BALLPARK MINING: REGION-BASED CONTRASTIVE DISTILLATION FOR EMBEDDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

When compressing large embedding models through knowledge distillation, aligning student and teacher representations alone can leave substantial gaps in downstream performance. Contrastive objectives over hard negatives mined from the teacher's embedding space are therefore commonly added to supply additional supervision. However, hard-negative mining suffers from a well-known limitation: the presence of false negatives. In particular, in duplicate-rich corpora, the nearest candidates to an anchor are often its paraphrases, so the contrastive loss pushes apart texts with the same meaning. Existing remedies can require similarity labels or auxiliary models. To address this, we propose *Ballpark Mining*, a label-free contrastive distillation method that uses teacher-derived regions to select negatives. Using the variability of a frozen teacher's embeddings under stochastic perturbation, we define for each anchor a region of representations that the teacher cannot reliably distinguish from it. Candidates within this region are treated as probable false negatives, and the nearest candidates outside it are selected as negatives. We evaluate our proposed method with teachers ranging from 110M to 27B parameters on a deduplicated Wikipedia corpus and a duplicate-rich Quora corpus, using semantic similarity and retrieval tasks from MTEB and BEIR. On Quora, distilling a 110M-parameter teacher into an 11M-parameter student with Ballpark Mining reduces the fraction of duplicate negatives from 21.9% to 2.3% and increases duplicate recall@10 from 0.919 to 0.973. The gains extend to a 27B teacher, where distillation on Quora improves the average score across semantic similarity and retrieval tasks from 0.211 to 0.285, while performance on Wikipedia remains comparable, which has lesser duplicates. In the 110M-to-11M setting, estimating regions at an intermediate teacher layer further improves the average score by approximately 3% on Quora and 6% on Wikipedia relative to final-layer regions. Together, our studies show that the perturbation behavior of the teacher provides a simple, label-free criterion for negative selection in embedding distillation, while identifying a corpus-transferable choice of region size as a key challenge.

1 INTRODUCTION

Text embeddings are used for a wide range of downstream cases such as semantic search, duplicate detection, and retrieval by representing texts in a space where similarity reflects their meaning (Reimers & Gurevych, 2019; Gao et al., 2021). Deploying these models under latency and throughput constraints motivates compressing large encoders into smaller models. Knowledge distillation provides a framework for this compression by training a compact student using the outputs of a pretrained teacher (Hinton et al., 2015). In this work, we consider distillation from an unlabeled text corpus, where a frozen teacher supplies the representations used to train the student. The objective is to preserve the semantic relationships needed for downstream similarity and retrieval tasks cost-effectively.

An obvious approach is to align each student embedding with the teacher's embedding of the same input. However, the capacity gap between teacher and student can limit the effectiveness of representation matching (Cho & Hariharan, 2019; Mirzadeh et al., 2020). For a compact embedding model, training must also convey which distinctions among nearby texts matter for downstream comparisons. Contrastive objectives provide this additional supervision by requiring the student to distinguish the matching teacher representation from representations of other texts (Tian et al., 2020). Hard-negative mining selects these alternatives from the anchor's neighborhood in the teacher's

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.

embedding space, exposing the student to comparisons among similar texts (Xiong et al., 2021). In our experiments, adding such comparisons improves retrieval over alignment-only distillation. Their usefulness, however, depends on whether the selected candidates are valid negatives.

In particular, proximity in the teacher's embedding space does not establish that two texts should be separated. On corpora containing repeated or paraphrased content, the nearest candidates to an anchor may express the same meaning. Treating these candidates as negatives introduces conflicting supervision: the contrastive objective pushes apart texts that should remain close. This false-negative problem is well established in dense retrieval, where approaches such as RocketQA use an auxiliary cross-encoder to denoise mined candidates (Qu et al., 2021). For distillation from an unlabeled corpus, a useful alternative would identify probable false negatives using the frozen teacher itself. A deterministic embedding provides a measure of proximity, but does not indicate how that proximity compares with the teacher's own variation when representing the same input.

We propose *Ballpark Mining*, a method that uses teacher perturbations to select negatives for contrastive embedding distillation. For each anchor, repeated forward passes with dropout active produce a collection of representations whose per-coordinate variation defines an ellipsoidal region, or *ballpark* (Figure 1). We use this region as an empirical criterion for identifying candidates whose proximity to the anchor is comparable to the teacher's variation under perturbation. Candidates inside the region are treated as probable false negatives and excluded; nearby candidates outside it are selected for contrastive training. The regions and negative sets are computed before student training, which then uses cosine alignment and an InfoNCE objective (van den Oord et al., 2018). Region construction and candidate filtering require neither similarity labels nor an auxiliary scoring model, and introduce no additional computation at student inference time.

We evaluate Ballpark Mining on deduplicated Wikipedia text and duplicate-rich Quora questions, using semantic similarity and retrieval tasks from MTEB and BEIR (Muennighoff et al., 2023; Thakur et al., 2021). We distill a 110M-parameter teacher into an 11M-parameter student and evaluate additional configurations with teachers up to 27B parameters. For 110M-to-11M distillation on Quora, Ballpark Mining reduces the fraction of selected negatives that are duplicates of the anchor from 21.9% with standard hard-negative mining to 2.3%, and increases duplicate recall@10 from 0.919 to 0.973. On deduplicated Wikipedia, the final-layer filter performs comparably to standard hard-negative mining. For this teacher–student pair, constructing regions at an intermediate teacher layer further improves the similarity–retrieval composite by approximately 3% on Quora and 6% on Wikipedia relative to final-layer regions. Across all four tested Quora configurations, spanning compression ratios from 3:1 to 2450:1, Ballpark Mining improves duplicate retrieval and the similarity–retrieval composite over standard hard-negative mining. For a 27B teacher distilled into an 11M student, filtering reduces duplicate-negative contamination from 21.8% to 0.2% and increases duplicate recall@10 from 0.741 to 0.942. Together, these results support teacher perturbations as a useful criterion for negative selection, while showing that the benefit depends on the corpus and the representation layer. Choosing a region size that transfers across corpora remains an important challenge. Our contributions are summarized as follows:

1. **Teacher-derived negative selection.** We introduce Ballpark Mining, which uses perturbation regions from a frozen teacher to filter probable false negatives before contrastive student training, without similarity labels or an auxiliary scorer (Section 2.2).
2. **Evaluation across corpora and model scales.** We show that Ballpark Mining improves duplicate recall and the similarity–retrieval composite over standard hard-negative mining on duplicate-rich Quora, while the final-layer filter performs comparably on deduplicated Wikipedia. The evaluation covers teachers from 110M to 27B parameters and compression ratios from 3:1 to 2450:1, with 110M-to-11M study as well as additional scale configuration (Sections 4.2, 4.3, and 4.5).
3. **Regions from intermediate teacher representations.** We show that constructing regions at an intermediate layer further improves distillation, reaching composite gains of approximately 15% on Quora and 22% on Wikipedia over the respective alignment-only controls. These findings use five seeds on Quora and two seeds on Wikipedia (Section 4.4).
4. **A geometric analysis of comparison stability.** We derive an exact bound on how ellipsoidal perturbations can change a fixed linear score difference, yielding a sufficient margin condition for preserving its ordering. The result characterizes comparisons within the specified regions; it does not guarantee the behavior of the trained student (Section 3).

1.1 Related work

Hard-negative mining and false negatives. Hard-negative mining strengthens contrastive training by selecting examples that are close under the current representation. ANCE retrieves negatives from an embedding index to train dense retrievers (Xiong et al., 2021), while RocketQA uses a cross-encoder to remove likely false negatives from mined candidates (Qu et al., 2021). These approaches illustrate both the value of nearby candidates and the difficulty of distinguishing informative negatives from relevant or semantically equivalent texts. Ballpark Mining addresses this selection problem in embedding distillation from an unlabeled corpus. Its filtering criterion comes from repeated perturbations of the frozen teacher, allowing probable false negatives to be identified without training or querying an auxiliary relevance model.

Representation and relational distillation. Knowledge distillation transfers information from a larger teacher to a smaller student (Hinton et al., 2015). Feature-based approaches include matching intermediate representations, as in FitNets (Romero et al., 2015), and transferring information through a variational objective, as in VID (Ahn et al., 2019). Other methods supervise relationships among representations: relational knowledge distillation matches pairwise distances and angles (Park et al., 2019), MiniLM transfers attention relations (Wang et al., 2020), and contrastive representation distillation learns from matched and mismatched pairs (Tian et al., 2020). For text embeddings, LEAF demonstrates that alignment with teacher representations can support effective compression (Vujanic & Rückstieß, 2026). These results motivate retaining alignment-only distillation as a baseline and evaluating the additional value of contrastive supervision. Ballpark Mining operates at the selection of examples for that supervision, and can be used with an alignment-plus-contrastive objective. Our experiments compare implemented feature and relational losses under a shared training setup; they are not reproductions of the complete published systems (Appendix C.3).

Teacher perturbations and uncertainty in distillation. Repeated stochastic forward passes have established uses in uncertainty estimation (Gal & Ghahramani, 2016). Related distillation methods use teacher variability to modify supervision: AvatarKD uses teacher-perturbation variance to scale feature errors (Zhang et al., 2023), and UET incorporates uncertainty estimated through repeated dropout passes (Yi et al., 2024). Ballpark Mining uses perturbation statistics to construct an exclusion region for negative selection before training. The measured variation describes the teacher's response to the chosen perturbation; it is not assumed to be a calibrated measure of semantic uncertainty. Whether this variation identifies probable false negatives is therefore evaluated empirically through duplicate contamination, duplicate recall, and downstream similarity and retrieval performance.

Preserving comparisons under compression. Embedding compression also requires understanding how representation errors affect downstream comparisons. Relational distillation explicitly preserves distances and angles (Park et al., 2019), while V_kD uses an orthogonal projection to preserve similarities (Miles et al., 2024). LEAF studies the tolerance of text comparisons to nonzero embedding error (Vujanic & Rückstieß, 2026), and Cui and Pei analyze the propagation of teacher variation to students (Cui & Pei, 2026). Our analysis characterizes the largest change in a fixed linear score difference permitted by two ellipsoidal regions. This yields a margin condition for preserving the ordering throughout those regions and clarifies the additional assumptions needed to transfer such a result across representation maps (Section 3).

2 Method

2.1 Setup and base loss

Let $\{x_i\}_{i=1}^N$ be an unlabeled corpus and $t_i \in \mathbb{R}^D$ the sentence embedding of x_i under a frozen pretrained teacher, computed with dropout (Srivastava et al., 2014) disabled. The student encoder f_θ mean-pools its final-layer token states and maps them through a learned affine layer to $p_i \in \mathbb{R}^D$. Let hats denote L_2-normalized vectors, $\widehat{p}_i = p_i/\|p_i\|_2$. For a minibatch $B \subset \{1, \dots, N\}$ and a set $\mathcal{N}_i$ of eight negative indices per anchor i, training minimizes

$$L = 0.7\, L_{\cos} + 0.2\, L_{\mathrm{HN}}, \tag{1}$$

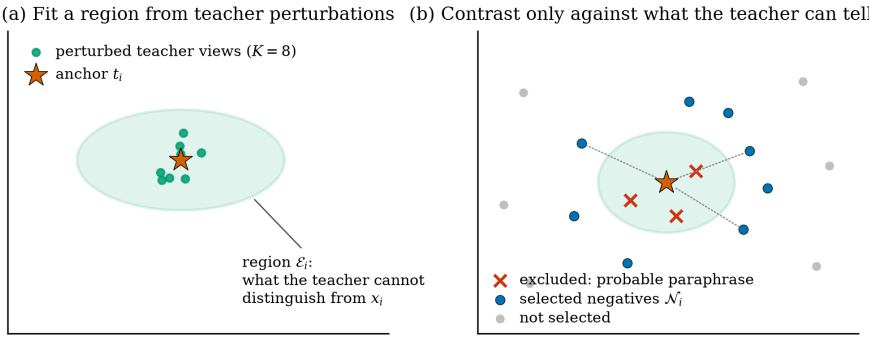


Figure 1: **Ballpark Mining.** (a) Passing input x_i through the frozen teacher $K{=}8$ times with dropout active yields perturbed views of its embedding; their spread defines a per-anchor ellipsoid $\mathcal{E}_i$, the *ballpark*: everything the teacher's own output wanders over when the input does not change. (b) Negative candidates inside $\mathcal{E}_i$ are texts the teacher cannot reliably distinguish from the anchor; they are excluded as probable paraphrases, and the nearest survivors become the InfoNCE negatives $\mathcal{N}_i$. Training uses a plain alignment+InfoNCE loss; the regions act only before training, in the miner.

where

$$L_{\text{cos}} = \tfrac{1}{|B|} \textstyle\sum_{i \in B} \big(1 - \widehat{p}_i^{\top} \widehat{t}_i\big),$$
$$L_{\text{HN}} = -\tfrac{1}{|B|} \textstyle\sum_{i \in B} \log \dfrac{\exp(\widehat{p}_i^{\top} \widehat{t}_i / \tau)}{\exp(\widehat{p}_i^{\top} \widehat{t}_i / \tau) + \sum_{j \in \mathcal{N}_i} \exp(\widehat{p}_i^{\top} \widehat{t}_j / \tau)}, \tag{2}$$

with temperature $\tau = 0.05$. The first term aligns each student vector with its teacher target; the second is InfoNCE (van den Oord et al., 2018), contrasting the target against the teacher's embeddings of the negatives. We call Eq. (1) the *base loss*. In the following we study how to construct $\mathcal{N}_i$ only with the teacher model.

2.2 Ballpark Mining

The teacher's nearest neighbors are the most informative negatives but also the most likely paraphrases. Without labels they are hard to separate. In many use cases labeling duplicates can be difficult without sufficient expert review. However, when a teacher model is known to already perform well on a task we can use the teacher knowledge to perform this task. Usually that is done via prompting or similarity matching which can pollute or require labels which we do not have. Our motivation was to use the model's self-confidence to identify what it can, and cannot distinguish.

On a corpus with paraphrases, a close candidate to "How do I lose weight fast?" is a question such as "What is the fastest way to lose weight?". However, an InfoNCE term that lists it as a negative asks the student to separate two texts the teacher places together. Thus, deciding which neighbors are contrastive negatives and which are paraphrases is a similarity judgment. An unlabeled setting lacks the information to perform this judgment.

The core idea of ballpark mining is that the teacher can make this judgment itself, because a teacher run with dropout active reports its own resolution. Passing one input through the frozen teacher several times with dropout active yields a cloud of embeddings, and the spread of that cloud is the scale of differences the teacher does not resolve at that point in its embedding space. A candidate that falls within the spread is, as far as the teacher can tell, the same text, see Corollary 1. Since the student learns only from the teacher's outputs, listing such a candidate as a negative asks it to separate two texts its own supervision places together.

Ballpark Mining fits a per-anchor ellipsoid to the perturbed embeddings and uses it to identify probable false negatives. Candidates inside the region are excluded, and the nearest remaining candidates under the region distance are selected as negatives. The ellipsoid's extent along each coordinate is determined by the teacher's perturbation variance for that anchor. The regions are used to select negatives before training; the student is then trained with the alignment and contrastive losses, without an additional region-based loss term.

We remark that we choose dropout as the noise source throughout, but any stochastic element of the teacher may be used as the sampling kernel.

Constructing the ballpark Ballpark mining builds the negative set $\mathcal{N}_i$ of each anchor text in three steps: the negatives are chosen relative to a region, the region is an ellipsoid with a calibrated radius, and the ellipsoid's metric is read off the teacher's dropout noise.

Let candidate negatives $\mathcal{C}_i$ be the C texts whose teacher embeddings, computed with dropout off, have the highest cosine similarity to t_i. Let $\mathcal{E}_i \subset \mathbb{R}^D$ be a region around the anchor with distance function d_i^2. The negatives are the M candidates outside the region that lie closest to the anchor under that distance,

$$\mathcal{N}_i = \operatorname*{arg\,min}_{S \subset \{j \in \mathcal{C}_i:\, t_j \notin \mathcal{E}_i\},\ |S|=M} \sum_{j \in S} d_i^2(t_j). \tag{3}$$

We use $C = 128$ candidates (excluding the anchor) and $M = 8$ negatives. If fewer than M candidates survive, we fill the remaining slots with excluded candidates in decreasing order of d_i^2; the number reintroduced is the *backfill count*. The region is the ellipsoid $\mathcal{E}_i = \{z : d_i^2(z) \le q_i\}$ whose squared radius q_i is calibrated on the teacher's own embedding of the anchor text. Let $t_i^{(1)}, \dots, t_i^{(K)}$ be K passes of x_i through the frozen teacher with active dropout, and let $s_i^{(k)} = d_{i,-k}^2\big(t_i^{(k)}\big)$ be the distance of view k under the metric fitted to the other $K-1$ views. Then

$$q_i = \alpha \cdot \mathrm{Q}_{0.95}\big\{s_i^{(1)}, \dots, s_i^{(K)}\big\}, \qquad \alpha \ge 1, \tag{4}$$

so that at $\alpha{=}1$ a held-out view of the anchor lands inside $\mathcal{E}_i$ 95% of the time. The multiplier α is the only tunable quantity.

The distance d_i^2 measures displacement from the anchor in units of the teacher's output noise, direction by direction. With μ_i and v_i the per-coordinate mean and variance of the K views, the raw variance at small K is a noisy estimate, so we regularize it toward the corpus mean $\bar{v} = \frac{1}{N}\sum_i v_i$ and bound it away from zero:

$$\tilde{v}_i = \max\big(\tfrac{1}{2}(v_i + \bar{v}),\ \epsilon\big) \quad \text{coordinatewise}, \qquad d_i^2(z) = (z - \mu_i)^\top \operatorname{diag}(\tilde{v}_i)^{-1}(z - \mu_i), \tag{5}$$

with $\epsilon > 0$. The leave-one-out metric $d_{i,-k}^2$ in Eq. (4) is the same construction from the $K-1$ views other than k.

2.3 Computational cost

Let N be the corpus size, K the number of perturbed passes, C the candidate-list length, M the number of negatives, D the embedding dimension, and $F_T(\ell)$ the cost of one teacher forward pass on a sentence of length ℓ. Ballpark mining is a one-time preprocessing step; its cost splits into teacher passes, candidate retrieval, and the table build.

Teacher passes. One deterministic and K perturbed passes over the corpus cost $O\big((K{+}1)\, N\, F_T(\ell)\big)$. The views are consumed as running sums, so only μ_i and $\tilde{v}_i$ are stored, $O(ND)$ memory in total.
Candidate retrieval. Exact top-C retrieval by cosine costs $O(N^2 D)$ time as a blocked matrix product and the lists occupy $O(NC)$ memory.
Table build. Region statistics and leave-one-out self-distances cost $O(NKD)$; scoring and filtering the candidates costs $O(NCD)$, and selecting the M nearest survivors $O(NC)$ with a partial sort.

The total is $O\big((K{+}1)\, N\, F_T(\ell) + N^2 D + N(K{+}C)D\big)$, and in practice the teacher passes dominate, since a forward pass through a Transformer costs far more than D operations per sentence. The miner therefore scales linearly in N at fixed K, with the retrieval step the only superlinear term and the one with standard sublinear substitutes.

3 Analysis: when is a contrast safe?

In retrieval, semantic text similarity, and linear classification tasks, an embedding z enters only through a score $w^\top z$ against some fixed vector w. Suppose the embeddings of two texts are known only up to their ballparks. When can we still be sure which of the two scores higher? Theorem 1 gives the margin condition; Corollary 1 identifies the candidates for which no linear score can satisfy it.

Theorem 1 (Preservation of score order). *Let $z_i \in \mathcal{E}_i(q_i)$ and $z_j \in \mathcal{E}_j(q_j)$, where $\mathcal{E}_i(q_i) = \{z : (z-\mu_i)^\top V_i^{-1}(z-\mu_i) \le q_i\}$ with $V_i \succ 0$. For any fixed w, the ordering $w^\top z_i > w^\top z_j$ holds whenever*

$$w^\top(\mu_i - \mu_j) \;>\; \sqrt{q_i\, w^\top V_i w} + \sqrt{q_j\, w^\top V_j w}. \tag{6}$$

Further, the right-hand side is the exact largest decrease in the score difference that the two regions permit.

The proof is a two-sided Cauchy–Schwarz argument, see Appendix A for a sketch and formal proof.

Corollary 1 (Point candidate). *Fix an anchor i with region $\mathcal{E}_i(q_i)$ and a candidate embedding t_j. Some linear score w satisfies the margin condition* (6) *with $q_j = 0$, i.e. orders t_j against every $z_i \in \mathcal{E}_i(q_i)$, if and only if $d_i^2(t_j) > q_i$. Equivalently, $t_j \in \mathcal{E}_i(q_i)$ if and only if for every w some admissible view $z_i \in \mathcal{E}_i(q_i)$ reverses or ties the order.*

We refer to Appendix A for a formal proof.

In ballpark mining, only the anchor has a region, whereas a candidate j is a single point with deterministic embedding t_j, which is the theorem's setting with $q_j = 0$. The corollary concludes that a score w ranking t_j against the anchor in its ballpark exists exactly when t_j lies outside the ballpark. Excluding the candidates inside the ballpark therefore removes exactly those that no score can reliably tell apart from the anchor.

4 Experiments

We evaluate how negative selection affects embedding distillation on corpora with different levels of duplication. We first test whether Ballpark Mining improves distillation on duplicate-rich Quora (Section 4.2) and compare its effect with standard hard-negative mining on deduplicated Wikipedia (Section 4.3). We then study the choice of teacher layer for constructing the regions (Section 4.4) and whether the findings hold across student sizes and teacher models (Section 4.5).

4.1 Experimental setup

We use two corpora of 200K texts each: deduplicated Wikipedia sentences and Quora questions. The corpora differ in the prevalence of duplicates, allowing us to study how this affects negative selection. Wikipedia contains few duplicates, whereas 74.8% of the Quora questions belong to a labeled duplicate cluster. We therefore use Wikipedia to assess the benefit of mined negatives when false negatives are uncommon, and Quora to measure their effect on training and the benefit of removing them. Duplicate labels are used only for evaluation and diagnostics. Details of corpus construction are provided in Appendix C.

We use frozen BGE and Qwen teachers (Xiao et al., 2024; Zhang et al., 2025), with the main comparisons based on distilling BGE-base-en-v1.5 (110M) into an 11.24M-parameter student. The student is randomly initialized and has four layers with a hidden dimension of 256. All methods in these comparisons follow the same training setup: 30,000 steps, a batch size of 64, and checkpoints at 2,000, 10,000, and 30,000 steps. We train each method with five random seeds, sharing the cached data, student initialization, and minibatch order within each seed so that the comparisons are paired.

Under this setup, we compare four methods that differ only in the InfoNCE term of Eq. (1). ALIGN omits this term and serves as the alignment-only baseline, which prior work has shown can support effective embedding compression (Vujanic & Rückstieß, 2026). The remaining methods each use eight negatives per anchor: RANDOM samples them uniformly, HARD selects the nearest neighbors by teacher cosine similarity, and BALLPARK (α) applies the region-based selection described in Section 2.2. These comparisons measure the benefit of adding contrastive supervision and the effect of negative selection. To compare against other forms of distillation supervision, we also replace the InfoNCE term with coordinate mean-squared error, Gaussian regression, or a relational feature loss. We refer to these baselines as KD-MSE, KD-Gauss, and KD-RKD, respectively; the names identify the implemented losses rather than reproductions of the complete published systems (Appendix C.3).

We evaluate the resulting students on semantic similarity and retrieval tasks, using Spearman correlation for STS-B and nDCG@10 for SciFact and NFCorpus (Thakur et al., 2021; Muennighoff

Table 1: **Duplicate-rich Quora: utility at 30k steps and duplicate diagnostics (mean ± sd).** Slot contamination is the fraction of negative slots holding a labeled duplicate of the anchor; Recall@10 (2,000 held-out duplicate-cluster anchors) and Pair AUC (2,000 positive, 2,000 negative held-out cosine pairs) are the duplicate diagnostics. Baselines first; BALLPARK last, the final-space radius sweep (Section 4.2) above the layer-10 regions (Section 4.4).

	Slot	Utility at 30k steps				Duplicate diagnostics	
Training method	contam. (%)	STS-B	SciFact	NFCorpus	Composite	Recall@10	Pair AUC
Baselines							
ALIGN	—	0.753±0.004	0.207±0.008	0.101±0.004	0.354±0.004	0.953±0.003	**0.775±0.001**
RANDOM	0.0	0.743±0.003	0.213±0.006	0.095±0.008	0.350±0.003	0.952±0.002	0.767±0.002
HARD (unfiltered NN)	21.9	0.727±0.004	0.250±0.012	0.105±0.004	0.361±0.005	0.919±0.004	0.718±0.002
KD-MSE	—	0.769±0.005	0.232±0.014	0.109±0.003	0.370±0.006	0.961±0.002	0.773±0.001
KD-Gauss	—	0.769±0.003	0.273±0.017	0.117±0.005	0.386±0.007	0.956±0.002	0.754±0.001
KD-RKD	—	0.750±0.004	0.165±0.016	0.087±0.002	0.334±0.006	0.952±0.004	0.774±0.001
Ours							
BALLPARK, α=1	19.1	0.729±0.003	0.263±0.016	0.112±0.010	0.368±0.007	0.924±0.004	0.722±0.002
BALLPARK, α=2	9.3	0.744±0.003	0.269±0.018	0.122±0.007	0.378±0.006	0.951±0.001	0.733±0.002
BALLPARK, α=3	2.3	0.764±0.002	0.291±0.014	0.129±0.005	0.395±0.007	**0.973±0.001**	0.749±0.001
BALLPARK, layer 10, α=1	17.3	0.764±0.001	0.297±0.009	0.140±0.003	0.400±0.002	0.968±0.001	—
BALLPARK, layer 10, α=2	8.7	**0.773±0.002**	0.300±0.016	**0.145±0.005**	0.406±0.005	0.971±0.001	—
BALLPARK, layer 10, α=3	1.5	0.771±0.003	**0.309±0.013**	0.143±0.005	**0.408±0.006**	0.971±0.002	—

et al., 2023). Our primary comparison metric, the *utility composite*, is the unweighted mean of these three task scores. We report a method as outperforming another only when the improvement holds across all five paired seeds; otherwise, we state how many seeds favor each method. Results on four additional BEIR datasets are provided in Appendix B.6.

4.2 DUPLICATE-RICH QUORA: FILTERING REPAIRS CONTRASTIVE DISTILLATION

Effect of standard hard-negative mining. Table 1 shows that 21.9% of the negatives selected by HARD are labeled duplicates of the anchor. Although HARD achieves a slightly higher mean utility composite than ALIGN (0.361 vs. 0.354), duplicate Recall@10 decreases from 0.953 to 0.919 and Pair AUC from 0.775 to 0.718. The reduction in duplicate recall persists through the end of training (Figure 3, Appendix B.1), consistent with the contrastive objective encouraging the student to separate duplicate questions.

Effect of region size. Ballpark Mining reduces the number of duplicate negatives as the region size increases. The fraction decreases from 19.1% at α=1 to 2.3% at α=3, where Recall@10 reaches 0.973 and exceeds the alignment-only baseline. The utility composite also increases to 0.395, the highest value among the final-layer configurations. Although α=3 was confirmed using duplicate labels, a post-hoc check identifies the same value without labels. The backfill count remains near zero until the region excludes most candidates, leaving too few negatives to select (Figure 2a). Choosing the largest α with fewer than one backfilled negative per anchor on average selects α=3 on Quora and α=2 on Wikipedia. This rule has only been checked on these two corpora.

Comparison with distillation baselines. Compared with the other distillation losses, BALLPARK at α=3 outperforms KD-MSE and KD-RKD on the utility composite. Its mean score also exceeds KD-Gauss, although the improvement holds in only three of five paired seeds and KD-Gauss performs better on STS-B. Ballpark Mining achieves higher duplicate Recall@10 than all three baselines. Across six BEIR retrieval tasks, it also achieves the highest mean score among the evaluated students, 0.161 ± 0.004 (Appendix B.6).

Limits of duplicate filtering. The improvement in duplicate recall does not fully extend to Pair AUC, which remains below ALIGN (0.749 vs. 0.775). The judge audit also identifies some excluded candidates as related texts rather than paraphrases: 76.0% of the sampled exclusions on Quora are judged to be paraphrases, compared with 97.5% for the conservative Wikipedia filter.

4.3 DISTILLATION ON DEDUPLICATED WIKIPEDIA

In this data, it allows us to assess the benefit of hard-negative mining and the effect of Ballpark Mining when duplicates are uncommon (Table 3, Appendix B.2). At 30,000 steps, mined negatives outperform alignment-only distillation and random negatives. Standard hard-negative mining increases the utility composite from 0.328 with alignment alone to 0.387, with gains concentrated in retrieval: SciFact increases from 0.167 to 0.308, while STS-B decreases by 0.035. These gains depend on training duration, as random negatives achieve higher mean utility at 2,000 steps before mined negatives perform better by 30,000 steps (Figure 4). Ballpark Mining excludes candidates for only 6.2% of anchors, requires no backfilling, and performs comparably to standard hard-negative mining, with a mean utility difference of -0.003. An independent judge identifies 97.5% of the sampled exclusions as genuine paraphrases. These results indicate that Ballpark Mining identifies probable false negatives on Wikipedia, but removing them has limited benefit when duplicates are uncommon.

4.4 REGIONS AT INTERMEDIATE TEACHER LAYERS

The regions require only perturbation statistics, which are available at every teacher layer, including layers where a cosine threshold cannot be set directly by hand. This allows us to vary the teacher layer independently of the radius: α controls how much the region excludes, while the layer determines the representation space in which proximity to the anchor is measured. We change only the region construction, keeping the candidate pool, loss, and training setup fixed.

Intermediate-layer regions on Wikipedia. We construct regions at teacher layers $\{2, 4, 6, 8, 10\}$ using the setup of Section 4.3 with two seeds. Every tested intermediate layer outperforms alignment-only distillation and has higher mean utility than the final-layer construction. Utility increases with depth, reaching 0.403 at layer 10 compared with 0.381 for the final layer (Table 4, Appendix B.3). The intermediate-layer regions exclude candidates for a similar proportion of anchors (6.5–7.7%, compared with 6.2% for the final layer), but select different negatives, with a median overlap of two out of eight negatives with standard hard-negative mining.

Intermediate-layer regions on Quora. With five seeds (Table 1, bottom block), layer-10 regions outperform final-layer regions at every matched radius in every seed: 0.400 vs. 0.368 at α=1, 0.406 vs. 0.378 at α=2, and 0.408 vs. 0.395 at α=3. The last configuration achieves the highest utility composite among the 11M students in this study. Changing the layer improves utility at each radius, while increasing the radius reduces duplicate-negative contamination from 17.3% to 8.7% and then to 1.5%. At α=3, Recall@10 reaches 0.971, compared with 0.953 for alignment-only distillation, and the label-free backfill rule again selects this radius, with 0.14 backfills per anchor. Together, the layer and radius choices yield a 15% improvement over alignment-only distillation on Quora and a 13% improvement over standard hard-negative mining. On Wikipedia, intermediate-layer regions improve utility by 22% over alignment-only distillation.

4.5 SCALE: ACROSS MODEL SIZES AND COMPRESSION RATIOS

Hard-negative mining on deduplicated Wikipedia. We vary student size, teacher family, and compression ratio under the same training setup on both corpora (Table 2). In the scale experiments, mined negatives outperform alignment-only distillation in every completed Wikipedia configuration and every seed, covering compression ratios from 3:1 to 2450:1 across three teacher families. The gains come mainly from retrieval, with SciFact roughly tripling at the three largest compression ratios (Table 5, Appendix B.4). BALLPARK also outperforms all three distillation baselines in the four configurations used for these comparisons, while alignment-only distillation falls below KD-Gauss in three of them (Table 6). The additional benefit over HARD remains small on this duplicate-poor corpus, with improvements of 0.006 in two configurations and comparable performance elsewhere.

Ballpark Mining on duplicate-rich Quora. HARD underperforms alignment-only distillation in three of four configurations, with a utility difference of -0.051 at 2450:1 compression, and reduces duplicate recall in all four. At α=3, BALLPARK outperforms HARD in every seed of every configuration, with utility gains from 0.024 to 0.074 (Figure 6, Appendix B.4). The largest gain occurs with the 27B teacher, where duplicate Recall@10 increases from 0.741 to 0.942 and the fraction of duplicate negatives decreases from 21.8% to 0.2%. Across all four configurations, duplicate recall recovers to at least the alignment-only level. The region size matters: at α=1, performance is

Table 2: **Full scale grid, utility composite at 30k steps on both corpora (report-only; two seeds per cell, A2 five seeds).** One frozen recipe throughout; on Quora, BALLPARK is the calibrated α=3 radius. "Warm" students start from a pretrained checkpoint, the two cold billion-parameter cells fail to train at the frozen 245M-token budget, and "—" marks a cell not run on that corpus.

				Wikipedia (dedup.)			Quora (dup.-rich)		
Cell	Teacher	Student	Ratio	ALIGN	HARD	BALLPARK	ALIGN	HARD	BALLPARK
Block 1: student size, teacher fixed (BGE 110M)									
A3	BGE 110M	37M (d512, L6)	3:1	0.462	0.463	0.469	0.439	0.437	0.461
A2	BGE 110M	11M (d256, L4)	10:1	0.328	0.387	0.384	0.354	0.361	0.395
A1	BGE 110M	4.7M (d128, L2)	24:1	0.222	0.265	0.268	0.241	0.234	0.268
Block 2: teacher family, student fixed (11M; the ladder reads A2→B1→C1)									
B1	Qwen3-Emb. 4B	11M (d256, L4)	364:1	0.258	0.324	0.321	—	—	—
C1	27B decoder	11M (d256, L4)	2450:1	0.234	0.264	0.261	0.262	0.211	0.285
Block 3: billion-parameter students (warm starts train, cold starts fail; Wikipedia only)									
B2W	Qwen3-Emb. 4B	Qwen2.5-0.5B (warm)	8:1	0.336	0.366	0.372	—	—	—
B2P	Qwen3-Emb. 4B	0.4B (d1280, L20)	10:1	0.197	0.051	—	—	—	—
C2W	27B decoder	Qwen2.5-3B (warm)	9:1	0.326	0.339	0.337	—	—	—
C2	27B decoder	2.7B (d2560, L34)	10:1	0.025	0.022	—	—	—	—

comparable to HARD outside the A2 configuration. At α=3, the improvement over HARD is larger on Quora than on Wikipedia in every matched configuration, consistent with the predeclared hypothesis.

Failed configurations and additional findings. The two large, randomly initialized students perform near chance under the fixed token budget, while the pretrained initializations specified as alternatives train successfully. With the student fixed at 11M parameters, utility decreases at each increase in teacher size, from 110M to 4B to 27B, for every method (Figure 5b). Thus, increasing teacher size does not improve the student under this training setup. Increasing the number of negatives to 64 per anchor adds approximately 0.03 to utility for both mining methods, indicating that both can improve beyond the results reported with eight negatives (Appendix B.5).

5 CONCLUSION

The effectiveness of contrastive embedding distillation depends on which examples the student is asked to separate. Our experiments show that hard-negative mining can improve retrieval while also teaching the student to separate semantically equivalent texts. This conflict persists across the tested student sizes and extends to distillation from a 27B teacher. Ballpark Mining uses the teacher's response to perturbations to identify probable false negatives before training. Its gains on duplicate-rich text support negative selection as an important part of distillation, while its comparable performance on deduplicated text shows where this additional selection has limited benefit. The intermediate-layer results further show that the teacher representation used to select negatives can affect the quality of the resulting student.

Several questions remain about the selection criterion. Cosine thresholds achieve similar duplicate filtering at matched exclusion rates, so the experiments do not establish that ellipsoidal geometry is necessary for the observed gains. Training comparisons against these thresholds and against corpus deduplication would help isolate the value of the perturbation-derived regions. Region size also remains sensitive to the corpus: the proposed label-free selection rule recovers the chosen radius on the two evaluated corpora, but has only been checked post hoc. Testing that rule on new corpora is the next step toward negative selection that can be applied without labeled calibration.

ETHICS STATEMENT

The experiments concern English sentence similarity and retrieval on public corpora with pretrained teachers. Compressed embeddings can inherit teacher biases and change rankings for underrepresented language varieties or subject matter; the reported benchmarks do not establish fitness for consequential decision-making. Student training uses no similarity labels; pretrained-teacher exposure to evaluation data cannot be ruled out.

AI USE STATEMENT

Generative AI tools assisted with literature search, manuscript drafting and editing, experiment implementation, mathematical derivations, and review of the experimental results. An LLM judge classified sampled text pairs as paraphrases, related texts, or unrelated texts when evaluating the excluded candidates. The reported numerical results were obtained from executed experiments or saved experimental records. The authors take responsibility for the final text, code, mathematical claims, and reported results.

REFERENCES

Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D. Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9163–9171, 2019.

Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *IEEE/CVF International Conference on Computer Vision*, 2019.

Ziyao Cui and Jian Pei. How is uncertainty propagated in knowledge distillation? *arXiv preprint arXiv:2601.18909*, 2026.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pp. 1050–1059, 2016.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *Conference on Empirical Methods in Natural Language Processing*, pp. 6894–6910, 2021.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Roy Miles, Ismail Elezi, and Jiankang Deng. V_kD: Improving knowledge distillation using orthogonal projections. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *AAAI Conference on Artificial Intelligence*, pp. 5191–5198, 2020.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In *Conference of the European Chapter of the Association for Computational Linguistics*, 2023.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3967–3976, 2019.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5835–5847, 2021.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing*, pp. 3982–3992, 2019.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2021.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2020.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. URL https://arxiv.org/abs/1807.03748.

Robin Vujanic and Thomas Rückstieß. LEAF: Knowledge distillation of text embedding models with teacher-aligned representations. In *Annual Meeting of the Association for Computational Linguistics*, pp. 43362–43383, 2026.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, 2020.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general Chinese embeddings. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*, 2021.

Junfei Yi, Jianxu Mao, Tengfei Liu, Mingjie Li, Hanyu Gu, Hui Zhang, Xiaojun Chang, and Yaonan Wang. Teaching with uncertainty: Unleashing the potential of knowledge distillation in object detection. *arXiv preprint arXiv:2406.06999*, 2024.

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. *arXiv preprint arXiv:2506.05176*, 2025.

Yuan Zhang, Weihua Chen, Yichen Lu, Tao Huang, Xiuyu Sun, and Jian Cao. Avatar knowledge distillation: Self-ensemble teacher paradigm with uncertainty. In *ACM International Conference on Multimedia*, pp. 5272–5280, 2023.

APPENDIX OVERVIEW

We provide the proofs in Appendix A. Appendix B presents additional results supporting Section 4, including the effects of region size and training duration, detailed Wikipedia results, comparisons across teacher layers and model sizes, comparisons with distillation baselines, the effect of the number of negatives, and results on six BEIR tasks. Appendix C describes the training configurations, corpora, and comparison losses.

A PROOFS

A.1 PROOF OF THEOREM 1

For any positive-definite V and $q > 0$, Cauchy–Schwarz gives $\sup_{e^\top V^{-1} e \le q} w^\top e = \sup_{\|u\| \le \sqrt{q}} (V^{1/2} w)^\top u = \sqrt{q\, w^\top V w}$, attained at $e = \sqrt{q}\, V w / \sqrt{w^\top V w}$ for $w \neq 0$. Minimizing the i contribution and maximizing the j contribution independently,

$$\inf_{z_i \in \mathcal{E}_i,\, z_j \in \mathcal{E}_j} w^\top (z_i - z_j) = w^\top (\mu_i - \mu_j) - \sqrt{q_i w^\top V_i w} - \sqrt{q_j w^\top V_j w},$$

so Eq. (6) guarantees a strictly positive margin, and the threshold is sharp for the stated sets (equality permits a tie).

A.2 PROOF OF COROLLARY 1

If $t_j \in \mathcal{E}_i(q_i)$, the admissible view $z_i = t_j$ ties every linear score, so no w can strictly order the candidate against every view. Otherwise, set $a = d_i^2(t_j) > q_i$ and $w = V_i^{-1}(\mu_i - t_j)$. Then $w^\top(\mu_i - t_j) = w^\top V_i w = a$, and the margin condition with candidate center $\mu_j = t_j$ and $q_j = 0$ becomes $a > \sqrt{q_i a}$, which holds. This also covers $q_i = 0$; boundary candidates with $a = q_i$ belong to the region and admit a tie, establishing both statements.

A.3 SEPARATE OUTPUT MAPS

Proposition 2 (Compatibility of output maps)**.** *Let $A \in \mathbb{R}^{r \times d}$ (full row rank) and $B \in \mathbb{R}^{D \times d}$ be the linear parts of a region map and a sentence output map on the same hidden state. A finite C with $\|B\delta\|_2 \le C\|V^{-1/2} A\delta\|_2$ for all δ exists iff $\ker A \subseteq \ker B$, and the smallest such constant is $C = \|BA^\dagger V^{1/2}\|_2$.*

Proof. If the bound holds, any $\delta \in \ker A$ zeroes its right side, forcing $B\delta = 0$. Conversely, $\ker A \subseteq \ker B$ gives $B\delta = BA^\dagger A\delta = BA^\dagger V^{1/2}(V^{-1/2} A\delta)$, so the operator norm is a valid constant; choosing $\delta = A^\dagger V^{1/2} u$ for a maximizing right singular vector u (using $AA^\dagger = I_r$) attains it. □

If the kernel inclusion fails, there is a direction v with $Av = 0$, $Bv \neq 0$: moving the hidden state along v leaves the region prediction fixed while changing the pooled output without bound. Our implemented architecture trains the two maps independently and does not enforce the inclusion; this is why Theorem 1 is stated for the supervised representation only.

A.4 GLOBAL VERSUS PER-ANCHOR VARIANCES

Proposition 3 (Approximation by a global variance)**.** *Let $V_i = \mathrm{diag}(v_i)$ and $\bar{V} = \mathrm{diag}(\bar{v})$ have positive entries with $(1-\delta)\bar{v}_j \le v_{ij} \le (1+\delta)\bar{v}_j$ for all j and some $0 \le \delta < 1$. Then $\bar{d}_i^2/(1+\delta) \le d_i^2 \le \bar{d}_i^2/(1-\delta)$ and $\bar{\mathcal{E}}_i((1-\delta)q) \subseteq \mathcal{E}_i(q) \subseteq \bar{\mathcal{E}}_i((1+\delta)q)$, with constants attained at the endpoints.*

Proof. Take reciprocals of the coordinate bounds, multiply by $(z_j - \mu_{ij})^2$, and sum; the inclusions follow directly. □

This motivates the miner's shrinkage toward the corpus-mean variance vector. We checked whether this condition holds for the cached teacher variances: on a fixed cache of teacher perturbation variances we measured ratios $v_{ij}/\bar{v}_j$ spanning $[0.05,\ 11.23]$, so no uniform $\delta < 1$ holds there and the proposition supplies a condition to test, not an established explanation.

B FURTHER PROBES AND SCALE-GRID DETAILS (REPORT-ONLY)

The probes in this appendix run under the frozen recipe (30k steps; the utility composite of Section 4), were predeclared before execution, and share caches and per-seed pairing with the main grids. Two-seed cells support directional consistency, not per-cell significance; we report every cell, including the failures.

B.1 QUORA PROBES: DOSE-RESPONSE AND TRAINING-HORIZON CURVES

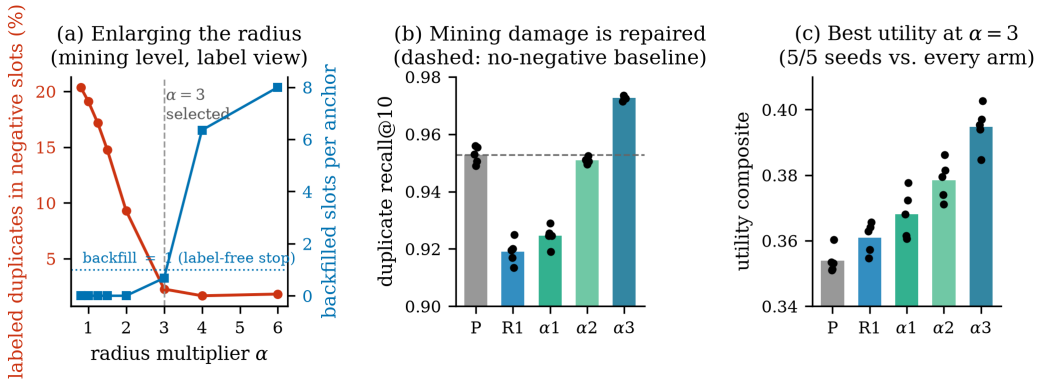


Figure 2: **Radius dose-response on Quora.** (a) Mining level: enlarging α removes labeled duplicates monotonically, while the backfill count, observable *without* labels, stays near zero until the region excludes most candidates; "largest α with backfill < 1" picks the same α=3 the labeled sweep chose. (b, c) Trained outcomes: HARD damages Recall@10; the enlarged radius repairs it above ALIGN and delivers the best Composite.

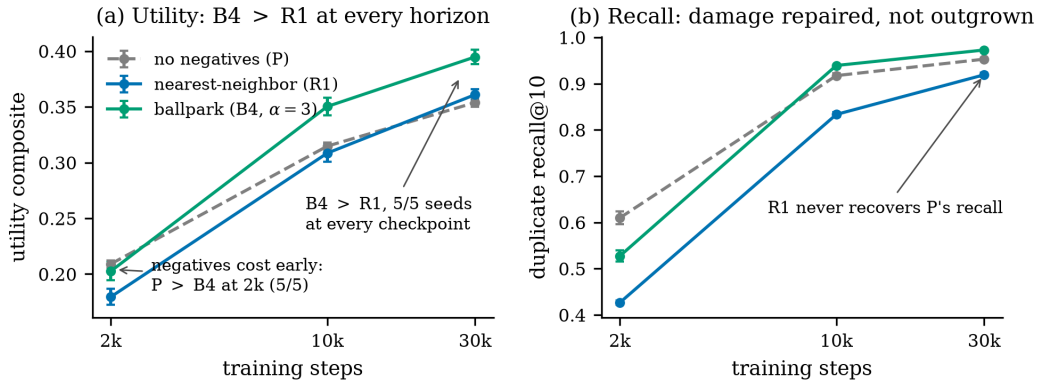


Figure 3: **How BALLPARK separates from HARD over the training horizon (Quora, α=3).** Five seeds per method; error bars one standard deviation. (a) BALLPARK at α=3 exceeds HARD's utility in 5/5 paired seeds at every checkpoint; against the no-negative reference it has lower utility than alignment-only distillation at 2,000 steps (ALIGN wins at 2k in 5/5 seeds) and overtakes by 10k (5/5). (b) The duplicate-recall damage of unfiltered mining is not a transient: HARD stays below ALIGN at every horizon, while the filtered method recovers and finishes above ALIGN.

Figure 2 plots the Quora radius dose-response behind Section 4.2; Figure 3 shows that the duplicate-recall damage of unfiltered mining persists to the end of training, while the filtered method leads it in 5/5 paired seeds at every checkpoint.

B.2 The Wikipedia grid: per-task numbers and training horizon

Table 3: **Deduplicated Wikipedia (200K sentences).** Utility composite at three checkpoints and per-task scores at 30k, mean ± sd. Mined negatives beat ALIGN and RANDOM at 30k; BALLPARK ties HARD (mean −0.003) because few anchors have candidates excluded. The final block adds the best-performing recipe measured on this same training setup: BALLPARK with the region fitted at teacher layer 10 (Section 4.4, Table 4).

	Utility composite			Per-task at 30k steps		
Method	2k	10k	30k	STS-B	SciFact	NFCorpus
ALIGN	0.150±0.006	0.272±0.008	0.328±0.009	**0.737±0.006**	0.167±0.018	0.078±0.005
RANDOM	**0.175±0.004**	0.273±0.005	0.332±0.004	0.715±0.005	0.189±0.004	0.093±0.004
HARD (unfiltered NN)	0.169±0.007	0.320±0.005	**0.387±0.003**	0.702±0.006	**0.308±0.010**	0.151±0.009
BALLPARK (filtered NN)	0.174±0.005	**0.322±0.005**	0.384±0.004	0.700±0.008	0.299±0.011	**0.152±0.008**
BALLPARK, teacher layer 10†	*0.186*±0.005	*0.347*±0.005	***0.403***±0.005	0.733±0.004	*0.314*±0.006	*0.163*±0.014

†Inner-layer region variant (two seeds), the best result measured on this training setup; *italic* marks values that exceed the five-seed methods above.

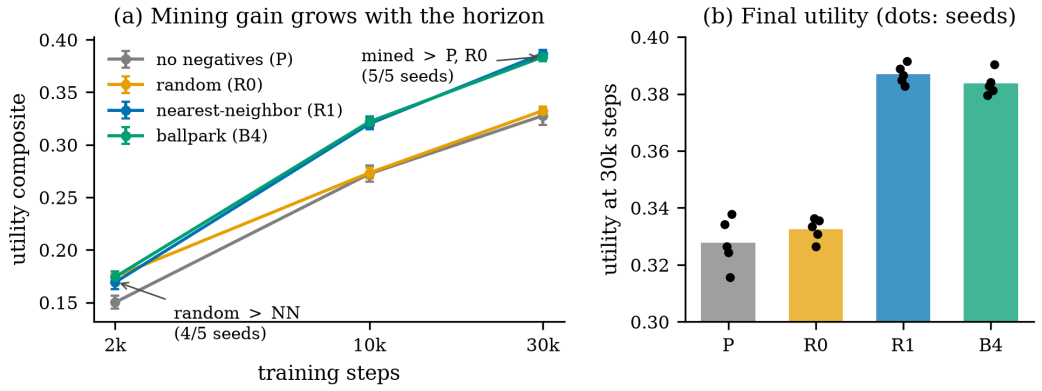


Figure 4: **The mining gain is horizon-dependent (deduplicated Wikipedia).** (a) Utility vs. training steps (mean ± sd): random negatives win at 2k steps (four of five seeds); the ordering reverses by 30k. (b) Final utilities with per-seed values; the two mined methods are inseparable on this corpus.

Table 3 records the five-seed Wikipedia grid behind Section 4.3, and Figure 4 its horizon dependence: random negatives win at 2k steps, the ordering reverses by 30k, and the two mined methods finish inseparable on this deduplicated corpus. The filter's 6.2% firing rate corresponds to 0.35 exclusions per anchor among the top-128 candidates.

B.3 Ellipsoids from inner teacher layers

Table 4 records the per-task numbers behind the depth sweep of Section 4.4; each inner layer filters with essentially no backfill. Per-layer truncated-teacher outputs were verified to match the full teacher exactly (maximum absolute error 0.0) before caching; the same zero-error check was repeated on Quora before the five-seed layer-10 radius sweep of Table 1.

B.4 Scale-grid details

The composite grid is Table 2 (Section 4.5); Table 5 records its per-task decomposition and Table 6 the multi-scale comparison with KD baselines. The predeclared teacher-scale claim fails outright: with

Table 4: **Per-task numbers behind the depth sweep, deduplicated Wikipedia (two seeds, 30k steps).** At an inner layer we run the same perturbation passes through the teacher but read the per-coordinate variance at the mean-pooled hidden state of layer n instead of the final embedding, and use *that* region to decide exclusions; the candidate pool stays in the final space, so only the geometry that defines "too close to the anchor" changes. Every inner layer beats ALIGN and, descriptively, the final-space filter; utility peaks at layer 10.

Region space	Filter rate (%)	STS-B	SciFact	NFCorpus	Utility
ALIGN (no filter, no negatives)	—	**0.742**	0.173	0.075	0.330
HARD (no filter, unfiltered NN)	—	0.707	0.311	0.141	0.387
BALLPARK, final space	6.2	0.705	0.295	0.144	0.381
BALLPARK, teacher layer 2	6.5	0.709	0.298	0.146	0.384
BALLPARK, teacher layer 4	7.1	0.716	0.306	0.151	0.391
BALLPARK, teacher layer 6	7.7	0.725	0.311	0.156	0.397
BALLPARK, teacher layer 8	7.6	0.725	**0.316**	0.161	0.401
BALLPARK, teacher layer 10	7.0	0.733	0.314	**0.163**	**0.403**

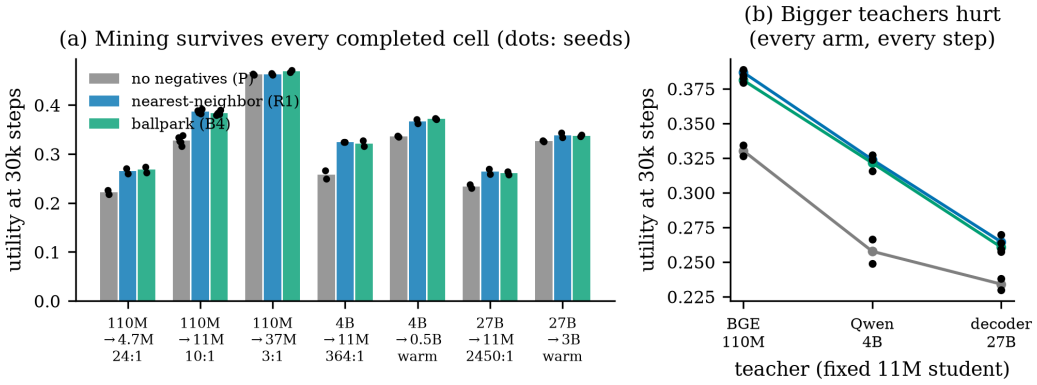


Figure 5: **The completed Wikipedia scale grid.** (a) Utility at 30k steps for the seven completed cells (dots: seeds): mined negatives beat alignment-only distillation in every seed of every cell. (b) The teacher-scale ladder at a fixed 11M student: utility falls at every step for every method; the predeclared monotonicity claim fails.

the 11M student fixed, every ladder step is negative in every seed for ALIGN, HARD, and BALLPARK alike (A2→B1: -0.052 to -0.078; B1→C1: -0.019 to -0.065; Figure 5b); nothing in this paper supports the claim that utility rises with teacher size. In the comparison with distillation baselines, HARD also beats every baseline in 2/2 seeds in every cell except a 1/2 split against KD-MSE at 3:1, and KD-Gauss remains the strongest baseline, consistent with the duplicate-rich comparison with distillation baselines (Section 4.2). At the extreme-compression cells the KD baselines' SciFact scores sit at 0.06–0.09, near-random retrieval, so their composites are carried almost entirely by STS-B.

B.5 HOW MANY NEGATIVES

The frozen recipe fixes eight negatives per anchor ($|\mathcal{N}_i|=8$); this probe rebuilds the negative tables at 2, 16, and 64 on the Wikipedia training setup (two seeds), after verifying that the frozen eight-negative table rebuilds bit-identically. Utility rises monotonically for filtered and unfiltered mining alike: from 0.349/0.348 (HARD/BALLPARK) at two negatives to 0.414/0.413 at 64, about $+0.03$ over the frozen recipe for each. Two consequences: the headline mining numbers in this paper are conservative, and the improvement is a property of the InfoNCE negative count, not of the ballpark filter: the two methods track each other at every count.

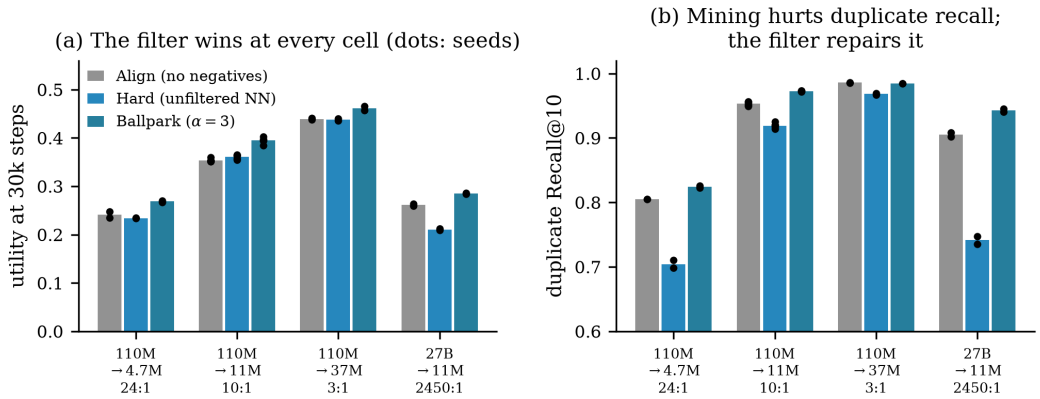


Figure 6: **Effect of Ballpark Mining across model sizes on Quora.** (a) Utility at 30k steps for the four completed cells (dots: seeds): BALLPARK (α=3) beats both baselines in every seed of every cell, while unfiltered HARD mining falls to or below alignment-only distillation. (b) Duplicate Recall@10 for the same methods: mining depresses duplicate retrieval at every scale and the calibrated radius restores it to the alignment-only level.

Table 5: **Per-task scores for the scale grid at 30k steps.** Two-seed means (single seed for the failed cells; †A2 is the five-seed grid of Table 3); STS-B Spearman, SciFact and NFCorpus nDCG@10, and the composite. The mining gain is carried by the retrieval tasks at every scale (SciFact roughly triples at 24:1, 364:1, and 2450:1) while STS-B moves little; only the warm students (B2W, C2W) improve STS-B as well.

		ALIGN (no negatives)				HARD (unfiltered)				BALLPARK (filtered)			
Cell	Ratio	STS	Sci	NFC	Comp.	STS	Sci	NFC	Comp.	STS	Sci	NFC	Comp.
A1	24:1	0.590	0.049	0.027	0.222	0.591	0.145	0.059	0.265	0.595	0.145	0.064	0.268
A2†	10:1	0.737	0.167	0.078	0.328	0.702	0.308	0.151	0.387	0.700	0.299	0.152	0.384
A3	3:1	0.815	0.374	0.198	0.462	0.791	0.382	0.217	0.463	0.790	0.398	0.218	0.469
B1	364:1	0.618	0.118	0.037	0.258	0.644	0.234	0.094	0.324	0.641	0.232	0.091	0.321
B2W	8:1	0.768	0.161	0.079	0.336	0.794	0.203	0.101	0.366	0.797	0.217	0.101	0.372
B2P	10:1	0.524	0.045	0.023	0.197	0.142	0.000	0.009	0.051			—	
C1	2450:1	0.635	0.049	0.018	0.234	0.597	0.146	0.050	0.264	0.590	0.147	0.045	0.261
C2	10:1	0.059	0.001	0.016	0.025	0.051	0.005	0.008	0.022			—	
C2W	9:1	0.791	0.123	0.065	0.326	0.793	0.145	0.077	0.339	0.795	0.137	0.079	0.337

B.6 FOUR FURTHER BEIR DATASETS

To test whether the two in-protocol retrieval tasks paint a representative picture, we scored every frozen five-seed final checkpoint of the headline methods (both grids, both α settings, all three KD baselines) plus the teacher on ArguAna, FiQA, SciDocs, and TREC-COVID under a predeclared, eval-only protocol (artifacts in Appendix **??**); no model was retrained or reselected, no gates were predeclared, and differences within about one standard deviation are not treated as wins. On the Wikipedia grid (Table 7), the mined methods' retrieval advantage is not an artifact of the two in-protocol datasets: HARD and BALLPARK also lead on FiQA and SciDocs, remain tied with each other everywhere (0.159 vs. 0.158 BEIR-6), and achieve lower scores on ArguAna, where the no-negative baseline is best. On the Quora grid (Table 8), the damage done by unfiltered mining is visible *out of domain* as well: HARD's BEIR-6 mean (0.138) falls to the no-negative baseline (0.140), while the filtered method at the confirmed operating point repairs the losses and posts the best student BEIR-6 of the study (0.161 ± 0.004); its edge over KD-Gauss (0.158 ± 0.004) is within overlapping standard deviations, so we note only that the composite's ordering carries over. The teacher's BEIR-6 mean (0.497) is roughly three times any student's: at this compression, trained on 200K sentences, out-of-domain retrieval remains far from teacher quality.

Table 6: **Multi-scale KD baseline comparison (report-only).** Mean utility difference from each baseline at 30k steps (paired-seed wins in parentheses), two seeds per cell. BALLPARK beats every distillation baseline in 2/2 seeds in all four completed cells; alignment-only distillation loses to KD-Gauss at three of them.

Cell	Baseline	Baseline mean	BALLPARK − base	HARD − base	ALIGN − base
A1 (24:1)	KD-Gauss	0.255	+0.013 (2/2)	+0.010 (2/2)	−0.033 (0/2)
	KD-MSE	0.248	+0.020 (2/2)	+0.017 (2/2)	−0.026 (0/2)
	KD-RKD	0.237	+0.031 (2/2)	+0.028 (2/2)	−0.015 (0/2)
A3 (3:1)	KD-Gauss	0.457	+0.012 (2/2)	+0.006 (2/2)	+0.005 (2/2)
	KD-MSE	0.461	+0.008 (2/2)	+0.002 (1/2)	+0.002 (1/2)
	KD-RKD	0.449	+0.020 (2/2)	+0.014 (2/2)	+0.014 (2/2)
B1 (364:1)	KD-Gauss	0.296	+0.026 (2/2)	+0.028 (2/2)	−0.038 (0/2)
	KD-MSE	0.239	+0.083 (2/2)	+0.085 (2/2)	+0.019 (2/2)
	KD-RKD	0.255	+0.067 (2/2)	+0.069 (2/2)	+0.003 (1/2)
C1 (2450:1)	KD-Gauss	0.243	+0.018 (2/2)	+0.022 (2/2)	−0.009 (0/2)
	KD-MSE	0.189	+0.072 (2/2)	+0.076 (2/2)	+0.045 (2/2)
	KD-RKD	0.229	+0.032 (2/2)	+0.036 (2/2)	+0.005 (2/2)

Table 7: **Six-task BEIR readout, Wikipedia grid (nDCG@10, mean ± sd over the five frozen final checkpoints).** Report-only: SciFact and NFCorpus are the in-protocol tasks; ArguAna, FiQA, SciDocs, and TREC-COVID were scored once on the frozen checkpoints under a predeclared eval-only protocol (BEIR ignore-identical-ids convention). BEIR-6 is the per-seed unweighted mean of all six. Best student per column in bold; teacher reference in the last row.

Method	SciFact	NFCorpus	ArguAna	FiQA	SciDocs	TREC-COVID	BEIR-6
ALIGN	0.167±0.018	0.078±0.005	**0.208±0.007**	0.038±0.001	0.035±0.002	0.189±0.009	0.119±0.005
RANDOM	0.189±0.004	0.093±0.004	0.207±0.009	0.039±0.003	0.036±0.002	0.195±0.015	0.126±0.003
HARD	**0.308±0.010**	0.151±0.009	0.198±0.005	**0.047±0.004**	**0.057±0.003**	0.191±0.011	**0.159±0.004**
BALLPARK, α=1	0.299±0.011	**0.152±0.008**	0.198±0.005	0.046±0.002	**0.057±0.003**	**0.197±0.017**	0.158±0.006
Teacher (109M)	0.730	0.375	0.589	0.363	0.227	0.700	0.497

Table 8: **Six-task BEIR readout, Quora grid (nDCG@10, mean ± sd over the five frozen final checkpoints).** Same protocol as Table 7. BeIR/quora itself is excluded because its text overlaps the training corpus; CQADupStack was excluded for cost in the predeclaration.

Method	SciFact	NFCorpus	ArguAna	FiQA	SciDocs	TREC-COVID	BEIR-6
ALIGN	0.207±0.008	0.101±0.004	0.199±0.008	0.072±0.003	0.053±0.002	0.210±0.022	0.140±0.003
RANDOM	0.213±0.006	0.095±0.008	0.204±0.009	0.067±0.006	0.053±0.004	0.212±0.021	0.141±0.004
HARD	0.250±0.012	0.105±0.004	0.189±0.010	0.073±0.001	0.052±0.004	0.158±0.009	0.138±0.004
KD-MSE	0.232±0.014	0.109±0.003	0.192±0.009	0.075±0.003	0.060±0.004	**0.233±0.023**	0.150±0.005
KD-Gauss	0.273±0.017	0.117±0.005	0.204±0.002	0.073±0.004	**0.062±0.003**	0.216±0.009	0.158±0.004
KD-RKD	0.165±0.016	0.087±0.002	0.168±0.006	0.067±0.002	0.048±0.003	0.190±0.026	0.121±0.006
BALLPARK, α=1	0.263±0.016	0.112±0.010	0.188±0.009	0.072±0.003	0.053±0.006	0.167±0.022	0.142±0.006
BALLPARK, α=2	0.269±0.018	0.122±0.007	0.190±0.009	0.076±0.003	0.057±0.005	0.180±0.022	0.149±0.006
BALLPARK, α=3	**0.291±0.014**	**0.129±0.005**	**0.205±0.005**	**0.080±0.004**	**0.062±0.005**	0.202±0.010	**0.161±0.004**
Teacher (109M)	0.730	0.375	0.589	0.363	0.227	0.700	0.497

C CONFIGURATIONS AND COMPARISON METHODS

C.1 TRAINING CONFIGURATION (ALL TRANSFORMER STUDIES)

Students use trainable WordPiece embeddings (vocabulary 30,522), learned positions, pre-norm Transformer blocks, GELU, feed-forward width $4\times$, dropout 0.1, mean pooling over non-padding tokens, and an affine output map to the teacher dimension; all are randomly initialized. The optimizer is AdamW (Loshchilov & Hutter, 2017) at initial learning rate $3 \cdot 10^{-4}$, weight decay 0.01, gradient

Algorithm 1 Ballpark mining and distillation

Require: Corpus $\{x_i\}$, frozen teacher, student parameters Θ, passes K, candidates C, radius multiplier $\alpha \geq 1$, steps T

1: Cache deterministic teacher embeddings t_i (normalized, mean-pooled).
2: Cache μ_i, v_i from K perturbed teacher passes; shrink and floor the variances; set each squared radius from the leave-one-out self-distances, scaled by α.
3: Retrieve top-C candidates per anchor; exclude those inside the region; set $\mathcal{N}_i$ to the eight nearest survivors (backfill if fewer, and count).
4: **for** $s = 0, \ldots, T-1$ **do**
5: Sample a minibatch B; compute student embeddings p_i.
6: Update Θ by AdamW on $L = 0.7L_{\cos} + 0.2L_{\mathrm{HN}}$ [Eq. (1)], gradient-clipped, cosine-decayed.
7: **end for**
8: **return** Student encoder and output map. Regions are used only by the miner.

clipping at norm 1, and cosine decay; minibatches have 64 examples. The mining-study recipe trains 30,000 steps. Within a comparison, all methods share the teacher cache, negative-candidate cache, per-seed initialization, and minibatch order.

C.2 Corpora

Wikipedia: 200K English sentences from `wikimedia/wikipedia` (config `20231101.en`), paragraph-filtered, split on sentence boundaries, length-filtered to 20–512 characters, truncated to 128 WordPieces, shuffled with seed 0. *Quora:* 200K questions sampled from Quora Question Pairs so that 74.8% of rows belong to a labeled duplicate cluster (149,571 of 200,000); 2,000 held-out anchors for duplicate recall@10 and 2,000 positive plus 2,000 negative held-out pairs for cosine AUC; construction seeds and the corpus SHA-256 are recorded in the readout. Duplicate labels are used only for evaluation and for the mining-level contamination diagnostics; no training method sees them.

C.3 Comparison losses

Every table label names an implemented loss, not a reproduction of a published system. ***KD-MSE***: $\mathbb{E}\|p_i - t_i\|^2/D$. ***KD-Gauss***: Gaussian feature regression, $\frac{1}{2}\sum_j[(z_j - \mu_j)^2/v_j + \log v_j]$ with cached variances (VID-inspired; original VID learns them; Ahn et al., 2019). ***KD-RKD***: distance- and angle-matching relational loss (Park et al., 2019) on pooled embeddings. In the Quora comparison with distillation baselines, KD losses *replace* the InfoNCE term on top of the same alignment loss, so the comparison isolates "mined contrasts" against "more feature supervision" at matched budgets.