NEIGHBOR CO-OCCURRENCE ALONE SHOULD NOT DEFINE SUPERVISION IN EMBEDDING DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Relational knowledge distillation for text embeddings is typically performed within mini-batches, making batch composition determine which teacher-defined relationships the student can observe. Random batches may miss a seed text's close teacher neighbors. Grouping these neighbors makes the relevant comparisons available, but grouping alone does not tell the student which neighbor the teacher considers closer to the seed text, or by how much. We introduce **Neighborhood Similarity Transfer Knowledge Distillation (NeST-KD)**, which turns each mini-batch into a shared teacher-neighborhood subgraph. NeST-KD samples seed texts, retrieves their nearest neighbors in the teacher embedding space, and then matches teacher and student conditional similarity distributions over the resulting subgraph. This transfers graded differences in similarity while allowing other eligible texts to provide additional supervision from the same encoded examples. An auxiliary objective also matches teacher and student similarity distributions from each seed text to the other examples in the shared subgraph. Extensive experiments on nine standard sentence-embedding datasets show that NeST-KD improves over the strongest evaluated baseline by 1.18 and 2.64 absolute points for the 66M and 22M MiniLM students, respectively. Controlled studies show that correctly assigning teacher similarities, beyond merely selecting neighbors, improves downstream performance and relational fidelity.

1 INTRODUCTION

Text embedding models map texts into a shared vector space, making semantic relationships available to retrieval, clustering, classification, and semantic textual similarity systems (Reimers & Gurevych, 2019; Muennighoff et al., 2023). Larger encoders improve representation quality, but their latency and memory costs limit deployment at high query volumes or under constrained compute (Zhang et al., 2025). Knowledge distillation addresses this gap by training a compact student to match individual teacher embeddings (Vujanic & Rückstieß, 2026) or relationships between examples (Zhang et al., 2024a; Truong et al., 2025). Relational supervision is attractive for embeddings, yet evaluating all $O(N^2)$ corpus pairs is impractical. Existing methods therefore compare examples within mini-batches, making batch composition part of the supervision because it determines which relationships the student can observe.

Random mini-batches rarely place multiple close teacher neighbors around the same text, limiting the student's opportunity to learn fine-grained similarity differences. Adding teacher-selected neighbors improves this exposure (Xu et al., 2022), but retrieval alone does not determine how the resulting structure should be used. A text added as a neighbor of one sampled seed may itself have several teacher neighbors in the expanded batch; treating it only as a candidate discards relationships that have already been made available by the same student encodings. This leads to our central question: *once semantic neighbors are available, how should their relationships be turned into effective supervision?* Answering it requires jointly deciding which candidate relations to preserve, how much teacher probability to assign to each relation, and which texts in the expanded batch should act as supervised queries.

We introduce **Neighborhood Similarity Transfer Knowledge Distillation (NeST-KD)** to answer this question. Rather than treating retrieved neighbors as extra batch members, NeST-KD treats the expanded batch as a shared pool of teacher neighborhoods, in which every text with enough

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

observed neighbors acts as a supervised query and receives the teacher's graded probabilities over them. Our contributions are threefold:

- **Co-occurrence is not supervision.** We show that random batches expose fewer than two teacher neighbors for about 92% of anchors, and that a sparse target's distortion splits into a candidate-selection term and a probability-assignment term. For a fixed budget, teacher top-k neighbors with their conditional probabilities minimize both, and matching them recovers the teacher's local similarity differences.
- **A novel solution for neighborhood supervision.** NeST-KD merges the teacher neighborhoods of sampled seeds into a shared pool and encodes each distinct text once. Every pooled text with at least two observed neighbors matches the teacher's graded conditional probabilities, and an auxiliary objective compares each seed with the broader pool.
- **Empirical gains.** Across three teacher–student pairs and nine datasets, NeST-KD achieves the highest overall averages among the evaluated methods for both compact students. Same-pool uniform and shuffled controls test graded probability assignment within neighborhood rows, while anchor-only and dense-row controls probe how the encoded pool is used for supervision.

2 Related Work

Text embedding distillation. Sentence encoders can be compressed through representation alignment, contrastive objectives, and retrieval-oriented constraints (Gao et al., 2021a; Xu et al., 2023; Kim et al., 2023). SimTDE uses a compact token-embedding block and a shallow encoder, with losses on token and sentence embeddings (Xie et al., 2023). Other approaches transfer retrieval scores through self-distillation (Chen et al., 2024), align student layers to teacher representations (Dao et al., 2026), or combine relational alignment with transport across tokenizers (Truong et al., 2025). CoSS aligns both feature representations and pairwise geometry (Singh & Wang, 2024). NeST-KD focuses on how the encoded examples jointly determine candidate exposure, conditional targets, and the texts eligible to supply supervision.

Candidate selection. BINGO forms teacher-guided bags for intra- and inter-instance distillation (Xu et al., 2022); text-embedding methods also synthesize negatives at multiple difficulty levels (Pan et al., 2025). KDMCSE uses multimodal teacher knowledge to identify noisy negatives and adaptive angular margins to shape contrastive comparisons (Nguyen et al., 2024). Its multimodal setting differs from our distillation between text encoders. These approaches motivate examining both candidate availability and the supervision assigned to each relation. NeST-KD studies this distinction using uniform and shuffled targets on unchanged neighborhood supports.

Relational and distributional supervision. Relational KD transfers distances or similarities across embedding spaces (Tung & Mori, 2019; Park et al., 2019), while probabilistic and graph-based methods match affinity distributions and neighborhood structure (Passalis et al., 2021; Yang et al., 2020). HRKD models hierarchical and inter-domain relations through domain graphs and representative prototypes (Dong et al., 2021). Other objectives preserve topology (Kim et al., 2024), transport geometry-aware distributions (Lv et al., 2024), or transfer confidence-weighted rankings (Bassam et al., 2025). NeST-KD organizes directed teacher neighborhoods into a shared pool, matches conditional similarities for every eligible text, and adds anchor-to-pool comparisons. Its contribution concerns this combination of candidate construction, graded target assignment, and reuse of encoded texts as supervision sources.

3 Motivation

3.1 Mini-Batches as a Supervision Design

Let $\mathcal{D} = \{x_i\}_{i=1}^{N}$ be a deduplicated text corpus with index set $\mathcal{I} = \{1, \ldots, N\}$. Random batching can provide limited access to a text's teacher neighborhood. For an anchor x_j, let $\mathcal{N}(j)$ denote the indices of its top-k non-self neighbors in the teacher embedding space. Let $B \subseteq \mathcal{I}$ be a batch of

b indices sampled uniformly without replacement. Conditional on $j \in B$, the number of teacher neighbors present in the batch, $X_j = |\mathcal{N}(j) \cap B|$, satisfies

$$\mathbb{E}[X_j \mid j \in B] = \frac{(b-1)k}{N-1}. \tag{1}$$

Each of the anchor's k neighbors has probability $(b-1)/(N-1)$ of appearing among the remaining batch members. For the illustrative setting in Figure 1, with $N = 13{,}553$, $b = 64$, and $k = 100$, the expected number of neighbors is only 0.465, and the probability of observing at least two is approximately 7.9%.

Random batches rarely expose comparable neighbors. Random batches rarely expose two members of an anchor's teacher neighborhood in the same row. The 7.9% figure in Figure 1 uses $b = 64$ as an illustrative setting; at the training batch size $b = 128$, the corresponding probability is approximately 24.1%. Two observed teacher neighbors are needed for a nontrivial *conditional neighborhood row*: a softmax over a single candidate has zero KL and zero gradient. Other relational objectives may still compare a teacher neighbor with a non-neighbor. Adding teacher-selected neighbors exposes local comparisons, including for some pooled texts that were not initially sampled as anchors. We next examine the supervision assigned to those comparisons. Appendix A.2 gives the exact distribution.

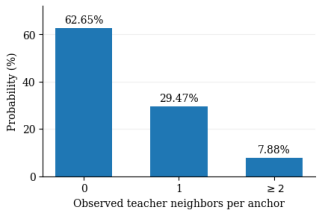


Figure 1: Neighbor exposure under random batching for $N = 13{,}553$, $b = 64$, and $k = 100$; the exact distribution is derived in Appendix A.2.

3.2 From Candidate Exposure to Conditional Supervision

For an anchor x_j, let $\mathcal{U}_j = \mathcal{I} \setminus \{j\}$ and let s^T_{ju} and s^S_{ju} be the teacher and student cosine similarities to candidate u. Consider a candidate set $C_j \subseteq \mathcal{U}_j$ containing at least two texts. For a common positive temperature τ_j, define the conditional teacher and student distributions

$$P^T_j(u \mid C_j) = \frac{\exp(s^T_{ju}/\tau_j)}{\sum_{v \in C_j} \exp(s^T_{jv}/\tau_j)}, \qquad P^S_j(u \mid C_j) = \frac{\exp(s^S_{ju}/\tau_j)}{\sum_{v \in C_j} \exp(s^S_{jv}/\tau_j)}, \quad u \in C_j. \tag{2}$$

These are the distributions matched by the neighborhood objective in Section 4.2, with C_j instantiated as the teacher neighbors observed in the shared pool. The choice of C_j determines which comparisons are available; the target probabilities determine what the student learns from them. For any target r_j on C_j, the row loss $\ell_j(r_j) = D_{\mathrm{KL}}\big(r_j \| P^S_j(\cdot \mid C_j)\big)$ satisfies

$$\frac{\partial \ell_j(r_j)}{\partial s^S_{ju}} = \frac{P^S_j(u \mid C_j) - r_j(u)}{\tau_j}, \qquad \frac{\partial\big[\ell_j(r_j) - \ell_j(P^T_j(\cdot \mid C_j))\big]}{\partial s^S_{ju}} = \frac{P^T_j(u \mid C_j) - r_j(u)}{\tau_j}. \tag{3}$$

Thus, even with identical candidates, target assignment matters. Uniform targets discard teacher similarity differences, while shuffled targets preserve their magnitudes but assign them to the wrong neighbors. Both distort the relational information transferred to the student.

What conditional KL preserves. Let $\delta_{ju} = s^S_{ju} - s^T_{ju}$ denote the student–teacher similarity error. Substituting into the KL yields

$$D_{\mathrm{KL}}\big(P^T_j(\cdot \mid C_j) \| P^S_j(\cdot \mid C_j)\big) = \log \mathbb{E}_{u \sim P^T_j(\cdot|C_j)} \exp(\delta_{ju}/\tau_j) - \frac{\mathbb{E}_{u \sim P^T_j(\cdot|C_j)} \delta_{ju}}{\tau_j}. \tag{4}$$

This identity shows that the loss is unchanged when the same error is added to every candidate. Conditional KL therefore penalizes differences in similarity errors across candidates, rather than a common shift in their similarities. In particular,

$$D_{\mathrm{KL}}\big(P^T_j(\cdot \mid C_j) \| P^S_j(\cdot \mid C_j)\big) = 0 \iff s^S_{ju} - s^S_{jv} = s^T_{ju} - s^T_{jv}, \quad \forall u, v \in C_j. \tag{5}$$

Thus, zero KL requires exact preservation of all pairwise similarity gaps within a row, without requiring absolute similarities to match. Appendix A.4 further expresses the KL as a nonnegative weighted integral of squared gap errors.

Why use teacher top-k candidates? There is a separate reason to choose the local support. Fix a temperature $\tau_j > 0$ and define a reference teacher distribution on all non-self texts:

$$\pi_j(u) = \frac{\exp(s^T_{ju}/\tau_j)}{\sum_{v \in \mathcal{U}_j} \exp(s^T_{jv}/\tau_j)}, \qquad u \in \mathcal{U}_j. \tag{6}$$

For a k-element support C_j and a distribution r_j supported there, the sparse-target distortion has the following decomposition.

Proposition 3.1 (Sparse teacher-target construction)**.** *For finite teacher similarities, a fixed $\tau_j > 0$, and $1 \le k \le N-1$, every $C_j \subseteq \mathcal{U}_j$ with $|C_j| = k$ and every distribution r_j supported on C_j satisfy*

$$D_{\mathrm{KL}}(r_j \| \pi_j) = -\log \pi_j(C_j) + D_{\mathrm{KL}}\big(r_j \| \pi_j(\cdot \mid C_j)\big). \tag{7}$$

The joint minimum is attained by selecting the teacher's top-k similarities and using $r_j = \pi_j(\cdot \mid C_j)$; ties at the selection boundary may yield multiple optimal supports.

The first term is minimized by selecting the teacher's top-k similarities, while the second is minimized, for any fixed support, by assigning the corresponding conditional teacher probabilities. This establishes optimal sparse-target construction under the stated KL criterion, without implying optimal student training or downstream performance. Proof and tie conditions are given in Appendix A.3. For non-anchor texts, conditional targets remain optimal on their observed supports, even when these contain only a subset of their top-k neighbors.

How local and broader rows allocate supervision. Partition a broader candidate set D_j into C_j and its nonempty complement O_j. For teacher and student distributions $P^T_j(\cdot \mid D_j), P^S_j(\cdot \mid D_j)$ computed at a common temperature, let $a = P^T_j(C_j \mid D_j)$ and $c = P^S_j(C_j \mid D_j)$. Then

$$\begin{aligned} D_{\mathrm{KL}}\big(P^T_j(\cdot \mid D_j) \| P^S_j(\cdot \mid D_j)\big) = \; & \mathrm{kl}(a\|c) + aD_{\mathrm{KL}}\big(P^T_j(\cdot \mid C_j) \| P^S_j(\cdot \mid C_j)\big) \\ & + (1-a)D_{\mathrm{KL}}\big(P^T_j(\cdot \mid O_j) \| P^S_j(\cdot \mid O_j)\big), \end{aligned} \tag{8}$$

where kl denotes binary KL. The first term measures whether the student assigns the correct total probability to the neighborhood. The remaining terms measure how well it matches teacher similarities within and outside that neighborhood. Crucially, dense-row KL weights the neighborhood term by its teacher probability mass a, which can weaken local supervision when a is small. NeST-KD instead supervises neighborhood relations separately, uses an auxiliary loss for broader comparisons, and includes eligible non-anchor texts as additional supervised queries. These differences motivate the controls in Section 5.3.

4 Proposed Method: NeST-KD

NeST-KD organizes each training step around a shared pool of teacher-selected neighborhoods. It first samples anchors and combines them with their teacher neighbors, encoding each distinct text once. It then matches neighborhood similarity distributions for every eligible text, including non-anchor texts, and adds broader anchor-to-pool supervision. Figure 2 summarizes this procedure.

4.1 Teacher-Guided Candidate Construction

Let f_T be the fixed teacher encoder and f_θ the student encoder with trainable parameters θ. For each text x_i in the corpus $\mathcal{D}$, define the normalized embeddings

$$z^T_i = \frac{f_T(x_i)}{\|f_T(x_i)\|_2}, \qquad z^S_i(\theta) = \frac{f_\theta(x_i)}{\|f_\theta(x_i)\|_2}. \tag{9}$$

We assume nonzero encoder outputs. Teacher and student embeddings may have different dimensions, since cosine similarities are computed within each model's own embedding space:

$$s^T_{ju} = (z^T_j)^\top z^T_u, \qquad s^S_{ju} = (z^S_j(\theta))^\top z^S_u(\theta). \tag{10}$$

Teacher embeddings are computed once and cached. All student similarities and distributions depend on θ; we omit this dependence below for readability. We use an anchor batch size $1 \le b \le N$ and a neighborhood size $2 \le k \le N-1$.

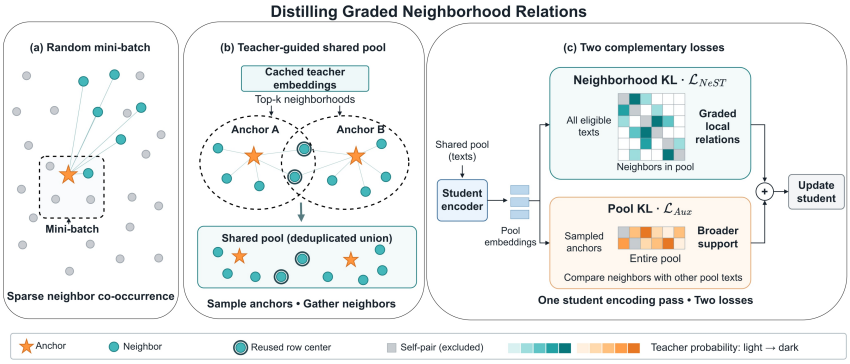


Figure 2: Overview of NeST-KD. (a) Random mini-batches may contain few teacher neighbors of each sampled text. (b) NeST-KD combines sampled anchors and their teacher-selected neighbors into a shared pool, encoding each distinct text once. (c) Neighborhood supervision matches teacher and student similarity distributions for all eligible texts. Auxiliary pool supervision compares each sampled anchor with all other texts in the pool.

Teacher neighborhoods. For each text x_j, retrieve the indices of its k nearest neighbors in the teacher embedding space, excluding j itself. With $\mathcal{U}_j = \mathcal{I} \setminus \{j\}$, this gives

$$\mathcal{N}(j) \in \arg\max_{\substack{C \subseteq \mathcal{U}_j \\ |C|=k}} \sum_{u \in C} s^T_{ju}. \tag{11}$$

This top-k selection attains the candidate-selection optimum in Proposition 3.1 for a fixed positive temperature. The resulting neighborhood graph is directed: $u \in \mathcal{N}(j)$ does not imply $j \in \mathcal{N}(u)$.

Shared candidate pool. At each training step, sample b anchor indices uniformly without replacement:

$$\mathcal{B}_b = \{B \subseteq \mathcal{I} : |B| = b\}, \qquad B \sim \mathrm{Unif}(\mathcal{B}_b). \tag{12}$$

We combine these anchors and their teacher neighbors into a shared pool. The pool and its induced neighborhood edges are

$$\begin{aligned} \mathcal{P}_B &= B \cup \bigcup_{i \in B} \mathcal{N}(i), \\ \mathcal{E}_B &= \{(j, u) \in \mathcal{P}_B \times \mathcal{P}_B : u \in \mathcal{N}(j)\}. \end{aligned} \tag{13}$$

The set union removes repeated indices, so the student encodes each distinct pooled text once, even when it belongs to multiple anchors' neighborhoods. For every text indexed by $j \in \mathcal{P}_B$, we identify which of its own teacher neighbors are present:

$$C_j(B) = \mathcal{N}(j) \cap \mathcal{P}_B, \qquad R_B = \{j \in \mathcal{P}_B : |C_j(B)| \geq 2\}. \tag{14}$$

Here, $C_j(B)$ is the observed candidate set, and R_B contains the indices of texts eligible for neighborhood supervision. At least two candidates are required because a singleton distribution cannot express a relative comparison. Every sampled anchor retains its full neighborhood: $C_i(B) = \mathcal{N}(i)$ for $i \in B$. A non-anchor text uses only its teacher neighbors already present in the pool; we do not expand the pool recursively. A text can therefore serve both as another text's candidate and as a source of its own neighborhood supervision. Since $k \geq 2$, all sampled anchors belong to R_B, ensuring that R_B is nonempty. We use $b = 128$ and $k = 100$ by default. The pool size varies with the overlap among the selected neighborhoods.

4.2 Similarity Supervision on the Shared Pool

The two objectives use different supervision sources and candidate sets. The neighborhood objective supervises every eligible text using its teacher neighbors present in the pool. The auxiliary objective supervises only sampled anchors, but compares each anchor with all other pooled texts. Within each objective, teacher and student distributions are normalized over the same candidate set.

Neighborhood similarity transfer. For each $j \in R_B$, we construct teacher and student similarity distributions over $C_j(B)$. Let $s^T_{j,(r)}$ denote the r-th largest teacher similarity between x_j and the texts indexed by $\mathcal{U}_j$. We define the positive temperature

$$\tau_j = \max\left\{\frac{s^T_{j,(1)} - s^T_{j,(k)}}{\log k}, 10^{-6}\right\}. \tag{15}$$

The lower bound of 10^{-6} follows the implementation and ensures a finite positive temperature when the top-k similarities are tied or nearly tied. If all top-k scores are equal, the teacher target is uniform. The stated $1/k$ ratio between the first and k-th unnormalized weights holds when the floor is inactive. The temperature is computed from the full teacher neighborhood and remains fixed throughout training, even when only a subset of that neighborhood is present in the pool. Teacher and student use the same τ_j:

$$P^\alpha_j(u \mid C_j(B)) = \frac{\exp(s^\alpha_{ju}/\tau_j)}{\sum_{v \in C_j(B)} \exp(s^\alpha_{jv}/\tau_j)}, \qquad u \in C_j(B), \quad \alpha \in \{T, S\}. \tag{16}$$

Here, P^T_j is the teacher target and P^S_j is the student distribution. The teacher target equals the conditional reference distribution $\pi_j(\cdot \mid C_j(B))$ from Section 3. By Proposition 3.1, it minimizes target distortion on this fixed support. For non-anchor texts, this statement concerns target assignment; their observed support need not be an optimal candidate set. The neighborhood loss averages teacher-to-student KL divergence over all eligible texts:

$$\mathcal{L}_{\mathrm{NeST}}(\theta; B) = \frac{1}{|R_B|} \sum_{j \in R_B} D_{\mathrm{KL}}\big(P^T_j(\cdot \mid C_j(B)) \,\big\|\, P^S_j(\cdot \mid C_j(B))\big). \tag{17}$$

Each eligible text contributes one equally weighted KL term, regardless of its number of observed candidates. This extends supervision to non-anchor texts using embeddings already computed for the shared pool. At zero divergence, teacher and student similarity differences agree for every pair of candidates in $C_j(B)$, as established in Equation (5).

Auxiliary pool supervision. The neighborhood loss only compares candidates within each observed neighborhood. For a sampled anchor, it does not directly compare the similarities of its neighbors with those of texts outside that neighborhood. We therefore add an objective over all other texts in the pool. For each anchor $i \in B$, define the broader candidate set and a single corpus-level temperature:

$$C^{\mathrm{pool}}_i(B) = \mathcal{P}_B \setminus \{i\}, \qquad \bar{\tau} = \mathrm{median}_{j \in \mathcal{I}}\, \tau_j. \tag{18}$$

The corresponding teacher and student distributions are

$$Q^\alpha_i(u \mid B) = \frac{\exp(s^\alpha_{iu}/\bar{\tau})}{\sum_{v \in C^{\mathrm{pool}}_i(B)} \exp(s^\alpha_{iv}/\bar{\tau})}, \qquad u \in C^{\mathrm{pool}}_i(B), \quad \alpha \in \{T, S\}. \tag{19}$$

Their divergence defines the auxiliary loss:

$$\mathcal{L}_{\mathrm{aux}}(\theta; B) = \frac{1}{|B|} \sum_{i \in B} D_{\mathrm{KL}}\big(Q^T_i(\cdot \mid B) \,\big\|\, Q^S_i(\cdot \mid B)\big). \tag{20}$$

This objective transfers each anchor's relative similarities to candidates from across the shared pool, including texts introduced by other anchors. For sampled anchors, the two objectives overlap on neighborhood comparisons. However, the neighborhood loss normalizes probabilities within each observed neighborhood, giving these local comparisons a dedicated training signal. It also supervises eligible non-anchor texts, which are not used as supervision sources in the pool loss.

Overall optimization objective. We train the student by minimizing

$$\min_\theta \; \mathbb{E}_{B \sim \mathrm{Unif}(\mathcal{B}_b)} \left[\mathcal{L}_{\mathrm{NeST}}(\theta; B) + \lambda_{\mathrm{aux}} \mathcal{L}_{\mathrm{aux}}(\theta; B)\right]. \tag{21}$$

The neighborhood loss has unit weight, and $\lambda_{\mathrm{aux}} > 0$ controls the auxiliary contribution, with a default value of 0.5. Teacher embeddings, neighbor sets, and temperatures are precomputed and fixed throughout training. Each sampled anchor batch determines the shared pool, observed candidate sets, and corresponding teacher target distributions. These targets do not depend on θ, and gradients flow only through the student embeddings. After training, inference requires only the student encoder.

5 EXPERIMENTS

5.1 SETUP

Baselines. We compare NeST-KD against SimCSE-unsup (Gao et al., 2021b), RKD (Park et al., 2019), CDM (Chen et al., 2025), DSKD (Zhang et al., 2024b), Jasper and Stella (Zhang et al., 2024a), DistillCSE (Xu et al., 2023), EMO (Truong et al., 2025), and TALAS (Dao et al., 2026). We use task-level TALAS results for headline comparisons and include frozen-teacher and pretrained-student references.

Training Data. We distill on 13,553 deduplicated, unlabeled training texts from Emotion, WiC, and STS-B, excluding labels and evaluation examples. We treat these three datasets as in-domain and the remaining six evaluation datasets as out-of-domain. Here, "out-of-domain" means that the corresponding benchmark training splits do not contribute texts to the main distillation corpus. Task-independent Wikipedia distillation is reported in Appendix B.2.

Evaluation. We evaluate frozen embeddings using macro-F1 (Banking77, Tweet, Emotion), cosine-similarity average precision (MRPC, SciTail, WiC), and Spearman correlation (SICK, STS12, STS-B). Scores are scaled by 100; their unweighted mean is reported as Avg. Results are mean $\pm$ sample standard deviation over seeds 42, 43, and 44 unless stated otherwise.

Implementation Details. We evaluate Qwen3-Embedding-4B→BERT-base (109M), BGE-M3→MiniLMv2-H768 (66M), and Qwen3-Embedding-0.6B→MiniLMv2-H384 (22M) (Zhang et al., 2025; Chen et al., 2024). Each run uses one NVIDIA H200 GPU, five epochs, batch size 128, and learning rate 3×10^{-5}. We set $k = 100$ and $\lambda_{\text{aux}} = 0.5$, precomputing teacher embeddings, neighborhoods, and temperatures. Further configurations and ablation definitions are in Appendix C.

5.2 MAIN RESULTS

Table 1: Qwen3-Embedding-0.6B → MiniLMv2-H384 (22M); ⋆ marks in-domain tasks. **Bold** denotes the best distilled result, and underlined denotes the second best.

	Classification (F1)			Pair Classification (AP)		
Method	Emotion⋆	Banking77	Tweet	WiC⋆	MRPC	SciTail
Teacher	66.53	93.21	71.33	66.85	83.37	90.00
Student base	69.90	68.29	41.07	59.58	78.39	64.56
SimCSE-unsup	52.77 ±0.31	84.52 ±0.12	68.55 ±0.18	64.16 ±0.15	80.50 ±0.09	72.98 ±0.17
RKD	55.48 ±0.35	87.39 ±0.07	70.88 ±0.21	66.76 ±0.22	79.64 ±0.03	77.85 ±0.20
CDM	52.92 ±0.95	87.21 ±0.21	70.36 ±0.20	69.10 ±0.11	79.36 ±0.14	71.07 ±0.30
DSKD	54.71 ±0.29	85.74 ±0.10	71.41 ±0.20	66.86 ±0.04	80.26 ±0.01	70.96 ±0.09
Jasper and Stella	50.99 ±0.28	85.52 ±0.04	70.29 ±0.44	**69.56** ±0.16	82.66 ±0.12	70.52 ±0.32
DistillCSE	53.04 ±0.24	86.21 ±0.15	70.80 ±0.19	66.83 ±0.12	81.23 ±0.08	76.53 ±0.21
EMO	55.90 ±0.27	82.26 ±0.31	71.27 ±0.06	68.86 ±0.09	82.32 ±0.13	71.07 ±0.16
TALAS	61.13 ±0.52	84.16 ±0.15	71.52 ±0.11	66.32 ±0.07	**84.85** ±0.10	80.91 ±0.27
NeST-KD	**64.34** ±0.40	**88.45** ±0.07	**71.58** ±0.24	65.78 ±0.06	83.91 ±0.12	**82.76** ±0.12

	STS (Spearman)			Aggregate		
Method	STS-B⋆	SICK	STS12	Avg-In	Avg-Out	Avg.
Teacher	84.60	80.64	77.11	72.66	82.61	79.29
Student base	22.08	47.60	21.95	50.52	53.64	52.60
SimCSE-unsup	63.01 ±0.22	67.69 ±0.17	60.76 ±0.16	59.98 ±0.18	72.50 ±0.10	68.33 ±0.12
RKD	69.33 ±0.21	71.00 ±0.18	65.67 ±0.19	63.86 ±0.20	75.41 ±0.09	71.56 ±0.10
CDM	65.02 ±0.24	65.60 ±0.36	61.17 ±0.22	62.35 ±0.43	72.46 ±0.19	69.09 ±0.27
DSKD	62.25 ±0.18	65.53 ±0.06	58.62 ±0.19	61.27 ±0.03	72.09 ±0.06	68.48 ±0.04
Jasper and Stella	64.94 ±0.20	66.42 ±0.19	59.69 ±0.12	61.83 ±0.13	72.52 ±0.13	68.95 ±0.10
DistillCSE	72.43 ±0.25	68.06 ±0.14	69.23 ±0.20	64.10 ±0.11	75.34 ±0.08	71.60 ±0.15
EMO	62.39 ±0.23	61.42 ±0.16	56.40 ±0.17	62.38 ±0.09	70.79 ±0.07	67.99 ±0.02
TALAS	72.39 ±0.31	70.72 ±0.15	67.80 ±0.08	66.61 ±0.24	76.66 ±0.07	73.31 ±0.10
NeST-KD	**78.67** ±0.04	**74.09** ±0.06	**73.95** ±0.12	**69.60** ±0.12	**79.12** ±0.08	**75.95** ±0.09

NeST-KD advances the compact-model frontier. Table 1 shows an overall average of 75.95 for Qwen3-Embedding-0.6B→MiniLMv2-H384, exceeding TALAS by 2.64 points and the strongest

other baseline (DistillCSE) by 4.35 points. The BGE-M3→MiniLMv2-H768 result in Appendix B.1.1 reaches 77.83, improving over TALAS by 1.18 points. Together, the two compact pairs indicate that the gain persists across different teachers and student capacities. Section 5.3 shows that this lead is not only greater exposure.

The lead survives the domain shift. NeST-KD improves the best out-of-domain average by 2.46 points for Qwen3-Embedding-0.6B→MiniLMv2-H384 and by 1.22 points for BGE-M3→MiniLMv2-H768. Relative to the best baseline in each column, SICK improves by 3.09 and 2.68 points, while STS12 improves by 4.72 and 6.04 points, respectively.

Teacher quality sets the student frontier. Stronger teachers are paired with stronger NeST-KD students: across Qwen3-Embedding-4B→BERT-base, BGE-M3→MiniLMv2-H768, and Qwen3-Embedding-0.6B→MiniLMv2-H384 (Tables 4, 3, and 1), teacher averages decrease from 81.22 to 79.76 to 79.29, while the corresponding students follow the same order at 78.66, 77.83, and 75.95. NeST-KD remains 1.93–3.34 points below its teacher, suggesting that teacher quality is largely transferred but not systematically exceeded.

5.3 Ablation Studies

Unless noted otherwise, controlled studies use the Qwen3-Embedding-0.6B→MiniLMv2-H384 pair. We test three possible sources of NeST-KD's gains: co-encoding teacher neighbors, assigning the correct graded teacher probability to each relation, and supervising non-anchor texts already present in the shared pool.

Table 2: Ablations of supervision design for Qwen3-Embedding-0.6B→MiniLMv2-H384. Appendix C.1.1 specifies the encoded pool, relational rows, targets, and loss for every control. Scores are mean ± sample standard deviation over three seeds; Δ is the change in Avg. relative to NeST-KD.

Method	Avg-In	Avg-Out	Avg.	Δ
NeST-KD	**69.60** ±0.12	**79.12** ±0.08	**75.95** ±0.09	–
Exposure and row coverage				
Pointwise, random	61.58 ±0.18	71.59 ±0.12	68.25 ±0.34	−7.70
Pointwise, neighbor	58.91 ±0.21	68.80 ±0.15	65.50 ±0.25	−10.45
In-batch, random	63.93 ±0.16	75.79 ±0.10	71.84 ±0.08	−4.11
In-batch, neighbor	67.11 ±0.14	78.21 ±0.09	74.51 ±0.04	−1.44
Anchor rows only	68.52 ±0.13	78.86 ±0.07	75.41 ±0.06	−0.54
Targets on the same support				
Uniform targets	67.13 ±0.15	78.52 ±0.09	74.72 ±0.07	−1.23
Shuffled targets	67.11 ±0.17	78.49 ±0.10	74.69 ±0.11	−1.26
Pool and density				
Random columns	69.12 ±0.14	78.77 ±0.08	75.55 ±0.06	−0.40
Random pool	69.09 ±0.13	78.91 ±0.07	75.64 ±0.05	−0.31
Dense rows	69.54 ±0.16	78.81 ±0.09	75.72 ±0.04	−0.23
Dense rows only	69.22 ±0.16	78.42 ±0.09	75.35 ±0.04	−0.60

Relational supervision turns neighbor exposure into useful signal. Table 2 shows that neighbor batches lower the pointwise score by 2.75 points but raise batch-local relational distillation by 2.67 points to 74.51, above TALAS at 73.31 with only 128 encoded texts per update; supervising all eligible rows adds another 0.54 points over anchor-only supervision. Together, the gains come from transferring teacher relations across more pooled queries.

Correct assignments matter more than denser support. Uniform and shuffled targets lose 1.23 and 1.26 points on the same candidates, although shuffling preserves the target histogram and entropy. Dense rows score 0.23 points below NeST-KD despite using approximately 67× more candidates, and removing the auxiliary loss lowers the dense-row score by another 0.37 points to 75.35. These controls encode the same pool as NeST-KD.

Candidate selection and target assignment create complementary errors. In Figure 3(a), replacing more top-k neighbors increasingly raises the selection term as teacher probability mass is discarded. Panel (b) isolates assignment: its KL remains zero without assignment noise, even when candidates change, but rises with noise at every replacement rate. The strongest joint perturbation produces the largest error, illustrating the two terms in Equation (7).

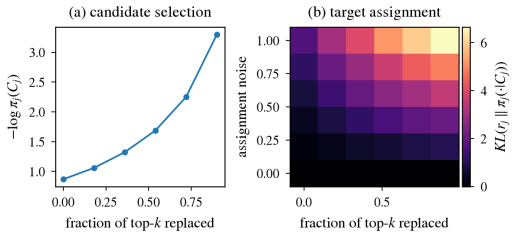


Figure 3: Selection and assignment perturbations increase distinct terms of target distortion.

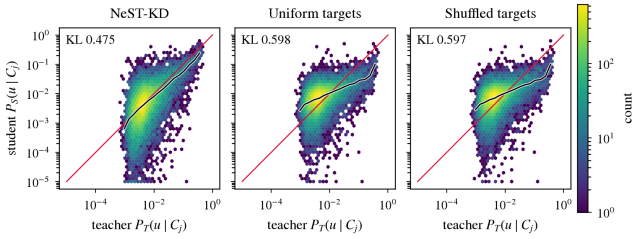


Figure 4: Conditional-probability fidelity on common teacher supports.

Graded targets improve probability fidelity. We evaluate conditional KL on 256 sampled rows using identical teacher top-100 supports across students. Figure 4 shows that NeST-KD more closely follows the agreement diagonal, while students trained with uniform or shuffled targets exhibit flatter probability trends, obscuring differences among neighbors. Despite preserving the teacher probability values, shuffling assigns them to the wrong neighbors and performs similarly to uniform targets. NeST-KD achieves a conditional KL of 0.475, compared with 0.598 and 0.597 for the two controls.

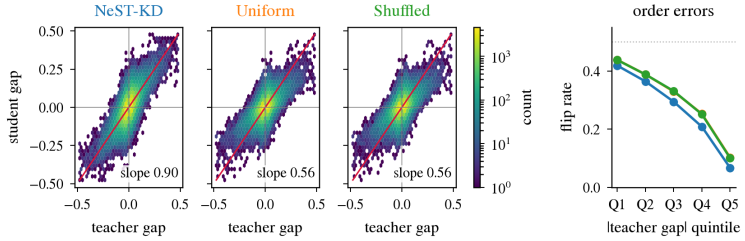


Figure 5: Teacher–student similarity gaps and pairwise order errors.

Probability fidelity carries over to local geometry. Using the same rows and supports, we sample 200,000 within-row pairs per model for the slope and order-error estimates, excluding duplicate candidates and teacher near-ties. In Figure 5, NeST-KD tracks teacher similarity gaps with slope 0.90; both controls compress positive and negative gaps toward zero with slope 0.56. Pairwise order errors decrease as teacher gaps grow across all quintiles, while NeST-KD stays below shuffled targets, especially in the middle quintiles. The appendix rank probe in Figure 8 agrees: teacher–student Spearman correlation rises from 0.43 to 0.50.

6 Conclusion

NeST-KD organizes teacher neighborhoods into a shared pool and transfers graded conditional similarities from every eligible text. Across the evaluated teacher–student pairs and training corpora, it achieves strong performance among the compared distillation methods. Ablation studies show that correctly assigning teacher probabilities within the neighborhood objective improves performance when the encoded pool and graded auxiliary supervision are held fixed. Probability, similarity-gap, and ranking analyses further indicate better preservation of local teacher relations.

AI USE STATEMENT

Generative AI tools were used to improve the clarity and readability of the manuscript, identify and summarize related literature to assess potential overlap, suggest experimental hyperparameters, and correct minor coding errors. They were not used to develop theoretical models or mathematical claims, write proofs, design the methodology or experiments, implement the proposed method, translate materials, process datasets, conduct qualitative analyses, or interpret experimental results. All AI-assisted material was reviewed by the authors, who take full responsibility for the theoretical claims, experimental results, code, and written content of this work.

REPRODUCIBILITY STATEMENT

Section 4 specifies the pool construction, candidate supports, and training objectives. Appendix A gives the assumptions and proofs, while Appendix C records the training configuration, pooling choices, validation protocol, and evaluation procedure. We report the seeds and reference runs used for controlled comparisons. Upon acceptance, we will release the training and evaluation code, configuration files, and instructions needed to reproduce all reported experiments.

ETHICS STATEMENT

This work studies knowledge distillation for text embedding models using existing research datasets and does not collect new human-subject data. As with other distillation approaches, the student may inherit biases present in the teacher representations and training data. We therefore encourage further evaluation of fairness and robustness before deployment in sensitive applications.

REFERENCES

Ejafa Bassam, Dawei Zhu, and Kaigui Bian. PLD: A choice-theoretic list-wise knowledge distillation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 136090–136112, 2025. doi: 10.52202/085713-4542. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/c6b2921f24f82dc05cda53bbe50bab90-Abstract-Conference.html.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.137. URL https://aclanthology.org/2024.findings-acl.137/.

Yijie Chen, Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. Enhancing cross-tokenizer knowledge distillation with contextual dynamical mapping. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 8005–8018. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.419. URL https://aclanthology.org/2025.findings-acl.419/.

Quoc Phong Dao, Hoang Son Nguyen, Pham Khanh Chi, Linh Ngo Van, Nguyen Thi Ngoc Diep, Thien Huu Nguyen, and Trung Le. TALAS: Teacher-anchored layer alignment with adaptive sharpness-aware minimization for embedding distillation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32705–32723. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1509. URL https://aclanthology.org/2026.acl-long.1509/.

Chenhe Dong, Yaliang Li, Ying Shen, and Minghui Qiu. HRKD: Hierarchical relational knowledge distillation for cross-domain language model compression. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3126–3136. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.250. URL https://aclanthology.org/2021.emnlp-main.250/.

Chaochen Gao, Xing Wu, Peng Wang, Jue Wang, Liangjun Zang, Zhongyuan Wang, and Songlin Hu. DistilCSE: Effective knowledge distillation for contrastive sentence embeddings. *CoRR*, abs/2112.05638, 2021a. doi: 10.48550/arXiv.2112.05638. URL https://arxiv.org/abs/2112.05638.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6894–6910. Association for Computational Linguistics, 2021b. doi: 10.18653/v1/2021.emnlp-main.552. URL https://aclanthology.org/2021.emnlp-main.552/.

Jungeun Kim, Junwon You, Dongjin Lee, Ha Young Kim, and Jae-Hun Jung. Do topological characteristics help in knowledge distillation? In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 24674–24693. PMLR, 2024. URL https://proceedings.mlr.press/v235/kim24aj.html.

Seungyeon Kim, Ankit Singh Rawat, Manzil Zaheer, Sadeep Jayasumana, Veeranjaneyulu Sadhanala, Wittawat Jitkrittum, Aditya Krishna Menon, Rob Fergus, and Sanjiv Kumar. Embeddistill: A geometric knowledge distillation for information retrieval. *arXiv preprint arXiv:2301.12005*, 2023. doi: 10.48550/arXiv.2301.12005. URL https://arxiv.org/abs/2301.12005.

Jiaming Lv, Haoyuan Yang, and Peihua Li. Wasserstein distance rivals kullback–leibler divergence for knowledge distillation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 65445–65475, 2024. doi: 10.52202/079017-2090. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/78526d7ad4a2532bd91416e948b9644c-Abstract-Conference.html.

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*, pp. 2014–2037. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.eacl-main.148. URL https://aclanthology.org/2023.eacl-main.148/.

Cong-Duy Nguyen, Thong Nguyen, Xiaobao Wu, and Anh Tuan Luu. KDMCSE: Knowledge distillation multimodal sentence embeddings with adaptive angular margin contrastive learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 733–749. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.42. URL https://aclanthology.org/2024.naacl-long.42/.

Tengyu Pan, Zhichao Duan, Zhenyu Li, Bowen Dong, Ning Liu, Xiuxing Li, and Jianyong Wang. Negative matters: Multi-granularity hard-negative synthesis and anchor-token-aware pooling for enhanced text embeddings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31102–31118. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1501. URL https://aclanthology.org/2025.acl-long.1501/.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3967–3976, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Park_Relational_Knowledge_Distillation_CVPR_2019_paper.html.

Nikolaos Passalis, Maria Tzelepi, and Anastasios Tefas. Probabilistic knowledge transfer for lightweight deep representation learning. *IEEE Transactions on Neural Networks and Learning Systems*, 32(5):2030–2039, 2021. doi: 10.1109/TNNLS.2020.2995884. URL https://doi.org/10.1109/TNNLS.2020.2995884.

Nils Reimers and Iryna Gurevych. 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*, pp. 3982–3992. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Aditya Singh and Haohan Wang. Simple unsupervised knowledge distillation with space similarity. In *Computer Vision – ECCV 2024*, pp. 147–164. Springer, 2024. doi: 10.1007/978-3-031-72627-9_9. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/195_ECCV_2024_paper.php.

Minh-Phuc Truong, Hai An Vu, Tu Vu, Nguyen Thi Ngoc Diep, Linh Ngo Van, Thien Huu Nguyen, and Trung Le. EMO: Embedding model distillation via intra-model relation and optimal transport alignments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 7594–7606. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.385. URL https://aclanthology.org/2025.emnlp-main.385/.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1365–1374, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Tung_Similarity-Preserving_Knowledge_Distillation_ICCV_2019_paper.html.

Robin Vujanic and Thomas Rückstieß. LEAF: Knowledge distillation of text embedding models with teacher-aligned representations. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 43362–43383, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.2008. URL https://aclanthology.org/2026.acl-long.2008/.

Jian Xie, Xin He, Jiyang Wang, Zimeng Qiu, Ali Kebarighotbi, and Farhad Ghassemi. SimTDE: Simple transformer distillation for sentence embeddings. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2389–2393. Association for Computing Machinery, 2023. doi: 10.1145/3539618.3592063. URL https://www.amazon.science/publications/simtde-simple-transformer-distillation-for-sentence-embeddings.

Haohang Xu, Jiemin Fang, Xiaopeng Zhang, Lingxi Xie, Xinggang Wang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Bag of instances aggregation boosts self-supervised distillation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=FBl7QNXWQc_.

Jiahao Xu, Wei Shao, Lihui Chen, and Lemao Liu. DistillCSE: Distilled contrastive learning for sentence embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 8153–8165. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.547. URL https://aclanthology.org/2023.findings-emnlp.547/.

Yiding Yang, Jiayan Qiu, Mingli Song, Dacheng Tao, and Xinchao Wang. Distilling knowledge from graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7074–7083, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Yang_Distilling_Knowledge_From_Graph_Convolutional_Networks_CVPR_2020_paper.html.

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. Jasper and stella: Distillation of SOTA embedding models. *CoRR*, abs/2412.19048, 2024a. doi: 10.48550/arXiv.2412.19048. URL https://arxiv.org/abs/2412.19048.

Songming Zhang, Xue Zhang, Zengkui Sun, Yufeng Chen, and Jinan Xu. Dual-space knowledge distillation for large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18164–18181, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1010. URL https://aclanthology.org/2024.emnlp-main.1010/.

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. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *CoRR*, abs/2506.05176, 2025. doi: 10.48550/arXiv.2506.05176. URL https://arxiv.org/abs/2506.05176.

A Theoretical Details and Proofs

This section supports the supervision-design argument in Section 3: it derives neighbor exposure under random batching, proves the sparse-target decomposition, derives the similarity constraints induced by conditional distribution matching, and applies them to the shared pool of Section 4.

A.1 Notation and Assumptions

We retain the notation of the main text: $\mathcal{I} = \{1, \dots, N\}$ indexes the corpus, $\mathcal{U}_j = \mathcal{I} \setminus \{j\}$ contains the non-self candidates of x_j, s^T_{ju} and s^S_{ju} are the teacher and student cosine similarities of Equation (10), and $\mathcal{N}(j) \subseteq \mathcal{U}_j$ holds the k indices with the largest teacher similarities (ties resolved arbitrarily). For a fixed temperature $\tau_j > 0$, the full-corpus teacher normalizer, reference distribution, and retained mass of a candidate set $C_j \subseteq \mathcal{U}_j$ are

$$Z^T_j = \sum_{u \in \mathcal{U}_j} \exp(s^T_{ju}/\tau_j), \qquad \pi_j(u) = \frac{\exp(s^T_{ju}/\tau_j)}{Z^T_j}, \qquad \pi_j(C_j) = \sum_{u \in C_j} \pi_j(u). \tag{22}$$

All similarities are finite, so $\pi_j(u) > 0$ for every u. A target $r_j \in \Delta(C_j)$ is a probability distribution on a nonempty C_j, extended by zero outside it, with the convention $0 \log(0/q) = 0$. The temperature is held fixed whenever candidate sets or targets of the same anchor are compared; the method supplies it through Equation (15).

A.2 Neighbor Coverage in a Random Batch

Let B be a uniformly sampled subset of $\mathcal{I}$ of size b. Conditional on $j \in B$, the remaining members $B \setminus \{j\}$ form a uniform subset of $\mathcal{U}_j$ of size $b-1$, and the number of observed teacher neighbors is $X_j = |\mathcal{N}(j) \cap B|$. Choosing x indices among the k neighbors and $b-1-x$ among the other $N-1-k$ candidates gives the hypergeometric law, and linearity of expectation over the k indicators $\mathbf{1}\{u \in B\}$, each with conditional probability $(b-1)/(N-1)$, gives the mean:

$$\begin{aligned} \Pr(X_j = x \mid j \in B) &= \frac{\binom{k}{x}\binom{N-1-k}{b-1-x}}{\binom{N-1}{b-1}}, \qquad \max\{0,\, b-N+k\} \le x \le \min\{k,\, b-1\}, \\ \mathbb{E}[X_j \mid j \in B] &= \sum_{u \in \mathcal{N}(j)} \Pr(u \in B \mid j \in B) = \frac{(b-1)k}{N-1}. \end{aligned} \tag{23}$$

This proves Equation (1). Relational supervision needs at least two neighbors in the same batch, an event with probability

$$\Pr(X_j \ge 2 \mid j \in B) = 1 - \frac{\binom{N-1-k}{b-1} + k\binom{N-1-k}{b-2}}{\binom{N-1}{b-1}}. \tag{24}$$

For $N = 13{,}553$, $b = 64$, and $k = 100$, these expressions give $\mathbb{E}[X_j \mid j \in B] \approx 0.465$, $\Pr(X_j = 0 \mid j \in B) \approx 0.626$, $\Pr(X_j = 1 \mid j \in B) \approx 0.295$, and $\Pr(X_j \ge 2 \mid j \in B) \approx 0.079$, so about 92% of anchors see fewer than two of their teacher neighbors in a random batch.

A.3 Optimal Sparse Teacher Targets

Proof of Proposition 3.1. Write $\pi_j(u \mid C_j) = \pi_j(u)/\pi_j(C_j)$ for $u \in C_j$; both the retained mass and every conditional probability are strictly positive. Since r_j vanishes outside C_j,

$$\begin{aligned} D_{\mathrm{KL}}(r_j \| \pi_j) &= \sum_{u \in C_j} r_j(u) \log \frac{r_j(u)}{\pi_j(u)} \\ &= \sum_{u \in C_j} r_j(u) \log \frac{r_j(u)}{\pi_j(C_j)\, \pi_j(u \mid C_j)} \\ &= D_{\mathrm{KL}}(r_j \,\|\, \pi_j(\cdot \mid C_j)) - \log \pi_j(C_j) \sum_{u \in C_j} r_j(u) \\ &= -\log \pi_j(C_j) + D_{\mathrm{KL}}(r_j \,\|\, \pi_j(\cdot \mid C_j))\,, \end{aligned} \tag{25}$$

which is Equation (7). Since KL divergence is nonnegative and vanishes only at equality, the unique minimizer on a fixed support is $r_j = \pi_j(\cdot \mid C_j)$, and the remaining problem is

$$C_j^\star \in \arg\max_{\substack{C_j \subseteq \mathcal{U}_j \\ |C_j|=k}} \pi_j(C_j). \tag{26}$$

If a set contains v but excludes u with $\pi_j(u) > \pi_j(v)$, swapping v for u raises the retained mass by $\pi_j(u) - \pi_j(v) > 0$; repeating such exchanges reaches a set of k largest probabilities, and any such set maximizes the sum. Because $\exp(\cdot/\tau_j)$ is strictly increasing, these are exactly the top-k teacher neighbors, so $\mathcal{N}(j)$ is an optimal support and

$$\min_{\substack{C_j \subseteq \mathcal{U}_j,\ |C_j|=k \\ r_j \in \Delta(C_j)}} D_{\mathrm{KL}}(r_j \| \pi_j) = -\log \pi_j(\mathcal{N}(j)). \tag{27}$$

□

The optimal support does not depend on the temperature: since $\exp(\cdot/\tau_j)$ is strictly increasing for every $\tau_j > 0$, the ordering of π_j coincides with the ordering of teacher similarities, so $\mathcal{N}(j)$ is optimal for all $\tau_j > 0$ and only the optimal target $\pi_j(\cdot \mid C_j)$ depends on τ_j. In particular, defining τ_j from the top-k similarities in Equation (15) introduces no circularity. The decomposition separates the two design choices: replacing candidates changes only the selection term, while changing the target on an unchanged support changes only the assignment term. This is the separation probed in Figure 3.

Uniform and shuffled targets. For a fixed support of size $m = |C_j|$, the uniform target $r_j^{\mathrm{unif}}(u) = 1/m$ incurs assignment error

$$D_{\mathrm{KL}}\big(r_j^{\mathrm{unif}} \,\big\|\, \pi_j(\cdot \mid C_j)\big) = -\log m - \frac{1}{m} \sum_{u \in C_j} \log \pi_j(u \mid C_j), \tag{28}$$

which is zero exactly when the conditional teacher distribution is uniform. A shuffled target $r_j^{\mathrm{shuf}}(u) = \pi_j(\sigma_j(u) \mid C_j)$ for a permutation σ_j of C_j keeps the histogram and entropy of the teacher target but not its assignment:

$$H(r_j^{\mathrm{shuf}}) = H(\pi_j(\cdot \mid C_j)), \qquad D_{\mathrm{KL}}\big(r_j^{\mathrm{shuf}} \,\big\|\, \pi_j(\cdot \mid C_j)\big) = 0 \iff r_j^{\mathrm{shuf}} = \pi_j(\cdot \mid C_j). \tag{29}$$

For instance, exchanging two unequal conditional probabilities p and q contributes $(p - q)\log(p/q) > 0$ to the assignment error, whereas permutations among equal probabilities leave the target unchanged.

Remarks. The optimal size-k support is unique when the k-th largest teacher probability strictly exceeds the $(k+1)$-th; ties across this boundary allow several supports with the same retained mass, but the optimal conditional target on each support remains unique. The criterion uses the reverse direction $D_{\mathrm{KL}}(r_j \| \pi_j)$ because, for a proper subset $C_j \subsetneq \mathcal{U}_j$, the forward divergence $D_{\mathrm{KL}}(\pi_j \| r_j)$ is infinite. This target-design criterion is distinct from the teacher-to-student KL optimized during training, and it establishes neither optimal downstream quality nor optimal allocation of a shared encoding budget across anchors.

A.4 Conditional Targets and Similarity Differences

For a fixed candidate set C_j, define restricted normalizers and conditional distributions consistently with Equation (16):

$$Z_j^\alpha(C_j) = \sum_{v \in C_j} \exp(s_{jv}^\alpha/\tau_j), \qquad P_j^\alpha(u \mid C_j) = \frac{\exp(s_{ju}^\alpha/\tau_j)}{Z_j^\alpha(C_j)}, \qquad \alpha \in \{T, S\}. \tag{30}$$

The full-corpus normalizer cancels in the ratio, so $P_j^T(\cdot \mid C_j) = \pi_j(\cdot \mid C_j)$ and the target requires only the teacher similarities on C_j.

Proposition A.1 (Restricted-row similarity gaps)**.** *Let $|C_j| \geq 2$, let teacher and student similarities be finite, and use the same temperature $\tau_j > 0$ for both distributions. Then*

$$\begin{aligned} & D_{\mathrm{KL}}\big(P_j^T(\cdot \mid C_j) \,\big\|\, P_j^S(\cdot \mid C_j)\big) = 0 \\ \iff \quad & s_{ju}^S - s_{jv}^S = s_{ju}^T - s_{jv}^T \quad \textit{for all } u, v \in C_j \\ \iff \quad & \exists c_j \in \mathbb{R} :\ s_{ju}^S = s_{ju}^T + c_j \quad \textit{for all } u \in C_j. \end{aligned} \tag{31}$$

Proof. Both distributions are strictly positive, so zero KL is equivalent to their equality. For either $\alpha \in \{T, S\}$, the ratio of two candidate probabilities cancels the restricted normalizer:

$$\log \frac{P_j^\alpha(u \mid C_j)}{P_j^\alpha(v \mid C_j)} = \frac{s_{ju}^\alpha - s_{jv}^\alpha}{\tau_j}. \tag{32}$$

Equality of the two distributions therefore implies equality of all similarity gaps. Conversely, fix any $v \in C_j$; equal gaps give $s_{ju}^S - s_{ju}^T = s_{jv}^S - s_{jv}^T =: c_j$ for every $u \in C_j$, hence $Z_j^S(C_j) = \exp(c_j/\tau_j) Z_j^T(C_j)$. The same factor multiplies each numerator and cancels, giving $P_j^S = P_j^T$ on C_j. □

Away from exact matching, Equation (32) gives $(s_{ju}^S - s_{jv}^S) - (s_{ju}^T - s_{jv}^T) = \tau_j \big[\log\big(P_j^S(u \mid C_j)/P_j^T(u \mid C_j)\big) - \log\big(P_j^S(v \mid C_j)/P_j^T(v \mid C_j)\big)\big]$, which links errors in relative probabilities to errors in similarity gaps and motivates the separate probability and gap diagnostics in Figures 4 and 5. The proposition permits an additive offset c_j within each row, so it identifies neither absolute similarities nor similarities to texts outside C_j.

Why a row needs at least two candidates. Writing the KL of one row as ℓ_j, its student-dependent part is $-\sum_{v \in C_j} P_j^T(v \mid C_j)\, s_{jv}^S/\tau_j + \log Z_j^S(C_j)$, so

$$\frac{\partial \ell_j}{\partial s_{ju}^S} = \frac{P_j^S(u \mid C_j) - P_j^T(u \mid C_j)}{\tau_j}, \qquad u \in C_j. \tag{33}$$

For a singleton C_j both probabilities equal one, the KL is identically zero, and the row supplies no gradient, which is the eligibility condition $|C_j(B)| \geq 2$ in Equation (14).

A.5 Application to the Shared Candidate Pool

For an anchor batch B, recall $\mathcal{P}_B = B \cup \bigcup_{i \in B} \mathcal{N}(i)$, the observed support $C_j(B) = \mathcal{N}(j) \cap \mathcal{P}_B$, and the eligible set $R_B = \{j \in \mathcal{P}_B : |C_j(B)| \geq 2\}$. Every anchor $i \in B$ has $\mathcal{N}(i) \subseteq \mathcal{P}_B$, so $C_i(B) = \mathcal{N}(i)$ is an optimal size-k support and $B \subseteq R_B$ whenever $k \geq 2$. A non-anchor row may observe a subset that is not optimal at its cardinality, but conditioning on the realized support still gives

$$\operatorname*{arg\,min}_{r_j \in \Delta(C_j(B))} D_{\mathrm{KL}}(r_j \| \pi_j) = \pi_j(\cdot \mid C_j(B)) = P_j^T(\cdot \mid C_j(B)), \tag{34}$$

so the fixed-support conclusion applies to every eligible row without recursive pool expansion. Since each row KL is nonnegative, $\mathcal{L}_{\mathrm{NeST}}(\theta; B) = 0$ exactly when every eligible row has zero KL, and Proposition A.1 then holds on each observed support. The auxiliary distributions $Q_i^\alpha(\cdot \mid B)$ use the broader support $\mathcal{P}_B \setminus \{i\}$ and the common temperature $\bar{\tau}$; the same proof yields similarity-gap constraints across that support when an auxiliary row has zero loss.

B Additional Experimental Results

We first report the two additional teacher–student pairs, then extend all three pairs to a larger corpus. The remaining sections provide controlled studies using Qwen3-Embedding-0.6B→MiniLMv2-H384. Scores summarize three seeds unless stated otherwise.

B.1 Additional Teacher–Student Pairs

B.1.1 MiniLMv2-H768 Student

A wide capacity gap does not erase the gain. Table 3 places NeST-KD at 77.83, 1.18 points above TALAS and 3.19 above DistillCSE, the strongest other baseline. Avg-Out rises by 1.22 points over TALAS, supported by clear gains on SICK and STS12. NeST-KD also improves Emotion, Banking77, STS-B, and SciTail, although it regresses on Tweet, MRPC, and WiC.

Table 3: BGE-M3 → MiniLMv2-H768 (66M); ⋆ marks in-domain tasks. **Bold** denotes the best distilled result, and underlined denotes the second best.

	Classification (F1)			Pair Classification (AP)		
Method	Emotion⋆	Banking77	Tweet	WiC⋆	MRPC	SciTail
Teacher	68.56	93.52	73.85	61.47	85.81	91.87
Student base	71.67	81.84	47.70	60.89	79.67	65.06
SimCSE-unsup	54.37 ±0.32	89.21 ±0.14	70.31 ±0.18	69.18 ±0.16	82.36 ±0.11	75.09 ±0.17
RKD	63.48 ±0.27	87.31 ±0.09	74.62 ±0.09	63.91 ±0.10	86.02 ±0.06	77.32 ±0.11
CDM	57.21 ±0.51	89.41 ±0.19	68.57 ±0.09	**71.13** ±0.02	82.11 ±0.07	77.38 ±0.40
DSKD	61.08 ±0.61	86.15 ±0.16	71.62 ±0.32	68.24 ±0.04	84.66 ±0.18	73.94 ±0.12
Jasper and Stella	55.42 ±0.74	88.05 ±0.28	68.73 ±0.12	70.91 ±0.05	81.39 ±0.12	76.19 ±0.12
DistillCSE	60.96 ±0.29	88.78 ±0.16	74.31 ±0.15	68.10 ±0.13	86.16 ±0.10	77.14 ±0.20
EMO	57.34 ±0.30	89.08 ±0.18	69.88 ±0.14	70.63 ±0.22	80.84 ±0.18	76.05 ±0.58
TALAS	65.83 ±0.58	89.72 ±0.19	**75.38** ±0.11	66.96 ±0.07	**87.58** ±0.02	82.90 ±0.12
NeST-KD	**68.47** ±0.09	**90.96** ±0.08	73.40 ±0.08	62.99 ±0.08	85.34 ±0.01	**83.91** ±0.06

	STS (Spearman)			Aggregate		
Method	STS-B⋆	SICK	STS12	Avg-In	Avg-Out	Avg.
Teacher	84.87	79.18	78.73	71.63	83.83	79.76
Student base	24.12	49.46	26.76	52.23	58.42	56.35
SimCSE-unsup	62.49 ±0.26	68.37 ±0.18	58.31 ±0.21	62.01 ±0.17	73.94 ±0.10	69.97 ±0.13
RKD	73.22 ±0.38	71.51 ±0.19	63.66 ±0.29	66.87 ±0.07	76.74 ±0.06	73.45 ±0.04
CDM	64.14 ±0.15	67.65 ±0.16	59.32 ±0.34	64.16 ±0.13	74.07 ±0.06	70.77 ±0.02
DSKD	66.44 ±0.18	64.12 ±0.37	62.68 ±0.31	65.25 ±0.27	73.86 ±0.10	70.99 ±0.11
Jasper and Stella	65.72 ±0.10	70.45 ±0.20	61.79 ±0.28	64.02 ±0.27	74.43 ±0.05	70.96 ±0.11
DistillCSE	72.16 ±0.24	75.48 ±0.17	68.67 ±0.19	67.07 ±0.15	78.42 ±0.09	74.64 ±0.16
EMO	62.66 ±0.03	63.32 ±0.55	55.48 ±0.27	63.54 ±0.19	72.44 ±0.10	69.48 ±0.12
TALAS	77.51 ±0.18	75.82 ±0.11	68.11 ±0.20	70.10 ±0.13	79.92 ±0.05	76.65 ±0.04
NeST-KD	**82.22** ±0.05	**78.50** ±0.04	**74.71** ±0.14	**71.23** ±0.04	**81.14** ±0.03	**77.83** ±0.01

Table 4: Qwen3-Embedding-4B → BERT-base (109M); ⋆ marks in-domain tasks. **Bold** denotes the best distilled result, and underlined denotes the second best.

	Classification (F1)			Pair Classification (AP)		
Method	Emotion⋆	Banking77	Tweet	WiC⋆	MRPC	SciTail
Teacher	68.94	93.63	72.22	66.80	84.79	91.77
Student base	68.02	84.86	47.79	59.22	77.36	67.74
SimCSE-unsup	54.03 ±0.28	88.23 ±0.14	70.03 ±0.19	71.07 ±0.16	83.36 ±0.11	77.66 ±0.18
RKD	59.26 ±0.05	90.39 ±0.11	70.26 ±0.31	69.85 ±0.18	80.79 ±0.26	76.19 ±0.13
CDM	53.41 ±0.50	87.60 ±0.23	69.58 ±0.20	70.91 ±0.09	84.00 ±0.06	76.56 ±0.27
DSKD	52.65 ±0.31	89.20 ±0.16	68.14 ±0.28	**72.29** ±0.08	82.28 ±0.15	77.98 ±0.59
Jasper and Stella	54.07 ±0.50	87.42 ±0.30	69.91 ±0.17	71.43 ±0.09	**84.71** ±0.10	74.33 ±0.54
DistillCSE	61.53 ±0.33	90.49 ±0.15	69.98 ±0.20	69.38 ±0.13	82.36 ±0.12	79.61 ±0.21
EMO	52.02 ±1.09	86.89 ±0.07	68.99 ±0.11	69.62 ±0.16	79.74 ±0.12	76.37 ±0.35
TALAS	69.01 ±0.52	90.44 ±0.13	74.13 ±0.17	68.84 ±0.15	84.60 ±0.10	81.78 ±0.29
NeST-KD	**69.46** ±0.34	**92.00** ±0.12	**74.21** ±0.17	66.85 ±0.10	84.62 ±0.04	**85.55** ±0.19

	STS (Spearman)			Aggregate		
Method	STS-B⋆	SICK	STS12	Avg-In	Avg-Out	Avg.
Teacher	86.87	83.43	82.55	74.20	84.73	81.22
Student base	20.30	42.43	21.54	49.18	56.95	54.36
SimCSE-unsup	69.73 ±0.24	69.01 ±0.19	60.82 ±0.22	64.94 ±0.16	74.85 ±0.10	71.55 ±0.12
RKD	74.75 ±0.08	72.35 ±0.07	68.25 ±0.18	67.95 ±0.05	76.37 ±0.07	73.57 ±0.06
CDM	68.29 ±0.42	68.50 ±0.37	60.63 ±0.37	64.20 ±0.28	74.48 ±0.12	71.05 ±0.18
DSKD	70.11 ±0.12	66.82 ±0.66	61.84 ±0.26	65.02 ±0.12	74.38 ±0.14	71.26 ±0.10
Jasper and Stella	70.51 ±0.36	70.78 ±0.64	63.62 ±0.21	65.34 ±0.17	75.13 ±0.13	71.86 ±0.06
DistillCSE	77.93 ±0.27	74.59 ±0.18	73.63 ±0.23	69.61 ±0.15	78.44 ±0.11	75.50 ±0.14
EMO	64.54 ±0.50	67.89 ±0.47	56.46 ±0.49	62.06 ±0.49	72.72 ±0.15	69.17 ±0.24
TALAS	79.24 ±0.10	76.53 ±0.10	73.12 ±0.30	72.36 ±0.19	80.10 ±0.01	77.52 ±0.07
NeST-KD	**82.60** ±0.03	**77.53** ±0.10	**75.08** ±0.07	**72.97** ±0.13	**81.50** ±0.05	**78.66** ±0.01

B.1.2 BERT-BASE STUDENT

The gain survives a change of student architecture. Table 4 places NeST-KD at 78.66, 1.14 points above TALAS. It leads seven of nine tasks, with its largest margins over the best baseline reaching 3.77 points on SciTail and 3.36 on STS-B. Avg-Out increases by 1.40 points, while Avg-In rises by 0.61, from 72.36 to 72.97. These gains show that the benefit extends beyond MiniLM students to a larger BERT-base student.

B.2 LARGE-CORPUS RESULTS

B.2.1 EXPANDED 100K CORPUS

Training Data. The 100K corpus contains 102,361 distinct unlabeled texts collected from the training splits of all nine evaluation benchmarks: Banking77, Tweet, Emotion, MRPC, SciTail, WiC, SICK, STS12, and STS-B. We use only the raw text; task labels, validation examples, and test examples are excluded from distillation. Unlike the main 13,553-text corpus, which draws from Emotion, WiC, and STS-B, this corpus expands both data volume and training-data coverage. Table 5 therefore measures the effect of broader text coverage rather than data volume alone.

Implementation Details. We retain the three teacher–student pairs and train every method for five epochs with maximum sequence length 256 and a nominal batch size of 128. For NeST-KD, this batch comprises 128 sampled anchors per optimizer step; their $k = 100$ teacher neighborhoods are merged and deduplicated before each distinct pooled text is encoded once. NeST-KD uses learning rate 3×10^{-5} and $\lambda_{\text{aux}} = 0.5$, with teacher embeddings, neighborhoods, and temperatures precomputed. All runs are conducted using a single NVIDIA H200 GPU, with the configuration fixed before test evaluation.

Table 5: Nine-task results on the expanded 100K corpus, grouped by teacher–student pair. **Bold** denotes the best distilled result in each column, and underlined denotes the second best.

	Classification (F1)			Pair Classification (AP)			STS (Spearman)			
Method	Banking77	Tweet	Emotion	MRPC	SciTail	WiC	SICK	STS12	STS-B	Avg.
(a) Qwen3-Embedding-4B → BERT-base (109M)										
Teacher	93.26	71.04	69.08	84.80	91.77	66.79	83.43	82.55	86.87	81.07
Student (pre-KD)	84.33	67.08	47.79	77.36	67.74	59.22	42.43	21.54	20.29	54.20
SimCSE-unsup	89.24	70.59	54.37	84.26	77.43	67.99	69.14	62.59	70.23	71.76
RKD	92.23	71.96	63.79	85.00	80.27	66.31	79.32	71.72	77.70	76.48
CDM	89.16	69.63	53.74	85.57	71.22	**70.04**	71.26	66.97	73.69	72.36
DSKD	88.87	69.05	54.02	85.90	70.02	69.63	71.07	66.52	73.52	72.07
Jasper and Stella	90.02	71.90	60.14	85.96	78.82	69.93	76.53	69.79	77.22	75.59
DistillCSE	92.32	73.92	67.45	**86.64**	83.48	68.03	79.52	77.95	84.02	79.26
EMO	84.96	68.93	54.80	85.05	65.75	68.80	62.00	61.75	69.54	69.07
TALAS	93.12	**74.67**	71.42	86.30	87.06	68.09	80.29	76.34	82.26	79.95
NeST-KD	**93.43**	72.98	**71.44**	84.86	**89.82**	65.22	**80.51**	**78.55**	**84.42**	**80.14**
(b) BGE-M3 → MiniLMv2-H768 (66M)										
Teacher	93.49	73.75	69.03	85.80	91.87	61.46	79.18	78.74	84.86	79.80
Student (pre-KD)	81.18	71.30	48.85	79.66	65.06	60.89	49.46	26.76	24.12	56.36
SimCSE-unsup	87.50	71.29	59.51	84.76	69.45	66.49	65.83	60.90	63.70	69.94
RKD	91.68	74.57	62.45	85.81	79.81	63.57	76.91	69.54	75.56	75.54
CDM	87.64	70.82	56.37	85.67	67.69	66.89	70.15	66.68	69.49	71.27
DSKD	86.69	69.76	56.98	84.92	66.75	**67.23**	68.26	66.28	68.31	70.58
Jasper and Stella	88.72	73.43	61.63	86.39	74.13	66.98	73.32	68.30	72.66	73.95
DistillCSE	92.29	**75.74**	67.85	86.77	83.78	64.89	76.72	75.39	82.01	78.38
EMO	87.17	71.07	59.32	84.25	67.89	66.09	69.04	62.26	66.47	70.40
TALAS	92.44	75.58	65.13	**87.09**	85.39	64.39	78.14	72.56	81.03	77.97
NeST-KD	**92.94**	74.54	**69.45**	86.11	**89.72**	61.83	**78.59**	**78.28**	**84.11**	**79.51**
(c) Qwen3-Embedding-0.6B → MiniLMv2-H384 (22M)										
Teacher	92.51	71.28	65.80	83.38	90.09	66.84	80.65	77.11	84.59	79.14
Student (pre-KD)	67.03	69.45	42.90	78.39	64.55	59.58	47.60	21.96	22.08	52.62
SimCSE-unsup	85.69	70.29	54.26	84.45	69.91	65.49	67.89	62.43	67.23	69.74
RKD	89.74	72.34	57.99	84.13	75.20	67.49	74.48	68.12	72.55	73.56
CDM	85.18	69.14	52.30	84.02	68.24	66.57	69.15	63.05	66.55	69.36
DSKD	85.04	69.21	52.85	84.34	67.33	65.98	69.01	63.46	66.42	69.29
Jasper and Stella	85.74	70.25	53.48	84.43	71.23	66.51	70.96	63.02	67.42	70.34
DistillCSE	91.32	73.33	66.23	**86.26**	79.90	**67.61**	**78.03**	74.89	80.76	77.59
EMO	85.08	68.11	53.96	84.38	65.48	65.23	68.61	60.00	65.38	68.47
TALAS	89.79	**73.87**	61.75	85.99	82.07	65.79	75.67	73.91	78.31	76.35
NeST-KD	**91.66**	73.31	**66.81**	84.82	**86.12**	65.79	77.26	**77.81**	**80.88**	**78.27**

A broader training corpus yields stronger students. With the expanded 100K corpus, NeST-KD reaches 80.14, 79.51, and 78.27 across the three teacher–student pairs, improving over the corresponding 15K results by 1.48, 1.68, and 2.32 points. It exceeds the strongest non-NeST row by 0.19, 1.13, and 0.68 points, respectively, with TALAS strongest for BERT-base and DistillCSE strongest for both MiniLM students. Improvements are especially consistent on SciTail, STS12, and STS-B, indicating that broader neighborhood coverage benefits relation-sensitive evaluations.

B.2.2 WIKI100K CORPUS

Training Data. Wiki100K contains 100,000 unlabeled texts sampled from English Wikipedia. Unlike the expanded corpus in Section B.2.1, it does not draw from the downstream benchmark training sets and provides a task-independent source of general-domain text for distillation.

Table 6: Qwen3-Embedding-0.6B → MiniLMv2-H384 trained on Wiki100K. Values are mean ± sample standard deviation over seeds 42, 43, and 44. **Bold** denotes the best result, and underlined denotes the second best.

	Classification (F1)			Pair Classification (AP)		
Method	Emotion	Banking77	Tweet	WiC	MRPC	SciTail
SimCSE-unsup	55.65 ±0.41	86.06 ±0.32	70.59 ±0.12	63.88 ±0.16	82.96 ±0.19	74.81 ±0.88
RKD	51.88 ±0.48	87.66 ±0.17	70.64 ±0.12	66.81 ±0.11	**83.96** ±0.02	75.83 ±0.31
Jasper and Stella	51.43 ±0.63	85.29 ±0.23	69.75 ±0.30	66.84 ±0.20	83.51 ±0.16	74.32 ±0.44
DistillCSE	56.13 ±0.89	88.15 ±0.12	71.03 ±0.28	67.15 ±0.09	82.87 ±0.16	82.33 ±0.26
TALAS	54.46 ±0.63	87.54 ±0.13	71.09 ±0.19	65.74 ±0.15	83.28 ±0.11	82.53 ±0.16
NeST-KD	**56.83** ±0.95	**88.78** ±0.15	**71.24** ±0.05	**67.23** ±0.08	82.60 ±0.06	**83.72** ±0.12

	STS (Spearman)			Aggregate
Method	STS-B	SICK	STS12	Avg.
SimCSE-unsup	64.03 ±0.91	67.11 ±0.73	57.55 ±2.14	69.18 ±0.45
RKD	62.66 ±0.06	66.70 ±0.11	56.76 ±0.35	69.21 ±0.05
Jasper and Stella	61.68 ±0.40	67.36 ±0.36	56.23 ±0.25	68.49 ±0.08
DistillCSE	71.56 ±0.15	72.31 ±0.21	67.20 ±0.19	73.19 ±0.14
TALAS	**73.81** ±0.09	71.96 ±0.04	**67.34** ±0.07	73.08 ±0.14
NeST-KD	71.77 ±0.05	**72.46** ±0.03	65.71 ±0.12	**73.37** ±0.12

NeST-KD retains its edge on general-domain text. Table 6 places NeST-KD first at 73.37, followed by DistillCSE at 73.19 and TALAS at 73.08. NeST-KD leads Emotion, Banking77, Tweet, WiC, SciTail, and SICK, while TALAS is strongest on STS-B and STS12 and RKD leads MRPC. Overall, NeST-KD improves over RKD by 4.16 points and exceeds the strongest competing average by 0.18 points under task-independent Wikipedia supervision.

B.2.3 WIKI1M CORPUS

Training Data. Wiki1M contains 974,010 unique unlabeled sentences sampled from English Wikipedia. Like Wiki100K, it is collected independently of the downstream benchmark training splits and supplies task-independent, general-domain text for distillation. Its nearly one-million-text scale is approximately 9.7× larger than Wiki100K, allowing us to test whether neighborhood-based supervision remains effective as the corpus expands substantially.

Table 7: Qwen3-Embedding-0.6B → MiniLMv2-H384 trained on Wiki1M (974,010 unique Wikipedia sentences). All runs use seed 42. **Bold** denotes the best result, and underlined denotes the second best.

	Classification (F1)			Pair Classification (AP)			STS (Spearman)			
Method	Emotion	Banking77	Tweet	WiC	MRPC	SciTail	STS-B	SICK	STS12	Avg.
SimCSE-unsup	18.47	58.04	49.49	55.07	80.07	67.40	36.91	51.01	36.98	50.38
RKD	27.79	76.51	56.88	57.79	81.51	72.08	54.93	65.53	58.98	61.33
Jasper and Stella	55.97	88.79	71.72	65.96	84.87	81.03	71.44	72.38	64.46	72.96
TALAS	56.88	90.00	**71.88**	**66.72**	**84.94**	86.20	**78.15**	**73.00**	68.85	75.18
NeST-KD	**57.69**	**90.96**	71.14	65.98	82.87	**87.69**	78.14	72.88	**70.21**	**75.28**

NeST-KD keeps its lead at million-text scale. Table 7 shows that NeST-KD reaches an average of 75.28 while training on 974,010 task-independent Wikipedia sentences. It exceeds TALAS by 0.10 points and outperforms RKD and SimCSE-unsup by 13.95 and 24.90 points, respectively. NeST-KD also leads on Emotion, Banking77, SciTail, and STS12, showing that its shared-neighborhood supervision remains effective when the training corpus grows to nearly one million texts.

B.3 Retrieval Benchmarks

Metrics. We evaluate zero-shot retrieval on ArguAna, SciFact, SciDocs, NFCorpus, and FiQA-2018 from BEIR, reporting nDCG@10 for each benchmark and their unweighted mean. No retrieval labels or relevance judgments are used for distillation or checkpoint selection.

Implementation Details. We train Qwen3-Embedding-0.6B→MiniLMv2-H384 on the same 974,010-sentence Wiki1M corpus as Section B.2.3, using five epochs, batch size 128, maximum sequence length 256, and seed 42. NeST-KD uses $k = 100$, $\lambda_{\text{aux}} = 0.5$, and learning rate 3×10^{-5}; every final checkpoint is evaluated without retrieval-specific fine-tuning. Because training the full baseline suite on Wiki1M is computationally costly, we restrict the comparison to methods compatible with black-box teacher embeddings and include SimCSE-unsup as a teacher-free reference.

Table 8: Zero-shot BEIR retrieval (nDCG@10) for Qwen3-Embedding-0.6B → MiniLMv2-H384 trained on Wiki1M (974,010 unique Wikipedia sentences). All runs use seed 42. **Bold** denotes the best distilled result, and underlined denotes the second best.

Method	ArguAna	SciFact	SciDocs	NFCorpus	FiQA-2018	Avg.
Teacher	68.11	68.26	15.72	29.73	38.71	44.11
Student (pre-KD)	9.85	0.66	0.32	1.28	0.05	2.43
SimCSE-unsup	1.82	6.44	1.69	6.17	0.90	3.40
RKD	13.87	0.67	1.09	1.94	1.38	3.79
Jasper and Stella	44.90	30.88	4.81	10.57	7.55	19.74
TALAS	**49.27**	37.75	7.80	11.24	10.54	23.32
NeST-KD	46.90	**51.74**	**8.76**	**20.93**	**17.39**	**29.14**

The distilled geometry transfers to zero-shot retrieval. Table 8 shows that NeST-KD reaches 29.14 average nDCG@10, improving over TALAS by 5.82 points and outperforming RKD and SimCSE-unsup by more than 25 points. It leads four of the five retrieval benchmarks, including gains over TALAS of 13.99 points on SciFact, 9.69 on NFCorpus, and 6.85 on FiQA-2018, while TALAS remains 2.37 points higher on ArguAna. These results show that the learned neighborhood structure transfers beyond the nine sentence-embedding benchmarks to zero-shot retrieval.

B.4 Sensitivity Analysis

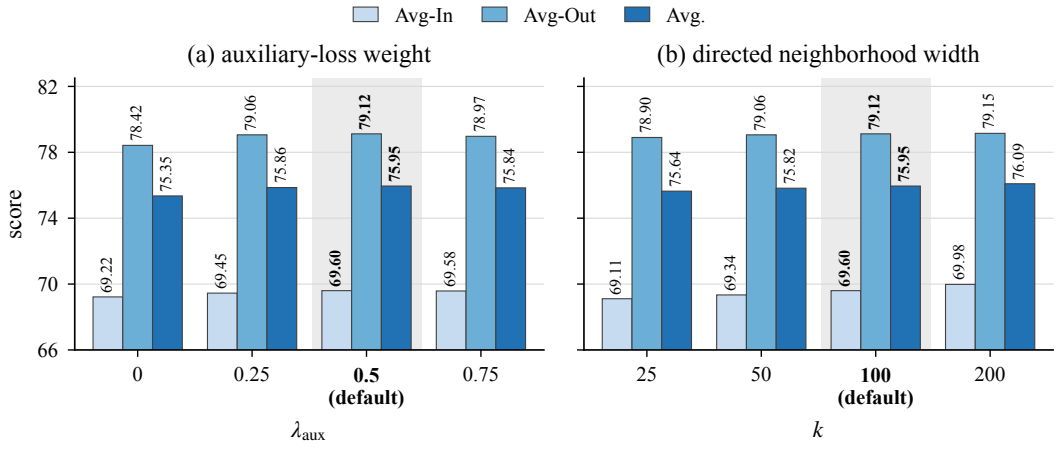


Figure 6: Sensitivity analyses for Qwen3-Embedding-0.6B→MiniLMv2-H384. Panel (a) varies the auxiliary-loss weight, and panel (b) varies the directed neighborhood width. Shaded groups mark the default settings.

Both hyperparameters are stable around their defaults. In Figure 6(a), $\lambda_{\text{aux}} = 0.5$ gives the highest average of 75.95, while 0.25 and 0.75 remain within 0.11 points; removing the auxiliary term reduces the average to 75.35. Figure 6(b) shows gradual saturation as k increases: moving

from $k = 100$ to $k = 200$ changes the average by only 0.14 points. We therefore retain $\lambda_{\text{aux}} = 0.5$ and $k = 100$ as stable default settings.

B.5 Robustness

Table 9: Robustness controls for Qwen3-Embedding-0.6B→MiniLMv2-H384.

Configuration	Avg-In	Avg-Out	Avg.
Default	69.60	79.12	75.95
Inclusion reweighting	69.58	79.10	75.92
Student kNN	69.28	78.98	75.75
Mutual kNN	69.05	79.18	75.81
Global temperature	**69.74**	**79.26**	**76.09**

NeST-KD is robust to graph and temperature choices. Table 9 shows that inclusion reweighting remains close to the default, trailing by 0.03 points. Student-derived neighborhoods and mutual filtering reduce Avg. by 0.20 and 0.14 points, respectively, while a global temperature improves it by 0.14 points. The narrow 0.34-point range shows that the gain does not hinge on one graph construction or temperature policy.

B.6 Computation Efficiency

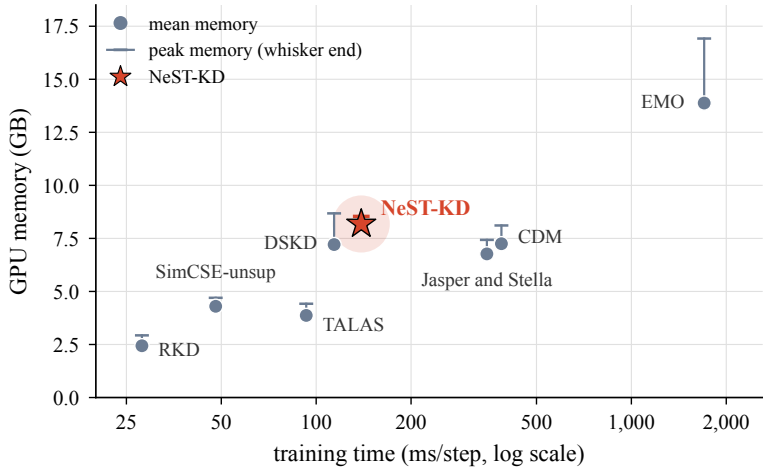


Figure 7: Training compute and GPU-memory usage for Qwen3-Embedding-0.6B→MiniLMv2-H384 on one NVIDIA H200. Markers show mean memory and whiskers extend to peak memory; the time axis is logarithmic. All methods use batch size 128, maximum length 256. One-time model loading, teacher-cache construction, and graph construction are excluded.

NeST-KD maintains a practical efficiency profile. As shown in Figure 7, at 139 ms/step it is only 25 ms slower than DSKD and 46 ms slower than TALAS, while remaining substantially faster than Jasper and Stella, CDM, and EMO at 348, 387, and 1,702 ms/step. Its mean and peak memory are 8.18 and 8.54 GB, respectively, close to the corresponding values for DSKD and CDM and well below those of EMO.

B.7 Neighborhood-Rank Fidelity

Graded targets better preserve the teacher's neighbor ordering. Figure 8 concentrates more probability near the agreement diagonal for NeST-KD than for uniform targets. Teacher-to-student Spearman correlation consequently rises from 0.43 to 0.50, indicating fewer large rank displacements within the retrieved neighborhoods. Together with the probability and similarity-gap results in Figures 4 and 5, this shows that graded supervision preserves both relation strength and candidate order. The diagnostic is descriptive and was not used for method or hyperparameter selection.

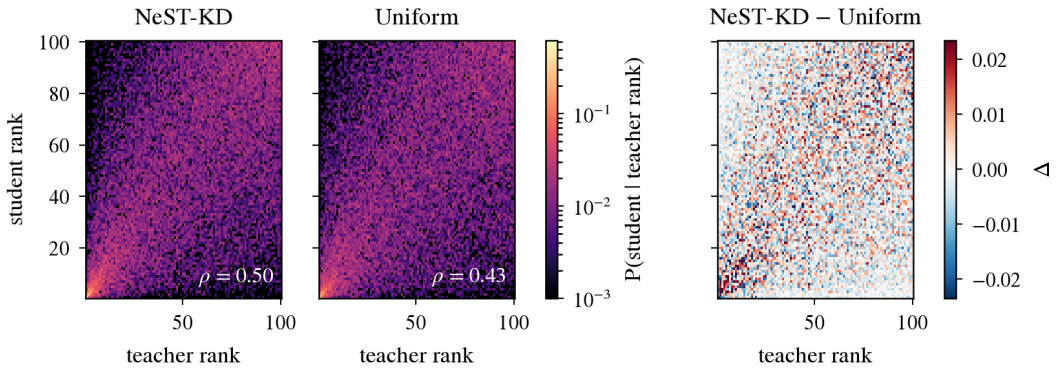


Figure 8: Teacher-to-student neighborhood-rank transitions for Qwen3-Embedding-0.6B→MiniLMv2-H384. The left panels show transition probabilities for NeST-KD and uniform targets; the right panel shows their difference.

C Implementation and Reproducibility Details

C.1 Training Configuration

Table 10 records the default NeST-KD configuration. The neighborhood and auxiliary weights correspond to Equations (17) and (21). Teacher embeddings and graph artifacts are cached separately for each teacher–student pair. Student initialization uses BERT-base-uncased with Qwen3-Embedding-4B, MiniLMv2-L6-H768-distilled-from-BERT-Base with BGE-M3, and MiniLMv2-L6-H384-distilled-from-BERT-Base with Qwen3-Embedding-0.6B.

Table 10: Default NeST-KD training configuration. Ablations change only their specified controls.

Parameter	Value
Hardware per run	One NVIDIA H200 GPU
Optimizer	AdamW
Learning rate / minimum	3×10^{-5} / 3×10^{-6}
Schedule / warmup	Cosine decay / 6% of updates
Training epochs	5
Anchor batch size b	128
Maximum training sequence length	256 tokens
Student encoding chunk size	256 texts
Teacher neighborhood width k	100, directed, self excluded
Neighborhood loss weight	1.0
Auxiliary weight λ_{aux}	0.5
Auxiliary temperature $\bar{\tau}$	Median teacher row temperature
Eligible row centers	At least two observed teacher neighbors
Inclusion reweighting	Disabled by default
Checkpoint evaluated	Final epoch
Training seeds	42, 43, 44

The shared pool changes exposure while retaining a fixed training configuration. Table 10 separates the 128 sampled anchors from the variable number of pooled texts encoded at each update. Each distinct text is encoded once, in chunks, and the losses jointly update the student. Controlled studies use the same optimizer settings and final-epoch evaluation unless their captions specify otherwise.

Pooling. Teacher caches use the last non-padding token for Qwen3-Embedding models and the first token for BGE-M3. The student uses its first-token representation; teacher and student vectors are normalized before computing training similarities.

C.1.1 How the Table 2 Controls Differ

All controls use the same corpus, five epochs, anchor batch size, optimizer schedule, number of anchor draws, and number of optimizer updates. They do not all encode the same number of texts: pointwise and in-batch controls encode 128 texts per update, whereas every graph-pool control encodes the same pool $\mathcal{P}_B$ as NeST-KD, and Random pool matches $|\mathcal{P}_B|$ at every update. Two comparisons therefore control for the student-encoding budget. First, In-batch, neighbor encodes 128 texts per update, the same as In-batch, random and fewer than TALAS, which runs four student forward passes over 128-example batches per update; yet it reaches 74.51, above both In-batch, random at 71.84 and TALAS at 73.31. Second, Uniform and Shuffled targets encode the same $\mathcal{P}_B$ as NeST-KD and trail it by 1.23 and 1.26 points, so these gaps reflect target assignment rather than greater exposure. We vary how texts are grouped and supervised. A random batch and its neighbor-composed counterpart each contain 128 distinct texts. The neighbor-composed version starts from a randomly selected seed, adds its unused teacher neighbors, and fills any remaining positions with unused random texts.

For the graph-pool controls, let B be the sampled anchors and $\mathcal{P}_B = B \cup \bigcup_{i \in B} \mathcal{N}(i)$ be their deduplicated neighborhood pool. For a pooled text j, the observed neighborhood is $C_j = \mathcal{N}(j) \cap \mathcal{P}_B$, and $R_B = \{j \in \mathcal{P}_B : |C_j| \geq 2\}$ contains the texts eligible to serve as relational row centers. Table 11 groups the controls by the question they test.

Table 11: Exact construction of the Table 2 controls.

Setting	Encoded texts	Supervision
Does neighbor exposure help?		
Pointwise, random	Random batch B	No relational rows. Each $i \in B$ minimizes $1 - \cos(Wz_i^S, z_i^T)$.
Pointwise, neighbor	Neighbor batch B	The same pointwise loss; only batch composition changes.
In-batch, random	Random batch B	Every $i \in B$ is a center with support $B \setminus \{i\}$. The teacher target uses $\bar{\tau}$, and the loss is $0.5\mathcal{L}_{\text{aux}}$.
In-batch, neighbor	Neighbor batch B	The same in-batch relational loss; only batch composition changes.
Which rows and target assignments matter?		
Anchor rows only	Graph pool $\mathcal{P}_B$	Only eligible anchors $j \in B \cap R_B$ are row centers. Each uses support C_j, graded teacher targets, and $0.5\mathcal{L}_{\text{aux}}$.
Uniform targets	Graph pool $\mathcal{P}_B$	Centers R_B and supports C_j are unchanged, but each row target is uniform. The auxiliary target remains the teacher distribution.
Shuffled targets	Graph pool $\mathcal{P}_B$	Centers and supports are unchanged, but teacher probabilities are permuted within each C_j, preserving the row's histogram and entropy. The auxiliary target is unchanged.
Do the pool and candidate support matter?		
Random columns	Graph pool $\mathcal{P}_B$	Centers remain R_B, but each row replaces C_j with $\|C_j\|$ uniformly sampled non-self columns. The teacher softmax is recomputed at τ_j, and the auxiliary loss is unchanged.
Random pool	Random pool of size $\|\mathcal{P}_B\|$	A sampled text is a center when at least two of its teacher neighbors are present. The observed graph support, graded targets, and auxiliary loss otherwise follow NeST-KD.
Dense rows	Graph pool $\mathcal{P}_B$	Every pooled text is a center with support $\mathcal{P}_B \setminus \{j\}$. The teacher target uses τ_j, with $0.5\mathcal{L}_{\text{aux}}$.
Dense rows only	Graph pool $\mathcal{P}_B$	The same dense centers, support, and teacher targets, with $\lambda_{\text{aux}} = 0$.
NeST-KD	Graph pool $\mathcal{P}_B$	Centers R_B use supports C_j and graded teacher targets at τ_j, together with $0.5\mathcal{L}_{\text{aux}}$.

Random pool matches the realized graph-pool size separately at every update rather than only in expectation. The remaining graph-pool controls encode the same $\mathcal{P}_B$ once and reuse those student representations. Their differences therefore come from row eligibility, candidate support, target construction, or the auxiliary term rather than additional encoder forward passes.

C.2 Optimization Procedure

Algorithm 1 makes explicit which quantities are precomputed and which are constructed at every optimization step. The teacher embeddings, neighborhoods, and temperatures remain fixed; gradients pass only through the student embeddings computed for the current shared pool.

Algorithm 1 NeST-KD optimization

Require: Corpus $\mathcal{D}$, teacher f_T, student f_θ, neighborhood width k, anchor batch size b, auxiliary weight λ_{aux}, epochs E

Cache normalized teacher embeddings $\{z_i^T\}_{i\in\mathcal{I}}$
Construct $\mathcal{N}(j)$ and τ_j for every $j \in \mathcal{I}$; set $\bar{\tau} \leftarrow \text{median}_{j\in\mathcal{I}}\, \tau_j$
for $e = 1, \ldots, E$ **do**
  for each of the $\lfloor N/b \rfloor$ anchor batches B of a random partition of $\mathcal{I}$ **do**
    $\mathcal{P}_B \leftarrow B \cup \bigcup_{i\in B} \mathcal{N}(i)$ and remove duplicate indices
    Encode each x_j, $j \in \mathcal{P}_B$, once with f_θ and normalize the resulting embeddings
    $C_j(B) \leftarrow \mathcal{N}(j) \cap \mathcal{P}_B$ for every $j \in \mathcal{P}_B$
    $R_B \leftarrow \{j \in \mathcal{P}_B : |C_j(B)| \geq 2\}$
    Compute $\mathcal{L}_{\text{NeST}}$ on $\{(j, C_j(B)) : j \in R_B\}$ using τ_j
    Compute $\mathcal{L}_{\text{aux}}$ for $i \in B$ over $\mathcal{P}_B \setminus \{i\}$ using $\bar{\tau}$
    $\mathcal{L} \leftarrow \mathcal{L}_{\text{NeST}} + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}}$
    Update θ with AdamW using $\nabla_\theta \mathcal{L}$
  end for
end for
return f_θ

C.3 Evaluation Procedure

Frozen-embedding classification. Banking77, Tweet, and Emotion use a logistic-regression probe trained on their labeled training splits, with frozen student embeddings as features. The probe uses scikit-learn's `LogisticRegression` with `random_state=42`, `max_iter=200`, and otherwise default settings. It receives the unnormalized first-token vectors and predicts labels on the evaluation split; we report macro-F1. A convergence warning is recorded if the solver reaches the iteration limit. Probe fitting uses task labels only for evaluation of the frozen encoder, not for distillation.

Pair and similarity tasks. MRPC, SciTail, and WiC use average precision from cosine similarity; this metric does not require choosing a classification threshold. SICK, STS12, and STS-B use Spearman correlation between cosine-based scores and gold similarity labels. Classification inputs are truncated to 512 tokens and pair inputs to 128 tokens. Evaluation uses batches of 256 texts or pairs with the encoder in evaluation mode. The reported main-run checkpoints are the final-epoch weights, and task scores are scaled by 100 before computing the descriptive nine-task mean.