
Experimental Record for “From Plato to Aristotle: Toward a General Theory of Representation Hypotheses”

Dai Shi¹ Andi Han² Xiaoyu Li³
Luke Thompson² José Miguel Hernández-Lobato¹

¹University of Cambridge ²University of Sydney ³University of New South Wales

This record holds the full text of Appendix F of the paper before we condensed that appendix. The condensed Appendix F of the paper summarizes each subsection of this record. The text is unchanged, and only the numbering changes from F to R. References to results, sections, figures, and tables of the paper print the paper’s own numbers.

R Experimental Details and Full Results

This appendix gives the protocol, definitions, compute, and preregistration history behind Section 5, and then the full decision tables, figures, and details of every failed preregistered prediction of each experiment in the order of that section. Table R.1 lists every preregistered prediction with its outcome. One script generated every table from the frozen result files and recorded the sha256 of each input. The exceptions are entries in Tables R.1, R.2, R.13, R.14, R.23, and R.24. These entries come from the sentence-encoder follow-up and its two exploratory follow-ups on the depth of the vision features, the layer maximum along OpenLLaMA and Pythia, the second network experiment, and the finite Gibbs model. We transcribed these entries from the frozen summary files and logs of these experiments and checked every value against its source.

R.1 Protocol

Appendix F.1 of the paper summarizes this subsection.

Reporting Rules. A prediction is frozen before scoring if we computed it from single-model features before any cross-model score of the new models. An analysis is exploratory if it lacks a decision rule fixed in advance or if its plan labels it exploratory, and post hoc if we formulated its tested statement after seeing the result. We present both kinds as descriptions of the measured results, not as tests. We report effect sizes, directions, and intervals with the exact p of each family, and we label a joint p over families as exploratory. A clear trend that misses a preregistered threshold is reported with both the outcome of its test and its size.

Models, Data, and Criteria. The language models compared with the 17 vision transformers of ARH are BLOOMZ at five sizes from 0.56B to 7.1B parameters, OpenLLaMA 3B and 7B, and LLaMA-1 7B. In Section 5.4, we add LLaMA-1 13B and 30B, OpenLLaMA 13B, and seven Pythia models from 70M to 6.9B. Features are preprocessed by the clip and ℓ_2 normalization of PRH (Appendix B). Our implementations of the criteria of PRH, ARH, and SOG agree with the published code of these studies to within 10^{-6} on the same float64 features, except for differences caused by two WIT inputs with identical captions (Appendix R.4). Appendix R.10 compares our measurements with the published results. The size trend Tr of a score is the mean over vision models of its Spearman correlation with language-model size, with an exact one-sided p of at least $1/120$ along BLOOMZ, $1/6$ along LLaMA-1 and OpenLLaMA, and $1/5040$ along Pythia. Plain ℓ_2 normalization serves as secondary preprocessing, and we apply the layer maximum of PRH, ARH, and SOG

Experiment, item	Preregistered rule or prediction	Outcome
<i>Pretrained models (fixed; randomness from permutations or input draws)</i>		
COCO, G1	Crossing-law predictor beats two baseline predictors out of sample, after an estimator passes a preliminary check against these predictors on synthetic couplings	Not run , since every candidate estimator failed this preliminary check; predictor reported as unvalidated
COCO, secondary	Slopes ordered seed $\geq$ vision $\geq$ vision-language; elasticities in k and m cancel, without a tolerance	Ordering holds at $k = 1, 10, 100$, mirroring the levels; per-pair sums of the elasticities have class medians 0.03 to 0.06, and the correlation of 0.99 between the elasticity in k and minus the elasticity in m is nearly guaranteed by the range of pairs
WIT, PE2 (Prop. 2)	“SOG-global $\approx$ 1-NN” if forest share of shared MST edges ≥ 0.70 in $\geq 90\%$ of pairs, $ s_T - s_{N,1} \leq 0.03$ in $\geq 80\%$, and bridge trend absent	“SOG-global $\approx$ 1-NN” at mid depth and the last layer; the first two conditions hold nearly automatically, and at the last layer the bridge trend is 0.57 (Holm $p = 0.23$), a power-limited non-rejection
WIT, diagnostics	A difference by a global monotone link refuted if a global monotone fit explains < 0.10 beyond an affine fit; anchors certified by Equation (D.1) expected near 0	Refuted (0.0004, 0.0012); 0 anchors at $k = 10$ in 544 configurations
WIT, PE3	“Width artifact” if equalizing width to 128 components raises the trend of CKA minus its permutation mean by ≥ 0.3 ; “genuine global loss” if by ≤ 0.1	Rule returns “genuine global loss” (change 0.00 and -0.07); the claim in this name is false as worded , since CKA minus its permutation mean keeps its fixed-depth trend (0.70, 0.92), and the test shows that width does not drive this trend
WIT, layer max	Layer-maximum protocol of ARH applied with decision rules fixed in a frozen plan	ARH-style pooled category “ARH pattern absent”; BLOOMZ category “layer max loses the CKA trend”; with LLaMA-1 13B and 30B added, pooled category “partial” and LLaMA-1 category “intermediate” (mKNN trend $0.38 < 0.60$); with OpenLLaMA 13B added, pooled category not decidable (interval covers the boundary of the rule), difference 0.10 $[-0.02, 0.24]$ with “direction not established”, and OpenLLaMA category “layer max loses the CKA trend”
WIT, LLaMA-1, P-L1 to P-L4 (Prop. 4, Cor. 1)	Frozen from features before any score of LLaMA-1 13B or 30B: calibration lowers the layer-max CKA trend by ≥ 0.10 (P-L2); median threshold within 0.01 of the predicted baseline, with the predicted order in ≥ 34 of 51 comparisons (P-L2a); mKNN trend moves by < 0.10 (P-L1); debiased CKA trend drops by < 0.10 and its pooled correlation by ≤ 0.05 (P-L4); fixed-depth CKA keeps its sign (P-L3, marked at risk)	All hold; CKA $0.68 \rightarrow -0.85$; median gap 0.0018, order 51 of 51; mKNN -0.03 ; debiased CKA 0.00, pooled $0.37 \rightarrow 0.36$; fixed depth 0.97 and 1.00; every $p \geq 1/6$
WIT, LLaMA-1, P-L5	Calibration of each layer pair before the maximum keeps the sign of the CKA trend	Reversed ($0.68 \rightarrow -0.88$)
WIT, OpenLLaMA, P-O1 to P-O5 (Prop. 4, Cor. 1)	Frozen from features before any 13B score, all descriptive: features fix boundaries on the raw 13B score that forecast whether 13B ranks below a smaller model under calibration, with 0 wrong among ≥ 10 decided (P-O2b); median threshold offset in $[0.0009, 0.0029]$ and predicted order in ≥ 22 of 33 comparisons (P-O2a); mKNN trend moves by < 0.10 (P-O1); debiased CKA trend drops by < 0.10 and its pooled correlation by ≤ 0.05 (P-O4); calibration before the maximum does not raise the CKA trend (P-O5); fixed-depth CKA trend reported without a prediction (P-O3)	All hold; 30 of 34 decided, all correct; offset 0.0019, order 33 of 33; CKA $0.88 \rightarrow 0.18$ (trend lost); mKNN 0.00; debiased CKA $+0.06$, pooled $0.37 \rightarrow 0.36$; calibrate then max $0.88 \rightarrow 0.21$; fixed depth keeps its sign (P-O3); every $p \geq 1/6$
WIT, Pythia, H1 to H4 and P-P2 (Prop. 4, Cor. 1)	Fixed before any Pythia weight was downloaded: calibration lowers the layer-max CKA trend by ≥ 0.10 consistently across the seven sizes (H1, at level 0.05); if H1 is confirmed, three secondary tests H2 to H4 with a Holm correction; frozen from features before any Pythia score: median threshold offset in the predicted band for ≥ 6 of 7 models and the predicted order in ≥ 277 of 307 comparisons (P-P2a); 0 wrong among ≥ 30 decided forecasts of the order after calibration (P-P2b)	H1 not confirmed : CKA $0.93 \rightarrow 0.89$, drop 0.034, $p_{\text{perm}} = 1612/5040$; H2 to H4 not tested, since H1 was not confirmed; mKNN and debiased CKA move by 0.00; P-P2a holds, 7 of 7 in band, order 307 of 307; P-P2b holds, 352 of 357 decided, all correct

Table R.1: Preregistered predictions and their outcomes, grouped by the experiments of Section 5, with failed items in bold. The outcome of G1 is in bold because every candidate estimator failed its preliminary check. The LLaMA-1, OpenLLaMA, and Pythia rows concern language models added after the first preregistration (LLaMA-1 13B and 30B, OpenLLaMA 13B, and seven Pythia models), with predictions frozen before scoring (Appendix R.1). Two continuations list the sentence-encoder follow-up, the linear autoencoder, and the first network experiment, and then the second network experiment, the finite Gibbs model, and the contrastive networks.

Experiment, item	Preregistered rule or prediction	Outcome
<i>Sentence-encoder follow-up on the COCO evaluation sets (pretrained models; randomness from input draws)</i>		
COCO, sentence encoders, P-S1 (Conj. D.1)	Elasticities in k and m on $k \in \{10, 100\}$ sum to a median within ± 0.03 and to within ± 0.05 in $\geq 80\%$ of the pairs (≥ 8) whose elasticity in m has size ≥ 0.1 ; $k \geq 10$ and tolerances set after the primary pairs	Holds; 15 of 15 pairs qualify and lie within ± 0.05 , median 0.006; holds in every ordering, in all 11 further arms, and for the primary pairs under centering
COCO, sentence encoders, P-S2b, P-S2c	(P-S2b) The slope at $k = 10$ follows the line fitted on the primary pairs against the log level, not the language-model slope of its vision model; (P-S2c) within a pair, averaging one to five captions moves the slope with the log level at the rate of this line	Undecided, both labeled uninformative; predicted slopes differ by a median 0.092, below the required 0.113 (P-S2b); five captions raise the level by 19% over the first, below the required 35% (P-S2c)
<i>Controlled training (randomness from initialization, data, order, and minibatches)</i>		
Linear AE, P1	Gradient descent: neighbor sets, G_s , CKA unreliable at every N , with mean at $N = 2^{16}$ inside the 95% interval at $N = \infty$	Falsified ; unreliable in 12 of 12 cells, plateau test fails for G_s and CKA (paired differences about 10^{-4})
Linear AE, P2, P3	Agreement on neighbor sets, G_s , and CKA reliable at $N = 2^{16}$ under weight decay, tied weights, and small initialization; agreement under the projector criterion reliable for every training procedure	Confirmed
Network, P1	Without weight decay, u unreliable on $T_{N,10}$ and G_s at both N	Failed ; $T_{N,10}$ 96 of 96 at $N = 250k$, and the result meets the preregistered falsification condition; G_s unreliable at 65 of 96 under a tolerance set at 100k (same-size test in row “Second network, A3”)
Network, P2, P4	Without weight decay, y and Δ_{II} reliable at 250k; trunk $s_N(10)$ equal across arms within 0.05	Passed (largest trunk difference 0.011)
Network, P3	Weight decay and small initialization make u reliable at 250k	Held under weight decay; failed under small initialization (G_s 82 of 96)
Network, P5	Fraction of agreeing pairs under small initialization within 0.1 of that under weight decay	Failed (0.146 on G_s at 250k)

Table R.1: Preregistered predictions and their outcomes (continued), for the sentence-encoder follow-up of Appendix R.5, the linear autoencoder of Appendix R.11, and the first network experiment of Appendix R.12, with failed items in bold. The first network experiment sets its tolerances from pilot runs at 100k, whereas the second network experiment sets them at each sample size. At $N = 25k$ the first experiment therefore finds 0 of 96 pairs within tolerance for u and y on every criterion under each of its three training procedures with independent training subsets. In the rows on training, an entry such as “ u reliable on $T_{N,10}$ ” states that agreement of disjoint pairs of runs under this criterion is reliable (Appendix R.1).

under a preregistered plan. The fixed depths are relative depth 0.5 (mid depth) and 1 (the last layer). In the learning experiments, we evaluate criteria ordered by Proposition 1, from the neighbor-set criterion $T_{N,10}$ through the order criteria T_O and T_E and the spanning-tree criterion T_{MST} to the scale-normalized criteria G_s and $1 - \text{CKA}$ on $T_{G/\mathbb{R}_+}$. Appendix R.2 defines each discrepancy.

Decisions in Controlled Training. In a learning experiment we train disjoint pairs of runs, count the pairs whose discrepancy is at most a frozen tolerance ε , and decide with a one-sided 95% Clopper–Pearson bound. Agreement is reliable if the lower bound exceeds 0.9, unreliable if the upper bound is below 0.8, and inconclusive otherwise, which instantiates Definitions 3 and C.1 with $\delta = \eta = 0.1$. With 96 pairs, these rules return “reliable” with at least 92 pairs within tolerance and “unreliable” with at most 69 such pairs. With 64 pairs these counts are 62 and 45, and with 32 pairs they are 32 and 21.

What Is Random. The pretrained models are fixed, and a score on them is random through the draw of evaluation inputs (Lemma D.1) and through the $n_\pi = 500$ permutations of its calibration. The comparisons on WIT use one draw of 1,024 inputs, and our intervals resample the 17 vision models, so they cover the choice of vision models and the permutations but not the draw of inputs. In the experiment of Section 5.2, we measure the variation over input draws directly through five random nested orderings of the COCO pool (Appendix R.5). In the linear autoencoder the randomness is the initialization, the training sample at finite N , and the minibatches, with common random numbers across training procedures. In the finite Gibbs model the randomness is the training sample of each run, drawn as exact multinomial counts from its own hashed seed, and each learner returns a certified minimizer of its empirical risk. In the networks of Sections 5.5 and 5.6 the randomness is the initialization, the training subset, its order, and the augmentation, each drawn from its own hashed seed. The two Gram matrices of each disjoint pair of runs therefore form a draw from $\mu_{r_A} \otimes \mu_{r_B}$, with $r_A = r_B$ for two runs of one training procedure and objective, and draws of different pairs are

Experiment, item	Preregistered rule or prediction	Outcome
<i>Second network experiment (fresh runs; tolerances from pilot pairs of the same arm and size)</i>		
Second network, C3, C4a (Rem. 1)	Momentum from $C_0 = 0$: median relative imbalance above 0.05 (lower 97.5% bound) and at least 0.02 in $\geq 90\%$ of runs; momentum from the default initialization: median change of C above its initial norm, and median cosine of the final and the initial C below 0.5 (upper bound)	All hold at both N ; 0.28 and 0.33, 64 of 64 runs; 9.3 and 8.8 times; -0.017 and -0.005 (upper bounds 0.011, 0.012)
Second network, I3	From a balanced or small initialization, momentum adds a larger imbalance part than head flow on G_s and $1 - \text{CKA}$ at both N ; 8 tests, Holm at 0.05	8 of 8 rejected; raw differences 0.0012 to 0.0111
Second network, I4 (part 2 of Thm. 2)	Head flow from the default initialization, a decision between two outcomes without a predicted direction: (i) the mechanism of part 2 is realized if the imbalance part on G_s and its run-to-run part exceed 0 at both N ; (ii) it is not detectable if the imbalance part lies within ± 0.5 SD on G_s and $1 - \text{CKA}$ at both N ; simultaneous 98.75% intervals	Outcome (i), 0.54 and 0.59 SD with run-to-run parts 0.29 and 0.46 SD; condition (ii) does not hold, upper bounds 0.59 to 0.74 SD
Second network, A1 (part 3 of Thm. 2)	Weight decay: u reliable on $T_{N,10}$ and G_s at both N	Holds; 96 of 96 in all four cells, also under the pooled tolerance
Second network, A3	Momentum without weight decay: u and its common-gauge counterfactual unreliable on G_s at both N	Tolerance-sensitive, reported without a categorical claim; u 0 and 5 of 96, counterfactual 1 and 21 of 96; counterfactual 95 of 96 at 250k under the pooled tolerance, 7.8% larger
Second network, A7	Balanced or small initialization under head flow: (a) u reliable on G_s at both N and not tolerance-sensitive; (b) balanced part above 0.5 SD at both N (lower 95% bound)	(a) fails in both arms (u 20 and 35 of 96 balanced, 8 and 93 of 96 small); (b) holds, 2.7 to 3.9 SD; the preregistered sentence on weight decay is not added, since the balanced part under weight decay lies outside ± 0.5 SD (-1.23 SD at 25k)
Second network, A2	Where y is reliable, u not unreliable on $T_{N,10}$ and the 95% interval of its mean excess within ± 0.002	Fails in 4 of 10 cells, all at 250k, with u agreeing better than y by 0.0017 to 0.0023; u reliable on $T_{N,10}$ in every cell
Second network, A4, A8, A9	Momentum without weight decay: common-gauge share of the G_s excess in $[0.5, 0.9]$ at 250k with the intervals of the mismatch, balanced, and imbalance parts excluding 0; G_s excess at 250k over that at 25k in $[0.5, 1.5]$; $\Delta_E(u) < \Delta_E(y)$ in $\geq 90\%$ of pairs	All hold; 0.71; 0.83; 96 of 96 at both N
Second network, K4, K5, K6 (Cor. 2)	At $k = 10$, for the outputs of one run under the gauges of both runs: median bound over observed slot change in $[5, 25]$ per cell, certified fraction ≤ 0.10 ; momentum without weight decay: the coordinates of the balanced head in place of u move ≥ 0.01 of the slots; cross-run excess over the function alone at most 0.0005	All hold; 10.9 to 12.7; at most 0.047; 0.052 and 0.070; at most 0.0002
<i>Finite Gibbs model (certified minimizers inside the premises of Theorem 3)</i>		
Gibbs, F0 (Thm. 3, parts 1 to 4 of the full version, and one prediction of the model)	Nine confirmatory tests with a Holm correction: convergence of both learners from 2^{34} to 2^{40} ; exact agreement across objectives on the global order of eight probes reliable at 2^{40} ; agreement across objectives on mean-normalized distances unreliable at 2^{34} and 2^{40} , $1 - \text{CKA}$ at its limit, and the difference after the link is undone converging; agreement under the isotonic discrepancy reliable; mean discrepancy across objectives within the sum of the two mean discrepancies within objectives on order criteria and above it on geometric ones	All 9 confirmed, p from 7.9×10^{-68} to 5×10^{-5} ; order 96 of 96, geometry 0 of 96, isotonic 96 of 96; details in Table R.24
Gibbs, secondary	Monte Carlo predictions of the ratio of the mean discrepancy across objectives to the mean of the two mean discrepancies within objectives, at 2^{40} and within a factor of 1.5 (P5b); the Gaussian limit and the transition of Thm. E.1 (P1b, P2b)	P5b failed for neighbor sets and spanning trees; P1b and P2b pass
<i>CIFAR-5m contrastive networks (outside the premises of Theorem 3)</i>		
Contrastive, link	Link confirmed if items (i) to (v) hold; falsified if the warp share is < 0.2 or an order gap > 0.1	Falsified ; warp share 0.147, order gaps 1.0; (i), (iii), (iv), (v) fail, (ii) passes by its letter
Contrastive, secondary	Fitted link exponent in $[2.5, 10]$; co-occurrence prediction derived from Lemma E.1	Both failed

Table R.1: Preregistered predictions and their outcomes (continued), for the second network experiment of Appendix R.12.1, the finite Gibbs model of Appendix R.13, and the contrastive networks of Appendix R.14, with failed items in bold. SD is the standard deviation of the output discrepancy over the main pairs of a cell in the second network experiment, whose checks and estimates appear in Appendix R.12.1.

independent. The exception is one arm of the first network experiment, whose runs share training corpora (Appendix R.12). An arm is the set of runs or measurements under one setting of an experiment, such as one training procedure and step size or one objective and temperature. The networks are evaluated on a fixed set of 2048 held-out images.

R.2 Criteria and Discrepancies

Appendix F.1 of the paper summarizes this subsection.

All criteria are computed in float64 on squared distances D of the preprocessed features, with ties broken by input index.

- $T_{N,k}$: the discrepancy $1 - s_N$ at k neighbors, with chance level $k/(m - 1)$ for s_N .
- T_O : the fraction of discordant pairs between the two rankings of the other inputs by distance to an anchor, with ties counted as one half, averaged over anchors.
- T_E : $(1 - \tau_b)/2$ for Kendall’s τ_b between the off-diagonal entries of D_A and D_B .
- T_{MST} : $1 - s_{\mathcal{T}}$, with the forest and bridge split of Proposition 2 reported for the forest F_A .
- $\text{SOG}_{0.1}$: one minus the length-weighted 10-nearest-neighbor overlap of Equation (B.1) at $\tau = 0.1$.
- $T_{G/\text{Aff}}$: one minus the Pearson correlation of the off-diagonal entries of D_A and D_B .
- G_s on $T_{G/\mathbb{R}_+}$: the root mean square difference of the off-diagonal entries of D_A and D_B after division of each matrix by its mean off-diagonal entry.
- $1 - \text{CKA}$: one minus linear (biased) CKA.
- Debiased CKA (pretrained models): linear CKA with the unbiased estimator of the Hilbert-Schmidt independence criterion in place of the biased one, the estimator that ARH calls unbiased CKA, matched to its code.
- Δ_{iso} (contrastive networks): the cross-fitted root mean square residual of the best global nondecreasing map from one distance matrix to the other, in units of the target’s mean distance and averaged over both directions. The warp share is the fraction of the affine residual removed by this map.
- Δ_{Π} : the projector distance $\|\Pi_A - \Pi_B\|_F^2 / (\text{rank } \Pi_A + \text{rank } \Pi_B)$ between the orthogonal projectors Π_a onto the column spaces of the two centered feature matrices, which equals one minus the mean squared canonical correlation when the ranks are equal (part 4 of Theorem 2).

R.3 Hardware, Software, and Compute

Appendix F.1 of the paper summarizes this subsection.

Every experiment ran on one machine with an NVIDIA RTX PRO 6000 Blackwell Workstation Edition GPU (97,887 MiB), an Intel Core Ultra 9 285K CPU with 24 cores, and 125 GiB of memory, under Ubuntu 24.04.3. The training code used PyTorch 2.13.0 with CUDA 13.2 and bf16 autocast, and the metric code used float64 NumPy. GPU kernels are seed-deterministic but not bitwise reproducible. Feature extraction for the pretrained models took 38.7 minutes on the GPU, and the preregistered computation of the layer maximum took 21 minutes. Its extension to LLaMA-1 took 5.3 minutes of forward passes for the two new models and 14 minutes for two independent implementations of the maximum. The OpenLLaMA experiment took 61 seconds of forward passes for 13B and 9.7 minutes for its two implementations. The Pythia experiment took 15.0 GPU-minutes in total, 2.0 of them for the forward passes of the seven models. The finite Gibbs experiment took 12.71 GPU-hours, 2.71 of them for computing the Gaussian limits of both learners before the first sample and the rest for the solves of its InfoNCE learner. Its spectral learner ran on the CPU (Appendix R.13). The 1,888 network training runs took 40.3 hours of summed run time with up to four runs sharing the GPU, and 96 of these runs were excluded pilots. The second network experiment trained 4,352 further runs, 1,984 of them pilots, in 20.92 hours of summed run time, and its excluded engineering runs took 0.43 hours. The linear autoencoder took about 1.5 CPU-hours, and the neighbor computations over evaluation sizes took 159 CPU-hours. The sentence-encoder follow-up took

0.38 GPU-hours to embed the captions and 1.5 to 1.9 CPU-hours for its neighbor computations and checks. The two exploratory follow-ups on the depth of the vision features took about 1.5 CPU-hours, and the first also took 0.04 GPU-hours to extract new features of DINOv2-B. Several jobs shared the machine, so wall-clock times include contention. Table R.2 lists the compute of each job.

Job	Device	Compute
Feature extraction, 66 model and dataset jobs	GPU	38.7 min
Known-answer checks on synthetic couplings	CPU	54.0 min wall, 10 processes
Estimator selection for the evaluation-size test	CPU	31.1 + 2.7 + about 1.5 CPU-h
Pretrained models at fixed depth	CPU	3.93 CPU-h ($m = 1024$), 0.38 ($m = 2047$), 1.84 (layer-rule sensitivity)
Preregistered layer maximum	GPU	1,284 s, peak 6.9 GB; independent implementation 384 + 425 s
Evaluation-size test, 1,879 units	CPU	158.7 CPU-h
Sentence-encoder extraction, 3 encoders	GPU	22.3 min run and 0.7 min smoke test (0.38 GPU-h); embedding loops 194, 165, and 968 s; peak 2.4 GB
Sentence-encoder follow-up, caches, scores, and checks	CPU	1.5 to 1.9 CPU-h on at most 6 cores
Depth follow-ups, DINOv2-B extraction (exploratory)	GPU	2.3 min run and about 0.3 min smoke test (0.04 GPU-h)
Depth follow-ups, caches, scores, and checks (exploratory)	CPU	about 1.5 CPU-h on at most 6 cores
Linear autoencoder	CPU	about 1.5 CPU-h
First network experiment, 48 pilot + 1,344 runs	GPU	0.17 + 4.74 h summed run time; 2.34 h wall with 4 workers
Contrastive networks, 48 pilot + 448 runs	GPU	5.05 + 30.38 h summed run time; 1.57 + 7.52 h wall
Second network experiment, 1,984 pilot + 2,368 runs	GPU	9.91 + 11.01 h summed run time, 0.43 h engineering; 6 h 38 min wall for the whole pipeline, including the tolerance freeze and analysis
LLaMA-1 13B and 30B extraction	GPU	forward 53 + 263 s, load 37 + 27 s, peak 52.5 and 68.7 GB
LLaMA-1 layer maximum, two implementations	GPU	321 + 195 s and 170 + 169 s; check against the code of ARH 4,598 s
OpenLLaMA 3B and 7B re-extraction, 13B extraction	GPU	37 + 53 s; forward 61 s, load 33 s, peak 52.6 GB
OpenLLaMA layer maximum, two implementations	GPU	244 s and 155 + 181 s; check against the code of ARH on a reduced grid 111 s
Pythia extraction, 7 models	GPU	2.0 min forward passes (6.7 to 38.8 s per model)
Pythia layer maximum, two implementations	GPU	2.41 + 2.46 min and 2.25 + 2.38 min (primary and secondary preprocessing); 15.0 GPU-min for the whole Pythia experiment with checks
Finite Gibbs model	GPU and CPU	12.71 GPU-h (Gaussian limits before the first sample 2.71; pilots and gates 1.37 and main runs 8.63 for the InfoNCE learner); spectral learner on the CPU

Table R.2: Compute per job on the single machine of Appendix R.3.

R.4 Preregistration, Gates, and Deviations

Appendix F.2 of the paper summarizes this subsection.

Documents. The Stage 1 preregistration (pretrained models and the linear autoencoder) was frozen before any experiment ran, and an amendment (v1.1) was frozen before the COCO features were opened. Separate analysis plans were frozen before each computation. The plan of the fixed-depth comparison preceded this comparison, and the plan of the layer maximum preceded any real-data layer-max score. The plan of the evaluation-size test followed the estimator selection and preceded the opening of the COCO features, and the plan of the linear autoencoder preceded its

confirmatory runs. The Stage 2 preregistration (first network experiment and contrastive networks) was frozen before any main run, and its one amendment changed the time criterion of a pilot gate from per-run wall time to throughput, before that gate was evaluated. The preregistration of the second network experiment was frozen together with all of its code before its engineering, pilot, and main runs (Appendix R.12.1). The preregistration of the finite Gibbs experiment was frozen together with all of its code before any of its computations, and each of its four amendments preceded every computation affected by this amendment (Appendix R.13). For the LLaMA-1 and OpenLLaMA families, the analysis plans were frozen before the download of any weights of LLaMA-1 13B and 30B or OpenLLaMA 13B. The OpenLLaMA analysis code was frozen before the first byte of the 13B weights reached the machine. For the Pythia family, we fixed an analysis plan with one primary and three secondary confirmatory tests before the download of any Pythia weight, and we froze its predictions from features before any Pythia score. The predictions computed from features alone were frozen before any cross-model score of the new models was computed, an order confirmed by the creation times of every output file. For the sentence-encoder follow-up, the analysis plan and then its analysis and extraction code were frozen before the download of any sentence-encoder weight, and its one amendment preceded the first cross-model score (Appendix R.5). Every hash is recorded in the event log.

Gates. Before any real data were analyzed, the metric code passed four of its six known-answer checks on synthetic couplings, and the amendments made in response to the two failures preceded the opening of the COCO features.

- The first of the two failed known-answer checks used a coupling built by an exponential link. On this coupling, one cross-fitted isotonic fit explained 0.9978 of the variance against a required 0.999, because one evaluation pair fell below the fitted range and received a constant extrapolation. The discrepancy under every neighbor, component, and spanning-tree criterion was exactly 0 in all 27 cells. The second failed check used Gaussian couplings, on which the first coupling estimator missed the observed s_N at $k = 1$ by up to 0.057 against a tolerance of 0.02. This error comes from the first-order approximation behind the estimator. The amendment based the isotonic check on the in-sample fit and specified a rebuild of the exponential coupling to remove float64 noise on its diagonal. This rebuild was not executed, so the pass rests on the existing run, in which both order discrepancies are 0 and the in-sample isotonic fraction is 1.000000. For $k \leq 3$ the amendment replaced the approximation by the exact expected discrepancy.
- Agreement with the reference implementations of PRH and ARH passed, with a largest difference of 1.67×10^{-8} against a tolerance of 10^{-6} .
- In a separate check, we ran the published code of PRH, ARH, and SOG on our float64 features of three pairs of models, at mid depth and at the last layer. The pairs join two language models (LLaMA-1 7B, OpenLLaMA 7B), two vision models (OpenCLIP ViT-B/16, DINOv2-B), and a language model with a vision model (BLOOMZ-1.7B, DINOv2-L). Under the preprocessing of PRH, the code of SOG returned the same minimum spanning trees and MST overlaps as our implementation on all three pairs. Its length-weighted overlaps lay within 10^{-8} of ours on the two pairs with a vision model. Every difference above 10^{-6} came from two WIT inputs whose identical captions give them identical features in every language model. The code of SOG computes the distance between these two inputs as a rounding error near 10^{-8} instead of 0, which changes the length-weighted overlaps of the pair of language models by up to 9.8×10^{-4} . Over both preprocessings and five values of k , the mKNN code shared by ARH and SOG breaks the tie between these inputs differently from ours in one of 60 values, a difference of 3.3×10^{-5} .
- The first coupling estimator failed a real-data sanity check on held-out WIT items, predicting s_N near zero on a vision pair and a language pair where 0.13 to 0.48 was observed. Under the amendment, we re-selected among at most four pre-listed estimators, three of them local to each anchor. Every candidate failed the preliminary check of the out-of-sample test G1 against two baseline predictors on synthetic couplings, so G1 was not run.
- The pilot gates of both Stage 2 experiments and of the second network experiment passed.

Deviations.

1. The power gate of the spanning-tree test required a rerun on an evaluation set of $m = 3072$ inputs, which could not be built from the extracted inputs, so we used the largest available disjoint evaluation set, with $m = 2047$ inputs. The outcome is the same on $m = 1024$ under the same correction.
2. In the linear autoencoder the preregistered step 0.05 changes the imbalance C by a median 53%, so arms with step 0.002 were added as secondary arms before the confirmatory runs. The preregistration left the learning rates of stochastic gradient descent and Adam open, and we fixed these rates on excluded pilots.
3. In the contrastive experiment, training with the unconstrained spectral loss diverged at every tested learning rate. The spectral arm therefore uses the ball projection of [HaoChen et al. \(2021\)](#), which was active on every evaluated input, so this arm trains on the unit sphere.
4. The main run of the layer maximum started before the check against the code of ARH finished, and its outputs were opened after this check passed.
5. In the first network experiment the analysis code was hashed after the tolerances were frozen, and an interim analysis on two runs per cell preceded the freeze. An aborted interim analysis left cached per-pair results, which the final analysis reused after an independent recomputation matched all 252 decision cells. The code computes the projector criterion on u , where the preregistration names y , and the projector distances computed on u and on y agree to 10^{-7} .
6. The fixed-depth comparison ran before the recording of the amendment responding to the two failed known-answer checks, although the preregistration required these checks to pass before any real-data analysis. This comparison uses neither failed check, and the cross-fitted isotonic value carries a flag at each occurrence.
7. A script of the evaluation-size test was edited after its outputs existed, without an entry in the event log, and an independent script reproduces every number of the analysis plan. A verification run printed the frozen predictions of the first coupling estimator for three WIT pairs, which are not an endpoint of any preregistered analysis.
8. The LLaMA-1 experiment adds LLaMA-1 13B and 30B to the preregistered comparisons under its own plan and predictions. The automated order check of the predictions inspected a directory left unused by the two implementations, so the order rests on the creation times of the output files, all later than the freeze of the predictions. For about 77 minutes, concurrent validation jobs of the two implementations exceeded the cap of six CPU threads set by the plan; every computation is deterministic, and the LLaMA-1 7B results of one implementation reproduce the frozen run bitwise.
9. In the OpenLLaMA experiment, we saw the joint statistic of the three families in a log of one implementation before the two implementations passed their agreement check. This statistic is exploratory, and its inputs, code, and rule were frozen beforehand. Read-only check scripts were added after the code freeze, with every hash in the event log. The plan stated that the fixed-depth baseline could not change across sizes by enough to enter a prediction. This statement holds under the preprocessing of PRH and fails at the last layer under plain ℓ_2 , and under both preprocessings we report the fixed-depth CKA trend (P-O3) without a prediction.
10. In the Pythia experiment, the two implementations differ in 163 pair-level and 3 descriptive floating-point values, by at most 9.1×10^{-14} , and agree on every integer, p -value, decision, and category, so the preregistered agreement check failed by its letter. After seeing the category, we accepted this check with an added absolute tolerance of 10^{-12} , a post hoc ruling. A third implementation reproduces the results of both from their pair files with 0 mismatches. A smoke test of the GPU was rewritten after the code freeze and recorded with its hash before its first run.
11. The finite Gibbs experiment ran under four hashed amendments, two of which fixed programming errors in the frozen analysis code; their order and content are given in [Appendix R.13](#).
12. The sentence-encoder follow-up ran under one hashed amendment, which fixed a programming error in the file-writing step of its frozen evaluation code ([Appendix R.5](#)). The plan of this follow-up describes the pool caption as drawn at random by the dataset author, whereas

the pool holds the last caption of every image, and every computation used this stored caption. The review recomputed the labels ahead of the independent reimplementations, in the reverse of the order listed in the plan, and we wrote the summary of the follow-up after the agreement of both.

Design Errors in the Stage 2 Preregistration. We record seven errors in our own Stage 2 design, each of which leaves every number and decision above unchanged. Prediction P1 of the first network experiment extended the obstruction to the neighbor-set criterion, a case not covered by part 2 of Theorem 2. Prediction P3 stated that agreement of u under small initialization would be reliable, although the same document expected the balance premise of this initialization to fail under stochastic gradient descent. Both experiments set the tolerance at a pilot size between the two tested sizes, and with a fixed budget the smaller size also meant more epochs. At the smaller size of the first network experiment, 0 of 96 pairs therefore lie within tolerance for u and y on every criterion, under each of its three training procedures with independent training subsets. The second network experiment sets each tolerance from pilot pairs of the same arm and size (Appendix R.12.1). A decision on u at the tolerance of y depends on both the gauge shared by all runs and the run-to-run mismatch, without separating these two parts. In the contrastive experiment, item (v) assumed that a change of InfoNCE temperature is a pure rescaling. Remark 2 gives this rescaling for a unique positive definite minimizer over all unit-diagonal positive semidefinite matrices. The networks restrict the Gram matrix to rank at most 128. Even without this restriction, the empirical risk has no positive definite minimizer at any temperature, since two augmented views of one image are each other’s only positive partners and other images supply the negatives. The falsification threshold of 0.2 on the warp share rules out the sharp link of the theorem but not a general monotone link. The sharp link has $\beta = 5$, the inverse of the InfoNCE temperature 0.2, and an expected warp share of about 0.75, whereas even an exact link with $\beta = 1$ gives a warp share near 0.12. The general conclusion therefore rests on the order criteria and on Δ_{iso} . Finally, the fitted link exponent $\hat{\beta}$ equals minus the logarithm of the mean cosine of the spectral arm to within 0.008, so $\hat{\beta}$ is fixed by the spectral embedding independently of the InfoNCE temperature.

R.5 Evaluation Size

Appendix F.3 of the paper summarizes this subsection.

This subsection gives the protocol and the full results of the two preregistered experiments of Section 5.2 on growing COCO evaluation sets, and of two exploratory follow-ups on the depth of the vision features. The evaluation sets are prefixes of one pool of 16,384 COCO validation images, each paired with the caption stored by coco-30-val-2014, i.e., its last caption in the COCO annotation file. Five hashed orderings of the pool define nested prefixes of sizes $m \in \{256, 512, \dots, 16,384\}$. In each ordering, the anchors are the images of one hashed half of the first 1024 images (515 in the first ordering), and the neighbors of each anchor are searched in the whole prefix. The slope of a pair at fixed k is the ordinary least-squares slope of $\ln s_N$ on $\ln m$ over the seven sizes, averaged over the five orderings. A slope is flat if its absolute value is below 0.1, and falling or rising otherwise, according to its sign. At $k = 10$ the standard deviation of the s_N of a pair over the five orderings is at most 0.031. The vision models are DINOv2-B, OpenCLIP ViT-B/16 (LAION-2B), MAE-B, AugReg ViT-B/16 (ImageNet-21k), and SigLIP-B/16, and the language models are PolyPythia-410M seeds 1 to 9, BLOOMZ-1.7B, and OpenLLaMA-3B, all at mid depth. The 29 primary pairs are 4 disjoint pairs of PolyPythia-410M seeds, the 10 pairs of vision models, and the 15 pairs of a vision model with BLOOMZ-1.7B, OpenLLaMA-3B, or seed 1 of PolyPythia-410M. The secondary set holds all 36 pairs of the nine seeds. Tables R.3 to R.8 give the per-class results, and Figure R.1 compares the observed class medians and slopes with the crossing-law predictions.

Slope Classes. All 4 seed pairs have flat slopes at $k = 1$ and $k = 10$, as do all 36 seed pairs of the secondary set. All 10 vision pairs have falling slopes at $k = 10$ and $k = 100$, although after correction for chance the slope stays falling for 7 of 10 at $k = 100$ (exploratory). All 15 vision–language pairs have falling slopes at every k . At $m = 16,384$ and $k = 10$, the s_N of every vision–language pair still exceeds its chance level by a factor of at least 8.75. The preregistered secondary ordering of slopes, seed above vision above vision–language, holds at all three values of k . Since the levels of s_N have the same order, this ordering repeats the ordering of the levels rather than adding a second observation. The rising class of Conjecture D.1 appears in 0 of the 87

combinations of a primary pair and a value of k . At these levels of s_N , the preregistered slope rule fails to detect this class. It classifies as flat all 9 units of a synthetic conformal coupling used in the selection of the coupling estimator, with slopes from -0.016 to 0.057 . The absence of rising slopes therefore leaves the conformal case of the conjecture untested. In a post hoc analysis of conformal couplings at $k = 10$, the slope of $\ln(1 - s_N)$ on $\ln m$ rises from -0.46 at dimension 2 to -0.07 at dimension 8. For the seed pairs this slope lies between 0.003 and 0.047 , so their $1 - s_N$ stays level as m grows. At the intrinsic dimension of the seed models, estimated at 17 to 19 with the maximum-likelihood estimator of [Levina and Bickel \(2004\)](#), a conformal coupling would give a slope near -0.05 . The steepest seed pair reaches -0.026 at $k = 1$, so the separation between the seed pairs and the conformal case is moderate.

Stability at Fixed k/m . At a shared k/m , the class medians for $k = 100$ stay within 4% of those for $k = 10$ (Figure 2b). The class medians for $k = 1$ lie 13 to 16% below those for $k = 10$ for seed pairs and 17 to 19% below for vision pairs, and they stay within 11% for vision-language pairs. As BPC does in one of its neighbor criteria, [Schnaus et al. \(2026\)](#) hold k/m fixed by setting $k = \max(1, \lfloor m/100 \rfloor)$ in their comparison of alignment across granularities. The elasticities in k and m of a pair are the two coefficients of a least-squares fit of $\ln s_N$ on $\ln k$ and $\ln m$. Their per-pair sums have class medians of 0.03, 0.06, and 0.03 for seed, vision, and vision-language pairs, against -0.02 , -0.19 , and -0.56 for the elasticity in m alone (Table R.5). The preregistered secondary prediction that the elasticities in k and m cancel was stated without a tolerance, so we report the per-pair sums of Table R.5 as its outcome. Across the 29 pairs, the elasticity in k and minus the elasticity in m have correlation 0.99. This value is nearly guaranteed by the spread of s_N across the pairs, since identical representations give zero for both elasticities and pairs at chance level give one and about minus one. For vision-language pairs the stability at large k/m is partly trivial, since at $k = 100$ and $m = 256$ their class median sits 1.10 times above chance, against a median of 13.9 times at $k = 10$ and $m = 16,384$.

The Crossing-law Estimate. The preregistered out-of-sample test G1 required a coupling estimator that passed a real-data sanity check and a preliminary check on synthetic couplings against two baseline predictors, the nulls N0 and N1. Every candidate estimator failed this preliminary check. In three of its four comparisons, the null generates the synthetic coupling exactly or approximately. The failure means that each estimator lost to a null on the coupling generated by this null, not that the crossing law is wrong. G1 was therefore not run, and the predictor is reported as unvalidated, following the preregistered outcome rule. In a descriptive comparison, the predictor recovers the slope class of all 29 pairs at $k = 10$, against 4 for the no-change null N0 and 25 for the affine null N1. Its slope signs are right in 69 of 69 non-flat combinations, as are those of N1 (Figure R.1d at $k = 10$). Its slope magnitudes are wrong on vision-language pairs, where it predicts a median slope of -0.24 against -0.57 observed. On these pairs its median level error is a factor of 2.86, larger than the 2.26 of N0 (Table R.6). At $k = 10$ the class median slopes are -0.01 , -0.21 , and -0.57 observed, -0.02 , -0.19 , and -0.24 under the crossing law, and -0.54 , -0.80 , and -0.88 under N1, for seed, vision, and vision-language pairs (Figure R.1a–c). The crossing law predicts too shallow a slope for all 15 vision-language pairs, by 0.29 to 0.40. N1 predicts too steep a slope for all 29 pairs at $k = 10$, but at $k = 1$ for 5 of the 15 vision-language pairs. N1 is undefined in 2 of the 145 combinations of a pair and an ordering and is replaced there by N0, following the analysis plan. The slopes are measured over a 64-fold range of m , not in the limit of large m . At $k = 1$ the class median of the vision pairs levels off at 0.082 to 0.088 for $m \geq 2048$, so the falling class of these pairs at larger k may be pre-asymptotic.

Sentence Encoders. We call the second preregistered experiment of Section 5.2 the sentence-encoder follow-up. This follow-up pairs the five vision models with all-mpnet-base-v2 (MPNet), Contriever, and Qwen3-Embedding-0.6B at pinned revisions, and each encoder embeds a caption with the pooling of its model card. Its main arm keeps the evaluation sets, orderings, anchors, caption, and clip of the 29 primary pairs, with a sentence encoder in place of the language model of each vision-language pair. The language models embed this caption alone, so the comparisons between sentence encoders and language models use the arms with this caption. Eleven further arms vary the caption or the normalization, and six of them average the first one to five captions of each image by annotation identifier or all of its captions. Two further arms combine this averaging with centering. The remaining three arms center the normalized features in place of the clip, use plain ℓ_2 normalization, and drop duplicate captions, respectively.

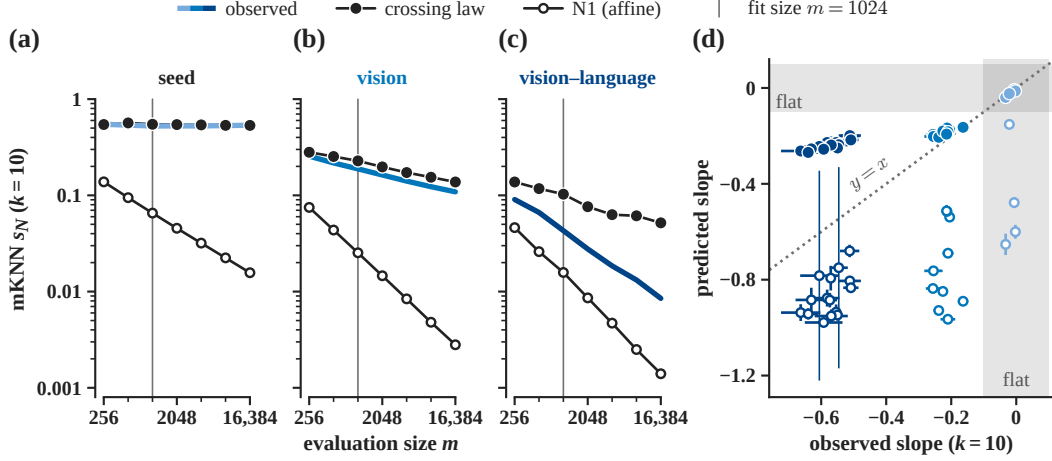


Figure R.1: **Descriptively, at $k = 10$ the crossing law recovers the slope class of every pair but predicts too shallow vision-language slopes, whereas the affine null N1 predicts every slope too steep (Theorem 1).** (a–c) Class medians of mKNN at $k = 10$ on the growing COCO evaluation sets, observed and as predicted from the measurements at the fit size $m = 1024$. (d) Predicted against observed slope of $\ln s_N$ on $\ln m$ per pair, ± 1 s.d. over the five orderings of the images. Predicted and observed slope classes coincide where the grey bands cross or in the white lower left. The two long thin bars come from orderings where N1 is undefined. Tables R.3 to R.8 give the values.

Levels and Slopes with Sentence Encoders. In the main arm, s_N falls with m for all 15 pairs at $k = 1, 10$, and 100, as the thin lines of Figure R.2a show for $k = 10$. At $k = 10$ their class median falls from 0.105 at $m = 256$ to 0.013 at $m = 16,384$. At $k = 10$, $m = 1024$, and mid depth, the median s_N of the sentence-encoder pairs is 0.059, between the 0.043 of vision-language pairs with language models and the 0.189 of vision pairs. At $k = 10$ and $m = 1024$, the s_N of each vision model with every sentence encoder exceeds its s_N with any language model. For SigLIP-B/16, s_N ranges from 0.116 to 0.137 with the sentence encoders, against 0.074 to 0.098 with the language models (Table R.7). As for the primary pairs, at every k the four pair classes have the same order by median slope as by median level, with the sentence-encoder pairs third (descriptive).

Stability with Sentence Encoders. The preregistered prediction of stability at a shared k/m (P-S1 in Table R.1) required at least 8 pairs whose elasticity in m has absolute value at least 0.1. Among these pairs, the median sum of the elasticities in k and m on $k \in \{10, 100\}$ had to lie within ± 0.03 , with at least 80% of the sums within ± 0.05 . We fixed the range $k \geq 10$ and these tolerances after scoring the primary pairs and before any sentence-encoder score. The range leaves out $k = 1$, where the s_N of the primary pairs departs from a function of k/m and where the bound of Theorem 1 is loosest relative to κ_i . The tolerances come from the sums of the primary pairs on $k = 10$ and $k = 100$, so the sums of the primary vision-language pairs lie inside the band of Figure R.2b by construction, with a median of 0.004. All 15 new pairs qualify, with elasticities in m from -0.60 to -0.39 , and their sums lie between -0.009 and 0.016 around a median of 0.006 (Figure R.2b). P-S1 holds in the main arm, in each of its five orderings, in all 11 further arms, and for the primary pairs under centering. With $k = 1$ included, the median sum of the new pairs rises to 0.058, against 0.030 for the primary vision-language pairs, and 5 of the 15 new sums stay within ± 0.05 , against 12 of 15 (exploratory). The cells of the new pairs at $k = 100$ sit near chance at $m = 256$, as for the primary vision-language pairs. On $m \geq 2048$, where these cells lie at least 2.25 times above chance, the elasticities give a median sum of 0.005 (exploratory). The training data of MPNet include COCO caption pairs, probably from the 2014 training images outside our validation pool (unverified), and the pairs of MPNet have the largest median sum, 0.012. The sums of the ten pairs of Contriever and Qwen3-Embedding-0.6B keep a median of 0.006 and all lie within ± 0.05 , so this prediction also holds without the pairs of MPNet. For $k = 10$ and $k = 100$, this result extends the interpretation of the evaluation size as a scale parameter to sentence encoders as the text side.

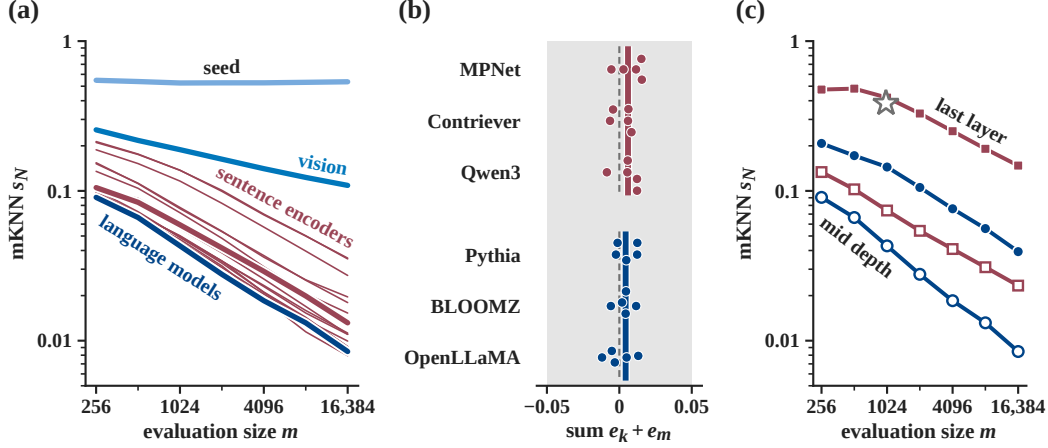


Figure R.2: **Neighbor overlap between vision models and sentence encoders falls with the evaluation size but stays close to stable at a fixed k/m (Conjecture D.1), and in exploratory follow-ups vision features from the last layer in place of mid depth raise this overlap.** (a) Class medians at $k = 10$ as in Figure 2a, with sentence-encoder pairs as thin lines. (b) Sums of the elasticities in k and m on $k \in \{10, 100\}$, with group medians as bars and the preregistered tolerance in grey. (c) Medians at $k = 10$ with vision features from mid depth (open) or from the last layer (filled, exploratory), for sentence encoders with centering and five averaged captions (squares) and for language models (circles). The star marks the value of Schnaus et al. (2026) for DINOv2-B and MPNet at equal k/m .

Undecided Slope Predictions. Two further preregistered predictions concern the slope at $k = 10$, the first between pair classes and the second within a pair (P-S2b and P-S2c in Table R.1). On the primary pairs, we fitted a line of the slope against the logarithm of the level of s_N . The prediction between classes stated that the slope of each new pair follows this line, rather than the slope of the same vision model with the language models. The prediction within a pair stated that, when one to five captions are averaged, the slope of a pair moves with the logarithm of its level at the rate of this line. Its alternative was the smaller rate seen within the classes of the primary pairs. The frozen plan counted each prediction as decidable if its two alternatives lay far enough apart, a condition called the guard. For the prediction between classes, the guard required a median gap of at least 0.113 between the two predicted slopes. For the prediction within a pair, it required a median rise of the level by at least 35% from the first caption to the average of the first five. The gap reached 0.092 and the rise 19%, so both predictions carry the frozen label “uninformative”. For P-S2b, the frozen operating characteristics gave this label in 84 to 88% of simulated runs at a median level near 0.06, whereas its power was tabulated at median levels of 0.12 to 0.30. The plan fixes two values to report whatever the label. The weight of the level line in a least-squares mixture of the two predictions of P-S2b is 0.44, and it stays between 0.39 and 0.56 in leave-one-out fits by encoder and by vision model. The median rate of P-S2c is 0.435, against 0.228 for the line and 0.080 within the classes of the primary pairs, and the unmet guards leave both values without an interpretation. Of the 12 secondary cells of P-S2b, 9 are uninformative and the arm with plain ℓ_2 normalization gives “inconclusive”. The two arms that average five or all captions under the clip pass the guard by less than 0.01 and give “holds” with weights of 0.96. In both arms this weight results from averaging slopes flatter than the line for DINOv2-B, OpenCLIP ViT-B/16, and MAE-B with steeper slopes for SigLIP-B/16. Applied descriptively at the other values of k , the rule of P-S2c would give “holds” at $k = 1$, where averaging raises the level further, and “uninformative” at $k = 100$. Since both predictions concern $k = 10$, these descriptive cells leave both labels unchanged.

Absolute Level. Schnaus et al. (2026) center each side and average the captions of an image before they compare two models. At $k = 10$ and $m = 1024$, centering in place of the clip raises the level of the new pairs by a median 9%, and the average of five captions in place of the pool caption raises it by 17%. With both changes, the s_N of DINOv2-B and MPNet reaches 0.077 at $k = 100$ and $m = 10,000$, against about 0.38 read from the figure of Schnaus et al. (2026) for the same two models on 10,000 COCO images. We interpolate every value at $m = 10,000$ between our

two largest evaluation sets, of 8,192 and 16,384 images. In two exploratory follow-ups, each with decision rules fixed before its first score, vision features from the last layer in place of mid depth close the gap between 0.077 and about 0.38. When we take the CLS token of DINOv2-B from its last layer instead of mid depth and keep every other setting, its s_N with MPNet rises to 0.429. With the pooling of Schnaus et al. (2026), which averages all output tokens at 224 pixels, s_N is 0.410, 8% above their value. At $k = 100$ and $m = 10,000$, vision features from the last layer raise the s_N of all 15 pairs of a vision model and a sentence encoder by a median factor of 4.3 (range 2.3 to 7.4). At the last layer, the median s_N with the three sentence encoders lies between 0.40 and 0.45 for OpenCLIP ViT-B/16, AugReg ViT-B/16, DINOv2-B, and SigLIP-B/16, against 0.15 for MAE-B.

Depth of the Vision Features. With language models as the text side, last-layer vision features raise the s_N of all 15 pairs at $k = 10$ and $m = 1024$ by a median factor of 2.8, from a median of 0.043 to 0.144. Across layers of the language model, this level changes by at most a factor of 1.4. At the last layer, s_N still falls with m at $k = 10$ for all 30 pairs, with median slopes of -0.42 with language models and -0.30 with sentence encoders, against -0.57 and -0.43 at mid depth (Figure R.2c). This panel shows the medians of these 30 pairs at $k = 10$ next to their medians at mid depth. In this panel, s_N is computed with centering and five averaged captions for the sentence-encoder pairs and with the protocol of Section 5.2 for the language-model pairs. The elasticities of these 30 pairs in k and m on $k \in \{10, 100\}$ sum to at most 0.02 in absolute value. In a post hoc comparison at a shared k/m , the values for $k = 100$ lie within 9% of those for $k = 10$. Neighbor overlap thus stays close to stable at a fixed k/m also at the higher levels reached with last-layer features. Since the values at a shared k/m are this close, we draw the value measured by Schnaus et al. (2026) at $k = 100$ on 10,000 images as the star of Figure R.2c. The star sits at the same k/m on the axis for $k = 10$. In a post hoc check, the line fitted on the primary pairs and its counterpart for centered features predict too shallow slopes at the last layer. The first gives a median of -0.31 against -0.42 observed with language models, and the second -0.05 against -0.30 observed with sentence encoders. These follow-ups rest on COCO, five ViT-B models, and language models of at most 3B parameters, and we did not score the vision pairs at the last layer.

Order of the Sentence-encoder Follow-up. Before the download of any sentence-encoder weight, we froze the analysis plan, with its three confirmatory predictions and their guards, and then the analysis and extraction code. We froze this code once it had reproduced the scores of the primary pairs bitwise. The three encoders then embedded the pool captions and the 81,981 official captions of the pool images, a step on each encoder alone that precedes every cross-model score. During the embedding run, we logged one hashed amendment to the file-writing step of the frozen evaluation, which would have stopped on four tables written by an earlier step. The amended step reuses such a table when its content is identical and keeps every rule and statistic of the plan. Before the first score of a new pair, a checking script verified the order and hashes of these steps and found the output folders free of files of the new pairs. The creation times of all 87 such files follow this check. A review recomputed every label with its own code. An independent reimplementation then rebuilt the neighbor sets from the stored embeddings and features. It reproduced the overlap counts in all 9,450 cells of the main arm and of the five arms with the first one to five captions, and all 41 labels. For each of the two exploratory follow-ups on the depth of the vision features, we fixed the decision rules in a hashed file before computing any of its scores. Independent reimplementations reproduced all overlap counts and labels of both follow-ups. We wrote the rules of the second follow-up after the first follow-up had produced the scores of the three last-layer pairs of DINOv2-B with the sentence encoders. These rules include the three pairs, and the s_N of the other 12 sentence-encoder pairs also rises at the last layer.

R.6 Pretrained Models at Fixed Depth

Appendix F.4 of the paper summarizes this subsection.

This subsection gives the details of the fixed-depth comparisons of Section 5.3 on WIT and of a preregistered test of whether model width drives the CKA trend. Table R.9 gives the spanning-tree test on the evaluation sets of $m = 1024$ and $m = 2047$ inputs, and Table R.10 the width test, with the rules of Table R.1. Unless stated otherwise, the trend Tr in this subsection is computed on s_{obs} minus its permutation mean, not on s_{cal} of Section 2.1. Its p -values are exact over the 120 relabelings of the BLOOMZ sizes and Holm-corrected across the four primary tests where stated.

Pair class	pairs	$k = 1$		$k = 10$		$k = 100$	
		r/f/d	median	r/f/d	median	r/f/d	median
seed	4	0/4/0	0.009	0/4/0	-0.014	0/3/1	-0.055
vision	10	0/7/3	-0.076	0/0/10	-0.214	0/0/10	-0.278
vision-language	15	0/0/15	-0.522	0/0/15	-0.571	0/0/15	-0.598
seed (all)	36	0/36/0	0.012	0/36/0	-0.018	0/28/8	-0.061

Table R.3: Fixed- k slope classes per pair class: the numbers of pairs with rising (r), flat (f), and falling (d, for decreasing) slopes and the median slope of $\ln s_N$ on $\ln m$. The last row covers all 36 seed pairs of the secondary set.

Pair class		$m = 256$	$m = 512$	$m = 1024$	$m = 2048$	$m = 4096$	$m = 8192$	$m = 16384$
$k = 1$	seed	0.445	0.458	0.462	0.474	0.478	0.487	0.500
	vision	0.127	0.112	0.098	0.083	0.082	0.085	0.088
	vision-language	0.026	0.017	0.010	0.0075	0.0047	0.0040	0.0036
	chance $k/(m-1)$	0.0039	0.0020	0.0010	0.0005	0.0002	0.0001	0.0001
$k = 10$	seed	0.548	0.537	0.525	0.527	0.527	0.531	0.535
	vision	0.256	0.217	0.189	0.163	0.141	0.123	0.109
	vision-language	0.091	0.066	0.043	0.028	0.018	0.013	0.0085
	chance $k/(m-1)$	0.0392	0.0196	0.0098	0.0049	0.0024	0.0012	0.0006
$k = 100$	seed	0.674	0.602	0.566	0.552	0.542	0.538	0.539
	vision	0.566	0.418	0.332	0.273	0.236	0.205	0.175
	vision-language	0.432	0.258	0.161	0.107	0.073	0.049	0.033
	chance $k/(m-1)$	0.3922	0.1957	0.0978	0.0489	0.0244	0.0122	0.0061

Table R.4: Class medians of s_N over the primary pairs (4 seed, 10 vision, 15 vision-language) by k and m , with the s_N of each pair averaged over the 5 orderings, and the chance level $k/(m-1)$.

Pair class	pairs	elasticity in k	elasticity in m	sum	fit RMSE
seed	4	0.047	-0.019	0.027	0.050
vision	10	0.250	-0.194	0.061	0.122
vision-language	15	0.583	-0.565	0.030	0.131

Table R.5: Class medians of the two elasticities, their sum, and the RMSE of a per-pair least-squares fit of $\ln s_N$ on $\ln k$ and $\ln m$, whose coefficients are the elasticities in k and m . The sum measures how far s_N departs from a function of k/m alone. Across the 29 primary pairs, the elasticity in k and minus the elasticity in m have correlation 0.99, a value nearly guaranteed because the pairs range from the high s_N of seed pairs to s_N near chance.

Predictor	seed		vision		vision-language	
	level error	sign	level error	sign	level error	sign
N0 (no change)	0.036 ($\times 1.04$)	0/1	0.213 ($\times 1.24$)	0/23	0.816 ($\times 2.26$)	7/45
N1 (affine fit with noise)	2.156 ($\times 8.64$)	1/1	2.094 ($\times 8.12$)	23/23	1.388 ($\times 4.01$)	45/45
crossing law (unvalidated)	0.020 ($\times 1.02$)	1/1	0.150 ($\times 1.16$)	23/23	1.050 ($\times 2.86$)	45/45

Table R.6: Descriptive accuracy of the unvalidated crossing-law predictor and of the nulls N0 (the value of the pair at $m = 1024$, predicted for every m) and N1 (an affine fit with Gaussian noise, replaced by N0 when the slope of the fit is not positive). Level error is the median $|\ln(\text{predicted}/\text{observed})|$ over cells with $m \neq 1024$ and observed s_N above twice chance, with the corresponding factor. Sign gives the number of correctly predicted slope signs among the combinations of a pair and a value of k with a non-flat observed slope. These accuracies are descriptive, since the preregistered out-of-sample test G1 was not run.

Pair	class	$s_N(10)$ at $m = 256 / 1024 / 16384$	slope at $k = 1 / 10 / 100$	class
pythia410m_s1 $\times$ pythia410m_s2	seed	0.561 / 0.537 / 0.544	0.028 / -0.006 / -0.046	0/0/0
pythia410m_s3 $\times$ pythia410m_s4	seed	0.401 / 0.360 / 0.348	-0.008 / -0.033 / -0.130	0/0/-
pythia410m_s5 $\times$ pythia410m_s6	seed	0.534 / 0.514 / 0.527	0.025 / -0.003 / -0.050	0/0/0
pythia410m_s7 $\times$ pythia410m_s8	seed	0.720 / 0.692 / 0.663	-0.006 / -0.021 / -0.059	0/0/0
augreg_b $\times$ siglip_b	vision	0.368 / 0.263 / 0.156	-0.123 / -0.205 / -0.238	-/-/-
clip_b $\times$ augreg_b	vision	0.368 / 0.266 / 0.156	-0.067 / -0.213 / -0.238	0/-/-
clip_b $\times$ mae_b	vision	0.252 / 0.160 / 0.085	-0.075 / -0.256 / -0.325	0/-/-
clip_b $\times$ siglip_b	vision	0.229 / 0.146 / 0.079	0.023 / -0.255 / -0.342	0/-/-
dinov2_b $\times$ augreg_b	vision	0.259 / 0.210 / 0.133	-0.091 / -0.164 / -0.224	0/-/-
dinov2_b $\times$ clip_b	vision	0.218 / 0.160 / 0.092	-0.012 / -0.210 / -0.278	0/-/-
dinov2_b $\times$ mae_b	vision	0.191 / 0.129 / 0.069	-0.039 / -0.238 / -0.341	0/-/-
dinov2_b $\times$ siglip_b	vision	0.234 / 0.167 / 0.092	-0.078 / -0.226 / -0.287	0/-/-
mae_b $\times$ augreg_b	vision	0.303 / 0.214 / 0.125	-0.106 / -0.210 / -0.267	-/-/-
mae_b $\times$ siglip_b	vision	0.324 / 0.228 / 0.132	-0.120 / -0.214 / -0.278	-/-/-
augreg_b $\times$ bloomz_1b7	vision-language	0.131 / 0.060 / 0.012	-0.468 / -0.571 / -0.539	-/-/-
augreg_b $\times$ openllama_3b	vision-language	0.116 / 0.052 / 0.011	-0.539 / -0.582 / -0.568	-/-/-
augreg_b $\times$ pythia410m_s1	vision-language	0.106 / 0.048 / 0.010	-0.478 / -0.573 / -0.589	-/-/-
clip_b $\times$ bloomz_1b7	vision-language	0.087 / 0.039 / 0.008	-0.494 / -0.554 / -0.627	-/-/-
clip_b $\times$ openllama_3b	vision-language	0.078 / 0.035 / 0.008	-0.538 / -0.546 / -0.652	-/-/-
clip_b $\times$ pythia410m_s1	vision-language	0.076 / 0.033 / 0.007	-0.496 / -0.550 / -0.668	-/-/-
dinov2_b $\times$ bloomz_1b7	vision-language	0.100 / 0.050 / 0.009	-0.592 / -0.569 / -0.567	-/-/-
dinov2_b $\times$ openllama_3b	vision-language	0.091 / 0.043 / 0.007	-0.723 / -0.606 / -0.598	-/-/-
dinov2_b $\times$ pythia410m_s1	vision-language	0.089 / 0.041 / 0.008	-0.579 / -0.592 / -0.614	-/-/-
mae_b $\times$ bloomz_1b7	vision-language	0.089 / 0.039 / 0.006	-0.405 / -0.630 / -0.631	-/-/-
mae_b $\times$ openllama_3b	vision-language	0.081 / 0.036 / 0.005	-0.516 / -0.663 / -0.661	-/-/-
mae_b $\times$ pythia410m_s1	vision-language	0.077 / 0.033 / 0.005	-0.458 / -0.640 / -0.673	-/-/-
siglip_b $\times$ bloomz_1b7	vision-language	0.179 / 0.098 / 0.021	-0.534 / -0.513 / -0.428	-/-/-
siglip_b $\times$ openllama_3b	vision-language	0.158 / 0.083 / 0.019	-0.577 / -0.512 / -0.459	-/-/-
siglip_b $\times$ pythia410m_s1	vision-language	0.144 / 0.074 / 0.018	-0.522 / -0.509 / -0.484	-/-/-

Table R.7: Neighbor overlap $s_N(10)$ at three evaluation sizes, slopes at three values of k , and slope classes of the 29 primary pairs. Model codes: pythia410m_sj is PolyPythia-410M seed j , augreg_b, clip_b, dinov2_b, mae_b, and siglip_b are AugReg ViT-B/16, OpenCLIP ViT-B/16, DINOv2-B, MAE-B, and SigLIP-B/16, and bloomz_1b7 and openllama_3b are BLOOMZ-1.7B and OpenLLaMA-3B. The last column gives the slope class at $k = 1, 10$, and 100 , flat (0), falling (-), or rising (+).

Configuration	units	median $ \Delta\text{slope} $	max $ \Delta\text{slope} $	class changes
ℓ_2 preprocessing	87	0.010	0.169	2 (dinov2_b__augreg_b@k=1: flat→falling; mae_b__augreg_b@k=1: falling→flat)
clip threshold per layer	87	0.008	0.080	2 (augreg_b__siglip_b@k=1: falling→flat; mae_b__augreg_b@k=1: falling→flat)
duplicates removed	87	0.001	0.017	0

Table R.8: Sensitivity of the slopes to three changes of configuration, over the 87 combinations of a primary pair and a value of k , with the resulting changes of slope class.

Spanning Trees. On the evaluation set of $m = 2047$ inputs, the forest share of the shared MST edges exceeds the threshold of 0.70 in every pair. With the other two conditions also met, the preregistered rule returns “SOG-global $\approx$ 1-NN” at both depths, i.e., it classifies MST overlap as mostly one-nearest-neighbor overlap (Figure R.3). The forest share has medians 0.875 and 0.889, against 0.850 and 0.852 expected if the shared edges were a random subset of the language-model tree, and its minimum over pairs is 0.76 and 0.78. MST overlap exceeds one-nearest-neighbor overlap on this evaluation set by a median factor of 1.20 and 1.32 at the two depths, in 129 and 136 of the 136 pairs (Figure 2c). The bound of Proposition 2 holds in all 816 configurations, with $s_{\mathcal{T}}$ between 1.58 and 4.15 times the bound. On the evaluation set of $m = 2047$ inputs, chance is $1/(m-1)$ for $s_{N,1}$ and about 0.001 for $s_{\mathcal{T}}$, the cross in Figure R.3a. Bridge overlap exceeds its permutation null with medians 0.016 and 0.035 against 0.001, and the bridge trend within BLOOMZ is $\text{Tr} = -0.16$ at mid depth and 0.57 at the last layer ($p = 0.058$, Holm-adjusted 0.23). The grey band of Figure R.3c reaches the median over pairs of the 95% permutation quantile of bridge overlap minus its null mean. Two of the three conditions of the rule hold nearly automatically. The forest already makes up a median 84 to 85% of each language-model tree across configurations, and at least 82% in every pair. This fraction is the forest share expected among the shared edges under

random relabeling, so a random subset of tree edges would pass the threshold of 0.70. With s_T and $s_{N,1}$ between 0.02 and 0.05, their difference stays within 0.03 almost by construction. The data add a modest enrichment over the forest fraction, with forest edges shared at 1.25 and 1.40 times the bridge rate (exploratory). At the last layer, the rule counts the bridge trend as absent through a failure to reject. The point estimate 0.57 exceeds the preregistered bar of 0.5, but the Holm correction across the four primary tests requires $p \leq 0.0125$, and the exact test over five sizes gives $p = 0.058$. The half term $m s_{N,1}/2$ attains the maximum of the bound in every configuration, so these pretrained pairs leave the branch $m s_{N,1} - n_A^{\text{mut}}$ of the bound untested. On the evaluation set of $m = 2047$ inputs, the mean number of shared language-model bridges per pair is 5.32 at mid depth, below the power target of 10, and 10.72 at the last layer. At the last layer, the outcome of the rule also depends on the forest. On this evaluation set, the rule returns “mixed” when the forest of the vision model replaces the primary forest of the language model.

Warps and Margins. We split the squared-distance discrepancy into the shares explained by four increasingly flexible fits, a scale fit, an affine fit, a global monotone fit, and one monotone fit per row. Each share counts what its fit adds beyond the previous fits. The shares are in-sample, since each fit is estimated and scored on the same $m = 1024$ inputs. Their medians are 0.681, 0.190, 0.0004, and 0.029 at mid depth and 0.683, 0.148, 0.0012, and 0.034 at the last layer, with residual shares 0.102 and 0.123. After the affine fit, the cumulative share reaches a median 0.87 at mid depth and 0.84 at the last layer. The bound of Equation (D.1) certifies 0 anchors at $k = 10$ in all 544 configurations, and every variant of the bound gives the same count. This count follows from the neighbor gap, which is a median 0.0017 and 0.0026 times twice the largest discrepancy in its row. At $k = 1$ the largest certified fraction in any pair under any variant is 1.07%. On these pretrained pairs, neighbor overlap therefore depends on the typical crossings of Theorem 1 rather than on the worst-case margin.

Preprocessing and Families. With the clip of PRH and after subtraction of the permutation mean, mKNN and CKA rise with size at the last layer along BLOOMZ and LLaMA-1, whereas the scores with OpenLLaMA 7B exceed those with OpenLLaMA 3B for 12% of the vision models on mKNN and 29% on CKA (exploratory). The direction of a step in size can therefore differ between families at one layer, and plain ℓ_2 normalization reverses several BLOOMZ trends (Table R.10). At mid depth and after subtraction of the permutation mean, the BLOOMZ trends with the clip are 0.70 for CKA (exact one-sided $p = 0.117$) and 0.84 for mKNN ($p = 0.025$). Under plain ℓ_2 they are -1.00 (two-sided $p = 0.017$) and -0.92 ($p = 0.033$), and CKA falls strictly with size for 17 of the 17 vision models (Figure 3b,c). In Figure 3b,c each line is divided by its value at 0.56B. Before this division, the means over vision models range from 0.113 to 0.130 for CKA and from 0.034 to 0.041 for mKNN with the clip. Under plain ℓ_2 normalization the ranges are 0.007 to 0.014 and 0.018 to 0.024. With 13B added, the OpenLLaMA trends of s_{cal} at the last layer are 0.59 for mKNN and 0.56 for CKA (Table R.13), so this difference between families rests on the step from 3B to 7B. Each trend is measured over five sizes of BLOOMZ or three of LLaMA-1 or OpenLLaMA, with an exact p of at least $1/120$ along BLOOMZ and $1/6$ along the other two families, without a correction for multiplicity. On these models, a size trend depends on the model family, the layer, and the preprocessing together. Schnaus et al. (2026) report a similar dependence on preprocessing for one encoder applied to MR and CT images. Under the identity map, the mean share of candidates closer than the true partner falls from 0.178 to 0.108 when the embeddings are centered and normalized.

Width. In the preregistered width test, we project every model to 128 principal components. This projection changes the fixed-depth trend of CKA minus its permutation mean along BLOOMZ by 0.00 at mid depth and -0.07 at the last layer. Width therefore does not drive this trend, in line with Proposition 4, which ties the CKA baseline to effective dimension rather than width. The preregistered rule then returns “genuine global loss” at both depths, a name whose claim is false as worded, since CKA minus its permutation mean keeps its fixed-depth trend at 0.70 ($p = 0.117$) and 0.92 ($p = 0.017$).

R.7 Layer Aggregation

Appendix F.5 of the paper summarizes this subsection.

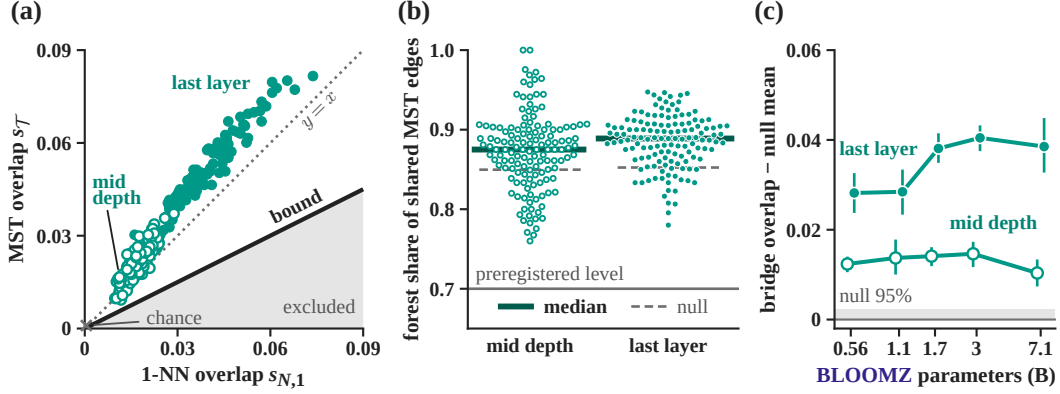


Figure R.3: **Spanning-tree overlap mostly depends on the one-nearest-neighbor scale, and bridge overlap exceeds its permutation null without an established size trend (Proposition 2).** Vision–language pairs of WIT at mid depth (open) and the last layer (filled). (a) MST overlap against one-nearest-neighbor overlap for each pair. (b) Per pair, the share of the shared MST edges that lie in the nearest-neighbor forest of the language model, with its value for a random subset of tree edges (null). (c) Bridge overlap minus its permutation null mean along BLOOMZ, as means over vision models with 95% intervals. Table R.9 gives the values.

m	depth	(a) frac. ≥ 0.70	median share	(b) frac.	(c,d) $\text{Tr}_{\text{bridge}}(p)$	(e) frac. shared bridges	rule outcome
1024	mid	0.99	0.875	1.00	$-0.12 (0.617)$	0.78	SOG-global $\approx$ 1-NN
1024	last	1.00	0.917	1.00	$0.72 (0.017)$	0.93	Genuine global convergence
2047	mid	1.00	0.875	1.00	$-0.16 (0.742)$	0.93	SOG-global $\approx$ 1-NN
2047	last	1.00	0.889	1.00	$0.57 (0.058)$	1.00	SOG-global $\approx$ 1-NN

Table R.9: Spanning-tree test on the 136 vision–language pairs, with the clip and ℓ_2 normalization of PRH (prh_clip+l2) and the forest of the language model. Columns give the fraction of pairs with forest share at least 0.70 (a) and the median share, the fraction with $|s_{\mathcal{T}} - s_{N,1}| \leq 0.03$ (b), the bridge trend with its exact one-sided p (c, d), the fraction of pairs whose bridge overlap exceeds its permutation null (e), and the mean number of shared bridges. The primary evaluation set has $m = 2047$ inputs. The last column applies the rule with the unadjusted p , which gives “genuine global convergence” on $m = 1024$ at the last layer ($p = 0.017$). Under the Holm correction of the primary family, both evaluation sets give “SOG-global $\approx$ 1-NN” at both depths.

Depth	pre	Tr, native (p)	Tr, 128 PCs (p)	Tr, mKNN (p)	change	rule outcome
mid	prh_clip+l2	0.70 (0.117)	0.70 (0.117)	0.84 (0.025)	0.00	Genuine global loss
mid	l2	$-1.00 (1.000)$	$-1.00 (1.000)$	$-0.92 (1.000)$	0.00	Genuine global loss
last	prh_clip+l2	0.92 (0.017)	0.85 (0.025)	0.98 (0.008)	-0.07	Genuine global loss
last	l2	0.63 (0.125)	0.63 (0.125)	0.68 (0.117)	0.00	Genuine global loss

Table R.10: Width test for biased CKA. All trends use s_{obs} minus its permutation mean. The column “pre” gives the preprocessing, the clip and ℓ_2 normalization of PRH (prh_clip+l2) or plain ℓ_2 normalization (l2). “Change” is the trend at 128 principal components minus the trend at native width. The last column gives the branch returned by the preregistered rule, “genuine global loss”, a name whose claim is false as worded, since the trend of CKA minus its permutation mean with the clip of PRH stays positive. Changes of at most 0.07 in absolute value show that width does not drive this trend.

This subsection gives the details of the layer-maximum experiments of Section 5.4 on BLOOMZ and LLaMA-1. In the preregistered experiment on WIT ($m = 1024$), we evaluate all 80,960 layer pairs of 136 vision–language pairs with $n_\pi = 500$ permutations per pair. Each permutation applies the same relabeling to every layer, and we recompute the maximum inside each permutation. Table R.11 gives the BLOOMZ trends under the three scores, and Figure R.4 the full version of Figure 4. In panels (d) to (i) of Figures 4 and R.4, every line and marker gives a mean over the 17 vision models, and the baseline of a language-model layer is its largest baseline over the layers of a vision model. Two implementations written independently agree on all 136×2 pairs to 10^{-14} , and both agree with the published code of ARH and PRH to within 6×10^{-9} . Across the three families of Figure R.4, the measured baselines plotted in panels (g–i) agree with the baselines predicted by Proposition 4 to a relative error of at most 4.4×10^{-4} . In panels (d–f), the standard error over vision models of each mean is at most 0.018. Over the 120 relabelings of the BLOOMZ sizes, the one-sided 5% critical values of the layer-max trend after subtraction are -0.80 and 0.80 . On BLOOMZ, the permutation means of CKA match the baseline of part 2 of Proposition 4 entrywise on all 80,960 layer pairs, with a median relative error of 0.078%. A calibrated trend of biased CKA after a layer maximum is interpreted together with the trend of the baselines that set its threshold, whichever layers enter the comparison.

BLOOMZ Trends. The uncalibrated layer-max CKA trend is 0.89 ($p = 0.042$), and after subtraction of the permutation mean the trend is -0.29 ($p = 0.68$; 95% interval over vision models $[-0.41, -0.18]$). Under s_{cal} the trend is -0.02 ($[-0.19, 0.12]$). After subtraction of the permutation mean, the layer-max trend of mKNN is 0.99 ($p = 0.008$), and that of debiased CKA is 0.84 ($p = 0.033$, exploratory). Part 3 of Proposition 4 gives this score a zero baseline at every layer pair. At fixed depth, with the permutations drawn for the layer maximum, the threshold of the fixed layers rises with size at mid depth and falls at the last layer ($\text{Tr} = 0.99$ and -0.91). Calibration moves the CKA trend opposite to the trend of the threshold, without a reversal. After subtraction of the permutation mean, the trend falls from 0.88 to 0.69 ($p = 0.10$) at mid depth and rises from 0.76 to 0.92 ($p = 0.017$) at the last layer.

BLOOMZ Thresholds. In post hoc analyses of the preregistered BLOOMZ experiment, the thresholds lie near the largest baseline, as assumed in Corollary 1. Over the 85 BLOOMZ pairs, the threshold c_α of the layer maximum lies 0.0017 to 0.0028 above the largest predicted baseline $\max_{\ell, \ell'} \sqrt{\text{PR}_A \text{PR}_B} / (m - 1)$, and its mean rises from 0.135 at 0.56B to 0.184 at 7.1B. Over the same sizes the observed maximum rises from 0.336 to 0.384, and the threshold of the layer pair attaining the maximum stays between 0.107 and 0.121. For every size except 1.1B the largest baseline sits in the first two language-model layers, where the participation ratio jumps from 134 at 0.56B to between 242 and 249 at 1.7B and above. The maximum and its threshold thus come from different layer pairs. From 0.56B to 7.1B the threshold rises by 0.048 and the largest baseline by 0.049, as much as the maximum (0.048). Between 1.1B and 1.7B the threshold jumps from 0.131 to 0.181, while the maximum moves from 0.359 to 0.355. The calibrated score therefore loses its upward trend, the outcome allowed by the corollary when the baselines grow as fast as the score. Two further analyses support the corollary on BLOOMZ, the first fixed in the frozen plan and the second post hoc. A subtraction of the threshold of the layer pair attaining the maximum gives $\text{Tr} = 0.76$. A restriction of the maximum to language-model layers at relative depth 0.5 or more removes the early high-baseline layers and restores the trend after subtraction of the permutation mean to 0.89 ($p = 0.017$).

Layer Maximum along LLaMA-1. Along LLaMA-1, we evaluate the 7B, 13B, and 30B models against the same 17 vision models under the protocol used for BLOOMZ, in 51 pairs with 47,520 layer pairs. The re-extraction of LLaMA-1 7B with the copied pipeline is bitwise equal to the stored features, and the three tokenizers give identical token identifiers on all 1,024 captions. Two independent implementations agree on every quantity of the plan, with a largest difference of 4.6×10^{-15} over 25,704 floating-point values. They reproduce the frozen LLaMA-1 7B results bitwise and to a relative 2.4×10^{-15} , respectively. In the reproduction by the second implementation, the relative deviation exceeds the strict tolerance of 10^{-12} in one field, a debiased-CKA permutation mean whose expectation is zero by part 3 of Proposition 4. The absolute deviation of 6.7×10^{-18} in this field lies within the mixed tolerance of the plan. The permutation means of biased CKA on the new layer pairs match the baseline of part 2 of this proposition with a median relative error of

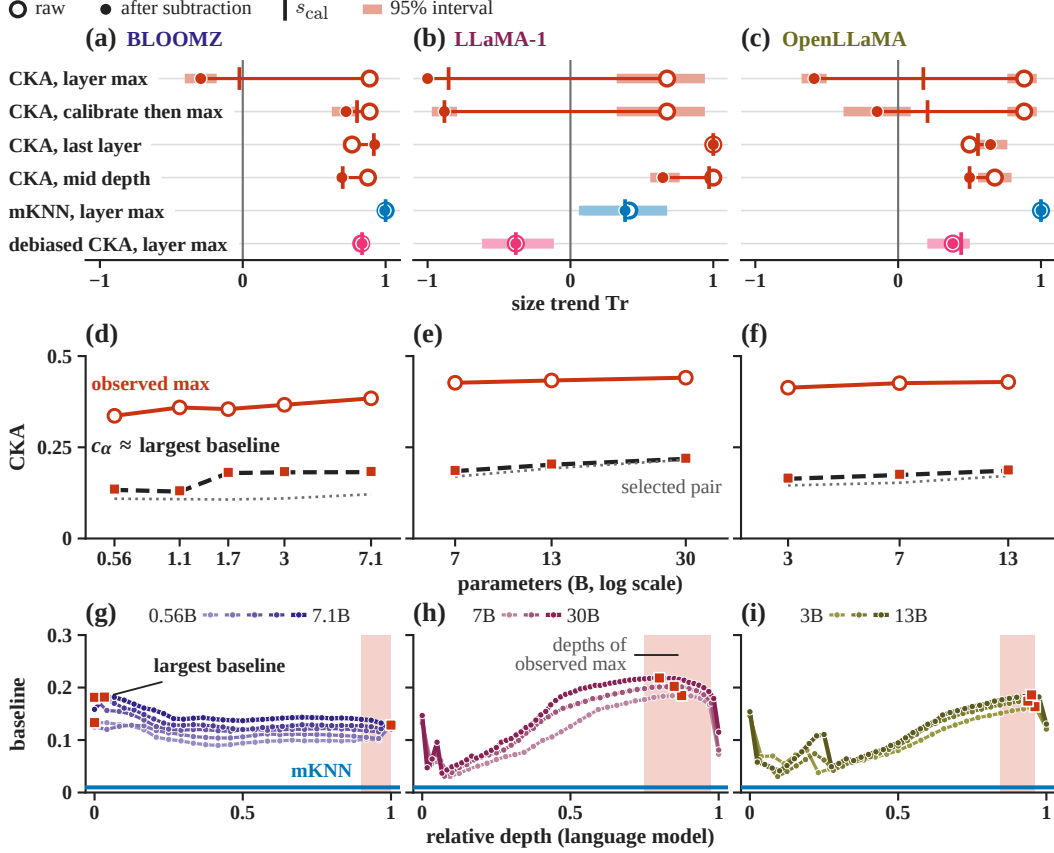


Figure R.4: **After a maximum over layers, the calibration threshold follows the largest random-pairing baseline along BLOOMZ, LLaMA-1, and OpenLLaMA (Proposition 4, Corollary 1).** This full version of Figure 4 adds the rows for calibration before the maximum, CKA at the last layer, and debiased CKA. The trends of debiased CKA and the BLOOMZ trends of calibration before the maximum are exploratory. (a–c) Size trend of each score in its raw form, after subtraction of the permutation mean, and under s_{cal} . (d–f) Squares give the threshold c_α , dashed lines the largest baseline predicted from features, and dotted lines the threshold of the layer pair attaining the maximum (“selected pair”). (g–i) CKA baseline of each language-model layer, predicted (lines) and measured (dots). Appendices R.7 and R.8 give the values.

0.07%, and every deviation lies within four Monte Carlo standard errors. The permutation means of debiased CKA are consistent with zero (mean $z = -0.02$). Table R.12 gives the trends.

LLaMA-1 Thresholds. On LLaMA-1 the thresholds come from the same late block as the maximum, and over the 51 pairs they lie 0.0015 to 0.0021 above the largest baseline predicted from features. Computed from single-model participation ratios before any cross-model score, these baselines have means 0.1847, 0.2023, and 0.2182 for 7B, 13B, and 30B, and the thresholds lie a median 0.0018 above them and follow their order in 51 of 51 comparisons. This baseline sits at relative depth 0.80 to 0.88, where the observed maximum also lies (mean depth 0.81 to 0.83). The threshold of the layer pair attaining the maximum rises from 0.170 to 0.216 and at 30B lies 0.004 below the threshold of the maximum. The two analyses that restored the trend on BLOOMZ therefore leave the LLaMA-1 trend reversed. A subtraction of this threshold gives $Tr = -0.91$, and a restriction to layers at relative depth 0.5 or more leaves s_{cal} at -0.85 . The growing number of layer pairs does not produce the reversal. The threshold of the maximum equals the threshold of the single highest-baseline layer pair to within 3.3×10^{-6} , and a maximum over 17 layers of each LLaMA-1 model gives a trend of -0.91 under s_{cal} . From 7B to 30B the threshold of the maximum gains 0.033 against 0.014 for the observed maximum. Calibration turns the layer-max CKA trend of 0.68

into -0.85 under s_{cal} (95% interval $[-1.00, -0.65]$) and into -1.00 after subtraction, while mKNN moves from 0.41 to 0.38 .

Scope of the LLaMA-1 Result. With three sizes the smallest exact one-sided p is $1/6$, so each LLaMA-1 outcome is a directional result on one family, and its intervals cover the choice of vision models but not of language models. In all 51 comparisons the thresholds follow the predicted order, but this order is the same for each of the 17 vision models. The vision model enters the predicted baseline as a common factor, so the order of these baselines is the order of the largest participation ratios of the three LLaMA-1 models. Part of the prediction that calibration lowers the CKA trend holds almost by construction. Proposition 4 fixes the permutation mean of biased CKA, matched on the new pairs to a median relative error of 0.07% , and the calibrated score s_{cal} of Section 2.1 does not increase with the threshold. Thresholds that rise with size for every vision model therefore lower or keep each per-model correlation, here without any score censored at zero. The raw gain falls below the threshold gain for 17 of 17 vision models at each pair of sizes, with medians of 0.006 against 0.019 from 7B to 13B. Depth and width grow with size along these three models, from 32 blocks of width 4096 at 7B to 60 blocks of width 6656 at 30B, and 30B was trained on more tokens than the two smaller models (Touvron et al., 2023). The growth of the baseline is therefore a property of the larger models of this family, and the comparison along these three models leaves parameter count, depth, width, and training data unseparated.

Fixed Depth and Calibration before the Maximum on LLaMA-1. At fixed depth the calibrated LLaMA-1 CKA trend keeps its sign, as predicted, at 0.97 at mid depth and 1.00 at the last layer. We had marked this prediction as at risk in advance, because the baselines of the fixed layers also rise with size, with means 0.050 , 0.065 , and 0.075 at mid depth. The sign survives because the raw scores of these layers rise faster, from 0.184 to 0.227 between 7B and 13B at mid depth against a threshold rise from 0.052 to 0.067 . Even so, the trend after subtraction of the permutation mean falls from 1.00 to 0.65 . Calibration of each layer pair before the maximum reverses the trend to -0.88 , although we had predicted in advance that the trend would keep its sign under this protocol. The layer pairs selected under this protocol stay on the late block, at mean relative depth 0.81 to 0.82 , where their own thresholds rise from 0.146 to 0.167 with size. On BLOOMZ, where the largest baseline sits away from the layers of the maximum, calibration under this protocol keeps the trend, at 0.80 for s_{cal} and 0.72 after subtraction. This result is exploratory, since we defined the protocol in the LLaMA-1 plan after seeing the BLOOMZ result.

Joint Test over BLOOMZ and LLaMA-1. The analysis plan for LLaMA-1 fixed the joint trend of the two families as descriptive, and the BLOOMZ half was known before the plan, so the following test is exploratory. The joint trend is the mean of the BLOOMZ and LLaMA-1 trends, and under calibration of the layer maximum it falls by 1.22 for s_{cal} and by 1.43 after subtraction of the permutation mean. Over the $120 \times 6 = 720$ joint relabelings of the sizes, these falls have exact one-sided p of $6/720$ and $2/720$.

Pooled Statistic of ARH. On our models, the loss of the CKA trend under calibration appears in the pooled statistic of ARH at most in a weaker form. Over the 136 pairs of the preregistered experiment, the Pearson correlation between score and language-model rank moves from 0.49 to 0.43 under calibration for CKA and stays at 0.52 for mKNN. The preregistered decision rule returns “ARH pattern absent”. Over the 170 pairs with LLaMA-1 13B and 30B, CKA moves from 0.51 to 0.40 and mKNN from 0.52 to 0.52 , and the difference of 0.12 between the two drops (95% interval over vision models $[0.003, 0.26]$) gives the category “partial”. With OpenLLaMA 13B added, the 187 pairs give 0.50 to 0.40 for CKA and 0.51 to 0.51 for mKNN, and the difference falls to 0.10 , with an interval of $[-0.02, 0.24]$ that contains zero. This interval leaves the direction of the loss not established over our 187 pairs (Appendix R.8). Since the interval also covers the boundary 0.10 of the preregistered decision rule, the category returned by this rule changes within the interval, and we report this category as not decidable with 187 pairs. The loss appears most clearly along the sizes within BLOOMZ, LLaMA-1, and OpenLLaMA, with five, three, and three sizes and unadjusted p -values. Along the seven Pythia sizes, calibration keeps the trend, and the preregistered prediction of its loss was not confirmed (Appendix R.9). Each of these size trends rests on one draw of evaluation inputs. In a secondary analysis along the sizes of the vision models at the last layer, CKA minus its permutation mean has $\text{Tr} = -0.04$ ($p = 0.55$) against 0.71 ($p = 0.005$) for mKNN, so the contrast reported by ARH appears there at fixed depth. Such a loss can come from baselines of the

fixed vision layers that grow with size faster than the score, as in part 2 of Corollary 1, and our experiments leave this explanation untested.

Metric	layers	raw s_{obs}	s_{cal}	$s_{\text{obs}} - \bar{s}_{\pi}$
CKA	layer-max	0.89 (0.042)	−0.02 (0.508)	−0.29 (0.683)
	fixed, mid depth	0.88 (0.042)	0.70 (0.117)	0.69 (0.100)
	fixed, last layer	0.76 (0.067)	0.92 (0.017)	0.92 (0.017)
mKNN	layer-max	1.00 (0.008)	1.00 (0.008)	0.99 (0.008)
	fixed, mid depth	0.85 (0.025)	0.84 (0.025)	0.84 (0.025)
	fixed, last layer	0.99 (0.008)	0.98 (0.008)	0.99 (0.008)
debiased CKA (expl.)	layer-max	0.83 (0.050)	0.84 (0.033)	0.84 (0.033)
	fixed, mid depth	0.70 (0.117)	0.70 (0.117)	0.70 (0.117)
	fixed, last layer	0.90 (0.025)	0.91 (0.017)	0.91 (0.017)

Table R.11: Trends Tr along BLOOMZ under the clip and ℓ_2 normalization of PRH, with exact one-sided p over 120 relabelings, unadjusted. The last column gives the trend of the uncensored score, s_{obs} minus the mean of its 500 permutation values. The fixed-depth rows are computed with the permutations drawn for the layer maximum, and the trends of debiased CKA are exploratory.

Metric	layers	raw s_{obs}	s_{cal}	$s_{\text{obs}} - \bar{s}_{\pi}$
CKA	layer-max	0.68 (0.167)	−0.85 (1.000)	−1.00 (1.000)
	fixed, mid depth	1.00 (0.167)	0.97 (0.167)	0.65 (0.333)
	fixed, last layer	1.00 (0.167)	1.00 (0.167)	1.00 (0.167)
	calibrate then max	n/a	−0.88 (1.000)	−0.88 (1.000)
mKNN	layer-max	0.41 (0.333)	0.38 (0.333)	0.38 (0.333)
	fixed, mid depth	0.82 (0.167)	0.82 (0.167)	0.82 (0.167)
	fixed, last layer	0.91 (0.167)	0.91 (0.167)	0.91 (0.167)
	calibrate then max	n/a	0.38 (0.333)	0.38 (0.333)
debiased CKA (expl.)	layer-max	−0.38 (1.000)	−0.38 (1.000)	−0.38 (1.000)
	fixed, mid depth	1.00 (0.167)	1.00 (0.167)	1.00 (0.167)
	fixed, last layer	1.00 (0.167)	1.00 (0.167)	1.00 (0.167)
	calibrate then max	n/a	−0.38 (1.000)	−0.38 (1.000)

Table R.12: Trends Tr along LLaMA-1 7B, 13B, and 30B under the clip and ℓ_2 normalization of PRH, with exact one-sided p over 6 relabelings, unadjusted, so the smallest attainable p is $1/6$. Columns follow Table R.11. For calibration before the maximum, the column s_{cal} gives the trend of the maximum of the calibrated layer-pair scores, and the last column the trend of its uncensored version. The trends of debiased CKA are exploratory.

Post Hoc Analyses of the 136 Preregistered Pairs. For biased CKA under the clip and ℓ_2 normalization of PRH, the null mean of the layer maximum exceeds the largest single-pair null mean by 6×10^{-5} on average. In contrast, the threshold of the maximum exceeds the threshold of the layer pair attaining the maximum by 0.040. The level of the highest-baseline layer pair therefore sets the threshold, rather than the number of layer pairs. For mKNN, whose baselines are equal across layers, the null mean of the layer maximum exceeds the largest single-pair null mean by 0.0027. Of the 68,000 null maxima of biased CKA, 33,884 fall at relative language-model depth below 0.1, while 112 of the 136 observed maxima fall at depth 0.9 or above. Unlike the restriction to relative depth 0.5 or more, which restores the uncensored trend to 0.89 ($p = 0.017$), a maximum without the embedding layer leaves this trend at −0.29. Under plain ℓ_2 normalization the uncensored layer-max CKA trend is −0.09 against 0.43 before calibration, and the mKNN trend is 0.69.

Secondary Preprocessing on LLaMA-1. Before any score was read, the plan declared the secondary preprocessing by plain ℓ_2 normalization not material, i.e., its predicted baselines differ too little between sizes to support a directional prediction (0.0023 between 13B and 30B). This preprocessing therefore receives a descriptive analysis. With ℓ_2 normalization, calibration moves the layer-max CKA trend from −0.50 to −0.56. The thresholds lie a median 0.0021 above the predicted baselines and follow their order in 32 of 32 qualifying comparisons, and the outcomes for mKNN and debiased CKA match their predictions. The prediction that the calibrated CKA trend keeps its

sign at fixed depth does not apply at mid depth, where the raw trend is 0.03, and holds at the last layer. Calibration before the maximum keeps the sign of the trend at -0.50 . Under ℓ_2 , the largest baseline sits at the embedding layer for all three sizes. Under the clip and ℓ_2 normalization of PRH, it sits on a late plateau, at layers 28, 34, and 48 (relative depth 0.875, 0.850, and 0.800). There the second highest layer has a participation ratio within 0.1 to 0.8% of the highest. The baselines of these neighboring layers differ by less than the standard deviation of the permuted maximum, 0.0010 to 0.0011. The threshold still equals the threshold of one layer pair, which suggests that adjacent layers move together under each permutation, an interpretation still without a direct measurement (exploratory).

R.8 Layer Maximum along OpenLLaMA

Appendix F.5 of the paper summarizes this subsection.

Along OpenLLaMA, we evaluate the 3B, 7B, and 13B models against the same 17 vision models under the protocol used for LLaMA-1 (Appendix R.7), in 51 pairs with 14,432 layer pairs for the new 13B model. Re-extractions of OpenLLaMA 3B and 7B after a driver update are bitwise equal to the stored features, and the 13B weights match their published checksums. The three tokenizers give identical token identifiers on all 1,024 captions. Two independent implementations agree on every quantity of the plan, with a largest difference of 4.2×10^{-15} over 25,704 floating-point values and equal values for all 17,238 exact items. One implementation reproduces the frozen 3B and 7B results bitwise, and the other to a relative 2.4×10^{-15} . A review recomputation from the pair files of each implementation reproduces all 5,709 decision values. The permutation means of biased CKA on the new layer pairs match the baseline of part 2 of Proposition 4 with a median relative error of 0.08%, and the permutation means of debiased CKA are consistent with zero. The published code of ARH agrees with our implementations on a reduced grid to 2.2×10^{-8} . Table R.13 gives the trends.

OpenLLaMA Thresholds. Under the clip and ℓ_2 normalization of PRH, the largest participation ratio over layers grows from 202 at 3B through 229 at 7B to 261 at 13B. The largest baseline grows with it for each of the 17 vision models, with means 0.1636, 0.1741, and 0.1860. This baseline sits on late layers, at relative depth 0.94 to 0.96, as on LLaMA-1. The offsets of the observed thresholds above this baseline range from 0.0016 to 0.0023 around their median of 0.0019, and the threshold of the layer pair attaining the maximum rises from 0.146 to 0.172. The offsets lie inside the frozen band of 0.0009 to 0.0029 for 17 of 17 vision models. From 7B to 13B the raw gain falls below the threshold gain for 17 of 17 vision models, with medians of 0.0018 against 0.0130, an outcome outside the predictions fixed in advance. From 3B to 13B the threshold and the largest baseline gain 0.022, against 0.016 for the observed maximum. The CKA trend of 0.88 falls to 0.18 under s_{cal} and to -0.59 after subtraction (95% interval $[-0.68, -0.50]$). Although the lines of Figure 4f are nearly parallel, the trend after subtraction is strongly negative, because Tr averages per-vision-model rank correlations over three sizes and the small excess of baseline growth recurs in all 17 vision models. The 30 correct forecasts of the order after calibration cover 16 of the 17 orders reversed by calibration, 10 of the 13 kept orders, and 4 comparisons in which 13B already scores lower before calibration. As on LLaMA-1, the thresholds follow the predicted order in 33 comparisons, but this order is the same for each of the 17 vision models, since the vision model enters the predicted baseline as a common factor. Once the thresholds track the largest baseline, the forecast of the order after calibration holds almost by construction, and a failure of this forecast would come from the raw gain. Under calibration, 17 of the 34 comparisons involving OpenLLaMA 13B turn from concordant to discordant, against 0 in the other direction.

Scope of the OpenLLaMA Result. Before the download of any 13B weight, the OpenLLaMA plan declared every outcome descriptive, and with three sizes the two drops of the trend have exact one-sided p of $2/6$ and $1/6$. Under calibration the mKNN trend stays at 1.00 and the debiased CKA trend moves by 0.06 or less, as predicted. Calibration before the maximum lowers the CKA trend from 0.88 to 0.21, in line with the prediction that this protocol does not raise the trend. The thresholds of the layer pairs selected under this protocol follow the predicted order in 30 of 33 comparisons, in line with the premise of this prediction. At both fixed depths, the CKA trend keeps its sign under calibration (descriptive). With an uncensored mKNN trend of 1.00, the decision rule fixed in the plan assigns OpenLLaMA the category “layer max loses the CKA trend”, whereas the

mKNN trend of 0.38 left LLaMA-1 in the category “intermediate”. The three OpenLLaMA models were trained on 1T tokens of the same corpus with the training recipe of LLaMA-1 (Geng and Liu, 2023). Their depth and width grow from 26 blocks of width 3200 at 3B through 32 blocks of width 4096 at 7B to 40 blocks of width 5120 at 13B. Unlike the three LLaMA-1 models, these three models share their training data, and as along LLaMA-1, their parameter count, depth, and width grow together.

Metric	layers	raw s_{obs}	s_{cal}	$s_{\text{obs}} - \bar{s}_{\pi}$
CKA	layer-max	0.88 (0.167)	0.18 (0.500)	−0.59 (0.833)
	fixed, mid depth	0.68 (0.333)	0.50 (0.500)	0.50 (0.500)
	fixed, last layer	0.50 (0.500)	0.56 (0.333)	0.65 (0.333)
	calibrate then max	n/a	0.21 (0.500)	−0.15 (0.667)
mKNN	layer-max	1.00 (0.167)	1.00 (0.167)	1.00 (0.167)
	fixed, mid depth	0.68 (0.333)	0.68 (0.333)	0.68 (0.333)
	fixed, last layer	0.56 (0.333)	0.59 (0.333)	0.56 (0.333)
	calibrate then max	n/a	1.00 (0.167)	1.00 (0.167)
debiased CKA (expl.)	layer-max	0.38 (0.500)	0.44 (0.500)	0.38 (0.500)
	fixed, mid depth	0.50 (0.500)	0.50 (0.500)	0.50 (0.500)
	fixed, last layer	0.59 (0.333)	0.53 (0.333)	0.56 (0.333)
	calibrate then max	n/a	0.38 (0.500)	0.38 (0.500)

Table R.13: Trends Tr along OpenLLaMA 3B, 7B, and 13B under the clip and ℓ_2 normalization of PRH, with exact one-sided p over 6 relabelings, unadjusted, so the smallest attainable p is $1/6$. Columns follow Table R.12. Every entry is descriptive, and the trends of debiased CKA are exploratory.

Robustness. The forecasts of the order after calibration stay correct for every band width examined, with correct forecasts for 34 of 34 comparisons at half-width 0.0003 and for all 27 comparisons forecast at 0.002. After subtraction of the permutation mean, a band of ± 0.0002 yields forecasts for 33 of the 34 comparisons, all correct. Under plain ℓ_2 normalization the largest baseline sits at the embedding layer for all three sizes, and its participation ratio of 178.45 at 13B lies below the 180.02 of 7B. The plan therefore declared this preprocessing not material before any score was read. With this normalization the thresholds lie a median 0.0021 above the largest baseline, and the bands yield a predicted order after calibration in 32 of the 34 comparisons, all correct. Calibration moves the layer-max CKA trend from -0.26 to -0.62 . Under this preprocessing, the difference between the drops of CKA and mKNN in the pooled statistic of ARH is 0.081 with interval $[0.006, 0.165]$, a secondary result that the plan excluded from the comparison with ARH (Appendix R.10).

Pooled and Joint Statistics. Over the 187 pairs of all eleven language models, the difference between the drops of CKA and mKNN in the pooled statistic of ARH is 0.103, with a 95% interval over vision models of $[-0.017, 0.239]$. Its exact one-sided p over the $11!$ relabelings of the language models is 0.12, a descriptive value since 10 of the 11 models were known before the plan. Before OpenLLaMA 13B was added, the same difference was 0.116 with interval $[0.003, 0.26]$. In post hoc checks, 50 further bootstrap seeds give lower ends between -0.021 and -0.0004 , and a leave-one-out over vision models gives differences between 0.088 and 0.161. A joint permutation test over BLOOMZ, LLaMA-1, and OpenLLaMA compares the weakest of the three drops in Tr with its value under all 4,320 joint relabelings and gives $p = 48/4320$ for s_{cal} and $2/4320$ after subtraction (exploratory). The BLOOMZ result generated the tested hypothesis, and the trend of the largest BLOOMZ baseline fails the materiality rule of the plan. The BLOOMZ and LLaMA-1 results were also known before the OpenLLaMA plan. With the BLOOMZ and LLaMA-1 drops held at their observed values, a rule rejecting at $p \leq 0.05$ rejects under 2 of the 6 relabelings of the OpenLLaMA sizes. This joint value therefore carries little evidence about OpenLLaMA itself.

R.9 Layer Maximum along Pythia

Appendix F.5 of the paper summarizes this subsection.

Calibration kept the layer-max CKA trend along the seven Pythia models, from 0.93 to 0.89 under s_{cal} . The primary preregistered prediction of a drop of at least 0.10 consistent across the

sizes was not confirmed. The trend changes by -0.034 under calibration, with a 95% interval of $[-0.046, -0.023]$ over vision models. The permutation p of this change for consistency across sizes is 1612/5040, and its bootstrap p for a drop of at least 0.10 is 1.0. Since the plan made the secondary tests conditional on a confirmed primary prediction, we did not carry out these tests. Descriptively, calibration before the maximum also gives 0.89, and the trends of mKNN and debiased CKA move by 0.00. Every comparison of these two scores keeps its order under calibration. At fixed depth the CKA trend keeps its sign, from 0.98 to 0.96 at mid depth and from 0.67 to 0.80 at the last layer. Table R.14 gives the trends.

Metric	layers	raw s_{obs}	s_{cal}	$s_{\text{obs}} - \bar{s}_{\pi}$
CKA	layer-max	0.93 (0.0020)	0.89 (0.0042)	0.89 (0.0042)
	fixed, mid depth	0.98 (0.0004)	0.96 (0.0014)	0.96 (0.0014)
	fixed, last layer	0.67 (0.0423)	0.80 (0.0073)	0.90 (0.0012)
	calibrate then max	n/a	0.89 (0.0042)	0.90 (0.0042)
mKNN	layer-max	1.00 (0.0002)	1.00 (0.0002)	1.00 (0.0002)
	fixed, mid depth	0.99 (0.0002)	0.99 (0.0002)	0.99 (0.0002)
	fixed, last layer	0.94 (0.0004)	0.95 (0.0004)	0.95 (0.0004)
	calibrate then max	n/a	1.00 (0.0002)	1.00 (0.0002)
debiased CKA (expl.)	layer-max	0.90 (0.0042)	0.90 (0.0042)	0.90 (0.0042)
	fixed, mid depth	0.97 (0.0004)	0.96 (0.0014)	0.96 (0.0014)
	fixed, last layer	0.84 (0.0024)	0.84 (0.0024)	0.84 (0.0024)
	calibrate then max	n/a	0.90 (0.0042)	0.90 (0.0042)

Table R.14: Trends Tr along Pythia 70M to 6.9B under the clip and ℓ_2 normalization of PRH, with exact one-sided p over 5040 relabelings of the sizes, unadjusted. Columns follow Table R.12. The preregistered confirmatory test required a drop of at least 0.10 in the layer-max CKA trend under s_{cal} , and the trend fell by 0.034. Every other entry is descriptive, and the trends of debiased CKA are exploratory.

Thresholds. To forecast the offset of each threshold above its predicted baseline, the analysis plan fixes for model i the band $f(\text{PR}_i^*) \pm 0.0008$, where PR_i^* is the largest participation ratio over the layers of the language model. In this plan, we fitted $f(\text{PR}^*) = 0.001069 + 0.01292/\sqrt{\text{PR}^*}$ on the 187 known pairs of the 11 language models of the other families. The values of PR^* are 91.5, 118.5, 129.1, 144.6, 156.3, 142.4, and 166.6 from 70M to 6.9B, so the bands of 70M and 160M extrapolate below the fitted range, which starts at 124.5. The layer of PR^* is the embedding output for 70M to 1B and lies at relative depth 0.031 to 0.094 for the three larger models. The median offset of the thresholds above the predicted baseline lies between 0.0020 and 0.0021 for every model, inside its band, and the threshold order matches the predicted order in 307 of the 307 comparisons with a material predicted gap. Given the raw scores, the bands yield a predicted order after calibration in 352 of the 357 comparisons, all correct. These predicted orders include 11 of the 14 orders reversed by calibration. The permutation means of CKA on the 53,856 new layer-pair entries match the baseline of part 2 of Proposition 4 with a median relative error of 0.09%. From size to size, the raw score of the layer maximum gains a median 0.024 per doubling of parameters against 0.0063 for the threshold.

Scope. With seven sizes, the exact one-sided p of a trend can reach 1/5040. The confirmatory test combined a permutation test of the consistency of the drop across sizes with a bootstrap test of its magnitude. The seven models share their training data and its order (Biderman et al., 2023), and their width grows from 512 at 70M to 4096 at 6.9B. Their depth does not grow monotonically, since 1B has 16 blocks against 24 at 410M and 1.4B. Along these seven models, the training data are thus fixed, and size and width grow together. The plan fixed the wording of Section 5.4 for each outcome before any Pythia weight was downloaded, and that section reports a subset of these numbers in a different order, a post hoc choice of wording.

Robustness. With one size left out, the change of the trend lies between -0.054 and -0.017 , and the five sizes from 410M give -0.065 ($p = 0.35$), so every subset stays short of the required 0.10. Without the censoring of s_{cal} the change is -0.038 ($p = 1520/5040$), and with one of the four vision-model families left out it lies between -0.041 and -0.027 . The category returned by the

decision rule stays the same for every required drop from 0.06 to 0.20 in place of the preregistered 0.10, and changes at 0.03. Under plain ℓ_2 normalization, a secondary and descriptive preprocessing, calibration turns the trend from 0.68 into -0.06 ($p_{\text{perm}} = 47/5040$). There the raw score gains 0.0015 per doubling against 0.0049 for the threshold, as the condition of part 2 of Corollary 1 requires. This drop rests on 6.9B, without which p_{perm} is 0.12.

Weights, Validation, and Compute. The 2.8B model was loaded from its final training checkpoint (step 143000), as the frozen rule requires when the main weight file differs from this checkpoint. The main file holds 32 rotary buffers outside the names allowed by the rule, and one of its layer-norm biases equals the checkpoint in every value but differs in its bytes, while the other 387 shared tensors are bitwise equal. We froze the plan at 00:17 UTC on 2026-10-10 and the code at 04:05. The first byte of Pythia weights reached the machine at 07:52, and the predictions from features were frozen at 08:22, earlier than every score file. Two independent implementations agree on every integer, p -value, decision, and category and differ in 163 pair-level and 3 descriptive floating-point values by at most 9.1×10^{-14} , a difference accepted by a post hoc ruling. A third implementation reproduces both (Appendix R.4, Deviations). The Pythia experiment used 15.0 GPU-minutes.

R.10 Comparison with the Published Results

Appendix F.6 of the paper summarizes this subsection.

This subsection compares our measurements with the published results of PRH, ARH, BPC, and SOG. On the models shared with these studies, our measurements agree with a central observation of PRH, of ARH, and of BPC. With the preprocessing of PRH, layer-max mKNN rises with BLOOMZ size ($\text{Tr} = 1.00$ before calibration), in line with the cross-modal trend of Huh et al. (2024). For every metric, the scores of all 136 pairs are significant at $p \leq 0.05$ against their permutation null, as are those of the 204 pairs of Gröger et al. (2026). Under s_{cal} , the layer-max CKA trend is lost along BLOOMZ and OpenLLaMA (-0.02 and 0.18) and reverses along LLaMA-1 (-0.85), in line with the loss reported by ARH. ARH plots this loss for RBF kernel alignment in its main figure and reports the same loss for linear CKA, the version of CKA in our measurements. Along Pythia, whose raw scores outgrow their baselines, the trend stays at 0.89 (Appendix R.9). BPC reports that mKNN at $k = 10$ for its primary vision-language pair falls from 0.135 on 1,024 WIT pairs to 0.008 on fifteen million LAION pairs, with the layer pair fixed. On our COCO evaluation sets, the median vision-language mKNN at $k = 10$ falls in the same direction, from 0.091 to 0.0085 over a 64-fold range of m .

Our theory gives a mechanism for three published observations, by SOG, ARH, and BPC, and we state each mechanism at the strength of its supporting result. SOG reports that spanning-tree overlap tracks mKNN with a Spearman correlation of 0.968 over 204 pairs (You et al., 2026). By Proposition 2, the MST contains the nearest-neighbor forest for pairwise distinct distances, and on our models most shared MST edges are shared nearest-neighbor edges, as expected from the forest share of each tree. This correlation therefore comes at least partly from the nearest-neighbor forest. On our models, MST overlap and one-nearest-neighbor overlap have a correlation of 0.96 over both depths (exploratory), computed from uncalibrated scores at fixed depth rather than from the calibrated layer-max scores of SOG. We derived Corollary 1 after seeing the BLOOMZ result. By this corollary, the loss of the calibrated CKA trend can come from the growth of the baselines that set the threshold, which under the layer maximum are the largest baselines across layers. From these baselines, computed from features before any score of the new models, we predicted the drop of the LLaMA-1 trend. Given the raw scores, the thresholds forecast from features yield a predicted order after calibration in 30 of the 34 comparisons involving OpenLLaMA 13B and in 352 of the 357 Pythia comparisons. Every predicted order is correct. The heuristic first-order law behind Conjecture D.1 predicts the stability of scores at fixed k/m observed by BPC at $k = m/100$. Descriptively, the crossing law with an estimated coupling recovers the direction of the collapse of fixed- k mKNN for vision-language pairs, as does an affine null, but underestimates its size (Appendix R.5).

Our results differ from those of these studies in four respects. First, the BLOOMZ trends at mid depth reverse under plain ℓ_2 normalization, so their sign depends on the clip of PRH, and preprocessing belongs to the criterion. Second, the pooled statistic of ARH changes less under calibration on our models than on the models of ARH. Over our 136 pairs it moves from 0.49 to 0.43 for CKA and from 0.52 to 0.52 for mKNN, against 0.86 to 0.45 and 0.85 to 0.84 on the models of ARH.

With LLaMA-1 13B and 30B and OpenLLaMA 13B added, the 187 pairs give 0.50 to 0.40 and 0.51 to 0.51. The difference between the two drops is 0.10, with a 95% interval over vision models of $[-0.02, 0.24]$ that contains zero, so the direction of the loss reported by ARH is not established on our models. The corresponding difference of 0.40 in ARH lies above this interval, so this difference is smaller on our models, which do not include LLaMA-65B. ARH also reports a loss for unbiased CKA, from 0.65 to -0.01 . Our debiased CKA uses the same estimator and moves from 0.37 to 0.36 over our 187 pairs (Appendix R.2). Part 3 of Proposition 4 gives this score a zero baseline at every layer pair, and its permutation means on the layer pairs of LLaMA-1 and OpenLLaMA 13B are consistent with zero (Appendices R.7 and R.8). This zero baseline places debiased CKA under part 1 of Corollary 1, by which calibration does not reverse the order of model pairs with well-separated maxima (Section 3.4). The source of the difference from ARH therefore remains open. Third, ARH attributes the inflation of layer-max scores to the number of layer pairs (Gröger et al., 2026). For biased CKA the null mean of the maximum exceeds the largest single-pair null mean by 6×10^{-5} on our 136 pairs and by at most 10^{-5} on LLaMA-1. The level of the highest-baseline layer pair therefore sets the threshold, rather than the number of layer pairs (exploratory). On BLOOMZ this layer pair differs from the pair attaining the maximum, whereas on LLaMA-1 both pairs lie on the same late block. Over the 136 pairs, the threshold of the pair attaining the maximum lies 0.040 below the threshold of the maximum on average. Fourth, SOG interprets its spanning-tree trends as convergence of global structure, whereas on our models the shared spanning-tree edges are mostly nearest-neighbor edges and the size trend of bridge overlap stays unresolved.

Our controlled runs add two results to this comparison. In the CIFAR-5m networks with a linear head, the linear map of the head adds an excess to the geometric discrepancy without lowering agreement on neighbor sets, and weight decay on the head removes this excess. This result suggests one candidate mechanism for why representations agree on neighbor sets but not on global geometry, a question left open by ARH. Under momentum without weight decay, most of this excess on G_s at 250,000 images comes from a gauge shared by all runs trained this way, and the rest from the run-to-run mismatch addressed by Theorem 2 (preregistered, Appendix R.12.1). Our contrastive networks also disagree on neighbor sets across objectives (Appendix R.14), so these networks are not described by the idealized contrastive learner of PRH, which recovers one kernel up to an offset.

R.11 Linear Autoencoder

Appendix F.7 of the paper summarizes this subsection.

In the linear-autoencoder experiment of Section 5.5, the inputs $x \in \mathbb{R}^{16}$ are Gaussian with mean zero and covariance $U\Lambda U^\top$, $\Lambda = \text{diag}(4, 3, 2, 1.5, 0.2, \dots, 0.2)$ with twelve eigenvalues 0.2, and U fixed by a hash. The latent dimension is $n_h = 4$, and $m = 256$ probes drawn independently of every training stream form the rows of $\mathbf{X}$ and define the probe Gram matrix $K = \mathbf{X}E^\top E\mathbf{X}^\top$ of Theorem 2. Every criterion uses the unnormalized features Ex of the probes, without the unit normalization of Section 2, and the training loss uses the second-moment matrix $N^{-1} \sum_i x_i x_i^\top$ of the run’s sample without centering, or $U\Lambda U^\top$ at $N = \infty$. Runs start from $\mathcal{N}(0, 1/\text{fan-in})$, or from entries of scale 0.01 under small initialization, and train until the relative gap to the closed-form optimum or the relative gradient falls below 10^{-8} . Weight decay uses $\lambda_{\text{wd}} = 0.075$. Stochastic gradient descent and Adam use minibatches of 16 for 4×10^5 steps at a constant rate and 10^5 steps of decay, and every stochastic run ends with full-batch gradient steps until the relative gap or the relative gradient falls below 10^{-8} . For each training procedure we trained 48 disjoint pairs of runs at each $N \in \{2^8, 2^{12}, 2^{16}\}$ and 500 pairs at $N = \infty$, the population limit. The tolerances were 0.05 for $1 - s_N$ at $k = 10$, 0.02 for G_s , and 0.01 for $1 - \text{CKA}$ and for the projector distance. A pair with a diverged run fails every criterion. Table R.15 gives every decision, and Figure R.5 shows every criterion, training procedure, and sample size. Run r draws its own training sample and minibatches at each N , so the two runs of a pair differ in initialization, training sample, and minibatches. The initialization and training sample of run r are the same under every training procedure.

Gradient Flow and Other Optimizers. In the linear autoencoder, each arm holds the runs of one training procedure at one step size. The preregistered gradient-descent arm uses a step of 0.05, which changes the imbalance C of part 1 of Theorem 2 by a median 53%. A secondary arm declared before the confirmatory runs uses a step of 0.002 and conserves C to 0.18%, so this arm stands in for gradient flow. In that arm the spectrum of the learned $EE^\top$ equals $f(\text{spec } C_0)$ of part 1 to a median relative error of 0.1%, and the condition number $\text{cond}(EE^\top)$ has median 4.66, with 10th and 90th

percentiles of 2.98 and 7.43 across initializations. Part 1 predicts a median of 4.67 for this arm. At both steps, 0 of 48 pairs of gradient-descent runs at every finite N and 0 of 500 at $N = \infty$ lie within tolerance on 10-nearest-neighbor sets, G_s , and CKA (Figure 5 for the step-0.002 arm). Under weight decay, the median condition number of the learned $EE^\top$ is 1.0330, against 1.0329 predicted by part 3, a comparison made for Figure R.5e outside the preregistered tests. Tied weights, small initialization, and weight decay use the preregistered step of 0.05, since part 3 holds at any step and small initialization keeps the imbalance C of part 1 near zero. In Figure R.5e, the predictions for the step-0.002 arm and small initialization come from part 1, and those for weight decay and tied weights from part 3. Minibatch stochastic gradient descent and Adam lie outside the premises of the theorem. They change C during training and leave the condition number of the learned $EE^\top$ at an intermediate value (medians 1.19 and 1.54). On the three criteria whose outcomes depend on the gauge, agreement between runs under these optimizers stays unreliable at every N , with 322 and 172 of 500 pairs within tolerance on CKA at $N = \infty$. On neighbor sets and G_s , 164 and 12 of 500 pairs lie within tolerance under stochastic gradient descent and 13 and 0 of 500 under Adam. Under the training procedures whose runs agree reliably, every failing pair at $N \geq 2^{12}$ contains a diverged run.

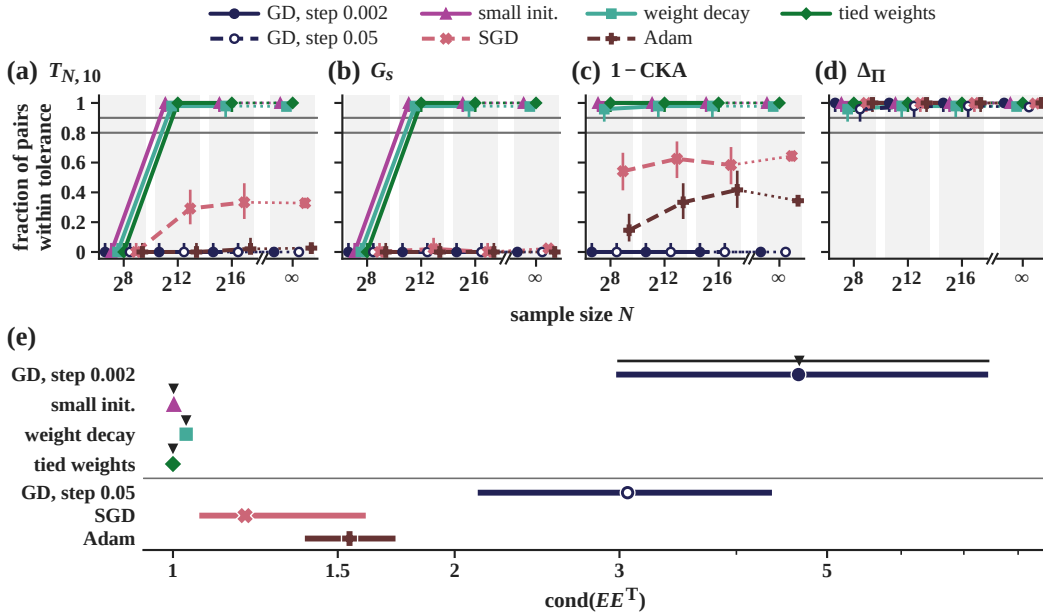


Figure R.5: **On the criteria whose outcomes depend on the gauge, agreement between runs trained by gradient descent without a regularizer stays unreliable at every sample size, whereas agreement between runs trained with weight decay, tied weights, or small initialization becomes reliable as N grows (Theorem 2).** (a–d) Fraction of disjoint pairs of linear-autoencoder runs within tolerance against N , with one-sided 95% Clopper–Pearson bounds and grey decision levels. Dashed lines mark training procedures outside the premises of the theorem, and dotted segments cross the axis break. The projector criterion (d) is a control whose outcome is invariant to the gauge. (e) Condition number of the learned $EE^\top$ at $N = \infty$ as the median and the 10th to 90th percentile of runs, with the predictions of parts 1 and 3 for the training procedures covered by the theorem in black. Table R.15 gives every decision.

Failed Prediction P1. For gradient descent at step 0.05, P1 required unreliable agreement on neighbor sets, G_s , and CKA at every N , and a mean discrepancy at $N = 2^{16}$ inside the two-sided 95% interval of the mean at $N = \infty$. The first part holds in 12 of 12 cells. The second part holds for neighbor sets (0.2787 against $[0.2699, 0.2806]$) and fails for G_s (0.4287 against $[0.4072, 0.4286]$) and CKA (0.1496 against $[0.1319, 0.1447]$). The paired differences between the same initializations at $N = 2^{16}$ and $N = \infty$ are 8.6×10^{-4} , 1.3×10^{-4} , and -1.0×10^{-4} , each with a 95% interval containing zero. The plan stated before the runs that this literal test passes with probability about 0.46 when the two means are equal, because the 48-pair mean is far noisier than the 500-pair interval. For the secondary arm at step 0.002, the second part fails for all three criteria. P1 remains falsified,

Training procedure	N	$1 - s_N(10) \leq 0.05$	$G_s \leq 0.02$	$1 - \text{CKA} \leq 0.01$	projector ≤ 0.01
GD, step 0.05 (prereg.)	2^8	0/48 U	0/48 U	0/48 U	46/48 I
	2^{12}	0/48 U	0/48 U	0/48 U	47/48 R
	2^{16}	0/48 U	0/48 U	0/48 U	47/48 R
	∞	0/500 U	0/500 U	0/500 U	486/500 R
GD, step 0.002 (secondary)	2^8	0/48 U	0/48 U	0/48 U	48/48 R
	2^{12}	0/48 U	0/48 U	0/48 U	48/48 R
	2^{16}	0/48 U	0/48 U	0/48 U	48/48 R
	∞	0/500 U	0/500 U	0/500 U	500/500 R
weight decay	2^8	0/48 U	0/48 U	46/48 I	46/48 I
	2^{12}	47/48 R	47/48 R	47/48 R	47/48 R
	2^{16}	47/48 R	47/48 R	47/48 R	47/48 R
	∞	488/500 R	488/500 R	488/500 R	488/500 R
weight decay, step 0.002 (sec.)	2^8	0/48 U	0/48 U	48/48 R	48/48 R
	2^{12}	48/48 R	48/48 R	48/48 R	48/48 R
	2^{16}	48/48 R	48/48 R	48/48 R	48/48 R
	∞	500/500 R	500/500 R	500/500 R	500/500 R
tied weights	2^8	0/48 U	0/48 U	48/48 R	48/48 R
	2^{12}	48/48 R	48/48 R	48/48 R	48/48 R
	2^{16}	48/48 R	48/48 R	48/48 R	48/48 R
	∞	500/500 R	500/500 R	500/500 R	500/500 R
small init	2^8	0/48 U	0/48 U	48/48 R	48/48 R
	2^{12}	48/48 R	48/48 R	48/48 R	48/48 R
	2^{16}	48/48 R	48/48 R	48/48 R	48/48 R
	∞	500/500 R	500/500 R	500/500 R	500/500 R
SGD, batch 16	2^8	0/48 U	0/48 U	26/48 U	48/48 R
	2^{12}	14/48 U	1/48 U	30/48 U	48/48 R
	2^{16}	16/48 U	0/48 U	28/48 U	48/48 R
	∞	164/500 U	12/500 U	322/500 U	498/500 R
Adam, batch 16	2^8	0/48 U	0/48 U	7/48 U	48/48 R
	2^{12}	0/48 U	0/48 U	16/48 U	48/48 R
	2^{16}	1/48 U	0/48 U	20/48 U	48/48 R
	∞	13/500 U	0/500 U	172/500 U	500/500 R

Table R.15: Linear autoencoder: pairs within tolerance out of 48 disjoint pairs (500 at $N = \infty$), with the decision reliable (R), inconclusive (I), or unreliable (U). For $N \geq 2^{12}$, every failing pair under weight decay and every pair failing the projector criterion contains a diverged run. These runs diverged at the step size 0.05 under gradient descent and weight decay, and in the minibatch phase for the two failing pairs of stochastic gradient descent at $N = \infty$.

and our conclusion that more data leaves agreement between gradient-descent runs unreliable rests on the first part.

R.12 CIFAR-5m Networks with a Linear Head

Appendix F.8 of the paper summarizes this subsection.

In the first network experiment of Section 5.5, the trunk is the airbench94 network (Jordan, 2024) on 32×32 CIFAR-5m images (Nakkiran et al., 2021) with trunk features $h \in \mathbb{R}^{256}$, followed by the bias-free head $u = Mh$, $y = Wu$ with $u, y \in \mathbb{R}^9$, and logits equal to $y/9$ mapped by a fixed 10×9 matrix with orthonormal columns. Each run follows the airbench94 recipe (Nesterov momentum 0.85, label smoothing 0.2, flips and translations) for 488 steps of 1024 images. Three arms differ in the head alone, without weight decay on M and W (noWD), with weight decay on both (WD), or with a tenfold smaller initialization of both (small init). Each arm was trained on $N = 25,000$ and 250,000 images at a fixed budget of 500,000 images seen, in 192 runs per arm and size, which gives 96 disjoint pairs. Each run draws its own training subset from five million images, so two subsets overlap by 0.5% at 25,000 and 5% at 250,000 in expectation, and features are compared without normalization on 2048 held-out images. The two runs of a pair differ in initialization, training

subset, data order, and augmentation. The tolerance of each criterion is the largest discrepancy between the outputs y over 24 pairs of pilot runs at $N = 100,000$, eight per arm, and u is judged against the tolerance of y . Table R.16 gives the tolerances, Table R.17 every decision, and Figure R.6 the fractions of agreeing pairs and the paired excess for every training procedure. Appendix R.12.1 reports a second, preregistered experiment with a gradient-flow head and tolerances set at each size.

Preregistered Decisions. At $N = 250,000$, weight decay makes u agree in 96 of 96 pairs under all nine criteria, as predicted by part 3, and y agrees in 96 of 96 pairs under every criterion in every arm. Without weight decay, u agrees on 10-nearest-neighbor sets in 96 of 96 pairs, against our preregistered prediction P1 of unreliable agreement, so this result meets the preregistered falsification condition. The neighbor criterion is loose here, since the largest discrepancy of the pilot outputs sets its tolerance at 0.84 on $1 - s_N$. Agreement of u fails instead on the scale-normalized criteria, where it is unreliable under G_s at 65 of 96 pairs and inconclusive under $1 - \text{CKA}$ at 73 of 96 (Table R.17). Both decisions depend on a tolerance set at $N = 100,000$. The mean G_s discrepancy of u is 0.2050 , just below the tolerance $\varepsilon = 0.2073$. The fraction of passing pairs is 0 at 0.9ε , 0.68 at ε , and 1 at 1.1ε , so a tolerance 4.7% larger would make the cell reliable (Figure R.6b). The mean $1 - \text{CKA}$ of u is 0.0812 against a tolerance of 0.0833 . With tolerances from the same arm and size, the second network experiment finds agreement of u on G_s unreliable at both sizes, with 0 and 5 of 96 pairs within tolerance. For the common-gauge counterfactual of u , 1 and 21 of 96 pairs lie within tolerance, and its agreement is also unreliable (Appendix R.12.1). Under the pooled tolerance, which is 7.8% larger, 95 of 96 pairs of this counterfactual lie within tolerance at $250,000$, so we report these decisions as tolerance-sensitive, without a categorical claim. Under momentum the final imbalance no longer reflects its initial value, so the imbalance that raises the geometric discrepancy of u in this arm arises during training rather than from the initialization. In the second network experiment, momentum creates such an imbalance even from a balanced or small start. Under head flow, the initial imbalance still raises the G_s discrepancy of u , by about one fifth of the amount under momentum from the default initialization (Appendix R.12.1). Under a tenfold smaller initialization, 82 of 96 pairs of u lie within tolerance on G_s , so small initialization does not replace weight decay under this optimizer. Prediction P5 and the part of prediction P3 on small initialization fail at this count (Failed Predictions below). At $N = 25,000$, 0 of 96 pairs lie within tolerance for u and y on every criterion under each of the three training procedures with independent training subsets, because the tolerance came from pilot runs at $N = 100,000$. With tolerances set at each size, the second network experiment finds agreement of u under weight decay reliable at both sizes.

The Paired Excess. An exploratory comparison of the discrepancies of u and y within each pair shows the criteria on which the gauge of the head changes the discrepancy (Figure R.6c). Without weight decay, u exceeds y in G_s discrepancy in 96 of 96 pairs at both sizes, by a median 0.023 at $25,000$ and 0.020 at $250,000$, and by 0.019 in the pilot pairs at $100,000$. On $T_{N,10}$ the median excess of u over y lies between -0.0011 and 0.0001 in every cell, and between -0.0011 and 0.0003 with the pilot pairs included. Weight decay removes the G_s excess (-0.0001 at $250,000$), and small initialization keeps most of the excess (0.016). Applied to the outputs of both runs of a pair, one common gauge produces about three quarters of the excess at $250,000$ (0.015 of 0.020). The preregistered decision on u therefore depends both on a gauge shared by all runs and on the run-to-run mismatch addressed by Theorem 2. In the second network experiment, we take the mean gauge of its pilot runs as the common gauge and judge agreement under the output discrepancy plus the mismatch part, with tolerances from the same arm and size. Under momentum without weight decay at $250,000$, this agreement is inconclusive, so its categorical status depends on the tolerance (Appendix R.12.1). Order criteria stronger than the neighbor sets can detect a gauge, and on these criteria the excess of u over y has the opposite sign. On T_E the pairs of u agree more closely than the pairs of y (median -0.023 at $250,000$, 96 of 96 pairs), a difference predicted by neither Theorem 2 nor Corollary 2.

Optimizer, Projector, and Trunk. These runs fall in the case of Remark 1, in which finite steps and momentum change C during training. Stochastic gradient descent with momentum raised the median imbalance $\|MM^\top - W^\top W\|_F$ of the head from 0.98 to 8.59 at $N = 250,000$, about ninefold, and from 0.0098 to 6.92 under small initialization, about seven-hundredfold. Weight decay lowered the imbalance to 0.43 . Under momentum, the imbalance at the end of training no longer reflects its initial value, since in the second network experiment the median cosine between the final and the initial C is -0.017 and -0.005 (Appendix R.12.1). In the first experiment, we based this

conclusion on the fraction of runs whose final gauge lies closer to their own initial gauge than to that of another run (46% to 52%). Under head flow, which conserves C , this fraction stays at the chance rate, so the statistic is insensitive to memory of the initialization. Agreement of u under the projector criterion is reliable in every arm at $N = 250,000$, because $\Pi_u = \Pi_y$ (largest deviation 4.9×10^{-8} over 1,344 runs) makes this criterion compare the column spaces of y . The mean $s_N(10)$ of the trunk features differs across arms by at most 0.011, as predicted by P4. Weight decay on the head also lowers the mean G_s discrepancy of the trunk features h from 0.239 to 0.180 at 250,000 (exploratory). In an exploratory cell with a shared training corpus every discrepancy fell, consistent with the distinction between independent and shared corpora drawn in Appendix C.

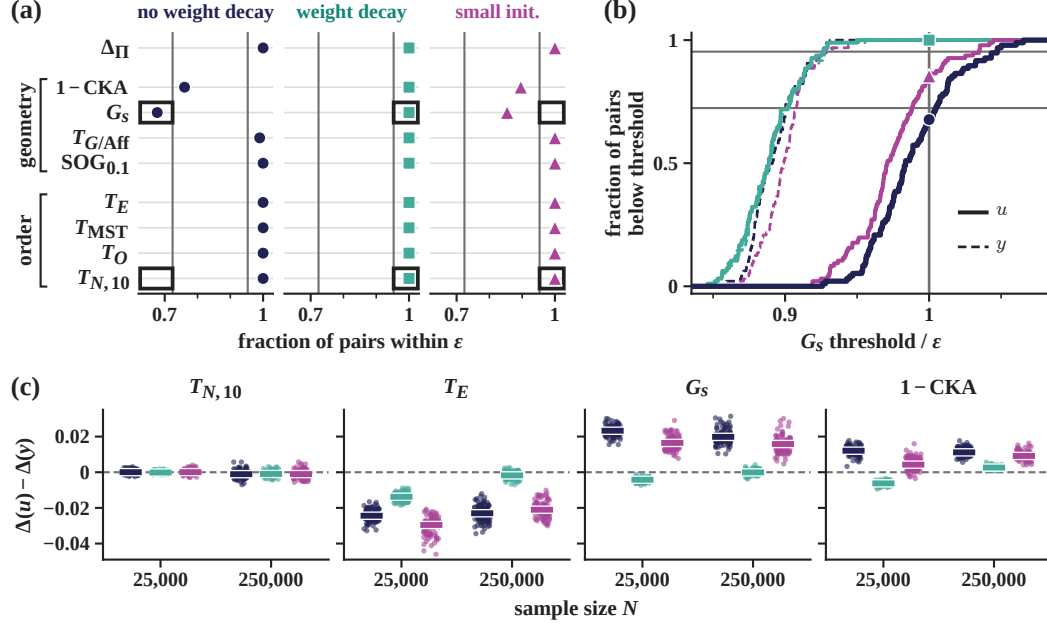


Figure R.6: **Without weight decay, the gauge of a linear head raises the discrepancy of the head coordinates u above that of the outputs y on scale-normalized geometry but not on neighbor sets, and weight decay makes agreement of u reliable under every criterion at $N = 250,000$ (Theorem 2, Remark 1).** Most of this excess comes from a gauge shared by all runs (Table R.21). (a) Fraction of disjoint pairs of CIFAR-5m runs at $N = 250,000$ whose discrepancy in the head coordinates u lies within the tolerance ϵ fixed in advance. Black boxes mark the ranges predicted in the preregistration, so a marker outside its box shows a failed prediction. (b) Fraction of pairs whose G_s discrepancy lies below a threshold, against this threshold over ϵ , for u (solid) and the outputs y (dashed). Grey lines at the fractions 69/96 and 92/96 in (a) and (b) mark the decision levels, and the vertical line in (b) marks ϵ . (c) Excess $\Delta(u) - \Delta(y)$ of the discrepancy of u over that of y per pair of runs at both sample sizes, with median bars and the training procedures ordered and colored as in (a). Panels (b) and (c) are exploratory, and Table R.17 gives every decision.

Rep.	$T_{N,10}$	T_O	T_E	T_{MST}	$SOG_{0.1}$	$T_{G/Aff}$	G_s	1 - CKA	$\Delta\Pi$
y and u	0.8399	0.2463	0.2433	0.9541	0.9659	0.1463	0.2073	0.0833	0.0923
h	0.7404	0.1787	0.1775	0.8793	0.9054	0.1577	0.3032	0.0643	0.7040
h, ℓ_2	0.7145	0.1736	0.1599	0.8578	0.9014	0.0610	0.1137	0.0552	0.7042

Table R.16: Tolerances of the first network experiment, fixed in advance as the largest discrepancy of each representation over 24 pooled pilot pairs at $N = 100,000$ (eight per training procedure). Rows give the tolerances for the outputs y , the trunk features h , and h after ℓ_2 normalization, and the head coordinates u are judged against the tolerance of y .

Sensitivity to the Tolerance. Table R.18 gives the fraction of passing pairs at 0.9, 1, and 1.1 times the tolerance, and Table R.19 the counts under the preregistered secondary tolerance, the 0.9 quantile of the pilot discrepancies. Without weight decay, the G_s discrepancy of u nearest to the tolerance

Arm	N	Rep.	$T_{N,10}$	T_O	T_E	T_{MST}	$SOG_{0.1}$	$T_{G/Aff}$	G_s	$1 - CKA$	Δ_{Π}
noWD	25k	y	0 U	0 U	0 U	0 U	0 U	0 U	0 U	0 U	0 U
		u	0 U	0 U	0 U	0 U	0 U	0 U	0 U	0 U	0 U
		h	0 U	31 U	1 U	0 U	0 U	0 U	0 U	0 U	0 U
		h, ℓ_2	0 U	2 U	4 U	0 U	0 U	0 U	0 U	0 U	0 U
WD	25k	y	0 U	0 U	0 U	0 U	0 U	0 U	0 U	0 U	0 U
		u	0 U	0 U	0 U	0 U	0 U	0 U	0 U	0 U	0 U
		h	0 U	0 U	0 U	0 U	0 U	0 U	96 R	0 U	0 U
		h, ℓ_2	0 U	0 U	0 U	0 U	0 U	0 U	0 U	0 U	0 U
small init	25k	y	0 U	0 U	0 U	0 U	0 U	0 U	0 U	0 U	0 U
		u	0 U	0 U	0 U	0 U	0 U	0 U	0 U	0 U	0 U
		h	0 U	43 U	37 U	0 U	0 U	0 U	0 U	0 U	0 U
		h, ℓ_2	0 U	1 U	1 U	0 U	0 U	0 U	0 U	0 U	0 U
noWD, shared	25k	y	0 U	0 U	0 U	0 U	0 U	0 U	0 U	0 U	0 U
		u	0 U	15 U	31 U	0 U	0 U	0 U	0 U	0 U	0 U
		h	0 U	96 R	96 R	0 U	0 U	70 I	42 U	53 U	0 U
		h, ℓ_2	0 U	86 I	89 I	0 U	0 U	85 I	96 R	86 I	0 U
noWD	250k	y	96 R	96 R	96 R	96 R	96 R	96 R	96 R	96 R	96 R
		u	96 R	96 R	96 R	96 R	96 R	95 R	65 U	73 I	96 R
		h	96 R	96 R	96 R	96 R	96 R	96 R	96 R	96 R	96 R
		h, ℓ_2	96 R	96 R	96 R	96 R	96 R	96 R	96 R	96 R	96 R
WD	250k	y	96 R	96 R	96 R	96 R	96 R	96 R	96 R	96 R	96 R
		u	96 R	96 R	96 R	96 R	96 R	96 R	96 R	96 R	96 R
		h	96 R	96 R	96 R	96 R	96 R	96 R	96 R	96 R	96 R
		h, ℓ_2	96 R	95 R	94 R	96 R	96 R	96 R	96 R	96 R	96 R
small init	250k	y	96 R	96 R	96 R	96 R	96 R	96 R	96 R	96 R	96 R
		u	96 R	96 R	96 R	96 R	96 R	96 R	82 I	86 I	96 R
		h	96 R	96 R	96 R	96 R	96 R	95 R	93 R	91 I	96 R
		h, ℓ_2	96 R	96 R	96 R	96 R	96 R	96 R	96 R	96 R	96 R

Table R.17: Full decision table of the first network experiment: pairs within tolerance out of 96, with the one-sided 95% Clopper–Pearson decision reliable (R, at least 92), inconclusive (I), or unreliable (U, at most 69). In the rows “noWD, shared”, the runs are trained on four shared corpora with 24 pairs each, so the pairs within a corpus are dependent. The reliable decision for h under weight decay at 25k on G_s comes from a tolerance set by one pilot pair trained with small initialization.

differs from the tolerance by 1.0×10^{-4} . Since the mean G_s discrepancy of u without weight decay lies just below the tolerance, the decision of this cell turns on the spread of the pairs around the tolerance. Under a normal model of the 24 pilot discrepancies, a fresh pilot of the same size would have given a tolerance large enough to make this cell inconclusive with probability 0.04 and reliable with probability below 10^{-6} (exploratory).

Exploratory Analyses. Table R.20 gives the median paired excess $\Delta(u) - \Delta(y)$ on each criterion, and Table R.21 splits the excess on G_s and CKA into a part produced by a gauge common to both runs of a pair and a part produced by their mismatch. We chose these analyses after seeing the preregistered decisions. The mean excess on G_s without weight decay fell from 0.0233 to 0.0200 between 25,000 and 250,000, a change of -0.0033 with 95% interval $[-0.0044, -0.0023]$, while the discrepancy of y itself fell from 0.275 to 0.185. At 25,000 and 250,000, the median relative imbalance $\|MM^\top - W^\top W\|_F / (\|MM^\top\|_F + \|W^\top W\|_F)$ of the head at the end of training is 0.27 and 0.31 without weight decay, 0.030 and 0.040 with weight decay, and 0.24 and 0.28 under small initialization. Table R.22 gives the condition number of W . In the shared-corpus cell, every mean discrepancy fell relative to independent subsets, by 0.033 on G_s for y , 0.039 for u , and 0.056 for h , with every 95% interval excluding zero.

Neighbor Sets under the Gauge. Corollary 2 certifies the neighbor set of an anchor whose ratio of the $(k+1)$ th to the k th smallest distance exceeds cond_{AB} , and an exploratory analysis of the stored runs counted the anchors covered by this certificate. We formulated the corollary after the results of the first network experiment, and the second network experiment tests it prospectively (Appendix R.12.1). Because the corollary is proved and its certificate holds under its hypotheses, we use the counts below to check our code and to count the anchors covered by the certificate, not to

Criterion	Arm	Rep.	$\hat{p}(0.9\varepsilon)$	$\hat{p}(\varepsilon)$	$\hat{p}(1.1\varepsilon)$	$\varepsilon_{\text{rel}}/\varepsilon$
G_s	noWD	u	0.00	0.68	1.00	1.05
	noWD	y	0.69	1.00	1.00	0.93
	WD	u	0.72	1.00	1.00	0.93
	WD	y	0.72	1.00	1.00	0.93
	small init	u	0.00	0.85	1.00	1.03
	small init	y	0.52	1.00	1.00	0.93
$1 - \text{CKA}$	noWD	u	0.04	0.76	1.00	1.04
	noWD	y	0.95	1.00	1.00	0.90
	WD	u	1.00	1.00	1.00	0.81
	WD	y	1.00	1.00	1.00	0.79
	small init	u	0.19	0.90	1.00	1.01
	small init	y	0.99	1.00	1.00	0.88
$T_{N,10}$	noWD	u	0.00	1.00	1.00	0.98
	noWD	y	0.00	1.00	1.00	0.98
	WD	u	0.00	1.00	1.00	0.96
	WD	y	0.00	1.00	1.00	0.96
	small init	u	0.00	1.00	1.00	0.97
	small init	y	0.00	1.00	1.00	0.97

Table R.18: Fraction of the 96 pairs at $N = 250,000$ whose discrepancy lies within 0.9, 1, and 1.1 times the tolerance fixed in advance, and the ratio $\varepsilon_{\text{rel}}/\varepsilon$ of the smallest tolerance giving a reliable decision to this tolerance.

Arm	Criterion	$\varepsilon_{0.9}$	pairs at $\varepsilon_{0.9}$	pairs at ε
noWD	G_s	0.2030	39 U	65 U
noWD	$1 - \text{CKA}$	0.0810	45 U	73 I
WD	G_s	0.2030	96 R	96 R
WD	$1 - \text{CKA}$	0.0810	96 R	96 R
small init	G_s	0.2030	58 U	82 I
small init	$1 - \text{CKA}$	0.0810	79 I	86 I

Table R.19: Pairs out of 96 at $N = 250,000$ whose discrepancy in the head coordinates u lies within the preregistered secondary tolerance $\varepsilon_{0.9}$ or within the primary tolerance ε , with decisions as in Table R.17. The secondary tolerance is the 0.9 quantile of the pilot discrepancies.

Arm	N	$T_{N,10}$	T_O	T_E	T_{MST}	$T_{G/\text{A}\#}$	G_s	$1 - \text{CKA}$
noWD	25k	0.0000 (50)	-0.0260 (0)	-0.0244 (0)	0.0000 (40)	0.0135 (95)	0.0233 (96)	0.0121 (96)
WD	25k	-0.0002 (40)	-0.0150 (0)	-0.0138 (0)	0.0000 (42)	-0.0135 (0)	-0.0042 (0)	-0.0062 (0)
small init	25k	0.0001 (49)	-0.0310 (0)	-0.0296 (0)	0.0000 (43)	-0.0011 (41)	0.0165 (96)	0.0043 (89)
noWD, shared	25k	-0.0001 (44)	-0.0267 (0)	-0.0246 (0)	-0.0005 (36)	0.0068 (88)	0.0173 (96)	0.0081 (96)
noWD	250k	-0.0011 (24)	-0.0193 (0)	-0.0230 (0)	-0.0017 (33)	0.0100 (96)	0.0198 (96)	0.0111 (96)
WD	250k	-0.0010 (24)	0.0002 (50)	-0.0018 (23)	-0.0007 (30)	0.0021 (87)	-0.0001 (44)	0.0026 (96)
small init	250k	-0.0010 (31)	-0.0190 (0)	-0.0210 (0)	-0.0020 (34)	0.0074 (91)	0.0158 (96)	0.0091 (96)

Table R.20: Median paired excess $\Delta(u) - \Delta(y)$ of the discrepancy of the head coordinates over that of the outputs in the 96 pairs of each cell, with the number of pairs in which $\Delta(u)$ exceeds $\Delta(y)$ in parentheses (exploratory).

Arm	N	Criterion	total excess	common gauge	mismatch
noWD	25k	G_s	0.0233	0.0198	0.0035
noWD	25k	$1 - \text{CKA}$	0.0121	0.0079	0.0042
WD	25k	G_s	-0.0042	-0.0052	0.0009
WD	25k	$1 - \text{CKA}$	-0.0063	-0.0071	0.0008
small init	25k	G_s	0.0166	0.0121	0.0045
small init	25k	$1 - \text{CKA}$	0.0046	0.0000	0.0046
noWD, shared	25k	G_s	0.0176	0.0137	0.0039
noWD, shared	25k	$1 - \text{CKA}$	0.0081	0.0041	0.0040
noWD	250k	G_s	0.0200	0.0148	0.0051
noWD	250k	$1 - \text{CKA}$	0.0113	0.0073	0.0040
WD	250k	G_s	-0.0002	-0.0005	0.0004
WD	250k	$1 - \text{CKA}$	0.0026	0.0025	0.0001
small init	250k	G_s	0.0156	0.0086	0.0070
small init	250k	$1 - \text{CKA}$	0.0093	0.0037	0.0056

Table R.21: Mean excess $\Delta(u) - \Delta(y)$ split into a common-gauge part, obtained by applying one run’s gauge to the outputs of both runs of a pair, and a mismatch part from the difference between the gauges of the two runs (exploratory, from an independent recomputation). Judged against the tolerance of this experiment, the output discrepancy plus the mismatch part lies within tolerance in 96 of 96 pairs in every cell at 250,000. The second network experiment (Appendix R.12.1) judges a mismatch part defined by the mean gauge of pilot runs under tolerances from the same training procedure and size.

Arm	N	median $\text{cond}(W)$	range
noWD	25k	1.55	1.45 to 1.65
WD	25k	1.16	1.12 to 1.22
small init	25k	1.63	1.48 to 1.76
noWD, shared	25k	1.53	1.44 to 1.67
noWD	250k	2.09	1.96 to 2.24
WD	250k	1.41	1.37 to 1.47
small init	250k	1.99	1.82 to 2.25

Table R.22: Condition number of the head matrix W at the end of training.

confirm a prediction. Two scripts written independently produced the analysis from the stored runs without retraining. Both compute cond_{AB} from a generalized eigenproblem, and in the network this value agrees with the squared condition number of $W_B^{-1}W_A$ to 2.1×10^{-15} . In the network, u is recomputed in float64 as $W^{-1}y$, since the stored u agrees with this product to within a relative 3.7×10^{-4} , the precision of TF32 arithmetic.

In the linear autoencoder at $N = \infty$, where both runs of a pair minimize the same population loss, every certified anchor kept its neighbor set at $k = 1$ and $k = 10$ in 3,972 pairs from eight training procedures. Counted over both directions of every pair, the certified anchors number 1,282,589, 1,007,192, and 963,155 at $k = 1, 10$, and 30, with 2 violations at $k = 30$. Both violations come from one anchor of one tied-weights pair, counted in both directions. The saved encoders of this pair leave the span of $\hat{U}$ by a relative 6×10^{-5} , and both violations disappear when the distances are computed from the projected minimizer. At finite N the two runs of a pair see different samples and lie outside the premises of the corollary, and certified anchors then change, for instance 6,877 of 12,288 for tied weights at $N = 2^8$.

The fraction of anchors covered by the certificate depends on how close cond_{AB} lies to one. Under weight decay, tied weights, and small initialization, $\text{cond}_{AB} - 1 < 0.008$, and the certificate covers at least 94% of the anchors at $k = 10$. From the default initialization, the median cond_{AB} is 8.9 under gradient descent at step 0.002, which stands in for gradient flow. The certificate then covers 0 anchors at $k = 10$, and 99% of the 10-nearest-neighbor sets differ between the two runs of a pair, with a mean $1 - s_N(10)$ of 0.35. In the model of Theorem 2 the gauge thus prevents agreement on neighbor sets together with agreement under CKA, and 0 of the 3,972 pairs met the neighbor tolerance while missing the CKA tolerance. Evaluated with the learned Ω_a , the inequality

$\text{cond}_{AB} \leq \text{cond}(\Omega_A) \text{cond}(\Omega_B)$ from the proof of the corollary holds in every pair of every training procedure. The bound through the initial imbalance in the proof of the corollary holds in 500 of 500 pairs of the step-0.002 arm, with a median of 22.3 against a median cond_{AB} of 8.9. Its weaker form $f(\lambda_0)^4$ has a median of 2,334. Outside gradient flow both expressions can fail, as Remark 1 allows. Under discrete gradient descent from a small initialization, cond_{AB} exceeds the first expression in 281 of 500 pairs and the second in 90, by at most 0.007, and under gradient descent at step 0.05 it exceeds the first expression in 1 of 486 pairs.

In the CIFAR-5m networks, which lie outside the premises of Theorem 2, the certificate covers few anchors and the gauge barely changes the decisions on neighbor sets. The condition number of W stays at most 2.25, and the median cond_{AB} lies between 1.14 and 1.45 across cells, so the certificate covers at most 5% of the anchors at $k = 10$. Counted over both directions of every pair, the certificate covers 1,074,449, 29,208, and 1,220 anchors under an exchange of the head gauge at $k = 1, 10$, and 30, with 0 violations, the count implied by construction. Applied to the outputs of one run, an exchange of the head gauge alone moves 2 to 6% of the 10-nearest-neighbor slots and changes at least one slot at 21 to 52% of the anchors. Without weight decay at 250,000 the outputs of two runs share a mean of 0.19 of their 10 nearest neighbors, and each anchor moved by the gauge is already moved by this difference between the learned functions. Added to that difference, the gauge changes the mean cross-run discrepancy on neighbor sets by at most 0.0002 in every cell. The decisions of this network on neighbor sets therefore reflect the learned functions and are not evidence for the corollary, and the same holds for the failed prediction P1.

Failed Predictions. P1 predicted that agreement of u without weight decay is unreliable on $T_{N,10}$ and G_s at both sizes. At 250,000, agreement on $T_{N,10}$ is reliable at 96 of 96 pairs, and agreement on G_s is unreliable at 65 of 96 with Clopper–Pearson bounds 0.590 and 0.756. P3 predicted reliable agreement of u under small initialization at 250,000, and agreement on G_s is inconclusive at 82 of 96 with lower bound 0.781. P5 predicted that the fractions of agreeing pairs under small initialization and under weight decay differ by at most 0.1, and the difference on G_s at 250,000 is 0.146; its other three comparisons are 0 against 0 or 96 against 96 pairs. The 25,000 parts of P1, P3, and P5 hold or fail vacuously, since at this size 0 of 96 pairs lie within tolerance for y and u on every criterion under the three training procedures with independent training subsets (Preregistered Decisions above).

R.12.1 Second Experiment: Head Flow and Same-size Tolerances

Appendix F.8.1 of the paper summarizes this subsection.

Design. The second network experiment trains fresh runs of the network above, with new seeds for every source of randomness and the head computed in float64. It sets each tolerance from pilot pairs of the same arm and size and separates a gauge shared by the runs of an arm from the run-to-run gauge. Five core arms have 96 main pairs per size and 96 pilot pairs (64 for FLOW). MOM trains the head with the Nesterov momentum of the trunk, as in the arm without weight decay above, and WD adds weight decay on the head. FLOW, BAL-FLOW, and SMALL-FLOW remove the head from the optimizer, while the trunk keeps its momentum. After each trunk step, these arms hold the trunk features fixed and integrate gradient flow of the head on the minibatch loss over the time of one momentum-equivalent step (the learning rate of the trunk divided by $1 - 0.85$). They start from the default initialization, from a balanced initialization with $C_0 = 0$, and from a tenfold smaller initialization. In an engineering phase under the frozen code, a frozen rule chose an adaptive Heun–Euler integrator with relative tolerance 10^{-4} . Three secondary arms with 32 main and 16 pilot pairs per size train the head with momentum from the balanced and the small initialization (BAL-MOM, SMALL-MOM) or with one plain Euler step per minibatch (SGD-K). At $N = 250,000$, MOM-BF16 repeats MOM with the bfloat16 head of the first experiment in 32 runs, and KS-4 and KS-16 split the Euler step into 4 and 16 substeps in 16 runs each. These three cells serve to estimate the drift of C . The balanced initialization gives W a condition number near 1.34 against about 26 for the default. SMALL-FLOW controls for this confound, since it keeps the default condition number at one tenth of the scale.

Tolerances and Decompositions. Each tolerance is the largest discrepancy of y over the pilot pairs of the same arm and size, and u is judged against this tolerance. As a sensitivity check, a pooled tolerance takes the largest discrepancy over the pilot pairs of all arms that pass the accuracy

gate at that size. A decision entry that is reliable under one of the two tolerances and unreliable under the other is reported as tolerance-sensitive, without a categorical claim. A decision on u requires reliable agreement of y in the same cell and is otherwise inconclusive, and the decision rule is that of Appendix R.1. Per pair and criterion, the excess $\Delta(u) - \Delta(y)$ splits in two ways. The common-gauge part is the excess of a counterfactual that gives every run of the arm one common gauge, and the mismatch part is the rest. This common gauge is the log-Euclidean mean of the determinant-normalized metrics $(WW^\top)^{-1}$ of the pilot runs. The balanced part is the excess of the coordinates $(WMM^\top W^\top)^{-1/4}y$ of the balanced head computing the same outputs, and the imbalance part is the rest. Its run-to-run component equals the mismatch part minus the mismatch part of the balanced coordinates. Every part is given in standard deviations (SD) of $\Delta(y)$ over the main pairs of the cell (Table R.23).

Order of the Experiment. The preregistration, the training, gate, and analysis code, and the results of unit and known-answer tests were hashed and frozen at 20:10:53 UTC on 2026-10-09, before any engineering, pilot, or main run. Every program refused to run on a changed hash. Pilot runs followed from 20:24 to 22:53 UTC, the gate file was written at 22:54:42, and the tolerances were frozen from pilot pairs alone at 23:38:49, during the training of the main runs. The main runs ended at 01:40:11 on 2026-10-10, the main measurements began at 01:43:56, after the main cohort was complete, and one analysis run finished at 03:02:39. All 4,352 runs completed, with 0 runs flagged as failed or diverged and 0 retries, in 21.35 GPU-hours including engineering, against an approved 31. An independent recomputation reproduced every tolerance, all 704 decision entries, and all 880 means of the decompositions without importing the analysis code. It also re-measured 30 pairs from the stored head weights to within 3.3×10^{-16} and confirmed this order in 15 of 15 checks. We wrote a second recomputation without reading the analysis code or the first recomputation, and it reproduced every preregistered decision and every estimate up to Monte Carlo error, as the preregistration requires before any claim.

Checks and Gates. The relative drift of C is $\|C_{\text{end}} - C_0\|_F$ divided by $\|MM^\top\|_F + \|W^\top W\|_F$ at the end of training, and the relative imbalance is $\|C\|_F$ at the end of training divided by the same sum. The pilot runs passed every gate, with a relative drift of C at most 2.79×10^{-5} under head flow and 0 diverged core runs. Their median relative imbalance was 0.273 and 0.314 under MOM and 0.030 and 0.040 under WD at the two sizes, as required by the pilot gate on the relative imbalance. On the main runs every check passed, with a largest relative drift of C under head flow of 2.75×10^{-5} against a limit of 2×10^{-4} . The exact drift identity of a plain Euler step held to 3.4×10^{-15} , and the median cosine between the final and the initial C under FLOW was 0.99999995. The distortion caused by the imbalance is the value of cond_{AB} between the coordinates u of a run and the coordinates of its balanced head. Two bounds on this distortion fixed in the preregistration held in the 2,337 of 2,368 runs where the first bound is non-vacuous and in all 1,152 runs of the flow arms, and a third bound held in every pair. The certificate of Corollary 2 had 0 violations, the count implied by construction, with its bound at or above the observed change at every anchor.

Estimates. Under FLOW, whose C is conserved, the final gauge lies closer to the run’s own initial gauge than to that of another run in 95 and 102 of 192 runs, and both 95% intervals contain 0.5. Since this statistic stays at the chance rate even when C is conserved, the conclusion of the first experiment on this memory rests on the cosine between the final and the initial C under momentum (C4a in Table R.1). The median relative drift of C is 2.4×10^{-5} and 2.5×10^{-5} under FLOW, 0.276 and 0.314 under MOM, and 0.343 and 0.376 under SGD-K, so a plain Euler step of the same size changes C slightly more than momentum. KS-4 and KS-16 give 0.137 and 0.028, a log-log slope of -0.935 in the number of substeps, and at 250,000 the drift under momentum is 0.76 times the extrapolated drift of a plain step. This ratio lies below the preregistered factor of 3 for naming momentum the main cause, so Remark 1 names both finite steps and momentum as causes of the change of C . The bfloat16 head gives a median relative imbalance of 0.3121 against 0.3119 in float64, so the ninefold growth of the first experiment is not an artifact of its precision.

Relative to the balanced start, the default start raises the imbalance part more under head flow than under momentum, with every difference-in-differences interval above zero. The distortion caused by the imbalance rises with the imbalance relative to the smallest singular value of the head. The Spearman correlations are 0.89 and 0.73 under FLOW for the initial value and 0.33 and 0.32 under MOM for the final value (unadjusted $p < 10^{-5}$). Under head flow the imbalance part is 0.19 and

0.18 times that under momentum on G_s at the two sizes and 0.39 and 0.48 times on $1 - \text{CKA}$. Weight decay removes the excess on G_s , at -0.91 and -0.04 SD, whereas on $1 - \text{CKA}$ the excess under weight decay is -1.65 SD at 25,000 and $+1.02$ SD at 250,000, with all 96 pairs above zero (Figure R.7). At 250,000 images the imbalance part under FLOW is positive in 98% of pairs (Figure 5e). Table R.23 gives the decomposition on G_s (A4). Under MOM at 250,000, agreement under

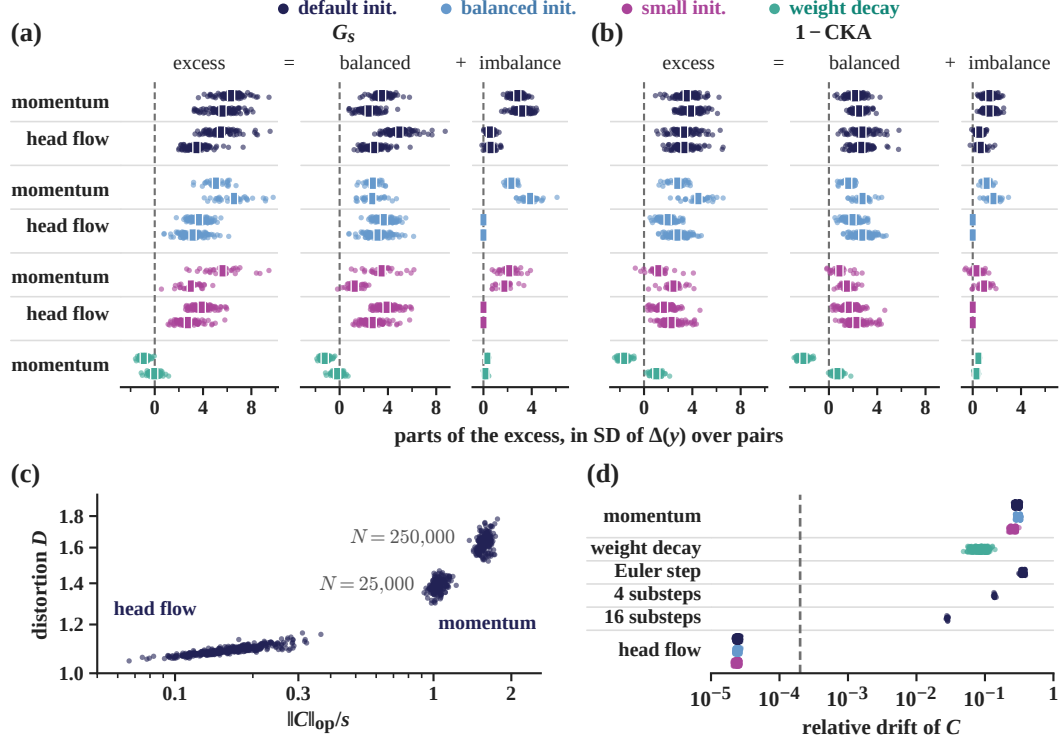


Figure R.7: **The initial imbalance kept by head flow contributes a smaller part of the excess discrepancy of the head coordinates than the imbalance created by momentum, and weight decay removes the excess on G_s (Theorem 2, Remark 1).** (a, b) Excess $\Delta(u) - \Delta(y)$ of each pair of runs in the second network experiment and its balanced and imbalance parts. Dots are pairs, bars means, and each row shows $N = 25,000$ above $N = 250,000$. (c) Distortion D caused by the imbalance in each run against the imbalance relative to the smallest singular value s of the head, initial under head flow and final under momentum (exploratory). (d) Relative drift of C per run against the limit set in advance for head flow (dashed). Appendix R.12.1 gives the tests.

the output discrepancy plus the mismatch part is inconclusive at 78 of 96 pairs, with a mismatch part of 1.63 SD (95% interval $[1.46, 1.80]$). Agreement of u under the tolerance of the common-gauge counterfactual is inconclusive at 90 of 96. Judged against its own pilot tolerance, agreement of u is reliable in every cell of MOM, WD, FLOW, BAL-FLOW, and SMALL-FLOW, a nearly tautological check, and these tolerances are 0.95 to 1.15 times those of y . The mean $s_N(10)$ of the trunk features differs across arms by at most 0.010 and 0.012. We also exchange the coordinates of each run for those of its balanced head. This imbalance-only exchange moves 0.052 and 0.070 of the 10-nearest-neighbor slots under MOM (K5), 0.014 and 0.016 under FLOW, and at most 0.0002 under BAL-FLOW and SMALL-FLOW.

Sensitivity to the Pooled Tolerance and a Collapsed Pilot Run. At 250,000, the tolerance-sensitive outcomes of predictions A3 and A7 (Table R.1) rest on one pilot pair of SMALL-FLOW, which sets the pooled tolerance on G_s (0.2098), $1 - \text{CKA}$, and $T_{N,10}$. Its G_s discrepancy lies 5.7 SD above the main pairs of the arm, although both runs are healthy (accuracy about 0.897). At the next pooled value on G_s (0.2017), 79 of 96 pairs of the common-gauge counterfactual of MOM and 72 of 96 pairs of u under BAL-FLOW lie within tolerance, and both decisions are inconclusive. The same pair sets the tolerance of SMALL-FLOW at 250,000, where 93 of 96 pairs of u lie within tolerance on G_s . This count falls to 52 of 96 at the second-largest pilot value of the arm and to 22

Arm	N	excess	common gauge	mismatch	balanced	imbalance	run-to-run imbalance
MOM	25k	6.32	5.09	1.23	3.51	2.81	0.44
	250k	5.61	3.98	1.63	2.41	3.21	0.76
WD	25k	-0.91	-1.19	0.28	-1.23	0.33	0.00
	250k	-0.04	-0.11	0.07	-0.21	0.17	0.00
FLOW	25k	5.48	3.99	1.50	4.95	0.54	0.29
	250k	3.45	1.74	1.71	2.86	0.59	0.46
BAL-FLOW	25k	3.66	2.85	0.82	3.66	0.00	0.00
	250k	3.14	1.95	1.19	3.14	0.00	0.00
SMALL-FLOW	25k	3.89	2.80	1.10	3.89	0.00	0.00
	250k	2.75	1.38	1.37	2.74	0.00	0.00
BAL-MOM	25k	5.07	4.28	0.79	2.76	2.31	0.30
	250k	6.56	4.79	1.78	2.70	3.86	0.96
SMALL-MOM	25k	5.61	4.01	1.60	3.48	2.13	0.59
	250k	2.99	1.53	1.46	1.25	1.73	0.78
SGD-K	25k	7.31	6.51	0.80	2.16	5.15	0.33
	250k	4.57	4.05	0.52	0.89	3.68	0.09

Table R.23: Mean decomposition of the excess $\Delta(u) - \Delta(y)$ on G_s in the second network experiment, in standard deviations of $\Delta(y)$ over the main pairs of each cell (96 pairs under MOM, WD, FLOW, BAL-FLOW, and SMALL-FLOW, and 32 under BAL-MOM, SMALL-MOM, and SGD-K). The excess equals the common-gauge part plus the mismatch part and also the balanced part plus the imbalance part, whose run-to-run component is listed last. Entries of 0.00 have absolute value below 0.005.

of 96 under the 0.9-quantile tolerance. We therefore do not read this cell as a selection of the gauge by small initialization. One SGD-K pilot run at 250,000 collapsed without triggering the divergence flag, with accuracy 0.089, $\|C\|_F$ near 7.6×10^7 , and a condition number of W near 2.1×10^5 . This run alone makes the tolerances of its cell vacuous (1.17 on G_s against 0.20 at the 0.9 quantile). It alone also triggers the accuracy flag that removed the cell from the pooled tolerance (mean accuracy 0.875, or 0.901 without this run). The primary endpoints exclude this cell, and the estimates of the drift of C in the Estimates paragraph use main runs, whose final condition number of W is at most 2.5.

R.13 Contrastive Objectives in the Finite Gibbs Model

Appendix F.9 of the paper summarizes this subsection.

Inside the premises of Theorem 3, all nine preregistered confirmatory predictions held after Holm correction, and by the preregistered rule the pattern of the theorem is present on the order criteria (Table R.24).

Design. The model has $|\mathcal{X}| = 256$ inputs with uniform masses, in 16 clusters of 16 points grouped into 4 superclusters, and the temperature is $t = 0.3$. Its K_* is built from a Gaussian kernel on fixed coordinates in $\mathbb{R}^8$ and adjusted so that the masses are balanced. The smallest eigenvalue of K_* is 0.129, and its off-diagonal distances are pairwise distinct. An example is an anchor, a positive partner, and one negative, and each run draws its own N examples as exact multinomial counts. Learner A minimizes the empirical InfoNCE risk over unit-diagonal positive semidefinite matrices. Learner B minimizes the empirical spectral contrastive risk over features of dimension at least $|\mathcal{X}|$ and reports their Gram matrix normalized to unit diagonal. Both return minimizers with numerical certificates of optimality, and a run whose certificate fails after one re-solve is kept and counted against the prediction of every test that uses the run. Each learner trains 192 runs at each of the sample sizes $2^{20}, 2^{22}, 2^{25}, 2^{28}, 2^{31}, 2^{34}, 2^{37}$, and 2^{40} , which form 96 disjoint blocks of two runs of each learner. Exact criteria compare the order of all distances among eight probe inputs, one in each of eight clusters, and the criteria of Appendix R.2 compare all 256 inputs. The tolerance of the largest difference of mean-normalized distances and of the isotonic discrepancy $\inf_g \|g(D_A) - D_B/\psi(D_B)\|_{\max}$, which is that of part 4 of the full version of Theorem 3 in units of $\psi(D_B)$, is half the limit $\varepsilon_\psi = 0.3477$ of part 3, with ψ the mean off-diagonal distance. These criteria, CKA, and ε_ψ take all 256 inputs as the probes of Theorem 3, and the exact criteria take the

eight probe inputs. Before the first training sample, the frozen code computed the Gaussian limits of both learners, with that of InfoNCE given by part 1 of Theorem E.1. From these limits, a Monte Carlo simulation predicted that the share of pairs across objectives with exact agreement on the order of the probes reaches 90% at $N^* = 2^{33.75}$ examples, so the grid is centered on 2^{34} . A secondary arm pairs learner A with InfoNCE at temperature $t/2$ at 2^{20} and 2^{22} examples. Positive-control pairs and planted controls check the pipeline on the link of the theorem applied to a run, on a pure rescaling, and on two planted breaks of the order.

Results. Exact agreement across objectives on the order of the probes held in 0 of 96 pairs up to 2^{28} examples, 7 of 96 at 2^{31} , 94 of 96 at 2^{34} , and 96 of 96 at 2^{37} and 2^{40} . The predicted shares were 0.073 at 2^{31} and 0.944 at 2^{34} . Within one objective the same agreement was 5 of 96 at 2^{31} for both learners and 96 and 89 of 96 at 2^{34} . From 2^{28} on, each eightfold increase of the sample shrank the recovery errors by factors of 2.77 to 2.91 for learner A and 2.80 to 2.87 for learner B, against $\sqrt{8} = 2.83$. Across objectives the largest difference of mean-normalized distances fell from a median 0.519 at 2^{20} to 0.348 at 2^{40} , and 0 of 96 pairs came within the tolerance of 0.174 at any size. Within one objective the same difference fell from 0.31 and 0.53 to 0.00058 and 0.00043. One minus CKA across objectives on all 256 inputs reached 0.14135, its limit, against 2.8×10^{-7} and 5.9×10^{-8} within objectives. With the link undone on the spectral run, the median difference fell from 0.606 at 2^{20} to 0.00497 at 2^{34} and 0.000618 at 2^{40} . Applied without a fit, the link of the theorem removes at least 0.99999992 of the squared-distance discrepancy in every pair at 2^{40} . At 2^{20} , the mean $s_N(10)$ was 0.806 and 0.735 within the two objectives and 0.763 across them, against 0.715 within InfoNCE in the networks. At this size the mean discrepancy across objectives stayed below the sum of the two mean discrepancies within objectives on every order criterion and on Δ_{iso} . Its ratio to the mean discrepancy within objectives was 1.01 to 1.03 on the order criteria and 1.14 on Δ_{iso} , against 3.2 in the networks. The InfoNCE minimizers lie on the boundary of the positive semidefinite cone up to 2^{25} examples and in its interior at every larger size of the grid. At 2^{20} and 2^{22} , where the rescaling of Remark 2 does not apply, the mean discrepancy across the temperatures t and $t/2$ exceeded the mean discrepancy within each temperature by at most 6.1% on every criterion (descriptive). In Figure 6a and c, the grey band runs between the values within InfoNCE and within the spectral loss. Figure 6b shares the axes and color scale of Figure R.8b and shows the pair at 2^{40} examples that ranks 48th of 96 in increasing order of its difference after the link is undone, a choice fixed before plotting.

Secondary Failures and Certificates. The noise ratio is the mean discrepancy across objectives divided by the mean of the two mean discrepancies within objectives. Its Monte Carlo prediction at 2^{40} was 0.557 against an observed 0.973 on neighbor sets and 0 against 1.111 on spanning trees. Both lie outside the factor of 1.5 required by the plan, and we attribute the first failure to the Monte Carlo error of 48 simulated blocks. Of the 3,456 main solves, 44 narrowly missed the tolerance 10^{-6} of their optimality certificate, by relative amounts up to 7.8×10^{-5} and with duality-gap bounds of at most 9.1×10^{-7} . Thirty-nine of them are InfoNCE runs at 2^{25} , a size without a test, and five belong to the secondary arm at 2^{22} . Counted against the descriptive conclusion of that arm, these five runs leave the conclusion unchanged. All 44 runs were kept and flagged.

Order of the Experiment. The preregistration and the 37 code files of this experiment were hashed and frozen at 20:06:42 UTC on 2026-10-09, before any computation, and every entry point refused to run on a changed hash. A first launch stopped during the pilot solves, when another job on the shared GPU slowed them about tenfold. Under amendment A1, we resumed the experiment on a free GPU and excluded the slowed timings from the frozen projection of GPU time. The main phase then stopped at a soft limit of 11.5 GPU-hours with 156 runs of the secondary arm left. We had set this limit and a hard limit of 12 GPU-hours ourselves when writing the preregistration. Amendment A2 was frozen at 08:32:28 UTC on 2026-10-10 and removed both limits, after which the remaining runs ran to completion. Two programming errors in the frozen analysis code then stopped the analysis, a cache that lost its newest matrix on overflow and a division by a Monte Carlo prediction equal to zero. Amendment A3 fixed the cache without changing any computed value, and the seven cache files left by the crash equal their recomputation byte for byte. Amendment A4 defined the one secondary comparison that divided by zero, a boolean outside every decision. We wrote and froze A3 and A4 during the analysis run. The preregistration allows such code fixes when they are recorded before the affected data are opened. The file metadata do not show whether the 418 pair caches written before the A4 freeze were opened before this freeze, and every decision is independent of

Item (source)	Preregistered prediction	Outcome
P1, learner A (part 1)	Recovery error falls from 2^{34} to 2^{40} by a ratio below 0.35	Confirmed; 0.124 [0.122, 0.126], against 0.125 at the rate $N^{-1/2}$
P1, learner B (part 1)	As for learner A	Confirmed; 0.126 [0.121, 0.130]
P2 (part 2)	Exact agreement across objectives on the global order of the eight probes reliable at 2^{40}	Confirmed; 96 of 96, lower bound 0.969; 96 of 96 within each objective
P3a (part 3)	Agreement across objectives within 0.174 on the largest difference of mean-normalized distances unreliable at 2^{34} and 2^{40}	Confirmed; 0 of 96 at both sizes; 96 of 96 within each objective
P3b (part 3)	$1 - \text{CKA}$ across objectives within 0.01 of its limit 0.1414	Confirmed; 0.14135
P3c (part 3)	With the link undone, the difference across objectives converges from 2^{34} to 2^{40}	Confirmed; ratio 0.124 [0.121, 0.127]
P4a (part 4)	Agreement across objectives within 0.174 under the isotonic discrepancy reliable at 2^{40}	Confirmed; 96 of 96
Order excess (model)	Mean discrepancy across objectives at most the sum of the two mean discrepancies within objectives on neighbor sets, per-anchor rankings, and the global order, at 2^{40}	Confirmed; excess below zero on all three, ratio 0.97 to 1.00; spanning trees (secondary) 1.11
Geometry excess (part 3)	Mean discrepancy across objectives above that sum on G_s and $1 - \text{CKA}$	Confirmed; excess 0.113 and 0.141
Holm	The nine items above at family level 0.05	9 of 9 rejected, p from 7.9×10^{-68} to 5×10^{-5}
Positive controls and planted breaks (design)	Link of the theorem and a rescaling recognized; two planted breaks detected	All match; the link applied to a run's own Gram matrix 96 of 96 at every size, rescaling within 4.4×10^{-16} , both planted breaks detected
Thm. E.1 (secondary)	Gaussian limit of part 1 on 16 frozen contrasts (P1b); transition of the agreement within InfoNCE (P2b)	Both pass; P1b 192 of 192 checks
Noise ratio, MC (secondary)	Monte Carlo ratio at 2^{40} within a factor of 1.5 of the observed ratio	Failed ; 0.557 against 0.973 (neighbor sets), 0 against 1.111 (spanning trees)
Matched noise (model, descriptive)	At 2^{20} , mean order discrepancy across objectives at most the sum of the two mean discrepancies within objectives	Holds; ratio 1.01 to 1.03, Δ_{iso} 1.14
Half temperature (model, descriptive)	Ratio of the mean discrepancy across the temperatures t and $t/2$ to the mean discrepancy within each temperature at 2^{20} and 2^{22} , reported descriptively	Ratio 1.00 to 1.06 on six criteria; five failed certificates counted against
Certificates	Every solve certified; failures kept and counted against	44 of 3,456 narrowly failed (39 at 2^{25} , 5 in the secondary arm at 2^{22})

Table R.24: Items of the finite Gibbs experiment with their preregistered predictions and outcomes, at 2^{40} examples unless stated, with failed items in bold. Parts refer to the full version of Theorem 3 (Appendix E.5), and “model” marks predictions computed from the finite model rather than stated by the theorem. The excess is the mean discrepancy of pairs across objectives minus the sum of the two mean discrepancies of pairs within one objective, and the ratio divides the first by the mean of the other two. Intervals are 95% block-bootstrap intervals over the 96 blocks.

this question, since A4 changes one secondary boolean. The results were written at 09:31:58 UTC on 2026-10-10.

Independent Recomputation. A recomputation written without importing any code of the experiment re-derived 4,416 pairs, 3,072 recovery errors, and 384 projections of the planted controls with its own solver (KKT residual at most 10^{-13}). It equals the frozen values bitwise on the neighbor, order, and spanning-tree discrepancies, on $1 - \text{CKA}$, and on the isotonic discrepancy of part 4 of the full version of Theorem 3, and to within 7×10^{-16} on the scale-normalized differences and 3×10^{-17} on Δ_{iso} . On the planted controls, whose frozen projections are accurate to 6×10^{-8} , it agrees to within 1.2×10^{-7} , and it reproduces every decision. Drawn with its own seeds, its bootstrap intervals differ in the fourth or fifth significant digit. A check of 53 file creation times against the event log confirmed the order above, and a separate recomputation of about 80 plotted values of Figure 6a–c from the stored Gram matrices found every value equal.

Compute and Scope. This experiment used 12.71 GPU-hours, 2.71 of them for the Gaussian limits and the rest for the solves of learner A, against an estimate of about 8 GPU-hours made when the plan was frozen. The frozen projection underestimated the GPU time of the solves at 2^{25} about fivefold, since the InfoNCE minimizer lies on the boundary of the cone at this size. Learner B and the analysis ran on the CPU. One preregistered gate used the rescaling identity of Remark 2 as a check of the solver, with a residual of at most 1.1×10^{-16} on the 32 samples where its premises hold. This check is not a test of the remark. The preregistration also fixes a second stage that removes one premise of the theorem at a time; we have not run this stage. A departure that breaks the pattern

inside the model would show that the departure suffices there, not that it causes the failed prediction in the networks of Appendix R.14.

R.14 Contrastive Networks

Appendix F.10 of the paper summarizes this subsection.

In our preregistration, we extended the prediction of Theorem 3 to CIFAR-5m networks, where learners trained with InfoNCE and with the spectral contrastive loss would learn one relation through different increasing links. Their representations would then agree on every order criterion and disagree on scale-normalized geometry. The trunk is the airbench94 network (Jordan, 2024) with features $h \in \mathbb{R}^{256}$, followed by a projection head to a 128-dimensional embedding. InfoNCE uses a cosine critic at temperature 0.2 (arm A) and 0.5 (arm A'), and the spectral contrastive loss (arm B) uses the ball projection of HaoChen et al. (2021) with unit parameter. All arms share stochastic gradient descent with Nesterov momentum, weight decay 5×10^{-4} , and batches of 512 images with two views each. Each arm trains 64 runs on $N = 50,000$ and 500,000 images at a budget of five million images seen, so 50,000 images mean 100 epochs and 500,000 images 10. Embeddings are compared with the cosine kernel on 2048 held-out images. The tolerance of each criterion is the largest discrepancy over 24 pooled within-arm pilot pairs at $N = 150,000$, and pairs within an arm (32 per arm) and across arms (64) are disjoint. As a check of the pipeline, A runs are paired with the exponential link $\exp(5(K - 1))$ of their own Gram matrix (self, 64 pairs) and of the Gram matrix of another A run (peer, 32 pairs). Table R.25 gives every decision. Figure R.8 compares InfoNCE at temperature 0.2 with the spectral loss at 500,000 images. Its panel (a) adds the projector distance Δ_{Π} and the isotonic discrepancy Δ_{iso} of Appendix R.2 to the criteria of Section 2, with grey boxes spanning all pairs within the three arms. Among the 64 pairs of an InfoNCE run at 0.2 and a spectral run at this size, Figure R.8b shows the pair that ranks 32nd of 64 in increasing order of its warp share, a choice fixed before plotting.

Outcome. These networks lie outside the premises of the theorem, since the unconstrained spectral loss diverged at every tried learning rate and the spectral arm therefore trains on the unit sphere. The embedding dimension also falls far below the size of the input domain. With two augmented views as each other's only positive partners and other images as negatives, the empirical InfoNCE risk has no positive definite minimizer at any temperature (Appendix E.5). All 64 pairs of InfoNCE and spectral runs fall outside the tolerance on neighbor sets, per-anchor rankings, the global order, and spanning trees at both sizes. The tolerance would have to grow 2.53-fold for agreement under every order criterion to become reliable. At 500,000 images their median discrepancies on these criteria are 1.34 to 2.51 times the tolerance, whereas pairs with a common objective meet every criterion (Figure R.8a). On all ten criteria of Figure R.9, the discrepancy of every pair across objectives exceeds that of every pair within an objective, at both sizes and for both temperatures. At 50,000 images the median discrepancy across objectives on the criteria other than Δ_{Π} is 1.21 to 6.30 times the tolerance for InfoNCE at 0.2 and 1.19 to 2.78 times for InfoNCE at 0.5. At this size, the preregistered requirement of reliable agreement within each objective fails on four criteria for InfoNCE at 0.2 and on six for the spectral loss. At 500,000 images, $s_N(10)$ is 0.313 across objectives against 0.715 and 0.640 within the two objectives. The isotonic discrepancy across objectives is 3.2 times its value within an objective, against a limit of 1.5. Applied without a fit, the link of the theorem explains a share of -0.04 of the discrepancy of the median pair (Figure R.8b). By our preregistered rule, the link prediction is falsified for these networks. Theorem 3 remains a statement about the finite Gibbs model, where its pattern appears at 2^{34} , 2^{37} , and 2^{40} examples.

Link Checks. Our analysis pipeline detects an exact link, since every pair of an InfoNCE run and the exponential link of the theorem applied to its own Gram matrix agrees on every order criterion and disagrees on every geometric criterion (diamonds in Figure R.8a,c). The median warp share is 1.00 for the link of the theorem applied to a run's own Gram matrix and 0.66 and 0.75 at the two sizes for the link applied to another InfoNCE run. Applied to the Gram matrix of another InfoNCE run, the link leaves the discrepancy under every order criterion at its value for the two InfoNCE runs and moves the discrepancy under every geometric criterion outside the tolerance (Figure R.9a,b). At 50,000 images, every median cumulative share of Figure R.9c lies within 0.004 of its value at 500,000 images, except for the link applied to another InfoNCE run, which explains 0.86 against

0.90 at the global monotone step. In every panel of Figure R.9, the range from the 5th to the 95th percentile over pairs is narrower than the mark of its median.

Temperatures. At each size, InfoNCE runs agree on neighbor sets in 32 of 32 pairs within each temperature. All 64 pairs of InfoNCE runs across the two temperatures lie within the tolerance on Δ_{Π} at both sizes and outside the tolerance on every other criterion, with median discrepancies of 1.05 to 3.58 times the tolerance at 500,000 images. Between these temperatures, a fitted monotone link removes a median 11% of the residual left by an affine map. For these pairs, the scale, affine, and global monotone fits together explain a median 0.39 of the squared-distance discrepancy, against 0.62 for pairs of InfoNCE at 0.2 and the spectral loss. The two InfoNCE arms therefore lie outside the rescaling case of Remark 2, which remains a statement about exact minimizers over unit-diagonal positive semidefinite matrices.

N	Pair set	$T_{N,10}$	T_O	T_E	T_{MST}	$SOG_{0.1}$	$T_{G/Aff}$	G_s	1 - CKA	Δ_{iso}	Δ_{Π}
50k	InfoNCE at 0.2, within	32/32 R	0/32 U	0/32 U	32/32 R	32/32 R	0/32 U	32/32 R	0/32 U	32/32 R	32/32 R
	InfoNCE at 0.5, within	32/32 R	32/32 R	32/32 R	32/32 R	32/32 R	32/32 R	32/32 R	32/32 R	32/32 R	32/32 R
	spectral, within	1/32 U	32/32 R	32/32 R	7/32 U	2/32 U	32/32 R	19/32 U	32/32 R	19/32 U	0/32 U
	InfoNCE at 0.2 vs spectral	0/64 U	0/64 U	0/64 U	0/64 U	0/64 U	0/64 U	0/64 U	0/64 U	0/64 U	0/64 U
	InfoNCE at 0.2 vs 0.5	0/64 U	0/64 U	0/64 U	0/64 U	0/64 U	0/64 U	0/64 U	0/64 U	0/64 U	64/64 R
	InfoNCE at 0.5 vs spectral	0/64 U	0/64 U	0/64 U	0/64 U	0/64 U	0/64 U	0/64 U	0/64 U	0/64 U	0/64 U
	link check, self	64/64 R	64/64 R	64/64 R	64/64 R	0/64 U	0/64 U	0/64 U	0/64 U	64/64 R	n/a
	link check, peer	32/32 R	0/32 U	0/32 U	32/32 R	0/32 U	0/32 U	0/32 U	0/32 U	32/32 R	n/a
500k	InfoNCE at 0.2, within	32/32 R	32/32 R	32/32 R	32/32 R	32/32 R	32/32 R	32/32 R	32/32 R	32/32 R	32/32 R
	InfoNCE at 0.5, within	32/32 R	32/32 R	32/32 R	32/32 R	32/32 R	32/32 R	32/32 R	32/32 R	32/32 R	32/32 R
	spectral, within	32/32 R	32/32 R	32/32 R	32/32 R	32/32 R	32/32 R	32/32 R	32/32 R	32/32 R	32/32 R
	InfoNCE at 0.2 vs spectral	0/64 U	0/64 U	0/64 U	0/64 U	0/64 U	0/64 U	0/64 U	0/64 U	0/64 U	0/64 U
	InfoNCE at 0.2 vs 0.5	0/64 U	0/64 U	0/64 U	0/64 U	0/64 U	0/64 U	0/64 U	0/64 U	0/64 U	64/64 R
	InfoNCE at 0.5 vs spectral	0/64 U	0/64 U	0/64 U	0/64 U	0/64 U	0/64 U	0/64 U	0/64 U	0/64 U	0/64 U
	link check, self	64/64 R	64/64 R	64/64 R	64/64 R	0/64 U	0/64 U	0/64 U	0/64 U	64/64 R	n/a
	link check, peer	32/32 R	32/32 R	32/32 R	32/32 R	0/32 U	0/32 U	0/32 U	0/32 U	32/32 R	n/a

Table R.25: Full decision table of the contrastive networks on the embedding with the cosine kernel: pairs within tolerance and the one-sided 95% Clopper–Pearson decision (R reliable, I inconclusive, U unreliable; with 64 pairs, reliable needs at least 62 and unreliable means at most 45, and with 32 pairs, all 32 and at most 21). InfoNCE runs are labeled by their temperature. The link checks pair an InfoNCE run with the exponential link of Theorem 3 applied to its own Gram matrix (self) or to the Gram matrix of another InfoNCE run (peer).

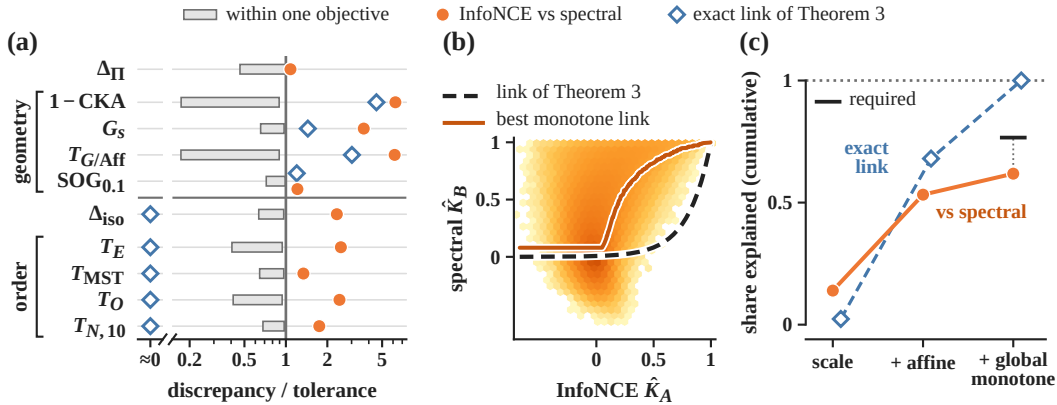


Figure R.8: In CIFAR-5m networks, InfoNCE and spectral learners disagree on the order of distances as well as on geometry, and a monotone link explains little beyond an affine map. (a) Discrepancy over the tolerance of each criterion; Theorem 3 predicts agreement on the criteria below the horizontal line, and pairs of an InfoNCE run and the exact link of the theorem applied to its Gram matrix (diamonds) show this pattern. (b) Density of kernel entries of one pair of runs. (c) Share of the squared-distance discrepancy explained by successive maps, with the share required by the preregistered prediction. Marks in (a) and (c) are medians over pairs. All panels are at $N = 500,000$, and each pair across objectives joins a run of InfoNCE at temperature 0.2 with a spectral run. Figure R.9 gives both sample sizes and both temperatures.

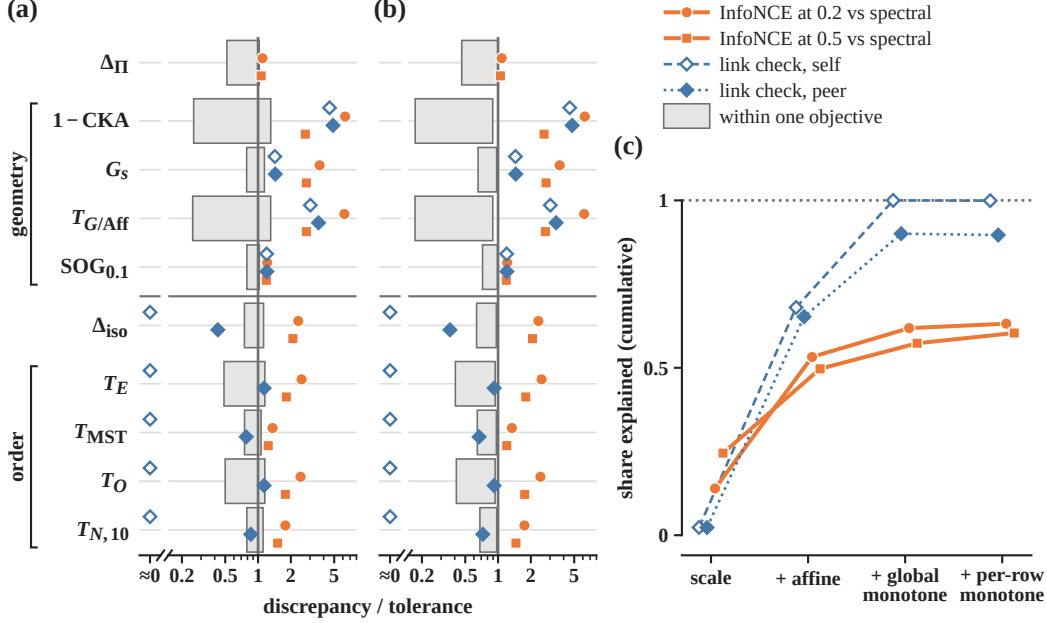


Figure R.9: In CIFAR-5m networks, runs trained with InfoNCE at either temperature and runs trained with the spectral loss disagree even on the order of distances at both sample sizes, whereas an InfoNCE run and the exact link of Theorem 3 applied to its own Gram matrix agree on the order of distances. (a, b) Discrepancy over tolerance per criterion at $N = 50,000$ (a) and $500,000$ (b), with agreement predicted on the criteria below the horizontal line and grey boxes spanning the pairs within one objective. Diamonds mark pairs of an InfoNCE run and the link of the theorem applied to its own Gram matrix (open) or to that of another InfoNCE run (filled). (c) Cross-fitted share of the squared-distance discrepancy explained by successive maps at $N = 500,000$. Marks are medians over pairs, and Table R.25 gives every decision.

Items (i) to (v) of the Link Prediction. Item (i) required agreement across objectives under each order criterion to come within 0.1 of the within-objective ceiling of Remark C.1 at $500,000$ and to improve with N . Every gap is 1.0, and the improvement from $50,000$ to $500,000$ is significant after Holm correction for $T_{N,10}$ (-0.0125 , 95% interval $[-0.0145, -0.0105]$) and the MST (-0.0119) but not for T_O ($p = 0.037$ against 0.025) or T_E ($p = 0.25$). Item (ii) required agreement across objectives under the geometric criteria to be unreliable at both N , with reliable within-objective agreement at $500,000$. The item passes by its letter, but agreement across objectives is unreliable under every criterion, so the item does not single out geometry. At $50,000$, the required reliable agreement within each objective fails for all three criteria. Item (iii) required the isotonic discrepancy across objectives to stay within 1.5 times its within-objective value, and the ratio is 3.198 (0.1767 against 0.0553). Item (iv) required a median warp share of at least 0.5 across objectives, and the median is 0.147 at $500,000$ (Figure R.8c) and 0.142 at $50,000$, against 0.012 and 0.002 within the two objectives. Item (v) required InfoNCE runs at the two temperatures to agree on G_s , CKA, and $\text{SOG}_{0.1}$, and every one of the 64 pairs fails all three, while its clause on raw distances passes.

Secondary Predictions. The fitted link $K_B \approx \exp(\beta(K_A - 1))$ has median exponent $\hat{\beta} = 2.04$ across objectives, outside the predicted interval $[2.5, 10]$, and the ratio of $\hat{\beta}$ between the two temperatures is 1.006 against a predicted $[1.5, 3.5]$. The co-occurrence prediction derived from Lemma E.1 stated that more data would shrink the log-distance error of near pairs faster than that of far pairs in both InfoNCE arms. In all three arms, more data shrinks the error of near pairs less than that of far pairs. The predicted ordering between the two temperatures holds, with a difference of 0.023 in the log ratio (interval $[0.010, 0.035]$).

Exploratory Analyses. Across objectives at $500,000$, the mean $s_N(10)$ of 0.313 (Outcome above) lies above its value of 0.164 between two networks at initialization. The global-order discrepancy

across objectives (0.466) exceeds the discrepancy between two networks at initialization (0.186). The participation ratio of the centered Gram matrix is 67.7 for InfoNCE at temperature 0.2, 27.9 at temperature 0.5, and 11.7 for the spectral arm. On four pairs of runs, an exponential link of the InfoNCE Gram matrix raises this ratio above 300, so the link predicts variance spread over many directions, whereas the spectral embedding concentrates its variance on few directions. Two explanations remain open, a change of rank or spectrum in the sense of the model-class row of Table E.1 and a common path of uniformity temperatures for the three arms. The uniformity temperature of a contrastive loss is the temperature t that weights the term spreading the embeddings over the sphere against the term pulling the two views of an image together. A heuristic second-order expansion in $1/t$ makes the spectral contrastive loss of unit-norm embeddings resemble InfoNCE at $t = 1$ without its penalty on the mean embedding, so the three arms may lie at $t = 0.2, 0.5$, and about 1 on one path. These runs do not separate the two explanations.

References

- Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR, 2023. URL <https://proceedings.mlr.press/v202/biderman23a.html>.
- Xinyang Geng and Hao Liu. OpenLLaMA: An open reproduction of LLaMA. GitHub repository, May 2023. URL https://github.com/openlm-research/open_llama.
- Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the platonic representation hypothesis: An aristotelian view. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL <https://arxiv.org/abs/2602.14486>.
- Jeff Z. HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL <https://arxiv.org/abs/2106.04156>.
- Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 20617–20642. PMLR, 2024. URL <https://arxiv.org/abs/2405.07987>.
- Keller Jordan. 94% on CIFAR-10 in 3.29 seconds on a single GPU, 2024. URL <https://arxiv.org/abs/2404.00498>.
- Elizaveta Levina and Peter J. Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems*, volume 17, 2004.
- Preetum Nakkiran, Behnam Neyshabur, and Hanie Sedghi. The deep bootstrap framework: Good online learners are good offline generalizers. In *International Conference on Learning Representations*, 2021. URL <https://arxiv.org/abs/2010.08127>.
- Dominik Schnaus, Thomas Dagès, Daniel Cremers, Xi Wang, and Phillip Isola. Shared geometry as a Rosetta stone: Cross-modal alignment without paired data, 2026. URL <https://arxiv.org/abs/2610.09411>.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models, 2023. URL <https://arxiv.org/abs/2302.13971>.
- Junwon You, Mihyun Jang, Sangwoo Mo, and Jae-Hun Jung. What converges in the platonic representation hypothesis? structure over geometry, 2026. URL <https://arxiv.org/abs/2609.27252>.