

Label-Free Post-hoc Alignment of Frozen Decoder-Only LLM Hidden States for Symmetric Sentence-Similarity Retrieval

GangHo Lee Yong-Gu Lee*

AI Graduate School

Gwangju Institute of Science and Technology

ianleegh20@gmail.com, lygu@gist.ac.kr

Abstract

Decoder-only large language models (LLMs) carry rich semantic information in their hidden states, but these states are not directly usable for retrieval: they exhibit strong anisotropy and excessive dimensionality, producing distorted cosine-similarity rankings and prohibitive index sizes. A standard solution is to train a dedicated embedding model, which requires positive pairs and substantial compute. We ask a different question: *can we make frozen LLM hidden states retrieval-usable purely through post-hoc transformation, without fine-tuning the LLM backbone or using labeled positive pairs?* We propose **NCWP** (Neighbor-Contrastive Whitening Projection), a post-hoc alignment method for symmetric sentence-similarity retrieval that combines ZCA-shrink whitening for global isotropy correction with augmentation-free neighbor-contrastive projection learning, requiring only a single linear projection at inference. On the official STS-Benchmark test split and BEIR-Quora across three decoder-only LLM backbones (Llama-3.1-8B, Qwen1.5-4B, Qwen2-7B), NCWP substantially *exceeds the raw Base hidden states* at aggressive-compression dimensions (e.g., +23.8 Spearman over Qwen1.5-4B Base on STSBenchmark at $8\times$ compression; +13.6 nDCG@10 over Qwen1.5-4B Base on BEIR-Quora at $32\times$ compression), simultaneously delivering retrieval-quality improvements and index-size reductions; the pattern also holds in breadth analyses on STS12-16+SICK-R and CQADupStack. All main results are reported as mean over three random seeds. We characterize a *dimensional crossover* in which NCWP’s advantage over PCA-whitening widens monotonically down to $256\times$ compression on Qwen1.5-4B and Qwen2-7B, and provide geometric evidence consistent with the proposed mechanism via anisotropy measurements: NCWP reduces anisotropy on all three LLM backbones by roughly an order of mag-

nitude, to a near-zero residual on Qwen1.5-4B and Llama-3.1-8B and a smaller reduction on Qwen2-7B. NCWP is also complementary to prompt-based extraction: applied on top of PromptEOL it improves over full-dimensional PromptEOL while compressing up to $64\times$. Code is available at <https://github.com/ianleeGH20/NCWP>.

1 Introduction

Recent large language models (LLMs) such as Llama-3 (Grattafiori et al., 2024) and Qwen (Qwen Team, 2024; Yang et al., 2024) expose high-dimensional internal representations. We study whether these representations can be repurposed for *retrieval*: in a pipeline that already serves such a model for generation (Lewis et al., 2020), reusing it as the embedding backbone would avoid maintaining a separate embedding model and simplify versioning.

The core challenge: LLM hidden states are not directly retrieval-usable. Naively mean-pooling the last-layer hidden states of a decoder-only LLM performs poorly on retrieval. Two structural issues are well documented. First, contextual representations from decoder-only models are known to be strongly *anisotropic* at the token level (Ethayarajh, 2019), echoing the non-isotropy reported for static word representations (Mu and Viswanath, 2018); we measure the corresponding sentence-level effect directly in Section 4.4, and it distorts cosine-similarity neighborhood structure, the primitive driving top- k retrieval. Second, modern LLM hidden states are *high-dimensional* (3584–4096 for Qwen2-7B and Llama-3.1-8B): a flat index stores $O(ND)$ floats and an exhaustive cosine scan costs $O(ND)$ per query, so both grow linearly in D , which is precisely the pressure that motivates approximate and compressed indexes (Johnson et al., 2021; Malkov and Yashunin, 2020). Together, raw mean-pooled hidden states substantially underper-

* Corresponding author.

form what the semantic information they contain would suggest.

The standard solution and its cost. The dominant response to this gap is to *convert* the LLM into a dedicated embedding model through additional contrastive training. Such training can achieve strong retrieval quality, but requires labeled, synthetic, or mined positive pairs and substantial compute—resources that may be unavailable when only an inference endpoint is exposed, when the deployed LLM must remain fixed for compatibility, or when retraining is prohibited for safety or compliance reasons. The natural question is whether a *post-hoc* alternative exists that keeps the LLM backbone frozen and uses no labeled pairs, while still substantially closing the gap.

Our approach: NCWP. We propose **Neighbor-Contrastive Whitening Projection (NCWP)**, a post-hoc method that keeps the LLM fully frozen and learns only a single linear projection $\mathbf{W} \in \mathbb{R}^{D \times r}$ on an unlabeled corpus ($r \ll D$). NCWP combines two stages: (1) ZCA-shrink whitening (Kessy et al., 2018; Ledoit and Wolf, 2004) to obtain an approximately isotropic starting geometry, and (2) augmentation-free neighbor-contrastive projection learning, in which positives are mined from each anchor’s k -nearest neighbors in the whitened space and training is stabilized by hard-negative selection and periodic QR retraction onto the Stiefel manifold. Unlike SimCSE-style methods (Gao et al., 2021) that rely on dropout-based augmentation, NCWP uses the geometric structure of the whitened space directly. The closest prior work, Dwibedi et al. (2021)’s NN-CLR, also mines nearest-neighbor positives but operates end-to-end on visual encoders with strong augmentations—our post-hoc, frozen-backbone, sentence-retrieval setting differs along all three axes (Section 2).

Empirical findings. Across three decoder-only LLM backbones (Llama-3.1-8B, Qwen1.5-4B, Qwen2-7B), NCWP exhibits four consistent properties. (i) **It substantially exceeds raw Base hidden states at aggressive compression:** +23.8 Spearman on Qwen1.5-4B STSBenchmark at $r=320$ ($8\times$) and +13.6 nDCG@10 on Qwen1.5-4B BEIR-Quora at $r=80$ ($32\times$). (ii) **A dimensional crossover holds against PCA-whitening across all backbones:** NCWP’s advantage widens from the largest tested dimensions through the aggressive-compression range, while both meth-

ods eventually degrade at the most extreme dimensions. (iii) **The pattern is not specific to two testbeds:** it replicates in breadth analyses on STS12–16+SICK-R and CQADupStack, and a projection fit on one corpus transfers to a related one without refitting (Section 4.6). (iv) **NCWP substantially reduces anisotropy**, from 0.67–0.80 at the raw Base to near zero on Qwen1.5-4B and Llama-3.1-8B. All main NCWP numbers are mean $\pm$ std over three random seeds.

Contributions.

- We propose **NCWP**, a post-hoc alignment method for the symmetric sentence-similarity retrieval setting: it makes frozen decoder-only LLM hidden states retrieval-usable through a single learned linear projection, with the LLM backbone fully frozen, no labeled positive pairs, and inference cost identical to PCA-whitening.
- We empirically show that on three decoder-only LLM backbones, NCWP *exceeds raw LLM hidden states* at the reported aggressive-compression dimensions, simultaneously delivering retrieval-quality improvements and index-size reductions of $8\times$ – $32\times$.
- We characterize a *dimensional crossover* in compressed symmetric retrieval and provide geometric evidence consistent with the proposed mechanism via anisotropy measurements, near-zero on Qwen1.5-4B and Llama-3.1-8B.
- We show that NCWP *composes with* rather than competes against prompt-based extraction: stacked on PromptEOL it beats full-dimensional PromptEOL while compressing up to $64\times$ (Section 4.7).

2 Related Work

Sentence embeddings via encoder training.

A dominant line of work *trains* an encoder for retrieval-friendly geometry, either via supervised Siamese or triplet objectives over labeled pairs (Reimers and Gurevych, 2019), self-supervised contrastive objectives based on dropout, augmentation, prompting, or auxiliary difference prediction (Gao et al., 2021; Yan et al., 2021; Chuang et al., 2022; Jiang et al., 2022), or large-scale weakly supervised, unsupervised, and multi-stage retrieval training (Wang et al., 2022;

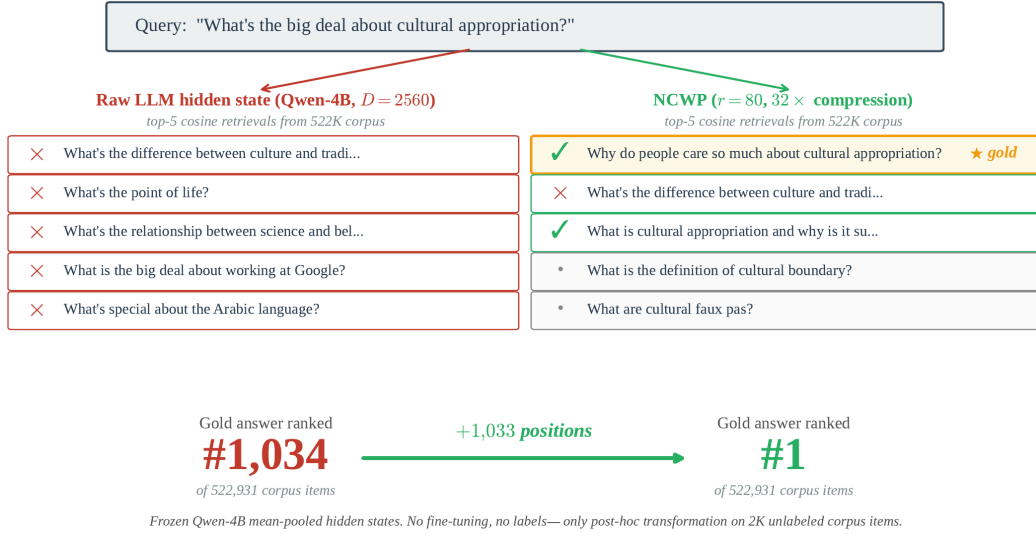


Figure 1: **NCWP makes frozen LLM hidden states retrieval-usable.** A duplicate-question retrieval example from BEIR-Quora (522,931-item corpus): with raw mean-pooled Qwen1.5-4B hidden states ($D=2560$) the gold paraphrase is buried at rank 1,034 and the top-5 cosine neighbors are dominated by the surface pattern “What’s the ...” regardless of topic. After post-hoc NCWP alignment to $r=80$ ($32\times$ compression) the same query retrieves it at rank 1, with the backbone frozen and no labeled pairs. More examples: Appendix M.

Xiao et al., 2024; Izacard et al., 2022; Ni et al., 2022). Dense encoders of this kind are the standard retrieval component in open-domain QA and retrieval-augmented pipelines (Karpukhin et al., 2020; Xiong et al., 2021; Lewis et al., 2020), but obtaining them requires encoder training on labeled, weakly supervised, synthetic, or mined pairs. NCWP targets the complementary setting in which the encoder cannot be retrained.

LLM-as-retriever. A recent line investigates whether decoder-only LLMs can serve as text encoders. LLM2Vec (BehnamGhader et al., 2024) unlocks bidirectional attention and applies unsupervised contrastive fine-tuning; E5-Mistral (Wang et al., 2024) fine-tunes contrastively on synthetic triplets, and Qwen3-Embedding (Zhang et al., 2025) combines unsupervised pretraining with supervised fine-tuning. These methods generally produce stronger retrieval than the approach we study, but all require either labeled or synthetic pairs or substantial compute for fine-tuning. We position NCWP as a *backbone-frozen, label-free* alternative for settings where such resources are unavailable.

Training-free extraction from frozen LLMs. A separate family changes only the input template or the pooling rule, leaving the backbone untouched and requiring no training at all. PromptEOL (Jiang et al., 2024) elicits a representation through an ex-

plicit one-word-summary prompt, and Echo embeddings (Springer et al., 2025) repeat the input and read from the second occurrence; both operate zero-shot without fine-tuning. These are the strongest label-free competitors to the plain mean pooling our Base uses, and they are *orthogonal* to NCWP: they change how the D -dimensional vector is produced, NCWP how it is transformed and compressed. We evaluate NCWP both against and stacked on them (Section 4.7).

Post-hoc embedding alignment. A complementary, lighter-weight line treats the encoder as fixed and applies a post-hoc transformation. Anisotropy diagnostics (Ethayarajh, 2019; Mu and Viswanath, 2018) motivated whitening-based corrections such as PCA-whitening and WhiteningBERT (Su et al., 2021), distribution-matching approaches such as BERT-flow (Li et al., 2020), and simple post-processing baselines that remove one or more top principal components (Arora et al., 2017; Mu and Viswanath, 2018). Within this family, NCWP uses ZCA-shrink whitening—the unique *symmetric* whitening transform, and the one minimizing the expected squared discrepancy between whitened and original centered vectors among covariance-whitening transforms (Kessy et al., 2018). In our experiments, combining this initialization with neighbor-contrastive learning preserves useful neighborhood structure under ag-

gressive truncation rather than stopping at the closed-form whitening step. We compare against PCA-whitening, Soft-whitening, All-But-The-Top (ABTT) (Mu and Viswanath, 2018), LPP (He and Niyogi, 2003), and random projection (Bingham and Mannila, 2001; Johnson and Lindenstrauss, 1984).

Compression-aware embeddings. Matryoshka Representation Learning (Kusupati et al., 2022) jointly trains an embedding model to perform well at multiple truncated dimensions, providing native dimensional flexibility. This approach is complementary to ours but operates in a fundamentally different regime: it modifies the encoder’s training objective. NCWP applies to any frozen encoder without modification and is therefore the appropriate choice when the encoder cannot be retrained.

Nearest-neighbor contrastive learning. Dwibedi et al. (2021)’s NN-CLR replaces standard augmentation-based positives with nearest-neighbor positives from a feature memory bank for self-supervised visual representation learning. Our neighbor-contrastive objective shares this high-level intuition but differs in setting: NN-CLR trains the visual encoder end-to-end with strong augmentations, whereas we keep the encoder frozen, learn only a single linear projection without augmentation, and explicitly construct an isotropic starting geometry via ZCA-shrink whitening so that base-space k -NN already approximates relevance.

3 Method

We formalize the post-hoc, frozen-backbone setting (Section 3.1), then describe the two stages of NCWP: ZCA-shrink whitening (Section 3.2) and neighbor-contrastive projection learning (Section 3.3). Figure 3 in Appendix G illustrates the intuition behind NCWP through the lens of top- k cosine retrieval.

3.1 Problem Setting

Frozen encoder, post-hoc projection. Let $f : \mathcal{X} \rightarrow \mathbb{R}^D$ be a frozen pretrained encoder. For decoder-only LLMs, f extracts last-layer hidden states, mean-pools over valid tokens, and ℓ_2 -normalizes. Given an unlabeled corpus $\mathcal{C} = \{x_1, \dots, x_N\}$, our goal is to learn a single linear

projection $\mathbf{W} \in \mathbb{R}^{D \times r}$ ($r \ll D$) such that

$$g(\mathbf{x}) = \frac{(\mathbf{x} - \boldsymbol{\mu}) \mathbf{S} \mathbf{W}}{\|(\mathbf{x} - \boldsymbol{\mu}) \mathbf{S} \mathbf{W}\|_2} \in \mathbb{R}^r \quad (1)$$

yields better cosine retrieval than raw embeddings, especially at $r \ll D$. Here $\boldsymbol{\mu}$ and $\mathbf{S}$ are closed-form from corpus statistics; $\mathbf{W}$ is the only learned component. At inference, the entire post-hoc transformation reduces to a single matrix multiplication plus ℓ_2 normalization.

Symmetric retrieval and the neighbor-mining assumption. NCWP exploits a property of *symmetric retrieval*—tasks where queries and indexed items come from the same distribution (paraphrase mining, duplicate question detection, STS-style lookup). In this regime, two items judged relevant are likely to lie among each other’s nearest neighbors under f , even when f ’s ranking is imperfect. This is the empirical foundation that lets us use k -NN as a label-free positive-pair signal. We do not claim transfer to *asymmetric retrieval* (short queries against long abstracts), which lies outside our scope.

3.2 Stage 1: ZCA-Shrink Whitening

Let $\mathbf{X} \in \mathbb{R}^{N \times D}$ be the matrix of frozen corpus embeddings. We compute the centering $\boldsymbol{\mu} = \frac{1}{N} \sum_i \mathbf{x}_i$, the centered matrix $\mathbf{X}_c = \mathbf{X} - \mathbf{1}\boldsymbol{\mu}^\top$, and the sample covariance $\hat{\boldsymbol{\Sigma}} = \frac{1}{N-1} \mathbf{X}_c^\top \mathbf{X}_c$. To prevent inverse-square-root amplification of small eigenvalues, we apply Ledoit–Wolf-style linear shrinkage (Ledoit and Wolf, 2004) toward the scaled identity, with λ fixed a priori rather than set by the Ledoit–Wolf optimal coefficient:

$$\hat{\boldsymbol{\Sigma}}_\lambda = (1 - \lambda) \hat{\boldsymbol{\Sigma}} + \lambda \frac{\text{tr}(\hat{\boldsymbol{\Sigma}})}{D} \mathbf{I}_D, \quad \lambda \in [0, 1]. \quad (2)$$

With the eigendecomposition $\hat{\boldsymbol{\Sigma}}_\lambda = \mathbf{U} \boldsymbol{\Lambda} \mathbf{U}^\top$, the ZCA whitening matrix is

$$\mathbf{S} = \mathbf{U} \boldsymbol{\Lambda}^{-1/2} \mathbf{U}^\top, \quad (3)$$

and the whitened, ℓ_2 -normalized embeddings $\tilde{\mathbf{x}}_i = \text{norm}((\mathbf{x}_i - \boldsymbol{\mu})\mathbf{S})$ are passed to Stage 2. The key distinction from PCA-whitening is the rotation back by $\mathbf{U}^\top$: PCA-whitening uses $\mathbf{S}_{\text{PCA}} = \boldsymbol{\Lambda}^{-1/2} \mathbf{U}^\top$, which discards the original axis alignment, whereas ZCA preserves it. ZCA is the symmetric whitening transform and, among whitening transforms, minimizes the expected squared discrepancy between the whitened and original

centered vectors (Kessy et al., 2018). Combined with shrinkage, this preserves more local geometric structure under aggressive truncation, which our ablation (Section 4.5) shows to be quantitatively important at low target dimensions.

3.3 Stage 2: Neighbor-Contrastive Projection Learning

The second stage learns $\mathbf{W} \in \mathbb{R}^{D \times r}$ on the whitened embeddings. The output is $\mathbf{z}_i = \tilde{\mathbf{x}}_i \mathbf{W} / \|\tilde{\mathbf{x}}_i \mathbf{W}\|_2 \in \mathbb{R}^r$. Note that since $\tilde{\mathbf{x}}_i = \text{norm}((\mathbf{x}_i - \boldsymbol{\mu}) \mathbf{S})$ has unit ℓ_2 norm, normalizing it again before projection is unnecessary, and the training-time map $\mathbf{x}_i \mapsto \mathbf{z}_i$ coincides exactly with the inference map of Equation 1: cosine-based ranking is invariant to the intermediate normalization. We initialize $\mathbf{W}_{ij} \sim \mathcal{N}(0, 1/D)$ entrywise and optimize with AdamW.

Neighbor mining with cosine threshold. For each anchor i , we form $\mathcal{N}_k(i)$, its k nearest neighbors in the whitened space (excluding itself), and filter by a cosine threshold $\tau \in [-1, 1]$:

$$\mathcal{P}(i) = \{j \in \mathcal{N}_k(i) \mid \tilde{\mathbf{x}}_i^\top \tilde{\mathbf{x}}_j \geq \tau\}. \quad (4)$$

If $\mathcal{P}(i)$ is empty, we fall back to $\mathcal{N}_k(i)$. At each step we sample one positive i^+ uniformly from $\mathcal{P}(i)$.

Contrastive objective with memory bank and hard negatives. For a mini-batch $\mathcal{B}$ of anchors and their sampled positives, all $|\mathcal{B}|$ positives serve as in-batch negatives for the other anchors. To enrich the negative pool, we additionally maintain a FIFO memory bank $\mathcal{M}$ of past projected embeddings (size M) and select the top- K hardest negatives (highest cosine similarity to the anchor) from the union of in-batch and memory-bank negatives. In the main STS trainer, the $M=4096$ bank saturates, expanding the candidate pool from 512 in-batch negatives to 4,608 candidates; across nine instrumented conditions, 87.9–88.4% of selected top- K negatives come from the bank. With temperature T , we use a symmetrized InfoNCE loss (van den Oord et al., 2018; Chen et al., 2020):

$$\mathcal{L}_{\text{NCE}} = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \log \frac{\exp(\mathbf{z}_i^\top \mathbf{z}_{i^+}/T)}{\sum_j \exp(\mathbf{z}_i^\top \mathbf{z}_j/T)}, \quad (5)$$

where j ranges over $\mathcal{B}_i^- \cup \{i^+\}$ and $\mathcal{B}_i^-$ is the per-anchor hard-negative set, averaged over both anchor-positive directions. Equation 5 is the entire training objective: $\mathcal{L} = \mathcal{L}_{\text{NCE}}$.

On soft regularizers. Earlier versions of NCWP added a VICReg-inspired covariance penalty (Bardes et al., 2022), $\|\text{Cov}(\mathbf{Z}) - \mathbf{I}_r\|_F^2$, and an additional orthogonality penalty, $\|\mathbf{W}^\top \mathbf{W} - \mathbf{I}_r\|_F^2$, as safeguards against variance collapse under aggressive compression. An ablation prompted by review (Appendix J) shows that both are redundant once QR retraction is in place: removing either or both leaves Spearman within seed noise and leaves the conditioning and collapse diagnostics unchanged. We therefore drop them from the default objective and report them only as optional stabilizers.

QR retraction. To stabilize the projection geometry during training, every N_{qr} steps we apply a QR retraction of $\mathbf{W}$ onto the Stiefel manifold: $\mathbf{W} \leftarrow \mathbf{Q}$, where $\mathbf{W} = \mathbf{Q}\mathbf{R}$ is the (thin) QR decomposition. Periodically enforcing orthonormal columns exactly provides a small but consistent stabilization across target dimensions (see ablation, Section 4.5), and it is what makes the soft regularizers above unnecessary: removing QR retraction instead degrades the conditioning of $\mathbf{W}$ (condition number $1.46 \rightarrow 1.62$) together with the off-diagonal covariance and residual anisotropy.

Inference cost. At inference, NCWP stores $\boldsymbol{\mu} \in \mathbb{R}^D$, $\mathbf{S} \in \mathbb{R}^{D \times D}$, and $\mathbf{W} \in \mathbb{R}^{D \times r}$; these can be precomposed into a single $D \times r$ matrix $\mathbf{S}\mathbf{W}$ for deployment, so inference is one matrix multiplication and one ℓ_2 normalization per embedding—identical in cost to PCA-whitening at the same target r . Whitening is a one-time $O(D^3)$ eigendecomposition (seconds for $D \leq 4096$); training $\mathbf{W}$ converges in tens of minutes on a single GPU. Full details are in Appendix C.

4 Experiments

We evaluate NCWP on aggressive compression of frozen LLM hidden states (Sections 4.2, 4.3), anisotropy mechanism evidence (Section 4.4), component ablation (Section 4.5), generalization beyond the two headline testbeds (Section 4.6), and comparison against training-free extraction recipes (Section 4.7).

4.1 Experimental Setup

Data, metrics, backbones. Our two headline testbeds are **STSBenchmark** (Cer et al., 2017) and **BEIR-Quora** (Thakur et al., 2021) (522,931 corpus items, 10,000 test queries); Section 4.6 adds

the **STS12–16+SICK-R** suite (Agirre et al., 2012, 2013, 2014, 2015, 2016; Marelli et al., 2014) (six datasets, 22.5K evaluated pairs, including all 9,927 SICK-R pairs) and three **CQADupStack** subforums (Hoogeveen et al., 2015) (4.2K queries), all accessed through MTEB/BEIR (Muennighoff et al., 2023; Thakur et al., 2021) and all public. We report **Spearman** $\times 100$ on STS tasks and **nDCG@10** on retrieval tasks. The three frozen decoder-only backbones are **Llama-3.1-8B** = meta-llama/Meta-Llama-3.1-8B (Grattafiori et al., 2024) ($D=4096$), **Qwen1.5-4B** = Qwen/Qwen1.5-4B (Qwen Team, 2024) ($D=2560$), and **Qwen2-7B** = Qwen/Qwen2-7B (Yang et al., 2024) ($D=3584$). Embeddings are last-layer hidden states, mean-pooled over valid tokens and ℓ_2 -normalized, in bf16 with maximum length 128; Section 4.7 varies pooling and layer.

Baselines. Within the label-free, frozen-backbone family we compare against **PCA-whitening** (Su et al., 2021), **Soft-whitening** (PCA-whitening with eigenvalue shrinkage), **ZCA-only** (Stage 1 alone), **LPP** (He and Niyogi, 2003), **Random projection** (Bingham and Mannila, 2001; Johnson and Lindenstrauss, 1984), and the training-free recipes **PromptEOL** (Jiang et al., 2024) and **Echo** (Springer et al., 2025). On STSBenchmark we additionally evaluate **ABTT** (Mu and Viswanath, 2018) and dimension-matched **ABTT+PCA**, which we did not run on Quora. Every other baseline is run at every target dimension on both tasks; Appendix B gives the exact pipelines. The uncompressed **Base** is an unadapted reference point, not a competitor.

Fitting protocol. All post-hoc methods are fit on the same unlabeled subsample per setting, and headline test queries never enter fitting. Fit sizes, collected in one place: 2,910 unique STSBenchmark validation sentences (de-duplicated from the 1,500 validation pairs; an implementation-level count, released as an explicit index with our code), 2,000 BEIR-Quora corpus items for Qwen1.5-4B and 1,000 for the two larger backbones, and a 50,000-item Quora fit for the positive-source control of Section 4.5. The full NCWP configuration was fixed once on the STSBenchmark development/fitting split and reused without per-dataset hyperparameter tuning; see Appendix B. The breadth-only STS12–16/SICK-R analysis instead fits each transform on the unlabeled sentences of the evaluated benchmark and is therefore transductive; we

report it separately from the headline inductive test results.

4.2 Main Results: LLM Backbones on STSBenchmark

We use the official STSBenchmark split and MTEB metric (Cer et al., 2017; Muennighoff et al., 2023): 2,910 unique validation sentences are used as unlabeled fitting data for NCWP/PCA/ZCA/ABTT, and Spearman correlation is computed only on the 1,379 official test pairs (rescaled to $[0, 100]$). Target dimensions are scaled to each backbone’s native D . Table 1 reports results on Qwen1.5-4B (the smallest backbone); Llama-3.1-8B and Qwen2-7B are in Appendix D.

Table 1: **STSBenchmark official test split** (Spearman $\times 100$) on Qwen1.5-4B ($D = 2560$, Base = 31.31). NCWP reports mean over 3 random seeds (std ≤ 1.70); best in each column in **bold**. Random projection, LPP, and the additional $r=5$ point are in Table 7.

$r =$	1280	640	320	160	80	40	20	10
PCA	55.5	53.6	51.9	49.9	46.0	40.4	34.3	23.4
Soft-W	51.7	49.3	47.0	44.4	40.8	36.1	30.4	21.5
ZCA	55.8	54.0	51.3	47.6	42.5	38.4	37.3	32.3
ABTT+PCA	47.3	46.2	45.0	43.4	40.9	37.9	33.7	28.2
NCWP	56.5	55.6	55.1	54.2	52.5	49.0	44.9	37.4

Full D : raw Base = 31.31; ABTT ($m=5$) = 48.06.

NCWP substantially exceeds raw Base. NCWP outperforms every dimension-matched post-hoc baseline, including ABTT+PCA, and *substantially exceeds the raw $D=2560$ Base* (31.31): +23.8 Spearman at $r=320$ ($8\times$), +21.2 at $r=80$ ($32\times$), and +13.6 at $r=20$ ($128\times$). PCA-whitening remains above Base through $r=20$ but falls below it at $r=10$. Thus, the right post-hoc transformation can simultaneously *compress and improve* the geometry; Figure 2 shows the same pattern on the other backbones.

Dimensional crossover and a second whitening observation. As shown in Figure 2, NCWP exhibits a consistent dimensional crossover against PCA-whitening: on Qwen1.5-4B the advantage grows from +1.0 at $r=1280$ to +14.0 at $r=10$; at the still smaller $r=5$, both methods degrade and the gap narrows to +12.7. ZCA-only also differs meaningfully from PCA-whitening, but the ordering is not described by a single threshold on the official test split: ZCA is higher at $r=20$ and 10, whereas PCA is higher at $r=40$ and 80.

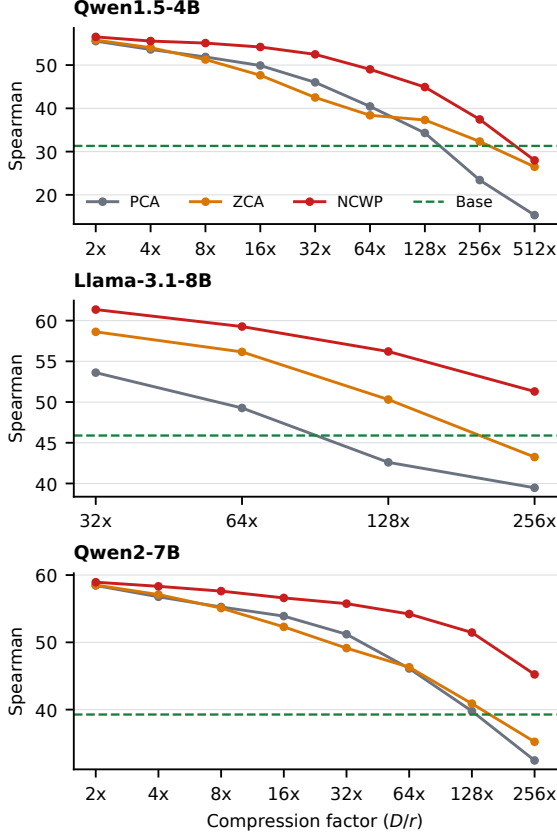


Figure 2: **The dimensional crossover holds across all three backbones.** STSBenchmark official-test Spearman against compression factor D/r . NCWP (red) increasingly separates from PCA-whitening (gray) through the aggressive-compression range; the dotted green line is the uncompressed Base. Tabulated in Appendix D.

4.3 Main Results: LLM Backbones on BEIR-Quora

We evaluate on the full BEIR-Quora benchmark (522,931 corpus items, 10,000 test queries), fitting all post-hoc methods on a fixed random subsample of 2,000 corpus items. Quora is a duplicate-question retrieval task and falls within our symmetric retrieval scope. Table 2 shows results on Qwen1.5-4B; Llama-3.1-8B and Qwen2-7B follow the same pattern (Appendix E).

Compression and retrieval quality simultaneously. At $r=80$ ($32\times$ compression of Qwen1.5-4B’s 2560-dim hidden state), NCWP achieves 57.28 nDCG@10—+13.57 over the raw $D=2560$ Base (43.71). Even at $r=40$ ($64\times$ compression), NCWP remains above Base by +1.26 nDCG, whereas PCA-whitening falls to 33.14. This is a clean demonstration of NCWP’s intended use case:

Table 2: **nDCG@10 ($\times 100$) on BEIR-Quora** with Qwen1.5-4B ($D = 2560$). All post-hoc methods fit on the same 2,000 corpus items; test queries are excluded. NCWP reports mean over 3 random seeds (std ≤ 0.34), exceeding Base by +13.6 at $r=80$ and +1.3 at $r=40$. Best in each column in **bold**. The full sweep, down to $512\times$ compression, is in Table 10.

Method	$r=320$	160	80	40
PCA	56.26	49.33	41.43	33.14
Soft-W	49.69	43.42	37.07	30.47
Random	43.62	42.35	40.98	35.14
LPP	33.76	28.87	24.33	19.45
ZCA	56.72	49.89	41.85	32.93
NCWP	64.86	62.46	57.28	44.97
Raw Base ($D=2560$, uncompressed): 43.71.				

for frozen decoder LLM hidden states, NCWP can deliver index-size reductions and absolute retrieval improvements simultaneously.

Crossover sharpens, but the usable range is narrower. In the moderate range the NCWP-over-PCA gap is larger on Quora than on STSBenchmark (+15.9 vs. +6.5 at $32\times$), consistent with ranking a 522,931-item corpus under binary relevance rather than scoring 1,379 labeled pairs. At the extreme end the corpus size acts in the opposite direction: on STSBenchmark NCWP leads every baseline at every dimension we tested (to $1024\times$), whereas on Quora it leads only to $128\times$, below which every method collapses (Appendix E).

4.4 Mechanism: Anisotropy Reduction

To complement the retrieval-quality results with geometric evidence, we measure two sentence-level diagnostics. These are inspired by the anisotropy analysis of Ethayarajh (2019) but are not the same quantities: that work measures self-similarity of one word across contexts and an anisotropy baseline over random word representations, whereas we define **anisotropy_mean** as the average cosine similarity between an embedding and the corpus-mean direction, and **self_sim** as the average cosine similarity between distinct sentence embeddings. For a representation well-suited to cosine retrieval, both should be near zero.

NCWP substantially reduces anisotropy across LLM backbones. The raw mean-pooled hidden states show strong anisotropy across all three backbones (anisotropy_mean 0.67–0.80), consistent with prior token-level anisotropy findings on decoder-only models (Ethayarajh, 2019). After NCWP, anisotropy_mean falls by roughly an order of magnitude—to ≤ 0.03 on Qwen1.5-4B and

Table 3: **Anisotropy diagnostics across three LLM backbones.** Measured on the 2,910 STSBenchmark validation sentences (L2-normalized), which are also the fitting sentences, so these are in-sample geometric diagnostics rather than held-out performance. For NCWP we report the range over the full sweep of target dimensions r (Appendix F). NCWP substantially reduces anisotropy across all backbones; the residual is near zero on Qwen1.5-4B and Llama-3.1-8B.

Backbone	D	Base	NCWP
<i>anisotropy_mean</i>			
Qwen1.5-4B	2560	0.797	[0.008, 0.014]
Llama-3.1-8B	4096	0.666	[0.010, 0.027]
Qwen2-7B	3584	0.771	[0.069, 0.089]
<i>self_sim</i>			
Qwen1.5-4B	2560	0.639	[−0.003, 0.008]
Llama-3.1-8B	4096	0.444	[−0.004, 0.006]
Qwen2-7B	3584	0.597	[0.005, 0.010]

Llama-3.1-8B, and 0.07–0.09 on Qwen2-7B—and self_sim drops to within ± 0.01 of zero on all three backbones. The residual anisotropy on Qwen2-7B is the only deviation; we revisit it in the Limitations. Full per- r measurements: Appendix F.

4.5 Ablation

We ablate each NCWP component against simpler variants. The component run evaluated the 7,128 STSBenchmark train+test pairs—every pair outside the fitting split—so we keep it separate from the official-test headline table. Three of its four rows (PCA-whitening, ZCA-only, NCWP full) also appear on the official split in Table 7, with the same ordering and the same ZCA-over-PCA gap at $r=20$ (+3.0); only ZCA+contrastive was not re-run there.

Table 4: **Component diagnostic on Qwen1.5-4B STSBenchmark train+test pairs** (7,128 pairs; Spearman $\times 100$). Separate from the official-test headline evaluation. Last two rows report mean over 3 random seeds. NCWP (full) uses ZCA-shrink whitening, neighbor-contrastive projection, hard-negative-augmented InfoNCE, and periodic QR retraction.

Method	$r=20$	$r=80$	$r=320$
PCA-whitening	33.6	46.5	53.4
ZCA-only	37.9	46.2	52.9
ZCA + contrastive	46.9	53.0	54.3
NCWP (full, +QR)	47.0	53.4	55.8

Component-level contributions. Table 4 traces the path from a closed-form baseline to the full

NCWP. (i) Replacing PCA-whitening with ZCA-shrink alone lifts the most aggressive setting ($r=20$: 33.6 $\rightarrow$ 37.9), consistent with ZCA preserving more local geometric structure. (ii) Adding the neighbor-contrastive projection (without QR) gives the dominant lift at $r=20$ (37.9 $\rightarrow$ 46.9), so the learned $\mathbf{W}$ extracts retrieval-relevant structure beyond closed-form whitening. (iii) Periodic QR retraction adds a small but consistent improvement, largest at moderate dimensions (+0.4 at $r=80$, +1.5 at $r=320$). The same ranking carries over to BEIR-Quora. The covariance and orthogonality penalties of earlier versions contribute nothing measurable once QR retraction is present and are not part of the default objective (Table 18).

The gains come from *neighbor* mining, and the mining premise holds. Any positive-generation scheme could drive the contrastive stage, so we retrain $\mathbf{W}$ with an identical loss, schedule, and budget, changing only the positives to randomly paired items. On a 50,000-item Quora fit the difference is decisive at aggressive compression: k -NN positives reach 44.4 nDCG@10 at $r=20$ against 23.2 for random, matching the full-dimensional raw Base (43.71) where random collapses. Whitened-space k -NN also approximates relevance against gold labels: the nearest mined neighbor is a true positive for 57% of Quora and 71% of STSBenchmark anchors, against 37% and 56% in the raw space (Appendices I and K).

4.6 Generalization Beyond STSBenchmark and Quora

Results on two testbeds could reflect properties of those two datasets rather than of the setting. We therefore add STS12–16+SICK-R and a second symmetric duplicate-retrieval corpus, CQADupStack (Table 5), and test transfer without refitting. For this breadth analysis, each STS transform is fit on the unlabeled sentences of the evaluated benchmark (including all 9,927 SICK-R pairs), so these values are transductive and should not be read as official test-only scores.

The crossover replicates; the exceeds-Base margin is dataset-dependent. On both additional benchmarks NCWP is ahead of PCA-whitening in the aggressive-compression regime and the two converge at the largest r , as on STSBenchmark. The exceeds-Base finding is not unconditional, though: on CQADupStack the raw states are near-degenerate (Base nDCG@10 = 5.8), so the relative

Table 5: **Breadth analyses beyond the two headline testbeds** (Qwen1.5-4B). STS12–16/SICK-R uses transductive unlabeled fitting on each benchmark; retrieval rows are nDCG@10. Base is dimension-free; r is the target dimension for PCA and NCWP. Cross-corpus transfer is in Appendix H.

Benchmark	Base	PCA	NCWP
STS12–16+SICK-R avg ($r=40$)	43.8	47.5	53.1
CQADupStack avg ($r=320$)	5.8	15.4	17.7

margin is wide while absolute numbers stay low and the crossover sits at moderate rather than extreme dimensions.

Refitting is helpful but not always required. A projection fit on STSBenchmark and applied unchanged to STS12–16/SICK-R stays above Base in 17 of 18 backbone $\times$ dataset cells, roughly 3.5 Spearman below in-domain refitting on average; Quora $\rightarrow$ CQADupStack also stays above Base on the three reported subforums (Table 16). This supports transfer within a task family but is not a general cross-task guarantee.

4.7 Comparison with Training-Free Extraction Recipes

Because our Base uses mean pooling, we also compare NCWP with training-free pooling, layer, and prompt-based extraction from the same frozen backbone.

Pooling and layer choice. Mean pooling beats last-token pooling at the raw Base on all three backbones, so our Base is not a straw baseline. On Qwen1.5-4B the second-to-last layer improves the raw Base (36.6 vs. 31.3), and NCWP stays effective on alternative layer and pooling choices (Appendix L).

Prompt-based extraction. NCWP on mean-pooled states does *not* beat PromptEOL, a stronger training-free recipe. The two are nevertheless complementary: adding NCWP improves full-dimensional PromptEOL at $64\times$ compression on all three backbones (Table 6). On Echo the composition is weaker—NCWP matches full-dimensional Echo+ZCA at $8\times$ compression rather than exceeding it (Appendix L).

5 Discussion

Why post-hoc neighbor-contrastive alignment works on frozen LLMs. That a backbone-frozen

Table 6: **PromptEOL + NCWP** on the STSBenchmark test (Spearman $\times 100$). NCWP uses $64\times$ compression ($r=40/64/56$).

Method	Qwen1.5	Llama-3.1	Qwen2
PromptEOL, full D	60.2	72.3	72.3
+ NCWP, $64\times$	67.7	77.0	76.0

linear post-hoc transformation can compress *and* exceed raw Base follows from a three-step chain. (i) Cosine ranking on raw hidden states is distorted by anisotropy (mean cosine 0.67–0.80), producing the low raw-Base scores (31.31 on STSBenchmark test, 43.71 on Quora). (ii) In the ZCA-shrink-whitened space, k -NN approximates relevance—a usable label-free signal, measured in Section 4.5. (iii) A projection trained against it preserves those neighborhoods, so the output ranking exceeds raw Base at $r \ll D$.

What “exceeding Base” does and does not mean.

Raw Base is not adapted to the target corpus in any way, so exceeding it is not a claim that NCWP beats a task-tuned system. It establishes something narrower: at these dimensions the binding constraint on frozen hidden states is their geometry, not their semantic content. The group we claim to compete in is the label-free, frozen-backbone family of Section 4.1. Within it, PromptEOL and Echo act on a different part of the pipeline—they change how the D -dimensional vector is produced, NCWP how it is transformed and compressed—which is why stacking them can beat either alone under compression. A second, practically actionable corollary: PCA-whitening should not be the only closed-form whitening baseline, since ZCA-only matches or beats it at several aggressive dimensions (though without a single backbone-independent threshold), so post-hoc proposals reporting gains over PCA-whitening at low projected dimensions should also compare against ZCA-shrink.

6 Conclusion

We presented **NCWP**, a label-free post-hoc alignment method for symmetric retrieval: the backbone stays frozen and one linear projection is learned from unlabeled text. Across three decoder-only backbones it improves over raw Base and standard post-hoc baselines under aggressive compression, reduces anisotropy, and composes with prompt-based extraction.

Limitations

Symmetric retrieval, corpus-adaptive scope. NCWP’s neighbor-mined positives require base-space k -NN to approximate relevance (queries and documents from the same distribution); we do not claim transfer to asymmetric retrieval. NCWP is also corpus-adaptive—the projection $\mathbf{W}$ is fit on the target corpus distribution. Section 4.6 shows transfer is usable within a task family, but the evidence is limited to the reported STS and duplicate-question corpora and should not be read as a general guarantee.

Dataset-dependent margins; baselines not run.

The exceeds-Base result is strongest where the raw states are informative; on CQADupStack the raw geometry is close to degenerate and absolute numbers remain low for every method. We did not include BERT-flow (Li et al., 2020) in the evaluated baselines, and the Echo comparison is limited to STSBenchmark.

Regime and recipe limits. NCWP’s advantage over closed-form alternatives is concentrated at small target dimensions; near native D , NCWP, ZCA-only, and PCA-whitening sit within a narrow band. It also has a task-dependent lower bound. On STSBenchmark NCWP leads every baseline at every dimension we tested, to $1024\times$ compression; on BEIR-Quora it leads only to $128\times$, and at $256\times$ and beyond every method collapses to unusable $n\text{DCG}@10$ with NCWP degrading faster than PCA-whitening (Appendix E). The usable range must therefore be read off the retrieval task, not the STS task. On mean-pooled states NCWP also does not exceed PromptEOL at full dimension: our claim is compositional—NCWP improves whatever frozen representation it is given, under compression—not that mean pooling plus NCWP is the best training-free recipe available.

Residual anisotropy on Qwen2-7B; no fine-tuning comparison. NCWP drives anisotropy_mean to near zero on Qwen1.5-4B and Llama-3.1-8B; the residual on Qwen2-7B is larger (0.07–0.09), though retrieval quality remains strong (Appendix D). We do not directly compare against fine-tuning-based LLM embedding models, which lie at a different operating point from our frozen-backbone, label-free setting.

Acknowledgments

This work was partially supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-01842, Artificial Intelligence Graduate School Program (GIST)); Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea Government (MOTIE) [P0020535, The Competency Development Program for Industry Specialist]; the InnoCORE program of the Ministry of Science and ICT (25-InnoCORE-01); and by the ANCHOR program through the Gwangju ANCHOR Center, funded by the Ministry of Education (MOE) and the Jeonnam-Gwangju Special Metropolitan City, Republic of Korea (2026-ANCHOR-05-001).

References

- Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. [SemEval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability](#). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263. Association for Computational Linguistics.
- Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. [SemEval-2014 task 10: Multilingual semantic textual similarity](#). In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91. Association for Computational Linguistics.
- Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. [SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation](#). In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pages 497–511. Association for Computational Linguistics.
- Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. [SemEval-2012 task 6: A pilot on semantic textual similarity](#). In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pages 385–393. Association for Computational Linguistics.
- Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. [*SEM 2013 shared task: Semantic textual similarity](#). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Simi-*

- larity, pages 32–43. Association for Computational Linguistics.
- Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. [A simple but tough-to-beat baseline for sentence embeddings](#). In *International Conference on Learning Representations (ICLR)*.
- Adrien Bardes, Jean Ponce, and Yann LeCun. 2022. [VICReg: Variance-invariance-covariance regularization for self-supervised learning](#). In *International Conference on Learning Representations (ICLR)*.
- Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. [LLM2Vec: Large language models are secretly powerful text encoders](#). In *Conference on Language Modeling (COLM)*.
- Ella Bingham and Heikki Mannila. 2001. [Random projection in dimensionality reduction: Applications to image and text data](#). In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 245–250.
- Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. [SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation](#). In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. [A simple framework for contrastive learning of visual representations](#). In *International Conference on Machine Learning (ICML)*, pages 1597–1607.
- Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljačić, Shang-Wen Li, Wen-tau Yih, James Glass, and Yoon Kim. 2022. [DiffCSE: Difference-based contrastive learning for sentence embeddings](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4207–4218.
- Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. 2021. [With a little help from my friends: Nearest-neighbor contrastive learning of visual representations](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9588–9597.
- Kawin Ethayarajh. 2019. [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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65.
- Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. [SimCSE: Simple contrastive learning of sentence embeddings](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.
- Aaron Grattafiori and 1 others. 2024. [The Llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Xiaofei He and Partha Niyogi. 2003. [Locality preserving projections](#). In *Advances in Neural Information Processing Systems*, volume 16.
- Doris Hoogeveen, Karin M. Verspoor, and Timothy Baldwin. 2015. [CQADupStack: A benchmark data set for community question-answering research](#). In *Proceedings of the 20th Australasian Document Computing Symposium*, pages 3:1–3:8.
- Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. [Unsupervised dense information retrieval with contrastive learning](#). *Transactions on Machine Learning Research*.
- Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. 2024. [Scaling sentence embeddings with large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3182–3196, Miami, Florida, USA. Association for Computational Linguistics.
- Ting Jiang, Jian Jiao, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Denvy Deng, and Qi Zhang. 2022. [Prompt-BERT: Improving BERT sentence embeddings with prompts](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8826–8837.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. [Billion-scale similarity search with GPUs](#). *IEEE Transactions on Big Data*, 7(3):535–547.
- William B. Johnson and Joram Lindenstrauss. 1984. [Extensions of Lipschitz mappings into a Hilbert space](#). *Contemporary Mathematics*, 26:189–206.
- Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.
- Agnan Kessy, Alex Lewin, and Korbinian Strimmer. 2018. [Optimal whitening and decorrelation](#). *The American Statistician*, 72(4):309–314.
- Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. 2022. [Metryoshka representation learning](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 30233–30249.
- Olivier Ledoit and Michael Wolf. 2004. [A well-conditioned estimator for large-dimensional covariance matrices](#). *Journal of Multivariate Analysis*, 88(2):365–411.

- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive NLP tasks](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.
- Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. [On the sentence embeddings from pre-trained language models](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130.
- Yu A. Malkov and D. A. Yashunin. 2020. [Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):824–836.
- Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. [A SICK cure for the evaluation of compositional distributional semantic models](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 216–223.
- Jiaqi Mu and Pramod Viswanath. 2018. [All-but-the-Top: Simple and effective postprocessing for word representations](#). In *6th International Conference on Learning Representations (ICLR)*.
- Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. [MTEB: Massive text embedding benchmark](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037.
- Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. 2022. [Large dual encoders are generalizable retrievers](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855.
- Qwen Team. 2024. Introducing Qwen1.5. <https://qwenlm.github.io/blog/qwen1.5/>. Model configuration: <https://huggingface.co/Qwen/Qwen1.5-4B/blob/main/config.json>.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.
- Jacob Springer, Suhas Kotha, Daniel Fried, Graham Neubig, and Aditi Raghunathan. 2025. [Repetition improves language model embeddings](#). In *International Conference on Learning Representations*.
- Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. [Whitening sentence representations for better semantics and faster retrieval](#). *arXiv preprint arXiv:2103.15316*.
- Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. [BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models](#). In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. [Representation learning with contrastive predictive coding](#). *arXiv preprint arXiv:1807.03748*.
- Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. [Text embeddings by weakly-supervised contrastive pre-training](#). *arXiv preprint arXiv:2212.03533*.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. [Improving text embeddings with large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916.
- Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. [C-Pack: Packed resources for general Chinese embeddings](#). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 641–649.
- Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. [Approximate nearest neighbor negative contrastive learning for dense text retrieval](#). In *International Conference on Learning Representations (ICLR)*.
- Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. [ConSERT: A contrastive framework for self-supervised sentence representation transfer](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5065–5075.
- An Yang and 1 others. 2024. [Qwen2 technical report](#). *arXiv preprint arXiv:2407.10671*.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. [Qwen3 Embedding: Advancing text embedding and reranking through foundation models](#). *arXiv preprint arXiv:2506.05176*.

A Algorithm: Full NCWP Procedure

Algorithm 1: Neighbor-Contrastive Whitening Projection.

1. **Input:** frozen embeddings $\{\mathbf{x}_n\}_{n=1}^N$, target dimension r , and hyperparameters $\lambda, k, \tau, T, M, K, N_{\text{qr}}$.
2. Compute $\boldsymbol{\mu}$, sample covariance $\hat{\boldsymbol{\Sigma}}$, shrunk covariance $\hat{\boldsymbol{\Sigma}}_\lambda$, and ZCA matrix $\mathbf{S} = \mathbf{U}\boldsymbol{\Lambda}^{-1/2}\mathbf{U}^\top$.
3. Whiten $\tilde{\mathbf{x}}_i \leftarrow \text{norm}((\mathbf{x}_i - \boldsymbol{\mu})\mathbf{S})$; build $\mathcal{N}_k(i)$ and form $\mathcal{P}(i)$ using threshold τ .
4. Initialize $\mathbf{W} \sim \mathcal{N}(0, 1/D)$ and an empty FIFO memory bank $\mathcal{M}$ of capacity M .
5. For each mini-batch $\mathcal{B}$, sample $i^+ \in \mathcal{P}(i)$, select the top- K hard negatives from in-batch negatives and $\mathcal{M}$, minimize $\mathcal{L}_{\text{NCE}}$ with AdamW, and push the current projections into $\mathcal{M}$.
6. Every N_{qr} steps, compute $\mathbf{W} = \mathbf{Q}\mathbf{R}$ and set $\mathbf{W} \leftarrow \mathbf{Q}$.
7. **Output:** $(\boldsymbol{\mu}, \mathbf{S}, \mathbf{W})$.

B Hyperparameters

Development setting. The full configuration below was fixed once on the STSBenchmark development/fitting split and then held constant across every backbone, dataset, and target dimension reported in this paper; there is no per-dataset hyperparameter tuning. The headline STSBenchmark test split and Quora test queries are disjoint from fitting. The breadth-only STS analysis uses transductive unlabeled fitting as stated in Section 4.6, and projection training never uses relevance labels.

NCWP. Shrinkage $\lambda = 0.08$; neighborhood size $k = 10$; cosine threshold $\tau = 0$ in the whitened space; temperature $T = 0.12$ with cosine annealing; hard-negative selection $K = 256$ per anchor; memory bank size $M = 4096$ (FIFO); QR retraction every $N_{\text{qr}} = 10$ steps; one refinement round; AdamW with cosine schedule and 100-step linear warmup, learning rate 8×10^{-3} , weight decay 1×10^{-4} ; batch size 256; at most $E = 20$ epochs. Instrumentation of the main STS trainer confirms that the bank reaches 4096/4096 occupancy, expands the negative-candidate pool from 512 to 4,608, and supplies 87.9–88.4% of the selected top- K hard negatives across nine measured conditions.

Optional stabilizers. When the covariance and orthogonality penalties of Appendix J are enabled, we use $\lambda_{\text{cov}} = 0.05$ and $\lambda_{\text{orth}} = 0.02$. They are disabled in every result reported in this paper.

Backbone extraction. Last hidden state, mean pooling over valid tokens, ℓ_2 normalization, bf16, maximum sequence length 128. Checkpoints: Qwen/Qwen1.5-4B, Qwen/Qwen2-7B, meta-llama/Meta-Llama-3.1-8B.

Baselines. Random projection uses i.i.d. Gaussian entries $\mathcal{N}(0, 1/r)$. PCA-whitening and ZCA-only share the eigendecomposition and shrinkage of Stage 1, differing only in the rotation. The remaining baselines each have one free parameter, fixed once and reused across all backbones and target dimensions. **Soft-whitening** applies PCA-whitening with eigenvalue shrinkage $\alpha = 0.1$ toward the scaled identity, which softens the inverse-square-root amplification of small eigenvalues; it is thus the PCA-rotation counterpart of our ZCA-shrink Stage 1, so the gap between the two isolates the effect of the rotation choice. **LPP** uses a symmetric k -NN graph with neighborhood size 10 and binary (0/1) affinity. For **ABTT**, we estimate the fitting-set mean and top $m=5$ principal components, remove those components, and ℓ_2 -normalize. **ABTT+PCA** then recenters the normalized ABTT-transformed fitting embeddings, fits a rank- r PCA projection, and applies the same ABTT, recentering, projection, and final ℓ_2 normalization to test embeddings. Its output dimension is therefore exactly r .

Release. Code, the exact fit indices, and the random seeds $\{42, 43, 44\}$ used for all three-seed means are released at <https://github.com/ianleeGH20/NCWP>.

C Computational Cost

On a single A100 GPU with LLM hidden states ($D = 2560$ – 4096) and the corpus subsamples used in our experiments (2,910 STSBenchmark validation sentences for STSBenchmark; 1,000–2,000 BEIR-Quora corpus items), total post-hoc fitting takes approximately 5–15 minutes (whitening eigendecomposition: seconds; projection training: 5–15 minutes). PCA-whitening on the same corpus completes in under one second. Inference cost is identical to PCA-whitening at the same target r : a single $D \times r$ matrix multiplication and ℓ_2 normalization per embedding. Memory: NCWP

retains $(\mu, \mathbf{S}, \mathbf{W})$, totaling $D + D^2 + Dr$ floats; for deployment, μ and $\mathbf{SW}$ ($D + Dr$ floats) suffice.

D Full LLM STSBenchmark Results

Tables 7, 8, and 9 report STSBenchmark official-test results on Qwen1.5-4B, Llama-3.1-8B, and Qwen2-7B. All NCWP rows report mean over 3 random seeds; per-seed std is shown alongside each cell. ABTT and ABTT+PCA are deterministic. Full-dimensional ABTT scores 48.06, 56.65, and 48.78 on the three backbones, respectively; the tables report the dimension-matched ABTT+PCA results. Each table sweeps the full grid we ran, down to $512\times$, $1024\times$, and $512\times$ compression respectively.

NCWP is the best of the seven methods in every column of all three tables, at every target dimension on every backbone, including the most extreme. Exceeding the *uncompressed* Base is a separate and narrower claim: NCWP clears it by wide margins in the aggressive range (Qwen1.5-4B +23.8 at $r=320$; Llama-3.1-8B +15.5 at $r=128$; Qwen2-7B +18.3 at $r=448$) but falls below it at the very smallest dimensions ($r=5$ on Qwen1.5-4B; $r=8$ and 4 on Llama-3.1-8B; $r=7$ on Qwen2-7B), where every method has lost most of the signal. Random projection and LPP stay below NCWP everywhere but are far flatter in r than the whitening-based methods, so at each backbone’s smallest dimensions random projection overtakes PCA-whitening while remaining well below NCWP.

E Full LLM Quora Results

Tables 10–12 give the complete BEIR-Quora sweeps on all three backbones, extending the main-paper Table 2 down to the smallest dimensions we ran. Qwen1.5-4B is fit on 2,000 corpus items, Llama-3.1-8B and Qwen2-7B on 1,000. All NCWP rows report mean $\pm$ std over 3 random seeds; the best entry in each column is in **bold**, including the columns where that entry is not NCWP.

The operating range ends at $128\times$. Read as compression factors, the three sweeps behave the same way. NCWP is the best method at every compression factor up to and including $128\times$ ($r=20$, 32, and 28 respectively). At $256\times$ and beyond, all methods fall to nDCG@10 that is unusable for a 522,931-item corpus—below 26 on every backbone, and near zero at the extreme—and NCWP

degrades faster there than PCA-whitening and random projection, so it is no longer the top method. The ordering in that collapsed regime is between systems none of which would be deployed, but $128\times$ is the boundary of the range we can recommend.

On STSBenchmark NCWP instead stays ahead down to $512\times$ on Qwen1.5-4B (Table 7). Scoring 1,379 sentence pairs tolerates a far smaller subspace than ranking a half-million-item corpus, so the usable compression range is task-dependent and should be read off the retrieval task, not the STS task. Exceeding raw Base is also dimension-dependent: at the most extreme Llama-3.1-8B settings NCWP falls below that backbone’s raw Base of 52.86.

F Full Anisotropy Measurements

Table 3 in the main paper reports the range of NCWP anisotropy diagnostics across target dimensions. Tables 13–15 report the per- r values.

G Method Schematic

H Cross-Corpus Transfer

Section 4.6 summarizes what happens when an NCWP projection fit on one corpus is applied unchanged to a related corpus. *Transfer* reuses a projection fit elsewhere, whereas *refit* fits on the target corpus. We report only within-task-family transfers here; the STS breadth protocol is transductive as described in Section 4.6.

I Positive-Source Control

J Optional Soft Regularizers

Earlier versions of NCWP added a VICReg-inspired covariance penalty (Bardes et al., 2022), $\mathcal{L}_{\text{cov}} = \|\text{Cov}(\mathbf{Z}) - \mathbf{I}_r\|_F^2$, and an additional orthogonality penalty, $\mathcal{L}_{\text{orth}} = \|\mathbf{W}^\top \mathbf{W} - \mathbf{I}_r\|_F^2$, weighted by small α, β and intended as safeguards against variance collapse under aggressive compression. We audited their effect under the final STSBenchmark test-only and Quora corpus-sample protocols.

On STS-B, the maximum standard deviation across these cells is 1.01, so the observed changes fall within seed variation. Quora values differ in the third or fourth decimal place before rounding, confirming that the setting is applied but has negligible effect on top-10 rankings. Removing QR retraction, by contrast, worsens the conditioning

Table 7: STSBenchmark official test (Spearman $\times$ 100) on **Qwen1.5-4B** ($D = 2560$, Base = 31.31). NCWP reports mean $\pm$ std over seeds {42, 43, 44}.

Method	$r=1280$	640	320	160	80	40	20	10	5
PCA-white	55.54	53.61	51.87	49.90	46.01	40.44	34.31	23.42	15.31
Soft-white	51.68	49.32	46.99	44.44	40.76	36.13	30.42	21.50	14.49
ZCA-only	55.77	54.02	51.28	47.64	42.51	38.40	37.31	32.33	26.49
ABTT+PCA	47.27	46.18	44.99	43.38	40.93	37.89	33.75	28.20	21.53
Random	37.81	38.39	37.34	37.74	36.31	36.24	31.83	30.47	22.91
LPP	38.01	35.76	33.84	31.94	30.32	28.36	24.22	17.77	14.37
NCWP	56.52 $\pm$ 0.15	55.55 $\pm$ 0.26	55.08 $\pm$ 0.38	54.18 $\pm$ 0.51	52.48 $\pm$ 0.19	49.02 $\pm$ 1.70	44.92 $\pm$ 0.77	37.44 $\pm$ 1.33	27.97 $\pm$ 1.32

Table 8: STSBenchmark official test (Spearman $\times$ 100) on **Llama-3.1-8B** ($D = 4096$, Base = 45.89). NCWP reports mean $\pm$ std over 3 random seeds.

Method	$r=2048$	1024	512	256	128	64	32	16	8	4
PCA-white	62.97	60.97	58.90	56.75	53.61	49.27	42.59	39.47	28.26	23.54
Soft-white	60.54	58.08	55.66	53.15	50.07	46.39	40.84	37.10	27.16	22.98
ZCA-only	63.07	61.97	60.64	61.43	58.62	56.16	50.31	43.24	39.66	27.98
ABTT+PCA	55.53	54.39	53.19	51.72	50.07	47.89	45.43	43.19	38.26	29.49
Random	49.85	49.92	49.99	48.85	47.13	48.13	44.27	43.00	39.52	29.00
LPP	48.70	45.84	44.52	43.31	41.74	39.57	36.97	33.70	27.28	20.13
NCWP	63.35 $\pm$ 0.11	63.38 $\pm$ 0.13	62.94 $\pm$ 0.49	62.58 $\pm$ 0.20	61.36 $\pm$ 0.22	59.27 $\pm$ 0.08	56.21 $\pm$ 0.54	51.30 $\pm$ 0.81	43.97 $\pm$ 1.21	32.79 $\pm$ 2.39

Table 9: STSBenchmark official test (Spearman $\times$ 100) on **Qwen2-7B** ($D = 3584$, Base = 39.27). NCWP reports mean $\pm$ std over 3 random seeds.

Method	$r=1792$	896	448	224	112	56	28	14	7
PCA-white	58.45	56.77	55.25	53.90	51.20	46.11	39.78	32.44	22.66
Soft-white	54.64	52.67	50.86	48.78	45.99	41.95	37.05	31.20	23.02
ZCA-only	58.52	57.10	55.09	52.30	49.14	46.31	40.89	35.21	28.50
ABTT+PCA	48.03	47.20	46.31	45.26	43.79	41.58	38.45	35.84	31.33
Random	41.38	41.19	40.53	40.32	41.29	40.32	40.40	35.00	29.70
LPP	40.81	38.87	37.99	37.04	35.84	34.89	32.47	29.35	24.92
NCWP	58.94 $\pm$ 0.02	58.33 $\pm$ 0.26	57.62 $\pm$ 0.11	56.61 $\pm$ 0.14	55.75 $\pm$ 0.82	54.21 $\pm$ 0.47	51.46 $\pm$ 0.83	45.23 $\pm$ 0.51	37.76 $\pm$ 1.27

and residual-anisotropy diagnostics in the separate component analysis. We therefore use only the InfoNCE objective in all camera-ready results and retain the two terms as optional stabilizers.

K Quality of the Mined Positives

Section 4.5 summarizes the label precision of the mined sets; Table 19 gives the full numbers, and the three supporting analyses follow.

Gold-STs distribution of mined neighbors. Table 20 compares the gold STS score of mined pairs against randomly drawn pairs. Mined neighbors are substantially more semantically related than random pairs in every space. Note that this distribution does not by itself separate the whitened space from the raw one—that separation is visible in the precision numbers of Table 19, not here.

Sensitivity to k and τ . Precision decreases with k , as expected when each anchor has roughly one gold neighbor: on Quora in the ZCA-whitened space, $P@1 = 0.57$ falls to $P@10 = 0.11$. The cosine threshold is far less consequential: varying τ from 0 to 0.3 changes downstream retrieval

nDCG@10 by less than 0.3, so we keep $\tau = 0$ throughout.

Failure cases. The mining rule can also select pairs that are lexically close but semantically far. Screening the mined sets for gold $STS \leq 2$ combined with token-level Jaccard overlap ≥ 0.5 (and, on Quora, non-duplicate pairs with high overlap) yields 321 such candidates on STSBenchmark. They fall into the same two patterns the qualitative examples of Appendix M illustrate: a shared syntactic frame with different content, and a single shared content word in different predicates. Because these pairs enter training as positives, they are the mechanism by which the projection can retain a residual lexical bias; that they number in the hundreds out of a 2,910-sentence fit is consistent with the mined-set precision in Table 19. The full list is released with our code.

L Full Training-Free Extraction Results

Section 4.7 summarizes the comparison against pooling, layer selection, and prompt-based extraction. This appendix gives the underlying numbers,

Table 10: nDCG@10 ($\times 100$) on **BEIR-Quora** with **Qwen1.5-4B** ($D = 2560$, Base = 43.71). Fit on 2,000 corpus items; NCWP reports mean $\pm$ std over 3 seeds. Compression factors run from $2\times$ ($r=1280$) to $512\times$ ($r=5$).

Method	$r=1280$	640	320	160	80	40	20	10	5
PCA-white	65.22	61.64	56.26	49.33	41.43	33.14	23.10	11.99	1.73
ZCA-only	65.12	61.89	56.72	49.89	41.85	32.93	21.00	7.72	0.86
Soft-white	60.02	55.10	49.69	43.42	37.07	30.47	21.68	11.56	1.71
Random	44.82	44.36	43.62	42.35	40.98	35.14	24.42	9.50	1.00
LPP	44.47	39.06	33.76	28.87	24.33	19.45	14.71	8.10	1.24
NCWP	66.65 ± 0.06	66.11 ± 0.03	64.86 ± 0.10	62.46 ± 0.13	57.28 ± 0.34	44.97 ± 0.18	25.03 ± 0.26	7.14 ± 0.05	0.50 ± 0.03

Table 11: nDCG@10 ($\times 100$) on **BEIR-Quora** with **Llama-3.1-8B** ($D = 4096$, Base = 52.86). Fit on 1,000 corpus items; NCWP reports mean $\pm$ std over 3 seeds.

Method	$r=2048$	1024	512	256	128	64	32	16	8	4
PCA-white	69.62	66.37	61.21	55.20	48.69	41.50	32.85	22.40	11.57	1.07
ZCA-only	66.25	66.25	60.71	54.84	47.65	40.15	30.48	18.15	5.06	0.21
Soft-white	66.29	61.10	55.33	50.02	44.62	38.92	31.75	22.23	11.42	1.00
Random	53.11	53.06	52.80	51.84	49.68	47.06	39.87	25.10	6.58	0.23
LPP	53.30	49.41	44.97	39.95	35.04	29.52	23.21	15.10	5.14	0.24
NCWP	69.87 ± 0.01	69.59 ± 0.07	68.76 ± 0.18	67.05 ± 0.07	63.52 ± 0.48	57.21 ± 0.18	43.74 ± 0.61	21.40 ± 0.25	4.68 ± 0.13	0.17 ± 0.04

split so that each table covers exactly the configurations we ran.

Pooling rule. Table 21 compares the two pooling rules at the raw Base on all three backbones. Mean pooling is stronger at the raw Base on every backbone, by 1.9–6.5 Spearman. On Qwen1.5-4B at $r=80$, NCWP reaches 52.48 from mean-pooled states (Table 7) and 48.46 from last-token states; both improve substantially over their corresponding raw Base.

Layer selection. Table 22 sweeps the layer at the raw Base on Qwen1.5-4B with mean pooling. The second-to-last layer is better and a middle layer is substantially weaker. We keep the last layer throughout the main results for a fixed extraction recipe; applying NCWP to the second-to-last layer reaches 56.60 at $r=320$, compared with its raw Base of 36.63.

Echo embeddings. Table 23 reports the Echo comparison on the STSBenchmark official test split with Qwen1.5-4B. Whitening already helps substantially, and the learned projection approaches the full-dimensional Echo+ZCA score at $8\times$ compression while exceeding Echo+PCA-whitening at $r=80$. The Echo runs are STSBenchmark-only; the corresponding Quora sweep is left to future work.

M Qualitative Retrieval Examples

Figure 1 in the main paper shows one representative BEIR-Quora retrieval example where NCWP recovers the gold paraphrase from rank 1,034 to rank 1. This appendix provides additional qualitative ex-

amples on both BEIR-Quora and STSBenchmark, illustrating the two failure modes that NCWP corrects: (i) on retrieval, raw cosine similarity ranks lexically-overlapping but semantically-unrelated items highly, burying the gold answer; (ii) on STSBenchmark pairwise scoring, raw cosine similarity is compressed into a narrow high-similarity band (the anisotropy phenomenon), so even completely unrelated sentences score above 0.85. All examples use the same frozen Qwen1.5-4B mean-pooled hidden states and NCWP at $r=80$ ($32\times$ compression) fit on 2,000 unlabeled corpus items.

M.1 Additional BEIR-Quora Examples

We show three further BEIR-Quora examples spanning a range of base-space failure modes: surface lexical-template matching (Example G.1), syntactic-frame matching at the cost of semantic content (Example G.2), and topical retrieval that is just barely too coarse (Example G.3).

Example G.1. Query: “What is the easiest way to learn PPC?” (query ID 135,674); gold paraphrase: “How can I learn PPC?” (doc ID 135,673). Raw retrieves 5 unrelated “learn . . .” questions (rank of gold: #288); NCWP places the gold at rank #2.

Example G.2. Query: “What’s the best way to treat roach bites?” (query ID 254,177); gold paraphrase: “How do I treat a sore roach bite?” (doc ID 254,178). Raw retrieves on the syntactic frame “What is the best way to . . .” regardless of topic (rank of gold: #1,842). NCWP recovers the gold at rank #2 alongside other roach-related items.

Table 12: nDCG@10 ($\times 100$) on **BEIR-Quora** with **Qwen2-7B** ($D = 3584$, Base = 39.67). Fit on 1,000 corpus items; NCWP reports mean $\pm$ std over 3 seeds.

Method	$r=1792$	896	448	224	112	56	28	14	7
PCA-white	66.12	62.56	57.30	51.06	44.13	36.11	27.99	16.52	5.48
ZCA-only	63.23	62.47	57.25	50.99	43.00	34.88	25.02	12.86	2.38
Soft-white	60.28	54.87	49.12	43.66	38.55	32.90	26.15	16.10	5.44
Random	40.27	40.18	39.81	38.40	38.67	33.37	26.55	16.84	3.35
LPP	39.98	35.31	31.18	28.02	25.27	22.16	17.36	9.76	2.54
NCWP	67.02 ± 0.04	66.49 ± 0.06	65.57 ± 0.03	63.64 ± 0.05	59.77 ± 0.29	52.22 ± 0.29	37.04 ± 0.42	15.43 ± 0.25	2.33 ± 0.12

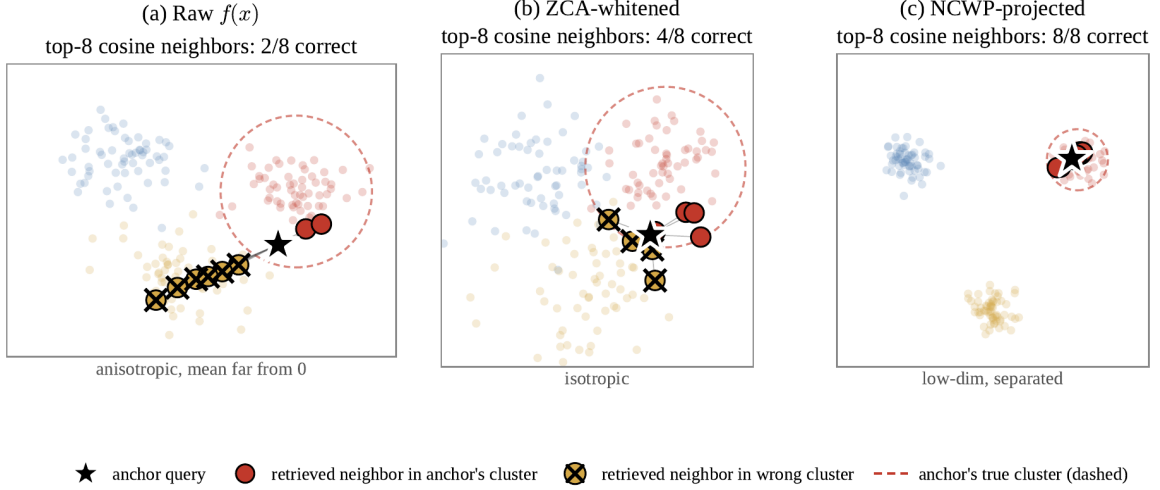


Figure 3: **Why post-hoc neighbor-contrastive alignment helps** (illustrative 2-D schematic, three semantic clusters; the anchor’s true cluster is the red one, marked by the dashed circle). **(a)** On raw anisotropic hidden states, cosine similarity is dominated by the corpus-mean direction, so most of the anchor’s top- k neighbors come from *wrong* clusters: the k -NN signal is largely uninformative. **(b)** After ZCA-shrink whitening, the corpus is centered and isotropic, and k -NN already *approximates* relevance—a substantial fraction of the top- k now land in the anchor’s true cluster, even though boundaries remain blurry. This is the label-free signal NCWP exploits. **(c)** A learned linear projection $\mathbf{W} \in \mathbb{R}^{D \times r}$ trained against this k -NN signal sharpens cluster separation, so top- k cosine retrieval becomes reliable in the low-dimensional output.

Table 13: **Qwen1.5-4B** ($D = 2560$). Base: anisotropy_mean = 0.7973, self_sim = 0.6388.

r	anisotropy_mean	self_sim
1280	0.0079	−0.0003
640	0.0079	0.0002
320	0.0078	0.0011
160	0.0083	−0.0003
80	0.0083	0.0008
40	0.0079	−0.0033
20	0.0084	−0.0033
10	0.0116	0.0008
5	0.0140	0.0052

Table 14: **Llama-3.1-8B** ($D = 4096$). Base: anisotropy_mean = 0.6658, self_sim = 0.4443.

r	anisotropy_mean	self_sim
2048	0.0098	−0.0004
1024	0.0108	−0.0000
512	0.0106	−0.0015
256	0.0101	0.0007
128	0.0098	−0.0012
64	0.0110	−0.0011
32	0.0104	0.0056
16	0.0114	−0.0018
8	0.0137	−0.0043
4	0.0270	0.0040

Example G.3. Query: “*Is the Ancient Alien Theory Correct?*” (query ID 153,281); gold paraphrase: “*What are the evidences to support the ancient alien theory?*” (doc ID 338,886). Raw retrieves a mix of unrelated “Is the ...” questions; NCWP recovers the gold at rank #3 together with topically-related items about ancient aliens.

Pattern. Across these examples (Figure 1, Examples G.1–G.3), raw retrieval is dominated by surface lexical overlap or syntactic templates (“*What is the best way to ...*”, “*Is the ...*”); the actual paraphrase, which uses different surface forms, is buried hundreds to thousands of positions deep. NCWP’s neighbor-contrastive projection appears

Table 15: **Qwen2-7B** ($D = 3584$). Base: anisotropy_mean = 0.7708, self_sim = 0.5971.

r	anisotropy_mean	self_sim
1792	0.0834	0.0061
896	0.0888	0.0076
448	0.0822	0.0075
224	0.0884	0.0064
112	0.0822	0.0065
56	0.0753	0.0050
28	0.0841	0.0101
14	0.0690	0.0052
7	0.0744	0.0103

Table 16: **Within-family cross-corpus transfer** (Qwen1.5-4B, $r=80$). STS is Spearman $\times 100$; retrieval is nDCG@10.

Fit $\rightarrow$ test	Base	transfer	refit
STS-B $\rightarrow$ STS12-16/SICK-R	43.8	51.6	55.1
Quora $\rightarrow$ CQADupStack	5.8	9.9	6.8

to suppress these lexical-template attractors and surface the topical match.

M.2 STSBenchmark: Raw Cosine Similarity Is Broken on Unrelated Pairs

The same anisotropy that produces poor retrieval rankings on BEIR-Quora is directly visible in raw cosine similarities on STSBenchmark: for sentence pairs whose gold similarity score is near 0 (i.e. semantically unrelated), raw cosine similarity is nevertheless above 0.85. The three examples in Table 24 are typical: a one-word overlap (“cutting”), a shared syntactic frame (“Two . . . in . . .”), or no overlap at all, all yield raw cosine above 0.86. After NCWP, cosine similarity matches the gold semantic judgment, dropping to within ± 0.15 of zero.

Connection to anisotropy. These examples make the anisotropy diagnostics in Table 3 concrete: when *anisotropy_mean* is near 0.8, every embedding is close (in cosine) to the corpus-mean direction, which mechanically forces all pairwise cosine similarities into a narrow band well above zero. NCWP whitens the corpus distribution, which decompresses this band and restores the dynamic range needed for cosine to discriminate related from unrelated pairs.

Table 17: **Positive source controls the aggressive-compression regime.** nDCG@10 on BEIR-Quora with Qwen1.5-4B, fit on 50,000 corpus items, identical loss/schedule/budget; only the positive source differs. At $r=20$, k -NN positives match the full-dimensional raw Base (43.71) while random positives collapse.

Positives	$r=80$ (32 $\times$)	$r=40$ (64 $\times$)	$r=20$ (128 $\times$)
Random	57.6	43.5	23.2
k -NN	60.5	55.1	44.4
Δ	+2.9	+11.6	+21.2

Table 18: **Removing both soft regularizers causes no measurable loss.** Differences are “without both” minus “with both” across 3 backbones, 4 dimensions, and 3 seeds.

Protocol	Mean Δ	Max. $ \Delta $	$ \Delta > 1$
STS-B test	-0.029	0.88	0/12
Quora corpus	-0.001	0.01	0/12

Table 19: **Precision of the mined positive sets $\mathcal{P}(i)$ against gold labels** (Qwen1.5-4B, $k=10$, $\tau=0$). Quora uses duplicate pairs; STSBenchmark uses pairs with gold score ≥ 4 . Ceiling for P@1 is 1.0; P@10 is bounded by the roughly one gold neighbor per anchor.

Dataset	Space	P@1	P@10	R@10
Quora (dup)	raw	0.37	0.07	0.53
	ZCA-whitened	0.57	0.11	0.80
STS-B (≥ 4)	raw	0.56	0.08	0.74
	ZCA-whitened	0.71	0.10	0.93

Table 20: **Gold STS scores of mined neighbors** (Qwen1.5-4B, STSBenchmark, $k=10$, $\tau=0$).

Pair source	mean gold STS	% with gold ≥ 4.0
Random pairs	2.58	22%
Raw k -NN	3.25	33%
ZCA k -NN	3.23	32%
NCWP k -NN	3.33	35%

Table 21: **Mean pooling versus last-token pooling at the raw Base** (STSBenchmark, Spearman $\times 100$, full D).

Pooling	Qwen1.5-4B	Llama-3.1-8B	Qwen2-7B
Last-token	28.74	39.43	37.37
Mean	31.31	45.89	39.27
Δ	+2.57	+6.46	+1.90

Table 22: **Layer choice at the raw Base** (Qwen1.5-4B, mean pooling, STSBenchmark, Spearman $\times$ 100, full D).

Layer	Base
Middle	13.9
Last	31.3
Second-to-last	36.6

Table 23: **NCWP on top of Echo embeddings** (Qwen1.5-4B, STSBenchmark, Spearman $\times$ 100).

Method	dim.	STS-B
Echo	full D	42.94
Echo + ZCA	full D	63.55
Echo + PCA-whitening	$r=80$	54.37
NCWP-on-Echo	$r=320$	63.17 ± 0.29

Raw top-5 (rank of gold: #288)

- × What is the easiest way to learn good photography?
- × What is the best way to learn programming?
- × What is the best way to start learn hacking?
- × What is the best way to learn SAS?
- × What is the best way to learn analytics?

NCWP top-5 (rank of gold: #2)

- × What is the easiest way to learn good photography?
- ✓ *How can i learn PPC?* ★ gold
- × What is the easiest way to become wealthy?
- × What is the best way to learn electrical engineering?
- × What is the easiest way to get in shape?

Raw top-5 (rank of gold: #1,842)

- × What is the best way to kill house flies?
- × What is the best way to cure of anxiety disorder?
- × What is the best way to get rid of acne?
- × What is the best way to get rid of a scar?
- × What is the best way to get rid of annoying room-mates?

NCWP top-5 (rank of gold: #2)

- × How do you get rid of roaches?
- ✓ *How do I treat a sore roach bite?* ★ gold
- × What's the best way to get sleep?
- × What is the best way to treat the flu?
- × What is the best way to cure of anxiety disorder?

Raw top-5 (rank of gold: #1,092)

- × Is Scott Gordon's "The God Entity: The Theory of Everything" a valid premise?
- × Is the New York Times failing?
- × Is the Bible a lie?
- × Is the Moon hollow?
- × Does anyone take the Church of the SubGenius, a satire, seriously?

NCWP top-5 (rank of gold: #3)

- × Is the Bible a lie?
- × Is Scott Gordon's "The God Entity: The Theory of Everything" a valid premise?
- ✓ *What are the evidences to support the ancient alien theory?* ★ gold
- × How historically accurate is the evidence shown in "Ancient Aliens" from History Channel?
- × Is the Big Bang actually possible, or not?

Table 24: **Raw cosine is not aligned with semantic similarity**; NCWP corrects this. Three STSBenchmark sentence pairs with low gold similarity (≤ 0.6). Raw Qwen1.5-4B cosine is uniformly above 0.86; NCWP cosine matches the gold judgment.

Sentence pair	Gold	Raw	NCWP
"A red haired women in black is posing for a man while resting on a tree."	0.0	0.918	-0.005
"A man in a black shirt is holding a child wearing sunglasses."			
completely unrelated scenes			
"A man is cutting paper with a sword."	0.6	0.894	-0.138
"A woman is cutting a tomato."			
one-word match ("cutting"), different actions			
"Two men standing in grass staring at a car."	0.4	0.865	-0.083
"Two women sitting in chairs in a shop."			
syntactic pattern match, different content			