Geometric Drift Metrics are Insufficient: A Matched-Magnitude Dissociation Between Aligned and Anti-Aligned Fine-Tuning

Abdul Basit Tonmoy*
Wabash College

May 12, 2026

Abstract

Geometric similarity metrics—Centered Kernel Alignment, neighborhood preservation, isotropy—are widely used to compare neural representations and infer functional similarity. We show this inference fails in two complementary directions. On E5-base-v2, we hold geometric drift fixed (Neighborhood Preservation Score ≈ 0.48, CKA ≈ 0.83) and vary only the fine-tuning objective: contrastive LoRA on MS MARCO leaves retrieval within base 95% CI, while masked-language-modeling LoRA on the *same corpus* drops MS MARCO nDCG@10 by 10.2 ± 0.8 pp and NFCorpus by 12.7 ± 0.8 pp across three seeds. A matched-Frobenius Gaussian perturbation at comparable NPS and CKA produces only -1.3 / -6.6 pp. The same geometric drift magnitude thus corresponds to preserved, modestly damaged, or substantially damaged downstream behavior depending on the objective. On BERT, the dissociation reverses direction: MLM fine-tuning barely moves the geometry (NPS 0.836 ± 0.002) yet produces a markedly different functional outcome from contrastive LoRA on the same corpus (+1.17 vs. +15.95 pp NFCorpus gain, CI-disjoint), because BERT's pretraining objective is itself MLM. Geometric drift magnitude predicts neither the sign nor the scale of functional change without knowledge of the gradient–pretraining alignment. We additionally document a probe-methodology artifact: linear probes used for embedding evaluation are dominated by probe-convergence artifacts unless trained on the full available data with cross-validated regularization, producing accuracy swings up to 28.5 pp at *identical* embeddings. We argue that label-free geometric metrics should be treated as change detectors rather than functional- similarity proxies, and that downstream effect estimation requires either calibration data or knowledge of the gradient–pretraining alignment.

1 Introduction

Geometric similarity metrics—Centered Kernel Alignment [Kornblith et al., 2019], neighborhood-preservation statistics [Venna et al., 2010], isotropy measures [Mu and Viswanath, 2018]—are foundational tools for comparing neural representations. They are widely used in interpretability research to assess representational similarity across model variants, training stages, and architectures, and to draw inferences about functional similarity from observed geometric similarity. The same metrics underpin a class of *label-free drift-detection* methods that flag representation change as a proxy for downstream task degradation [Muennighoff et al., 2023], particularly when ground-truth evaluation data is unavailable for the deployment task.

The implicit assumption underlying both uses is that geometric similarity predicts functional similarity: if two representations are geometrically close, they should support similar downstream

*Correspondence: `atonmoy27@wabash.edu`. Code and pre-registrations: `https://github.com/abtonmoy/semantic-sentry`.

behavior; if they are geometrically distant, they should support different behavior. We show this assumption fails in two distinct ways. Holding geometric drift fixed and varying only the fine-tuning objective, the downstream effect on a fixed evaluation task ranges from indistinguishable from baseline to a 12 pp degradation. And on a different base model, varying the fine-tuning objective produces qualitatively swapped functional rankings between geometric drift and downstream outcome. A label-free geometric metric cannot distinguish these cases.

Contributions.

1. **A matched-magnitude dissociation** between gradient-aligned and gradient-anti-aligned fine-tuning at fixed Neighborhood Preservation Score (NPS) on E5-base-v2, demonstrated with three random seeds per condition and a pre-registered hypothesis.
2. **A second, structurally distinct dissociation on BERT.** On a pretraining-objective-aligned model, MLM fine-tuning barely shifts the geometry (NPS 0.836) yet produces a markedly different functional outcome from contrastive fine-tuning on the same corpus (+1.17 vs. +15.95 pp NFCorpus gain, CI-disjoint). Geometric drift magnitude underpredicts functional difference, the opposite failure mode from the E5 result. Together the two findings show that geometric drift magnitude is uninformative about functional outcome without knowledge of the gradient–pretraining alignment.
3. **A corpus control** ruling out distribution shift as the driver: masked-language-modeling LoRA trained on the *same corpus* contrastive LoRA was trained on damages retrieval by ∼10 pp.
4. **Two structural controls** ruling out alternative explanations: random rank-4 perturbation matched to LoRA's per-layer Frobenius norm, and full-rank fine-tuning matched to LoRA's loss.
5. **A methodological finding** for embedding evaluation: linear probes trained on small subsets with fixed regularization produce accuracy artifacts up to 28.5 pp at identical embeddings, dominating any fine-tuning signal below this threshold.

We do not claim the existence of objective-dependent fine-tuning effects is novel; the catastrophic forgetting and negative-transfer literatures have established this qualitatively [McCloskey and Cohen, 1989, Kirkpatrick et al., 2017]. The contribution here is the controlled *matched-magnitude* demonstration, its converse on a pretraining-aligned model, and the consequence for representation-similarity analysis: geometry-only metrics are change detectors, not functional-similarity proxies.

2 Related Work

Geometric similarity metrics in representation analysis. Geometric similarity metrics for representation comparison have a long history in the analysis of neural networks. CKA [Kornblith et al., 2019] and its variants [Nguyen et al., 2021] measure global structural alignment; neighborhood-based metrics [Venna et al., 2010, Lee and Verleysen, 2009] measure local preservation; isotropy measures [Mu and Viswanath, 2018, Ethayarajh, 2019] quantify spectral spread. These metrics are routinely used in interpretability research to argue functional similarity between models, training checkpoints, and architectural variants, and as label-free proxies for representation quality and stability. Our results do not contest their utility as *change detectors*; we contest their use as *predictors of functional similarity*, both in the sense of downstream task effect and in the sense of representational content inferences.

Fine-tuning, adaptation, and forgetting. The catastrophic-forgetting literature [McCloskey and Cohen, 1989, French, 1999] establishes that sequential fine-tuning damages prior task performance. Negative transfer [Wang et al., 2019] shows that auxiliary objectives can hurt rather than help a target task. Recent work on task arithmetic [Ilharco et al., 2023, Ortiz-Jiménez et al., 2023] treats fine-tuning updates as directional vectors in weight space, with task performance varying smoothly along these directions. Our finding is consistent with this literature; the contribution is the controlled matched-magnitude comparison against a label-free geometric metric, which no prior work we are aware of has directly evaluated, and the converse demonstration on a pretraining-aligned model.

LoRA and low-rank adaptation. LoRA [Hu et al., 2022] confines updates to a low-rank subspace, which has been argued to preserve base-model behavior on unrelated tasks [Biderman et al., 2024]. Our results show this preservation is contingent on the loss being gradient-aligned to the evaluation task. Low-rank confinement alone—demonstrated via random rank-4 perturbation matched to LoRA's Frobenius norm—is essentially inactive on downstream metrics, because random subspaces produce negligible drift. The relevant factor is not the rank of the update but the direction the gradient selects.

Linear probing methodology. Linear probing for representation evaluation has known sensitivities to training data size and regularization [Alain and Bengio, 2017, Hewitt and Liang, 2019]. We add a quantitative demonstration of how large the artifact can be in a standard MTEB-style setup—28.5 pp on Banking77 at identical embeddings—and a concrete recipe to avoid it.

3 Setup

Model. We use `intfloat/e5-base-v2` [Wang et al., 2022] as the base model for the matched-magnitude experiments. E5 is a transformer text encoder pretrained with a contrastive objective on a large corpus and is widely deployed for retrieval. We use `bert-base-uncased` [Devlin et al., 2019] as the second base model for the pretraining- objective-dependence experiment in §4.3. Where noted, we additionally replicate selected conditions on `openai/clip-vit-large-patch14` [Radford et al., 2021].

Drift metrics. We report two geometric similarity metrics in the main text. **Neighborhood Preservation Score (NPS).** For a fixed anchor set $\mathcal{A} = \{x_i\}_{i=1}^n$, let $\mathcal{N}_k^{(0)}(x_i)$ and $\mathcal{N}_k^{(1)}(x_i)$ be the k-nearest neighbors of x_i under the base and updated models. The Neighborhood Preservation Score is

$$\mathrm{NPS}_k(\mathcal{A}) = \frac{1}{n}\sum_{i=1}^{n} \frac{|\mathcal{N}_k^{(0)}(x_i) \cap \mathcal{N}_k^{(1)}(x_i)|}{k}.$$

NPS is in $[0, 1]$, equal to 1 iff every k-neighborhood is preserved. It is invariant under orthogonal rotation of either embedding space and under k to within a small range (± 4.6 pp across $k \in \{5, 10, 25, 50\}$ on our anchors; §A.2). We report $k = 10$ throughout. **Centered Kernel Alignment (CKA).** We report linear CKA [Kornblith et al., 2019] computed on the same anchor set as NPS, to test whether the dissociation depends on the choice of geometric similarity metric. We additionally compute an isotropy delta as a supplementary metric, reported in the supplementary code; it patterns identically with NPS and CKA.

Anchor set. The anchor set is fixed across all conditions on a given model and drawn from MS MARCO dev queries to ensure cross-condition NPS and CKA values are comparable. We document in §A.3 that NPS magnitude is anchor-dependent; all NPS and CKA values reported in the main text use the same anchor per model and are therefore directly comparable within-model.

Evaluation tasks. We evaluate on two retrieval benchmarks (MS MARCO dev nDCG@10 [Nguyen et al., 2016]; NFCorpus nDCG@10 [Boteva et al., 2016]) and one classification probe (Banking77 top-1 accuracy [Casanueva et al., 2020]). Confidence intervals are bootstrap 95% CIs over query/example resampling with $B = 1000$. We describe a probe specifically as "CI-disjoint" from base when its CI does not overlap the base-model CI.

Drift conditions. We construct the following drift conditions:

- **Contrastive LoRA on E5 (aligned).** LoRA rank-4 fine-tune on MS MARCO query-passage contrastive pairs (InfoNCE), 50 epochs, AdamW with cosine schedule, 3 seeds.
- **MLM LoRA on E5, Wikipedia (anti-aligned, OOD corpus).** Same architecture and recipe; loss replaced by masked-language modeling (15% masking, 80/10/10 split); corpus replaced by 2,000 Wikipedia paragraphs. 3 seeds.
- **MLM LoRA on E5, MS MARCO (anti-aligned, matched corpus).** Identical to MLM-Wikipedia except the corpus is the same MS MARCO passages used by Contrastive LoRA. 3 seeds. *Pre-registered.*
- **Gaussian noise on E5 (random).** Scale-relative perturbation $\mathcal{N}(0, (\sigma|w|)^2)$ at $\sigma \in \{0.005, 0.01, 0.02, 0.04, 0$ applied to base-model weights. We report $\sigma = 0.16$, which lands closest to the MLM NPS regime.
- **Random rank-4 on E5 (control).** For each LoRA layer, sample random A', B' and rescale so the per-layer Frobenius norm matches the contrastive-LoRA ΔW at epoch 50. 3 seeds.
- **Full fine-tune contrastive on E5 (control).** Full-rank update on MS MARCO contrastive pairs; same recipe as Contrastive LoRA except all parameters trainable. 3 seeds.
- **MLM LoRA on BERT, MS MARCO (cross-family).** Identical recipe to E5 MLM-on-MS-MARCO except the base model is `bert-base-uncased`. 3 seeds. *Pre-registered.* Evaluated on NFCorpus nDCG@10 (BERT is not a retrieval model; NFCorpus is the benchmark used by the BERT contrastive replication in Appendix B).

Pre-registration. The matched-corpus E5 MLM experiment, the multi-seed E5 contrastive LoRA experiment, the MLM training stability check, and the BERT MLM-on- MS-MARCO cross-family replication were pre-registered before runs in the project repository, including hypotheses, falsification criteria, and predicted outcomes. We refer to these documents at `experiments/*/PREREGISTRATION.md` in the supplementary code.

Table 1: Matched-magnitude dissociation on E5-base-v2. All conditions evaluated against the same E5-base-v2 checkpoint and the same MS MARCO dev anchor set; NPS and CKA are computed on identical embeddings per row. Multi-seed results reported as 3-seed mean ± std. CI-disjoint drops (95% bootstrap) marked with †.

Condition	**NPS**	**CKA**	**MS MARCO Δ**	**NFCorpus Δ**	**Regime**
Base (no drift)	1.000	1.000	0.0	0.0	—
Random rank-4 ($\|F\|$ matched)	0.973	0.9998	0.0	+3.4	control
Contrastive LoRA (3 seeds)	0.597 ± 0.013	0.884 ± 0.002	-2.1 ± 0.5	-1.4 ± 0.3	aligned
Full-FT contrastive (3 seeds)	0.569 ± 0.003	0.854 ± 0.003	-2.79 ± 0.23	-1.62 ± 0.19	aligned
Gaussian $\sigma = 0.16$	0.570	0.880	-1.3	$-6.6^{\dagger}$	random
MLM LoRA Wikipedia (3 seeds)	0.479 ± 0.002	0.820 ± 0.002	$-9.7 \pm 0.7^{\dagger}$	$-11.8 \pm 0.3^{\dagger}$	anti-aligned
MLM LoRA MS MARCO (3 seeds)	0.434 ± 0.009	0.807 ± 0.003	$-10.2 \pm 0.8^{\dagger}$	$-12.7 \pm 0.8^{\dagger}$	anti-aligned

4 Two Dissociations Between Geometric Drift and Functional Outcome

4.1 Matched-magnitude dissociation on E5

Figure 1 presents the central finding visually: panel (a) shows the four drift conditions sit in a comparable NPS band, while panel (b) shows their downstream retrieval drops span an order of magnitude. Table 1 reports the underlying numbers, including multi-seed standard deviations, confidence-interval status, and CKA values for every condition.

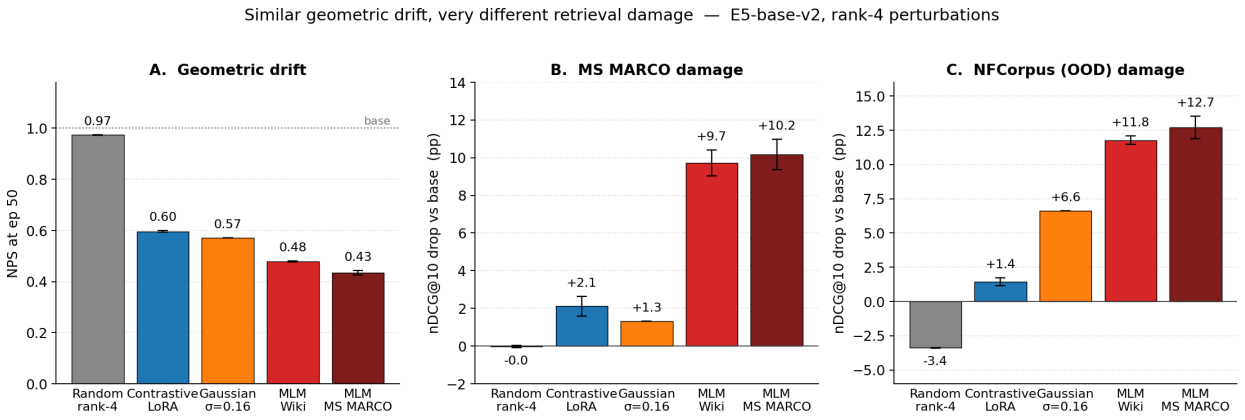

Figure 1: Matched-magnitude dissociation on E5-base-v2. **(a)** Neighborhood Preservation Score (NPS) across drift conditions; lower values indicate larger geometric drift. The four non-control conditions fall in a comparable NPS band. **(b)** Downstream retrieval drop on MS MARCO and NFCorpus for the same conditions. At comparable NPS, the downstream effect spans indistinguishable-from-baseline (Contrastive LoRA) to ∼13 pp (MLM LoRA on MS MARCO). NPS does not separate them. Error bars: 3-seed standard deviation where applicable.

The relevant comparison is between rows that sit in a comparable geometric-drift band. Gaussian $\sigma = 0.16$ sits at NPS 0.570 / CKA 0.880; Contrastive LoRA at NPS 0.597 / CKA 0.884; MLM LoRA Wikipedia at NPS 0.479 / CKA 0.820; MLM LoRA MS MARCO at NPS 0.434 / CKA 0.807. The MLM rows are at *slightly lower* NPS and CKA than the others, which is the conservative direction

for our claim: their downstream damage is larger despite their geometric drift being only modestly larger than Gaussian's. At matched-or-better NPS, random rank-4 preserves MS MARCO and produces a small (+3.4 pp) gain on NFCorpus (no-effect control); Contrastive LoRA preserves both retrieval benchmarks within base 95% CI; Gaussian $\sigma = 0.16$ preserves MS MARCO but produces a CI-disjoint −6.6 pp drop on NFCorpus; MLM LoRA on Wikipedia produces CI-disjoint drops on both benchmarks (−9.7 / −11.8 pp); and MLM LoRA on MS MARCO—the same corpus as Contrastive LoRA—produces equivalent damage (−10.2 / −12.7 pp). The four conditions cover an order-of-magnitude range in retrieval effect at a less than 0.16 NPS spread (and 0.08 CKA spread). Neither metric distinguishes them.

The dissociation is not metric-specific. A natural concern is that NPS, as a local neighborhood statistic, may miss global structural changes that a kernel-based metric like CKA would catch. Table 1 addresses this directly: CKA tracks NPS across all seven conditions, with the same rank order. The four perturbative conditions (Contrastive LoRA, Full-FT contrastive, Gaussian $\sigma = 0.16$, MLM-Wiki, MLM-MS-MARCO) cluster in CKA $\in [0.80, 0.89]$ while their MS MARCO damage spans 1 to 12 pp. The dissociation does not depend on which geometric similarity metric is used; it is a property of geometric similarity itself as a functional- similarity proxy. We additionally compute an isotropy delta (reported in the supplementary code) which patterns identically. This is the right outcome for the central claim: the failure to predict functional effect is not a limitation of any single metric but a structural property of the family of geometric similarity statistics.

4.2 Pre-registered corpus control

The most natural alternative explanation for "MLM damages retrieval" is distribution shift: Wikipedia is out-of-distribution relative to MS MARCO, and the model's representations adapt away from the retrieval evaluation distribution. We pre-registered an experiment to distinguish this explanation from an objective-driven one.

Method. We replace the Wikipedia corpus with the *same MS MARCO passages contrastive LoRA was trained on*. Every other variable is held constant: model, LoRA configuration, corpus size, learning rate, optimizer, schedule, epoch count, checkpoint schedule, anchor set, evaluation pipeline. Only the loss differs (MLM vs. contrastive InfoNCE).

Pre-registered hypotheses. (A) Objective-driven: MS MARCO retrieval drop ≥ 5 pp at epoch 50, supporting "the loss causes the damage regardless of corpus." **(B) Corpus-driven:** MS MARCO retrieval within base 95% CI, supporting "the previous result was about distribution shift." **(C) Mixed:** intermediate outcome between 1 and 5 pp.

Result. Hypothesis A is supported. MS MARCO nDCG@10 drops -10.2 ± 0.8 pp under MLM LoRA on MS MARCO passages, equivalent to the Wikipedia condition's -9.7 ± 0.7 pp. The realized drop exceeds the pre- registered 5 pp threshold by a factor of two, well outside the threshold-tuning concern. The damage is intrinsic to the objective. Distribution shift does not save it; identical corpora do not save it. The corpus control rules out the *class* of explanations that attribute MLM-induced damage to features of the training data; it does not rule out interactions between the objective and the corpus distribution of arbitrary fine-tuning data.

4.3 A second dissociation: pretraining-objective dependence on BERT

The E5 dissociation in §4.1 holds geometric drift fixed and varies the objective. We now run the converse experiment on a different model family to test whether the dissociation generalizes and to probe the role of the gradient–pretraining alignment. On `bert-base-uncased`, we apply the same MLM-on-MS- MARCO recipe used for the E5 anti-aligned condition (3 seeds, 50 epochs, LoRA rank-4, AdamW with cosine schedule). The result reveals a structurally different failure of geometric drift as a functional- similarity proxy.

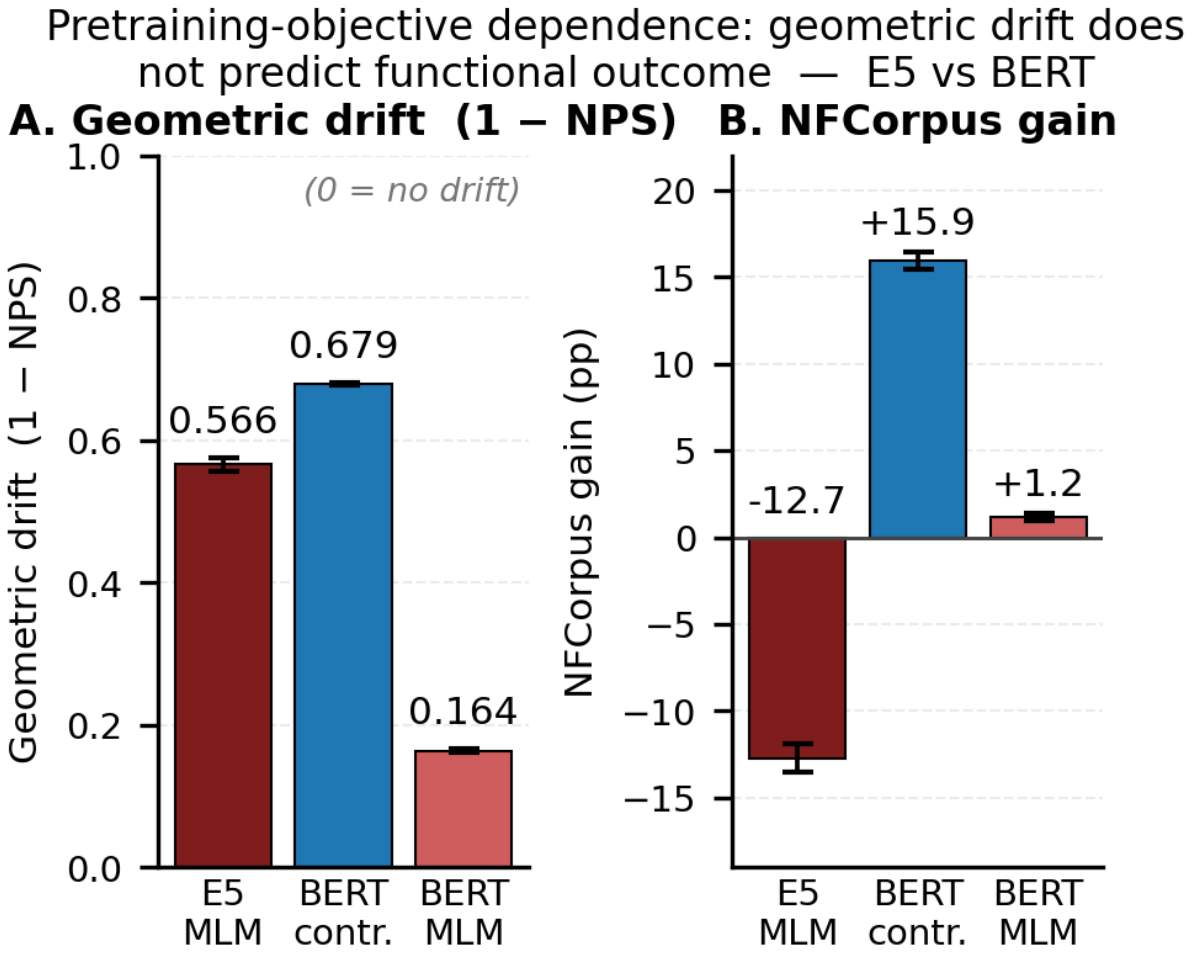

Figure 2: Pretraining-objective dependence on BERT. **(a)** Geometric drift, expressed as $1 - \text{NPS}$ at epoch 50; larger bars indicate larger drift. BERT MLM-on-MS-MARCO (rightmost) produces the *smallest* geometric drift of the three conditions, because BERT was pretrained with MLM and MLM fine-tuning is near a fixed point of the pretraining objective. **(b)** Functional outcome on the primary retrieval benchmark (NFCorpus gain in pp vs. base; MS MARCO drop in pp for the E5 reference condition). BERT contrastive LoRA produces the largest functional gain (+15.9 pp) despite having larger geometric drift than BERT MLM (+1.2 pp). Geometric drift magnitude does not predict the functional ranking. Error bars: 3-seed standard deviation.

Result. At epoch 50, BERT MLM LoRA produces NPS 0.836 ± 0.002—an order of magnitude smaller drift than E5 MLM at the same recipe (0.434 ± 0.009). The geometry barely moves (Figure 2, panel a). Yet the functional outcome is dramatically different from contrastive LoRA on the same corpus: BERT MLM gives $+1.17 \pm 0.19$ pp NFCorpus gain vs. base, while BERT contrastive LoRA at NPS 0.321 ± 0.002 gives +15.95 pp gain (panel b). The 95% CIs are disjoint. The lower-NPS condition (more drifted) is functionally *better*, not worse, contradicting the monotonicity assumption that underlies geometric-drift-as-proxy.

Why this happens. BERT was pretrained with MLM. Fine-tuning BERT with MLM on a new corpus is near a fixed point of the pretraining objective: gradients are small, geometric drift is minimal, and the representation barely changes because there is little for MLM to teach a model already trained on it. E5 was pretrained contrastively, so MLM fine-tuning on E5 is gradient-misaligned to pretraining, producing large gradients and large geometric drift. The relationship between fine-tuning loss and pretraining loss determines the magnitude of geometric drift; the relationship between fine-tuning loss and the evaluation task determines the functional outcome. These two relationships can vary independently.

Two failure modes, one conclusion. On E5, two conditions at matched NPS *and* matched CKA produce very different functional outcomes (§4.1): geometric similarity *overpredicts* functional similarity. On BERT, two conditions at very different NPS magnitudes produce a functional ranking opposite to what NPS would suggest: geometric drift magnitude *underpredicts* functional difference. Both findings rule out geometric-similarity-as-functional-proxy. They do so through different mechanisms but converge on the same conclusion: geometric drift magnitude is uninformative about functional outcome without knowledge of the gradient–pretraining alignment.

Pre-registration note. The BERT experiment was pre-registered with hypotheses (A) objective-driven, (B) corpus-driven, (C) mixed, and an explicit matched-NPS window of $[0.28, 0.36]$ derived from the BERT contrastive condition. The realized NPS of 0.836 falls outside this window, which the pre- registration specified as invalidating the strict matched-magnitude hypothesis test. The qualitative direction nonetheless supports hypothesis A: the loss does the work, not the corpus. The NPS-window violation is itself a finding—the matched-magnitude framework is well-defined only when the fine-tuning loss differs from the pretraining loss, a boundary condition the pre-registration did not anticipate. Per-checkpoint numbers are reported in Appendix B.

5 Structural Controls

The dissociation in §4 could in principle be explained by structural factors of the update: LoRA's low-rank confinement, the particular choice of rank, or the difference between low-rank and full-rank updates. Two controls rule these out.

Random rank-4 perturbation. If LoRA's preservation of retrieval were a property of low-rank confinement per se, then any rank-4 perturbation of comparable magnitude should also preserve retrieval. For each LoRA layer, we sample random A', B' matrices of the same shape as the trained LoRA matrices and rescale so the per-layer Frobenius norm of $\Delta W' = B'A'$ matches the per-layer Frobenius norm of the trained contrastive-LoRA ΔW at epoch 50. The result: random rank-4 perturbation produces NPS ≈ 0.97 (vs. base 1.00) and leaves all benchmarks within base 95% CI on both E5 and CLIP (Table 1). The same magnitude of weight-space perturbation, when produced

by gradient descent on a contrastive objective, drives NPS to 0.60. The factor that produces drift is not the rank, the magnitude, or any structural property of the update; it is the gradient-selected subspace direction.

Full fine-tune equivalence. If the aligned regime were a property of LoRA specifically—e.g., its implicit regularization toward sparse updates—then full fine-tuning with the same loss should pattern differently. We run full-rank fine- tuning on E5 with the same recipe as contrastive LoRA except all parameters are trainable. Full-FT NPS at epoch 50 is 0.569 ± 0.003 across 3 seeds, versus 0.597 ± 0.013 for LoRA. Retrieval benchmarks agree within 1.3 pp. The two conditions are statistically indistinguishable on every benchmark. The aligned-regime effect is a property of gradient-aligned fine-tuning, not of LoRA's rank constraint.

6 Probe-Configuration Artifacts in Embedding Evaluation

Linear probing is a standard interpretability technique for measuring what frozen representations encode [Alain and Bengio, 2017, Hewitt and Liang, 2019]. We document a probe- configuration artifact of up to 28.5 pp accuracy at *identical* embeddings on a standard MTEB-style evaluation, large enough to dominate any plausible representation-change signal in published comparisons. This is an interpretability illusion specific to embedding- evaluation practice.

The setup. Evaluating fine-tuning effects on classification tasks is commonly done by training a linear probe on top of frozen embeddings and reporting top-1 accuracy. The probe is typically a logistic regression classifier trained on a held-out subset of the task's training data with a fixed regularization constant.

The pitfall. Probe configuration choices—training set size and regularization—can produce accuracy swings that exceed any plausible fine-tuning signal, *at identical embeddings*. Figure 3 illustrates the effect on Banking77 with E5-base-v2 frozen embeddings; Table 2 summarizes the two endpoint configurations.

Table 2: Probe configuration sensitivity at identical E5-base-v2 embeddings on Banking77.

Probe configuration	**Train size**	**Regularization**	**Top-1 accuracy**
Weak	2,000	$C = 1.0$ fixed	63.5%
Strong	10,000 (full)	LogisticRegressionCV	92.05%
		Configuration-induced swing	**28.5 pp**

The 28.5 pp swing is driven entirely by probe configuration; the embeddings are byte-identical between the two rows. Any reported fine- tuning effect on Banking77 below this threshold is indistinguishable from probe noise unless the probe is configured strongly.

The artifact persists across fine-tuning. A subtler version of the same problem appears *within* a fine- tuning trajectory. Figure 3 shows that across the contrastive-LoRA fine-tuning of E5, the weak probe's accuracy on Banking77 moves from 63.5% at epoch 0 to ∼78% at epoch 50—a 14.6 pp swing—while the strong probe stays flat at ∼93%. The swing in the weak probe does not reflect a 14.6 pp improvement in the underlying representation; it reflects the weak probe slowly fitting structure that the strong probe already captured at epoch 0. A practitioner running the weak probe

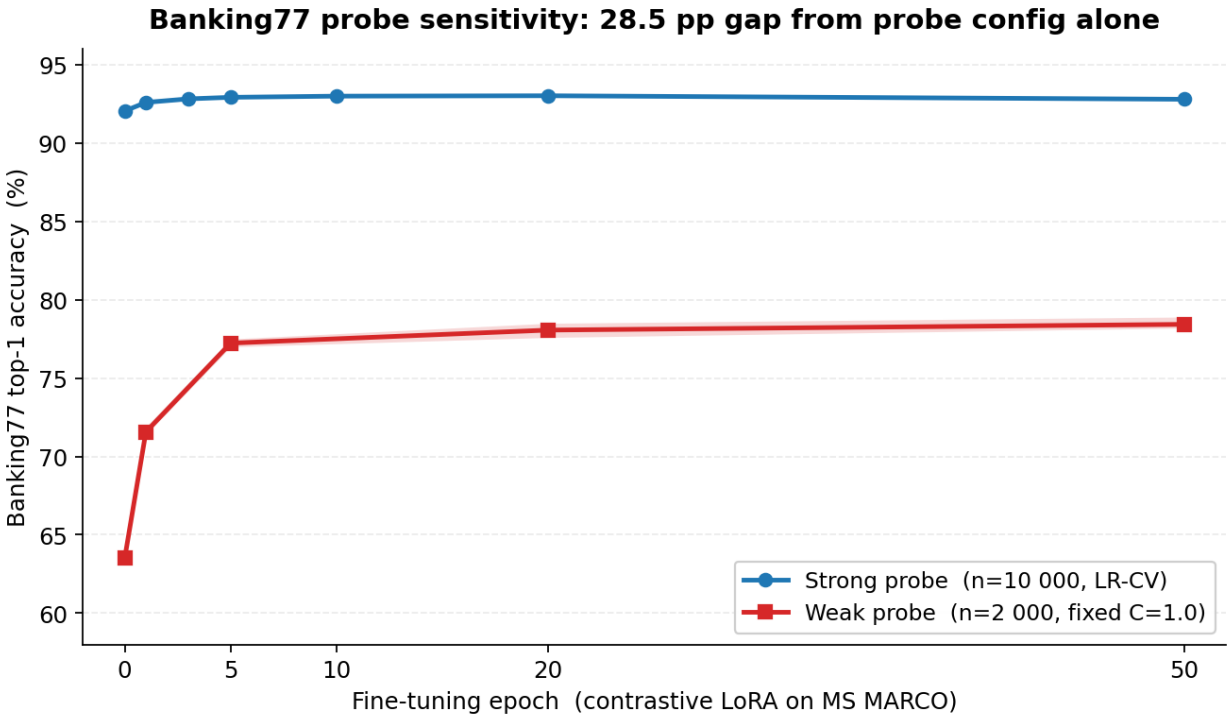

Figure 3: Probe-configuration sensitivity on Banking77 across the contrastive-LoRA fine-tuning trajectory. Both probes evaluate the *same* sequence of E5-base-v2 checkpoints. The strong probe (n=10,000, LogisticRegressionCV) is essentially flat at ∼93%; the weak probe (n=2,000, fixed C=1.0) sits ∼15–29 pp lower at every epoch and varies more across the trajectory. The 28.5 pp gap at epoch 0 is driven entirely by probe configuration; the embeddings being scored are byte-identical between the two curves.

and comparing epoch-0 to epoch-50 numbers would conclude that contrastive LoRA on MS MARCO substantially improves Banking77 classification. The strong probe shows no such improvement.

Practical implications.

1. Linear probes for embedding evaluation should be trained on the full available training set, not a subsample.
2. Regularization should be cross-validated (LogisticRegressionCV or equivalent), not fixed.
3. Reports should disclose probe training-set size, regularization selection procedure, and selected C.
4. When attributing a classification gap to representation change, the baseline probe should first be re-fit with strong hyperparameters.

We are not the first to note that probes are sensitive to their configuration [Hewitt and Liang, 2019]; we contribute a quantitative measurement of how large the artifact can be in a standard MTEB-style setup, and a concrete recipe to avoid it. We applied the strong configuration throughout the experiments in §4.

7 Discussion

What the dissociations show. The two dissociations in §4 make the same negative claim through different mechanisms. On E5, matched-NPS conditions produce order-of-magnitude differences in retrieval damage depending on the fine-tuning objective: NPS overpredicts functional similarity. On BERT, near-trivial geometric drift coexists with a large, CI-disjoint functional gap between MLM and contrastive fine- tuning at the same corpus, and the more-drifted condition is functionally better: NPS underpredicts functional difference. Together these rule out the use of NPS—and by extension any geometry-only label-free metric—as a predictor of functional outcome. They do not imply geometric metrics are uninformative: NPS reliably detects *that* representations have changed, and within a fixed gradient–pretraining alignment regime it correlates with effect magnitude. The failure is across-regime comparability, not within- regime sensitivity.

Implications for representation-similarity analysis. A common interpretability move is to compute CKA, NPS, or a related metric between two models—trained with different objectives, at different scales, or on different data—and infer from high similarity that the models encode similar information, or from low similarity that they encode different information. Our results show this inference is unsupported without additional information. Two models can have similar geometry and very different functional content (E5 matched-NPS comparison: $\sim$10 pp retrieval gap at the same NPS). Two models can have very different geometry and a functional ranking opposite to what the geometric metric would predict (BERT vs. E5 under matched recipes). The geometric similarity carries no licensure for the functional inference. Studies that use representation-similarity metrics to compare models should either ground the comparison in a downstream task or restrict their claims to geometric similarity proper.

Implications for label-free drift detection. A practical drift-detection system that takes NPS or CKA as input and emits a deployment decision is making an implicit assumption that the operating regime is fixed—typically the aligned regime, since that matches the common case of intentional fine-tuning on task-relevant data. This assumption is invisible in the metric. Production deployments where fine-tuning objectives drift away from the deployment task (continued pretraining, domain adaptation with mismatched objectives) can produce benchmark damage that the geometric metric does not flag as anomalous. We recommend that label-free drift detectors be deployed alongside either (a) calibration data tied to the deployment task, or (b) a registry of the training objective and pretraining objective used to produce each candidate update.

Relationship to catastrophic forgetting. The MLM-damages-retrieval finding overlaps qualitatively with catastrophic forgetting and negative transfer. The contribution of this work is the controlled *matched-magnitude* comparison: prior work documents that MLM continued pretraining can hurt downstream retrieval; this work shows that it does so at a geometric drift magnitude where random perturbation does not, and that the converse holds on a pretraining-aligned model where MLM produces near-trivial drift but still fails to recover the functional outcome of an aligned contrastive loss. The asymmetry between gradient-anti-aligned fine- tuning and matched-magnitude random perturbation, and the converse asymmetry on a pretraining-aligned model, have not to our knowledge been documented in the catastrophic-forgetting literature.

What we are not claiming. We do not claim that geometric drift metrics are always uninformative, that all label-free analysis is suspect, or that representation- similarity work in interpretability

is broadly invalid. The claim is narrower: geometric drift magnitude does not, on its own, predict the sign or scale of functional change across gradient–pretraining alignment regimes. Within a regime, geometric metrics behave as expected. The failures we document are systematic and predictable given knowledge of the alignment between fine-tuning loss, pretraining loss, and evaluation task; they are not evidence of metric noise. The matched-magnitude comparison itself is robust to anchor choice (§A.4).

Limitations. The matched-magnitude dissociation is demonstrated on E5-base-v2; the pretraining-objective-dependence dissociation extends the central claim to BERT, but a fully matched-NPS comparison on BERT is not achievable because MLM fine-tuning of an MLM-pretrained model produces near-trivial drift. Cross-family replication on a non-transformer encoder would strengthen the architectural generalization. We do not test multi-task losses (mixtures of contrastive and MLM); whether functional damage scales with mixing weight is a natural next experiment. While we verify the matched-magnitude framework is robust across three anchor sets (§A.4), absolute NPS and CKA magnitudes remain anchor-dependent (§A.3); cross-study comparisons of geometric similarity values should always cite the anchor used.

8 Conclusion

Geometric similarity metrics are foundational tools in the analysis of neural representations. We show, in controlled and pre-registered experiments, that they fail as proxies for functional similarity in two distinct ways. On E5-base-v2, matched-NPS conditions produce order-of- magnitude differences in retrieval damage depending on the fine-tuning objective. On BERT, near-matched fine-tuning recipes produce qualitatively swapped functional rankings between geometric drift and downstream outcome, because the gradient–pretraining alignment differs from E5. A corpus control rules out distribution shift as the cause; structural controls rule out low-rank confinement and full-rank/low- rank differences. We additionally document a probe-methodology artifact of up to 28.5 pp at identical embeddings, large enough to contaminate published fine-tuning evaluations that use weakly- configured linear probes.

The takeaway for interpretability practice is that representation- similarity metrics measure what they measure—geometric proximity in a chosen anchor space—and not, in general, functional similarity. Inferring functional similarity from geometric similarity requires either an evaluation on the function in question or knowledge of the gradient–pretraining alignment that produced the geometric state. Drift-detection systems and interpretability comparisons that treat these metrics as functional proxies will systematically miss a class of fine-tuning damage that looks geometrically routine but is gradient- aligned to features anti-correlated with the deployment task, and a mirror class of geometrically minimal fine-tuning that nonetheless fails to deliver functional outcomes that an aligned loss would produce on the same data.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2017.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P.

Cunningham. LoRA learns less and forgets less. In *Conference on Language Modeling (COLM)*, 2024.

Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. NFCorpus: A full-text learning to rank dataset for medical information retrieval. In *Advances in Information Retrieval – 38th European Conference on IR Research (ECIR)*, pages 716–722. Springer, 2016.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI (NLP4ConvAI)*, pages 38–45. Association for Computational Linguistics, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186, 2019.

Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 55–65, 2019.

Robert M. French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135, 1999.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2733–2743, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations (ICLR)*, 2023.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

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

John A. Lee and Michel Verleysen. Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing*, 72(7–9):1431–1443, 2009.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pages 109–165. Academic Press, 1989.

Jiaqi Mu and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations (ICLR)*, 2018.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 2014–2037, 2023.

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

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches at NeurIPS*, 2016.

Guillermo Ortiz-Jiménez, Alessandro Favero, and Pascal Frossard. Task arithmetic in the tangent space: Improved editing of pre-trained models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021.

Jarkko Venna, Jaakko Peltonen, Kristian Nybo, Helena Aidos, and Samuel Kaski. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research*, 11:451–490, 2010.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training, 2022.

Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11293–11302, 2019.

A NPS Sensitivity Analyses

A.1 NPS sanity checks

Identity. $\mathrm{NPS}(x, x) = 1.0$ at $k \in \{5, 10, 25, 50\}$. **Orthogonal rotation.** NPS under random rotation of the embedding space = 1.0. **Random permutation.** NPS under random permutation of points $\approx k/(n-1)$, matching the expected baseline.

A.2 k-sweep

NPS at E5 epoch-50 contrastive LoRA: 35.6% drop at $k = 5$ to 40.2% drop at $k = 50$, a 4.6 pp spread across a $10\times$ k-range. Conclusions in §4 are not artifacts of a specific k.

A.3 Anchor-set sensitivity: absolute magnitudes

NPS magnitude is anchor-dependent. On a CLIP rank-4 LoRA at epoch 50, an in-distribution anchor (drawn from the LoRA training data) gives a 67% NPS drop; an out-of-distribution anchor (general-purpose images) gives a 29% NPS drop on the same checkpoint. The qualitative drift signal is preserved but the magnitude differs by $\sim$2.3$\times$. All NPS values in the main text use a single fixed OOD anchor per model (MS MARCO dev for E5/BERT, MSCOCO val for CLIP), so within-model cross-condition comparisons are apples-to-apples.

A.4 Anchor-set robustness for the matched-magnitude comparison

The absolute-magnitude sensitivity in §A.3 raises a sharper question: is the matched-magnitude dissociation in Table 1 robust to anchor choice, or is the "comparable NPS band" an artifact of computing NPS on the contrastive-training anchor? We address this directly by recomputing NPS and CKA for every E5 condition under three anchor sets: **Anchor A** (MS MARCO probe, n=1000, the main-text reference), **Anchor B** (NFCorpus passages, n=1000, OOD relative to MS MARCO contrastive training), and **Anchor C** (Wikipedia paragraphs, n=1000, in-distribution for the MLM-Wiki training condition). All values are computed on existing embeddings (no retraining); each anchor uses the appropriate E5 `passage:` prefix.

Table 3: Anchor-set robustness for E5 Table 1 conditions. Each row reports (mean NPS / mean CKA) under three anchor sets. Multi-seed std omitted for compactness; full per-seed values are in the supplementary code.

Condition	**Anchor A (MS MARCO)**	**Anchor B (NFCorpus)**	**Anchor C (Wikipedia)**
Base	1.000 / 1.000	1.000 / 1.000	1.000 / 1.000
Random rank-4	0.974 / 1.000	0.981 / 1.000	0.978 / 1.000
Contrastive LoRA	0.596 / 0.884	0.701 / 0.908	0.602 / 0.879
Full-FT contrastive	0.570 / 0.854	0.679 / 0.893	0.590 / 0.865
Gaussian σ=0.16	0.570 / 0.880	0.674 / 0.864	0.592 / 0.876
MLM LoRA Wikipedia	0.479 / 0.820	0.514 / 0.768	0.491 / 0.830
MLM LoRA MS MARCO	0.434 / 0.807	0.456 / 0.738	0.456 / 0.829

Rank correlation across anchors. Spearman ρ between condition orderings under different anchors, computed across all 7 conditions:

Metric	**A vs. B**	**A vs. C**
NPS	$\rho = 0.964$ (p<0.001)	$\rho = 1.000$ (p<0.001)
CKA	$\rho = 0.964$ (p<0.001)	$\rho = 1.000$ (p<0.001)

The condition ranking is essentially anchor-invariant: a single tied- pair swap on Anchor B between Gaussian and Full-FT contrastive (which sit within 0.005 NPS of each other) is the only deviation from perfect rank agreement.

Matched-band condition under each anchor. The substantive claim of §4.1 is that the four non-control perturbative conditions (Contrastive LoRA, Gaussian σ=0.16, MLM-Wiki, MLM-MS-MARCO) sit in a comparable geometric-drift band relative to the spread of their downstream damage. We operationalize this as a ratio test: the MS MARCO damage ratio (max/min, in pp) across

these four conditions should be substantially larger than the drift-distance ratio $\max(1-\text{NPS})/\min(1-\text{NPS})$ across the same four. The test holds (margin $\geq 3\times$) under all three anchors:

Anchor	Drift ratio	Damage ratio	Holds?
A (MS MARCO)	1.40×	7.85×	yes
B (NFCorpus)	1.82×	7.85×	yes
C (Wikipedia)	1.37×	7.85×	yes

Verdict. The matched-magnitude dissociation in Table 1 is robust to anchor choice. Absolute NPS magnitudes shift by ∼0.10–0.18 across anchors (NPS is consistently higher under the OOD NFCorpus anchor, as neighborhoods of unrelated passages are easier to preserve), but the rank ordering is preserved and the matched-band ratio test holds under all three anchors with substantial margin. One caveat for absolute CKA values: CKA on the MLM conditions drops further under the NFCorpus anchor (MLM-MS-MARCO CKA 0.807 → 0.738), so CKA is more anchor-sensitive than NPS for those conditions and absolute CKA values across studies should always cite the anchor.

B Cross-Family Replication

B.1 Aligned-regime preservation across families

We replicate the contrastive-LoRA aligned condition on CLIP ViT-L/14 (4 seeds) and `bert-base-uncased` (3 seeds). Both preserve retrieval within base CI on their respective evaluation benchmarks (MSCOCO Karpathy and Flickr30K for CLIP; NFCorpus for BERT). The matched-magnitude Gaussian control on the same families produces CI- disjoint damage on out-of-distribution retrieval. Full per-seed numbers are provided in the supplementary code.

B.2 BERT MLM-on-MS-MARCO: full per-checkpoint numbers

The cross-family MLM condition reported in §4.3 trains BERT MLM LoRA on the same 2,000 MS MARCO passages used by BERT contrastive LoRA, with 3 seeds, 50 epochs, LoRA rank-4, AdamW with cosine schedule, 15% MLM masking. Per-checkpoint NPS (against base BERT, MS MARCO probe anchor) and NFCorpus nDCG@10 (3-seed mean ± std):

Epoch	NPS	NFCorpus nDCG@10
0	1.000 ± 0.000	6.97%
1	0.927 ± 0.003	$7.54\% \pm 0.03$
3	0.805 ± 0.007	$8.14\% \pm 0.06$
5	0.812 ± 0.008	$8.16\% \pm 0.26$
10	0.827 ± 0.009	$8.24\% \pm 0.15$
20	0.833 ± 0.007	$8.10\% \pm 0.10$
50	0.836 ± 0.002	$8.14\% \pm 0.19$

For comparison, BERT contrastive LoRA at epoch 50 produces NPS 0.321 ± 0.002 and NFCorpus nDCG@10 $22.92\% \pm 0.50$ pp (+15.95 pp gain vs. base 6.97%). The two conditions are CI-disjoint on NFCorpus (BERT contrastive 95% CI $[19.28, 26.56]$; BERT MLM 95% CI $[6.42, 10.08]$).

B.3 BERT MLM also gains modestly on MS MARCO

BERT is not pretrained as a retrieval model, so MS MARCO is not the primary evaluation for the BERT conditions. For completeness we note that BERT MLM on MS MARCO produces a $+6.70 \pm 0.27$ pp gain on MS MARCO nDCG@10 vs. base BERT (base 25.64%, MLM ep-50 $32.35\% \pm 0.27$). This gain is consistent with MLM fine-tuning recovering some domain-relevant masking patterns that aid passage scoring; it does not contradict the qualitative ranking in §4.3, where contrastive LoRA at the same corpus produces substantially larger gains on the primary NFCorpus benchmark.

C Pre-Registration Documents

The pre-registration documents for the multi-seed E5 experiment, the MLM training stability check, the matched-corpus E5 MLM control, and the BERT MLM-on-MS-MARCO cross-family replication are included in the supplementary code at `experiments/*/PREREGISTRATION.md`. Each document specifies the hypotheses, falsification criteria, and predicted outcomes prior to runs. The BERT pre-registration's matched- NPS window hypothesis $[0.28, 0.36]$ is discussed in §4.3; the realized NPS of 0.836 falls outside this window, a result the pre-registration explicitly flagged as invalidating the strict matched-magnitude hypothesis test.