SEMI-CONTINUOUS REPRESENTATION LEARNING: MIXED-WIDTH CONTRASTIVE TRAINING FOR CALIBRATED MATRYOSHKA EMBEDDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Matryoshka Representation Learning (MRL) has become a widely adopted approach for preserving semantic structure across embedding dimensionalities. Its premise is that each Matryoshka prefix is a coarser representation of the same input within one common embedding space. However, while all nested embeddings share the same space, their similarity scores are miscalibrated across prefix sizes.

We demonstrate theoretically and experimentally that the miscalibration stems from MRL's loss formulation where each prefix is optimised under its own partition function. Therefore, we propose *Semi-Continuous Representation Learning (SCRL)*, which replaces MRL's decomposed loss with a single contrastive objective over randomly sampled prefix sizes under a shared softmax. We prove that SCRL results in calibrated prefixes and show empirically that it ensures consistent similarity scores across nested dimensions, improving mixed-size retrieval performance by up to 8.8 nDCG@10 points on average across eight BEIR benchmarks. Furthermore, SCRL improves cross-size neighborhood consistency and increases the score distribution overlap between 64- and 768-dimensional embeddings from 71.7% to 84.9%.

SCRL can be applied as a direct training replacement for MRL, as a post-training step for existing MRL embedding models, and in MRL-adaptor settings on top of any general-purpose embeddings.

1 INTRODUCTION

Matryoshka Representation Learning (MRL) (Kusupati et al., 2022) has been established as a proven practice for creating nested embeddings of differing size in a shared space. The MRL training objective preserves global semantic structure at differing levels of granularity, enabling to *interchangeably* use prefixes of varying size without sacrificing top-line performance.

This versatility has resulted in its application to a range of domains such as retrieval (Yoon et al., 2024; Günther et al., 2025; Zhang et al., 2025a; Wang et al., 2025), recommendations (Wang et al., 2024; Lai et al., 2026), and various text classification tasks (Li et al., 2024; Phung et al., 2026). The MRL paradigm has furthermore emerged as a standard feature in general-purpose LLM-based embedding models (Lee et al., 2024; Hu et al., 2025; Zhang et al., 2025b; Lee et al., 2025; Li et al., 2026), enabling any client applications to *pick & choose* a task-appropriate embedding size.

However, in our work we find that the decomposed loss formulation in MRL (Kusupati et al., 2022) places no constraint on the relative scale of similarity scores across prefix sizes, and in practice the resulting distributions differ substantially. This undermines the promise of *interchangeability* of MRL prefixes across nesting levels, since any vector-based similarity scores are not guaranteed to follow a common scale. Concretely, this means that a similarity score of 0.8 in 128 dimensions *can mean something very different* from a score of 0.8 in 1024 dimensions, and that these scores cannot safely be compared directly.

These findings have direct and impactful practical consequences. For example, many modern retrieval pipelines leverage MRL-trained embeddings to use smaller prefixes for first-stage retrieval, and consecutively larger prefixes for later-stage ranking and re-ranking layers. That means that any

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.

ranker operates in a different query-item similarity space than the retriever, independent of whether the ranker consumes a query-item similarity score, or an MRL embedding of a different size, as feature input.

To address this shortcoming in MRL training, we propose *Semi-Continuous Representation Learning (SCRL)*, a method that replaces MRL's decomposed sum-of-softmax loss terms, with a single contrastive objective over randomly sampled prefix sizes under a unified softmax expression. This penalises differences in score scale between sizes, placing nested embeddings of *any* size on a common similarity scale. Figure 1 visualises the difference between MRL and *SCRL*, where MRL uses a

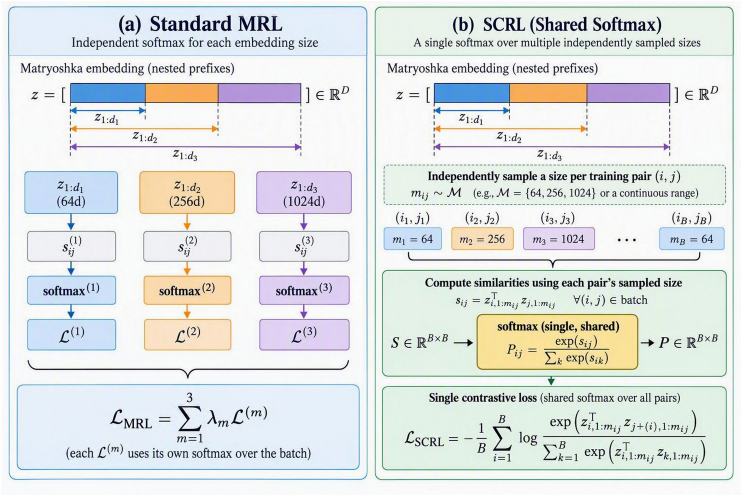


Figure 1: In standard MRL training, the final loss function is decomposed as a weighted sum where each prefix size is scored independently of all other prefixes. The objective therefore places no constraint on the similarity scales of nested representations. In *SCRL*, prefix sizes are sampled independently per query-item pair and subsequently combined under a single shared softmax. This penalises cross-size score offsets, so that nested embeddings share a similarity scale.

dedicated loss term for each prefix and later combines them in a weighted sum, *SCRL* independently samples a prefix size for each query-item pair in a batch, and combines their contribution under a shared loss term across all prefix sizes. This makes cross-size offsets identifiable under the loss, which the MRL objective cannot achieve, and penalises score drift between sizes during training. By directly addressing this miscalibration during training, SCRL allows a truncated index to reliably reproduce full-size rankings, bypassing the need to score all candidates at uniform widths.

Our introduction of sampling prefix sizes during training has a second impactful effect: it is no longer necessary to choose explicit prefix sizes, and their relative weighting in the loss, upfront. Instead, *SCRL* enables the use of *any* prefix size on the embeddings, ultimately creating a viable Matryoshka cutoff point at every index. Lastly, *SCRL* can be used flexibly either i) as a full replacement for the standard MRL training objective, ii) as a post-processing calibration method on top of existing MRL embeddings, and iii) in MRL-adaptor (Yoon et al., 2024) settings on top of any general purpose LLM embeddings which we explore in Appendix F.

Our contributions in this paper are thus as follows:

- We prove that the MRL objective is invariant to per-size offsets and show that standard MRL training leaves embeddings uncalibrated and propose *SCRL* as a remedy, proving that its objective identifies the cross-size offsets, which the MRL objective cannot.
- We empirically validate our theoretical findings and show that SCRL calibrates embeddings on a range of retrieval tasks across training, post-processing, and adaptor scenarios. In mixed-size retrieval, SCRL outperforms standard MRL by up to 8.8 nDCG@10 points on average, nearly matching the performance of a post-hoc per-width affine correction without requiring any extra fitting. Beyond score calibration, SCRL improves the underlying embedding geometry compared to post-hoc corrections. This results in superior quality at narrow prefixes, monotonic performance scaling, and higher cross-size neighbourhood

consistency, such as an 8.2% absolute increase in top-1 agreement between 32-dimensional prefixes and full-size embeddings.
- We extend Matryoshka training to a continuous range of cutoffs, with cross-size offsets identified at every cutoff in that range.

2 Related work

Matryoshka Representation Learning. Kusupati et al. (2022) introduced the concept of learning nested representations in a shared space and popularised the idea of the *funnel approach* by consecutively applying larger prefixes to capture more nuanced aspects of the embedded items. The versatility of MRL has led to a rich body of research to improve the learnt embeddings, by e.g. extending the Matryoshka principle to *model depth* in addition to *embedding width* (Li et al., 2024; Wang et al., 2025; Zhuang et al., 2025). Our work is complementary to that line of research by introducing the concept of completely flexible width, instead of pre-defined cutoff points.

Another line of research has focused on improving MRL via a deeper understanding of the learning dynamics across nested representations. For example, Talukder et al. (2026) study how full-prefix MRL breaks rotational invariance and produces a task-aligned ordering of nested representations. Zhang et al. (2025a) introduce sequential learning, adaptive dimension selection and cross-batch memory to improve the training of MRL embeddings. Phung et al. (2026) propose the use of self-distillation for improving token geometry across prefixes, and similarly Anonymous (2026), present a recipe for preserving global subspace geometry for distilling MRL embeddings. Our work is most closely related to this branch of MRL research by contributing the insight that the similarity space is incompatible across nested prefixes, and proposing *SCRL* as a solution to mitigate this issue.

Ordered and dynamic-size representations. The idea of ordering representation dimensions by importance predates Matryoshka Representation Learning. For example, Rippel et al. (2014) introduce nested dropout, which stochastically drops all dimensions after a sampled index during training, enforcing an ordering of the embedding dimensions. Cui et al. (2023) extends this idea by learning the nested dropout distribution. Yu et al. (2019); Yu & Huang (2019) propose *slimmable* networks that enable arbitrary runtime widths of a model during inference. Similarly, Devvrit et al. (2024) propose a dynamic-size transformer architecture that samples the width for a feedforward-block dynamically during training. Our work is methodologically related to this body of research as in *SCRL*, we independently sample a prefix-width for each query-item pair during training, thereby creating arbitrary cutoff points for the learnt representations.

3 Method and setup

Let $\mathcal{D}$ be a finite document corpus and $\mathcal{Q}$ be the query space. Consider a query $q \in \mathcal{Q}$, a positive document $d^+ \in \mathcal{D}$ and a negative document $d^- \in \mathcal{D}$. In standard MRL, the contrastive objective evaluates these vectors one dimensionality at a time. The objective only requires whether d^+ is assigned a higher cosine similarity than d^- where both vectors are represented using the same size. It never evaluates whether a relevant document stored at 768 dimensions scores higher than an irrelevant one stored at 32 dimensions. SCRL addresses this by assigning each candidate document a vector size drawn independently from a distribution and scoring all documents at their respective sizes within a single shared softmax.

Prefix scoring. An encoder maps queries and documents from the document corpus $\mathcal{D}$ and query space $\mathcal{Q}$ to full-size representations $u = f(q)$ and $v = g(d)$ in $\mathbb{R}^W$. For a size $w \in \mathcal{M} \subseteq \{1, \dots, W\}$, we let $x_{[:w]}$ denote the prefix truncation of x to its first w entries. The score of a document d stored at size w is $s_w(q, d) = \langle \bar{u}_{[:w]}, \bar{v}_{[:w]} \rangle$,where $\bar{x}_{[:w]} = x_{[:w]} / \|x_{[:w]}\|_2$ denotes the unit-normalized prefix vector.

3.1 Objective: cross-size contrastive optimization

Consider a batch of B candidate documents $\mathcal{B} = \{d_1, \dots, d_B\}$ for a query q, containing a single relevant target d_y. Let π be a distribution with full support over the admissible sizes $\mathcal{M}$.

Homogeneous partition (MRL). Adapting the Matryoshka framework (Kusupati et al., 2022) to contrastive retrieval, the objective applies one InfoNCE softmax (van den Oord et al., 2018) with temperature $\tau > 0$ per size of a predetermined ladder and sums them, $\mathcal{L}_{\mathrm{MRL}} = \sum_{w \in \mathcal{M}} \lambda_w \mathcal{L}_w$ with weights $\lambda_w \geq 0$, where $\mathcal{L}_w$ scores every candidate at size w. Each partition function is therefore size-homogeneous.

Heterogeneous partition (SCRL). SCRL instead samples a size $w_i \sim \pi$ independently for each candidate document, allowing representations of different sizes to compete within the same softmax,

$$\mathcal{L}_{\mathrm{SCRL}} = \mathbb{E}_{q,\mathcal{B},\{w_i\}}\left[-\log \frac{\exp\big(s_{w_y}(q,d_y)/\tau\big)}{\sum_{i=1}^{B}\exp\big(s_{w_i}(q,d_i)/\tau\big)}\right], \qquad w_1,\ldots,w_B \overset{\text{i.i.d.}}{\sim} \pi, \tag{1}$$

where the size w_y of the positive document is also sampled from π, which Section 4.3 shows is necessary.

The entire difference between the two methods lies in the selection of w. It is constant within each softmax in MRL, and varies within the softmax in equation 1. The ladder and the weights λ_w are thus replaced by a single choice of π, which can support a standard ladder such as $\{32, 64, 128, 256\}$ or on any subset of an arbitrary index range $\{l, \ldots, h\}$. This is the sense in which the method is semi-continuous. A single loss term can in principle supervise every vector size. Additionally, the dimensionality-sampling step may be applied more than once per batch to compute multiple loss terms, which trades supervision against compute. Unlike nested dropout (Rippel et al., 2014), which also truncates at random sizes, SCRL places candidates of different sizes in one softmax. As we do not ablate the two, we attribute the representation-level effects to SCRL as a whole.

3.2 Cross-size offset and calibration

The two objectives differ fundamentally in their sensitivity to score offsets across sizes. To illustrate this, consider augmenting the scoring function with free, size-dependent scalars b_w for $w \in \mathcal{M}$. The original scoring function $s_w(q,d)$ would then be replaced by $s_w(q,d) + b_w$.

Definition 1 (Offset invariance and cross-size calibration)**.** An objective $\mathcal{L}$ is *offset-invariant* if $\mathcal{L}(s+b) = \mathcal{L}(s)$ for every vector $b \in \mathbb{R}^{|\mathcal{M}|}$, where the shifted score function is defined as $(s+b)_w(q,d) := s_w(q,d) + b_w$. In a single softmax over candidates with prefix sizes all drawn i.i.d. from π, independently of the candidates, let P_w be the total softmax mass of candidates with prefix size w. Scores are defined to be cross-size calibrated if

$$\mathbb{E}[P_w] = \pi(w) \qquad \text{for every } w \in \mathcal{M}, \tag{2}$$

and the calibration gap is measured as $\Delta := \max_{w \in \mathcal{M}} \big|\mathbb{E}[P_w]/\pi(w) - 1\big|$.

With a homogeneous partition function as in MRL, the b_w terms cancel out in the numerator and denominator. As a consequence, the loss is unable to detect or penalize size-dependent score differences, leaving cross-size score shifts entirely unregularized. In contrast, SCRL's heterogeneous partition function couples candidates of varying sizes within the same denominator, explicitly penalizing baseline score drift across sizes. This identifies the offsets, and for a fixed encoder the minimiser has zero calibration gap (Theorem 10).

4 Theoretical Analysis

This section formalizes the effect of different partition functions, the resulting score discrepancy, and cross-size calibration. The analysis focuses on three main questions: (i) why contrastive MRL leaves cross-size score offsets unconstrained (Section 4.1); (ii) what score discrepancy naturally arises in the absence of supervision (Section 4.2); and (iii) how SCRL's shared heterogeneous softmax identifies the offsets and characterises its calibrated minimiser (Section 4.3). This section was motivated by the observation that similarity scores diverge sharply across MRL prefix sizes. We show why this happens, give an upper bound on how large the gap can get, and confirm the effect in a simulation in Appendix B.

4.1 Why Contrastive MRL Leaves Score Offsets Unconstrained

We begin by examining how the standard MRL cross-entropy loss optimizes each prefix size $w \in \mathcal{M}$ and why this partition method fails to penalize cross-size baseline differences. Specifically, we evaluate the sensitivity of the objective to the size-specific additive offsets $b = (b_w)_{w\in\mathcal{M}}$ introduced in Section 3.2, beginning with the elementary shift invariance of the softmax.

Lemma 2 (Shift invariance). *For logits $z \in \mathbb{R}^n$ and any $a \in \mathbb{R}$, $\mathrm{softmax}(z + a\mathbf{1}) = \mathrm{softmax}(z)$. Consequently, any cross-entropy objective is invariant under a shift of all logits within a softmax by a common constant.*

Each MRL term scores all candidates at the same size w, so b_w shifts every logit in that softmax equally. Applying Lemma 2 independently to each size immediately yields the following invariance.

Proposition 3 (Offset invariance of the Matryoshka cross-entropy objective). *Let $\mathcal{L}_{\mathrm{MRL}}(\theta, b)$ denote the Matryoshka objective computed with the augmented scores $s_w(q, d) + b_w$ of Definition 1. Then for every $b \in \mathbb{R}^{|\mathcal{M}|}$,*

$$\mathcal{L}_{\mathrm{MRL}}(\theta, b) = \mathcal{L}_{\mathrm{MRL}}(\theta, 0), \qquad \textit{hence} \qquad \nabla_b \mathcal{L}_{\mathrm{MRL}} \equiv 0, \quad \nabla_b^2 \mathcal{L}_{\mathrm{MRL}} \equiv 0. \tag{3}$$

The offsets lie in the null space of the objective, leaving the calibration gap Δ unconstrained. If $(\theta^*, 0)$ is a minimizer, then so is (θ^*, b) for every $b \in \mathbb{R}^{|\mathcal{M}|}$. The result is that no quantity of data, MRL hyperparameters, and no optimizer schedule creates pressure toward calibration.

Remark 4. The nested structure does couple the prefixes since they share coordinates, and Proposition 3 does not deny this. It only states that the objective is indifferent to the offsets b, so any cross-size alignment pressure comes from the architecture rather than the loss.

4.2 The Magnitude and Direction of the Calibration Gap

Proposition 3 establishes that any contrastive MRL loss leaves cross-size score offsets unconstrained in the objective. It does not establish that the resulting gap is large or practically harmful. We now show that without explicit cross-size supervision, standard spherical geometry induces a large systematic score advantage for low-dimensional representations. Contrastive training naturally pushes representations to spread out uniformly over the hypersphere (Wang & Isola, 2020). We therefore model negative candidates as uniform random unit vectors to analyse their behaviour.

Lemma 5 (Negative similarity concentration). *Let $u, v \sim \mathrm{Unif}(S^{w-1})$ be independent. Then $\mathbb{E}[\langle u, v\rangle] = 0$ and $\mathrm{Var}(\langle u, v\rangle) = 1/w$. Moreover, $\langle u, v\rangle$ is strictly sub-Gaussian with optimal variance proxy $\sigma^2 = 1/w$ (Marchal & Arbel, 2017):*

$$\mathbb{E}[\exp(t\langle u, v\rangle)] \le \exp\left(\frac{t^2}{2w}\right), \quad \forall t \in \mathbb{R}, \tag{4}$$

and hence, by the Chernoff bound, $\mathbb{P}(|\langle u, v\rangle| \ge \epsilon) \le 2\exp(-w\epsilon^2/2)$.

Because the variance is $1/w$, the standard deviation (or typical spread) of these negative similarities scales as $\sqrt{1/w} = w^{-1/2}$. Consequently, the distribution of negative scores is $\sqrt{256/32} \approx 2.83$ times wider at $w = 32$ than at $w = 256$. Since positives must outscore the highest-ranked negatives, a wider spread raises the bar they must clear at small w. This is the size-dependent offset that Proposition 3 leaves unconstrained.

Lemma 6 (Top-negative floor). *Let $u \in S^{w-1}$ be fixed and $v_1, \ldots, v_N \sim \mathrm{Unif}(S^{w-1})$. With the variance proxy of Lemma 5, the expected largest negative similarity at size w satisfies*

$$\mu_{w,N} = \mathbb{E}\Big[\max_{i\le N}\langle u, v_i\rangle\Big] \le \sqrt{\frac{2\ln N}{w}} \tag{5}$$

The bound equation 5 decreases in w, but a decreasing upper bound does not by itself establish that the floor is higher at small sizes.

Lemma 7 (Monotonicity in prefix size). *Let $u \in S^{w'-1}$ be fixed, let $v_1, \ldots, v_N \sim \mathrm{Unif}(S^{w'-1})$ be independent, and for $w \le w'$ let $\mu_{w,N}$ denote the expected maximum similarity when all vectors are truncated to their first w coordinates and renormalized. Then for all $w < w'$ and every $N \ge 2$,*

$$\mu_{w,N} > \mu_{w',N}$$

Proof. See Appendix A.1. □

By Lemma 7, the expected best score achieved by a negative document is strictly larger when that document is stored with a smaller prefix size, and this advantage grows as the prefix gets smaller. Simulation confirms the ordering at $N = 1000$ and $W = 256$, where the gap between the smallest and largest prefix is $\mu_{32,N} - \mu_{256,N} = 0.337$ in cosine units (Appendix B).

Remark 8 (Not hubness or distance concentration)**.** The effect is distinct from hubness (Radovanović et al., 2010) and distance concentration (Beyer et al., 1999), which worsen as dimensionality grows within a single space. Ours grows as dimensionality shrinks, appears only when scores from different sizes are compared, and follows from the $1/w$ variance of Lemma 5.

4.3 Why a Shared Softmax Calibrates Scores Across Prefix Sizes

Section 4.2 shows that the offset gap is meaningfully large and monotone. While the MRL cross-entropy objective leaves the offset gap unconstrained in the objective, we show that a shared softmax over mixed prefix sizes supplies the missing curvature, and that, for a fixed encoder, its minimiser is calibrated.

Where the curvature comes from. For brevity in the following analysis, we absorb the temperature hyperparameter τ into the scoring function s_w (equivalently, assuming $\tau = 1$), as it acts only as a positive scalar on the variance and does not alter the location of the optima. Consider the augmented scores $s_{w_i}(q, d_i) + b_{w_i}$ inside a single softmax over a batch of B candidates with sizes $w_1, \ldots, w_B$. Also denote the resulting cross entropy with ℓ and set the index of the positive example to y. Let p denote the softmax distribution over the batch, and $\mathbf{1}_w$ for the indicator of size w, which equals 1 for candidates scored at size w and 0 otherwise. Finally, recall the total softmax mass on size w from Definition 1, which for this batch is $P_w = \sum_{i:w_i=w} p_i = \sum_i p_i \mathbf{1}\{w_i = w\}$. In this section, w and w' range independently over $\mathcal{M}$ and are not ordered like they are in Section 4.2.

Theorem 9 (Curvature identity)**.** *With the notation from above,*

$$\frac{\partial \ell}{\partial b_w} = P_w - \mathbf{1}\{w_y = w\}, \qquad \frac{\partial^2 \ell}{\partial b_w \, \partial b_{w'}} = \mathrm{Cov}_p(\mathbf{1}_w, \mathbf{1}_{w'}), \tag{6}$$

and for any direction $a \in \mathbb{R}^{|\mathcal{M}|}$ in the offset space,

$$a^\top \nabla_b^2 \ell \, a = \mathrm{Var}_p(a_{w_i}) \tag{7}$$

Proof. See Appendix A.2. □

The curvature therefore does not depend on the label y, it is entirely determined by the composition of sizes in the softmax of the batch. In MRL, every candidate has the same size, $\mathbf{1}_w$ is constant, so the variance and the Hessian vanishes. In SCRL, it is positive whenever the batch mixes sizes.

The offsets are identified. Curvature alone does not say where the minimiser sits, or whether there is only one. The following theorem locates it.

Theorem 10 (Identifiability of the cross-size offsets under SCRL)**.** *Assume the sizes $w_1, \ldots, w_B$ are drawn i.i.d. from π with full support on $\mathcal{M}$, independently of any characteristics of the candidates and with the same π governing both the positive and the negative examples, and that $B \geq 2$. In this scenario, every pair of distinct sizes in $\mathcal{M}$ co-occurs in a batch with positive probability. Since s_w is a cosine similarity, the base scores are bounded: $|s_w| \leq S$. Note that cosine similarity implies $S = 1$, but this constraint is not necessary.*

Regard $\mathcal{L}_{\mathrm{SCRL}}$ as a function of the offsets b alone, with θ fixed. Then: (i) $b \mapsto \mathcal{L}_{\mathrm{SCRL}}$ is convex on $\mathbb{R}^{|\mathcal{M}|}$, invariant along $\mathrm{span}(\mathbf{1})$, and strictly convex on the quotient $\mathbb{R}^{|\mathcal{M}|}/\mathrm{span}(\mathbf{1})$; (ii) it is coercive on that quotient, hence attains its minimum there; and (iii) the minimiser is unique modulo the addition of a common constant, and is characterised by the mass-matching condition

$$\mathbb{E}[P_w] = \pi(w) \qquad \text{for every } w \in \mathcal{M}. \tag{8}$$

Proof. See Appendix A.3. □

The invariance along $\mathrm{span}(\mathbf{1})$ is the shift from Lemma 2 and leaves every ranking unchanged. What SCRL identifies, that standard MRL does not, are the differences between the b_w. At the optimum for a fixed encoder, each size receives softmax mass equal to its sampling rate, so the calibration gap of Definition 1 is $\Delta = 0$. This relies on the positive's size also being drawn from π. Otherwise, the label term in equation 6 averages to the positives' size distribution instead.

Corollary 11 (Any cutoff)**.** *If π assigns positive probability to every integer prefix size in $[w_{\min}, W]$, then Theorem 10 applies with $\mathcal{M} = \{w_{\min}, \ldots, W\}$. Therefore, the offsets are identified at every cutoff in that range. Calibration does not require a predetermined set of prefix sizes.*

5 Experimental setup

Training Setup. Retrieval is the main setting of our experiments. Our main comparison trains MRL and SCRL from the same BGE-base-en-v1.5 backbone (Xiao et al., 2023) ($W = 768$) on a collection of queries drawn from the training sets of MS MARCO, NQ, HotpotQA, FEVER, FiQA and SciFact. We also run experiments using E5-base-v2 (Wang et al., 2022), and the Qwen3-Embedding-0.6B backbone (Zhang et al., 2025b). Note that five of the eight evaluation domains also contribute training data, so this is not a zero-shot evaluation. The comparison runs share every recorded setting except the objective, and the best checkpoints are selected for both on the same development criterion. Both methods use one encoder pass per batch where MRL averages six fixed-width cross-entropies and SCRL averages five mixed-width ones (Appendix G), so their training cost is essentially the same. More details about hyperparameters, data and models are given in Appendices C and G.

SCRL Sampling. Each training example contains a query, a positive example, and a number of negatives. Every candidate in the softmax receives a size drawn independently from the same distribution π. As required by Theorem 10, the prefix sizes of the positive example and the negatives are sampled in the exact same way. We use two specifications for π. A *continuous* setting, uniform over every integer size in $\{32, \ldots, W\}$, and a *discrete* setting, uniform over the standard Matryoshka ladder $\{32, 64, 128, 256, \ldots, W\}$. Both have full support on their respective $\mathcal{M}$, as Theorem 10 requires. Since π enters the calibration condition equation 8 only as the sampling rate each size is compared against, shifting π during training acts as a curriculum over the optimization path rather than a modification of the target, provided full support is retained throughout.

Evaluation. We evaluate on eight BEIR domains against their full corpora, using exact retrieval to depth 100. Three protocols are used. In *fixed-width* retrieval every document is truncated to the same size, and the full ladder is evaluated separately. Since scores at a single width are never compared against scores at another, this protocol cannot reveal miscalibration. In practice, we may want to use different sizes for different documents. For example, a 100M-document index can cut memory from about 300 GB to 54 GB by keeping the most-retrieved 10% of documents at 768 dimensions and truncating the rest to 64. Every query then ranks both groups in a single list, and if 64-dimensional scores are biased upwards, truncated documents crowd out the relevant full-size ones. We therefore add two *mixed-width* protocols, in which each document is randomly assigned a prefix size, deterministically from its identifier and an assignment seed and held fixed across queries, so that documents stored at different sizes compete against each other in one global ranking. Random assignment is a controlled stress test since it keeps size independent of relevance, so differences between models reflect both calibration and per-width retrieval quality. The *discrete* protocol samples sizes uniformly from the ladder and the *continuous* protocol from every integer in $[32, W]$.

Metrics. The main metric reported is nDCG@10. The primary aggregate is an unweighted mean across eight BEIR benchmarks, computed separately for each evaluation protocol. Calibration is measured by the overlap between the similarity-score distributions at different prefix sizes, reported as the percentage of shared probability mass for the top-N documents, as well as KL-divergence. Cross-size consistency is measured by the Jaccard overlap between the top-N documents a model retrieves at prefix size w and those it retrieves at full size. Because this depends only on the ordering within a size, it is invariant to any per-size score transformation and therefore reflects the embedding geometry rather than the score scale.

6 RESULTS

We report results for BGE-base-en-v1.5, our main setting, in this section. Results for E5-base-v2 and for the adapter on a frozen Qwen3-Embedding-0.6B backbone are given in Appendix F.

Mixed-size retrieval. When documents at different sizes compete in one ranking (Table 1), SCRL improves over MRL by 8.8 and 5.8 points on average, and does so in every domain (Table 6). A post-hoc per-width correction of MRL recovers slightly more (45.84 and 52.79), confirming experimentally that MRL's failure stems from unconstrained score offsets and scale. The same correction adds only $+1.96$ and -0.03 points to SCRL, against $+9.10$ and $+6.55$ for MRL, so SCRL's scores are already close to calibrated. Fitted on FiQA and NFCorpus, the affine correction transfers unevenly, from $+24.0$ points on TREC-COVID to -5.5 on FEVER, while SCRL improves on MRL in every domain (Table 7).

Table 1: Mixed-size retrieval, nDCG@10, unweighted mean over eight BEIR domains and five width-assignment seeds, with sample standard deviation. Scores are uncalibrated unless stated.

Method	Discrete	Continuous
BGE-base-en-v1.5		
MRL	36.74 $\pm$0.19	46.24 $\pm$0.13
MRL + affine calibration	45.84 $\pm$0.19	52.79 $\pm$0.14
SCRL (uniform discrete π)	45.58 $\pm$0.17	52.00 $\pm$0.17
SCRL + affine calibration	47.54 $\pm$0.27	51.97 $\pm$0.15
SCRL (uniform continuous π)	42.48 $\pm$0.25	51.12 $\pm$0.14
SCRL post-training (uniform discrete π)	45.13 $\pm$0.31	51.06 $\pm$0.20

Fixed-size retrieval. Table 2 isolates representation quality. Because a positive per-size affine map preserves every ranking within a size, no score calibration can change these numbers, so differences here reflect the learned embeddings alone. No single configuration dominates across the ladder and the sampling distribution π determines where along the ladder quality is allocated. Uniform discrete sampling and self-distillation shift capacity towards the narrowest prefixes, improving on MRL by 1.1 and 2.2 points at 32d in exchange for at most 1.1 points at larger sizes, while continuous sampling and the width curriculum match MRL from 256d upward in exchange for 2.5 and 0.7 points at 32d.

Table 2: Fixed-size retrieval, nDCG@10, unweighted mean over eight BEIR domains. Scoring controls are omitted because a positive per-size affine map preserves rankings within a size. SCRL + self-distillation was trained with a different recipe (Appendix C) and is not a controlled ablation.

Method	32d	64d	128d	256d	512d	768d
BGE-base-en-v1.5						
MRL	34.53	46.86	52.14	54.76	56.04	56.50
SCRL (uniform discrete π)	35.66	46.87	51.72	54.32	55.48	55.85
SCRL (uniform continuous π)	32.00	45.89	51.76	54.69	56.07	**56.63**
SCRL (width curriculum)	33.83	46.99	**52.23**	**54.93**	**56.18**	56.61
SCRL post-training (uniform discrete π)	35.81	46.31	50.67	53.15	54.34	54.76
SCRL + self-distillation	**36.69**	**47.95**	52.13	54.20	55.08	55.39

Cross-size calibration and neighbourhood consistency. Figure 2b compares score distributions at 64 and 768 dimensions. Under MRL the two share 71.7% of their probability mass (KL $= 0.232$); under SCRL they share 84.9% (KL $= 0.078$), and Figure 2c shows the same pattern across widths. Calibration only concerns the scale of the scores, which a post-hoc correction could also adjust. However, a positive per-size affine map cannot change which documents a model retrieves at a given size, so the extent to which a truncated index reproduces the full-size ranking is a property of the embedding geometry alone. Figure 2a measures this as the Jaccard overlap between each model's top-N at prefix size w and its own top-N at 768 dimensions. SCRL is ahead at every size and both

cutoffs. Seven of the eight comparisons survive Holm correction across the family of tests, including all four at top-100, where the gain is 1.3 to 2.1 Jaccard points; the exception is 128d at top-1. The effect is largest where truncation is most severe. At 32d the top-1 result agrees with the full-size result for 54.3% of queries under SCRL against 46.1% under MRL, and among queries where the two models disagree, 160 favour SCRL against 78 favouring MRL. This bears directly on cascaded retrieval, where a small-prefix first stage is only useful if its shortlist contains the documents the full-size stage would rank highly.

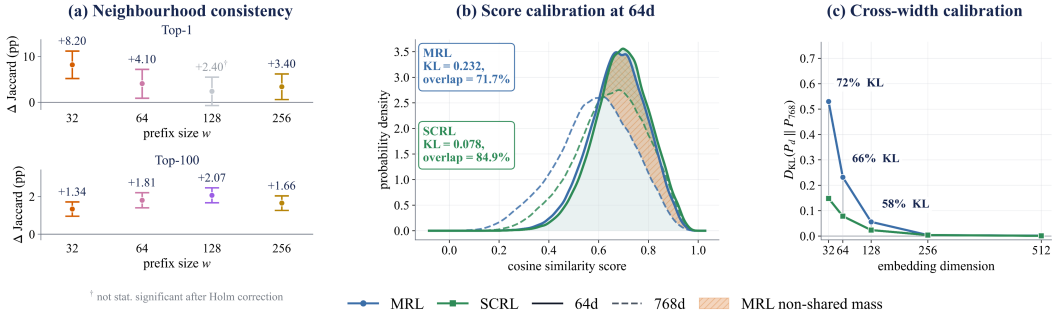


Figure 2: Cross-size behaviour of SCRL and MRL. (a) Gain of SCRL over MRL in neighbourhood consistency, measured as the Jaccard overlap between a model's top-N results at prefix size w and its own top-N at 768 dimensions, averaged over 1,000 MS MARCO development queries with 95% bootstrap intervals. † marks the one comparison that is not significant after Holm correction (Appendix D). (b) Distributions of the top-100 similarity scores at 64 and 768 dimensions. The shaded area shows the mass that MRL's two distributions do not share. (c) KL divergence between each width's score distribution and the 768-dimensional one.

What SCRL adds beyond score correction. A post-hoc correction and SCRL act on different objects. The correction rescales the scores of a fixed encoder, must be fitted separately for each checkpoint, and cannot change which documents a model retrieves at a given size. SCRL changes the encoder itself. Its scores are close to calibrated without a fitting step, and it provides other benefits that no score transform can, including higher quality at the narrowest prefixes (Table 2) and truncated neighbourhoods that agree more closely with the full size ranking (Figure 2). Additionally, while MRL already scales smoothly across all 737 widths from 32 to 768, the Kendall rank correlation between width and nDCG@10 rises from 0.889 under MRL to 0.913 under SCRL (Table 8), so quality under SCRL increases slightly more reliably with width.

7 Conclusion and Future Work

Standard Matryoshka training makes every prefix individually usable but leaves the sizes mutually incomparable because the cross-size offsets are an exact symmetry of its objective. SCRL removes the symmetry by placing candidates of independently sampled sizes inside one partition function, and in doing so also removes the need to fix a discrete size set and its loss weights in advance.

Empirically, SCRL nearly matches the mixed-size quality of a per-width affine correction fitted on held-out data, but does so with the embeddings themselves and without any extra fitting. It also improves cross-size neighbourhood consistency, which no score transformation can affect, most notably at narrow prefixes. Per-size quality is a trade-off rather than a uniform gain, and the sampling distribution π determines where along the ladder it falls. Mass concentrated on narrow prefixes improves those, while a curriculum or continuous sampling matches MRL from 256d upward.

Three directions follow. First, π is a design surface we have only begun to explore. A systematic search over allocations and schedules, rather than the three settings compared here, may recover both ends of the ladder at once. Second, the adapter setting lacks a matched MRL adapter on the same frozen backbone, so SCRL's contribution there is not yet isolated. Third, the storage, latency and quality trade-offs of a genuinely mixed-size production index remain to be measured, and depend on the deployment.

AI USE STATEMENT

In this work we used generative AI tools in the following way:

- As a writing aid in the form of re-formulating/paraphrasing individual statements, sentences, and paragraphs, i.e. as a a *glorified thesaurus*.
- As a coding assistant by creating draft implementations and reviewing our own code.
- Reviewing the proofs in our paper.
- Improving plots and figures.
- As literature-review assistant by finding additional similar work.
- For reviewing sections of our paper and identifying reasoning gaps.

The authors have checked **all** AI-assisted work. All theoretical claims were independently derived and verified by the authors, all bibliographic metadata was checked against primary sources, and all experimental results were produced and checked by the authors.

We take full responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies retrieval embedding efficiency and score calibration. It does not involve human subjects. Only public datasets and models were used as listed in Appendix C. As with dense retrieval in general, the models we train and evaluate may contain biases in their pretraining and fine-tuning data. SCRL does not change this in any way.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, the objective is specified in full in Section 3 and equation 1, and the training procedure requires only a size-sampling step and a single shared softmax. Proofs of all results in Section 4 are given in Appendix A. The models, benchmarks, training data and metrics are outlined in Section 5 with additional details in Appendices C and G.

REFERENCES

Anonymous. SUGAR: Spectral and geometry-aware alignment for matryoshka representation distillation. In *Submitted to ACL Rolling Review - May 2026*, 2026. URL https://openreview.net/forum?id=ifrlMRj2e7. under review.

Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In *International Conference on Database Theory (ICDT)*, LNCS 1540, pp. 217–235, 1999. doi: 10.1007/3-540-49257-7_15.

Yufei Cui, Yu Mao, Ziquan Liu, Qiao Li, Antoni B. Chan, Xue Liu, Tei-Wei Kuo, and Chun Jason Xue. Variational nested dropout. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10519–10534, August 2023. ISSN 1939-3539. doi: 10.1109/tpami.2023.3241945. URL http://dx.doi.org/10.1109/TPAMI.2023.3241945.

Devvrit, Sneha Kudugunta, Aditya Kusupati, Tim Dettmers, Kaifeng Chen, Inderjit Dhillon, Yulia Tsvetkov, Hannaneh Hajishirzi, Sham Kakade, Ali Farhadi, and Prateek Jain. Matformer: Nested transformer for elastic inference. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 140535–140564. Curran Associates, Inc., 2024. doi: 10.52202/079017-4461. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/fe066022bab2a6c6a3c57032a1623c70-Paper-Conference.pdf.

Michael Günther, Saba Sturua, Mohammad Kalim Akram, Isabelle Mohr, Andrei Ungureanu, Bo Wang, Sedigheh Eslami, Scott Martens, Maximilian Werk, Nan Wang, and Han Xiao. jina-embeddings-v4: Universal embeddings for multimodal multilingual retrieval, 2025. URL https://arxiv.org/abs/2506.18902.

Yue Hu, Haofei Gan, Jialong Wang, Xiaohan Zhang, and Yuxuan Qi. KaLM-Embedding: Superior training data brings a strong embedding model. *arXiv preprint arXiv:2501.13987*, 2025. doi: 10.48550/arXiv.2501.13987. URL https://arxiv.org/abs/2501.13987.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2205.13147.

Riwei Lai, Li Chen, Weixin Chen, and Rui Chen. Matryoshka representation learning for recommendation with layer- and hardness-adaptive negative sampling. *ACM Transactions on Intelligent Systems and Technology*, 17(4):1–25, August 2026. doi: 10.1145/3800942. URL https://doi.org/10.1145/3800942.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. NV-Embed: Improved techniques for training LLMs as generalist embedding models. *arXiv preprint arXiv:2405.17428*, 2024. doi: 10.48550/arXiv.2405.17428. URL https://arxiv.org/abs/2405.17428.

Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, Xiaoqi Ren, Shanfeng Zhang, Daniel Salz, Michael Boratko, Jay Han, Blair Chen, Shuo Huang, Vikram Rao, Paul Suganthan, Feng Han, Andreas Doumanoglou, Nithi Gupta, Fedor Moiseev, Cathy Yip, Aashi Jain, Simon Baumgartner, Shahrokh Shahi, Frank Palma Gomez, Sandeep Mariserla, Min Choi, Parashar Shah, Sonam Goenka, Ke Chen, Ye Xia, Koert Chen, Sai Meher Karthik Duddu, Yichang Chen, Trevor Walker, Wenlei Zhou, Rakesh Ghiya, Zach Gleicher, Karan Gill, Zhe Dong, Mojtaba Seyedhosseini, Yunhsuan Sung, Raphael Hoffmann, and Tom Duerig. Gemini embedding: Generalizable embeddings from gemini, 2025. URL https://arxiv.org/abs/2503.07891.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking, 2026. URL https://arxiv.org/abs/2601.04720.

Xianming Li, Zongxi Li, Jing Li, Haoran Xie, and Qing Li. 2d matryoshka sentence embeddings, 2024. URL https://arxiv.org/abs/2402.14776.

Olivier Marchal and Julyan Arbel. On the sub-Gaussianity of the Beta and Dirichlet distributions. *Electronic Communications in Probability*, 22(54):1–14, 2017. doi: 10.1214/17-ECP92.

Mervin E Muller. A note on a method for generating points uniformly on n-dimensional spheres. *Communications of the ACM*, 2(4):19–20, 1959.

Gia Huy Phung, Hai An Vu, Minh-Phuc Truong, Thang Duc Tran, Linh Ngo Van, Thanh Hong Nguyen, and Trung Le. MIPIC: Matryoshka representation learning via self-distilled intra-relational and progressive information chaining. *arXiv preprint arXiv:2604.24374*, 2026.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11:2487–2531, 2010.

Oren Rippel, Michael A. Gelbart, and Ryan P. Adams. Learning ordered representations with nested dropout. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, 2014. arXiv:1402.0915.

Arghamitra Talukder, Philippe Chlenski, and Itsik Pe'er. Objective-specific privileged bases via full-prefix matryoshka learning, 2026. URL https://arxiv.org/abs/2605.09160.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

Shuai Wang, Shengyao Zhuang, Bevan Koopman, and Guido Zuccon. 2d matryoshka training for information retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400715921. doi: 10.1145/3726302.3730330.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *PMLR*, pp. 9929–9939, 2020. arXiv:2005.10242.

Yueqi Wang, Zhenrui Yue, Huimin Zeng, Dong Wang, and Julian McAuley. Train once, deploy anywhere: Matryoshka representation learning for multimodal recommendation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13461–13472, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-emnlp.786/.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-pack: Packaged resources to advance general chinese embedding, 2023.

Jinsung Yoon, Rajarishi Sinha, Sercan O. Arik, and Tomas Pfister. Matryoshka-adaptor: Unsupervised and supervised tuning for smaller embedding dimensions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 10318–10336, 2024. 2024.emnlp-main.576; arXiv:2407.20243.

Jiahui Yu and Thomas Huang. Universally slimmable networks and improved training techniques. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1803–1811. IEEE, October 2019. doi: 10.1109/iccv.2019.00189. URL http://dx.doi.org/10.1109/ICCV.2019.00189.

Jiahui Yu, Linjie Yang, Ning Xu, Jianchao Yang, and Thomas Huang. Slimmable neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1gMCsAqY7.

Biao Zhang, Lixin Chen, Tong Liu, and Bo Zheng. SMEC: Rethinking matryoshka representation learning for retrieval embedding compression. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025a. 2025.emnlp-main.1332; arXiv:2510.12474.

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, 2025b. URL https://arxiv.org/abs/2506.05176.

Shengyao Zhuang, Shuai Wang, Fabio Zheng, Bevan Koopman, and Guido Zuccon. Starbucks-v2: Improved training for 2d matryoshka embeddings, 2025. URL https://arxiv.org/abs/2410.13230.

A Appendix

A.1 Proof of Lemma 7

Proof. Fix $w < w'$ and assume $u_{[:w]} \neq 0$. Since $\mu_{w,N}$ depends only on the marginal law of each $T_w^{(i)}$, and by rotational invariance this law is that of the first coordinate of a uniform vector on S^{w-1} for any u, we may take $u = e_1$ when computing it. This means that $\langle u, v\rangle = v_1$, the first coordinate of v. Instead of analysing two spheres in isolation, we couple the prefixes w' and w, using a shared random vector. Specifically, represent each $v_i = g^{(i)}/\|g^{(i)}\|$ using independent standard Gaussian vectors $g^{(i)} \sim \mathcal{N}(0, I_{w'})$ (Muller, 1959).

The similarity between u and the i-th negative sample at size w is then $T_w^{(i)} = g_1^{(i)}/\|g_{1:w}^{(i)}\|$. For size w', the similarity is $T_{w'}^{(i)} = g_1^{(i)}/\|g_{1:w'}^{(i)}\|$. Since $w' > w$, it is known that $\|g_{1:w'}\| \geq \|g_{1:w}\|$. Therefore $\operatorname{sign} T_{w'} = \operatorname{sign} T_w$ and $|T_{w'}| \leq |T_w|$.

Therefore, for $t \in (0, 1)$ the event $\{T_{w'} > t\}$ is contained in $\{T_w > t\}$. The containment is strict with positive probability. T_w does not depend on $g_{w+1:w'}$, while $T_{w'} \to 0$ as $\|g_{w+1:w'}\| \to \infty$, so the open set $\{T_w > t, T_{w'} < t\}$ is non-empty and since g has positive density everywhere, it has positive probability. Writing F_w for the cumulative distribution function of T_w this gives $F_w(t) < F_{w'}(t)$ for $t \in (0, 1)$.

For the maximum similarity $M_w = \max_i T_w^{(i)}$ of N independent draws, $\mathbb{P}(M_w \leq t) = F_w(t)^N$.

The expectation of M_w is given by the following.

$$\mu_{w,N} = \mathbb{E}[M_w] = \int_0^1 \mathbb{P}(M_w > t)\, dt - \int_{-1}^0 \mathbb{P}(M_w \leq t)\, dt$$

Substituting $\mathbb{P}(M_w \leq t) = F_w(t)^N$ into the above expression, and then mapping the second integral onto $[0, 1]$ via $t \mapsto -t$ and $F_w(-t) = 1 - F_w(t)$, we get

$$\mu_{w,N} = \int_0^1 \Big[1 - F_w(t)^N - \big(1 - F_w(t)\big)^N\Big] dt$$

Defining $h(s) := s^N + (1 - s)^N$ lets us write $\mu_{w,N} = \int_0^1 \big[1 - h(F_w(t))\big] dt$. Subtracting the same expression for w' gives

$$\mu_{w,N} - \mu_{w',N} = \int_0^1 \big[h(F_{w'}(t)) - h(F_w(t))\big] dt$$

The derivative of h with respect to s is given by $h'(s) = N\big[s^{N-1} - (1 - s)^{N-1}\big]$. For $N \geq 2$, $h'(s) > 0$ on $(\frac{1}{2}, 1]$, so h is strictly increasing on $[\frac{1}{2}, 1]$. Moreover, since F_w is symmetric around 0, we know $\frac{1}{2} \leq F_w(t)$ for $t \in (0, 1)$ since at least half of the distribution mass sits below 0. Coupling of the vectors means $F_w(t) < F_{w'}(t)$. Therefore, $\frac{1}{2} \leq F_w(t) < F_{w'}(t) \leq 1$ and the integrand is strictly positive and

$$\mu_{w,N} - \mu_{w',N} > 0$$

For $N = 1$, $h(s) = 1$ for all s and the sides cancel out. Therefore $N \geq 2$ is required. □

A.2 Proof of Theorem 9

Proof. Denote the augmented scores $s_{w_i}(q, d_i) + b_{w_i}$ as z_i and the softmax probability at index i as $p_i = e^{z_i}/\sum_j e^{z_j}$. The cross-entropy loss with positive index y is $\ell = -\log p_y$, which expands as

$$\ell = -\log \frac{e^{z_y}}{\sum_j e^{z_j}} = -\log e^{z_y} + \log \sum_j e^{z_j} = -z_y + \log \sum_j e^{z_j}.$$

Differentiating ℓ with respect to a single logit z_i gives the following,

$$\frac{\partial \ell}{\partial z_i} = -\mathbf{1}\{i = y\} + \frac{e^{z_i}}{\sum_j e^{z_j}} = p_i - \mathbf{1}\{i = y\},$$

and since b_w appears in z_i precisely when candidate i carries size w, $\partial z_i/\partial b_w = \mathbf{1}\{w_i = w\}$. The chain rule gives

$$\frac{\partial \ell}{\partial b_w} = \sum_i \frac{\partial \ell}{\partial z_i}\frac{\partial z_i}{\partial b_w} = \sum_i \big(p_i - \mathbf{1}\{i = y\}\big)\mathbf{1}\{w_i = w\}$$
$$= \underbrace{\sum_i p_i \mathbf{1}\{w_i = w\}}_{= P_w} - \underbrace{\sum_i \mathbf{1}\{i = y\}\mathbf{1}\{w_i = w\}}_{= \mathbf{1}\{w_y = w\}} = P_w - \mathbf{1}\{w_y = w\},$$

where the first sum restricted to size w is P_w by definition, and the second is nonzero only if the positive itself carries size w.

Writing $S = \sum_k e^{z_k}$, so that $p_i = e^{z_i}/S$, the numerator depends on z_j only when $i = j$ while the denominator always does, giving $\partial e^{z_i}/\partial z_j = \mathbf{1}\{i = j\}e^{z_i}$ and $\partial S/\partial z_j = e^{z_j}$. The quotient rule then gives

$$\frac{\partial p_i}{\partial z_j} = \frac{\mathbf{1}\{i = j\}\, e^{z_i} S - e^{z_i} e^{z_j}}{S^2} = \mathbf{1}\{i = j\}\frac{e^{z_i}}{S} - \frac{e^{z_i}}{S}\frac{e^{z_j}}{S} = p_i\big(\mathbf{1}\{i = j\} - p_j\big).$$

Differentiating again with respect to $b_{w'}$ this time, and using the above result and the fact that $\mathbf{1}\{w_y = w\}$ does not depend on $b_{w'}$ so that term vanishes,

$$\frac{\partial^2 \ell}{\partial b_w \partial b_{w'}} = \frac{\partial P_w}{\partial b_{w'}} = \sum_i \mathbf{1}\{w_i = w\}\sum_j \frac{\partial p_i}{\partial z_j}\frac{\partial z_j}{\partial b_{w'}}$$
$$= \sum_{i,j} p_i\big(\mathbf{1}\{i = j\} - p_j\big)\mathbf{1}\{w_i = w\}\mathbf{1}\{w_j = w'\} = \mathbb{E}_p[\mathbf{1}_w \mathbf{1}_{w'}] - P_w P_{w'}$$

and since $P_w = \mathbb{E}_p[\mathbf{1}_w]$, the above expression is $\mathbb{E}_p[\mathbf{1}_w \mathbf{1}_{w'}] - \mathbb{E}_p[\mathbf{1}_w]\mathbb{E}_p[\mathbf{1}_{w'}] = \mathrm{Cov}_p(\mathbf{1}_w, \mathbf{1}_{w'})$. Finally, for any $a \in \mathbb{R}^{|\mathcal{M}|}$, by bilinearity of covariance, we get the following.

$$a^\top \nabla_b^2 \ell\, a = \sum_{w,w'} a_w \mathrm{Cov}_p(\mathbf{1}_w, \mathbf{1}_{w'}) a_{w'} = \mathrm{Cov}_p\Big(\sum_w a_w \mathbf{1}_w, \sum_{w'} a_{w'} \mathbf{1}_{w'}\Big)$$

Since $\sum_w a_w \mathbf{1}_w(i) = a_{w_i}$ for every candidate i as all other indicators will be zero, and as w and w' are dummy indices which coincide, $\sum_w a_w \mathbf{1}_w = \sum_{w'} a_{w'} \mathbf{1}_{w'}$. The covariance is therefore a variance and,

$$a^\top \nabla_b^2 \ell\, a = \mathrm{Var}_p(a_{w_i})$$

□

A.3 Proof of Theorem 10

Proof. The proof is divided into multiple sections corresponding with the statements in the theorem.

(i) By Theorem 9, the Hessian of the per-batch loss in the b-directions satisfies $a^\top \nabla_b^2 \ell\, a = \mathrm{Var}_p(a_{w_i}) \geq 0$. Hence, the Hessian is positive semidefinite and so each per-batch loss is convex in b, and so is their expectation $\mathcal{L}_{\mathrm{SCRL}}$ since a convex combination of convex functions is convex. For $a = \mathbf{1}$ the quantity a_{w_i} is constant and thus there is no curvature. This is the same as the invariance of Lemma 2.

To show $b \mapsto \mathcal{L}_{\mathrm{SCRL}}$ is strictly convex on the quotient $\mathbb{R}^{|\mathcal{M}|}/\mathrm{span}(\mathbf{1})$, let $a \notin \mathrm{span}(\mathbf{1})$, so $a_w \neq a_{w'}$ for some $w \neq w'$. Since π has full support on $\mathcal{M}$ and $B \geq 2$, a batch contains candidates of both sizes with positive probability. In any such batch, a_{w_i} will take at least two distinct values. Since every softmax weight is strictly positive, $\mathrm{Var}_p(a_{w_i}) > 0$ for that batch. In any other batch, the variance is non-negative. Letting A denote the event that a batch contains candidates of both sizes w and w', so that $\mathbb{P}(A) > 0$, and splitting the expectation of the variance on A and its complement,

$$\mathbb{E}\big[\mathrm{Var}_p(a_{w_i})\big] = \underbrace{\mathbb{P}(A)\,\mathbb{E}\big[\mathrm{Var}_p(a_{w_i}) \mid A\big]}_{> 0} + \underbrace{\mathbb{P}(A^c)\,\mathbb{E}\big[\mathrm{Var}_p(a_{w_i}) \mid A^c\big]}_{\geq 0} > 0$$

Therefore, on the quotient $\mathbb{R}^{|\mathcal{M}|}/\mathrm{span}(\mathbf{1})$ the Hessian is positive definite since $a^\top \nabla_b^2 \mathcal{L}_{\mathrm{SCRL}}\, a = \mathbb{E}[\mathrm{Var}_p(a_{w_i})] > 0$.

(ii) Let $M = \max_w b_w - \min_w b_w$ be the spread of the offsets, attained at size indices w^+ and w^-. Since π has full support over $\mathcal{M}$ and $B \geq 2$, a batch has its positive at w^- and some negative at w^+ with positive probability $\rho > 0$. On such a batch, the following inequality holds,

$$(s_{w^+} + b_{w^+}) - (s_{w^-} + b_{w^-}) \;\geq\; M - 2S,$$

since the base scores s_{w^+} and s_{w^-} differ by at most $2S$. Writing $z_i = s_{w_i}(q, d_i) + b_{w_i}$ for the logits, in such a batch the negative at w^+ outscores the positive by some $g \geq M - 2S$. The loss then satisfies

$$\ell = -\log \frac{e^{z_y}}{\sum_j e^{z_j}} = \log \sum_j e^{z_j - z_y} \;\geq\; \log\bigl(1 + e^{g}\bigr) \;\geq\; g \;\geq\; M - 2S,$$

where the first inequality keeps only the positive's own term and that of the negative at w^+. On every other batch $\ell \geq 0$, so

$$\mathcal{L}_{\mathrm{SCRL}} \;\geq\; \rho\,(M - 2S),$$

which diverges as $M \to \infty$. The objective is therefore coercive on the quotient, and due to the strict convexity of (i) a unique minimum is attained there.

(iii) Taking the expectations of equation 6 and using that the positive's size is drawn from π independently of the candidates gives $\mathbb{E}[\mathbf{1}\{w_y = w\}] = \pi(w)$. Therefore,

$$\frac{\partial \mathcal{L}_{\mathrm{SCRL}}}{\partial b_w} = \mathbb{E}[P_w] - \pi(w).$$

Setting this to zero for every w gives equation 8, and by (i) and (ii) this stationary point is the unique minimiser modulo $\mathrm{span}(\mathbf{1})$. □

B Simulated top-negative floors

Proposition 12 (Numerical magnitude)**.** *Take $N = 1000$, $W = 256$ and $\tau = 0.05$. Direct simulation of μ_w, averaged over two thousand independent query draws, gives the values in Table 3, so the bound from equation 5 holds strictly at every size in this simulation. The 32-against-256 gap is $\mu_{32} - \mu_{256} = 0.337$ in cosine units.*

Table 3: Simulated top-negative floors μ_w against the bound $\sqrt{2 \ln N/w}$, at $N = 1000$, $W = 256$, averaged over two thousand query draws. Values are cosine similarities.

PREFIX SIZE w	**16**	**32**	**64**	**128**	**256**
$\mu_{w,N}$ (simulated)	0.713	0.538	0.394	0.283	0.201
$\sqrt{2 \ln N/w}$ (bound)	0.929	0.657	0.465	0.329	0.232

C Additional information about experimental setup

Training data. Rather than the common protocol of training on MS MARCO alone and evaluating on BEIR, we sample a mixture of BEIR-derived training sets together with MS MARCO, using each benchmark's corresponding training split where one exists. This makes the in-domain benefit visible rather than confounding it with domain shift. Three evaluation domains, NFCorpus, SciDocs and TREC-COVID, are held out of the training data so that generalisation can be measured out of domain. Held-out domains contribute no training data. The DEV split of NFCorpus is, however, used for checkpoint selection and for fitting the post-hoc calibration (Appendix G). Test splits are used only for evaluation. Table 4 shows the composition of the training data where all of the pairs were randomly sampled from the dataset. All BGE and E5 runs, including self-distillation, use the full mixture of 276,916 pairs.

Table 4: Composition of the training mixture, in query–positive pairs. Both BGE and E5 use the same complete mixture of 276,916 pairs.

Dataset	Pairs	Share
MS MARCO	150,000	54.17%
NQ	45,000	16.25%
HotpotQA	45,000	16.25%
FEVER	30,000	10.83%
FiQA	6,000	2.17%
SciFact	916	0.33%
Total	276,916	100%

Negatives. Only some of the source datasets ship mined negatives, so we use the same mining strategy across the mixture to avoid favouring datasets whose provided negatives happen to be of higher quality. For each domain, we encode all queries and documents with Qwen3-Embedding-4B (Zhang et al., 2025b). Using those vectors, we retrieve the 50 highest-scoring documents, and remove any known positives using the relevance labels provided by the datasets. That leaves every query-positive pair with 50 hard negatives, from which 16 are drawn per training step (Appendix G).

Models. Our main experiments use BGE-base-en-v1.5 (109M parameters), for which we compare MRL and all SCRL variants trained from the same checkpoint, data, batch size and training budget. To test whether the results transfer to a different encoder, we repeat the MRL and SCRL comparison on E5-base-v2 (110M), trained on the same mixture (Table 4). For the adapter setting, we train an SCRL output adapter on a frozen Qwen3-Embedding-0.6B backbone. As this model is already trained with MRL, the bare backbone is treated as the MRL baseline in this case. Qwen3-Embedding-4B is used only to mine negatives and is not among the evaluated encoders.

Description of self-distillation. In our experiments, we added an online cross-width distillation term to the SCRL loss inspired by Phung et al. (2026). At every step, the full-width (W) score distribution over the candidate set teaches every smaller width on the training ladder. For query i and widths $s < W$, teacher and student distributions are softmaxes over the prefix-normalised cosine scores. The distillation loss is $\mathrm{KL}(P \,\|\, U)$, where the teacher distribution P and student distribution U are softmaxes over the prefix-normalised cosine scores at widths W and s, scaled by $\kappa = 10$ (equivalent to a temperature of 0.1) and excluding known positives other than the target. The loss is averaged over queries and then over student widths, and then added to the SCRL loss with weight 1. To make it so only the smaller widths try to mimic the teacher and not the other way around, the teacher scores are detached from the gradient. Note that teacher and students are views of the same encoder, so no separate teacher model is trained.

Description of width curriculum. The width-curriculum model uses the discrete ladder $\mathcal{M} = \{32, 64, 128, 256, 512, 768\}$, but the sampling distribution π is altered during training. At the start, each width is sampled with a probability that is proportional to their size, so that larger prefixes are drawn more often. During the first half of training, the distribution gradually becomes uniform and then it stays uniform for the second half of training. The warmup is specified as follows. With t completed optimizer steps out of T, we set $u = \min\big(2t/T,\ 1\big)$ and $\alpha = \frac{1}{2}\big(1 - \cos(\pi u)\big)$, and sample widths from

$$\pi_t(w) = (1 - \alpha)\,\frac{w}{\sum_{v \in \mathcal{M}} v} + \frac{\alpha}{|\mathcal{M}|}$$

The schedule depends only on the optimizer step and does not reset between epochs, so in our two-epoch run the transition takes approximately the first epoch. Every width keeps nonzero probability throughout, so the full-support condition of Theorem 10 holds at every step.

D Neighbourhood consistency statistics

This section gives the full statistics behind Figure 2a.

Table 5: Within-model neighbourhood consistency: mean Jaccard overlap between each model's top-N at prefix size d and its own top-N at 768 dimensions, over 1,000 MS MARCO development queries. Deltas are paired per query, with 95% bootstrap intervals and Holm-adjusted p-values across the eight tests.

d	Cutoff	MRL	SCRL	Δ	95% CI	Holm p
32	top-1	.461	.543	.082	[.052, .112]	$2.3 \cdot 10^{-7}$
64	top-1	.607	.648	.041	[.009, .072]	.018
128	top-1	.727	.751	.024	[-.007, .055]	.071
256	top-1	.809	.843	.034	[.006, .062]	.020
32	top-100	.209	.223	.013	[.010, .017]	$3.4 \cdot 10^{-11}$
64	top-100	.360	.378	.018	[.014, .022]	$6.3 \cdot 10^{-18}$
128	top-100	.494	.515	.021	[.017, .025]	$4.2 \cdot 10^{-21}$
256	top-100	.631	.648	.017	[.013, .020]	$2.8 \cdot 10^{-15}$

E PER-DOMAIN RESULTS

This section contains results broken down by what dataset among the eight BEIR benchmarks was used.

Table 6: Per-domain nDCG@10 for *BGE-base-en-v1.5*. Domains above the line contribute training data with their respective training sets; those below are out of domain. Mixed-size results are means over five width-assignment seeds and are uncalibrated.

	Fixed 32d			Mixed discrete			Mixed continuous		
Domain	**MRL**	**SCRL**	**+post**	**MRL**	**SCRL**	**+post**	**MRL**	**SCRL**	**+post**
NQ	27.68	**29.16**	27.96	29.93	**37.08**	35.06	39.01	**43.23**	40.28
HotpotQA	31.79	**33.69**	33.26	30.23	45.16	**45.33**	43.94	56.88	**57.41**
FEVER	72.39	**74.40**	73.60	76.94	**81.64**	80.77	85.03	**86.86**	85.51
SciFact	49.03	49.82	**53.53**	56.00	63.16	**63.54**	65.00	68.87	**69.12**
FiQA	18.99	**19.52**	18.36	14.78	**26.88**	25.93	23.70	**33.91**	32.44
In-domain mean	39.98	41.32	**41.34**	41.58	**50.78**	50.13	51.34	**57.95**	56.95
SciDocs	10.97	**11.06**	10.20	13.66	**16.15**	15.45	18.23	**19.08**	18.43
NFCorpus	20.49	20.34	**22.37**	20.96	28.83	**29.28**	28.68	**32.71**	32.61
TREC-COVID	44.92	**47.31**	47.22	51.44	65.70	**65.71**	66.32	**74.46**	72.65
Out-of-domain mean	25.46	26.24	**26.60**	28.69	**36.89**	36.81	37.74	**42.08**	41.23

Table 7: Mixed-size retrieval per domain, discrete protocol, nDCG@10. The calibration is fitted once on FiQA and NFCorpus development data and frozen across all test sets; those two domains are marked $*$. Its benefit varies from -5.5 to $+24.0$ points.

Domain	MRL	MRL+aff.	SCRL	Calib. gain	SCRL − aff.
FEVER	76.94	71.43	**81.64**	−5.51	+10.21
SciFact	56.00	58.40	**63.16**	+2.40	+4.76
SciDocs	13.66	**16.50**	16.15	+2.84	−0.35
NQ	29.93	34.87	**37.08**	+4.95	+2.20
NFCorpus*	20.96	**30.68**	28.83	+9.73	−1.85
FiQA*	14.78	**27.69**	26.88	+12.90	−0.81
HotpotQA	30.23	**51.65**	45.16	+21.42	−6.49
TREC-COVID	51.44	**75.48**	65.70	+24.04	−9.78
Mean	36.74	**45.84**	45.58	+9.10	−0.26

Table 8: Monotonic predictability across all 737 evaluated dimensions ($d \in [32, 768]$), measured via Kendall's rank correlation (τ) between dimension budget and nDCG@10. Domains above the line contribute training data with their respective training sets; those below are out of domain.

Domain	**MRL**	**SCRL**	$\Delta\tau$ **(SCRL − MRL)**
NQ	**0.9571**	0.9373	−0.0197
HotpotQA	0.9725	**0.9807**	+0.0081
FEVER	**0.9303**	0.9114	−0.0189
SciFact	0.7384	**0.8753**	+0.1369
FiQA	0.8757	**0.9438**	+0.0681
In-domain mean	0.8948	**0.9297**	+0.0349
SciDocs	0.8941	**0.9330**	+0.0388
NFCorpus	**0.9249**	0.8818	−0.0432
TREC-COVID	0.8190	**0.8439**	+0.0249
Out-of-domain mean	0.8793	**0.8862**	+0.0069
Overall mean	0.8890	**0.9134**	+0.0244

F Full Results

This section extends the main results to all models and training variants. In addition to the results shown here, we attempted to reproduce the implementation of MIPIC (Phung et al. (2026)) but our attempts were not successful. Those results are therefore left outside the results table.

Table 9: Fixed-size retrieval, nDCG@10, unweighted mean over eight BEIR domains. Scoring controls are omitted because a positive per-size affine map preserves rankings within a size. Model families are not comparable to one another: the Qwen rows train only a residual output adapter on a frozen backbone, and Qwen has a native width of 1024.

Method	**32d**	**64d**	**128d**	**256d**	**512d**	**768d**	**1024d**
BGE-base-en-v1.5							
MRL	34.53	46.86	52.14	54.76	56.04	56.50	–
SCRL (discrete π)	35.66	46.87	51.72	54.32	55.48	55.85	–
SCRL (continuous π)	32.00	45.89	51.76	54.69	56.07	**56.63**	–
SCRL post-training	35.81	46.31	50.67	53.15	54.34	54.76	–
SCRL (width curriculum)	33.83	46.99	**52.23**	**54.93**	**56.18**	56.61	–
SCRL + self-distillation	**36.69**	**47.95**	52.13	54.20	55.08	55.39	–
E5-base-v2							
MRL	31.29	44.59	51.05	**54.38**	**55.89**	**56.22**	–
SCRL (discrete π)	31.58	44.35	50.71	53.27	54.51	54.96	–
SCRL + self-distillation	**34.43**	**46.03**	**51.26**	53.76	55.05	55.18	–
Qwen3-Embedding-0.6B, frozen backbone with residual output adapter							
No adapter	30.35	44.35	52.24	55.87	57.90	58.49	58.53
SCRL adapter	30.29	44.05	52.20	55.83	57.98	58.51	58.47

Table 10: Mixed-size retrieval, nDCG@10, unweighted mean over eight BEIR domains and five width-assignment seeds, with sample standard deviation. Scores are uncalibrated unless stated. The Qwen rows train only an output adapter on a frozen backbone and are not comparable to the fully fine-tuned BGE rows.

Method	**Discrete**	**Continuous**
BGE-base-en-v1.5		
MRL	36.74 $\pm$0.19	46.24 $\pm$0.13
MRL + affine calibration	45.84 $\pm$0.19	52.79 $\pm$0.14
SCRL (uniform discrete π)	45.58 $\pm$0.17	52.00 $\pm$0.17
SCRL + affine calibration	47.54 $\pm$0.27	51.97 $\pm$0.15
SCRL (uniform continuous π)	42.48 $\pm$0.25	51.12 $\pm$0.14
SCRL (width curriculum)	44.37 $\pm$0.04	52.30 $\pm$0.14
SCRL + self-distillation	45.87 $\pm$0.21	52.03 $\pm$0.17
SCRL post-training	45.13 $\pm$0.31	51.06 $\pm$0.20
E5-base-v2		
MRL	28.51 $\pm$0.19	36.65 $\pm$0.37
MRL + affine calibration	40.85 $\pm$0.46	50.13 $\pm$0.31
SCRL (uniform discrete π)	39.19 $\pm$0.32	48.28 $\pm$0.23
SCRL + self-distillation	41.90 $\pm$0.42	49.94 $\pm$0.21
Qwen3-Embedding-0.6B, frozen backbone with output adapter		
No adapter	24.33 $\pm$0.28	32.19 $\pm$0.30
SCRL adapter	29.50 $\pm$0.31	40.31 $\pm$0.19

G Training Hyperparameters

For the BGE-base MRL/SCRL comparison, both methods are initialized from BGE-base-en-v1.5 and fine-tuned on the same frozen mixture of 276,916 query–positive training pairs. The encoder parameters are shared between queries and documents and are updated during training. We use the final-layer CLS representation, truncate it to the required prefix width, and independently ℓ_2-normalize each prefix. Queries receive the BGE retrieval instruction, while documents are encoded without an instruction prefix.

Contrastive objective and temperature. Let $s_w(q_i, d_j)$ denote the cosine similarity between the normalized width-w prefixes of query q_i and document d_j, and let p_i index the designated positive. The training logits are

$$\ell_{ij} = \frac{s_{w_j}(q_i, d_j) - 0.05\,\mathbf{1}[j = p_i]}{0.05}.$$

Thus, the implementation's cosine-logit multiplier of 20 corresponds to temperature $\tau = 0.05$; the positive margin is subtracted before temperature scaling. MRL computes a separate contrastive cross-entropy at each of the six widths and averages these losses equally. SCRL instead averages five stochastic mixed-width cross-entropies. Neither self-distillation nor explicit score-calibration regularization is enabled in this matched comparison.

In-batch negatives and width assignment. Each local query contributes one positive and 16 mined negatives. All document occurrences are gathered across the eight ranks, giving

$$8 \times 16 \times (1 + 16) = 2,176$$

candidate columns per query before false-negative masking. Consequently, other queries' positives and mined negatives provide additional in-batch negatives. These are document occurrences, not necessarily unique document IDs. Known additional positives are excluded when matching qrels and document identifiers are available; the designated positive is retained. Remote document embeddings are gathered without gradients, while the local rank's document embeddings retain their gradients.

In each SCRL pass, every document column independently samples a width uniformly from $\{32, 64, 128, 256, 512, 768\}$. This includes positive, mined-negative, and other in-batch document

Hyperparameter	Value
Backbone	BGE-base-en-v1.5
Native embedding dimension	768
Training widths	$\{32, 64, 128, 256, 512, 768\}$
Query / document token limits	128 / 512
Training pairs	276,916
Training epochs	2
Optimizer updates	4,326
Query batch size per GPU	16
Number of GPUs	8
Global query batch size	128
Gradient accumulation steps	1
Optimizer	AdamW
Peak learning rate	2×10^{-5}
AdamW (β_1, β_2)	$(0.9, 0.999)$
AdamW ϵ	10^{-8}
Weight decay	0.01
Learning-rate schedule	Linear warm-up, then linear decay to zero
Warm-up updates	200
Gradient clipping	Global norm 1.0
Precision	BF16
Activation checkpointing	Enabled
Distributed optimizer	DeepSpeed ZeRO stage 2
Contrastive temperature τ	0.05
Positive cosine margin	0.05
Mined negatives per query	16 from a stored pool of 50
SCRL stochastic passes per batch	5
Training seed	42

Table 11: Shared training settings for the matched BGE-base MRL and SCRL models. AdamW betas, epsilon, and weight decay use the library defaults. Incomplete per-rank batches are dropped.

columns. A column's sampled width is shared across all query rows within that pass. Each query is truncated and renormalized to the corresponding document width when scoring that column; unequal-length vectors are never compared. Assignments are resampled across passes and batches. No width curriculum is used.

Negative-hardness curriculum. The stored negative pool is ordered from hardest to easiest. Throughout training, the number of mined negatives remains fixed at 16. A sliding window moves from the easiest retained candidates (positions 35–50) to the hardest candidates (positions 1–16). For training progress $u \in [0, 1]$, define

$$h(u) = \frac{1 - \cos(\pi \min(2u, 1))}{2}.$$

The zero-based window start is $\mathrm{round}((1 - h(u))(50 - 16))$. The curriculum therefore reaches the hardest window halfway through training and retains it thereafter. Negative hardness and embedding width are controlled independently. The same curriculum is applied to both MRL and SCRL.

Checkpoint selection. We evaluate every 300 updates and at the final update. Checkpoints are selected by the highest equal-domain mean of fixed-width $\mathrm{nDCG@10}$ over the DEV splits of MS MARCO, HotpotQA, FEVER, FiQA, and NFCorpus. Within each domain, scores are first averaged over the six widths. Periodic evaluation uses up to 1,000 queries per domain, their relevant documents, and up to 50,000 additional distractor documents, with encoding batch size 256. Downstream evaluation uses the selected checkpoint rather than necessarily the final-step weights. TEST metrics are not used for checkpoint selection.

Post-hoc affine score calibration. For each checkpoint to which calibration is applied, we fit a separate, label-free calibration using 50,000 DEV query–document pairs: 25,000 from FiQA and 25,000 from NFCorpus. Query and document indices are sampled independently and uniformly with replacement using seed 42. The same sampled pairs are scored at every fitted width.

For each integer width $w \in \{32, \ldots, 768\}$, let μ_w and σ_w denote the mean and standard deviation of the pooled DEV cosine scores. We use population variance with a variance floor of 10^{-12} and compute

$$a_w = \frac{\sigma_{768}}{\sigma_w}, \qquad b_w = \mu_{768} - a_w \mu_w, \qquad s'_w = a_w s_w + b_w.$$

The fitting criterion is matching each width's marginal score mean and standard deviation to the native-width reference. Parameters are obtained *independently per width in closed form*, not by joint multi-width optimization, supervised regression, or a ranking loss. The native-width transform is explicitly set to identity. Moment accumulation uses float64 arithmetic and processing blocks of 2,048 pairs. Relevance labels and TEST data are not used.

Here, "global" means that each width's parameters are pooled across DEV queries and the two calibration domains, not that all widths share one affine transform. The fitted parameters are frozen and applied unchanged across TEST datasets. Calibration operates on raw cosine scores, without the training temperature multiplier or positive margin.