IS THERE A PLATONIC TREE? CALIBRATING LATENT HYPERBOLICITY IN FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Hyperbolic representation learning rests on a premise we call latent hyperbolicity: standard models are already tree-like because their class geometry scores a low Gromov δ, a measure of tree-likeness that is zero for a tree. That low value is also what a random cloud of the same dimension and spread scores. We introduce a calibrated framework that reads tree-likeness as its excess over such structureless clouds, together with a second test, whose power we measure, designed to distinguish hierarchical organization from simple clustering. We apply it to 12 vision backbones, 6 datasets and 15 text models. On image features, where the premise is read, the calibrated reading is indistinguishable from a random cloud in most model–dataset cells and small elsewhere. Class centroids do carry structure that a random cloud of the same shape does not, in 44 of 72 cells, but so does a star of clusters. The second test finds no detectable hierarchy among superclasses in the 9 of 12 backbones where it detects a planted one. What it finds in 4 of them is how each cluster is oriented, which disappears when clusters are rotated at random. MERU, trained in hyperbolic space, shows the same structure as its Euclidean twin in a space that stays nearly flat. In text, the result depends on the model's training recipe and size. Across models, the trees agree on which classes group together well above chance, though less than two resampled versions of the same model, and not on distances; the self-supervised models organize classes by direction rather than by distance, which a comparison by distance misses. Read correctly, foundation models organize classes into clusters that they partly share and show no detectable hierarchy among the superclasses where the test has power. Their raw tree-likeness is not evidence for hyperbolic geometry.

1 INTRODUCTION

Semantic knowledge is hierarchical, and hyperbolic space is the natural geometric home for hierarchies (Nickel & Kiela, 2017; Sarkar, 2011; Ganea et al., 2018; Chami et al., 2019). A line of work therefore imposes it on representations at training time, from vision to vision–language and language (Khrulkov et al., 2020; Ermolov et al., 2022; Atigh et al., 2022; Bdeir et al., 2024; Desai et al., 2023; Pal et al., 2025; Chen et al., 2022; He et al., 2025; Sinha et al., 2024). These methods cite a premise, which we call *latent hyperbolicity*: standard models already organize concepts in a tree-like metric structure. The evidence is a single number, the Gromov δ of their features, which is zero for a metric tree and grows as a metric departs from one. Measured on CNN and ViT features (Khrulkov et al., 2020; Bdeir et al., 2024; Ermolov et al., 2022), on token embeddings (Yang et al., 2025) and on the word co-occurrence graph (Tifrea et al., 2019), it comes out low. Those readings are taken at the sample level, on individual images or tokens rather than on classes, and they are read against the ideal values of a tree and of a non-hyperbolic space, never against what a structureless cloud of the same dimension and spectrum, its spread along each direction, would score. Is the premise right, and what does it license? Figure 1 frames the question: an observer who sees only a shadow, the raw δ, asks which of three worlds casts it, a structureless cloud, a star of clusters or a tree. All three read alike.

Why is the reading low? Gromov δ assigns zero to trees (Gromov, 1987), but its reference level depends on dimension, spectrum and statistic, so a raw value cannot be called low on its own. High-dimensional clouds concentrate toward equidistance (Beyer et al., 1999; Aggarwal et al., 2001), which is a degenerate star tree: we call this the *dimension confound*, the geometric counterpart of the

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.

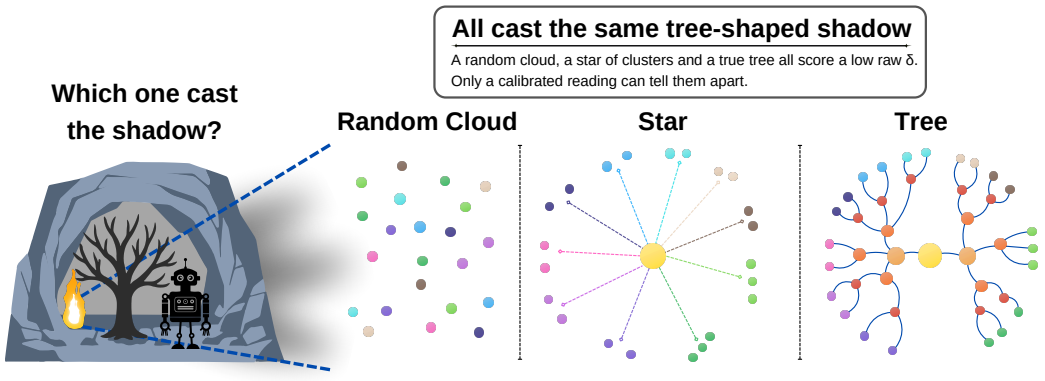

Figure 1: **Plato's cave: a random cloud, a star and a tree cast the same shadow.** An observer who sees only shadows cannot tell which object casts them, and a raw δ is such a shadow: points with no structure, clusters around a center, and a nested hierarchy all read as low δ. The instrument reads the objects instead of the shadow: the excess separates the random cloud from the other two, and the second test separates the star from the tree.

width confounder of Gröger et al. (2026). Anisotropic spectra lower the effective dimension and raise the reading: the *spectrum confound*. And computing the supremum exactly takes more than cubic time (Fournier et al., 2015), so it is taken over sampled quadruples, where it grows with the budget and does not converge: the *statistic confound*.

We build the instrument that reads δ against the right reference. It calibrates the estimator on reference geometries, expresses every reading as an *excess* over a random cloud of the same dimension and spectrum, ranks it against 200 such clouds, and adds a *second test* whose false alarms and ability to detect a planted hierarchy are measured on real clouds. Gröger et al. (2026) calibrate similarity across models; we calibrate geometry within a model, and comparing trees across models needs a calibration of its own. With it we ask of geometry what convergence work asks of similarity (Huh et al., 2024; Gröger et al., 2026): what survives calibration, and is it shared?

The answer has three parts. (i) *Where the premise is read,* on image features, the calibrated reading is indistinguishable from a random cloud of the same shape in most model–dataset cells and small elsewhere. The 4 values reported by Khrulkov et al. (2020) are reproduced and, calibrated, only MiniImageNet stays below the random cloud. CIFAR-10 and CUB-200 are indistinguishable from a random cloud, and CIFAR-100 falls below it only before correcting for testing four datasets at once. (ii) *Within each model,* class centroids carry structure that a random cloud does not have, in 44 of 72 cells and 30 of 36 on the 3 datasets with many classes (ImageNet, CIFAR-100 and DTD), but a star of clusters already produces it. The second test finds more in 4 of 12 ImageNet backbones: how each cluster is oriented relative to its superclass center, not a hierarchy among those centers. Rotating each cluster at random removes every verdict, whereas a hierarchy we plant ourselves, synthetic or trained into a model, survives. In the 9 of 12 backbones where a planted hierarchy is detected, none as strong is found; in the other 3 the test cannot see a planted hierarchy. A backbone trained in hyperbolic space shows the same structure as its Euclidean twin and stays nearly flat. (iii) *Across models,* the trees agree on which classes group together well above chance, though less than two resampled versions of the same model, and not on distances; the self-supervised models organize classes by direction rather than by distance, which a comparison by distance misses.

Our contributions are:

1. **A new instrument:** A calibrated reading of δ, compared with random clouds of the same shape, and a second test whose false-alarm rate and power are measured on real clouds.

2. **A new census:** The premise calibrated where it is read, including a published reading reproduced; structure beyond a random cloud in most models; non-random orientation of clusters relative to their superclass centers in a few; no detectable hierarchy among those centers where the test has power; and a hyperbolic-trained backbone as the control for imposing the geometry.

3. **A new map of trees:** Across models, the trees agree on which classes group together but not on distances, and the self-supervised models organize classes by direction, which a comparison by distance misses.

4. **Guidance for practice.** The rule that derives curvature from a raw δ assigns curvature to random clouds; we say what to measure before imposing curvature, and cosine collects most of the structure at no cost.

2 Related Work

Reading δ against a matched reference has precedent in network science, where the δ of real graphs is compared with that of random and model graphs (Narayan & Saniee, 2011; Kennedy et al., 2013; Adcock et al., 2013). We bring the idea to embedding clouds, where the reference must match dimension and spectrum. Curvature has been studied as a representation tradeoff (Sala et al., 2018) and as a mixed-curvature product to be learned (Gu et al., 2019). In few-shot learning Moreira et al. (2024) find that fixed-radius Euclidean prototypes match hyperbolic ones. The analysis of learned representations characterizes Euclidean structure and does not test for tree-likeness: representation similarity, intrinsic dimension and neural collapse (Kornblith et al., 2019; Ansuini et al., 2019; Pope et al., 2021; Papyan et al., 2020). Closest to us is the work on representational convergence: the Platonic Representation Hypothesis (Huh et al., 2024) argues that models converge toward shared representations, and its cross-modal convergence has been stress-tested at scale (Koepke et al., 2026); Gröger et al. (2026) show that permutation-calibrated similarity removes most of that convergence while local neighborhood structure survives, and Park et al. (2025) read concept hierarchy in language models from the orthogonality of concept directions. We are the geometric complement of this program: the same calibration applied within a model instead of across models.

3 Methodology

3.1 Background: Gromov δ

Tree-likeness is measured on quadruples. Take any four points w, x, y, z of a metric space with distance d. They can be paired in three ways, and each pairing has a sum of two distances:

$$d(w,x) + d(y,z), \qquad d(w,y) + d(x,z), \qquad d(w,z) + d(x,y). \tag{1}$$

Sort the three sums as $S_{\max} \geq S_{\text{mid}} \geq S_{\min}$. In a tree, the two largest sums are always equal, because the two pairings that cross the trunk traverse the same path, so the four-point defect $(S_{\max} - S_{\text{mid}})/2$ is zero. The further a metric is from a tree, the larger the defect can become. A metric tree has $\delta = 0$. Euclidean space has no bound at all: enlarging a square enlarges its defect in proportion. Hyperbolic space sits in between, with a bounded defect whose value for the hyperbolic plane is known, $\ln 2$ under this definition (Nica & Špakula, 2016).

Definition 1 (four-point defect)**.** For a quadruple (w, x, y, z) with the pairing sums of Eq. 1 sorted as $S_{\max} \geq S_{\text{mid}} \geq S_{\min}$, the four-point defect is $\delta(w,x,y,z) = \frac{1}{2}(S_{\max} - S_{\text{mid}})$, and the Gromov δ of a metric space is the supremum of the defect over its quadruples (Gromov, 1987).

3.2 Estimation and normalization

Three steps turn δ into a number that can be compared across models. First, a thousand centroids have far more quadruples than can be enumerated, and exact computation takes more than cubic time (Fournier et al., 2015). The defect is therefore computed on half a million random quadruples. Second, the largest defect in that sample is a poor estimate of the supremum: it belongs to a single extreme quadruple and keeps growing as more quadruples are drawn. We therefore read the 99.9th percentile of the sampled defects, $\hat{\delta}_{99.9}$, the value that 99.9 per cent of them fall below. It keeps the tail of the defects, as the supremum does, but it is set by hundreds of quadruples rather than by one. It also does not move with the sample size: across quadruple budgets the ImageNet reading drifts by less than half its own spread (Figure 6). Third, the defect grows with the size of the point set, so we divide by the diameter, the largest pairwise distance:

$$\delta_{\text{norm}} = \frac{\hat{\delta}_{99.9}}{\text{diam}}. \tag{2}$$

On this scale a tree reads zero, and a hyperbolic region of radius four or more reads below a sphere (Table 2).

Definition 2 (reading)**.** δ_{norm} of Eq. 2, with $\hat{\delta}_{99.9}$ computed over half a million sampled quadruples per seed and averaged over 10 seeds, is the statistic used throughout; we call it the *reading*.

Proposition 1 (range bound and dimension confound)**.** *Let $d_{\min}$ and $d_{\max}$ be the smallest and the largest pairwise distance of a finite metric space. (a) For any quadruple the four-point defect is at most $d_{\max} - d_{\min}$, hence $\delta_{norm} \leq 1 - d_{\min}/d_{\max}$. (b) For an iid Gaussian cloud of n points in dimension d, with n fixed, $d_{\min}/d_{\max} \to 1$ as d grows (Beyer et al., 1999; Aggarwal et al., 2001), so $\delta_{norm} \to 0$ without any structure.*

A raw value cannot be called low on its own. Proposition 1 makes the dimension confound precise; the calibration on reference geometries shows the spectrum and statistic confounds (Table 2).

3.3 Spectrum-matched null, excess and genuineness

Definition 3 (Haar null)**.** A null replicate is a cloud with the real shape and no structure. Let $X = U\Sigma V^\top$ be the singular value decomposition of the centered cloud; the replicate is $X' = Q\Sigma V^\top$ with Q the orthogonal factor of a column-centered Gaussian matrix, orthogonal to the all-ones vector and Haar-distributed on the centered subspace, so X' has exactly the centered singular values of X and structureless coefficients.

Of the two constructions, one reproduces the shape exactly. Of the two compared, Gaussian coefficients reproduce the spectrum only on average and inflate the excess of low-rank clouds. The centered Haar construction reproduces the centered spectrum exactly, so it is the null used throughout. The uncentered and the Gaussian readings are reported beside it (Table 5).

Definition 4 (excess)**.** With K null replicates $X'_1, \ldots, X'_K$ of Definition 3, the excess of a cell is:

$$e(X) = \delta_{\text{norm}}(X) - \frac{1}{K}\sum_{k=1}^{K} \delta_{\text{norm}}(X'_k). \tag{3}$$

Size and evidence. The excess is a size, the gap between a reading and its random cloud in Figure 2a. The evidence is the rank of the reading among its replicates and the left-tail p that the rank gives:

$$r = \#\{k : \delta_{\text{norm}}(X'_k) > \delta_{\text{norm}}(X)\}, \qquad p = \frac{1 + K - r}{K + 1}, \qquad K = 200. \tag{4}$$

Definition 5 (genuine)**.** Order the m left-tail p-values of a census as $p_{(1)} \leq \cdots \leq p_{(m)}$. A model–dataset cell is genuine when its p is at or below the Benjamini–Hochberg threshold (Benjamini & Hochberg, 1995) at level five per cent:

$$p \;\leq\; p_{(i^*)}, \qquad i^* = \max\{i : p_{(i)} \leq 0.05\, i/m\}, \tag{5}$$

and the word is used in this sense only.

What a genuine excess means. A negative genuine excess, more than chance would give, certifies structure beyond the second moments, that is, beyond what the mean and the covariance explain. Clustering is its most plausible reading, supported by the superclass recovery of Section 5.4. Two cells of the same dataset can have the same reading, 0.082, ViT-B on CIFAR-100 images and DINO-B on its centroids, and only the second is genuine.

3.4 Hierarchy test

Why a second test. A genuine excess can come from clusters alone, so a second test, which we call the hierarchy test, asks whether the clusters are themselves arranged hierarchically. It needs three constructions.

Definition 6 (hub null and hub excess)**.** Given a grouping of classes into superclasses, assigning each centroid x_i to a cluster $k(i)$ with hub $h_{k(i)}$, the cluster mean, write $x_i = h_{k(i)} + o_i$. The hub null keeps every offset o_i and replaces the hubs by a Haar resample with their spectrum, $x'_i = h'_{k(i)} + o_i$, and the excess of Eq. 3 against this null, e_{hub}, measures structure above the clusters of the grouping only.

Definition 7 (matched star)**.** The matched star of a cloud keeps the real clusters, each as a Haar sample with the cluster's own covariance, and places their hubs as Gaussian draws at the real hub radius: the clustering of the real cloud with no hierarchy above it. Its spectrum-matched-hub variant draws the hubs as a Haar resample of the real hubs.

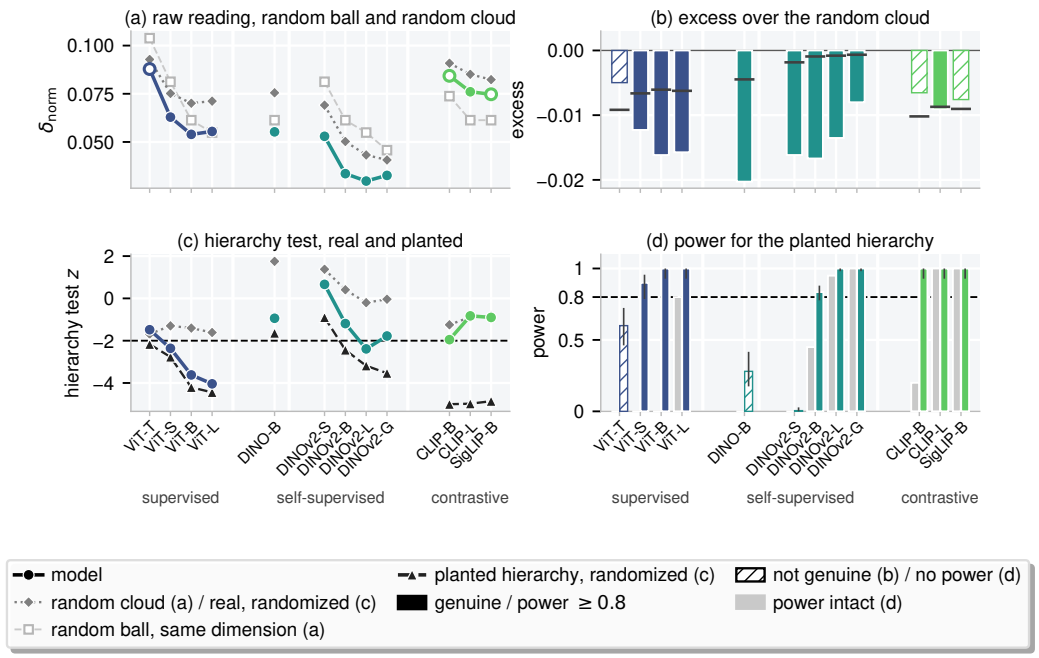

Figure 2: **The instrument in one figure.** (a) Raw reading, a random ball of the same dimension, and a random cloud of the same shape; the ball-to-cloud gap is the spectrum's share and the cloud-to-model gap is the excess. (b) Their difference, the excess, filled when genuine. (c) The hierarchy test on the real cloud, with its orientations randomized, and on a planted hierarchy randomized the same way. (d) The test's power for the planted hierarchy, intact and randomized; whiskers: 95 per cent intervals over 50 runs per backbone, 200 for the DINOv2 family.

Definition 8 (rotating each cluster)**.** This control keeps the hubs of the grouping and rotates each cluster's offsets by an independent Haar rotation. A verdict that survives the rotation is carried by the arrangement of the hubs; one that does not is carried by the orientation of the clusters relative to their hubs, which we call *alignment*.

The hierarchy test compares the real cloud with its matched star. With e_{hub} read on the real cloud and on its matched star, the depth statistic and its standardized form are:

$$\mathrm{depth} = e_{\mathrm{hub}}(X) - e_{\mathrm{hub}}(X_{\mathrm{star}}), \qquad z = \mathrm{depth} \,/\, \mathrm{s.d.}, \tag{6}$$

where the spread is taken over star seeds and null replicates, and a cloud is certified against its matched star when $z \leq -2$, a two-standard-deviation threshold whose false alarms on flat controls and power on planted hierarchies are measured on real clouds rather than assumed (Section 5.3).

3.5 Comparing trees across models

Trees are compared under a configuration chosen before the comparison. Two hierarchies are compared by the adjusted Rand index between their cut partitions and by the correlation of their cophenetic distances. They are also compared by their agreement on which pair of a random class triplet merges first. The configuration of metric and linkage is selected per dataset by a criterion fixed without reference to the agreements. It is the highest cophenetic fidelity among configurations whose reference cuts are non-degenerate for every model. Local agreement is read with the permutation calibration of Gröger et al. (2026) on mutual-kNN agreement (the overlap of nearest neighbours) and centroid CKA (a similarity index).

3.6 Projection and tasks

Three metrics are compared on the intact representation. A parameter-free map centers the features, scales the 95th-percentile norm to radius $1/\sqrt{2}$ and preserves directions. Euclidean, cosine and Poincaré metrics are then scored by nearest-centroid classification and 5-way 5-shot episodes on every cell of the task grid.

4 EXPERIMENTAL SETUP

The census, the reading of every model–dataset cell, has 12 backbones, 6 class sets and 15 text models. The vision backbones are 4 supervised ViTs (Dosovitskiy et al., 2021) trained on ImageNet-21k (Deng et al., 2009), DINO-B (Caron et al., 2021), 4 DINOv2 sizes (Oquab et al., 2024), CLIP-B, CLIP-L (Radford et al., 2021) and SigLIP-B (Zhai et al., 2023). The class sets are ImageNet (1000 classes) and CIFAR-100 (100 classes; Krizhevsky, 2009), which carry a real hierarchy; DTD (47 classes; Cimpoi et al., 2014) and CIFAR-10 (10 classes); and the flat FashionMNIST (FMNIST, 10 classes; Xiao et al., 2017) and MNIST (10 classes; LeCun et al., 1998). The text models read the thousand ImageNet class names: 4 GPT-2 sizes (Radford et al., 2019), 3 Pythia sizes (Biderman et al., 2023), OLMo-1B (Groeneveld et al., 2024) and 7 sentence embedders (Table 1).

Four controls test the alternatives. Two self-supervised ResNets (He et al., 2016; Zbontar et al., 2021; Grill et al., 2020) read the vision census. DeiT-B (Touvron et al., 2021) and the augreg ViT-B (Steiner et al., 2022), trained on ImageNet-1k leaf labels (the 1000 fine classes, with no superclasses), test whether a hierarchical label set explains the alignment. MERU and its Euclidean CLIP twin (Desai et al., 2023) test whether training in hyperbolic space creates hierarchy. The census ViT-B, fine-tuned with and without a hierarchical objective, is the trained control.

The premise is read on images and hierarchy on centroids. The sample-level reading uses about a thousand stratified training images per cell on CIFAR-100 and DTD. Class centroids average a hundred cached training images per class on ImageNet and the full cached split elsewhere. Text models are read one prompt at a time, since batching alters GPT-2's hidden states, each with its own pooling (Table 1).

Calibration budget and groupings. Each cell is read against 200 Haar replicates, with half a million sampled quadruples per seed and 10 seeds. With 10 classes there are only 210 quadruples, so $\hat{\delta}_{99.9}$ coincides with the supremum there. Benjamini–Hochberg runs over the vision census and, separately, over the text census. The grouping into superclasses is the WordNet (Miller, 1995) cut into 30 superclasses on ImageNet and the 20 coarse labels on CIFAR-100. A balanced grouping is the control, with 10 star seeds.

5 RESULTS

5.1 THE PREMISE AT THE SAMPLE LEVEL

The premise does not survive calibration where it is read. On per-image features, where the premise is read, the calibrated reading is not genuine in 14 of 24 cells, that is, no more than chance would give. The excess is conservative: the null keeps the spectrum, so a hierarchy carried by the spectrum alone would not show. What fails is the evidence the premise cites, not the possibility of a hierarchy. Where it is genuine, the excess removes at most two fifths of the null reading. Figure 3a and Table 7 give the 24 cells.

A published reading is reproduced and calibrated. The estimator of Khrulkov et al. (2020), run on our extraction, reproduces the raw $\delta_{\text{rel}} = 2\delta/\mathrm{diam}$ they report for ResNet-34 within 0.03. Calibrated, only MiniImageNet (Vinyals et al., 2016) is genuine. CIFAR-10 and CUB-200 (Wah et al., 2011) are indistinguishable from a random cloud, and CIFAR-100 falls below it only before correction ($p = 0.030$). Under their own statistic, the supremum from a fixed base point, CIFAR-10 and CUB-200 stay indistinguishable from a random cloud and MiniImageNet stays below it. The verdict for CIFAR-100 changes from trial to trial, as the statistic confound predicts. Figure 3b and Table 6 give the 4 datasets.

5.2 STRUCTURE BEYOND THE SECOND MOMENTS AT THE CLASS LEVEL

Most cells show structure beyond the second moments. It is genuine, beyond a random cloud, in 44 of 72 cells and 30 of 36 on the 3 datasets with many classes (ImageNet, CIFAR-100 and DTD). On ImageNet and CIFAR-100 the count is 18 of 24. Figure 4 and Table 4 give the census.

The count survives resampling. Folding image resampling and estimator seeds into the null spread, or repeating the census on resamples, keeps most of it. Table 8 gives the bootstrap.

A star of clusters already passes the census. A synthetic star with no hierarchy scores an excess as deep as any real cell, so a negative excess says structure beyond the second moments, not a hierarchy. Table 3 gives the star's excess.

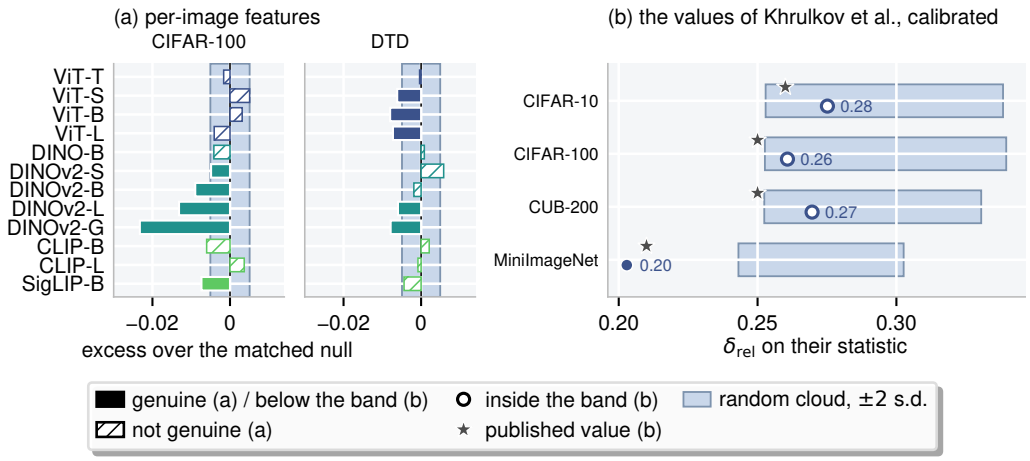

Figure 3: **The premise where it is read.** (a) Excess of the reading on per-image features; filled: genuine, 10 of 24 cells. (b) The values reported by Khrulkov et al. (2020), our reproduction with their estimator, and the random cloud of the same shape, all on their statistic, the supremum from a fixed base point: a dot inside the band is what chance gives. On their statistic CIFAR-10 and CUB-200 fall inside the band, MiniImageNet below it, and CIFAR-100 at its edge, its verdict changing from trial to trial (median rank 196/200); on our reading only MiniImageNet is genuine after correction, and CIFAR-100 falls below at $p = 0.030$ before it (Section 5.1).

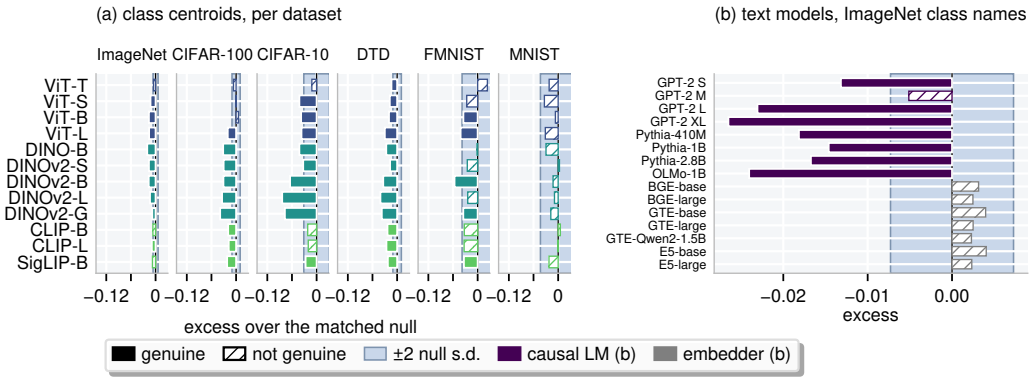

Figure 4: **Most cells show structure beyond the second moments.** (a) Excess of the reading over its matched null per dataset and backbone, filled when genuine; band: two null standard deviations. (b) The same reading for the 15 text models on the ImageNet class names; filled: genuine.

5.3 Hierarchy above the labelled clusters

The hierarchy test certifies that how clusters are oriented relative to their hubs is not random in 4 out of 12 backbones: ViT-S, ViT-B, ViT-L and DINOv2-L. The test compares each cloud with its matched star, the same clusters with no hierarchy above them. It was run on the 12 ImageNet centroid clouds under both stars, with the WordNet grouping into superclasses and its centers as hubs. Figure 2c shows the z per backbone and Tables 9 and 10 give both stars.

What it certifies is alignment, how each cluster is oriented. Once each cluster is rotated at random, the control of Definition 8, none of the 4 backbones fires. ViT-L then reads -1.61, within the -1.61 to -1.76 of the flat control: without its orientations the real cloud reads like a flat one. The alignment is relational: 0 of 60 runs fire with random hubs and 0 of 4 with random orientations, so it needs both. Removing the radial component of every offset leaves it in all 4 certified backbones, z from -2.08 to -3.99 under both stars; feature norms are not its source (Table 14). Under full L2

normalization, which also moves the hubs, it survives in ViT-B and ViT-L only. Its geometric form is not resolved here.

What the test can see. The noise level of a cloud is its within-cluster spread relative to the distance between its hubs. Given a planted two-level tree at this noise level, the test misses it; given a three-level hierarchy with ViT-L's spectrum and noise, it detects it in 4 out of 5 seeds. That detection survives orientation randomization, whereas none of the 4 real verdicts does. The same hierarchy was then built at each backbone's own spectrum and ratio. With clusters rotated the control detects it in 9 of 12 backbones (power 0.83–1.00) over 50 to 200 runs per backbone, and not in the other 3 (power 0.01–0.60). The 3 that cannot see a planted hierarchy, ViT-T, DINO-B and DINOv2-S, sit at noise levels 1.5 to 3.0, inside the 1.3 to 3.9 spanned by the 9 covered backbones.

A trained hierarchy survives, and the test leans toward firing with clusters rotated. ViT-B fine-tuned with a hierarchical cross-entropy keeps firing with its clusters rotated in 20 of 20 runs over 2 seeds, mean z -2.51, on the WordNet grouping. The leaf-only fine-tune fires in 6 of 20, mean z -1.37, and the frozen checkpoint in none. Fine-tuning itself adds some signal with clusters rotated and the hierarchical objective adds more, so the trained control separates the objectives by degree. A flat control with the same spectrum and ratio as the deep one fires in 0 of 5 intact runs and in 8 of 50 runs with clusters rotated. The test is therefore biased toward firing, so the failure of the 4 real backbones to fire once rotated is conservative evidence. The synthetic hierarchy at z -4.19 to -4.82 stands well clear of that bias. On the balanced grouping the matched flat control fires in 0 of 50 runs with clusters rotated. Tables 13 and 15 give the runs.

No hub hierarchy is detected where the test has power. In the 9 backbones, none is found as strong as the planted one. Rotating each cluster separates structure among the hubs of the grouping from structure carried by cluster orientations. A hierarchy below the grouping expressed in how clusters open would be removed with the orientations and counted as alignment. So, no hub hierarchy therefore means none among the tested superclass hubs. Figures 7 and 8 give the detection rates and the power.

The certified set depends on the grouping. Rerun under a balanced grouping, ViT-T joins the certified set and ViT-S and DINOv2-L leave it; other WordNet cuts change the set again (Table 9).

Leaf labels can produce the alignment but do not guarantee it. Read like the census models, the augreg ViT-B trained on ImageNet-1k leaf labels is certified, and DeiT-B, trained on the same labels with another recipe, is not (Table 11).

Training in hyperbolic space leaves the clustering unchanged. Read with the census and the hierarchy test, MERU shows the same clustering as its Euclidean twin. No hub structure is detected, although the test's power at MERU's spectrum was not measured. Its embeddings also stay nearly flat: 95 per cent lie within 0.29 curvature radii of the origin, where the space is nearly flat (Table 12).

5.4 WHOSE TREE

The naive map manufactures an island. Compared with Euclidean distances and average linkage, the DINOv2 models agree with none of the others, an island in the map of models. They agree with the 7 supervised and contrastive models far less than those 7 agree among themselves (DINO-B is not part of the island). The cause is chaining: DINOv2's Euclidean dendrograms put almost every class in one cluster, forcing the index to zero (Figure 5a).

Controlled, the trees share their topology, not their metric. Across models, triplet agreement reaches 0.77, against 0.90 for two resamples of the same model. Chance gives 0.33. The cophenetic and cut agreements reach about half their within-model ceiling (Table 16 and Figure 5b). Calibrated local agreement across models survives, as Gröger et al. (2026) find for similarity (Table 17).

The self-supervised tree is angular. Sibling triplets were resolved under cosine and under Euclidean distance: DINOv2 resolves them far better under cosine, so its semantics concentrate in the angular component and its Euclidean dendrograms chain. Repeated on cosine geometry, the census agrees with the Euclidean one in most cells, and DINOv2 on ImageNet is genuine, more than chance would give, under both. Figure 5c shows the triplets and Table 5 gives the cosine census.

Agreement with WordNet follows supervision. The correlation between inter-centroid and WordNet distances is highest for the contrastive vision–language models (VLMs, $+0.57$ to $+0.59$), then the supervised ViTs ($+0.49$ to $+0.53$), and lowest for DINOv2 ($+0.18$ to $+0.22$). On DBpedia

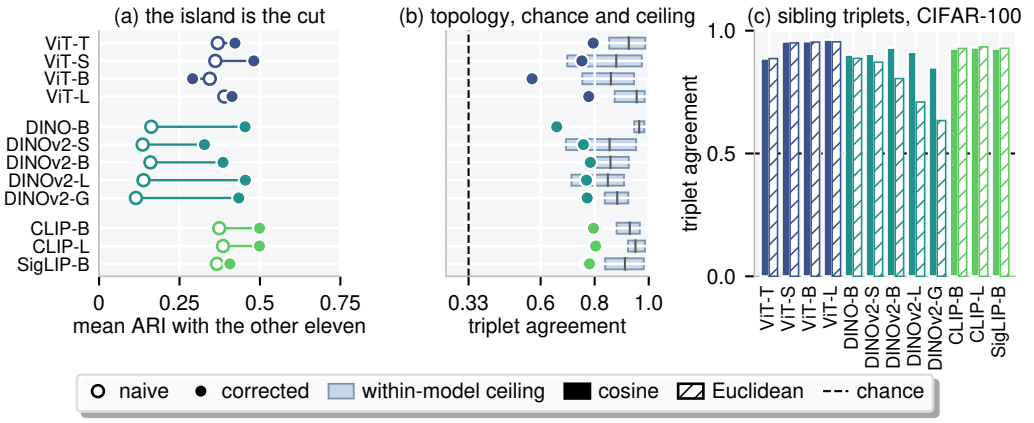

Figure 5: **The island is the cut; tree topology is shared well above chance, though short of the within-model ceiling; the self-supervised structure is angular.** (a) Mean agreement of each model's tree with the other 11, naive (hollow) and corrected (filled). (b) Triplet agreement under the corrected comparison against the within-model ceiling (band). (c) Sibling-triplet agreement under cosine and Euclidean distance.

Classes (Lehmann et al., 2015) and on HierarCaps (Alper & Averbuch-Elor, 2024), two hierarchies independent of WordNet, the text encoders read a genuine excess and resolve sibling triplets (two classes that share a WordNet parent and a third that does not) far above chance. Tables 18, 19 and 20 give the three readings.

Text depends on recipe and scale. 7 of 15 text models are genuine under the reading (Figure 4b). GPT-2 L and XL are genuine under every reading while S and M change with it. Pythia is genuine at every size. The sentence embedders are not genuine on class names (Table 21).

6 Implications for Hyperbolic Representation Learning

The raw reading cannot select a curvature. Khrulkov et al. (2020) set the curvature of the Poincaré ball from the relative hyperbolicity, $c = (0.144/\delta_{\text{rel}})^2$ with $\delta_{\text{rel}} = 2\delta/\text{diam}$. Applied, with the supremum it was calibrated for, to structureless Gaussian clouds, it assigns curvatures from 0.48 to 2.5. The raw and calibrated readings predict the gain of a non-Euclidean metric equally well, and the hierarchy verdict predicts none (Table 23). Calibration tells what a low δ means, not which metric to use. The Poincaré metric adds $+0.9$ to $+1.3$ % points over cosine for the contrastive VLMs on nearest-centroid and few-shot tasks and nothing consistent elsewhere (Table 22 and Figure 9). We test zero-cost metrics only; whether hyperbolic training helps for other reasons is outside this study.

7 Conclusion and Limitations

Conclusion. We revisited the premise that a low Gromov δ is evidence of latent hierarchy: structureless clouds of the same shape score as low, so it does not survive calibration. Foundation models do carry non-random class structure, but it is clustering rather than detectable hierarchy among the superclasses where the test has power, partly shared across models and angular in self-supervised ones. Raw tree-likeness should not, by itself, motivate hyperbolic geometry or curvature.

Limitations. (i) The null keeps the spectrum, so a hierarchy carried by the spectrum alone would not show. (ii) The verdict of no detectable hub hierarchy holds only in the 9 of 12 backbones where the test detects a planted hierarchy. (iii) The text census depends on the probe: template and batching move the reading, so marginal verdicts are fragile. (iv) The analysis rests on many choices; a pre-specified analysis is future work. (v) The census covers class centroids on 6 datasets, and the sample-level reading 2 datasets. (vi) Dividing by the diameter lets heavier tails lower δ_{norm}; the cosine census mitigates this. (vii) The hierarchy test is Euclidean, possibly conservative for the angular DINOv2 tree, and the objective–geometry link is correlational.

Reproducibility Statement

The supplementary material contains the scripts that produce every result in the paper, with fixed seeds, and a README that maps each result to its script; its `tool/` directory ships the instrument as one script that reproduces any cell of Table 4 from its centroid matrix.

Ethics Statement

This work analyzes publicly released models and standard public datasets, involves no human subjects and no personal data, and proposes an analysis methodology. We do not foresee ethical risks beyond those already associated with the underlying models and datasets.

AI Use Statement

We used generative AI tools to assist with writing and editing, retrieving references, and implementing and running the experimental code, and to give feedback on drafts and on the methodology. We reviewed and verified all AI-assisted work, including every number, figure and reference, and take full responsibility for the content of this paper.

References

Aaron B. Adcock, Blair D. Sullivan, and Michael W. Mahoney. Tree-like structure in large social and information networks. In *IEEE International Conference on Data Mining (ICDM)*, pp. 1–10, 2013. doi: 10.1109/ICDM.2013.77.

Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the surprising behavior of distance metrics in high dimensional space. In Jan Van den Bussche and Victor Vianu (eds.), *Database Theory - ICDT 2001, 8th International Conference, London, UK, January 4-6, 2001, Proceedings*, volume 1973 of *Lecture Notes in Computer Science*, pp. 420–434. Springer, 2001. doi: 10.1007/3-540-44503-X_27. URL https://doi.org/10.1007/3-540-44503-X_27.

Morris Alper and Hadar Averbuch-Elor. Emergent visual-semantic hierarchies in image-text representations. In *European Conference on Computer Vision (ECCV)*, 2024. doi: 10.1007/978-3-031-72943-0_13.

Alessio Ansuini, Alessandro Laio, Jakob H. Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 6109–6119, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/cfcce0621b49c983991ead4c3d4d3b6b-Abstract.html.

Mina Ghadimi Atigh, Julian Schoep, Erman Acar, Nanne van Noord, and Pascal Mettes. Hyperbolic image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4453–4462, June 2022. doi: 10.1109/CVPR52688.2022.00441.

Ahmad Bdeir, Kristian Schwethelm, and Niels Landwehr. Fully hyperbolic convolutional neural networks for computer vision. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=ekz1hN5QNh.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Kevin S. Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In Catriel Beeri and Peter Buneman (eds.), *Database Theory - ICDT '99, 7th International Conference, Jerusalem, Israel, January 10-12, 1999, Proceedings*, volume 1540 of *Lecture Notes in Computer Science*, pp. 217–235. Springer, 1999. doi: 10.1007/3-540-49257-7_15. URL https://doi.org/10.1007/3-540-49257-7_15.

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 Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pp. 2397–2430. PMLR, 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, October 2021. doi: 10.1109/ICCV48922.2021.00951.

Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 4869–4880, 2019.

Weize Chen, Xu Han, Yankai Lin, Hexu Zhao, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. Fully hyperbolic neural networks. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5672–5686, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.389. URL https://aclanthology.org/2022.acl-long.389/.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pp. 3606–3613. IEEE Computer Society, 2014. doi: 10.1109/CVPR.2014.461. URL https://doi.org/10.1109/CVPR.2014.461.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pp. 248–255. IEEE Computer Society, 2009. doi: 10.1109/CVPR.2009.5206848. URL https://doi.org/10.1109/CVPR.2009.5206848.

Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Ramakrishna Vedantam. Hyperbolic image-text representations. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023. URL https://proceedings.mlr.press/v202/desai23a.html.

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. URL https://openreview.net/forum?id=YicbFdNTTy.

Aleksandr Ermolov, Leyla Mirvakhabova, Valentin Khrulkov, Nicu Sebe, and Ivan V. Oseledets. Hyperbolic vision transformers: Combining improvements in metric learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 7399–7409. IEEE, 2022. doi: 10.1109/CVPR52688.2022.00726. URL https://doi.org/10.1109/CVPR52688.2022.00726.

Hervé Fournier, Anas Ismail, and Antoine Vigneron. Computing the Gromov hyperbolicity of a discrete metric space. *Information Processing Letters*, 115(6-8), 2015. doi: 10.1016/j.ipl.2015.02.002.

Octavian Ganea, Gary Becigneul, and Thomas Hofmann. Hyperbolic neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.,

2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/dbab2adc8f9d078009ee3fa810bea142-Abstract.html.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f3ada80d5c4ee70142b17b8192b2958e-Abstract.html.

Dirk Groeneveld, Iz Beltagy, Evan Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. Olmo: Accelerating the science of language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 15789–15809. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.841. URL https://doi.org/10.18653/v1/2024.acl-long.841.

Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the Platonic representation hypothesis: An Aristotelian view. In *International Conference on Machine Learning (ICML)*, 2026. doi: 10.48550/ARXIV.2602.14486. URL https://doi.org/10.48550/arXiv.2602.14486.

M. Gromov. Hyperbolic groups. In S. M. Gersten (ed.), *Essays in Group Theory*, pp. 75–263. Springer New York, New York, NY, 1987. ISBN 978-1-4613-9586-7. doi: 10.1007/978-1-4613-9586-7_3. URL https://doi.org/10.1007/978-1-4613-9586-7_3.

Albert Gu, Frederic Sala, Beliz Gunel, and Christopher Ré. Learning mixed-curvature representations in product spaces. In *International Conference on Learning Representations*, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. IEEE, 2016. doi: 10.1109/CVPR.2016.90.

Neil He, Rishabh Anand, Hiren Madhu, Ali Maatouk, Smita Krishnaswamy, Leandros Tassiulas, Menglin Yang, and Rex Ying. HELM: Hyperbolic large language models via mixture-of-curvature experts. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=RnbJPkakkm.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The Platonic representation hypothesis. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pp. 20617–20642. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/huh24a.html.

W. Sean Kennedy, Onuttom Narayan, and Iraj Saniee. On the hyperbolicity of large-scale networks. *arXiv preprint arXiv:1307.0031*, 2013. URL https://arxiv.org/abs/1307.0031.

Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. doi: 10.1109/CVPR42600.2020.00645.

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

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey E. Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, Proceedings of Machine Learning Research, pp. 3519–3529. PMLR, 2019. URL http://proceedings.mlr.press/v97/kornblith19a.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791. URL https://doi.org/10.1109/5.726791.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. DBpedia – a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6 (2):167–195, 2015. doi: 10.3233/SW-140134.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *CoRR*, abs/2308.03281, 2023. doi: 10.48550/ARXIV.2308.03281. URL https://doi.org/10.48550/arXiv.2308.03281.

George A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11): 39–41, 1995. doi: 10.1145/219717.219748.

Gabriel Moreira, Manuel Marques, João Paulo Costeira, and Alexander Hauptmann. Hyperbolic vs Euclidean embeddings in few-shot learning: Two sides of the same coin. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2082–2090, 2024. doi: 10.1109/WACV57701.2024.00208.

Onuttom Narayan and Iraj Saniee. Large-scale curvature of networks. *Physical Review E*, 84(6): 066108, 2011. doi: 10.1103/PhysRevE.84.066108.

Bogdan Nica and Ján Špakula. Strong hyperbolicity. *Groups, Geometry, and Dynamics*, 10(3), 2016. doi: 10.4171/GGD/372. arXiv:1408.0250.

Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 6338–6347, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/59dfa2df42d9e3d41f5b02bfc32229dd-Abstract.html.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Avik Pal, Max van Spengler, Guido Maria D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. Compositional entailment learning for hyperbolic vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=3i13Gev2hV.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020. doi: 10.1073/pnas.2015509117. URL https://www.pnas.org/doi/abs/10.1073/pnas.2015509117.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2406.01506.

Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=XJk19XzGq2J.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Frederic Sala, Christopher De Sa, Albert Gu, and Christopher Ré. Representation tradeoffs for hyperbolic embeddings. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4460–4469. PMLR, 2018. URL https://proceedings.mlr.press/v80/sala18a.html.

Rik Sarkar. Low distortion Delaunay embedding of trees in hyperbolic plane. In Marc J. van Kreveld and Bettina Speckmann (eds.), *Graph Drawing - 19th International Symposium, GD 2011, Eindhoven, The Netherlands, September 21-23, 2011, Revised Selected Papers*, Lecture Notes in Computer Science, pp. 355–366. Springer, 2011. doi: 10.1007/978-3-642-25878-7_34. URL https://doi.org/10.1007/978-3-642-25878-7_34.

Aditya Sinha, Siqi Zeng, Makoto Yamada, and Han Zhao. Learning structured representations with hyperbolic embeddings. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 91220–91259. Curran Associates, Inc., 2024. doi: 10.52202/079017-2895.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your ViT? data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research*, 2022. URL https://arxiv.org/abs/2106.10270.

Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. Poincaré GloVe: Hyperbolic word embeddings. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://arxiv.org/abs/1810.06546.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10347–10357. PMLR, 2021. URL https://proceedings.mlr.press/v139/touvron21a.html.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/hash/90e1357833654983612fb05e3ec9148c-Abstract.html.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD birds-200-2011 dataset. Computation & Neural Systems Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. URL https://authors.library.caltech.edu/records/cvm3y-5hh21.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *CoRR*, abs/2212.03533, 2022. doi: 10.48550/ARXIV.2212.03533. URL https://doi.org/10.48550/arXiv.2212.03533.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017. URL http://arxiv.org/abs/1708.07747.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packed resources for general Chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pp. 641–649, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657878. URL https://doi.org/10.1145/3626772.3657878.

Menglin Yang, Ram Samarth B B, Aosong Feng, Bo Xiong, Jiahong Liu, Irwin King, and Rex Ying. Hyperbolic fine-tuning for large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=TkEdQv0bXB.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320. PMLR, 2021. URL https://proceedings.mlr.press/v139/zbontar21a.html.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01100.

A PROOFS

Proposition 1 (range bound and dimension confound). *Let $d_{\min}$ and $d_{\max}$ be the smallest and the largest pairwise distance of a finite metric space. (a) For any quadruple, the four-point defect is at most $d_{\max} - d_{\min}$; hence $\delta_{norm} \le 1 - d_{\min}/d_{\max}$. (b) For an iid Gaussian cloud of n points in dimension d, with n fixed, $d_{\min}/d_{\max} \to 1$ as d grows, so $\delta_{norm} \to 0$ without any structure.*

Intuition. The defect is half the difference between two sums of distances. If all distances are nearly equal, every sum is nearly equal, and the difference cannot be large. Part (a) makes this exact; part (b) shows that a random cloud in high dimension has nearly equal distances.

Proof of (a). Take any four points and their three pairing sums (Eq. 1). Each sum adds two distances, and every distance lies between $d_{\min}$ and $d_{\max}$, so every sum lies between $2d_{\min}$ and $2d_{\max}$:

$$2d_{\min} \;\le\; S_{\min} \;\le\; S_{\mathrm{mid}} \;\le\; S_{\max} \;\le\; 2d_{\max}. \tag{7}$$

The defect is therefore bounded by the width of that interval:

$$\tfrac{1}{2}\big(S_{\max} - S_{\mathrm{mid}}\big) \;\le\; \tfrac{1}{2}\big(2d_{\max} - 2d_{\min}\big) \;=\; d_{\max} - d_{\min}. \tag{8}$$

The bound holds for every quadruple, so it holds for their supremum δ and for any percentile of the sampled defects, which never exceeds the supremum. The diameter of the set is $d_{\max}$; dividing by it gives

$$\delta_{\text{norm}} \;\le\; \frac{d_{\max} - d_{\min}}{d_{\max}} \;=\; 1 - \frac{d_{\min}}{d_{\max}}. \qquad \square \tag{9}$$

Proof of (b). Let $x_1, \dots, x_n \in \mathbb{R}^d$ have iid standard Gaussian coordinates. For one pair $i \neq j$, the squared distance is a sum of d independent terms:

$$\|x_i - x_j\|^2 \;=\; \sum_{k=1}^{d} \big(x_{ik} - x_{jk}\big)^2. \tag{10}$$

Each term is the square of a Gaussian of variance 2, so it has mean 2. By the law of large numbers, the average of the d terms tends to its mean:

$$\frac{\|x_i - x_j\|^2}{2d} \;\longrightarrow\; 1 \qquad \text{almost surely, as } d \to \infty. \tag{11}$$

With n fixed there are finitely many pairs, $n(n-1)/2$, so the convergence holds for all of them at once, and in particular for the smallest and the largest distance:

$$\frac{d_{\min}}{\sqrt{2d}} \to 1, \qquad \frac{d_{\max}}{\sqrt{2d}} \to 1, \qquad \text{hence} \qquad \frac{d_{\min}}{d_{\max}} \to 1. \tag{12}$$

Part (a) then gives $\delta_{\text{norm}} \le 1 - d_{\min}/d_{\max} \to 0$. The relative contrast between the farthest and the nearest pair vanishes at rate $d^{-1/2}$ (Beyer et al., 1999; Aggarwal et al., 2001): this is the dimension confound, with no structure in the cloud. $\square$

B IMPLEMENTATION

What this table answers. Every model named anywhere in the paper appears here once, with its size and its role: census, control or text. The notes below the table say how features are taken from each kind of model, which is part of the measurement rather than a detail of implementation.

type	model	params
Supervised ViT (vision)	ViT-T (i21k) (Dosovitskiy et al., 2021)	5M
Supervised ViT (vision)	ViT-S (i21k) (Dosovitskiy et al., 2021)	22M
Supervised ViT (vision)	ViT-B (i21k) (Dosovitskiy et al., 2021)	86M
Supervised ViT (vision)	ViT-L (i21k) (Dosovitskiy et al., 2021)	307M
SSL (vision)	DINO-B (Caron et al., 2021)	86M
SSL (vision)	DINOv2-S (Oquab et al., 2024)	22M
SSL (vision)	DINOv2-B (Oquab et al., 2024)	86M
SSL (vision)	DINOv2-L (Oquab et al., 2024)	307M
SSL (vision)	DINOv2-G (Oquab et al., 2024)	1.1B
Contrastive (vision)	CLIP-B (Radford et al., 2021)	86M
Contrastive (vision)	CLIP-L (Radford et al., 2021)	307M
Contrastive (vision)	SigLIP-B (Zhai et al., 2023)	86M
Causal LM	GPT-2 S (Radford et al., 2019)	117M
Causal LM	GPT-2 M (Radford et al., 2019)	345M
Causal LM	GPT-2 L (Radford et al., 2019)	774M
Causal LM	GPT-2 XL (Radford et al., 2019)	1.5B
Causal LM	Pythia-410M (Biderman et al., 2023)	410M
Causal LM	Pythia-1B (Biderman et al., 2023)	1.0B
Causal LM	Pythia-2.8B (Biderman et al., 2023)	2.8B
Causal LM	OLMo-1B (Groeneveld et al., 2024)	1.0B
Text embedder	BGE-base (Xiao et al., 2024)	110M
Text embedder	BGE-large (Xiao et al., 2024)	335M
Text embedder	GTE-base (Li et al., 2023)	110M
Text embedder	GTE-large (Li et al., 2023)	335M
Text embedder	GTE-Qwen2-1.5B (Li et al., 2023)	1.5B
Text embedder	E5-base (Wang et al., 2022)	110M
Text embedder	E5-large (Wang et al., 2022)	335M
Controls (vision)	DeiT-B/16 (IN-1k, no distillation) (Touvron et al., 2021)	86M
Controls (vision)	ViT-B/16 augreg (IN-1k) (Steiner et al., 2022)	86M
Controls (vision)	MERU ViT-S/B/L and CLIP twins (Desai et al., 2023)	22M/86M/307M
Controls (vision)	Barlow Twins (ResNet-50) (Zbontar et al., 2021)	24M
Controls (vision)	BYOL (ResNet-50) (Grill et al., 2020)	24M

Extraction: vision backbones use the encoder's default pooled embedding (timm, CLS or global pool per architecture); causal language models the last-token state of "a photo of a {class}", one prompt at a time, since batching alters GPT-2's hidden states; text embedders their recommended pooling, CLS for BGE, mean for GTE and E5, last token for GTE-Qwen2; the CLIP and SigLIP text towers their own text encoders.

Table 1: **The models and the extraction.** Every backbone, language model and text embedder of the census and of the controls, with its parameter count; the note below the table says how features are extracted and pooled.

C ADDITIONAL RESULTS

Every table is generated by script from the released per-cell result files, and no value is hand-transcribed; the tables are grouped by question, in the order in which the main text first cites each question, and the README of the supplementary material maps each table and figure to its script and its result file.

C.1 IS A LOW RAW READING EVIDENCE? CALIBRATION ON REFERENCE GEOMETRIES AND SYNTHETIC CLUSTERS

What this table answers. A raw reading means nothing until something known is read on the same scale. The first table reads geometries whose answer we know in advance, a tree, a hyperbolic region and a sphere. The second reads clouds we built ourselves, clusters with and without a hierarchy above them, against each null in turn.

space	absolute δ	δ_{norm}
balanced binary tree (depth 10)	0.000	0.000
$\mathbb{H}^2$ ($K=-1$), region radius R=2/4/8/16	0.65/0.69/0.69/0.69	0.162/0.087/0.043/0.022
uniform S^{99} (chord / geodesic)	0.25/0.37	0.143/0.179
iid Gaussian (n=1000), d=192, …, 1536	—	0.104/0.081/0.074/0.061/0.055/0.046

Table 2: **Known geometries read as they should on this scale.** A tree, hyperbolic regions of growing radius, a sphere and Gaussian clouds of growing dimension: the absolute reading and the normalized one.

synthetic cloud (n=1000, d=768)	$\hat{\delta}$	spectrum null A Gaussian	spectrum null A Haar	hub-randomizing null B Gaussian	hub-randomizing null B Haar
30-cluster star, tight (0.1)	0.055	−0.115	−0.115	−0.126	−0.058
30-cluster star, mid (0.3)	0.059	−0.104	−0.103	−0.114	−0.049
30-cluster star, loose (0.6)	0.061	−0.084	−0.082	−0.090	−0.033
2-level hierarchy 6×5 (0.3)	0.044	−0.143	−0.143	−0.113	−0.112

Table 3: **The excess certifies clusters, and a star of clusters passes it.** Synthetic clouds of clusters, with and without a hierarchy: the reading and the excess against the spectrum null and against the hub-randomizing null.

C.2 Is the structure beyond the second moments genuine? The vision census under every construction

What this table answers. This is the census itself, one row per model and class set. The first table gives the reading used in the paper, the second repeats the verdict under other constructions of the null and of the statistic, and the third repeats the census on cosine geometry. A cell is genuine when its excess survives the correction across cells.

The structure depends on the class set more than on the model. The smallest backbone is genuine on one dataset only, and the flat datasets have smaller and less consistent excesses; FMNIST is genuine in 5 of 12 backbones. Table 5 gives the details.

Neural collapse is the flat limit, not what the census sees. Neural collapse predicts that supervised class means converge to a simplex frame (Papyan et al., 2020), a star with no hierarchy. The census does not certify that picture: class means cluster around superclass hubs, which the superclass recovery of Section 5.4 shows.

	IN			C100			C10			DTD			FMNIST			MNIST		
model	exc.	r	p	exc.	r	p	exc.	r	p	exc.	r	p	exc.	r	p	exc.	r	p
ViT-T	-0.005°	169	.159	-0.006°	151	.249	-0.012°	148	.264	−0.014	195	.030	$+0.024^\circ$	19	.905	-0.022°	168	.164
ViT-S	−0.012	200	.005	-0.000°	112	.443	−0.042	199	.010	−0.018	199	.010	-0.027°	192	.045	-0.034°	183	.090
ViT-B	−0.016	200	.005	$+0.006^\circ$	42	.791	−0.038	196	.025	−0.020	200	.005	−0.035	198	.015	-0.007°	122	.393
ViT-L	−0.016	200	.005	−0.020	200	.005	−0.037	200	.005	−0.030	200	.005	−0.041	200	.005	-0.031°	187	.070
DINO-B	−0.020	200	.005	−0.031	200	.005	−0.042	199	.010	−0.026	200	.005	-0.002°	104	.483	-0.030°	182	.095
DINOv2-S	−0.016	200	.005	−0.030	200	.005	−0.033	200	.005	−0.017	200	.005	-0.025°	184	.085	$+0.003^\circ$	79	.607
DINOv2-B	−0.017	200	.005	−0.031	200	.005	−0.065	200	.005	−0.034	200	.005	−0.056	199	.010	-0.012°	142	.294
DINOv2-L	−0.014	200	.005	−0.034	200	.005	−0.084	200	.005	−0.041	200	.005	-0.024°	186	.075	-0.009°	135	.328
DINOv2-G	−0.008	200	.005	−0.039	200	.005	−0.077	200	.005	−0.037	200	.005	−0.035	197	.020	-0.018°	163	.189
CLIP-B	-0.007°	172	.144	−0.019	200	.005	-0.022°	184	.085	−0.023	200	.005	-0.032°	185	.080	$+0.005^\circ$	81	.597
CLIP-L	−0.009	197	.020	−0.018	200	.005	-0.020°	180	.104	−0.026	200	.005	-0.033°	186	.075	-0.001°	97	.517
SigLIP-B	-0.008°	189	.060	−0.022	200	.005	−0.027	197	.020	−0.023	200	.005	−0.033	195	.030	-0.022°	176	.124
Barlow-R50	−0.039	20/20	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
BYOL-R50	−0.020	20/20	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

Table 4: **Structure beyond the second moments is the rule at the class level.** Per cell: the excess over the matched null, the number r of replicates above it and the left-tail p; for Barlow Twins and BYOL, r of 20. $^\circ$: not genuine at the corrected threshold.

model	IN k/5	IN cos	C100 k/5	C100 cos	C10 k/5	C10 cos	DTD k/5	DTD cos	FMNIST k/5	FMNIST cos	MNIST k/5	MNIST cos
ViT-T	2/5	●	2/5	○	0/5	●	5/5	●	0/5	○	0/5	○
ViT-S	5/5	●	2/5	●	5/5	●	5/5	●	4/5	●	0/5	○
ViT-B	5/5	●	0/5	○	5/5	●	5/5	●	5/5	●	0/5	○
ViT-L	5/5	●	5/5	●	5/5	●	5/5	●	5/5	●	2/5	●
DINO-B	3/5	●	5/5	●	5/5	●	5/5	●	0/5	○	0/5	○
DINOv2-S	3/5	●	5/5	●	5/5	●	5/5	●	2/5	●	0/5	○
DINOv2-B	3/5	●	4/5	●	5/5	●	5/5	●	5/5	●	0/5	○
DINOv2-L	4/5	●	4/5	●	5/5	●	5/5	●	4/5	●	0/5	○
DINOv2-G	3/5	●	5/5	●	5/5	●	5/5	●	5/5	●	0/5	○
CLIP-B	2/5	●	5/5	●	4/5	●	5/5	●	0/5	○	0/5	○
CLIP-L	5/5	●	5/5	●	4/5	●	5/5	●	0/5	○	0/5	○
SigLIP-B	2/5	●	5/5	●	5/5	●	5/5	●	5/5	●	4/5	●

Table 5: **Does the verdict change with how it is measured?** Per cell, k/5: how many of five constructions call it genuine, centered Haar, uncentered Haar and Gaussian at $\hat{\delta}_{99.9}$, then Haar and Gaussian at the supremum; cos: the verdict on cosine geometry. All five agree on 37 of 72.

C.3 Does the premise survive at the sample level? The census on per-image features

What this table answers. The premise is read on per-image features, so this table reads the census there instead of on class centroids. Each cell gives the raw statistic beside the excess over its matched null. Where the excess survives, it removes a small fraction of the null reading.

(a) The census cells at the sample level.

model	CIFAR-100 (10 images/class, n=1000) supremum $\hat{\delta}$	$\hat{\delta}_{99.9}$	excess	exc./null	r/200	p	DTD (22 images/class, n=1034) supremum $\hat{\delta}$	$\hat{\delta}_{99.9}$	excess	exc./null	r/200	p
ViT-T	0.130	0.093	-0.0016°	-2%	134	0.333	0.124	0.087	-0.0004°	-0%	111	0.448
ViT-S	0.121	0.086	$+0.0051^{\circ}$	$+6\%$	21	0.896	0.103	0.070	-0.0062	-8%	197	0.020
ViT-B	0.121	0.082	$+0.0031^{\circ}$	$+4\%$	50	0.751	0.092	0.060	-0.0080	-12%	200	0.005
ViT-L	0.100	0.067	-0.0041°	-6%	185	0.080	0.092	0.056	-0.0073	-12%	200	0.005
DINO-B	0.104	0.070	-0.0042°	-6%	194	0.035	0.110	0.075	$+0.0008^{\circ}$	$+1\%$	80	0.602
DINOv2-S	0.100	0.065	-0.0049	-7%	200	0.005	0.145	0.090	$+0.0058^{\circ}$	$+7\%$	0	1.000
DINOv2-B	0.089	0.050	-0.0090	-15%	200	0.005	0.139	0.073	-0.0019°	-3%	162	0.194
DINOv2-L	0.087	0.043	-0.0131	-24%	200	0.005	0.131	0.064	-0.0060	-9%	200	0.005
DINOv2-G	0.102	0.036	-0.0232	-39%	200	0.005	0.136	0.065	-0.0079	-11%	200	0.005
CLIP-B	0.117	0.080	-0.0060°	-7%	184	0.085	0.125	0.089	$+0.0021^{\circ}$	$+2\%$	71	0.647
CLIP-L	0.132	0.084	$+0.0037^{\circ}$	$+5\%$	35	0.826	0.120	0.082	-0.0008°	-1%	124	0.383
SigLIP-B	0.113	0.078	-0.0074	-9%	199	0.010	0.114	0.079	-0.0044°	-5%	184	0.085

Table 7: **At the sample level, where the premise is read, the reading sits within null noise in most cells.** Per cell: the statistic, the excess, the excess as a fraction of the null, the rank and the left-tail p. $^{\circ}$: not genuine at the corrected threshold. r/200: replicates above the real value.

					the supremum from a fixed base point	
dataset	their δ_{rel}	ours, their estimator	excess	r/200, largest p	excess	r/200, p range
CIFAR-10	0.26	0.275	−0.0037	173/200, 0.333	−0.0205	167/200, 0.005–0.701
CIFAR-100	0.25	0.261	−0.0087	198/200, 0.030	−0.0353	196/200, 0.005–0.299
CUB-200	0.25	0.270	−0.0043	177/200, 0.249	−0.0218	174/200, 0.005–0.488
MiniImageNet	0.21	0.203	−0.0234	200/200, 0.005	−0.0700	200/200, 0.005–0.005

Table 6: **A published reading reproduced and calibrated.** ResNet-34 setting of Khrulkov et al. (2020): their relative hyperbolicity, ours with their estimator, then the excess and the rank of 200 with the largest left-tail p for the reading (99.9th percentile). For their own statistic, the supremum, against the same replicates: the mean excess, the median rank over 10 trials and the range of p across trials. CIFAR-100 is below the null in the median trial but not in every trial.

C.4 Is the reading robust? Quadruple budget and resampling

What this table answers. Two choices could have made the reading look structured: how many quadruples we sample and which images stand behind each centroid. Each is varied in turn and the same cells are read again. The verdicts move only where the excess was already at the edge of the null.

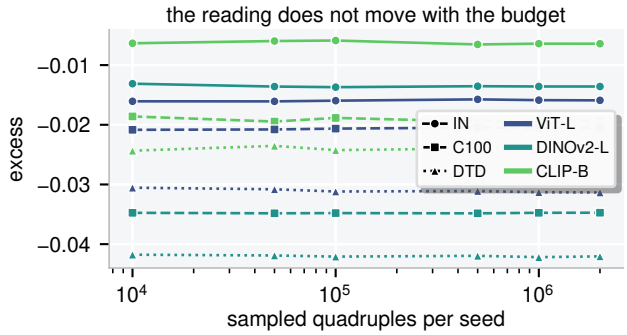

Figure 6: **The reading does not move with the quadruple budget.** Excess against the number of sampled quadruples per seed, one line per cell: 3 backbones on ImageNet, CIFAR-100 and DTD.

model	dataset	images/class	null rep.	excess (reference)	bootstrap mean	bootstrap s.d.	fraction negative
ViT-T	ImageNet	100	20	−0.0055	−0.0051	0.0011	1.00
ViT-S	ImageNet	100	20	−0.0126	−0.0122	0.0007	1.00
ViT-B	ImageNet	100	20	−0.0165	−0.0160	0.0006	1.00
ViT-L	ImageNet	100	20	−0.0160	−0.0160	0.0004	1.00
DINO-B	ImageNet	100	20	−0.0209	−0.0208	0.0003	1.00
DINOv2-S	ImageNet	100	20	−0.0164	−0.0162	0.0004	1.00
DINOv2-B	ImageNet	100	20	−0.0167	−0.0164	0.0002	1.00
DINOv2-L	ImageNet	100	20	−0.0136	−0.0134	0.0001	1.00
DINOv2-G	ImageNet	100	20	−0.0080	−0.0083	0.0004	1.00
CLIP-B	ImageNet	100	20	−0.0068	−0.0069	0.0009	1.00
CLIP-L	ImageNet	100	20	−0.0094	−0.0095	0.0006	1.00
SigLIP-B	ImageNet	100	20	−0.0080	−0.0081	0.0006	1.00
ViT-T	CIFAR-100	500	200	−0.0064	−0.0059	0.0018	1.00
ViT-S	CIFAR-100	500	200	−0.0003	−0.0004	0.0012	0.67
ViT-B	CIFAR-100	500	200	+0.0057	+0.0055	0.0011	0.00
ViT-L	CIFAR-100	500	200	−0.0205	−0.0205	0.0006	1.00
DINO-B	CIFAR-100	500	200	−0.0316	−0.0315	0.0006	1.00
DINOv2-S	CIFAR-100	500	200	−0.0306	−0.0306	0.0004	1.00
DINOv2-B	CIFAR-100	500	200	−0.0313	−0.0313	0.0004	1.00
DINOv2-L	CIFAR-100	500	200	−0.0348	−0.0348	0.0003	1.00
DINOv2-G	CIFAR-100	500	200	−0.0395	−0.0395	0.0002	1.00
CLIP-B	CIFAR-100	500	200	−0.0195	−0.0192	0.0007	1.00
CLIP-L	CIFAR-100	500	200	−0.0181	−0.0183	0.0009	1.00
SigLIP-B	CIFAR-100	500	200	−0.0220	−0.0218	0.0008	1.00
ViT-T	DTD	80	200	−0.0142	−0.0141	0.0016	1.00
ViT-S	DTD	80	200	−0.0186	−0.0189	0.0020	1.00
ViT-B	DTD	80	200	−0.0208	−0.0212	0.0017	1.00
ViT-L	DTD	80	200	−0.0311	−0.0302	0.0017	1.00
DINO-B	DTD	80	200	−0.0272	−0.0267	0.0024	1.00
DINOv2-S	DTD	80	200	−0.0179	−0.0171	0.0025	1.00
DINOv2-B	DTD	80	200	−0.0350	−0.0345	0.0016	1.00
DINOv2-L	DTD	80	200	−0.0420	−0.0417	0.0016	1.00
DINOv2-G	DTD	80	200	−0.0386	−0.0387	0.0026	1.00
CLIP-B	DTD	80	200	−0.0240	−0.0242	0.0021	1.00
CLIP-L	DTD	80	200	−0.0266	−0.0270	0.0022	1.00
SigLIP-B	DTD	80	200	−0.0233	−0.0234	0.0014	1.00

Table 8: **The genuine count survives resampling of the images behind each centroid.** Per cell: images per class, null replicates, the reference excess, the mean and spread over 30 bootstrap resamples, and the fraction of resamples whose excess stays negative.

C.5 Is the structure deeper than a star? The hierarchy test

What this table answers. The census certifies structure beyond the second moments, and this table asks whether the clusters are arranged hierarchically. Each backbone is read against a matched star under both constructions of the star. The controls follow: the models supervised on leaf labels, a backbone trained in hyperbolic space with its Euclidean twin, a fine-tuned pair and a control that removes the radial part of every offset. A pre-set criterion expecting no intact certification of the leaf and frozen models was not met, because they are aligned.

The certified set under other groupings. Only ViT-B and ViT-L are certified under every grouping. A flat cloud clustered under the WordNet grouping fires under the balanced grouping in 38 of 50 runs with clusters rotated. A real cloud read under a grouping that is not its own is therefore expected to fire. The frozen ViT-B firing in 7 of 10 there is consistent with that mismatch and is not evidence of hub structure.

	C100, K=20		ImageNet, matched star				bal. grouping
model	depth	z	K=10 z	K=30 depth	z	K=60 z	real z
ViT-T	-0.025	-1.43	-1.70	-0.018	-1.48	-2.53	-2.08
ViT-S	-0.008	-0.51	-1.72	-0.028	-2.36	-1.78	-1.72
ViT-B	$+0.003$	$+0.23$	-2.57	-0.033	-3.62	-2.23	-2.82
ViT-L	-0.017	-1.49	-2.73	-0.033	-4.04	-4.69	-5.57
DINO-B	-0.038	-4.24	-0.62	-0.009	-0.94	-1.11	-0.98
DINOv2-S	-0.020	-1.61	$+0.26$	$+0.008$	$+0.66$	$+0.61$	$+0.12$
DINOv2-B	$+0.010$	$+0.94$	-0.46	-0.016	-1.19	-0.75	-0.32
DINOv2-L	$+0.013$	$+1.20$	-1.20	-0.028	-2.39	-1.25	-1.70
DINOv2-G	-0.007	-0.68	-1.00	-0.029	-1.78	-0.48	-0.93
CLIP-B	-0.024	-2.00	-1.55	-0.016	-1.95	-2.10	-1.29
CLIP-L	-0.013	-2.34	-1.26	-0.008	-0.83	-1.40	-0.68
SigLIP-B	-0.021	-2.17	-0.76	-0.011	-0.90	-1.66	-1.12

Table 9: **How clusters are oriented relative to their hubs is not random in 4 of 12 backbones under the star with Gaussian hubs.** Per backbone: depth and z on CIFAR-100 and ImageNet, 3 WordNet cuts and the balanced grouping.

model	star e_{hub}	depth	z	r_{star}	z rotated	cert.
ViT-T	$+0.0011$	-0.0184	-1.45	0/10	-1.69 ± 0.28	1/10
ViT-S	$+0.0010$	-0.0257	-2.04	0/10	-1.30 ± 0.19	0/10
ViT-B	$+0.0012$	-0.0347	-3.01	0/10	-1.40 ± 0.25	0/10
ViT-L	$+0.0000$	-0.0337	-3.73	0/10	-1.61 ± 0.19	0/10
DINO-B	$+0.0002$	-0.0092	-1.02	0/10	$+1.75 \pm 0.16$	0/10
DINOv2-S	$+0.0007$	$+0.0083$	$+0.71$	10/10	$+1.38 \pm 0.14$	0/10
DINOv2-B	$+0.0018$	-0.0155	-1.18	0/10	$+0.41 \pm 0.11$	0/10
DINOv2-L	-0.0007	-0.0267	-2.21	0/10	-0.20 ± 0.14	0/10
DINOv2-G	-0.0021	-0.0263	-1.50	0/10	-0.04 ± 0.07	0/10
CLIP-B	$+0.0012$	-0.0171	-1.48	1/10	-1.25 ± 0.09	0/10
CLIP-L	$+0.0018$	-0.0090	-0.74	2/10	-0.88 ± 0.22	0/10
SigLIP-B	$+0.0024$	-0.0121	-0.78	1/10	-0.84 ± 0.18	0/10
flat control, balanced grouping	–	–	-0.46	0/5	-0.65	0/50
flat control, grouping mismatch	–	–	-3.13	5/5	-2.33	38/50

Table 10: **The same 4 backbones are certified under the star whose hubs are a Haar resample of the real ones.** Per backbone: the star's excess, the depth z, the star spread and the same test with clusters rotated, over 10 seeds.

model	IN excess	r	C100 excess	r	depth, Gaussian	z	depth, Haar	z
DeiT-B, IN-1k	-0.0073	200	-0.0257	200	-0.0213	-1.26	-0.0197	-1.21
ViT-B, augreg, IN-1k	-0.0070	199	-0.0117	187	-0.0249	-3.03	-0.0290	-2.65
ViT-B, augreg, IN-21k (census)	-0.0162	200	$+0.0057$	43	-0.0335	-3.62	-0.0347	-3.01

Table 11: **Supervision on leaf labels can produce the alignment but does not guarantee it.** The 2 ViT-B/16 trained on ImageNet-1k leaf labels, the excess and its rank on ImageNet and CIFAR-100, and the depth z under both stars (WordNet agreement: Table 18).

		CIFAR-100 (K=20, n=100: false alarms uncontrolled)				ImageNet (K=30, n=1000: false alarms controlled)			
model	input	exc.	r	$\hat{\delta}_{99.9}$ nat./Eucl.	depth z	exc.	r	$\hat{\delta}_{99.9}$ nat./Eucl.	depth z
MERU ViT-S	images	-0.020	199	0.117/0.117	-1.52	-0.008°	186	0.088/0.088	-1.54
CLIP ViT-S	images	-0.028	200	0.106/0.107	-3.95	-0.006°	171	0.110/0.088	-1.85
MERU ViT-B	images	-0.020	200	0.114/0.114	-2.16	-0.010	196	0.087/0.087	-0.86
CLIP ViT-B	images	-0.030	200	0.105/0.103	-3.28	-0.009°	187	0.107/0.085	-0.99
MERU ViT-L	images	-0.012°	195	0.118/0.118	-0.29	-0.007°	184	0.088/0.088	-0.52
CLIP ViT-L	images	-0.016	199	0.123/0.112	-1.68	-0.010	197	0.104/0.084	-1.21
MERU ViT-S	prompts	-0.022	200	0.086/0.086	-0.67	-0.003°	155	0.085/0.086	$+0.02$
CLIP ViT-S	prompts	-0.023	200	0.085/0.082	-0.78	$+0.006^{\circ}$	15	0.093/0.087	$+0.82$
MERU ViT-B	prompts	-0.012	200	0.096/0.096	-0.70	-0.001°	132	0.089/0.090	$+0.76$
CLIP ViT-B	prompts	-0.017	200	0.094/0.088	$+0.31$	$+0.008^{\circ}$	2	0.100/0.092	$+0.18$
MERU ViT-L	prompts	-0.006°	190	0.102/0.102	$+0.58$	$+0.001^{\circ}$	75	0.093/0.093	$+0.52$
CLIP ViT-L	prompts	-0.012	200	0.098/0.092	$+0.48$	$+0.006^{\circ}$	5	0.101/0.092	$+0.39$

Table 12: **Training in hyperbolic space shows no detected hub structure.** MERU and its Euclidean twin at 3 sizes, on images and on prompts: the excess, the rank r of 200, the reading in each model's own metric and the depth z. $^{\circ}$: not genuine at the corrected threshold.

ViT-B/16	excess	r	z, WordNet-30	z rotated	cert.	z, balanced	z rotated	cert.
frozen (census)	-0.0162	200	-3.62	-1.40	0.0	-2.82	-2.10	0.7
leaf cross-entropy, seed 0	-0.0243	200	-2.45	-0.72	0.0	-2.09	-1.38	0.0
leaf + hierarchical, seed 0	-0.0293	200	-2.97	-2.38	1.0	-3.57	-3.51	1.0
leaf cross-entropy, seed 1	-0.0242	200	-3.18	-2.02	0.6	-2.94	-3.02	1.0
leaf + hierarchical, seed 1	-0.0295	200	-4.22	-2.63	1.0	-5.82	-3.44	1.0

Table 13: **The trained control separates the objectives by degree.** ViT-B/16 fine-tuned with leaf cross-entropy and with an added hierarchical term, 2 seeds: excess, depth z intact and with clusters rotated, on both groupings.

	L2-normalized				radial component removed			
model	z aniso	z Haar-hub	z rotated	cert.	z aniso	z Haar-hub	z rotated	cert.
ViT-T	$+0.27$	-0.17	-0.78	0.0	-2.06	-2.07	-1.74	0.3
ViT-S	-0.76	-0.93	-1.16	0.0	-2.26	-2.08	-1.25	0.0
ViT-B	-2.45	-2.52	-0.57	0.0	-3.60	-3.57	-1.40	0.0
ViT-L	-2.00	-2.23	-1.41	0.1	-3.99	-3.47	-1.47	0.0
DINO-B	$+0.82$	$+0.95$	$+2.00$	0.0	-1.61	-1.89	$+1.91$	0.0
DINOv2-S	$+1.27$	$+1.38$	$+1.28$	0.0	$+0.61$	$+0.58$	$+1.43$	0.0
DINOv2-B	-0.07	-0.01	$+1.28$	0.0	-1.18	-1.19	$+0.46$	0.0
DINOv2-L	-0.46	-0.50	$+0.89$	0.0	-2.59	-2.28	-0.01	0.0
DINOv2-G	-0.78	-0.76	$+0.88$	0.0	-1.75	-1.66	$+0.11$	0.0
CLIP-B	-1.42	-0.93	-1.38	0.0	-1.53	-1.06	-1.19	0.0
CLIP-L	-1.57	-1.12	-0.82	0.0	-1.05	-0.89	-0.92	0.0
SigLIP-B	-1.13	-0.88	-0.51	0.0	-1.12	-1.07	-0.89	0.0

Table 14: **The alignment is not the radial spread of feature norms.** The same clouds after removing the radial part of every offset and after full L2 normalization: the depth z under both stars.

C.6 How much depth can the test see? Power on synthetic and planted hierarchies

What this table answers. A negative verdict is worth reading only where the test can see. Here hierarchies of known strength are planted in real and synthetic clouds and the test is run again, which gives its power; flat clouds are read the same way, which gives its false alarms. The power depends on how quiet the cloud is, so it is reported per backbone.

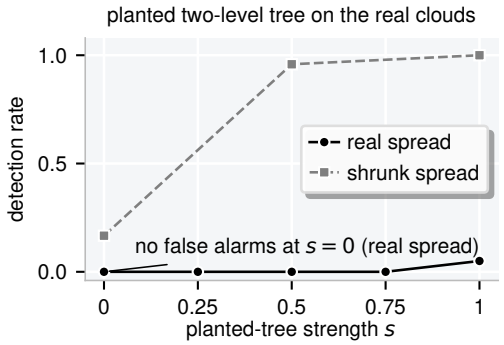

Figure 7: **A planted two-level tree on the real ImageNet clouds is missed at the real within-cluster spread and detected once the spread is shrunk.** Detection rate at $z \leq -2$ against planted-tree strength s with the real within-cluster spread (solid) and shrunk (dashed): at full strength 0.05 against 1.00, and 0.00 at $s = 0$ with the real spread.

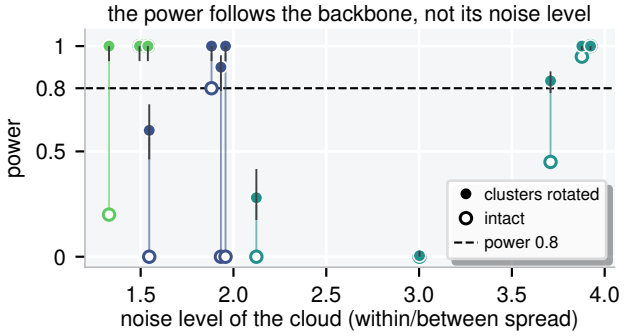

Figure 8: **The power follows the backbone, not the noise level of its cloud.** Power against the within/between spread, intact (hollow) and with clusters rotated (filled); whiskers: 95 per cent intervals over 50 runs per backbone, 200 for the DINOv2 family. ViT-S and DINOv2-B sit closest to the threshold, with 95 per cent intervals reaching 0.79 and 0.78.

cloud	seed	d	excess	r	depth z	z rotated	cert.
three-level hierarchy, ViT-L spectrum	0	1024	-0.0081	199	-1.80	-4.82 ± 0.73	1.0
three-level hierarchy, ViT-L spectrum	1	1024	-0.0095	199	-2.45	-4.44 ± 0.58	1.0
three-level hierarchy, ViT-L spectrum	2	1024	-0.0090	199	-2.38	-4.19 ± 0.60	1.0
three-level hierarchy, ViT-L spectrum	3	1024	-0.0099	199	-2.08	-4.48 ± 0.67	1.0
three-level hierarchy, ViT-L spectrum	4	1024	-0.0090	199	-2.20	-4.28 ± 0.54	1.0
flat control, ViT-L spectrum	0	1024	-0.0198	200	-0.08	-1.61 ± 0.27	0.1
flat control, ViT-L spectrum	1	1024	-0.0202	200	-0.32	-1.61 ± 0.31	0.1
flat control, ViT-L spectrum	2	1024	-0.0196	200	-0.20	-1.76 ± 0.37	0.2
flat control, ViT-L spectrum	3	1024	-0.0197	200	-0.69	-1.74 ± 0.33	0.2
flat control, ViT-L spectrum	4	1024	-0.0192	200	-0.27	-1.68 ± 0.39	0.2
WordNet Poincaré, d=10	0	10	$+0.0830$	0	$+2.36$	$+0.32 \pm 0.58$	0.0
WordNet Poincaré, d=50	0	50	$+0.0354$	0	$+1.53$	-2.22 ± 0.65	0.6

Table 15: **A deep hierarchy at the real noise level is detected, and a flat control is not.** Synthetic clouds with ViT-L's spectrum and ratio: the depth z intact and with clusters rotated, for the hierarchy and for the flat control.

C.7 Is the island real? The tree map and the within-model ceiling

What this table answers. Trees of different models are compared as objects, without labels. The comparison depends on the cut, so the first table gives the diagnostics of every admissible configuration and the criterion that selects one. The ceiling any cross-model agreement can reach is what two resamples of the same model agree on, 0.90 on triplets; Figure 5b carries it as the band behind each model.

dataset	configuration	ARI at the cut	cophenetic corr.	triplet agreement
ImageNet	euclid-average (deg.)	0.03 / 0.64	0.36 / 0.80	0.47 / 0.74
ImageNet	euclid-complete (deg.)	0.03 / 0.44	0.20 / 0.66	0.41 / 0.65
ImageNet	euclid-ward	0.16 / 0.53	0.62 / 0.84	0.67 / 0.85
ImageNet	cosine-average (selected)	0.39 / 0.48	0.48 / 0.78	0.77 / 0.76
ImageNet	cosine-complete	0.13 / 0.49	0.18 / 0.72	0.36 / 0.71
ImageNet	cosine-ward	0.17 / 0.55	0.59 / 0.85	0.69 / 0.86
CIFAR-100	euclid-average (deg.)	0.00 / 0.57	0.15 / 0.78	0.41 / 0.74
CIFAR-100	euclid-complete (deg.)	0.17 / 0.54	0.23 / 0.72	0.41 / 0.66
CIFAR-100	euclid-ward	0.37 / 0.63	0.63 / 0.77	0.70 / 0.79
CIFAR-100	cosine-average (deg.)	0.30 / 0.58	0.42 / 0.72	0.51 / 0.64
CIFAR-100	cosine-complete (selected)	0.39 / 0.63	0.46 / 0.64	0.50 / 0.61
CIFAR-100	cosine-ward	0.56 / 0.61	0.74 / 0.73	0.76 / 0.76

Table 16: **The self-supervised models sit apart on distances, not on topology.** The three agreement measures for DINOv2 against the supervised and contrastive models. Each cell: DINOv2 against the other models / the other models among themselves.

C.8 Is local structure shared? Permutation-calibrated agreement

What this table answers. Two models can place classes in the same neighborhoods and still disagree on distances. This table calibrates both readings against permutations of the class labels, so agreement is counted only where a shuffled pairing would not reach it.

Models share neighborhoods, not metrics. Permutation-calibrated mutual-kNN agreement and centroid CKA survive across all 66 model pairs, and the Poincaré metric improves neither (Gröger et al., 2026) (Table 17).

measure	raw	null mean	$\tau_{0.05}$	calibrated	null-centered	pairs $p < 0.05$
mutual-kNN (k=10), Euclidean	0.472	0.010	0.015	0.464	0.462	66/66
mutual-kNN, Poincaré	0.425	0.010	0.019	0.414	0.415	66/66
linear CKA, Euclidean	0.633	0.108	0.111	0.577	0.524	66/66
linear CKA, Poincaré	0.640	0.108	0.111	0.586	0.532	66/66

Table 17: **Cross-model agreement survives permutation calibration.** Columns: the raw measure, the null mean, the calibration threshold, the calibrated value, the null-centred value and the number of model pairs that survive, for mutual-kNN agreement and for centroid CKA.

C.9 Whose taxonomy? WordNet alignment, superclass recovery, DBpedia and HierarCaps

What this table answers. Here the trees are compared with a human taxonomy instead of with each other. The first two tables use WordNet on ImageNet and the coarse labels of CIFAR-100. The last two repeat the reading on an ontology built independently of WordNet and on a set of caption chains with no class set, so the recovery is not circular.

model	Spearman ρ, inter-centroid vs WordNet distance ρ_{WN}	shuffle	Gaussian	single image	raw $\hat{\delta}$ of WordNet / random groups
ViT-T	+0.52	+0.02	+0.01	–	0.146 / 0.159
ViT-S	+0.50	+0.00	−0.00	–	0.110 / 0.130
ViT-B	+0.49	+0.03	+0.01	–	0.092 / 0.120
ViT-L	+0.53	+0.01	+0.03	+0.25	0.104 / 0.117
DINO-B	+0.36	−0.01	−0.01	–	0.084 / 0.117
DINOv2-S	+0.22	+0.01	+0.01	–	0.077 / 0.100
DINOv2-B	+0.18	−0.01	+0.01	–	0.068 / 0.069
DINOv2-L	+0.18	−0.00	−0.03	+0.08	0.060 / 0.058
DINOv2-G	+0.20	−0.03	−0.01	+0.03	0.054 / 0.056
CLIP-B	+0.57	+0.00	+0.01	–	0.126 / 0.174
CLIP-L	+0.59	+0.01	+0.03	+0.26	0.117 / 0.162
SigLIP-B	+0.58	−0.01	+0.01	–	0.118 / 0.158
DeiT-B (IN-1k)	+0.08	+0.01	–	–	–
ViT-B (augreg, IN-1k)	+0.52	+0.01	–	–	–

Table 18: **Agreement with the human taxonomy follows supervision and recipe.** Per backbone on ImageNet: the correlation between inter-centroid and WordNet distances, the same under shuffled labels, a Gaussian cloud and single images, and the raw reading of WordNet against random groups.

model	supremum, Haar $\hat{\delta}_{\max}$	excess	census $\hat{\delta}_{99.9}$	excess	r/200	p	trip.	NC H−R	FS H−R
BGE-base	0.122	-0.032	0.082	-0.021	200	0.005	0.89	-0.59	+0.05±0.05
E5-base	0.127	-0.028	0.086	-0.017	200	0.005	0.89	-0.43	+0.04±0.03
GTE-base	0.132	-0.024	0.088	-0.017	200	0.005	0.89	-0.82	+0.08±0.06

Table 19: **The recovery is not WordNet circularity.** On DBpedia Classes, an ontology built independently: the reading, the excess, the triplet agreement and the metric gains per text embedder. r/200: replicates above the real value.

model	ρ(level, radius)	% monotone	trip. L1	trip. L2	$\hat{\delta}$ leaves	excess
BGE-base	+0.80±0.31	49.6	0.88	0.97	0.130	+0.003
E5-base	+0.87±0.19	53.9	0.90	0.97	0.139	+0.010
GTE-base	+0.47±0.55	24.0	0.90	0.97	0.135	+0.006
CLIP-B	+0.64±0.43	27.2	0.86	0.95	0.123	-0.011
CLIP-L	+0.51±0.49	18.7	0.86	0.96	0.117	-0.010
SigLIP-B	+0.32±0.55	9.8	0.83	0.91	0.138	+0.014

Table 20: **A caption hierarchy with no class set reads the same way.** On HierarCaps: the correlation between level and radius, the share of monotone chains, the triplet agreements and the leaf excess per encoder.

C.10 Is text the same? The text census under the census reading

What this table answers. The text census reads the thousand ImageNet class names through language models and sentence embedders. The verdicts follow the recipe and the scale of the model, not its family. Extraction is part of the measurement, so the second table varies batch size and precision on the same prompts.

				the reading					cosine	
model	type	d	$\hat{\delta}_{99.9}$	excess	exc./null	r/200	p	k/4	exc.	r
GPT-2 S	causal LM	768	0.077	-0.013	-15%	197	.020	2/4	$+0.000^\circ$	101
GPT-2 M	causal LM	1024	0.052	-0.005°	-9%	170	.154	0/4	-0.019	200
GPT-2 L	causal LM	1280	0.053	-0.023	-30%	200	.005	4/4	-0.016	200
GPT-2 XL	causal LM	1600	0.048	-0.027	-36%	200	.005	4/4	-0.018	200
Pythia-410M	causal LM	1024	0.067	-0.018	-21%	200	.005	4/4	-0.024	200
Pythia-1B	causal LM	2048	0.066	-0.015	-18%	199	.010	4/4	-0.026	200
Pythia-2.8B	causal LM	2560	0.066	-0.017	-20%	200	.005	4/4	-0.023	200
OLMo-1B	causal LM	2048	0.051	-0.024	-32%	200	.005	4/4	-0.032	200
BGE-base	embedder	768	0.078	$+0.003^\circ$	$+4\%$	26	.871	0/4	-0.002°	151
BGE-large	embedder	1024	0.078	$+0.002^\circ$	$+3\%$	49	.756	0/4	-0.003°	170
GTE-base	embedder	768	0.076	$+0.004^\circ$	$+6\%$	11	.945	0/4	$+0.001^\circ$	79
GTE-large	embedder	1024	0.075	$+0.003^\circ$	$+3\%$	36	.821	0/4	-0.001°	131
GTE-Qwen2-1.5B	embedder	1536	0.078	$+0.002^\circ$	$+3\%$	58	.711	0/4	-0.006	199
E5-base	embedder	768	0.074	$+0.004^\circ$	$+6\%$	15	.925	0/4	$+0.000^\circ$	97
E5-large	embedder	1024	0.071	$+0.002^\circ$	$+3\%$	44	.781	0/4	-0.001°	119

Table 21: **In text the verdict follows the recipe and the scale.** Per model: dimension, the reading, the excess, the excess as a fraction of the null, the rank, the left-tail p, k/4 constructions calling it genuine, and the cosine reading beside it. $^\circ$: not genuine at the corrected threshold. r/200: replicates above the real value.

C.11 What follows for practice? The corollary

What this table answers. The instrument certifies structure; this table asks what it buys downstream. Accuracies and zero-cost gains are read per cell, then correlated with the raw reading, with the calibrated one and with the hierarchy verdict. The gains are modest and limited to prototype tasks.

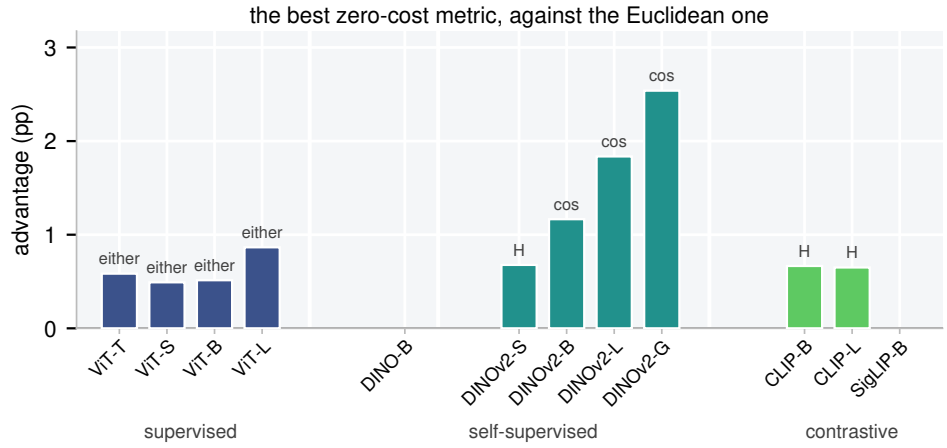

Figure 9: **The zero-cost advantage is small and largest for the self-supervised models.** Best few-shot advantage over the Euclidean metric, averaged over ImageNet, CIFAR-100, CIFAR-10 and DTD, with the metric that collects it; the 10 backbones evaluated, DINO-B and SigLIP-B excluded.

The zero-cost metric. In few-shot learning a fixed-radius Euclidean encoder does at least as well as hyperbolic prototypes (Moreira et al., 2024) (Table 22).

		NC acc. (%)		FS acc. (%)		FS advantage over Euclidean (pp)		
model	data	Eucl.	Poinc.	Eucl.	Poinc.	cosine	radial map	$\text{Poinc.}_{\text{N}} - \cos$
ViT-T	IN	68.1	67.2	97.5	97.6	-0.01	+0.01	+0.08
	C100	29.3	28.2	62.4	63.1	+1.01	+0.45	+0.55
	C10	87.5	87.4	87.6	88.1	+0.66	+0.19	+0.29
	DTD	63.7	62.5	83.3	83.9	-0.23	+0.32	+0.69
	FMNIST	78.5	78.4	81.6	81.9	+0.21	+0.10	+0.31
	MNIST	75.9	74.0	74.3	74.6	+0.64	+0.52	+0.20
ViT-S	IN	77.7	77.5	98.6	98.7	+0.17	-0.02	-0.07
	C100	74.9	74.6	92.2	92.4	+0.59	+0.16	-0.00
	C10	93.4	93.2	92.0	92.5	+0.56	+0.27	+0.19
	DTD	69.3	68.5	85.4	86.0	-0.03	+0.25	+0.73
	FMNIST	80.7	80.4	83.7	83.5	+0.28	+0.08	-0.19
	MNIST	84.8	84.1	80.5	81.4	+0.34	+0.65	+0.84
ViT-B	IN	81.6	81.7	98.9	99.0	+0.20	-0.04	-0.05
	C100	77.3	77.0	92.5	92.7	+0.56	+0.12	-0.05
	C10	94.5	94.0	92.7	93.0	+0.49	+0.18	+0.15
	DTD	70.8	70.8	85.4	86.2	+0.28	+0.19	+0.69
	FMNIST	81.3	81.6	83.4	83.3	+0.03	+0.09	-0.04
	MNIST	81.6	80.5	76.8	77.6	+0.19	+0.51	+0.66
ViT-L	IN	86.1	86.2	99.4	99.4	+0.11	-0.03	-0.02
	C100	85.3	85.2	94.8	95.5	+1.17	+0.31	+0.12
	C10	97.5	97.5	95.0	95.8	+1.25	+0.35	+0.59
	DTD	71.8	71.5	85.6	86.5	+0.13	+0.16	+1.02
	FMNIST	83.7	83.5	84.4	84.7	+0.00	+0.17	+0.23
	MNIST	84.2	83.6	80.8	81.6	+0.08	+0.42	+0.67
DINOv2-S	IN	74.7	74.4	97.4	97.9	+0.68	+0.03	-0.22
	C100	73.4	74.4	94.3	95.1	+0.80	+0.15	+0.06
	C10	95.0	95.2	92.9	93.5	-0.36	+0.13	+1.06
	DTD	73.4	73.4	89.8	90.5	+0.30	+0.04	+0.36
	FMNIST	82.2	82.2	85.4	85.5	+0.02	+0.07	+0.07
	MNIST	83.4	81.9	80.2	80.8	-0.07	+0.45	+0.67
DINOv2-B	IN	78.9	78.9	97.0	97.8	+1.40	+0.04	-0.63
	C100	87.4	87.9	96.5	97.2	+1.11	+0.09	-0.34
	C10	97.1	97.4	94.7	95.4	+0.27	+0.03	+0.43
	DTD	74.2	74.7	89.2	90.3	+1.48	+0.01	-0.35
	FMNIST	83.2	82.8	85.9	85.9	+0.21	+0.04	-0.18
	MNIST	80.6	78.8	76.6	77.1	+0.17	+0.48	+0.47
DINOv2-L	IN	79.8	80.4	96.2	97.1	+2.13	+0.02	-1.24
	C100	85.6	87.7	96.1	97.1	+1.80	+0.09	-0.74
	C10	97.9	98.3	94.1	95.1	+0.92	-0.00	+0.11
	DTD	74.4	75.2	87.1	88.4	+2.40	+0.03	-0.98
	FMNIST	84.4	84.5	87.8	87.9	+0.29	+0.03	-0.03
	MNIST	74.1	72.7	70.5	71.1	+0.25	+0.43	+0.46
DINOv2-G	IN	78.4	78.3	96.2	97.1	+2.57	+0.03	-1.65
	C100	89.8	91.0	95.9	97.2	+2.44	+0.13	-1.10
	C10	97.8	98.2	92.3	93.8	+2.51	-0.02	-1.01
	DTD	72.9	73.1	87.2	88.5	+2.63	+0.01	-1.38
	FMNIST	85.9	85.8	89.0	89.2	+0.38	+0.04	-0.07
	MNIST	79.3	78.6	72.9	73.8	+0.30	+0.56	+0.72
CLIP-B	IN	61.4	59.9	96.8	96.8	-0.04	-0.00	+0.14
	C100	65.6	65.1	87.7	88.9	-0.03	+0.44	+1.21
	C10	87.9	88.3	89.5	90.1	-0.19	+0.20	+0.90
	DTD	63.7	63.2	82.2	83.1	-0.23	+0.36	+1.29
	FMNIST	75.9	75.1	78.8	78.9	+0.06	+0.17	+0.10
	MNIST	70.1	66.6	70.9	69.7	+0.04	+0.23	-1.17
CLIP-L	IN	75.0	74.5	98.4	98.6	+0.01	+0.00	+0.22
	C100	74.6	74.5	90.0	90.8	+0.12	+0.39	+1.23
	C10	95.9	95.9	92.4	93.0	+0.16	+0.25	+0.91
	DTD	68.6	69.0	85.2	86.2	-0.08	+0.36	+1.24
	FMNIST	76.2	75.9	76.8	77.3	+0.12	+0.33	+0.68
	MNIST	85.2	86.1	83.7	84.7	+0.08	+0.58	+1.06

Table 22: **The zero-cost gains are modest and limited to prototype tasks.** The 10 backbones with cached per-dataset features, DINO-B and SigLIP-B excluded. Per cell: nearest-centroid and few-shot accuracy under the Euclidean (Eucl.) and the Poincaré (Poinc.) metric, then the advantage of each zero-cost metric in percentage points over the Euclidean metric.

gain	dataset	raw $\hat{\delta}_{99.9}$ r [CI95]	raw $\hat{\delta}_{99.9}$ dm	excess r [CI95]	excess dm	depth z r [CI95]
NC H−R	ImageNet	−0.84 [-0.96, -0.45]	−0.83	−0.73 [-0.93, -0.19]	−0.61	−0.28 [-0.77, +0.42]
	CIFAR-100	−0.92 [-0.98, -0.68]	−0.58	−0.79 [-0.95, -0.33]	−0.21	+0.52 [-0.16, +0.87]
	CIFAR-10	−0.63 [-0.90, +0.00]	−0.31	−0.67 [-0.91, -0.06]	−0.34	n/a
	DTD	−0.67 [-0.91, -0.07]	−0.67	−0.76 [-0.94, -0.26]	−0.60	n/a
FS H−R	ImageNet	−0.86 [-0.97, -0.50]	−0.56	−0.21 [-0.74, +0.48]	+0.35	+0.27 [-0.43, +0.77]
	CIFAR-100	−0.62 [-0.90, +0.02]	−0.51	−0.79 [-0.95, -0.31]	−0.80	−0.26 [-0.76, +0.45]
	CIFAR-10	−0.78 [-0.95, -0.31]	−0.64	−0.78 [-0.95, -0.30]	−0.66	n/a
	DTD	−0.87 [-0.97, -0.53]	−0.94	−0.98 [-1.00, -0.91]	−0.97	n/a
FS best−R	ImageNet	−0.83 [-0.96, -0.43]	−0.50	−0.08 [-0.68, +0.58]	+0.47	+0.13 [-0.54, +0.70]
	CIFAR-100	−0.77 [-0.94, -0.27]	−0.75	−0.75 [-0.94, -0.23]	−0.65	+0.22 [-0.48, +0.75]
	CIFAR-10	−0.63 [-0.90, -0.01]	−0.55	−0.62 [-0.90, +0.02]	−0.56	n/a
	DTD	−0.81 [-0.95, -0.36]	−0.78	−0.90 [-0.98, -0.63]	−0.84	n/a

Table 23: **The calibrated reading predicts the gain as well as the raw one, and the hierarchy verdict predicts none.** Per dataset and task: the same correlations for the raw statistic, the excess and the hierarchy-test z; dm is the same correlation with both variables centred within family.