experiments use PyTorch and HuggingFace Transformers. Total compute: approximately 60–80 A100 GPU-hours across the full Pythia suite, cross-architecture models, and multi-checkpoint training dynamics.