NOISE CEILINGS FOR REPRESENTATIONAL CONVERGENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

The Platonic Representation Hypothesis links increasing model alignment to a shared representation of the world. Yet a rising alignment score can also reflect more reproducible training outcomes. We develop a noise-ceiling analysis that separates training-seed reliability from consensus alignment. For mutual nearest-neighbor overlap and linear centered kernel alignment (CKA), normalized representation objects give an exact factorization of expected alignment, without assuming additive or isotropic noise. Across nine PolyPythia seeds at each of five scales, reliability accounts for 51.3% of the language–vision log-alignment trend; an independent OLMo suite gives 50.7%. Fixed-readout, capability-axis, and replicated-vision controls support a substantial reliability contribution. Fresh extraction of 105 training checkpoints reveals a sharper distinction: Pythia 14M raw CKA alignment rises by 89% while corrected consensus falls by 10%, and the measured training paths differ between Pythia and OLMo. Known-truth experiments expose the limits of correction with few seeds. Raw alignment also predicts held-out retrieval better than corrected alignment, distinguishing consensus from individual-model utility. The resulting protocol measures reproducibility, shared representational structure, and usefulness as separate properties, making convergence claims more precise and testable.

1 INTRODUCTION

The Platonic Representation Hypothesis (PRH) proposes that models trained on different data and modalities approach a shared representation of the world (Huh et al., 2024). Its empirical motivation is striking: better-performing models often agree more about which stimuli are similar. Yet a training recipe does not produce one representation. Different initializations and data sequences produce different models. A rising alignment score can therefore reflect either agreement between recipes or greater consistency within each recipe.

These explanations lead to different scientific conclusions. The intuition comes from noise ceilings in neuroscience: measured similarity depends on how reproducible the compared responses are. Here the repeated measurements are independently trained models. A useful geometric picture places every training run on a unit sphere. A recipe's mean lies inside that sphere: its length measures reproducibility, while its direction identifies the consensus. Figure 1 separates two changes that a single alignment curve confounds. Seed outcomes can tighten around an unchanged direction, or the direction itself can move toward another representation. Either change raises alignment.

We turn this ambiguity into a measurable decomposition. Mutual k-nearest-neighbor overlap is an inner product between normalized adjacency matrices; linear centered kernel alignment (CKA) is an inner product between normalized centered Gram matrices. In this space, a recipe has a mean object, whose direction describes its consensus and whose squared length measures its reliability across independent training runs. Expected cross-recipe alignment is the inner product of the two means. It therefore factors exactly into the cosine between their directions and the product of their lengths. This is a representation-level application of attenuation correction (Spearman, 1904), rather than a new general principle of correlation.

The contribution is an operational account of what seed-replicated alignment can identify. We establish the population identity, its finite-sample limitations, and the geometry of permutation calibration. We then measure how much reliability contributes to observed scaling trends, test

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

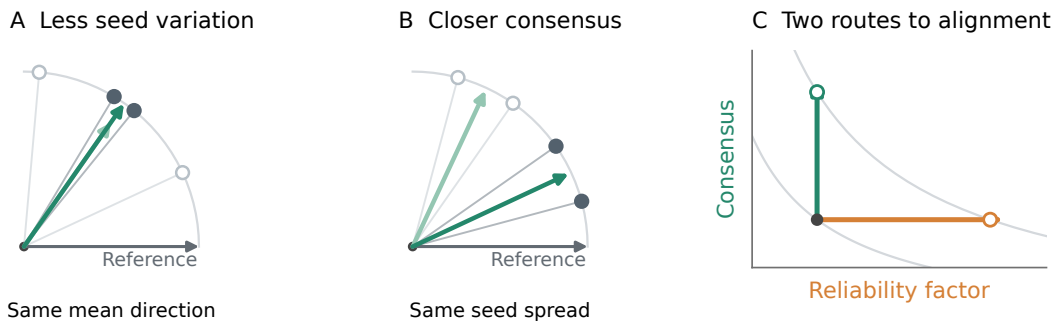

Figure 1: **Less noisy and more alike are different changes.** Illustrative unit-object geometry; each pair of dots denotes two equally likely seed outcomes, and its arrow denotes their mean. Pale open dots become dark filled dots. (A) Seed variation decreases while the mean direction stays fixed. (B) The mean moves toward the reference while seed spread stays fixed. (C) Greater reliability (right) and closer consensus (up) can produce the same increase in expected alignment. Gray contours satisfy $a = C\sqrt{\rho_A \rho_B}$. These schematics explain the measurements; the observed scale and training paths are shown separately.

whether the conclusions survive independent layer selection and fixed readouts, and examine what the correction means for a downstream task. The resulting protocol answers a question that a single alignment curve cannot: how much of a measured change concerns consensus, and how much concerns reproducibility?

The distinction matters empirically. In our main analysis, roughly half of the log-alignment trend from 14M to 410M parameters comes from language-model reliability. The remainder is positive, so reliability does not explain the entire observed trend. Across another language-model lineage and a replicated vision family, reliability remains a substantial contributor, but its share is not a universal constant. Our retrieval experiment also identifies a boundary of the method: a score that estimates consensus need not be the best predictor of the usefulness of individual trained models.

2 What alignment measures

Fix stimuli $x_1, \dots, x_n$, a feature readout, and a similarity metric. Let $F_f \in \mathbb{R}^{n \times d}$ contain one model's features. Its directed k-nearest-neighbor adjacency matrix A_f has exactly k ones per row and a zero diagonal. Neighbors are selected using cosine similarity, with a deterministic rule for ties. Its centered Gram matrix is $K_f = HF_fF_f^\top H$, where $H = I - \mathbf{1}\mathbf{1}^\top/n$. We exclude zero Gram matrices.

Lemma 1 (Unit representation objects)**.** *For a common n and k, mutual k-NN overlap equals $\langle A_f, A_g \rangle_F/(nk)$. Linear CKA equals $\langle K_f, K_g \rangle_F/(\|K_f\|_F \|K_g\|_F)$. Thus both scores are inner products $a(f,g) = \langle \mathbf{R}_f, \mathbf{R}_g \rangle$ of unit objects, with $\mathbf{R}_f = A_f/\sqrt{nk}$ or $K_f/\|K_f\|_F$.*

Proof. The adjacency inner product counts shared neighbors and $\|A_f\|_F^2 = nk$. The CKA statement follows directly from its centered-Gram definition. □

A recipe specifies architecture, training data, objective, optimizer, schedule, and checkpoint selection. Randomness in its training procedure induces a distribution over $\mathbf{R}$. For independent draws R, R' from recipe A, define

$$\mu_A = \mathbb{E}R, \qquad \rho_A = \mathbb{E}\langle R, R' \rangle = \|\mu_A\|^2. \tag{1}$$

The expectation is over fresh training randomness, not over different inputs to a fixed model. Reliability lies in $[0, 1]$ because the objects have unit norm. The consensus is an average of *normalized graphs or kernels*; it need not be the representation of a realizable model or an average feature vector.

Proposition 1 (Exact attenuation)**.** *For independent draws from two recipes with $\rho_A, \rho_B > 0$,*

$$\underbrace{\mathbb{E}\, a(R_A, R_B)}_{a_{AB}} = \langle \mu_A, \mu_B \rangle = \underbrace{\frac{\langle \mu_A, \mu_B \rangle}{\|\mu_A\| \|\mu_B\|}}_{C_{AB}} \sqrt{\rho_A \rho_B}. \tag{2}$$

Proof. Independence and bilinearity imply $\mathbb{E}\langle R_A, R_B \rangle = \langle \mathbb{E}R_A, \mathbb{E}R_B \rangle$. Applying the same argument to independent draws within a recipe gives Equation (1); dividing by the mean-object norms gives the result. □

The identity separates the direction of a recipe's mean from its concentration. It implies the *expected* ceiling $|a_{AB}| \le \sqrt{\rho_A \rho_B}$; an individual model pair can exceed this ceiling. At positive alignment and consensus, any scaling path satisfies

$$\log a_{AB}(N) = \log C_{AB}(N) + \tfrac{1}{2} \log \rho_A(N) + \tfrac{1}{2} \log \rho_B(N). \tag{3}$$

Applying the same linear regression operator to each term gives an additive decomposition of its fitted log slope. This separates components of an observed trend; it does not establish a causal effect of parameter count or determine the infinite-scale limit.

What a fixed reference identifies. For one fixed model and layer g, only

$$q_g(A) = \frac{\mathbb{E}_R \langle R, R_g \rangle}{\sqrt{\rho_A}} = \cos(\mu_A, R_g) \tag{4}$$

is identified. This is consensus-to-instance alignment, not a lower bound on the consensus of an unobserved reference recipe. If the reference layer changes with scale, $q_{g(N)}$ also follows a changing readout. We therefore report fixed final layers in addition to selected-layer results, and reserve C_{AB} for comparisons with seed replicates on both sides.

A complementary distance interpretation. The same construction separates expected disagreement into a consensus-distance term and two variance terms:

$$\mathbb{E}\|R_A - R_B\|^2 = \|\mu_A - \mu_B\|^2 + (1 - \rho_A) + (1 - \rho_B). \tag{5}$$

Furthermore, the mean of m independent objects satisfies $\mathbb{E}\|\bar{R}_m\|^2 = \rho + (1-\rho)/m$. These identities show why averaging runs reduces variation without making distinct recipe means coincide. They concern averages in object space, not the performance of a weight ensemble; their proofs are in Appendix A.

Permutation floors and the object being corrected. Chance similarity is separate from seed variability. Under a uniformly random permutation of stimulus labels in one object,

$$\mathbb{E}_\pi a(R_A, P_\pi R_B P_\pi^\top) = \begin{cases} k/(n-1), & \text{mutual } k\text{-NN}, \\ \operatorname{tr}(K_A)\operatorname{tr}(K_B)/[(n-1)\|K_A\|_F \|K_B\|_F], & \text{linear CKA}. \end{cases} \tag{6}$$

The CKA floor depends on effective rank. Floor calibration $\tilde{a} = (a - f)/(1 - f)$ and attenuation correction can be composed, but the calibrated consensus belongs to a *different* object space. Appendix A.1 gives an explicit unit-kernel construction, including varying CKA floors. It also distinguishes this calibration from normalized residualization. Degenerate zero-information cases and nonpositive estimated reliabilities require explicit handling; clipping them to a small positive number would manufacture a corrected score.

3 Estimation and experimental design

Finite samples and layer selection. For S seeds of recipe A, estimate reliability by the distinct-pair average

$$\hat{\rho}_A = \binom{S}{2}^{-1} \sum_{s<s'} \langle R_{A,s}, R_{A,s'} \rangle, \qquad \hat{a}_{AB} = \frac{1}{ST} \sum_{s,t} \langle R_{A,s}, R_{B,t} \rangle. \tag{7}$$

The component estimates are unbiased under independent draws, whereas their corrected ratio is generally biased. The $\binom{S}{2}$ pairs are dependent and do not constitute that many independent training runs. A finite-sample corrected estimate may exceed one. Appendix B gives the leading ratio-bias terms, including numerator–denominator covariance.

We select layer pairs by maximum mean raw alignment on a training subset of seeds, then evaluate alignment and reliability on disjoint seeds. With nine seeds we enumerate all 126 balanced 4/5 partitions, using the same partition at each model size. Each direction of a partition supplies one evaluation record. We average *logs* over records and references, so Equation (3) holds exactly for the reported finite-sample decomposition. The resulting estimand averages conditional readout-specific log quantities; it is not the log of a pooled arithmetic alignment. We also compute final-layer results with no selection.

Every range over seed partitions is a *sensitivity range*, not a confidence interval. Resampling seed identities before splitting them can put the same original run into both selection and evaluation; excluding diagonal reliability pairs does not repair that overlap. Our known-truth experiment tests a separate fixed-readout jackknife interval and explains why few-seed coverage requires empirical validation rather than an appeal to the population identity. Appendix B documents the corresponding diagnostic of the earlier analysis.

Models and stimuli. Our primary suite is PolyPythias: nine seeds at each of 14M, 31M, 70M, 160M, and 410M parameters (van der Wal et al., 2025; Biderman et al., 2023). Seeds 1–9 share the intended recipe; the original seed-0 models at the two smallest sizes use different recipes and are excluded. The seven fixed vision checkpoints are DINOv2-S/B/L (Oquab et al., 2024), OpenCLIP ViT-B/32 and ViT-L/14 trained on LAION-2B (Cherti et al., 2023), and ImageNet-21k-pretrained ViT-B/16 and ViT-L/16 (Dosovitskiy et al., 2021). We use WIT-1024 image–caption pairs, with DCI-1024 and WIT-4096 as stimulus-set checks (Srinivasan et al., 2021; Urbanek et al., 2024).

We further analyze three DataDecide seeds at each of eight OLMo scales from 4M to 1B, trained on Dolma 1.7 (Magnusson et al., 2025); five seeds of an ImageNet ViT-S/16 (Falk et al., 2025); GPT-2 small and medium reproductions with five seeds each (Karamcheti et al., 2021); and ten MultiBERTs seeds (Sellam et al., 2022). Parameter labels are nominal. Checkpoints are matched by step *within* each DataDecide size, not across sizes. DataDecide's approximately fixed token-to-parameter scaling path differs from Pythia's fixed training-token budget, so this is a replication across training regimes rather than a controlled architecture-only comparison.

Readout and protocol scope. We use first captions, masked mean pooling over language tokens, vision class-token readouts, and outlier clamping based on the mean per-stimulus 95th percentile of feature magnitudes. Mutual 10-NN uses row-normalized features. The checkpoint study additionally repeats language extraction in float32 with pinned commits. This is a PRH-style analysis, with explicit deviations from the released implementation: our cached language features use right padding and truncation at 256 tokens, vision readout sets can include a final normalization, and features are stored in half precision. The original CKA analysis did not normalize feature rows; we add a matched normalized-CKA control. These choices preclude claiming an exact reproduction of all PRH scores.

Status of the hypotheses. The earlier local protocol recorded tests of reliability (T1), lineage (T2), and modality (T3), together with disclosed pilot results. It was not a public, prospectively timestamped registration. The present revision changes aggregation and inference after inspecting those results, and labels the new validation, retrieval, and checkpoint studies exploratory. We report effect sizes and sensitivity rather than treating a corrected/raw slope ratio of 50% as a scientifically privileged acceptance threshold. The original protocol and results remain available in the supplement.

4 Results

4.1 Reliability explains a substantial part of the scaling trend

Table 1 collects the main controlled comparisons. Raw language–vision alignment rises from 0.0425 to 0.0778 between 14M and 410M parameters (Figure 2). Language-seed reliability rises from 0.238 to 0.432, with a maximum near 160M. The fitted log-alignment slope is 0.423 per decade of nominal

Table 1: **The reliability contribution survives independent checks.** The reliability column sums both available reliability terms. Corrected scores are consensus-to-instance for fixed vision references and two-sided consensus for the replicated ViT. The two blocks use different horizontal-axis units. Shares are descriptive point estimates; partition ranges are reported in the text and supplement.

Comparison / readout	Seeds	Raw	Reliability	Corrected	Share
Log slopes per decade of nominal parameters					
Pythia, independent layers	9	0.423	0.217	0.206	51.3%
Pythia, final layers	9	0.386	0.196	0.190	50.8%
OLMo, independent layers	3	0.441	0.224	0.217	50.7%
OLMo, final layers	3	0.271	0.180	0.091	66.4%
Pythia / ViT, independent layers	9×5	0.381	0.205	0.175	54.0%
Log slopes per 10 percentage points of OLMES-10 capability					
OLMo, independent layers	3	0.515	0.241	0.274	46.8%
OLMo, final layers	3	0.286	0.182	0.104	63.7%

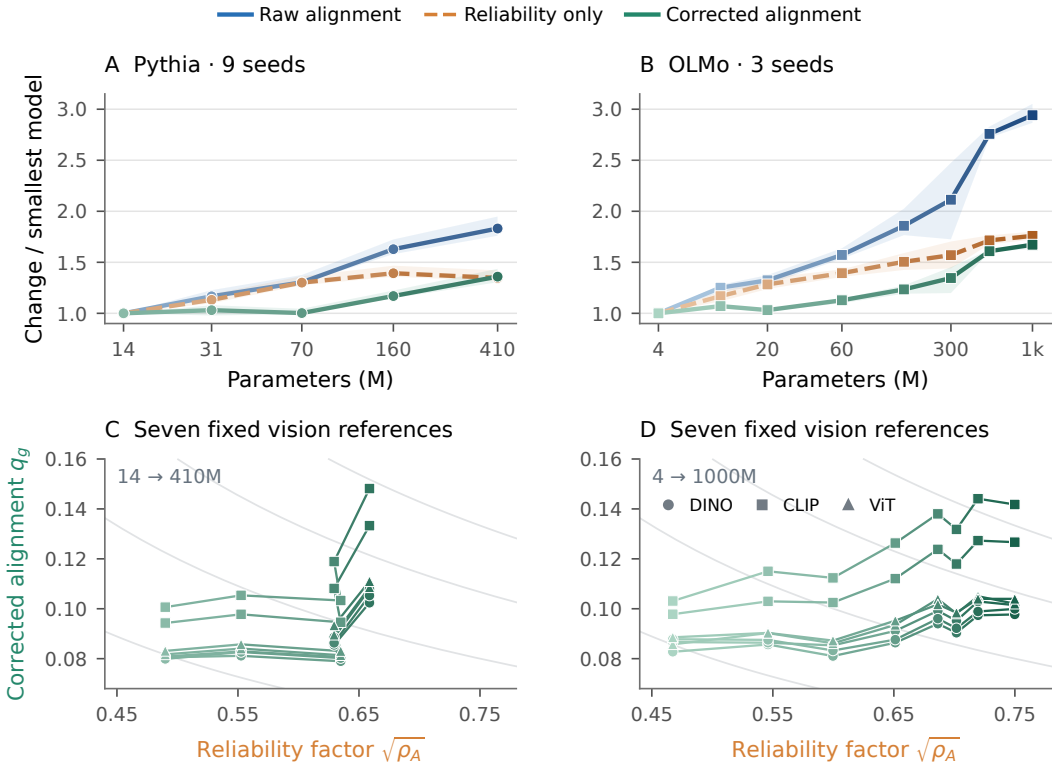

Figure 2: **How much apparent convergence survives the noise ceiling?** (A,B) Raw and corrected growth share the same axes and smallest-model baseline. The dashed curve holds corrected alignment fixed and changes only reliability: $a_{\mathrm{rel}}(N)/a(N_0) = \sqrt{\rho_A(N)/\rho_A(N_0)}$. The three curves obey raw growth = reliability growth × corrected growth. Shading is the full selector-partition range, not a confidence interval; central values average logs over folds, seven references, and partitions. (C,D) Seven separate scale paths at fixed final layers, one per vision checkpoint. Darker shades indicate larger models: blue is raw alignment, orange reliability, and green corrected alignment; shapes identify vision families. Early paths are mostly horizontal; later direction changes depend on the reference. Gray curves are equal-raw-alignment contours. Unlike (A,B), these panels do not select layers.

parameters. Its reliability component is 0.217, and its corrected component is 0.206. Thus reliability accounts for 51.3% of this trend. The full partition range is 46.7%–55.5%; every partition has a

positive corrected fitted slope. This is evidence that both factors matter in the measured suite, rather than a population significance test.

The fixed final-layer analysis gives a similar reliability share, 50.8%, without selecting a vision layer that can change with language-model size. The common-axis comparison makes the magnitude concrete: from 14M to 70M, raw alignment increases by about 30%, while corrected alignment changes by less than 1%. Across the full 14M–410M range, raw growth is about 83% and corrected growth about 36%. Thus *larger models are less noisy* explains the smaller range well, but does not explain the entire range.

An exploratory split at 70M assigns 99.1% of the fitted slope to reliability below that size and 9.7% above it; corrected slopes are 0.003 and 0.396. Each segment contains only three scales and does not identify a universal transition. Figure 2C–D further shows why one path is insufficient: fixed-reference trajectories separate by vision family, with CLIP references following different directions from DINO and supervised ViT references. The earlier layerwise profiles remain in Appendix C.1.

Independent language lineage. For DataDecide, we select layers on one DCI seed and evaluate them on the other two WIT seeds, repeating over all three possible selector seeds. Selection is therefore independent in seed identity as well as stimulus set. The reliability share is 50.7% (selector range 49.9%–52.4%). At fixed final layers it is 66.4%. A substantial reliability contribution replicates, but the numerical share depends on the readout. Three seeds support only a limited assessment of between-run variation; the narrow selector range is not evidence of a narrow population interval. Appendix D gives the full comparison.

Capability rather than parameter count. We match the published DataDecide OLMES-10 macro-average primary metric to the exact 24 analyzed seed/checkpoint combinations. Regressing log alignment on mean capability, rather than log nominal parameters, assigns 46.8% of the slope to reliability with independent layer selection and 63.7% at fixed final layers. The effect therefore persists on a direct capability axis. This observational comparison does not separate capability from size or compute; all eight scales and the published aggregate metric are retained (Appendix D.1).

Replicates on both sides. With nine language seeds and five vision seeds, we evaluate all $126 \times 10 = 1260$ combinations of balanced seed partitions, shared across language scales. The log-slope terms are 0.204 for language reliability, 0.001 for vision reliability, and 0.175 for consensus alignment, summing to the raw slope 0.381. Both reliability terms are retained because selected vision layers can differ across scales. The total reliability share is 54.0%, with a wider partition range of 34.7%–67.3%. The estimated two-sided consensus score rises from 0.098 to 0.126. Correction changes the interpretation of the score without making the cross-modal consensus close to one.

4.2 Training paths separate learning progress from alignment

Scale compares different recipes; training follows one recipe through time. We therefore extract 105 published checkpoints afresh: nine seeds each of Pythia 14M and 70M and three seeds of OLMo 60M, at five steps per run. We pin checkpoint commits, retain float32 outputs, and use fixed final layers for both modalities. This makes the trajectory a change in the measured recipe, without a moving layer-selection rule. Appendix E gives every checkpoint and the metric and precision controls.

The paths differ across recipes. Both Pythia sizes reach their largest sampled mutual-10-NN consensus at step 16,000, then retreat by 10.2% and 13.3%, respectively. OLMo instead reaches its largest value at the final checkpoint. The Pythia retreats remain without clamping and under normalized CKA, although their magnitudes and peak steps differ. For Pythia 70M, caption negative log likelihood improves from 5.349 at step 16,000 to 5.300 at the end while neighbor consensus falls. Better likelihood fit therefore need not move a language representation toward the visual consensus.

CKA gives a particularly direct demonstration of attenuation. From initialization to the final checkpoint, raw Pythia 14M alignment rises by 89%, while corrected consensus falls by 10%; for 70M, raw alignment rises by 49%, while corrected consensus rises by only 6% (Figure 3B). OLMo increases in both quantities. Training can thus change reliability, consensus, or both, and their balance depends on the recipe and metric. These five-checkpoint paths establish neither a universal learning phase nor an information bottleneck.

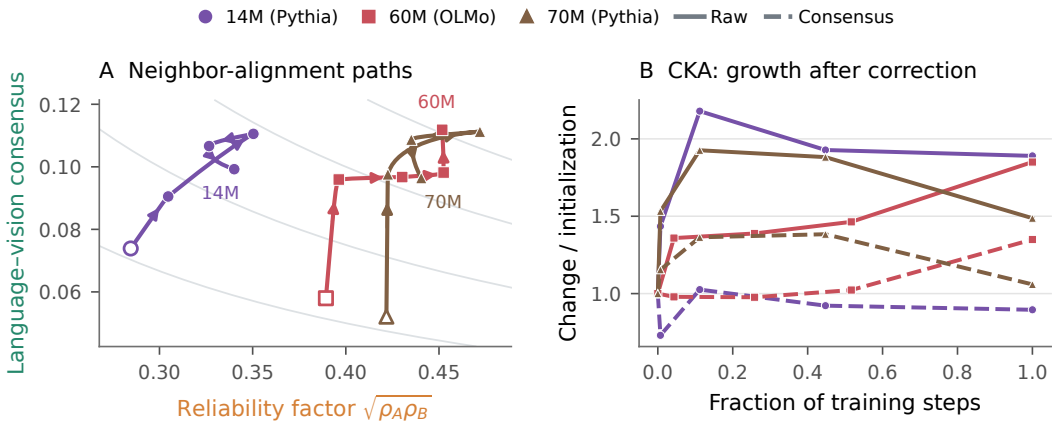

Figure 3: **Training traces a path, not a monotone convergence curve.** Fresh float32 readouts from five checkpoints of each of 21 training runs, compared with five fixed ViT seeds. (A) Mutual-10-NN paths: markers are measured checkpoints, open markers indicate initialization, and arrows follow time. Curves interpolate the observations for readability. Gray contours hold raw alignment constant. Colors and marker shapes identify the three recipes. (B) Normalized CKA: raw (solid) and corrected (dashed) changes share the same initialization baseline. Steps are divided by each recipe's final step, not matched by compute. All readouts are fixed final layers and all sampled checkpoints are retained.

The reference also matters. Cross-checkpoint consensus with each recipe's own final state increases across the sampled steps in all three families, even where agreement with vision recedes. For Pythia 70M, consensus with its final state rises from 0.774 at 16,000 steps to 0.940 at 64,000 steps. We exclude pairs sharing a seed identity to avoid coupling through training history; agreement with the final state itself is one by construction. Progress toward one endpoint can therefore coexist with movement away from another.

4.3 The result depends on which representational structure is measured

Across seven mutual-neighbor specifications, reliability contributes 37.8%–66.8% of the fitted trend (Table 3). Changing k, clamping, stimulus set, and permutation calibration changes the magnitude but leaves a positive corrected fitted slope. The contribution is appreciably smaller for CKA. At fixed final layers, row normalization changes CKA's reliability share from 14.9% to 14.3%. This matched comparison does not attribute the gap between CKA and nearest-neighbor overlap to the omitted normalization. Local neighbor identities and global kernel geometry have different seed sensitivities.

Increasing the stimulus set also changes the quantity being measured. At fixed k, WIT-4096 asks about a finer neighborhood fraction than WIT-1024. In the original matched-reference analysis at 410M, raw alignment falls from 0.078 to 0.053, reliability from 0.43 to 0.40, and the corrected score from 0.119 to 0.084. Reliability correction therefore does not eliminate this sample-set dependence. This supports a finite-sample sensitivity finding, not an extrapolation to millions of images or a claim that the entire effect has been isolated from caption and neighborhood-resolution changes.

Lineage and modality. Figure 4 compares complete recipes against a common anchor. The original selected-layer analyses found high agreement between adjacent sizes within a language-model lineage and lower agreement between lineages. We retain those comparisons, and add fixed final-layer estimates that avoid selection leakage in three-seed OLMo families and exclude equal published seed labels within each lineage. At fixed final layers, Pythia 160M versus 410M has $\hat{C} = 0.853$, while 410M Pythia versus GPT-2 small has $\hat{C} = 0.596$ and versus the replicated ViT has $\hat{C} = 0.127$. The original adjacent-size values near 0.95 thus describe selected readouts, not all layers. The complete matrix and the historical deduplication and initialization/order analyses are in Appendix F. These recipe comparisons jointly change training data, tokenizer, architecture, and objective; they are not causal estimates of isolated provenance effects.

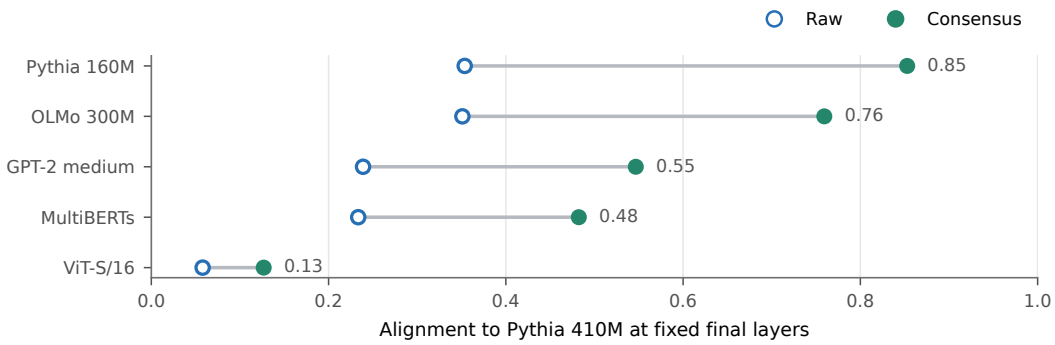

Figure 4: **Correction changes the levels without erasing recipe differences.** Every row compares a replicated partner with the same Pythia 410M anchor at fixed final layers. Open points are raw alignment; filled points are two-sided consensus estimates. Rows compare complete recipes rather than isolated causal changes in data, architecture, or modality. The full matrix is in Appendix F.

Table 2: **Raw alignment predicts this retrieval task better than consensus correction.** Leave-one-reference-out errors in percentage points (pp) of recall@1 over 35 dependent size–reference cells. Lower is better. All formulations are retained; no independence-based significance claim is made.

Predictor	MAE (pp)	RMSE (pp)
Mean only	3.87	5.13
Log parameters	3.10	3.84
Log raw alignment	1.80	2.41
Log corrected alignment	2.16	2.63
Log raw alignment + log reliability	1.85	2.37

4.4 An exact correction still needs finite-sample validation

We test the estimator on populations with analytically known means, reliabilities, and consensus alignment (Appendix G). At nine seeds and reliability 0.25, isotropic high-dimensional noise gives a median absolute error of 0.030 after correction, compared with 0.300 when raw alignment is used to estimate consensus. For angular noise on a circle with the same reliability and consensus, the corrected median error is 0.303, and nominal 95% interval coverage is only 70.7%. The corresponding high-dimensional coverage is 96.2%. Thus reliability and seed count alone do not determine estimation quality; the distribution of seed variation also matters. These controls motivate explicit validity checks and the conservative interpretation of our empirical partition ranges.

4.5 Consensus correction is not a universal predictor of utility

To connect representational measurements to a concrete use, we fit caption-to-image ridge maps at fixed final layers. Exact-caption deduplication leaves 1023 pairs, divided into 768 training and 255 test pairs. Each map uses train-fitted centering, row normalization, and a fixed ridge penalty of one, with no tuning. We measure alignment and seed reliability on the training stimuli only, then test whether these quantities predict seed-averaged recall@1 on the held-out pairs. Each prediction fold holds out one of the seven vision references and fits on the other six, giving 35 size–reference cells in total.

Raw alignment gives a mean absolute prediction error of 1.80 percentage points, versus 2.16 for the corrected score and 1.85 for raw alignment plus reliability. Mean-only and parameter-count baselines give 3.87 and 3.10 points (Table 2). We report all folds and do not treat these dependent cells as independent evidence for a significance test. The result does not support replacing raw alignment with corrected alignment for this retrieval task. It does support reporting both: an individual trained model's usefulness can depend on its recipe's reliability, while the scientific question about consensus deliberately factors reliability out.

5 Relationship to prior work

PRH links convergence to model capability (Huh et al., 2024). Recent work questions its metric calibration (Gröger et al., 2026), sample-size dependence (Koepke et al., 2026), distinction between structure and geometry (You et al., 2026), conceptual scope (Landau & Keren, 2026), and relation to reasoning (Usama & Chang, 2026). Identifiability arguments address idealized shared representations (Ziyin & Chuang, 2025), while dataset and task overlap can themselves induce similarity (Li et al., 2025). Our measurements address training-randomness attenuation within finite, specified recipes; they neither establish nor refute an unrestricted asymptotic PRH.

The closest conceptual connection is the signal-plus-noise account of Bangachev et al. (2026); Umar & Laio (2026) also study alignment through data signal-to-noise ratio and sample size. We directly estimate a training-seed mean and its concentration, rather than identifying seed variability with data noise. Earlier work documents seed-dependent representations and analyzes similarity measures (Raghu et al., 2017; Morcos et al., 2018; Kornblith et al., 2019; Mehrer et al., 2020; Nguyen et al., 2021; Ding et al., 2021; Davari et al., 2023; Klabunde et al., 2025). Decodable information and finite-sample CKA bias concern additional dimensions of measurement quality (Harvey et al., 2024; Murphy et al., 2024).

Attenuation correction and noise ceilings are established tools in representational similarity analysis (Kriegeskorte et al., 2008; Nili et al., 2014; Schütt et al., 2023; Lin & Kriegeskorte, 2024). In particular, Qian et al. (2025) use split-half attenuation correction for decision-variable correlations, including splits over model hidden units and biological neurons. Their work is not limited to trial noise. Our distinction is the use of independent training runs as the sampling unit, the normalized-object treatment of the alignment metrics, and an explicit decomposition of observed scaling trends. We claim this combination and its empirical findings, rather than priority for attenuation correction itself.

6 Discussion

An alignment curve mixes consensus with reproducibility. A high score can reflect reliable training outcomes, and a low score can hide similar consensuses. Our scaling results contain a substantial reliability contribution and a positive corrected trend. Both belong in an account of representational convergence.

A useful reporting standard follows directly. An alignment study should specify the representation object, readout, stimulus set, and training-randomness unit; report raw alignment and each measurable reliability; identify whether its corrected target is consensus-to-consensus or consensus-to-instance; and separate layer selection from evaluation. It should state the aggregation rule and distinguish seed sensitivity from calibrated statistical uncertainty. When the motivation is model utility, the relationship to a held-out task should also be measured rather than assumed.

The present evidence has clear limits. Replicated models reach 410M in Pythia and 1B nominally in DataDecide, well below the largest models motivating PRH. Our capability-axis check does not hold size or training compute constant. The stimulus sets are small, the cached features are quantized, and alternative pooling and tokenization choices remain open. Infinite-width theory can motivate increased kernel concentration under particular assumptions (Lee et al., 2018; Jacot et al., 2018; Yang & Hu, 2021), but nearest-neighbor stability additionally requires separation of competing neighbors; determinism is not a universal consequence of increasing scale. A high consensus score also need not imply shared semantics or equivalent task performance.

The scale and training studies suggest a practical diagnostic. When corrected alignment is stable, the reliability-only curve predicts the observed growth without requiring closer consensus directions. When both factors grow, a convergence interpretation survives attenuation correction. When corrected alignment falls despite rising raw alignment, greater consistency conceals movement away from the reference consensus. The experiments exhibit all three possibilities across their metrics and paths.

The next test is to match recipes by held-out capability and evaluate transfer to new stimulus distributions. Denser checkpoints can locate reversals; a verified initialization-by-data factorial design can identify their sources. The framework supplies measurable distinctions between convergence and consistency, with explicit controls on their interpretation.

AI USE STATEMENT

Generative AI assisted with conceptual development, math derivations, experimental design and implementation, data processing and annotation, and manuscript editing.

REPRODUCIBILITY STATEMENT

The supplement includes extraction and analysis code, cached alignment tensors, feature hashes, seed partitions, per-seed retrieval outcomes, known-truth simulations, figure sources, and tests of the identities and selection boundaries. Generated tables and new numerical macros read saved results directly. The dated revision amendment records changes made after the earlier local protocol; the original manuscript, code, and outputs are preserved in a deprecated snapshot.

REFERENCES

Kiril Bangachev, Guy Bresler, and Yury Polyanskiy. Representation alignment rests on linear structure. *arXiv preprint arXiv:2605.28870*, 2026.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 2023.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.

Lee J. Cronbach, Goldine C. Gleser, Harinder Nanda, and Nageswari Rajaratnam. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. Wiley, 1972. ISBN 9780471188506.

MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. Reliability of CKA as a similarity measure in deep learning. In *International Conference on Learning Representations*, 2023.

Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity through statistical testing. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1556–1568, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Damian Falk, Léo Meynent, Florence Pfammatter, Konstantin Schürholt, and Damian Borth. A model zoo of vision transformers. In *ICLR 2025 Workshop on Neural Network Weights as a New Data Modality*, 2025. arXiv:2504.10231.

Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the Platonic representation hypothesis: An Aristotelian view. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. arXiv:2602.14486.

Sarah E. Harvey, David Lipshutz, and Alex H. Williams. What representational similarity measures imply about decodable information. In *Proceedings of UniReps: the Second Edition of the Workshop on Unifying Representations in Neural Models*, volume 285 of *Proceedings of Machine Learning Research*, pp. 140–151. PMLR, 2024.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The Platonic Representation Hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642. PMLR, 2024.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Siddharth Karamcheti, Laurel Orr, Jason Bolton, Tianyi Zhang, Karan Goel, Avanika Narayan, Rishi Bommasani, Deepak Narayanan, Tatsunori Hashimoto, Dan Jurafsky, Christopher D. Manning, Christopher Potts, Christopher Ré, and Percy Liang. Mistral – a journey towards reproducible language model training. `https://github.com/stanford-crfm/mistral`, 2021. URL `https://crfm.stanford.edu/2021/08/26/mistral.html`.

Max Klabunde, Tassilo Wald, Tobias Schumacher, Klaus Maier-Hein, Markus Strohmaier, and Florian Lemmerich. ReSi: A comprehensive benchmark for representational similarity measures. In *International Conference on Learning Representations*, 2025.

A. Sophia Koepke, Daniil Zverev, Shiry Ginosar, and Alexei A. Efros. Back into Plato's cave: Examining cross-modal representational convergence at scale. *arXiv preprint arXiv:2604.18572*, 2026.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008.

Gilad Landau and Aviv Keren. The concept of representation in ML: Beyond Plato and Aristotle. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. arXiv:2607.17800.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as Gaussian processes. In *International Conference on Learning Representations*, 2018.

Zeyu Michael Li, Hung Anh Vu, Damilola Awofisayo, and Emily Wenger. Causes and consequences of representational similarity in machine learning models. *arXiv preprint arXiv:2505.13899*, 2025.

Baihan Lin and Nikolaus Kriegeskorte. The topology and geometry of neural representations. *Proceedings of the National Academy of Sciences*, 121(42):e2317881121, 2024. doi: 10.1073/pnas.2317881121.

Ian Magnusson, Nguyen Tai, Ben Bogin, David Heineman, Jena D. Hwang, Luca Soldaini, Akshita Bhagia, Jiacheng Liu, Dirk Groeneveld, Oyvind Tafjord, Noah A. Smith, Pang Wei Koh, and Jesse Dodge. DataDecide: How to predict best pretraining data with small experiments. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 42487–42502. PMLR, 2025. URL `https://proceedings.mlr.press/v267/magnusson25a.html`.

Johannes Mehrer, Courtney J. Spoerer, Nikolaus Kriegeskorte, and Tim C. Kietzmann. Individual differences among deep neural network models. *Nature Communications*, 11(1):5725, 2020. doi: 10.1038/s41467-020-19632-w.

Ari S. Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Alex Murphy, Joel Zylberberg, and Alona Fyshe. Correcting biased centered kernel alignment measures in biological and artificial neural networks. *arXiv preprint arXiv:2405.01012*, 2024.

Thao Nguyen, Maithra Raghu, and Simon Kornblith. Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. In *International Conference on Learning Representations*, 2021.

Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. A toolbox for representational similarity analysis. *PLOS Computational Biology*, 10 (4):e1003553, 2014. doi: 10.1371/journal.pcbi.1003553.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Yu (Eric) Qian, Wilson S. Geisler, and Xue-Xin Wei. Quantifying task-relevant representational similarity using decision variable correlation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 149725–149758, 2025. doi: 10.52202/085713-5008.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Heiko H. Schütt, Alexander D. Kipnis, Jörn Diedrichsen, and Nikolaus Kriegeskorte. Statistical inference on representational geometries. *eLife*, 12:e82566, 2023. doi: 10.7554/eLife.82566.

Thibault Sellam, Steve Yadlowsky, Ian Tenney, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Turc, Jacob Eisenstein, Dipanjan Das, and Ellie Pavlick. The MultiBERTs: BERT reproductions for robustness analysis. In *International Conference on Learning Representations*, 2022.

shoyammishra. pythia-160m-weight-seed{1,2} step¿=1 checkpoints appear inconsistent with their own step0 initialization. EleutherAI/pythia, GitHub issue #203, 2026. URL https://github.com/EleutherAI/pythia/issues/203. Unresolved user report; accessed 24 September 2026.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information, 2017. URL https://arxiv.org/abs/1703.00810.

Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. doi: 10.2307/1412159.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. WIT: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2443–2449, 2021. doi: 10.1145/3404835.3463257.

Ali Hussaini Umar and Alessandro Laio. Signal-to-noise ratio and sample size govern representational alignment in neural networks. *arXiv preprint arXiv:2605.26973*, 2026.

Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, and Adriana Romero-Soriano. A picture is worth more than 77 text tokens: Evaluating CLIP-style models on dense captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26700–26709, 2024.

Muhammad Usama and Dong Eui Chang. Convergence without understanding: When language models agree on representations but disagree on reasoning. *arXiv preprint arXiv:2605.23315*, 2026.

Oskar van der Wal, Pietro Lesci, Max Müller-Eberstein, Naomi Saphra, Hailey Schoelkopf, Willem Zuidema, and Stella Biderman. PolyPythias: Stability and outliers across fifty language model pre-training runs. In *International Conference on Learning Representations*, 2025.

Greg Yang and Edward J. Hu. Tensor programs IV: Feature learning in infinite-width neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11727–11737. PMLR, 2021.

Junwon You, Mihyun Jang, Sangwoo Mo, and Jae-Hun Jung. What converges in the Platonic representation hypothesis? structure over geometry. *arXiv preprint arXiv:2609.27252*, 2026.

Liu Ziyin and Isaac Chuang. Proof of a perfect Platonic representation hypothesis. *arXiv preprint arXiv:2507.01098*, 2025.

A Proofs and scope of the representation identities

Expected ceilings are not pointwise ceilings. Let U, V be orthogonal unit centered rank-one Gram matrices. Recipe A always produces U, and recipe B produces U with probability p and V otherwise. Then $\rho_A = 1$, $\rho_B = p^2 + (1-p)^2$, and $C_{AB} = p/\sqrt{p^2 + (1-p)^2}$. Conditional on observing $g = U$ from recipe B, the single-pair score and the one-sided corrected score are both one. At $p = 0.1$, the population consensus is about 0.110. This provides both a counterexample to a pointwise reliability ceiling and a counterexample to interpreting a fixed-reference correction as a lower bound on an unobserved recipe consensus. Averaging over fresh independent draws of g restores Equation (2).

Distance and ensemble identities. Writing $R_A = \mu_A + \epsilon_A$, the centered residual has expectation zero and $\mathbb{E}\|\epsilon_A\|^2 = 1 - \rho_A$. Expanding $\|R_A - R_B\|^2$ and taking expectations removes the cross terms under independence, yielding Equation (5). For an average of m independent unit objects,

$$\mathbb{E}\left\|\frac{1}{m}\sum_{s=1}^{m} R_s\right\|^2 = \frac{m + m(m-1)\rho}{m^2} = \rho + \frac{1-\rho}{m}. \tag{8}$$

This identity also gives the unbiased reliability estimator $(\|\sum_s R_s\|^2 - S)/(S(S-1))$. Neither this identity nor a ratio of expected ensemble norms gives the expected cosine of two finite ensembles; an expectation cannot generally be moved through that ratio.

Permutation floors. For fixed matrices M, N and a uniform stimulus permutation, diagonal terms in $\langle M, P_\pi N P_\pi^\top \rangle_F$ have expectation $\operatorname{tr}(M)\operatorname{tr}(N)/n$. The remaining terms have expectation

$$\frac{\left(\sum_{i\neq j} M_{ij}\right)\left(\sum_{k\neq l} N_{kl}\right)}{n(n-1)}. \tag{9}$$

Centered kernels have zero row sums, so each off-diagonal sum equals minus its trace. The combined expectation is $\operatorname{tr}(M)\operatorname{tr}(N)/(n-1)$, giving the CKA expression after normalization. A k-NN adjacency matrix has zero trace and off-diagonal sum nk, giving $k/(n-1)$ after normalization. For CKA, $\operatorname{tr}(K)^2/\|K\|_F^2$ is the participation ratio of the centered spectrum, explaining the role of effective rank in the floor.

A.1 Why floor and ceiling calibration compose

For CKA, let $u = H/\sqrt{n-1}$ and $\alpha_R = \langle R, u\rangle = \operatorname{tr}(K)/(\|K\|_F\sqrt{n-1})$. The permutation floor is $\alpha_R\alpha_S$. Set $v_R = R - \alpha_R u$, so that $\|v_R\|^2 = 1 - \alpha_R^2$. For $|\alpha_R| < 1$, define the Hilbert-space lift

$$\Phi(R) = \big(v_R, \alpha_R v_R, \alpha_R^2 v_R, \dots\big). \tag{10}$$

The convergent geometric series gives

$$\|\Phi(R)\|^2 = \frac{1-\alpha_R^2}{1-\alpha_R^2} = 1, \tag{11}$$

$$\langle \Phi(R), \Phi(S)\rangle = \frac{\langle R, S\rangle - \alpha_R\alpha_S}{1 - \alpha_R\alpha_S} = \tilde{a}(R, S). \tag{12}$$

The attenuation identity therefore applies to the means of $\Phi(R)$ when calibrated similarities are averaged over seeds. It does *not* identify the original centered-Gram consensus. At $\alpha = 1$, the

normalized kernel is isotropic on the centered subspace and contains no residual direction; this case is excluded rather than regularized silently.

For k-NN, take $u = (\mathbf{1}\mathbf{1}^\top - I)/\sqrt{n(n-1)}$. Then $\alpha = \sqrt{k/(n-1)}$ is constant across all graphs at the same n, k. Calibration reduces to the inner product between normalized residual graphs $(R - \alpha u)/\sqrt{1-\alpha^2}$. In contrast, for varying CKA floors, normalized residualization gives

$$\frac{\langle R, S\rangle - \alpha_R \alpha_S}{\sqrt{(1-\alpha_R^2)(1-\alpha_S^2)}}, \tag{13}$$

which is a different unit kernel. Both permit attenuation analysis, but they define different consensus objects. Calibrated scores and their population consensuses can be negative; a log decomposition is used only where its factors are positive.

B Estimation details and safeguards

Exact aggregation. An evaluation record contains $a_j, r_{A,j}, r_{B,j}$, with $r_{B,j} = 1$ for a fixed reference, and $c_j = a_j/\sqrt{r_{A,j} r_{B,j}}$. The revision averages the vector $(\log a_j, \frac{1}{2}\log r_{A,j}, \frac{1}{2}\log r_{B,j}, \log c_j)$ with equal weights over evaluation directions, references, and partitions. A common least-squares operator on log nominal size preserves the sum exactly. Averaging all four scores arithmetically before taking logs does not. The revised Pythia sum residual is below 10^{-12}; the original arithmetic-fold decomposition had a residual near -0.00167 per decade.

Cross-fitting makes a selected readout independent of its evaluation seeds conditional on its selection set. It does not make a ratio unbiased, remove uncertainty due to only a few seeds, or identify a single fixed layer across scales. With three seeds, choosing a layer on one and evaluating the other two is valid for independent selection but supplies only one within-recipe pair per evaluation fold. The three selector folds reuse the same runs and are not independent observations.

Ratio bias. For the one-sided estimator $\hat{q} = \hat{a}\hat{r}^{-1/2}$ with positive population a, r, a second-order expansion gives

$$\frac{\mathbb{E}\hat{q} - q}{q} \approx \frac{3}{8}\frac{\mathrm{Var}(\hat{r})}{r^2} - \frac{1}{2}\frac{\mathrm{Cov}(\hat{a}, \hat{r})}{ar}. \tag{14}$$

For two corrected sides, sum the two variance and numerator–reliability covariance terms and add $\mathrm{Cov}(\hat{r}_A, \hat{r}_B)/(4 r_A r_B)$. These expansions require concentration away from zero; they are not uniform guarantees near vanishing reliability. The original assertion that ratio bias was necessarily below 0.2% omitted the covariance term and is not used in the revised interpretation.

Bootstrap overlap and seed coupling. In a diagnostic of 100,000 nine-seed bootstrap draws under the earlier acceptance rule, approximately 89.2% of accepted draws put at least one original seed in both selection and evaluation halves. Excluding diagonal pairs from the reliability denominator does not remove that selection overlap. This observation identifies a violation of the intended independence; it alone does not establish the direction or magnitude of coverage error. The revision avoids that bootstrap for its empirical uncertainty display.

Across Pythia sizes, equal seed labels can share document-order randomness. For fixed-layer within-Pythia lineage comparisons we remove matched-label cross-size pairs. At final layers, matched-minus-unmatched raw overlap is approximately -0.00024 for 31M/70M, -0.00059 for 70M/160M, and 0.00613 for 160M/410M. These checks do not establish a large bias in every comparison; they explain why independent-pair bookkeeping matters. All seed partitions are shared across sizes, and all vision partitions are shared across sizes in the two-sided experiment.

Validity checks. The revised estimator rejects nonfinite input maps, overlapping selection and evaluation identities, fewer than two distinct evaluation seeds for a corrected side, and nonpositive inputs to the log decomposition. It records chosen layers and original seed identities. It does not clamp corrected estimates at one. Tests check the population identities, the floor-kernel lift, the three-seed boundary, and the independence of layer selection from evaluation. No assertion of general interval coverage follows from those unit tests.

Table 3: **Metric and protocol sensitivity.** Log slopes per decade of nominal Pythia parameters, using identical log aggregation. "Reliability" is the $\frac{1}{2}\log\rho_A$ slope. Each mutual-neighbor variant retains the seven reference checkpoints. Full seed-partition ranges appear in Appendix C.

Specification	Raw slope	Reliability	Corrected	Reliability share
WIT, 10-NN	0.423	0.217	0.206	51.3%
WIT, 5-NN	0.438	0.237	0.201	54.2%
WIT, 20-NN	0.381	0.190	0.190	50.0%
No clamping	0.308	0.206	0.102	66.8%
DCI, 10-NN	0.273	0.119	0.154	43.7%
WIT-4096, 10-NN	0.504	0.190	0.313	37.8%
Floor-calibrated 10-NN	0.510	0.224	0.287	43.8%
CKA	0.357	0.067	0.291	18.6%
Floor-calibrated CKA	0.444	0.058	0.386	13.1%

C Full scaling and robustness results

Table 4: Main Pythia estimates. Values are geometric averages over evaluation directions, seven references, and 126 partitions. The selected reference layer can vary across language sizes; the fixed-layer control addresses that variation.

Parameters	Raw alignment	Reliability ρ_A	Corrected q_g
14M	0.0425	0.238	0.0871
31M	0.0496	0.305	0.0898
70M	0.0554	0.403	0.0873
160M	0.0692	0.462	0.1018
410M	0.0778	0.432	0.1184

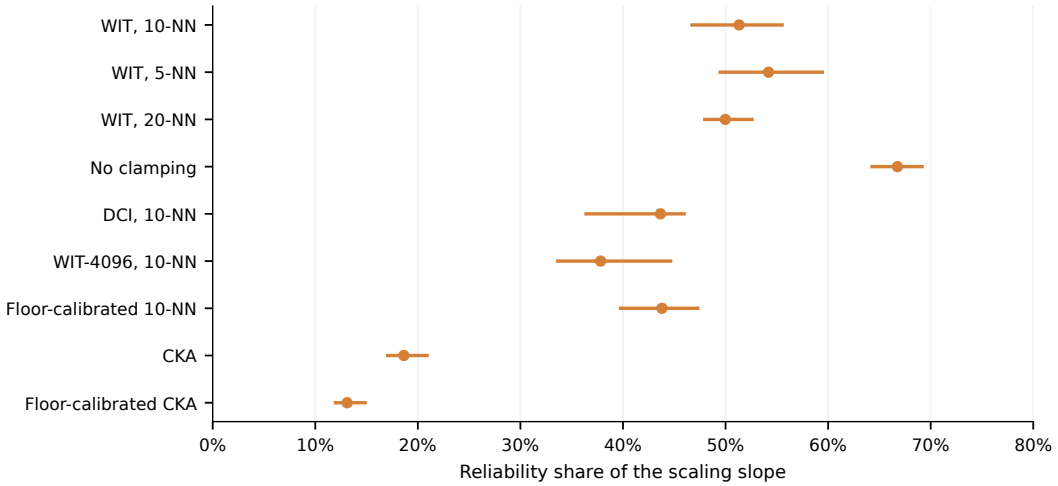

Figure 5: **Specification sensitivity of the reliability contribution.** Dots are ratios of the average log-slope components. Lines span all 126 seed partitions within each specification, rather than sampling confidence intervals. The choice of representational metric changes the estimated share more than several within-metric protocol changes.

The original arithmetic-fold Pythia estimate assigned 50.29% of the slope to reliability. Retaining its original 4/5 split but averaging fold logs gives 49.96%; averaging over all partitions gives 51.3%. Thus the main magnitude survives the aggregation correction, while a binary decision at exactly 50% is unstable. In the original arithmetic-fold sensitivity analysis, only 9 of 126 split schedules had a corrected/raw fitted-slope ratio at least 0.5. This is a count of point-estimate criteria, not a distribution of independent statistical tests.

The retained historical robustness shares were approximately 52.9% for 5-NN, 49.6% for 20-NN, 66.7% without clamping, 44.6% on DCI, 35.3% on WIT-4096, and 42.9% with the 10-NN floor removed. Table 3 reports their revised counterparts. All specifications use the same seven available vision references. WIT-1024 contains 1023 unique exact-caption strings and WIT-4096 contains 4090, with 1022 shared unique strings; the two sets are neither disjoint nor exactly identical in caption multiplicity. We therefore describe their comparison as a stimulus-set and neighborhood-resolution sensitivity check.

C.1 Depth profiles

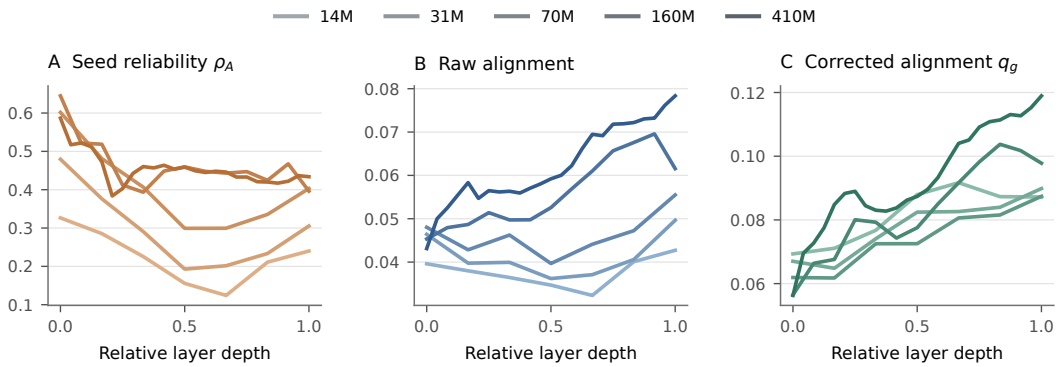

Figure 6: **Descriptive depth profiles.** At each language layer, the mean raw alignment is maximized over vision layers using all seeds and then geometrically averaged over the seven references. Panels show seed reliability, raw alignment, and reliability-corrected alignment. Shade deepens with size; orange, blue, and green denote reliability, raw alignment, and corrected alignment. These profiles retain the earlier exploratory analysis; their vision-layer maxima are in-sample and are not used for the cross-fitted headline estimates.

The smallest models have more similar corrected profiles than raw profiles, consistent with their small corrected scale slope. This is a description of the displayed readouts, not a test that layerwise consensuses are identical. Relative depth also does not guarantee functional correspondence between layers of different architectures.

D DataDecide and two-sided replication

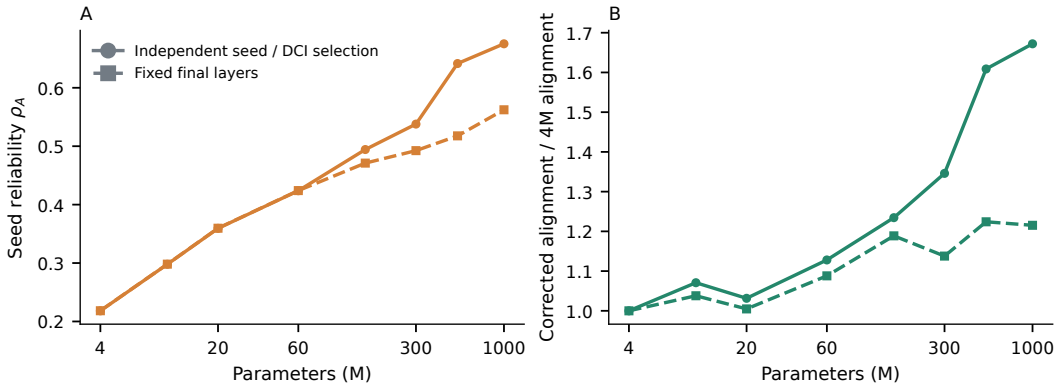

Figure 7: **DataDecide separates replication of an effect from replication of an exact share.** Three seeds at each of eight nominal scales. The selected-layer analysis trains its selector on one DCI seed and evaluates on the other two WIT seeds. Fixed final layers require no selector. Both analyses find a substantial reliability contribution, with different corrected trajectories.

The historical DataDecide analysis selected layers on all three DCI seeds and reused those seed identities on WIT. Its reliability share was 50.4%, with a reported 49–52% bootstrap interval. Dataset separation alone does not make training-seed selection independent, and that interval is not retained as calibrated uncertainty. The revised independent selector gives 50.7%, while the fixed final-layer share is 66.4%. The common within-size checkpoints and all three revision labels are recorded in the released configuration. Actual parameter counts differ from nominal labels, and the training-token budget varies with scale.

The historical replicated-vision analysis gave a raw slope of about 0.391, a language-reliability term of 0.215, and a corrected term of 0.169. It omitted the approximately 0.00758 vision-reliability slope that arose because the selected vision layer changed. The revision retains both terms and averages over 1260 joint partitions; every underlying cross-alignment and within-family tensor is saved. These many partitions do not increase the number of independent trained vision models beyond five.

D.1 Capability-axis extension

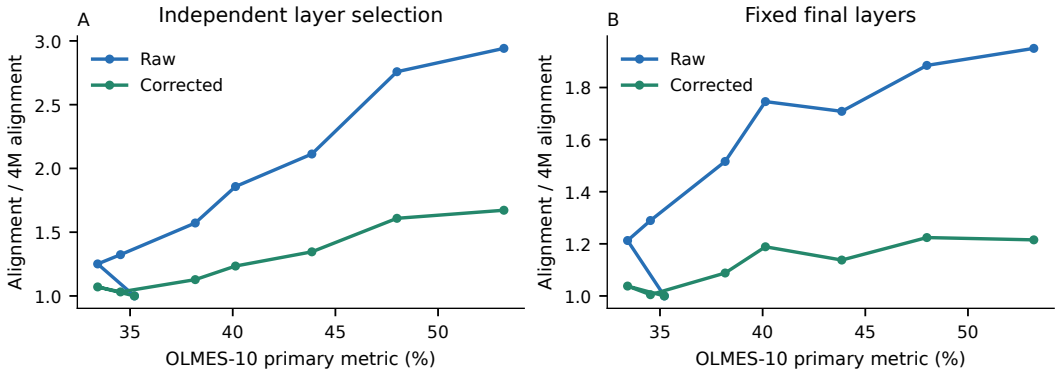

Figure 8: **Reliability remains relevant on a capability axis.** DataDecide alignment versus the published OLMES-10 primary metric, averaged over the exact three seed/checkpoint records at each scale. Lines connect increasing nominal model sizes; capability is not strictly monotone at the smallest scales. Raw and corrected values are indexed to 4M, with independent layer selection in (a) and fixed final layers in (b).

We use the published `olmes_10_macro_avg` row and its `primary_metric` field from DataDecide's evaluation release. Every analyzed scale has exactly one matching row for each of its three published seed labels at the exact common step used for feature extraction. The eight mean scores span 33.4–53.2 percent. The saved extension includes all matching task rows, seed/step identities, token counts, compute values, and a hash of the source parquet file. No nearest-checkpoint substitution or choice among task aggregates is used.

The fitted slopes are in natural-log alignment per ten percentage points of mean capability. Because regression on any common scalar predictor is linear, the four-term decomposition still holds algebraically. Reliability contributes 46.8% for the independent selector and 63.7% for fixed final layers. The capability-axis analysis was added after the parameter-axis results were inspected, and has no claim to prospective confirmation. It connects the measurement issue to actual benchmark capability but does not identify a causal effect of capability, size, or compute.

E Training-checkpoint study

The checkpoint study was specified after the original scale results, in response to the question of whether apparent convergence also changes during training. It is exploratory. Pythia 14M and 70M use seeds 1–9 at steps 0, $1{,}000$, $16{,}000$, $64{,}000$, and $143{,}000$. DataDecide Dolma 1.7 60M uses the default, small-aux-2, and small-aux-3 seeds at steps 0, $1{,}250$, $7{,}500$, $15{,}000$, and $29{,}042$. All 105 prescribed checkpoints are evaluated on the same WIT-1024 captions, including freshly extracted final checkpoints. No new pretraining is performed.

Each repository revision is resolved to an immutable commit before downloading. We record the commit, loaded parameter count, input-text hash, extraction time, readout, and numerical precision. Computation and feature storage both use float32. Captions are right-padded and truncated to 256 tokens; hidden-state readouts average only nonpadding tokens. The primary analysis applies the same outlier clamp and mutual-10-NN construction as the scale study. Seven vision checkpoints and the five-member ViT family are held at their fixed final layers. Existing vision features remain stored in half precision.

For each recipe and checkpoint, reliability averages distinct language-seed pairs. Cross-modal raw alignment averages all language–vision seed pairs; the two-sided correction divides by the square root of both reliabilities. No checkpoint or layer is chosen to maximize alignment. The lines interpolate five observed checkpoints and do not locate an exact turning time. Normalized linear CKA is a prespecified metric control. An unclamped mutual-10-NN control was added after inspecting preliminary Pythia paths; it retains every prescribed checkpoint and reference.

Table 5: **All fixed-layer training observations.** Raw and C compare against the five-seed ViT family; ρ_A is language-seed reliability. NLL is token-weighted caption next-token negative log likelihood, then averaged across seeds. PR is the mean participation ratio of centered, row-normalized language features. Tokenization differs across families, so absolute NLL is compared only within a family.

Recipe	Step	ρ_A	Raw	C	NLL	PR
Pythia 14M	0	0.168	0.0210	0.0739	11.022	30.7
	1,000	0.193	0.0276	0.0905	6.878	23.1
	16,000	0.255	0.0387	0.1105	6.055	35.1
	64,000	0.222	0.0348	0.1066	6.012	27.6
	143,000	0.240	0.0338	0.0992	6.059	26.6
OLMo 60M	0	0.315	0.0226	0.0580	10.902	54.3
	1,250	0.326	0.0380	0.0959	5.976	22.2
	7,500	0.385	0.0416	0.0967	5.525	18.7
	15,000	0.426	0.0444	0.0982	5.348	18.5
	29,042	0.424	0.0505	0.1118	5.078	24.1
Pythia 70M	0	0.370	0.0220	0.0521	11.030	83.5
	1,000	0.371	0.0412	0.0976	6.282	55.5
	16,000	0.463	0.0525	0.1113	5.349	80.9
	64,000	0.394	0.0473	0.1087	5.297	85.3
	143,000	0.403	0.0425	0.0964	5.300	54.6

Table 6: **Does final training retain peak consensus?** Final two-sided consensus as a percentage of its largest value among the five sampled checkpoints. Each metric uses its own sampled maximum. A value below 100% indicates a descriptive retreat from that maximum, not a statistically estimated phase boundary.

Recipe	10-NN	No clamping	Normalized CKA
Pythia 14M	89.8%	96.0%	87.3%
OLMo 60M	100.0%	100.0%	100.0%
Pythia 70M	86.7%	93.9%	76.6%

Caption likelihood is computed from causal next-token logits, summing negative log probabilities over nonpadding target tokens and dividing by their count. It uses the same captions as the representation analysis and is not an independent benchmark or proof of semantic improvement. For centered, row-normalized feature Gram matrix K, the participation ratio is $\mathrm{tr}(K)^2/\|K\|_F^2$. This describes spectral concentration. The two-coordinate presentation is inspired by information-plane trajectories (Shwartz-Ziv & Tishby, 2017), but neither coordinate estimates mutual information, and these measurements do not establish an information bottleneck.

We also save all per-layer profiles without selection and cross-checkpoint consensus matrices. Cross-checkpoint comparisons omit pairs with the same seed identity because their shared training history couples the two draws. For numerical controls, each fresh float32 readout is quantized to half precision and its neighbor graph compared with the float32 graph. Fresh final readouts are additionally

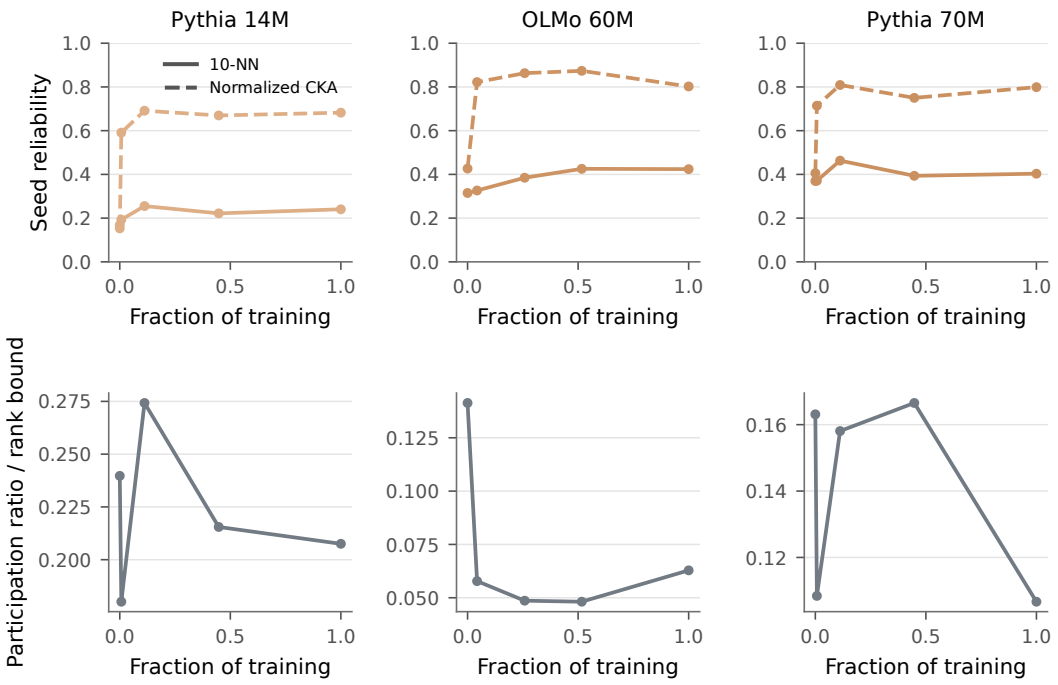

Figure 9: **Reliability and spectral diagnostics across training.** Top: mutual-10-NN and normalized-CKA seed reliability. Bottom: the mean participation ratio of centered, row-normalized language features, divided by its algebraic maximum $\min(d, n-1)$. Time is scaled within each recipe. A change in participation ratio measures spectral concentration, not mutual information or an information-bottleneck objective.

compared with the original cached final readouts. Across all cells, language storage quantization changes two-sided mutual-neighbor consensus by at most 0.180%; substituting the legacy final caches changes it by at most 0.261%. Individual graph overlaps can be lower: their minimum is about 98.0%, illustrating that small feature changes can alter near-tied neighbors while changing averaged alignment much less. These controls distinguish storage quantization from a changed extraction or checkpoint; source hashes and per-seed outcomes remain in the result artifact.

F Lineage, deduplication, and sources of seed variation

Historical T2 and T3 results. The earlier selected-layer analysis reported adjacent-size Pythia consensus values of 0.94–0.95 for 31M/70M, 70M/160M, and 160M/410M; GPT-2 small/medium gave 0.89. Pythia against GPT-2 small rose from 0.53 to 0.68 over 14M–410M. Comparable-size Pythia/OLMo pairs gave 0.73–0.86. Those OLMo comparisons used in-sample selection on the three-seed side. Figure 10 supplies a selection-free replacement for population interpretation while preserving the historical observations. A selected layer's high agreement should not be generalized to the complete network.

The original local prediction P4 had two parts. Its descriptive lineage-signature criterion, a corrected value below 0.95 at the largest Pythia scale, was met. Its predicted relationship between raw and corrected cross-lineage changes was not: the GPT-2 medium recipe was less reliable than the small recipe. The original modality comparison likewise found much smaller cross-modal scores. One-sided fixed-vision scores, however, cannot identify an unmeasured vision recipe's consensus and do not confirm a population modality-gap bound. The replicated ViT comparison provides the corresponding two-sided evidence within one supervised vision recipe.

Deduplicated Pile controls. The original deduplicated checkpoints provide only one model per size. Dividing their cross-alignment with standard Pythia by the standard recipe's reliability gave proxy

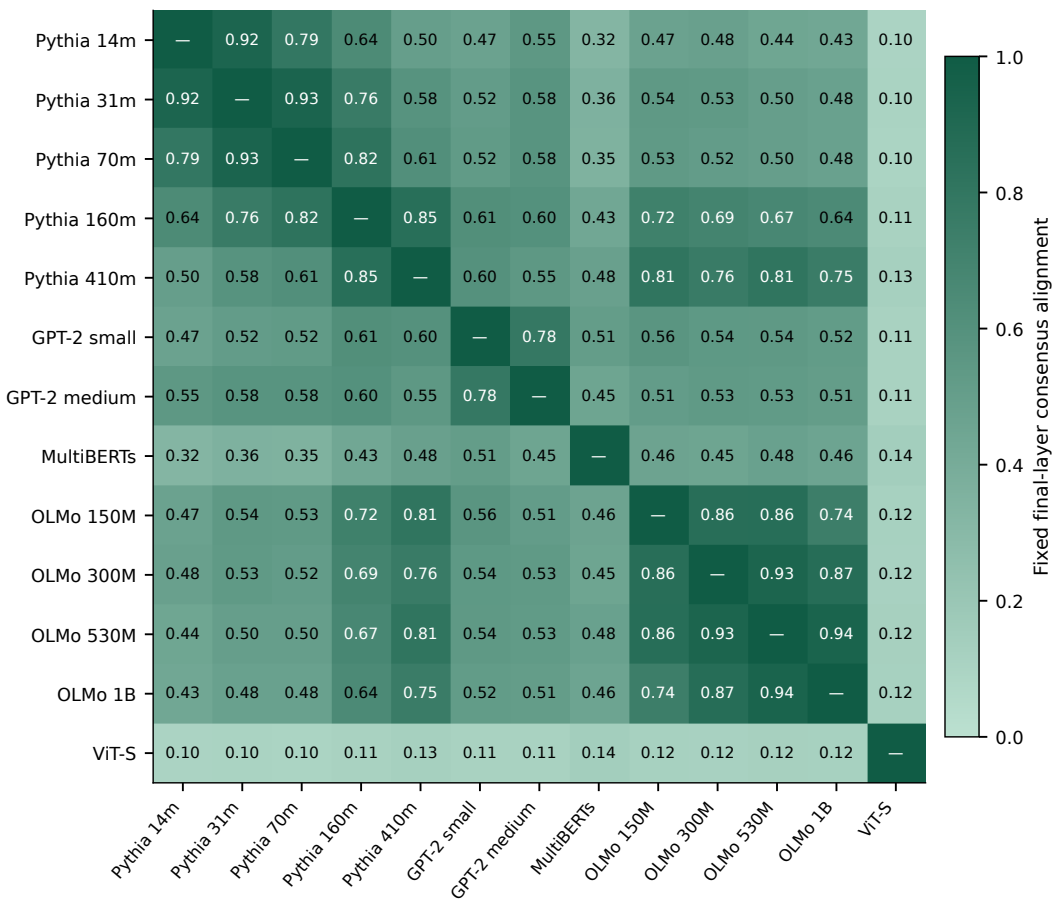

Figure 10: **Fixed-readout consensus comparisons.** Every off-diagonal entry uses final layers and seed replicates on both sides. Equal published seed labels within a lineage are excluded across sizes. The diagonal is a reference value and is not an estimate from comparing a seed with itself. Several readout choices can yield different magnitudes; the matrix is a descriptive comparison of complete recipes, not a causal ordering of corpus, tokenizer, objective, and modality effects.

values 1.01–1.06, slightly above one; correcting only the replicated standard side gave 0.78–0.81. The latter values are conditional consensus-to-instance scores, not lower bounds. The proxy assumes an unmeasured deduplicated reliability equal to the standard reliability, so values above one diagnose the inadequacy or uncertainty of that assumption rather than equivalence of the two recipe consensuses. Replicated deduplicated checkpoints are needed for that conclusion.

Initialization and data-sequence contrasts. The original 160M analysis retained a shared similarity component of 0.48, a data-order contrast of 0.02, an initialization contrast of 0.02, and a residual of 0.48, using six pairwise comparisons for each single-factor contrast. Those pairs reuse four runs and are not six independent replications. Under a genuinely crossed random-effects model $R_{io} = \mu + a_i + b_o + e_{io}$, independent zero-mean components imply

$$\mathbb{E}\langle R_{io}, R_{io'}\rangle = \|\mu\|^2 + \sigma_a^2, \tag{15}$$

$$\mathbb{E}\langle R_{io}, R_{i'o}\rangle = \|\mu\|^2 + \sigma_b^2, \tag{16}$$

$$\mathbb{E}\langle R_{io}, R_{i'o'}\rangle = \|\mu\|^2, \tag{17}$$

when the differing indices are independent. Unit norm then gives $\|\mu\|^2 + \sigma_a^2 + \sigma_b^2 + \sigma_e^2 = 1$. This is the relevant generalizability-theory interpretation (Cronbach et al., 1972), conditional on a valid design and the random-effects assumptions.

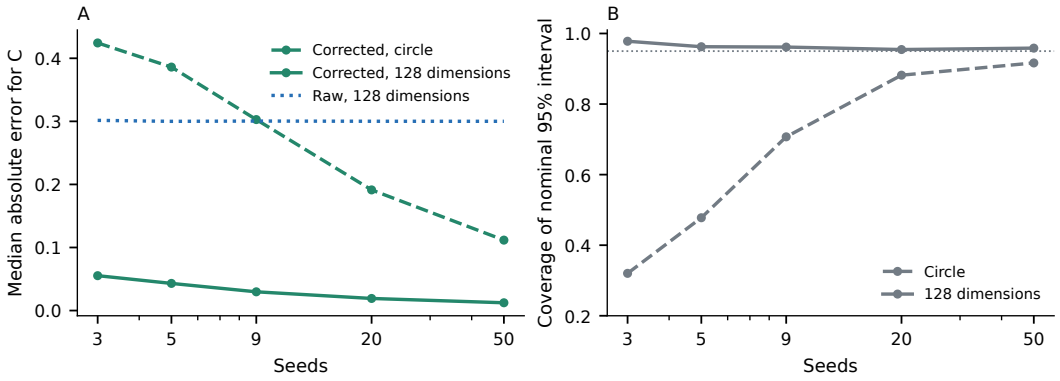

Figure 11: **Population correctness does not guarantee few-seed inference.** Two known-truth unit-vector populations have $\rho_A = 0.25$ and $C = 0.6$. Their noise is either angular on a circle or isotropic in a 127-dimensional orthogonal subspace. (A) Error of the corrected point estimate; undefined estimates are excluded and their frequencies are separately reported in the supplement. (B) Coverage of a nominal 95% fixed-readout jackknife-t interval, counting undefined intervals as uncovered. There are 10,000 repetitions per circle condition and 2,000 per high-dimensional condition. These are method checks, not additional pretrained models.

The checkpoint controls require caution. An unresolved upstream report questions the initialization identities of some PolyPythia 160M weight/data-seed variants (shoyammishra, 2026); we have not independently established the report's claim from all original weights. Moreover, document shuffling before packing can change batch composition as well as order. We retain these contrasts as exploratory measurements with unverified intervention provenance. They do not establish a causal variance decomposition or dynamical chaos. A complete crossed design, multiple initialization and sequence levels, preserved initial weights, and direct batch/step hashes would be required to make that stronger claim.

G KNOWN-TRUTH VALIDATION

For the angular population, $R = (\cos\theta, \sin\theta)$ with $\theta \sim \mathcal{N}(\arccos C, -\log\rho)$. Therefore $\mathbb{E}R = \sqrt{\rho}(C, \sqrt{1-C^2})$. Against fixed reference $g = (1, 0)$, raw population alignment is $C\sqrt{\rho}$ and the consensus-to-instance target is exactly C. We set $C = 0.6$, use $\rho \in \{0.25, 0.5, 0.8\}$, and evaluate $S \in \{3, 5, 9, 20, 50\}$ with 10,000 repetitions per condition.

The concentrated control draws $R = \sqrt{\rho}\, m + \sqrt{1-\rho}\, U$, where m is a unit vector and U is uniform on the unit sphere in the 127-dimensional subspace orthogonal to m. A reference g satisfies $m^\top g = C$. This construction has exactly the same population reliability and consensus target. It uses 2,000 repetitions per condition. This control was added after inspecting the circle results to test the role of concentration and is explicitly exploratory.

For each repetition we compute the distinct-pair reliability estimator and its one-sided corrected ratio. A delete-one-seed jackknife estimates the standard error; the nominal interval uses the $t_{S-1,0.975}$ critical value. It is defined only when the full estimate and all deletions have positive reliability. Undefined intervals count as uncovered in unconditional coverage, while point-estimate errors condition on a defined point estimate. Output files separately report the fraction of valid point estimates, fraction of defined intervals, unconditional and conditional coverage, Monte Carlo standard errors, and estimates outside the unit interval.

At $\rho = 0.25$ and nine seeds, only about 79.5% of circle repetitions have a defined interval, while essentially all concentrated-control repetitions do. For the same circle population, increasing to 50 seeds raises coverage to about 91.6%, still below nominal. In the concentrated control, coverage across the tested conditions is near nominal, with Monte Carlo uncertainty of roughly half a percentage point. These findings validate the need to examine the seed distribution; they do not calibrate intervals

for layer-selected empirical kNN graphs, nor do they prove that a corrected point estimate is always more accurate than raw alignment at a small seed count.

H HELD-OUT RETRIEVAL PROTOCOL AND INTERPRETATION

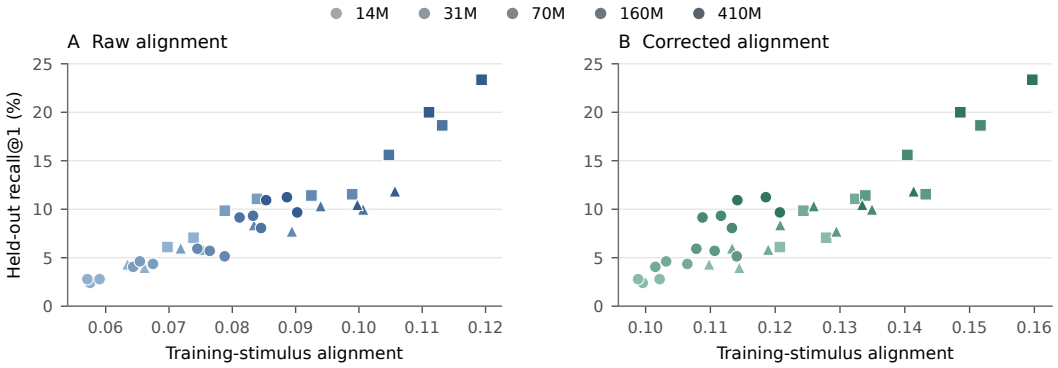

Figure 12: **Alignment and downstream utility are related but distinct.** Each point is a size–reference mean over nine language seeds. Raw (A) and corrected (B) alignment use the 768 training stimuli; recall@1 uses 255 held-out pairs. The horizontal axes have different scales because these panels show absolute scores, not relative growth. Blue denotes raw and green corrected alignment; shade deepens with size. Circles, squares, and triangles indicate DINO, CLIP, and supervised ViT families. Table 2 evaluates prediction across held-out references.

We retain the first occurrence of each exact caption string in WIT-1024, shuffle unique-caption indices with random seed 20260924, and allocate the first 768 to training. The remaining 255 form a held-out candidate set with one indexed image per caption. This removes exact-caption leakage but does not assert that all images or semantic concepts are distinct. No test labels are used to choose a layer, penalty, preprocessing parameter, or predictor.

Each final-layer feature set is centered by its training mean and normalized rowwise. This downstream protocol intentionally omits the whole-stimulus outlier clamp used in the main alignment analysis. For training caption features X, test captions Z, training image features Y, and penalty $\lambda = 1$, predictions are

$$\widehat{Y}_{\text{test}} = ZX^{\top}(XX^{\top} + \lambda I)^{-1}Y. \tag{18}$$

We normalize each predicted row and rank held-out image features by cosine similarity. Per-seed recall@1, recall@5, and the top-five candidate indices are saved for all $5 \times 9 \times 7 = 315$ model/reference fits. At each size, training-stimulus kNN reliability averages all 36 distinct language-seed pairs. Raw train alignment and retrieval recall are averaged over the nine seeds to produce each of the 35 family/reference cells.

Five ordinary least-squares predictors use an intercept and, respectively, no feature, log nominal parameters, log raw alignment, log corrected alignment, or both log raw alignment and log reliability. In each leave-one-reference-out fold, 30 cells fit the predictor and five evaluate it. The mean-only and parameter-count baselines were added as diagnostics while implementing the prespecified three alignment predictors. None uses regularization or clipping of its predicted recalls. All fold coefficients and predictions are retained so that a favorable aggregate cannot conceal a poor held-out reference.

This experiment concerns a simple learned linear transfer from caption representations to image representations. It is not zero-shot retrieval, does not establish that either representation encodes the same semantics, and is not an intervention on training reliability. The seven references are themselves related, and the same language seeds and test stimuli occur in many cells. Its role is to test a practical implication under a concrete protocol. Here, correction does not improve prediction over raw alignment; the distinction between consensus and utility is empirically consequential.

I REVISION RECORD AND REPRODUCIBILITY MAP

The supplement contains the unchanged local protocol, dated amendment, numerical results, and code for independent readouts, partition sensitivity, fixed-layer and CKA controls, replicated vision, simulation, capability matching, retrieval, and checkpoint dynamics. Tables and vector figures are generated from saved JSON/NPZ files; the README specifies execution order and external inputs. A checksummed local snapshot preserves the original manuscript and project materials, including the earlier NeurIPS draft.

Fresh float32 extraction covers the 105-checkpoint language study. Other suites retain cached precision; cache hashes establish input identity, not previously unrecorded checkpoint commits. Frontier-scale replication, matched-capability trajectories, verified initialization/order interventions, and alternate readouts remain future work. No revised outcome is inserted into the original protocol.