IMPROVING MULTILINGUAL SEMANTIC ALIGNMENT WITH MULTI-POSITIVE CONTRASTIVE LEARNING FOR INFORMATION RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Multilingual embedding models do not always retrieve all semantically equivalent relevant documents within the top-ranked results. For a given query, relevant documents in some languages are ranked substantially lower than parallel documents written in other languages, and retrieval performance also varies across query languages. In this paper, we introduce Multilingual Semantic Alignment (MUSA), a training method that aims to enhance semantic similarity among parallel document embeddings while jointly optimizing retrieval over multilingual positives associated with each query. Beyond query–document relevance, MUSA directly aligns the embeddings of multilingual parallel documents through a dimension-wise cross-correlation alignment objective. Experiments with 13 multilingual embedding models show that MUSA improves multilingual retrieval performance in settings with multiple references and reduces performance disparities across query languages observed in the baselines. Across the evaluated settings, MUSA reduces the depth required to retrieve all relevant documents by up to 84.62% relative to the corresponding baselines. Analysis shows that improvements in multilingual retrieval are associated with better alignment of parallel document embeddings.

1 INTRODUCTION

With the growing use of retrieval-augmented generation (RAG) (Lewis et al., 2020b; Gao et al., 2023) and multi-agent systems (MAS) (Chen et al., 2025), information retrieval (IR) has become a fundamental component for accessing external knowledge in modern NLP applications. Retrieved documents provide evidence for subsequent response generation and reasoning, making the retrieval of relevant documents essential to the use of external knowledge in these systems (Asai et al., 2024; Trivedi et al., 2023). In real-world scenarios, user queries and retrieval corpora often span multiple languages (Jang et al., 2026b). To support such settings, information retrieval has increasingly expanded to enable access to relevant information across languages (Asai et al., 2021). This requires models to assess semantic relevance between queries and documents despite differences in vocabulary and grammatical structure. Multilingual embedding models have been developed to support retrieval across languages by encoding text from a wide range of languages within a unified embedding space (Feng et al., 2022a; Chen et al., 2024).

Multilingual embedding models have achieved strong retrieval performance within individual languages (Wang et al., 2024; Zhang et al., 2025). However, this does not necessarily ensure that semantically equivalent relevant documents in different languages are retrieved consistently (Roy et al., 2020). Language-dependent retrieval biases and ranking disparities among relevant documents can arise when documents in multiple languages coexist in the same retrieval corpus (Yang et al., 2024c; Jang et al., 2026c). When a single query is associated with parallel relevant documents in multiple languages, some may be retrieved near the top, while others have lower similarity scores with the query and are ranked much lower despite containing the same information (Hong et al., 2026). We also observe that multilingual embedding models struggle to retrieve all parallel relevant documents within the top-ranked results, with retrieval performance varying across query languages. Such misalignment can limit access to relevant documents in certain languages, even when semantically equivalent documents in other languages are successfully retrieved.

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.

To address this limitation, we propose **Mu**ltilingual **S**emantic **A**lignment (MUSA), a training method that improves semantic alignment among multilingual relevant documents. For each query, MUSA groups multilingual relevant documents with shared semantic content into a unified positive set. By jointly optimizing query–document similarities over this set, MUSA encourages each positive document to receive a higher similarity score than negative documents. We further treat semantically equivalent positive documents in different languages as distinct linguistic views and directly align their document representations through a dimension-wise cross-correlation objective. As shown in Figure 1, our approach encourages consistent patterns across the embeddings of semantically equivalent documents in different languages. With these two objectives, MUSA aims to retrieve all parallel relevant documents more consistently across query languages.

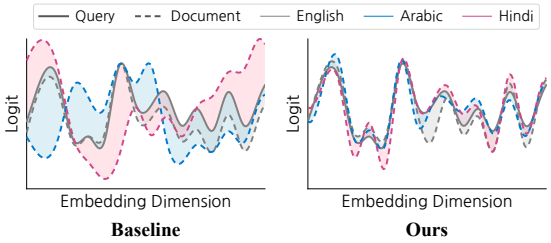

Figure 1: Illustration of multilingual alignment before (left) and after (right) applying MUSA.

To evaluate whether MUSA improves the retrieval of multilingual relevant documents and reduces performance disparities across query languages, we construct controlled retrieval settings using XQuAD and BELEBELE, which provide parallel queries and passages across languages. For each benchmark, we combine passages from all languages into a single retrieval corpus, where each query is associated with a set of parallel relevant passages. This setting requires models to retrieve all corresponding passages, rather than only a subset expressed in particular languages. We assess whether the entire set is included in the top-ranked results and measure the retrieval depth required to retrieve all relevant passages. We also compare performance across query languages to assess the consistency of multilingual retrieval.

Experiments across 13 multilingual embedding models spanning different scales and architectures show that MUSA improves multilingual retrieval performance across all evaluated models on both benchmarks. On XQuAD, Comp@20 increases by up to 60.69 points and Max@R decreases by up to 84.62% relative to the corresponding baselines. Language-wise analysis shows retrieval improvements across multiple query languages, while quantitative alignment analysis reveals that parallel documents become more closely aligned in the embedding space for most models. Ablation studies further show that directly aligning multilingual document embeddings with the BT objective improves retrieval performance beyond multi-positive optimization alone.

2 Diagnosing Multilingual Information Retrieval

We examine whether existing multilingual embedding models retrieve all semantically equivalent relevant documents from a shared multilingual corpus. Using parallel relevant documents across languages, we measure how often the complete set of relevant documents appears among the top-ranked results and the retrieval depth required to retrieve the entire set. We further compare these two metrics across query languages to examine performance differences.

Setup. We use XQuAD (Artetxe et al., 2020), which provides parallel queries and relevant documents in 12 languages, and combine documents from all languages into a single retrieval corpus. For each query q, we define the set of parallel relevant documents across the 12 languages as R_q and analyze retrieval performance by query language using eight publicly available multilingual embedding models. We evaluate performance using Comp@20 (Qu et al., 2024) and Max@R (Hong et al., 2026). Comp@K measures retrieval success based on whether all documents in R_q are included in the top-K results. Max@R is the rank of the lowest-ranked relevant document, representing the minimum retrieval depth required to retrieve the complete set of relevant documents.

2.1 Observations

Observation 1: Incomplete Retrieval of Multilingual Relevant Documents Multilingual embedding models struggle to retrieve all semantically equivalent relevant documents across languages within the top-ranked results. As shown in Figure 2 (left), Comp@20 ranges from 51.60 to 72.10

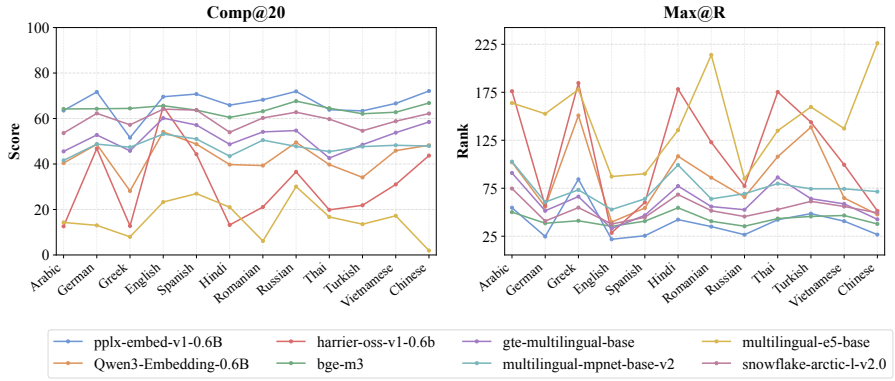

Figure 2: Retrieval performance of eight multilingual embedding models across 12 query languages on XQuAD. Left: Comp@20, where higher values indicate better performance. Right: Max@R, where lower values indicate better performance. Existing models frequently fail to retrieve all relevant documents within the top 20 results, with performance varying across query languages.

for pplx-embed-v1-0.6B and from 60.50 to 67.65 for bge-m3. For both models, at least one relevant document is missing from the top 20 results for at least 27.90% of queries in every query language. Figure 2 (right) shows that their Max@R values range from 21.8 to 84.3 and from 35.3 to 54.6, respectively, across query languages. Although each query has only 12 relevant documents, both models require substantially greater retrieval depths to retrieve all of them. The gap between the number of relevant documents and the required retrieval depth highlights how far down the ranking some relevant documents remain.

Observation 2: Retrieval Performance Varies Across Query Languages. The ability to retrieve all relevant documents varies substantially across query languages, even for the same model. For harrier-oss-v1-0.6b, Comp@20 and Max@R are 65.29 and 28.5 for English queries, compared with 12.77 and 184.6 for Greek queries. Even with a shared multilingual corpus, both the success rate of retrieving all relevant documents within the top 20 results and the depth required to retrieve them differ markedly by query language. However, smaller performance differences across query languages do not necessarily guarantee high success rates in retrieving all relevant documents. For example, bge-m3 exhibits relatively consistent performance across query languages, but its Comp@20 remains below 70 in every language. These findings show that consistency across query languages and the ability to retrieve all relevant documents are distinct aspects of retrieval performance.

Motivation. These observations highlight the need to account for multiple relevant documents across languages. For multilingual embedding models, this entails considering both the relevance of each document to the query and the relationships among semantically equivalent documents, with the aim of retrieving all of them consistently across query languages.

3 METHODS

To optimize representation learning for multilingual information retrieval, we propose MUSA, a contrastive learning method that integrates two complementary objectives: (i) a multi-positive optimization objective that enables the model to learn from multiple multilingual positive documents while reducing competition among positives, and (ii) a cross-correlation-based alignment penalty that encourages semantic invariance across semantically equivalent documents written in different languages. The overall loss $\mathcal{L}$ combines these two objectives:

$$\mathcal{L} = \mathcal{L}_{\text{LSEP}} + \lambda \mathcal{L}_{\text{BT}} \tag{1}$$

where $\mathcal{L}_{\text{LSEP}}$ denotes the LogSumExp Pairwise loss for multi-positive optimization, $\mathcal{L}_{\text{BT}}$ denotes the Barlow Twins loss for multilingual alignment, and λ controls the alignment penalty.

3.1 MULTI-POSITIVE OPTIMIZATION

The InfoNCE loss (Oord et al., 2018), widely used in contrastive learning, typically contrasts one relevant document against multiple negatives for each query. However, in realistic multilingual information retrieval scenarios, a query q can be associated with multiple relevant documents across languages. To fully exploit these multilingual positives, we employ the LogSumExp Pairwise (LSEPair) loss (Wang et al., 2026), which enables the model to learn from multiple positives while reducing competition among them. Specifically, let E_q and E_d denote the embeddings of a query q and a document d encoded by the embedding model, respectively. We define their similarity as $s(q, d) = \cos(E_q, E_d)/\tau$, where τ is a temperature parameter. Based on this similarity score, the LSEPair loss for a given query is defined as follows:

$$\mathcal{L}_{\text{LSEP}} = \log\left(1 + \sum_{n\in\mathcal{N}}\sum_{p\in\mathcal{P}} \exp\big(s(q,n) - s(q,p)\big)\right) \tag{2}$$

where $\mathcal{P}$ denotes the set of positive samples for query q, and $\mathcal{N}$ denotes the set of negative samples. This objective contrasts each positive sample against negative samples through pairwise comparisons, thereby alleviating the issue of multiple positives competing within a single softmax denominator. As a result, it reduces the tendency for one high-scoring positive sample to suppress the learning signals of other positives. For efficient computation with larger batch sizes, the loss can be expressed in an equivalent separable LogSumExp form with linear complexity:

$$\mathcal{L}_{\text{LSEP}} = \text{Softplus}\left(\log\sum_{n\in\mathcal{N}} \exp\big(s(q,n)\big) + \log\sum_{p\in\mathcal{P}} \exp\big(-s(q,p)\big)\right) \tag{3}$$

Consequently, we reformulate contrastive learning for multilingual information retrieval from one-to-one query-document optimization into a multi-positive objective over relevant documents across languages. This formulation reduces positive competition and stabilizes multi-positive learning.

3.2 MULTILINGUAL ALIGNMENT

While the preceding multi-positive objective optimizes query-document relevance, it does not explicitly align parallel positive documents written in different languages. To address this limitation, we utilize the Barlow Twins (BT) loss (Zbontar et al., 2021), which encourages semantically equivalent documents to form consistent representations across languages by aligning their dimension-wise correlation structures. Specifically, given the parallel positives for each query, we construct two distinct linguistic views from any pair of semantically equivalent documents.

Let these two batches of document embeddings be denoted as $E^{v_1}, E^{v_2} \in \mathbb{R}^{B\times D}$, where B is the batch size and D is the embedding dimension. The superscripts v_1 and v_2 represent two arbitrary linguistic views, ensuring that the alignment focuses on shared semantics rather than being tied to specific languages. For instance, a document batch of size $B = 3$ could be structured as follows: $v_1 = \{D_1^{\text{ko}}, D_2^{\text{en}}, D_3^{\text{es}}\}$ and $v_2 = \{D_1^{\text{en}}, D_2^{\text{fr}}, D_3^{\text{de}}\}$, where D_b^l denotes a document in language l, with the subscript b indicating its underlying semantic identity. Thus, documents sharing the same index b (e.g., D_1^{ko} and D_1^{en}) are semantically equivalent. We normalize each embedding dimension separately within each view to have zero mean and unit variance across all B instances in the batch, obtaining $\tilde{E}^{v_1}$ and $\tilde{E}^{v_2}$. The cross-correlation matrix $\mathcal{C} \in \mathbb{R}^{D\times D}$ and the corresponding BT loss are defined as:

$$\mathcal{C}_{ij} = \frac{1}{B}\sum_{b=1}^{B} \tilde{E}^{v_1}_{b,i}\tilde{E}^{v_2}_{b,j} \qquad \mathcal{L}_{\text{BT}} = \sum_i (1-\mathcal{C}_{ii})^2 + \mu\sum_i\sum_{j\neq i}\mathcal{C}_{ij}^2 \tag{4}$$

The first term (invariance term) drives the diagonal elements $\mathcal{C}_{ii}$ toward 1, encouraging semantically equivalent documents to have similar representations across corresponding dimensions. The second term (redundancy-reduction term), weighted by μ, penalizes the off-diagonal correlations $\mathcal{C}_{ij}$ by driving them toward 0, acting as a soft regularizer that discourages redundant correlations between distinct embedding dimensions across the two views. Consequently, the BT loss promotes semantic alignment among parallel positive documents, making their representations robust against language-specific surface forms.

4 EXPERIMENTAL SETUP

4.1 TRAINING

Models To assess the generalizability of MUSA across model scales and architectures, we evaluate it on 13 publicly available multilingual embedding models: multilingual-e5-base and multilingual-e5-large-instruct (Wang et al., 2024), paraphrase-multilingual-mpnet-base-v2 (Reimers & Gurevych, 2019), gte-multilingual-base (Zhang et al., 2024), KaLM-embedding-multilingual-mini-instruct-v2.5 (Zhao et al., 2026), bge-m3 (Chen et al., 2024), Solon-embeddings-large-0.1 (OrdalieTech, 2023), snowflake-arctic-embed-l-v2.0 (Yu et al., 2024), Qwen3-Embedding-0.6B (Zhang et al., 2025), pplx-embed-v1-0.6B (Eslami et al., 2026), harrier-oss-v1-270m (Microsoft, 2026b) and harrier-oss-v1-0.6b (Microsoft, 2026a), and llama-nemotron-embed-1b-v2 Moreira et al. (2024).

Dataset For training, we use mMARCO (Bonifacio et al., 2021), a parallel multilingual retrieval dataset covering 14 languages. We mine hard negatives from the English split and transfer the mined negative indices to the aligned instances in other languages to ensure consistent (q, p, n) pairing across multilingual versions; details are provided in Appendix B.2. As shown in Algorithm 1, we construct the training data through round-robin sampling over ordered language pairs (a, b), where $a \neq b$. For each language pair, we form a candidate pool of unused aligned instances and sample one index from the pool. The selected instance is converted into $(q_a^i, p_a^i, p_b^i, n_a^i)$, where q_a^i, p_a^i, and n_a^i are in language a, and p_b^i is the parallel positive passage in language b. The used-instance set $\mathcal{U}$ ensures that each underlying instance is selected at most once. This process continues until no language pair has an unused candidate.

Algorithm 1: MUSA Data Construction via Language-Pair Round-Robin Sampling

Require: Parallel dataset $\mathcal{M}$, languages $\mathcal{L}$
Ensure: Multi-positive dataset $\mathcal{D}$

$\mathcal{D} \leftarrow [\,]$, $\mathcal{U} \leftarrow \emptyset$
$\mathcal{T} \leftarrow \{(a, b) \mid a, b \in \mathcal{L},\ a \neq b\}$
repeat
  updated $\leftarrow$ **false**
  for all $(a, b) \in \mathcal{T}$ **do**
    $\mathcal{C}_{a,b} \leftarrow \{i \mid (q_a^i, p_a^i, p_b^i, n_a^i) \in \mathcal{M},\ i \notin \mathcal{U}\}$
    if $\mathcal{C}_{a,b} \neq \emptyset$ **then**
      Sample one index i uniformly from $\mathcal{C}_{a,b}$
      Append $(q_a^i, p_a^i, p_b^i, n_a^i)$ to $\mathcal{D}$
      $\mathcal{U} \leftarrow \mathcal{U} \cup \{i\}$
      updated $\leftarrow$ **true**
    end if
  end for
until updated $=$ **false**
return $\mathcal{D}$

4.2 EVALUATION

Benchmarks & Scenarios To evaluate our approach in realistic MLIR scenarios, we use XQuAD (Artetxe et al., 2020) and BELEBELE (Bandarkar et al., 2024) from MMTEB (Enevoldsen et al., 2025), which provide semantically equivalent queries and passages across multiple languages; our evaluation includes 12 and 18 languages from XQuAD and BELEBELE, respectively. Following Roy et al. (2020), we combine passages from all evaluated languages into a single unified corpus, requiring the model to retrieve all corresponding parallel passages for a given query regardless of language. We further include MLQA (Lewis et al., 2020a) to evaluate cross-lingual retrieval capability and use the Wikipedia dataset to measure multi-monolingual retrieval performance. A detailed list of evaluated languages is provided in Appendix B.3.

Metrics For a comprehensive assessment of the multi-positive retrieval setting, we employ three metrics. First, we use Mean Average Precision (MAP) as the standard metric to assess the ranking quality of relevant documents. Second, we adopt Complete@k (Qu et al., 2024) as a strict metric for measuring whether all parallel positive documents are retrieved within the top-k results. Finally, we use Max@R (Hong et al., 2026), which measures the cutoff rank required to retrieve the full set of relevant documents. Given a query q, let R_q denote the set of all positive documents, $\mathcal{D}_k(q)$ the top-k retrieved documents, and $\mathrm{rank}_q(d)$ the rank of document d in the retrieved list. The two metrics are defined as:

$$\mathrm{Comp@}k(q) = \mathbb{I}\left[R_q \subseteq \mathcal{D}_k(q)\right] \qquad \mathrm{Max@R}(q) = \max\left\{\mathrm{rank}_q(d) \mid d \in R_q\right\} \tag{5}$$

where $\mathbb{I}[\cdot]$ is an indicator function. We set $k = 20$ to cover the full set of parallel positives in our evaluated benchmarks, including 12 languages in XQuAD and 18 languages in BELEBELE. A lower Max@R indicates that the worst-ranked relevant document in R_q appears earlier in the ranked list, meaning that all relevant documents are retrieved at higher ranks.

Model	XQuAD (12)						BELEBELE (18)					
	Comp@20		MAP@20		Max@R (↓)		Comp@20		MAP@20		Max@R (↓)	
	Base	Ours	Base	Ours	Base	Ours	Base	Ours	Base	Ours	Base	Ours
Harrier-oss-270M-v1	16.91	**57.34**	41.06	**79.58**	180.7	**84.3**	2.88	**43.70**	24.82	**72.87**	474.1	**198.9**
mE5-base	16.02	**61.51**	45.41	**77.34**	147.0	**44.0**	5.54	**44.90**	40.44	**70.67**	376.2	**123.8**
mMPNet-base-v2	47.74	**54.73**	65.94	**72.24**	73.8	**55.0**	35.73	**48.86**	61.93	**71.25**	177.2	**122.4**
mGTE-base	51.86	**60.93**	75.63	**78.74**	60.5	**43.7**	47.91	**59.18**	73.27	**77.48**	121.6	**80.8**
KaLM-Embed-mini-v2.5	28.26	**57.44**	52.89	**77.50**	121.1	**48.3**	15.02	**47.21**	48.94	**73.50**	258.3	**125.3**
BGE-M3	64.15	**73.33**	81.97	**84.64**	42.5	**30.1**	50.07	**63.11**	76.99	**79.68**	103.8	**72.6**
mE5-large-instruct	11.69	**72.38**	29.78	**82.96**	194.4	**29.9**	3.93	**61.78**	26.38	**77.94**	370.5	**81.7**
Solon-Embed-large-0.1	58.95	**72.92**	75.78	**83.52**	47.8	**31.0**	42.10	**64.00**	72.93	**79.78**	129.1	**71.0**
Arctic-Embed-L-v2.0	59.44	**71.34**	78.00	**82.16**	53.1	**33.9**	44.83	**59.10**	73.71	**76.33**	130.0	**85.5**
Harrier-oss-0.6B-v1	30.75	**69.50**	50.04	**83.96**	112.9	**35.4**	4.13	**59.43**	28.15	**78.49**	408.5	**83.9**
Qwen3-Embed-0.6B	43.06	**50.47**	69.78	**73.47**	85.2	**67.0**	31.95	**41.45**	67.43	**70.71**	188.3	**146.3**
PPLX-Embed-0.6B-v1	66.60	**75.50**	84.06	**87.83**	39.4	**28.3**	55.98	**65.64**	80.94	**82.89**	97.5	**69.4**
Nemotron-Embed-1B-v2	81.37	**86.02**	90.75	**92.51**	24.7	**20.5**	57.99	**64.28**	81.21	**82.09**	86.6	**70.0**

Table 1: Multilingual information retrieval results on XQuAD (12 languages) and BELEBELE (18 languages). Base denotes the original embedding model, and Ours denotes the model trained with MUSA. Lower Max@R is better.

5 RESULTS

5.1 MAIN PERFORMANCE

Table 1 presents multilingual information retrieval performance in a mixed-language document pool, where all relevant documents corresponding to a given query must be retrieved. The results show that MUSA consistently improves all baseline models on both datasets. The simultaneous gains in Comp@20 and MAP@20 indicate that our method improves both the completeness and ranking quality of parallel relevant document retrieval within the top-20 results. Max@R, which measures the retrieval depth required to find all relevant documents, is also substantially reduced across most models. For example, Harrier-oss-v1-0.6B shows a large drop from 408.5 to 83.9 on BELEBELE.

These results indicate that our method improves the retrieval of positives that were previously placed far below the top results. The baseline models often fail to retrieve the complete set of relevant documents even when some of them are ranked highly, because other semantically equivalent positives still appear much lower in the ranked list. The reductions in Max@R show that these worst-ranked positives are promoted to higher ranks, while the gains in Comp@20 indicate that more complete sets of parallel relevant documents are recovered within the top-20 results. This pattern suggests that the proposed training signal improves the consistency of the positive set as a whole, rather than only strengthening already well-ranked query-document matches. Similar improvements across models with different baseline strengths, including BGE-M3, PPLX-Embed-0.6B-v1, and Nemotron-Embed-1B-v2, further suggest that the issue is not simply model capacity, but a mismatch between standard retrieval training and the multi-positive structure of multilingual retrieval.

5.2 LANGUAGE-WISE RETRIEVAL DISPARITIES

Figure 3 presents language-wise retrieval performance across 18 query languages on BELEBELE, using Comp@20 and Max@R for eight multilingual embedding models. The baseline models often fail to retrieve all parallel relevant documents within the top-ranked results and require much deeper retrieval for certain query languages. In contrast, MUSA increases Comp@20 and reduces Max@R across all 18 languages for every model shown. Specifically, for harrier-oss-v1-0.6b, Comp@20 averaged across query languages increases from 4.13 to 59.43. For Hindi queries, where the baseline requires the deepest retrieval for this model, Max@R decreases from 659.5 to 119.5. These improvements show that all relevant documents are retrieved within the top 20 results for more queries, while fewer documents need to be retrieved to recover the entire relevant set.

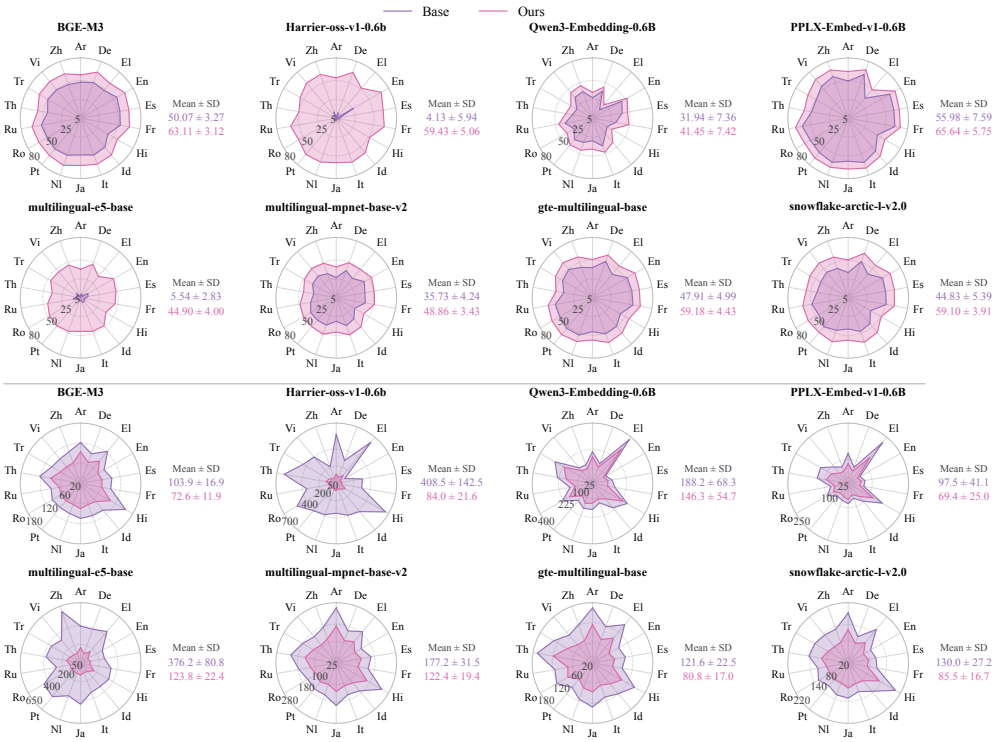

Figure 3: Language-wise retrieval performance on BELEBELE. **Top:** Comp@20, where values farther from the center indicate better retrieval performance. **Bottom:** Max@R, where values closer to the center indicate better retrieval performance; radial scales vary across models. Means and standard deviations across query languages are reported alongside each plot.

Beyond these performance improvements, MuSA narrows the gaps in retrieval depth across query languages. The standard deviation of Max@R across query languages decreases for all eight models; for harrier-oss-v1-0.6b, it falls from 142.5 to 21.6. This reduction indicates that the improvements extend beyond average performance, as the depth required to retrieve all relevant documents also becomes more consistent across query languages. The standard deviation of Comp@20 also decreases for most models, alongside higher average retrieval success rates. These results suggest that MuSA improves the retrieval of parallel relevant documents while mitigating language-wise performance disparities, reducing the extent to which access to the relevant set depends on the query language.

5.3 Cross-lingual and English Retrieval Performance

To further evaluate the effect of the proposed method on cross-lingual and monolingual retrieval, Table 2 reports NDCG@10 on MLQA and Wikipedia. On MLQA, our method improves performance for most backbone models, indicating that the learned multilingual alignment also benefits query-document matching across languages. This shows that the effect of MuSA extends beyond retrieving parallel relevant documents in the mixed-language setting. On Wikipedia, performance is largely preserved across models, with several cases showing slight improvements. This suggests that the proposed alignment objective does not substantially disrupt the retrieval representations learned by the original models. Overall, these results show that MuSA improves cross-lingual retrieval performance while largely maintaining English monolingual retrieval performance.

Model	MLQA (7)	Wikipedia (En)
Harrier-oss-270M-v1	64.72 (+7.43)	90.79 (+0.02)
mE5-base	60.67 (+1.67)	93.08 (+0.60)
mMPNet-base-v2	54.45 (+4.57)	87.90 (+0.69)
mGTE-base	64.99 (+2.41)	91.30 (+0.81)
KaLM-Embed-mini-v2.5	64.43 (-0.09)	92.78 (+0.35)
BGE-M3	66.89 (+1.88)	92.32 (-0.26)
mE5-large-instruct	64.00 (+3.16)	93.74 (+0.49)
Solon-Embed-large-0.1	64.89 (+0.52)	93.56 (-0.05)
Arctic-Embed-L-v2.0	62.47 (-0.32)	92.70 (-0.51)
Harrier-oss-0.6B-v1	66.98 (+2.49)	95.15 (+2.54)
Qwen3-Embed-0.6B	63.33 (+0.82)	93.02 (+0.49)
PPLX-Embed-0.6B-v1	71.53 (+1.07)	95.25 (+0.89)
Nemotron-Embed-1B-v2	71.83 (+0.36)	95.18 (+0.34)

Table 2: NDCG@10 on MLQA and Wikipedia (En). Parentheses indicate changes from the base models (pp).

5.4 REPRESENTATION ALIGNMENT

We analyze whether improvements in multilingual information retrieval performance are associated with better representation alignment in the embedding space. Specifically, we measure how closely parallel documents are mapped across languages.

Setup We perform this analysis on the 18 languages included in BELEBELE. To quantify multilingual representation alignment, we extend the notion of alignment from Wang & Isola (2020) to parallel documents across languages. Specifically, let $x_{i,\ell}$ denote the document in language ℓ belonging to the i-th semantic set, and let $f(\cdot)$ denote the embedding model. For each semantic set, we define the alignment score as the average squared L_2 distance between the representations of parallel documents expressed in different languages:

$$\text{Align}(i) = \mathbb{E}_{\ell_1 \neq \ell_2} \left[\|f(x_{i,\ell_1}) - f(x_{i,\ell_2})\|_2^2 \right]$$

Here, ℓ_1 and ℓ_2 denote different languages. The overall alignment score is obtained by averaging $\text{Align}(i)$ across all semantic sets. A lower alignment score indicates that parallel documents are closer in the embedding space.

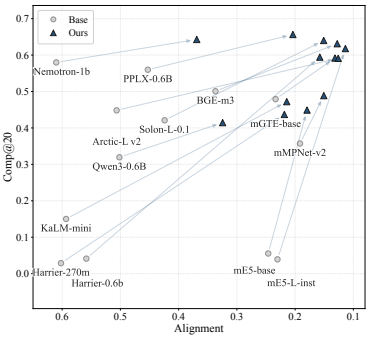

Figure 4: Alignment scores and Comp@20 on BELEBELE. Arrows point from Base to MUSA.

Results The results in Figure 4 demonstrate that, for most models, applying MUSA moves the points upward and to the right compared to their baselines. This indicates that parallel documents become closer in the embedding space while multilingual retrieval performance also improves. For example, multilingual-e5-base shows clear improvements in both alignment and Comp@20. This change is particularly pronounced for models with poor initial alignment and low retrieval performance, while relatively strong baselines also show improvements along both axes. These results suggest that the performance gains of MUSA are associated with improved alignment of multilingual parallel documents in the embedding space.

5.5 ABLATION

Method	XQuAD		BELEBELE		MLQA	Wiki.
	C@20	Max@R	C@20	Max@R	n@10	n@10
mMPNet-base-v2						
Baseline	47.74	73.80	35.73	177.18	49.88	74.83
XLCO	49.56	62.90	44.70	133.57	52.71	78.96
SupCon	51.24	60.20	45.28	132.53	52.04	78.53
MUSA	**54.73**	**55.00**	**48.86**	**122.42**	**54.45**	**80.77**
w/o BT	52.58	58.70	46.46	127.45	53.44	79.94
mGTE-base						
Baseline	51.86	60.50	47.91	121.63	62.58	84.39
XLCO	56.28	51.00	56.44	88.67	63.76	85.93
SupCon	**61.58**	44.80	58.54	80.97	64.93	**87.03**
MUSA	60.93	**43.70**	**59.18**	**80.85**	**64.99**	86.83
w/o BT	58.58	49.00	56.46	87.88	63.88	86.77
BGE-M3						
Baseline	64.15	42.50	50.07	103.85	65.01	89.85
XLCO	68.24	34.60	59.66	78.69	66.19	89.07
SupCon	72.66	30.50	60.80	73.93	66.22	90.04
MUSA	**73.33**	**30.10**	**63.11**	**72.58**	**66.89**	89.12
w/o BT	70.68	33.20	57.43	81.43	65.99	**90.52**
Qwen3-Embed-0.6B						
Baseline	43.06	85.20	31.95	188.25	62.51	88.08
XLCO	46.79	73.20	39.20	156.97	62.82	88.06
SupCon	45.13	84.40	33.84	178.32	60.15	87.73
MUSA	**50.47**	**67.00**	**41.45**	**146.25**	**63.33**	88.17
w/o BT	49.92	69.80	39.63	153.19	61.95	**88.85**
PPLX-Embed-0.6B-v1						
Baseline	66.60	39.40	55.98	97.49	70.46	91.30
XLCO	71.52	31.50	63.40	72.45	71.35	91.63
SupCon	74.78	30.30	62.85	80.05	70.94	91.49
MUSA	**75.50**	**28.30**	**65.64**	**69.38**	**71.53**	92.13
w/o BT	75.46	28.90	65.08	70.70	71.50	**92.52**

Table 3: Comparison of alignment objectives and ablation results. C@20 and n@10 denote Comp@20 and nDCG@10, respectively.

In this section, we compare MUSA with XLCO (Chi et al., 2021), a representative cross-lingual alignment objective, SupCon (Khosla et al., 2020), a supervised contrastive objective for multiple positives, and a w/o BT variant that removes the auxiliary alignment objective. As shown in Table 3, XLCO and SupCon improve over the baseline in most cases, while MUSA yields the strongest gains for multilingual and cross-lingual retrieval. For MLIR, MUSA achieves the highest Comp@20 in most cases and consistently obtains the lowest Max@R across all backbone models, indicating that directly optimizing multilingual positives is effective for retrieving the full set of parallel relevant documents. MUSA also improves or remains competitive on CLIR, while maintaining comparable performance in multi-monolingual retrieval.

The contribution of the BT objective is most evident in MLIR and CLIR. Removing BT from MUSA decreases Comp@20 and increases Max@R across all backbone models on MLIR, suggesting that multi-positive contrastive learning alone cannot sufficiently align parallel relevant documents. The auxiliary alignment objective further improves the consistency among semantically equivalent documents across languages, leading to more complete and stable multilingual retrieval. A similar tendency appears on CLIR, where w/o BT generally reduces

nDCG@10. However, this effect is less consistent in multi-monolingual retrieval, where the ablated variant sometimes performs comparably or better, possibly because the BT objective primarily encourages cross-lingual consistency among parallel documents.

6 Related Work

6.1 Multilingual Embedding Models

Recent multilingual embedding models encode dozens, or even hundreds, of languages into a shared embedding space and have substantially improved retrieval performance within individual languages (Zhang et al., 2025; Lee et al., 2025; Zhao et al., 2026; Akram et al., 2026). In particular, some recent models leverage cross-lingual parallel and multilingual data during training to reduce representational gaps across languages (Chen et al., 2024; Awasthy et al., 2025; Li et al., 2025). However, these models are primarily trained to optimize pairwise query-document relevance (Hong et al., 2026). As a result, when multilingual documents share the same meaning, these models do not explicitly ensure that such documents are consistently aligned in the embedding space regardless of language (Lee et al., 2026). Consequently, models may become more sensitive to language than to intrinsic semantics. This can lead to substantial disparities in retrieval performance across languages (Park & Lee, 2025; Goworek et al., 2025).

6.2 Multilingual Alignment for MLIR

Multilingual information retrieval (MLIR) addresses the task of retrieving semantically relevant documents from a multilingual document pool. In contrast, cross-lingual information retrieval (CLIR) typically considers a query in one language and documents in a different language, focusing on retrieval between two languages. A large body of work has studied methods for improving cross-lingual retrieval performance (Chi et al., 2021; Feng et al., 2022b; Yang et al., 2024b; Hong et al., 2026; Lee et al., 2026). These methods include contrastive learning, translation-based knowledge distillation, reverse training, and retrieval-oriented alignment objectives. Unlike this two-language setting, MLIR operates over a document pool spanning multiple languages. In this setting, retrieval models must identify relevant documents regardless of the language in which they are written (Yang et al., 2024c). Prior work has explored several directions for improving multilingual retrieval, including parallel-corpus-based contrastive learning, neural retrieval methods, knowledge-enhanced pretraining, and teacher-model distillation (Hu et al., 2023; Lawrie et al., 2023; Yang et al., 2024a). Within this line of work, multilingual query and document representations have been aligned to an English teacher embedding space through distillation jointly optimized with cross-lingual contrastive learning (Jang et al., 2026a).

In this work, we formulate multilingual relevant documents associated with each query as a unified positive set for retrieval learning. We jointly optimize query–document relevance over this set and directly align the embeddings of the same parallel positive documents through a separate dimension-wise cross-correlation objective. This formulation encourages more consistent retrieval of semantically equivalent documents and mitigates language-dependent ranking disparities.

7 Conclusion

In this paper, we introduce MuSA, a training method for improving multilingual information retrieval by reducing language-dependent ranking disparities among relevant documents across languages. We showed that multilingual embedding models can exhibit uneven rankings among parallel relevant documents when documents from multiple languages coexist in the same retrieval pool. To address this issue, MuSA reformulates multilingual retrieval as a multi-positive learning problem and encourages cross-language relevant documents to form more consistent representations. Experiments across diverse multilingual embedding models and retrieval benchmarks demonstrate that our method improves the retrieval of parallel relevant documents while largely preserving monolingual retrieval performance. Further analyses show that these gains are closely associated with improved representation alignment among parallel documents. Overall, our findings highlight the importance of modeling relationships among multilingual positives, rather than optimizing each query-document pair independently, for building more reliable multilingual retrieval systems.

AI USE STATEMENT

We used generative AI tools to assist with translating parts of our text into English. All coauthors reviewed the AI-assisted translations to ensure that they accurately reflected the intended meaning and were clearly expressed. The final wording was reviewed and approved by all coauthors. We take full responsibility for the content of the final manuscript, including all AI-assisted text.

ETHICS STATEMENT

This work uses publicly available datasets and models, and does not involve collecting personal information or conducting studies with human subjects. The goal of MUSA is to improve multilingual information retrieval by reducing language-dependent ranking disparities among relevant documents across languages. However, multilingual embedding models may inherit biases or uneven language coverage from their pretraining data. Improved alignment across languages does not necessarily eliminate these underlying biases or limitations. Moreover, better retrieval consistency across the evaluated languages does not guarantee equal performance for all languages, especially low-resource languages. Real-world deployment should therefore be accompanied by careful validation across target languages, domains, and user groups.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide detailed descriptions of the datasets, evaluation protocols, and implementation settings used in our experiments in Section 4. Additional details on training data construction, training configurations, hyperparameters, and computational resources are provided in Appendix B.

REFERENCES

Mohammad Kalim Akram, Saba Sturua, Nastia Havriushenko, Quentin Herreros, Michael Günther, Maximilian Werk, and Han Xiao. jina-embeddings-v5-text: Task-targeted embedding distillation. *arXiv preprint arXiv:2602.15547*, 2026.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4623–4637, 2020.

Akari Asai, Xinyan Yu, Jungo Kasai, and Hanna Hajishirzi. One question answering model for many languages with cross-lingual dense passage retrieval. *Advances in Neural Information Processing Systems*, 34:7547–7560, 2021.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations*, volume 2024, pp. 9112–9141, 2024.

Parul Awasthy, Aashka Trivedi, Yulong Li, Meet Doshi, Riyaz Bhat, Vishwajeet Kumar, Yushu Yang, Bhavani Iyer, Abraham Daniels, Rudra Murthy, et al. Granite embedding r2 models. *arXiv preprint arXiv:2508.21085*, 2025.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 749–775, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.44. URL https://aclanthology.org/2024.acl-long.44/.

Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. mmarco: A multilingual version of the ms marco passage ranking dataset. *arXiv preprint arXiv:2108.13897*, 2021.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024.

Zehui Chen, Kuikun Liu, Qiuchen Wang, Jiangning Liu, Wenwei Zhang, Kai Chen, and Feng Zhao. Mindsearch: Mimicking human minds elicits deep ai searcher, 2025. URL https://arxiv.org/abs/2407.20183.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, He-Yan Huang, and Ming Zhou. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3576–3588, 2021.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, et al. Mmteb: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*, 2025.

Sedigheh Eslami, Maksim Gaiduk, Markus Krimmel, Louis Milliken, Bo Wang, and Denis Bykov. Diffusion-pretrained dense and contextual embeddings. *arXiv preprint arXiv:2602.11151*, 2026.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic bert sentence embedding. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 878–891, 2022a.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 878–891, Dublin, Ireland, May 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.62. URL https://aclanthology.org/2022.acl-long.62/.

Luyu Gao, Yunyi Zhang, Jiawei Han, and Jamie Callan. Scaling deep contrastive learning batch size under memory limited setup. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pp. 316–321, 2021.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, Haofen Wang, et al. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1):32, 2023.

Roksana Goworek, Olivia Macmillan-Scott, and Eda B Özyiğit. What drives cross-lingual ranking? retrieval approaches with multilingual language models. *arXiv preprint arXiv:2511.19324*, 2025.

Seongtae Hong, Youngjoon Jang, Jungseob Lee, Hyeonseok Moon, and Heuiseok Lim. Improving semantic proximity in information retrieval through cross-lingual alignment. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=NvKvW5k6Kk.

Xiyang Hu, Xinchi Chen, Peng Qi, Deguang Kong, Kunlun Liu, William Yang Wang, and Zhiheng Huang. Language agnostic multilingual information retrieval with contrastive learning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 9133–9146, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.581. URL https://aclanthology.org/2023.findings-acl.581/.

Youngjoon Jang, Seongtae Hong, and Heuiseok Lim. Mimo: Multilingual information retrieval via monolingual objectives. *arXiv preprint arXiv:2605.31171*, 2026a.

Youngjoon Jang, Seongtae Hong, Hyeonseok Moon, and Heuiseok Lim. Mlaire: Multilingual language-aware information retrieval evaluation protocal. *arXiv preprint arXiv:2605.07249*, 2026b.

Youngjoon Jang, Seongtae Hong, Hyeonseok Moon, and Heuiseok Lim. Shift: Semantic harmonization via index-side feature transformation for multilingual information retrieval. *arXiv preprint arXiv:2606.18801*, 2026c.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, et al. Matryoshka representation learning. *Advances in Neural Information Processing Systems*, 35:30233–30249, 2022.

Dawn Lawrie, Eugene Yang, Douglas W Oard, and James Mayfield. Neural approaches to multilingual information retrieval. In *European Conference on Information Retrieval*, pp. 521–536. Springer, 2023.

Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, et al. Gemini embedding: Generalizable embeddings from gemini. *arXiv preprint arXiv:2503.07891*, 2025.

Seungyoon Lee, Minhyuk Kim, Seongtae Hong, Youngjoon Jang, Dongsuk Oh, and Heuiseok Lim. Clear: Cross-lingual enhancement in alignment via reverse-training, 2026. URL https://arxiv.org/abs/2604.05821.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. Mlqa: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 7315–7330, 2020a.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020b.

Shiyu Li, Yang Tang, Ruijie Liu, Shi-Zhe Chen, and Xi Chen. Conan-embedding-v2: Training an LLM from scratch for text embeddings. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 15000–15016, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.758. URL https://aclanthology.org/2025.emnlp-main.758/.

Microsoft. harrier-oss-v1-0.6b, 2026a. URL https://huggingface.co/microsoft/harrier-oss-v1-0.6b.

Microsoft. harrier-oss-v1-270m, 2026b. URL https://huggingface.co/microsoft/harrier-oss-v1-270m.

Gabriel de Souza P Moreira, Radek Osmulski, Mengyao Xu, Ronay Ak, Benedikt Schifferer, and Even Oldridge. Nv-retriever: Improving text embedding models with effective hard-negative mining. *arXiv preprint arXiv:2407.15831*, 2024.

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

OrdalieTech. Solon-embeddings-large-0.1, 2023. URL https://huggingface.co/OrdalieTech/Solon-embeddings-large-0.1.

Jeonghyun Park and Hwanhee Lee. Investigating language preference of multilingual RAG systems. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5647–5675, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.295. URL https://aclanthology.org/2025.findings-acl.295/.

Changle Qu, Sunhao Dai, Xiaochi Wei, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Jun Xu, and Ji-Rong Wen. Towards completeness-oriented tool retrieval for large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 1930–1940, 2024.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL http://arxiv.org/abs/1908.10084.

Uma Roy, Noah Constant, Rami Al-Rfou, Aditya Barua, Aaron Phillips, and Yinfei Yang. LAReQA: Language-agnostic answer retrieval from a multilingual pool. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5919–5930, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.477. URL https://aclanthology.org/2020.emnlp-main.477/.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 10014–10037, 2023.

Benben Wang, Minghao Tang, Hengran Zhang, Jiafeng Guo, and Keping Bi. Training dense retrievers with multiple positive passages. *arXiv preprint arXiv:2602.12727*, 2026.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*, 2024.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pp. 9929–9939. PMLR, 2020.

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

Eugene Yang, Dawn Lawrie, and James Mayfield. Distillation for multilingual information retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2368–2373, 2024a.

Eugene Yang, Dawn Lawrie, James Mayfield, Douglas W. Oard, and Scott Miller. Translate-distill: Learning cross-language dense retrieval by translation and distillation, 2024b. URL https://arxiv.org/abs/2401.04810.

Jinrui Yang, Fan Jiang, and Timothy Baldwin. Language bias in multilingual information retrieval: The nature of the beast and mitigation methods. In Jonne Sälevä and Abraham Owodunni (eds.), *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pp. 280–292, Miami, Florida, USA, November 2024c. Association for Computational Linguistics. doi: 10.18653/v1/2024.mrl-1.23. URL https://aclanthology.org/2024.mrl-1.23/.

Puxuan Yu, Luke Merrick, Gaurav Nuti, and Daniel Campos. Arctic-embed 2.0: Multilingual retrieval without compromise. *arXiv preprint arXiv:2412.04506*, 2024.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pp. 12310–12320. PMLR, 2021.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, et al. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1393–1412, 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.

Xinping Zhao, Xinshuo Hu, Zifei Shan, Shouzheng Huang, Yao Zhou, Xin Zhang, Zetian Sun, zhenyu liu, Dongfang Li, Xinyuan Wei, Youcheng Pan, Yang Xiang, Meishan Zhang, Haofen Wang, Jun Yu, Baotian Hu, and Min Zhang. KaLM-embedding-v2: Superior training techniques and data inspire a versatile embedding model. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Y7qzhvWhcz.

A LIMITATIONS

A limitation of MUSA is that it requires multilingual positives associated with the same query. In our experiments, these positives are obtained from publicly available parallel retrieval datasets, where semantically corresponding documents across languages are already provided. However, such supervision may be difficult to obtain in real-world domains or low-resource languages. In practice, applying the method to low-resource languages not covered by existing parallel datasets may require constructing multilingual positives through translation or other data augmentation procedures. Extending MUSA to settings with noisy, incomplete, or weakly aligned multilingual positives remains an important direction for future work.

Our main evaluation focuses on retrieval settings in which parallel relevant documents coexist across multiple languages, using the language configurations provided by the selected benchmarks. In real-world multilingual retrieval, document pools may vary in their language combinations and the proportion of documents in each language, and equivalent content may not be available in every language. Our evaluation therefore does not cover all such document pool configurations. Nevertheless, we use parallel settings with semantically corresponding documents across languages to examine whether all relevant documents are included in the top-ranked results regardless of their language and how this retrieval performance varies across query languages.

B EXPERIMENTAL DETAILS

B.1 MODELS

To assess the generalizability of MUSA across model scales and architectures, we evaluate it on 13 publicly available multilingual embedding models. We summarize the evaluated models below, focusing on their architectures, language coverage, sequence lengths, and retrieval-relevant characteristics. For both training and evaluation, we follow the original retrieval instruction format of each model: query and passage prompts are applied when provided by the model, and omitted otherwise. The prompt formats used for all models are summarized in Table 4.

Model	Query	Passage
Harrier-oss-v1-270m	`Instruct: Given a web search query, retrieve relevant passages that answer the query \nQuery:`	-
KaLM-Embed-mini-v2.5	`Instruct: Given a query, retrieve documents that answer the query \n Query:`	-
Solon-Embed-large-0.1	`query:`	-
Arctic-Embed-l-v2.0	`query:`	-
Harrier-oss-v1-0.6b	`Instruct: Given a web search query, retrieve relevant passages that answer the query \nQuery:`	-
Qwen3-Embed-0.6B	`Instruct: Given a web search query, retrieve relevant passages that answer the query\nQuery:`	-
Nemotron-Embed-1b-v2	`query:`	`passage:`

Table 4: Sentence Transformers prompt configurations for each evaluated multilingual embedding model based on their official configurations. Empty fields (−) indicate that the raw text is used without any prefix.

• **Harrier-oss-v1-270m**: A compact multilingual embedding model with a maximum sequence length of 32,768 tokens.

• **mE5-base**: Based on the XLM-RoBERTa-base architecture, this model supports 94 languages and is trained with weakly supervised multilingual retrieval data. It supports a maximum sequence length of 512 tokens.

• **mMPNet-base-v2**: Built upon an XLM-RoBERTa based MPNet architecture, this model supports 50 languages. It supports a maximum sequence length of 512 tokens.

• **mGTE-base**: A multilingual Transformer-based embedding model that supports 75 languages. It supports Matryoshka Representation Learning (Kusupati et al., 2022) (MRL) and a maximum sequence length of 8,192 tokens.

• **KaLM-Embed-mini-v2.5**: A multilingual embedding model that uses bidirectional attention and mean pooling. It supports dimension reduction through MRL and a maximum sequence length of 32,768 tokens.

• **BGE-m3**: A multilingual embedding model based on a modified XLM-RoBERTa architecture. It supports more than 100 languages, dense retrieval, multi-vector retrieval, and sparse retrieval, with a maximum sequence length of 8,192 tokens.

• **mE5-large-instruct**: Built upon the XLM-RoBERTa-large architecture, this model supports over 100 languages and is instruction-tuned for retrieval tasks. It supports a maximum sequence length of 512 tokens.

• **Solon-Embed-large-0.1**: An XLM-RoBERTa-large-based embedding model trained for French and English retrieval settings. It supports a maximum sequence length of 512 tokens.

• **Arctic-Embed-l-v2.0**: A multilingual embedding model based on a modified XLM-RoBERTa architecture. It supports 74 languages, MRL, and a maximum sequence length of 8,192 tokens.

• **Harrier-oss-v1-0.6b**: A multilingual embedding model from the Harrier model family with a maximum sequence length of 32,768 tokens.

• **Qwen3-Embed-0.6B**: Adapted from the Qwen3 model family, this embedding model supports over 100 natural and programming languages. It uses RoPE and GQA, and has a maximum sequence length of 32,768 tokens.

• **PPLX-Embed-v1-0.6B**: A multilingual embedding model that supports a maximum sequence length of 32,768 tokens.

• **Nemotron-Embed-1b-v2**: A Llama-3-based embedding model for long-context retrieval. It supports MRL and a maximum sequence length of 8,192 tokens.

B.2 TRAINING

Dataset. We train our models using mMARCO (Bonifacio et al., 2021). We use the released small training split for 14 languages[1]: Ar, De, En, Es, Fr, Hi, Id, It, Ja, Nl, Pt, Ru, Vi, and Zh; language codes follow Table 5. This split contains approximately 420K aligned query–positive instances across the 14 languages. We perform hard negative mining using a top-k retrieval strategy with a relative threshold to the positive score. We encode English queries and passages with `bge-large-en-v1.5`[2] (Xiao et al., 2023) and retrieve the top-200 candidate passages for each query. We select the highest-ranked candidate whose similarity score is no greater than 90% of the query–positive score, with a random fallback using seed 42 when no candidate satisfies this condition. Since the mMARCO instances are aligned across languages, the mined English negative index is reused for the corresponding aligned instance in each language, ensuring consistent (q, p, n) pairing across multilingual versions. To adapt the mined data to our training objective, we reorganize the aligned query–positive–negative instances into the multi-positive training format described in Algorithm 1.

Code	Language	Code	Language	Code	Language
Ar	Arabic	De	German	El	Greek
En	English	Es	Spanish	Fr	French
Hi	Hindi	Id	Indonesian	It	Italian
Ja	Japanese	Nl	Dutch	Pt	Portuguese
Ro	Romanian	Ru	Russian	Th	Thai
Tr	Turkish	Vi	Vietnamese	Zh	Chinese

Table 5: Mapping between language codes and language names used in our experiments.

Hyperparams. All models are trained for one epoch with a maximum sequence length of 512 and a global batch size of 2,048. Training is conducted on 8 NVIDIA A6000 GPUs, each with 48GB memory. We use mixed-precision training with FP16, a warmup ratio of 0.1, and a weight decay of 0.01. The temperature scaling factor is set to 50. For the loss weights, we set the BT regularization coefficient to 5×10^{-5} and the alignment loss weight to 0.02. The masking margin for self-guided masking is set to 0.05. The learning rate is set to 2×10^{-6} for most models, while models in the Harrier family are trained with a larger learning rate of 2×10^{-5}.

[1] `https://huggingface.co/datasets/unicamp-dl/mmarco`

[2] `https://huggingface.co/BAAI/bge-large-en-v1.5`

Usage	Dataset	# Lang.	Task	Languages
Training	mMARCO	14	–	Ar, De, En, Es, Fr, Hi, Id, It, Ja, Nl, Pt, Ru, Vi, Zh
Evaluation	XQuAD	12	MLIR	Ar, De, El, En, Es, Hi, Ro, Ru, Th, Tr, Vi, Zh
	BELEBELE	18	MLIR	Ar, De, El, En, Es, Fr, Hi, Id, It, Ja, Nl, Pt, Ro, Ru, Th, Tr, Vi, Zh
	MLQA	7	CLIR	Ar, De, En, Es, Hi, Vi, Zh
	WIKIPEDIA	6	Multi-Mono	De, En, Hi, It, Nl, Pt

Table 6: Datasets and language subsets used for training and evaluation. Language identifiers are presented using two-letter display codes for consistency.

Details. We use gradient caching (Gao et al., 2021), where each GPU processes 256 examples, resulting in a global batch size of 2,048 across 8 GPUs. To enlarge the pool of in-batch negatives, we gather representations across devices during contrastive learning. Following prior work, we apply self-guided masking to reduce the effect of potential false negatives among in-batch negatives (Eslami et al., 2026; Zhang et al., 2025): since each query has two positive passages, p_a and p_b, we use the larger positive similarity, $\max(\mathrm{sim}(q_a, p_a), \mathrm{sim}(q_a, p_b))$, as the reference score and set the masking margin to $m = 0.05$. We mask any in-batch negative n_j when $\mathrm{sim}(q_a, n_j) > \max(\mathrm{sim}(q_a, p_a), \mathrm{sim}(q_a, p_b)) - m$, while keeping the explicitly mined negative unmasked.

B.3 EVALUATION

We construct our MLIR evaluation benchmarks from multilingual datasets with parallel instances. This parallel structure enables controlled comparisons across language versions while preserving the same underlying content. XQuAD (Artetxe et al., 2020) provides manually translated query–passage pairs across multiple languages, while BELEBELE (Bandarkar et al., 2024) provides massively multilingual parallel query–passage instances across 122 languages. For our experiments, we use selected language subsets from each dataset. Each query is paired with its corresponding passage, which is used as the relevant document. We then combine passages from all evaluated languages into a single pooled corpus, requiring the model to retrieve all parallel positives for a given query. Table 6 summarizes the language subsets used in our experiments, and Table 7 reports the statistics of the constructed MLIR benchmarks. For MLQA[3] and Wikipedia[4], we follow the retrieval benchmark format provided by MMTEB (Enevoldsen et al., 2025) without additional reconstruction.

Dataset	# Queries / Lang.	Pooled Corpus	Pos./Query
BELEBELE	900	8,784	18
XQuAD	1,190	2,880	12

Table 7: Statistics of the constructed MLIR benchmarks. # Queries / Lang. denotes the number of queries per language, Pooled Corpus denotes the shared multilingual passage pool, and Pos./Query denotes the number of parallel positives per query.

C COMPARISON OBJECTIVES

We compare MUSA with two representative objectives that capture different aspects of multilingual retrieval learning.

XLCO (Chi et al., 2021) is used as a cross-lingual contrastive objective. It applies the InfoNCE objective to cross-lingual query–positive pairs, where the query and positive passage are written in different languages. This encourages the model to align query and passage representations across language boundaries through contrastive learning. In our setup, XLCO is trained with cross-lingual pairs of the form (q_a, p_b, n_a).

SupCon (Khosla et al., 2020) is used as a multi-positive contrastive objective. It uses the query as an anchor and treats the same-language positive and the cross-lingual positive as two positives for the query. The similarities to these positives are averaged to form the positive score, which is

[3] https://huggingface.co/datasets/mteb/MLQARetrieval

[4] https://huggingface.co/datasets/mteb/WikipediaRetrievalMultilingual

then contrasted against in-batch candidates and the mined negative passage. In our setup, SupCon is trained with tuples of the form (q_a, p_a, p_b, n_a).

D T-SNE VISUALIZATION

Figure 5 provides a qualitative example of how our method changes the local embedding geometry for multilingual retrieval. We use the same BELEBELE instance and Harrier-oss-v1-0.6B model for both the baseline and our method, and visualize query and document embeddings from four languages: Arabic, English, Spanish, and Chinese. For each model, we project the corresponding query embeddings and multilingual positive document embeddings into two dimensions using t-SNE. However, because t-SNE is a nonlinear dimensionality reduction method, its two-dimensional layout should not be interpreted as direct quantitative evidence. In particular, the visual arrangement can vary depending on hyperparameters such as perplexity and random initialization. Therefore, we repeat the visualization under multiple t-SNE configurations while keeping the underlying embeddings, model, languages, and selected example fixed to verify that the qualitative pattern is not merely an artifact of a single configuration.

The upper two rows of Figure 5 vary the perplexity value among 30, 50, and 70 with a fixed random seed of 42, where the first and second rows correspond to the baseline and our method, respectively. The lower two rows vary the random seed among 0, 42, and 123 with a fixed perplexity of 50, again comparing the baseline and our method. Each panel also includes a zoomed inset around the multilingual queries and their corresponding positive documents to make the local alignment pattern easier to inspect. Across these settings, the absolute layout of points changes, as expected from t-SNE. Nevertheless, we consistently observe the same qualitative trend: compared to the baseline, our method places the multilingual query representations and their corresponding positive documents in a more compact local neighborhood in the two-dimensional projections. This suggests that the observed qualitative pattern remains consistent across the tested perplexity values and random seeds.

We emphasize that the t-SNE plots are used only as qualitative evidence to illustrate the representational behavior of the models. The main conclusions of our work are supported by the retrieval results reported in the main experiments.

E RESULTS

The detailed language-wise results for BELEBELE are provided as follows:

- **Comp@20**: Table 8
- **MAP@20**: Table 9
- **Max@R**: Table 10

The detailed language-wise results for XQuAD are provided as follows:

- **Comp@20**: Table 11
- **MAP@20**: Table 12
- **Max@R**: Table 13

The detailed language-wise results for objective comparison and ablation are provided as follows:

- **BELEBELE**: Table 14
- **XQuAD**: Table 15
- **Wikipedia**: Table 16

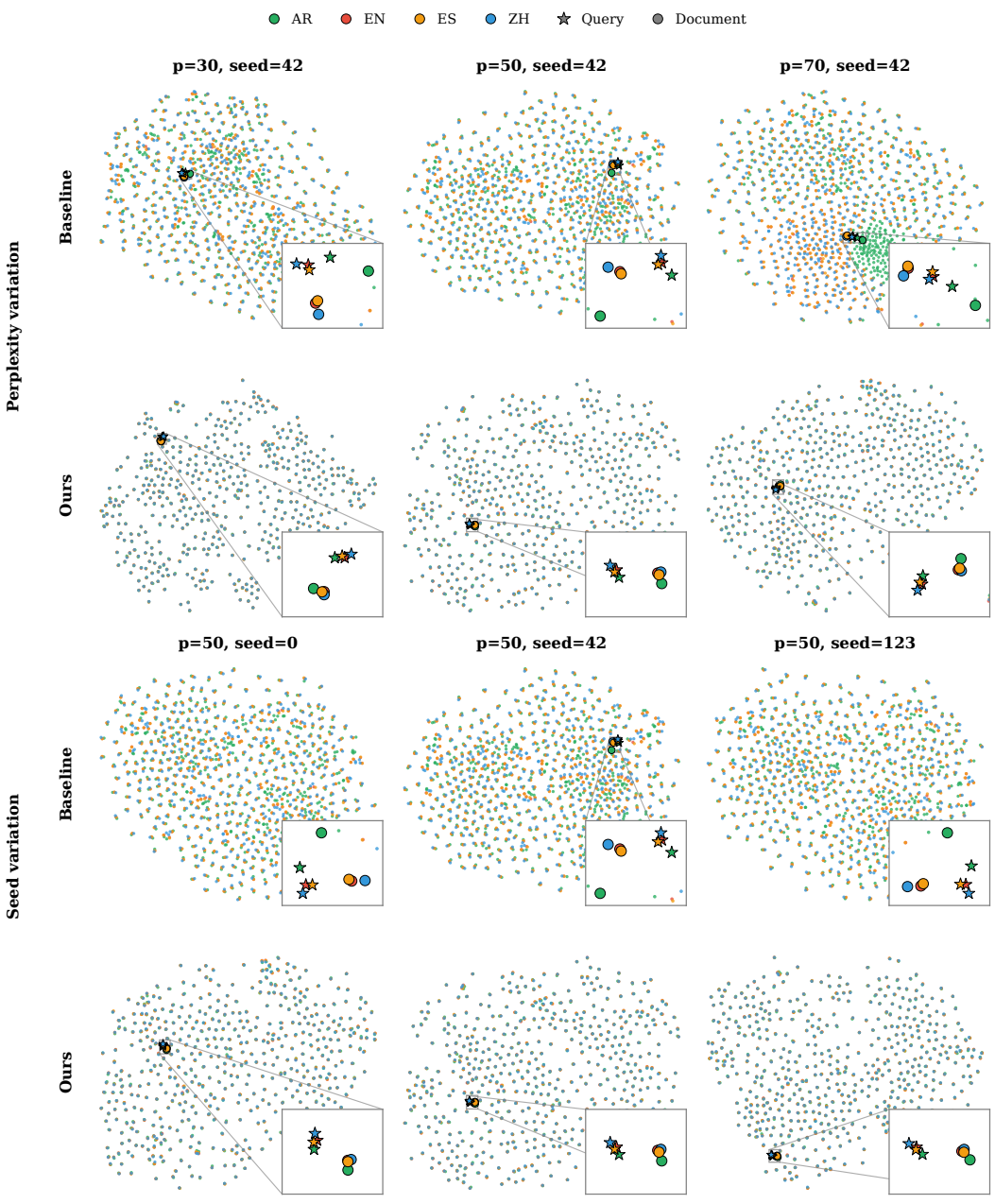

Figure 5: Robustness analysis of the t-SNE visualization under different perplexity values and random seeds. The upper two rows vary the perplexity while fixing the random seed to 42, and the lower two rows vary the random seed while fixing the perplexity to 50. Each panel includes a zoomed inset around the multilingual queries and their corresponding positive documents. Although the absolute t-SNE layouts vary across settings, the qualitative pattern remains consistent: our method places multilingual queries and their corresponding positive documents in a more compact neighborhood than the baseline.

Model		Ar	De	El	En	Fr	Hi	Id	It	Ja	Nl	Pt	Ro	Ru	Es	Th	Tr	Vi	Zh	Avg.
Harrier-oss-270M-v1	base	0.22	3.00	0.00	26.44	2.78	0.67	1.11	1.67	0.67	1.11	1.89	0.33	1.22	2.00	0.00	0.67	1.56	6.56	2.88
	ours	39.11	47.44	38.22	51.33	49.56	38.33	44.22	46.33	43.89	45.33	46.67	41.00	45.56	46.56	35.22	41.56	43.22	43.00	43.70
mE5-base	base	3.78	5.33	2.44	11.22	6.00	5.67	7.11	6.44	2.56	4.56	2.67	3.22	10.11	9.89	5.89	4.78	7.67	0.33	5.54
	ours	37.89	47.22	36.89	50.44	46.44	38.67	48.33	47.22	44.56	47.22	48.56	47.56	44.22	47.56	38.22	45.89	45.00	46.22	44.90
mMPNet-base-v2	base	26.89	38.33	33.44	42.89	39.22	28.00	36.44	39.33	31.11	38.56	38.78	39.22	34.33	38.56	30.33	36.22	37.56	34.00	35.73
	ours	41.00	49.44	49.33	54.11	51.56	42.22	50.67	52.33	45.44	51.33	50.22	50.22	48.78	51.00	43.89	48.56	50.78	48.67	48.86
mGTE-base	base	40.00	51.22	40.56	53.33	54.56	42.67	49.44	50.89	44.89	52.56	53.11	50.11	50.00	52.89	38.67	45.78	48.89	42.89	47.91
	ours	50.78	60.22	54.89	65.22	64.33	55.78	60.89	63.11	56.00	61.56	63.78	61.89	59.56	63.11	49.22	56.67	60.78	57.44	59.18
KaLM-Embed-mini-v2.5	base	14.89	20.33	1.22	39.89	19.00	4.78	10.78	13.89	13.44	8.44	12.22	3.78	18.00	23.33	7.78	3.44	19.78	35.44	15.02
	ours	41.67	50.67	29.78	57.22	55.11	34.56	48.22	52.33	46.56	50.22	52.78	45.67	52.67	52.78	39.56	37.89	50.67	51.33	47.21
BGE-M3	base	47.78	50.33	48.22	55.56	53.44	41.89	49.89	51.89	48.44	52.22	53.11	51.22	52.89	52.89	44.44	50.22	49.33	47.56	50.07
	ours	57.56	65.00	60.44	67.67	65.89	56.11	63.56	64.78	62.56	66.44	64.44	63.89	65.33	65.00	57.56	63.33	63.67	62.67	63.11
mE5-large-instruct	base	0.56	4.44	0.44	26.22	3.44	0.11	4.11	4.11	1.56	4.44	4.33	4.56	3.22	4.67	0.11	2.11	1.78	0.56	3.93
	ours	56.56	63.33	59.56	67.67	64.33	54.44	62.67	65.67	61.11	64.33	62.11	62.78	61.89	62.56	58.11	60.00	62.89	62.11	61.78
Solon-Embed-large-0.1	base	33.67	47.56	37.89	48.67	51.11	37.78	45.44	48.22	40.67	48.00	45.89	44.11	44.22	49.56	36.11	41.89	42.44	14.56	42.10
	ours	59.33	65.67	61.56	69.44	67.67	55.89	65.44	66.11	63.00	64.56	64.33	66.56	65.44	64.00	60.00	64.33	65.89	62.78	64.00
Arctic-Embed-L-v2.0	base	33.44	51.22	40.78	52.67	51.22	34.22	44.78	47.44	41.33	46.00	50.00	45.00	48.78	50.33	42.00	42.00	44.22	41.44	44.83
	ours	50.22	62.78	56.11	64.44	62.22	50.44	60.00	62.78	56.22	62.00	60.00	61.11	61.00	62.22	56.44	58.44	60.11	57.33	59.10
Harrier-oss-0.6B-v1	base	0.33	7.78	0.33	26.56	4.78	0.11	2.33	2.67	2.33	3.44	1.89	0.67	2.22	5.11	0.44	2.33	2.11	8.89	4.13
	ours	53.56	64.78	52.00	68.89	64.67	51.11	59.78	61.89	58.89	61.67	62.67	57.33	61.22	63.11	50.11	55.22	61.11	61.78	59.43
Qwen3-Embed-0.6B	base	27.22	38.22	14.56	44.67	31.56	22.22	26.78	39.00	29.78	34.00	37.22	28.22	36.78	40.00	27.11	23.89	36.11	37.67	31.95
	ours	34.22	44.00	23.33	53.11	49.11	29.22	41.89	47.89	40.89	45.00	46.00	39.11	45.56	47.00	35.89	33.22	44.56	46.11	41.45
PPLX-Embed-0.6B-v1	base	49.33	61.78	34.67	63.89	62.89	44.56	56.56	62.22	57.00	62.11	59.44	58.00	61.78	61.56	49.11	48.89	55.00	58.78	55.98
	ours	62.00	68.44	49.22	72.00	70.22	56.11	65.78	70.22	67.33	69.22	68.89	66.56	70.78	68.78	59.11	61.56	67.22	68.11	65.64
Nemotron-Embed-1B-v2	base	48.00	62.78	50.78	70.78	62.11	48.44	56.67	62.11	59.11	61.22	61.67	58.67	62.89	60.22	46.67	53.44	58.89	59.33	57.99
	ours	54.67	68.44	55.67	75.22	68.67	54.56	64.33	69.89	63.78	65.44	68.78	64.67	68.33	67.78	53.67	60.67	67.56	64.89	64.28

Table 8: BELEBELE Comp@20 results across languages.

Model		Ar	De	El	En	Fr	Hi	Id	It	Ja	Nl	Pt	Ro	Ru	Es	Th	Tr	Vi	Zh	Avg.
Harrier-oss-270M-v1	base	10.84	34.85	6.83	66.57	34.01	11.14	27.66	28.79	12.67	26.13	29.22	17.47	19.44	33.16	7.77	19.70	23.03	37.42	24.82
	ours	67.36	76.02	68.01	80.46	79.37	67.22	71.94	77.67	72.59	75.21	75.75	70.20	74.26	76.59	62.93	70.30	73.61	72.08	72.87
mE5-base	base	28.96	47.77	24.72	60.60	48.60	33.76	49.32	51.20	24.51	45.80	40.03	41.18	45.08	54.88	34.23	41.87	42.17	13.16	40.44
	ours	62.67	73.65	64.05	78.23	74.09	62.82	72.79	75.08	68.11	72.72	74.06	72.90	71.97	74.17	62.85	70.12	71.01	70.70	70.67
mMPNet-base-v2	base	49.87	65.50	58.39	70.34	66.71	50.69	64.25	66.70	56.37	66.36	65.48	66.61	61.72	65.33	54.32	62.35	64.75	58.95	61.93
	ours	61.32	74.05	69.41	76.57	75.78	62.48	73.06	74.79	67.93	74.50	73.66	74.74	72.62	72.72	64.82	70.77	73.36	70.00	71.25
mGTE-base	base	65.94	75.17	69.51	78.40	78.42	69.46	73.70	75.52	71.31	75.30	77.90	74.84	75.09	76.89	64.99	71.34	73.84	71.21	73.27
	ours	71.00	78.86	74.11	83.00	81.85	73.11	78.06	79.75	76.15	79.12	81.75	77.81	78.93	80.27	68.89	75.69	78.99	77.26	77.48
KaLM-Embed-mini-v2.5	base	45.01	61.70	12.23	80.36	64.84	26.38	46.23	53.79	49.17	43.74	52.93	29.45	58.95	67.66	33.21	23.13	54.94	77.11	48.94
	ours	67.31	77.48	50.68	84.37	81.15	60.07	74.08	80.03	74.46	75.84	80.55	70.41	79.70	80.44	66.23	62.69	76.91	80.57	73.50
BGE-M3	base	72.27	79.21	73.88	81.21	80.09	69.29	77.82	79.61	75.80	80.15	80.16	78.63	79.70	78.52	71.64	77.34	76.44	74.11	76.99
	ours	74.94	80.98	77.77	82.62	81.80	72.88	80.53	81.57	78.94	81.90	80.83	81.47	81.79	81.21	75.16	80.44	80.09	79.30	79.68
mE5-large-instruct	base	10.73	35.69	9.81	65.73	34.88	6.35	30.90	35.73	15.80	37.68	35.44	34.04	25.20	36.08	8.21	22.14	17.38	13.02	26.38
	ours	73.71	80.09	74.94	82.91	79.86	70.93	78.57	80.33	76.40	80.61	79.42	79.34	78.88	79.04	74.28	76.90	78.89	77.88	77.94
Solon-Embed-large-0.1	base	65.35	78.48	71.33	78.29	80.06	65.82	77.06	79.15	71.92	79.10	78.01	77.97	73.77	78.46	64.96	73.51	74.22	45.21	72.93
	ours	75.25	81.31	77.33	84.35	82.03	72.58	81.08	81.78	78.28	81.70	81.96	81.70	81.47	81.01	75.58	79.12	80.57	78.99	79.78
Arctic-Embed-L-v2.0	base	62.11	79.80	68.17	80.96	79.53	62.89	73.64	78.92	70.80	76.78	78.56	75.97	77.12	79.03	69.73	71.63	71.26	69.92	73.71
	ours	67.20	79.43	73.44	81.33	79.85	67.23	77.07	79.89	74.34	78.72	78.60	78.39	78.38	78.80	72.71	76.27	76.77	75.55	76.33
Harrier-oss-0.6B-v1	base	9.11	42.33	9.25	66.42	38.55	7.75	29.33	37.28	20.69	32.74	29.87	24.35	20.01	40.53	10.43	25.74	22.38	39.91	28.15
	ours	72.92	82.40	71.09	85.72	83.26	70.99	79.07	82.45	78.60	80.81	82.39	76.09	80.84	81.86	70.33	74.62	79.59	79.76	78.49
Qwen3-Embed-0.6B	base	62.94	74.65	41.02	80.50	69.36	58.13	61.68	75.31	68.07	70.68	75.21	63.50	72.52	76.08	58.92	58.23	73.02	73.92	67.43
	ours	64.07	75.44	48.08	81.66	78.47	56.94	73.44	77.07	72.07	73.28	77.10	67.77	75.62	77.48	62.96	60.13	74.83	76.41	70.71
PPLX-Embed-0.6B-v1	base	76.33	84.48	63.92	87.54	85.50	71.56	81.25	85.54	82.88	84.74	84.69	81.62	85.55	84.58	75.07	77.48	81.15	83.04	80.94
	ours	79.28	85.51	67.49	88.83	86.73	72.88	83.66	87.37	84.54	86.11	86.59	83.39	86.63	85.77	76.91	80.67	84.73	84.92	82.89
Nemotron-Embed-1B-v2	base	72.55	84.32	74.81	89.43	85.42	74.50	80.42	86.12	80.79	83.28	84.91	82.46	84.27	83.69	72.90	78.73	82.09	81.04	81.21
	ours	72.84	85.04	75.31	89.78	86.21	74.43	82.65	87.13	80.34	83.68	85.70	83.30	84.11	84.86	74.55	80.78	84.93	81.94	82.09

Table 9: BELEBELE MAP@20 results across languages.

Model		Ar	De	El	En	Fr	Hi	Id	It	Ja	Nl	Pt	Ro	Ru	Es	Th	Tr	Vi	Zh	Avg.
Harrier-oss-270M-v1	base	571.0	372.8	730.7	152.8	350.7	572.4	528.1	434.3	445.6	497.6	437.7	585.6	394.6	394.1	716.8	587.2	467.8	293.2	474.1
	ours	221.7	196.0	216.5	165.6	163.0	228.1	214.2	180.8	199.9	192.5	197.9	204.7	195.4	183.3	236.6	211.0	179.8	192.3	198.9
mE5-base	base	397.2	386.4	442.4	270.4	333.6	349.7	330.2	335.7	445.1	375.6	473.1	455.0	263.2	261.6	377.8	363.8	320.4	589.8	376.16
	ours	163.8	106.2	161.0	88.7	103.5	167.5	118.8	109.0	130.8	118.7	113.3	120.2	113.2	107.5	155.8	126.6	114.2	109.6	123.78
mMPNet-base-v2	base	255.0	158.1	190.2	133.6	150.3	243.7	169.0	156.2	197.6	157.1	157.3	152.0	175.1	164.2	214.2	180.4	159.0	176.2	177.18
	ours	170.8	110.7	128.3	99.3	103.2	163.9	119.8	112.4	132.4	105.4	114.9	106.8	120.1	118.1	147.1	122.5	108.9	119.0	122.42
mGTE-base	base	163.9	115.2	149.2	90.2	89.9	144.5	119.9	108.1	131.5	115.1	95.0	123.3	113.7	99.6	167.5	133.0	108.5	121.3	121.6
	ours	114.5	74.5	97.0	56.3	60.8	100.8	81.3	70.1	86.5	78.7	62.2	84.9	74.6	67.3	118.2	86.5	70.4	70.6	80.8
KaLM-Embed-mini-v2.5	base	241.7	189.6	577.1	81.0	159.2	402.3	260.7	191.7	222.6	287.3	219.3	413.7	174.7	154.3	334.8	468.1	165.2	105.5	258.26
	ours	144.3	107.8	259.4	68.9	85.8	194.2	118.3	87.9	117.8	110.1	89.9	144.3	94.8	92.5	161.1	194.7	99.6	84.3	125.30
BGE-M3	base	122.2	94.4	125.2	80.8	91.5	154.7	104.0	102.1	104.2	95.4	94.3	97.5	89.1	94.7	123.7	99.0	95.7	100.9	103.8
	ours	94.4	66.9	86.8	60.1	64.1	102.2	70.5	66.5	75.2	66.1	67.8	66.0	64.6	65.2	90.7	68.8	65.5	65.1	72.6
mE5-large-instruct	base	457.5	294.3	502.9	135.8	311.8	642.8	328.0	285.8	470.7	281.3	286.2	302.1	281.5	280.7	578.7	362.5	382.0	484.4	370.49
	ours	93.4	73.2	94.8	64.4	71.9	122.9	83.2	71.6	87.0	71.7	76.0	72.8	77.9	72.8	101.6	82.5	75.7	77.3	81.70
Solon-Embed-large-0.1	base	158.7	105.4	164.1	79.5	95.9	160.2	108.9	113.6	137.9	103.4	114.5	118.6	103.7	106.7	172.1	137.1	128.3	214.8	129.05
	ours	88.1	67.8	79.9	49.2	59.5	107.9	69.2	64.4	78.2	64.0	61.8	63.1	65.9	63.7	94.9	73.1	62.9	64.3	71.00
Arctic-Embed-L-v2.0	base	182.3	102.3	159.4	94.1	106.7	199.0	134.7	110.9	128.4	123.5	108.9	124.9	111.8	103.1	145.5	140.6	133.2	131.5	130.03
	ours	121.7	75.9	96.7	64.8	71.3	131.0	83.8	79.8	91.6	77.2	77.4	74.9	79.1	74.5	98.5	84.5	76.3	80.1	85.51
Harrier-oss-0.6B-v1	base	569.6	283.2	622.7	113.4	300.6	659.5	425.6	391.7	344.2	384.4	412.8	523.8	346.6	308.6	613.4	444.2	401.6	207.0	408.5
	ours	106.2	71.0	124.9	51.5	63.7	119.5	85.7	67.0	83.4	76.7	65.9	98.5	69.4	69.0	120.7	99.4	72.1	66.6	83.9
Qwen3-Embed-0.6B	base	210.0	140.0	379.4	100.8	151.5	265.5	210.1	136.3	172.7	174.8	146.1	228.6	151.1	131.7	236.3	286.9	135.4	131.3	188.25
	ours	176.4	114.9	296.4	84.0	102.0	232.9	132.8	108.9	130.9	140.8	105.7	176.6	115.9	104.5	177.4	217.4	111.3	103.9	146.25
PPLX-Embed-0.6B-v1	base	124.8	71.0	220.7	56.4	65.8	163.0	95.7	65.2	83.2	74.0	69.0	99.8	72.3	70.6	130.2	131.2	89.1	72.9	97.49
	ours	81.7	54.3	140.0	44.7	49.1	119.3	70.1	48.8	67.4	57.3	50.5	70.4	56.6	55.6	92.2	80.7	55.9	54.0	69.38
Nemotron-Embed-1B-v2	base	134.8	74.3	119.5	48.0	72.8	120.5	91.3	65.6	84.9	83.3	76.2	81.9	70.5	78.4	119.6	98.0	71.4	68.4	86.65
	ours	108.9	61.7	95.1	40.7	56.0	100.6	68.0	53.0	70.2	66.7	62.2	65.4	64.6	62.1	94.9	76.3	55.3	57.4	69.96

Table 10: BELEBELE Max@R results across languages.

Model		Ar	De	El	En	Es	Hi	Ro	Ru	Th	Tr	Vi	Zh	Avg.
Harrier-oss-270M-v1	base	5.97	24.45	1.09	50.92	25.13	8.49	8.07	18.82	4.12	4.29	13.53	37.98	16.91
	ours	53.70	60.92	53.61	62.69	61.51	55.38	57.31	59.66	54.54	53.61	57.31	57.82	57.34
mE5-base	base	14.29	13.03	7.98	23.28	26.97	21.01	6.13	30.08	16.72	13.53	17.23	1.93	16.02
	ours	55.71	62.77	57.90	66.39	63.78	59.33	62.35	64.96	58.66	57.73	62.35	66.13	61.51
mMPNet-base-v2	base	41.60	48.74	47.39	53.19	51.01	43.45	50.50	47.73	45.46	47.65	48.24	47.90	47.74
	ours	51.09	54.96	54.03	58.66	56.64	49.83	56.22	56.47	52.61	53.45	56.55	56.22	54.73
mGTE-base	base	45.55	52.77	45.80	60.17	57.14	48.66	54.12	54.71	42.61	48.49	53.78	58.49	51.86
	ours	53.28	63.28	57.06	69.16	65.71	56.47	61.93	63.11	52.61	58.74	62.94	66.81	60.93
KaLM-Embed-mini-v2.5	base	32.61	34.37	7.23	48.91	38.49	17.90	12.10	35.04	19.83	9.50	35.88	47.31	28.26
	ours	54.71	61.01	45.97	65.71	61.68	54.12	55.46	62.44	55.46	50.92	58.82	63.00	57.44
BGE-M3	base	64.20	64.29	64.45	65.63	63.70	60.50	63.19	67.65	64.54	62.10	62.77	66.81	64.15
	ours	70.50	74.71	73.36	75.80	75.21	67.82	72.61	75.13	72.52	72.52	73.45	76.30	73.33
mE5-large-instruct	base	2.35	15.97	0.84	56.97	18.24	0.17	13.36	15.46	0.42	9.58	6.30	0.67	11.69
	ours	68.49	74.29	70.76	76.39	74.37	68.82	73.70	73.78	70.08	70.17	73.11	74.54	72.38
Solon-Embed-large-0.1	base	51.43	64.20	54.79	61.09	64.20	59.66	61.76	62.94	55.46	58.40	60.17	53.28	58.95
	ours	69.24	74.54	71.60	77.65	74.37	69.24	74.45	73.03	70.59	71.76	74.62	73.95	72.92
Arctic-Embed-L-v2.0	base	53.61	62.27	57.23	64.12	63.70	53.95	60.25	62.77	59.75	54.62	58.82	62.18	59.44
	ours	65.97	74.87	69.83	75.63	73.11	66.97	72.61	73.11	70.08	69.66	71.43	72.77	71.34
Harrier-oss-0.6B-v1	base	12.61	46.81	12.77	65.29	44.29	13.19	21.09	36.55	19.83	21.85	31.01	43.70	30.75
	ours	65.88	72.86	67.39	76.13	73.03	66.97	68.32	69.50	65.38	66.22	70.00	72.35	69.50
Qwen3-Embed-0.6B	base	40.42	48.66	28.15	54.12	48.74	39.75	39.33	49.50	39.75	34.12	45.88	48.24	43.06
	ours	46.64	54.96	38.49	59.66	56.97	44.45	50.76	55.97	47.48	41.60	52.44	56.22	50.47
PPLX-Embed-0.6B-v1	base	63.53	71.68	51.60	69.58	70.76	65.88	68.24	71.93	63.87	63.36	66.64	72.10	66.60
	ours	72.35	79.33	67.31	79.33	79.41	72.77	77.23	78.32	73.11	72.35	76.05	78.40	75.50
Nemotron-Embed-1B-v2	base	76.97	83.70	78.74	85.97	83.78	78.15	83.45	83.36	77.06	78.15	82.94	84.20	81.37
	ours	80.50	88.32	84.20	90.92	88.91	83.53	88.91	88.32	81.43	83.03	86.47	87.65	86.02

Table 11: XQuAD Comp@20 results across languages.

Model		Ar	De	El	En	Es	Hi	Ro	Ru	Th	Tr	Vi	Zh	Avg.
Harrier-oss-270M-v1	base	24.50	55.29	16.62	79.19	55.48	28.69	36.54	43.13	22.69	32.50	40.33	57.73	41.06
	ours	75.39	82.81	77.72	84.72	83.60	78.35	80.58	80.66	75.07	77.43	79.15	79.42	79.58
mE5-base	base	37.84	52.30	36.14	63.15	59.84	44.94	43.33	53.94	41.07	47.35	47.74	17.30	45.41
	ours	70.95	79.93	74.54	83.54	81.49	73.46	79.82	78.59	72.81	75.87	78.38	78.66	77.34
mMPNet-base-v2	base	58.70	68.63	65.37	72.55	69.59	59.96	69.33	65.52	62.41	66.66	67.81	64.69	65.94
	ours	66.72	73.49	71.70	76.91	75.08	67.67	74.79	73.05	68.85	72.22	73.56	72.85	72.24
mGTE-base	base	68.25	78.09	72.51	83.56	80.93	72.80	77.63	77.74	68.20	73.95	76.38	77.49	75.63
	ours	71.27	81.09	76.86	85.81	83.39	75.79	80.21	80.57	71.81	77.04	79.80	81.24	78.74
KaLM-Embed-mini-v2.5	base	52.86	65.97	22.26	78.71	70.01	41.01	36.16	61.09	41.27	30.24	59.10	76.05	52.89
	ours	74.14	82.04	65.72	85.17	82.59	73.49	78.17	81.73	73.95	70.61	79.61	82.82	77.50
BGE-M3	base	79.34	83.42	81.56	85.26	84.34	79.69	83.03	84.16	80.77	81.63	80.29	80.13	81.97
	ours	81.50	86.26	84.28	87.36	86.57	81.48	85.48	85.74	83.68	84.13	84.23	84.93	84.64
mE5-large-instruct	base	15.94	43.87	15.71	75.99	43.34	10.13	37.11	35.04	12.15	27.71	26.73	13.62	29.78
	ours	79.01	85.47	82.18	86.56	85.04	80.43	84.58	83.56	80.41	81.80	83.16	83.27	82.96
Solon-Embed-large-0.1	base	68.45	81.36	75.44	76.17	80.89	74.56	80.54	76.27	73.02	77.99	76.88	67.80	75.78
	ours	79.44	85.70	82.47	87.85	85.18	81.04	85.34	84.00	81.03	83.06	83.68	83.39	83.52
Arctic-Embed-L-v2.0	base	69.51	82.94	75.65	84.98	83.19	73.78	80.44	80.59	76.37	76.09	76.40	76.04	78.00
	ours	75.92	85.30	81.01	87.26	84.82	78.49	84.10	83.53	80.00	81.42	82.26	81.82	82.16
Harrier-oss-0.6B-v1	base	29.64	69.06	34.15	85.01	67.18	29.70	45.59	51.66	37.15	45.08	51.17	55.12	50.04
	ours	80.22	87.26	82.28	89.20	87.16	81.41	83.82	85.22	78.62	81.95	84.47	85.91	83.96
Qwen3-Embed-0.6B	base	66.26	77.17	51.70	81.11	78.13	66.11	68.42	73.80	62.12	62.80	73.50	76.21	69.78
	ours	68.41	79.42	60.42	82.93	80.35	67.21	74.29	77.59	68.46	66.61	76.11	79.82	73.47
PPLX-Embed-0.6B-v1	base	80.56	88.54	72.51	89.23	88.50	82.64	85.19	87.97	81.65	81.88	83.52	86.58	84.06
	ours	84.94	90.90	79.88	91.98	90.86	86.13	88.93	90.64	85.65	86.23	88.15	89.62	87.83
Nemotron-Embed-1B-v2	base	86.81	93.95	88.65	95.22	94.38	89.92	92.80	92.38	84.74	88.81	89.96	91.32	90.75
	ours	88.03	94.93	91.41	96.39	95.40	90.97	94.29	93.64	87.66	91.55	92.62	93.25	92.51

Table 12: XQuAD MAP@20 results across languages.

Model		Ar	De	El	En	Es	Hi	Ro	Ru	Th	Tr	Vi	Zh	Avg.
Harrier-oss-270M-v1	base	241.5	115.8	327.9	49.8	102.4	204.8	187.5	140.5	289.3	251.4	177.7	79.5	180.69
	ours	127.9	63.9	103.6	42.7	57.5	105.1	91.9	68.5	127.7	94.0	70.2	57.9	84.25
mE5-base	base	163.8	152.6	177.8	87.2	90.1	135.6	213.9	84.9	135.0	159.7	137.2	226.1	146.98
	ours	56.5	37.6	53.4	30.3	38.7	59.7	39.8	39.3	51.1	46.5	39.8	36.0	44.03
mMPNet-base-v2	base	102.7	60.6	73.3	52.8	63.9	99.2	63.9	69.3	79.8	74.4	74.4	71.5	73.82
	ours	70.7	48.7	55.1	44.3	50.4	68.4	49.1	49.9	57.6	58.9	57.0	49.6	54.97
mGTE-base	base	90.9	51.5	66.3	32.9	46.8	77.3	55.9	52.6	86.3	64.1	58.7	42.6	60.47
	ours	63.9	37.4	47.0	25.4	33.8	54.0	39.7	38.5	64.4	45.3	44.0	31.4	43.74
KaLM-Embed-mini-v2.5	base	111.4	76.2	251.9	42.0	62.1	163.6	170.0	74.8	155.0	217.1	80.2	48.7	121.08
	ours	57.2	36.1	73.8	29.7	36.0	60.4	44.5	37.4	61.7	61.7	46.4	34.4	48.27
BGE-M3	base	50.1	38.5	41.1	35.3	40.7	54.6	40.6	35.4	43.4	45.6	46.6	37.8	42.46
	ours	37.2	27.1	30.9	24.8	27.1	37.3	28.4	26.8	32.9	30.6	31.0	26.7	30.06
mE5-large-instruct	base	272.4	104.0	297.8	35.5	105.9	302.4	140.8	114.1	352.2	183.4	187.5	237.2	194.43
	ours	39.0	24.5	35.4	23.5	26.3	34.6	26.2	28.6	33.0	28.9	29.3	29.4	29.89
Solon-Embed-large-0.1	base	64.5	34.0	57.0	38.4	38.5	50.8	42.0	41.9	52.7	45.8	46.9	61.0	47.79
	ours	38.7	26.4	35.1	24.3	29.0	35.0	28.1	29.0	32.8	30.9	30.7	31.7	30.96
Arctic-Embed-L-v2.0	base	74.7	40.8	54.8	37.7	44.7	68.4	51.6	45.5	52.8	61.3	56.0	49.4	53.14
	ours	47.0	28.5	37.7	24.9	29.8	42.3	29.5	30.7	35.5	37.0	33.4	30.3	33.87
Harrier-oss-0.6B-v1	base	176.1	56.8	184.6	28.5	60.2	178.2	122.9	77.4	175.3	143.8	99.5	51.4	112.88
	ours	42.74	28.71	39.26	25.27	29.34	42.04	33.31	30.42	48.14	40.37	35.61	28.96	35.35
Qwen3-Embed-0.6B	base	102.1	55.7	150.8	39.7	54.4	108.3	86.0	65.9	107.9	138.6	64.8	47.7	85.15
	ours	82.9	48.4	103.2	38.0	47.9	91.2	65.3	55.1	80.9	95.4	55.3	40.3	66.97
PPLX-Embed-0.6B-v1	base	54.7	24.7	84.3	21.8	25.5	42.3	35.0	26.6	42.2	48.5	40.7	26.7	39.41
	ours	35.2	22.1	46.8	20.1	22.1	30.6	25.6	22.1	33.3	31.0	28.5	22.4	28.31
Nemotron-Embed-1B-v2	base	35.9	19.8	25.1	17.3	20.6	30.6	21.6	19.7	32.7	28.2	24.2	20.4	24.67
	ours	28.9	17.5	20.5	15.5	17.3	24.1	18.2	17.2	25.9	22.0	20.4	18.0	20.45

Table 13: XQuAD Max@R results across languages.

Model	Method	Ar	De	El	En	Fr	Hi	Id	It	Ja	Nl	Pt	Ro	Ru	Es	Th	Tr	Vi	Zh	Avg.
BELEBELE [Comp@20] ↑																				
mMPNet-base-v2	Baseline	26.89	38.33	33.44	42.89	39.22	28.00	36.44	39.33	31.11	38.56	38.78	39.22	34.33	38.56	30.33	36.22	37.56	34.00	35.73
	XLCO	35.33	46.44	43.44	50.11	47.22	38.89	46.33	49.00	39.33	47.22	47.22	47.89	45.67	46.89	38.78	45.33	45.89	43.67	44.70
	SupCon	36.22	47.44	42.78	50.33	48.22	38.33	47.56	50.00	41.44	48.00	48.11	48.44	45.56	47.44	38.56	45.78	46.56	44.33	45.28
	MUSA	41.00	49.44	49.33	54.11	51.56	42.22	50.67	52.33	45.44	51.33	50.22	50.22	48.78	51.00	43.89	48.56	50.78	48.67	48.86
	w/o BT	36.89	48.00	45.56	51.33	49.44	39.78	47.78	50.11	42.67	49.44	48.89	49.00	46.67	49.11	40.44	46.56	47.89	46.67	46.46
mGTE-base	Baseline	40.00	51.22	40.56	53.33	54.56	42.67	49.44	50.89	44.89	52.56	53.11	50.11	50.00	52.89	38.67	45.78	48.89	42.89	47.91
	XLCO	48.11	58.78	51.89	62.44	62.22	53.22	57.78	59.56	54.11	58.89	61.89	58.00	57.44	60.33	47.00	54.33	57.00	53.00	56.44
	SupCon	49.22	59.67	53.67	64.89	64.00	55.00	59.67	61.78	55.67	60.89	64.56	61.00	59.78	63.44	48.89	56.44	58.67	56.56	58.54
	MUSA	50.78	60.22	54.89	65.22	64.33	55.78	60.89	63.11	56.00	61.56	63.78	61.89	59.56	63.11	49.22	56.67	60.78	57.44	59.18
	w/o BT	48.00	57.44	50.67	62.22	62.11	53.44	58.00	59.44	54.44	59.11	62.44	58.56	57.67	61.44	46.56	54.11	57.33	53.22	56.46
BGE-M3	Baseline	47.78	50.33	48.22	55.56	53.44	41.89	49.89	51.89	48.44	52.22	53.11	51.22	52.89	52.89	44.44	50.22	49.33	47.56	50.07
	XLCO	54.89	60.67	58.33	63.22	61.67	53.00	60.56	61.44	58.33	62.78	61.00	61.56	62.67	61.78	54.67	59.89	60.00	57.44	59.66
	SupCon	55.00	62.67	57.67	65.00	63.78	54.33	61.56	62.33	60.22	64.00	62.44	62.33	63.44	63.11	55.89	61.56	60.44	58.56	60.80
	MUSA	57.56	65.00	60.44	67.67	65.89	56.11	63.56	64.78	62.56	66.44	64.44	63.89	65.33	65.00	57.56	63.33	63.67	62.67	63.11
	w/o BT	53.67	58.44	55.00	61.22	60.00	50.44	58.11	58.44	57.78	60.11	59.33	58.44	60.22	60.33	51.44	58.11	57.33	55.33	57.43
Qwen3-Embed-0.6B	Baseline	27.22	38.22	14.56	44.67	31.56	22.22	26.78	39.00	29.78	34.00	37.22	28.22	36.78	40.00	27.11	23.89	36.11	37.67	31.95
	XLCO	33.11	41.78	21.56	49.00	44.89	28.33	39.33	45.67	39.56	42.22	44.56	36.00	45.11	44.00	34.56	30.89	42.00	43.11	39.20
	SupCon	25.89	37.56	16.33	45.56	41.56	21.44	36.67	39.11	33.67	35.33	38.56	31.11	38.56	39.78	29.44	25.00	35.56	38.00	33.84
	MUSA	34.22	44.00	23.33	53.11	49.11	29.22	41.89	47.89	40.89	45.00	46.00	39.11	45.56	47.00	35.89	33.22	44.56	46.11	41.45
	w/o BT	32.22	42.44	22.44	51.11	46.56	27.56	41.33	45.56	39.22	42.00	43.67	37.44	43.89	45.33	35.11	31.78	41.67	44.00	39.63
PPLX-Embed-0.6B-v1	Baseline	49.33	61.78	34.67	63.89	62.89	44.56	56.56	62.22	57.00	62.11	59.44	58.00	61.78	61.56	49.11	48.89	55.00	58.78	55.98
	XLCO	59.67	66.78	46.56	69.67	68.22	54.33	64.44	67.00	64.78	66.33	66.00	65.44	68.89	66.33	55.67	58.67	66.22	66.11	63.40
	SupCon	58.00	66.56	43.78	71.11	68.22	52.56	63.22	67.78	63.44	66.11	66.33	64.67	68.44	66.89	55.78	58.33	64.22	65.78	62.85
	MUSA	62.00	68.44	49.22	72.00	70.22	56.11	65.78	70.22	67.33	69.22	68.89	66.56	70.78	68.78	59.11	61.56	67.22	68.11	65.64
	w/o BT	61.56	68.44	47.11	71.89	70.44	55.44	65.78	70.22	66.56	67.89	68.22	66.44	70.33	68.56	58.00	60.67	66.78	67.11	65.08
BELEBELE [Max@R] ↓																				
mMPNet-base-v2	Baseline	255.0	158.1	190.2	133.6	150.3	243.7	169.0	156.2	197.6	157.1	157.3	152.0	175.1	164.2	214.2	180.4	159.0	176.2	177.18
	XLCO	191.7	119.6	141.6	107.2	112.2	177.3	130.0	123.2	145.2	117.6	124.4	115.6	129.2	126.5	161.3	134.7	118.8	128.3	133.57
	SupCon	191.5	119.0	140.3	104.6	110.7	177.7	128.9	121.0	146.2	115.8	122.5	114.4	128.5	125.3	160.6	134.3	116.9	127.3	132.53
	MUSA	170.8	110.7	128.3	99.3	103.2	163.9	119.8	112.4	132.4	105.4	114.9	106.8	120.1	118.1	147.1	122.5	108.9	119.0	122.42
	w/o BT	180.4	114.4	133.7	102.3	107.0	170.0	124.5	118.6	138.3	111.4	118.9	111.5	123.8	121.9	153.3	128.2	113.7	122.4	127.45
mGTE-base	Baseline	163.9	115.2	149.2	90.2	89.9	144.5	119.9	108.1	131.5	115.1	95.0	123.3	113.7	99.6	167.5	133.0	108.5	121.3	121.63
	XLCO	120.1	84.4	107.4	63.7	69.0	107.0	88.0	77.8	93.2	85.1	69.0	94.5	82.5	74.2	121.8	97.1	78.1	83.2	88.67
	SupCon	112.7	75.6	98.0	56.5	59.8	99.7	80.4	68.3	86.5	77.9	62.6	85.1	74.8	67.8	118.2	90.5	71.6	71.7	80.97
	MUSA	114.5	74.5	97.0	56.3	60.8	100.8	81.3	70.1	86.5	78.7	62.2	84.9	74.6	67.3	118.2	86.5	70.4	70.6	80.85
	w/o BT	120.0	83.3	107.0	62.6	67.0	107.7	87.5	76.2	93.0	84.6	67.9	92.8	82.0	73.3	123.9	95.0	77.0	81.1	87.88
BGE-M3	Baseline	122.2	94.4	125.2	80.8	91.5	154.7	104.0	102.1	104.2	95.4	94.3	97.5	89.1	94.7	123.7	99.0	95.7	100.9	103.85
	XLCO	99.4	71.2	92.4	65.6	70.7	113.4	76.7	71.7	80.7	72.4	72.3	72.0	69.5	70.9	97.4	74.8	72.3	73.1	78.69
	SupCon	96.1	66.6	89.6	60.2	64.3	108.9	71.7	67.1	77.3	67.3	67.0	64.9	65.0	66.1	93.4	71.1	67.3	66.9	73.93
	MUSA	94.4	66.9	86.8	60.1	64.1	102.2	70.5	66.5	75.2	66.1	67.8	66.0	64.6	65.2	90.7	68.8	65.5	65.1	72.58
	w/o BT	102.0	73.4	97.8	66.2	72.9	120.9	78.3	75.6	85.7	74.4	74.0	72.5	70.3	73.5	101.3	76.7	74.9	75.4	81.43
Qwen3-Embed-0.6B	Baseline	210.0	140.0	379.4	100.8	151.5	265.5	210.1	136.3	172.7	174.8	146.1	228.6	151.1	131.7	236.3	286.9	135.4	131.3	188.25
	XLCO	185.2	121.9	316.4	90.6	114.2	236.4	149.4	115.9	138.6	150.6	113.6	193.7	122.9	112.2	189.1	242.9	119.5	112.6	156.97
	SupCon	217.4	144.1	362.9	100.2	124.1	281.6	158.7	134.0	156.9	172.6	132.9	209.9	143.3	127.9	209.4	272.3	136.2	125.5	178.32
	MUSA	176.4	114.9	296.4	84.0	102.0	232.9	132.8	108.9	130.9	140.8	105.7	176.6	115.9	104.5	177.4	217.4	111.3	103.9	146.25
	w/o BT	185.3	121.9	308.3	87.7	107.0	242.2	137.5	115.0	137.5	147.7	112.9	183.3	121.6	110.0	182.5	229.5	117.8	109.7	153.19
PPLX-Embed-0.6B-v1	Baseline	124.8	71.0	220.7	56.4	65.8	163.0	95.7	65.2	83.2	74.0	69.0	99.8	72.3	70.6	130.2	131.2	89.1	72.9	97.49
	XLCO	82.4	56.6	149.6	48.5	53.5	119.7	73.0	50.2	68.2	60.0	52.6	75.7	60.3	57.7	94.4	86.7	59.6	55.6	72.45
	SupCon	98.0	60.7	175.1	44.8	57.0	142.0	79.9	53.6	73.1	61.3	57.8	78.4	61.8	61.4	112.2	99.0	65.9	59.0	80.05
	MUSA	81.7	54.3	140.0	44.7	49.1	119.3	70.1	48.8	67.4	57.3	50.5	70.4	56.6	55.6	92.2	80.7	55.9	54.0	69.38
	w/o BT	82.2	55.2	148.1	45.2	49.7	121.4	71.0	48.5	67.4	57.8	50.6	71.7	58.1	55.7	95.6	82.7	57.2	54.5	70.70

Table 14: Ablation results on BELEBELE across languages. The upper block reports Comp@20, where higher is better, and the lower block reports Max@R, where lower is better.

Model	Method	Ar	De	El	En	Es	Hi	Ro	Ru	Th	Tr	Vi	Zh	Avg.
							XQuAD [Comp@20] ↑							
mMPNet-base-v2	Baseline	41.60	48.74	47.39	53.19	51.01	43.45	50.50	47.73	45.46	47.65	48.24	47.90	47.74
	XLCO	45.80	50.59	49.92	53.61	50.67	45.29	51.09	51.09	46.55	48.74	50.50	50.92	49.56
	SupCon	46.81	52.52	51.18	55.55	52.27	45.88	54.03	52.35	48.40	50.67	52.35	52.86	51.24
	MuSA	51.09	54.96	54.03	58.66	56.64	49.83	56.22	56.47	52.61	53.45	56.55	56.22	54.73
	w/o BT	48.40	54.37	51.85	56.81	53.87	47.06	54.37	54.29	50.08	52.10	53.36	54.37	52.58
mGTE-base	Baseline	45.55	52.77	45.80	60.17	57.14	48.66	54.12	54.71	42.61	48.49	53.78	58.49	51.86
	XLCO	50.08	56.72	51.76	62.69	60.00	53.70	57.65	59.08	48.74	54.29	58.07	62.52	56.28
	SupCon	52.86	63.19	57.39	70.08	66.89	57.65	64.20	63.11	52.77	59.33	63.95	67.56	61.58
	MuSA	53.28	63.28	57.06	69.16	65.71	56.47	61.93	63.11	52.61	58.74	62.94	66.81	60.93
	w/o BT	50.67	59.41	54.37	66.13	62.77	55.29	60.59	60.84	51.34	56.05	60.67	64.79	58.58
BGE-M3	Baseline	64.20	64.29	64.45	65.63	63.70	60.50	63.19	67.65	64.54	62.10	62.77	66.81	64.15
	XLCO	67.82	68.82	67.82	69.41	69.08	64.03	66.97	71.26	68.40	66.97	68.07	70.25	68.24
	SupCon	70.92	74.29	72.77	75.21	74.20	67.31	72.35	74.20	71.60	71.68	73.19	74.20	72.66
	MuSA	70.50	74.71	73.36	75.80	75.21	67.82	72.61	75.13	72.52	72.52	73.45	76.30	73.33
	w/o BT	69.08	71.26	70.50	73.45	71.85	66.22	70.59	73.03	70.34	69.66	70.17	72.02	70.68
Qwen3-Embed-0.6B	Baseline	40.42	48.66	28.15	54.12	48.74	39.75	39.33	49.50	39.75	34.12	45.88	48.24	43.06
	XLCO	43.53	51.34	35.29	55.21	52.10	42.77	44.79	51.85	44.96	39.16	48.40	52.10	46.79
	SupCon	39.92	49.24	32.18	54.29	51.01	39.50	43.70	51.68	45.71	35.38	47.82	51.09	45.13
	MuSA	46.64	54.96	38.49	59.66	56.97	44.45	50.76	55.97	47.48	41.60	52.44	56.22	50.47
	w/o BT	44.45	54.20	38.24	58.57	56.30	45.13	49.41	54.96	48.91	40.67	52.18	55.97	49.92
PPLX-Embed-0.6B-v1	Baseline	63.53	71.68	51.60	69.58	70.76	65.88	68.24	71.93	63.87	63.36	66.64	72.10	66.60
	XLCO	70.08	74.62	61.93	74.03	75.13	69.08	72.94	75.38	69.41	68.49	70.92	76.22	71.52
	SupCon	71.26	79.08	64.03	79.83	79.24	72.94	77.73	78.15	72.18	71.34	75.13	76.47	74.78
	MuSA	72.35	79.33	67.31	79.33	79.41	72.77	77.23	78.32	73.11	72.35	76.05	78.40	75.50
	w/o BT	73.03	78.99	66.81	79.66	79.08	73.28	76.72	78.91	72.94	71.93	75.71	78.40	75.46
							XQuAD [Max@R] ↓							
mMPNet-base-v2	Baseline	102.7	60.6	73.3	52.8	63.9	99.2	63.9	69.3	79.8	74.4	74.4	71.5	73.8
	XLCO	80.5	54.8	62.3	49.8	58.3	78.6	56.8	57.0	65.5	66.8	66.2	57.6	62.9
	SupCon	78.6	52.1	59.5	47.1	55.6	75.9	54.5	55.0	63.7	62.8	62.3	55.3	60.2
	MuSA	70.7	48.7	55.1	44.3	50.4	68.4	49.1	49.9	57.6	58.9	57.0	49.6	55.0
	w/o BT	75.7	51.4	58.1	47.1	54.4	73.2	52.8	53.2	61.4	62.7	61.0	53.1	58.7
mGTE-base	Baseline	90.9	51.5	66.3	32.9	46.8	77.3	55.9	52.6	86.3	64.1	58.7	42.6	60.5
	XLCO	76.6	43.8	56.0	29.3	41.0	63.1	47.1	44.8	73.2	52.2	50.0	35.5	51.0
	SupCon	67.7	38.4	48.6	25.0	34.9	55.6	40.5	38.5	68.0	46.0	44.0	30.7	44.8
	MuSA	63.9	37.4	47.0	25.4	33.8	54.0	39.7	38.5	64.4	45.3	44.0	31.4	43.7
	w/o BT	73.1	41.9	53.7	28.0	39.0	61.0	45.3	43.2	70.5	50.1	48.3	34.2	49.0
BGE-M3	Baseline	50.1	38.5	41.1	35.3	40.7	54.6	40.6	35.4	43.4	45.6	46.6	37.8	42.5
	XLCO	41.7	32.0	33.8	29.4	31.8	43.9	33.1	30.4	36.3	35.2	36.3	31.5	34.6
	SupCon	38.2	28.6	31.0	25.2	27.7	37.5	29.4	27.3	32.2	31.5	30.7	26.8	30.5
	MuSA	37.2	27.1	30.9	24.8	27.1	37.3	28.4	26.8	32.9	30.6	31.0	26.7	30.1
	w/o BT	40.7	30.8	33.4	28.0	30.4	40.9	31.8	29.2	34.8	34.0	34.7	29.8	33.2
Qwen3-Embed-0.6B	Baseline	102.1	55.7	150.8	39.7	54.4	108.3	86.0	65.9	107.9	138.6	64.8	47.7	85.2
	XLCO	88.3	53.2	114.1	42.0	52.4	95.7	71.0	61.0	89.5	106.9	58.9	45.1	73.2
	SupCon	109.6	59.3	134.5	44.8	59.3	117.3	78.5	68.1	99.0	125.6	68.6	48.1	84.4
	MuSA	82.9	48.4	103.2	38.0	47.9	91.2	65.3	55.1	80.9	95.4	55.3	40.3	67.0
	w/o BT	87.2	50.9	107.1	39.6	50.5	94.7	67.6	58.0	82.6	98.9	57.9	42.4	69.8
PPLX-Embed-0.6B-v1	Baseline	54.7	24.7	84.3	21.8	25.5	42.3	35.0	26.6	42.2	48.5	40.7	26.7	39.4
	XLCO	39.9	23.7	54.3	22.0	23.8	34.1	28.2	23.8	37.0	35.3	32.0	23.8	31.5
	SupCon	38.7	22.4	55.2	20.1	22.3	32.4	26.9	22.5	34.9	35.1	29.2	23.5	30.3
	MuSA	35.2	22.1	46.8	20.1	22.1	30.6	25.6	22.1	33.3	31.0	28.5	22.4	28.3
	w/o BT	36.3	22.2	48.7	20.1	22.1	31.2	25.9	22.3	34.5	31.8	28.9	22.6	28.9

Table 15: Ablation results on XQuAD across languages. The upper block reports Comp@20, where higher is better, and the lower block reports Max@R, where lower is better.

Model	Method	De	En	Hi	It	Nl	Pt	Avg.
				Wikipedia [nDCG@10]				
mMPNet-base-v2	Baseline	77.39	87.21	56.35	75.49	76.68	75.84	74.83
	XLCO	80.69	85.90	67.11	79.12	81.41	79.50	78.96
	SupCon	80.16	86.38	65.89	78.78	81.06	78.93	78.53
	MuSA	82.25	87.90	69.54	81.25	82.81	80.86	80.77
	w/o BT	81.73	86.81	68.69	80.20	82.12	80.06	79.94
mGTE-base	Baseline	84.26	90.49	81.40	83.85	83.98	82.33	84.39
	XLCO	85.91	90.79	83.61	85.34	85.69	84.24	85.93
	SupCon	87.39	91.47	83.84	86.86	86.82	85.79	87.03
	MuSA	87.41	91.30	84.66	85.89	85.94	85.76	86.83
	w/o BT	86.85	91.43	83.89	86.56	86.62	85.25	86.77
BGE-M3	Baseline	90.28	92.58	87.37	90.13	90.11	88.65	89.85
	XLCO	89.56	91.99	86.49	89.55	89.48	87.36	89.07
	SupCon	90.90	92.76	86.78	90.27	90.32	89.20	90.04
	MuSA	90.35	92.32	85.20	89.32	89.27	88.25	89.12
	w/o BT	91.14	93.07	87.69	91.04	90.79	89.41	90.52
Qwen3-Embed-0.6B	Baseline	89.04	92.53	83.50	88.28	88.12	87.02	88.08
	XLCO	89.26	92.69	82.58	88.67	88.13	87.04	88.06
	SupCon	89.48	92.74	80.50	88.41	87.60	87.66	87.73
	MuSA	89.54	93.02	81.07	88.72	88.32	88.37	88.17
	w/o BT	90.28	93.37	81.85	89.56	89.23	88.80	88.85
PPLX-Embed-0.6B-v1	Baseline	91.57	94.36	87.55	92.21	91.32	90.78	91.30
	XLCO	91.88	94.53	87.54	92.67	91.90	91.24	91.63
	SupCon	92.03	94.97	85.72	92.38	92.49	91.36	91.49
	MuSA	92.67	95.25	86.60	93.17	92.97	92.11	92.13
	w/o BT	93.06	95.53	87.33	93.57	93.11	92.53	92.52

Table 16: Ablation results on Wikipedia multi-monolingual retrieval across languages.