CLEAR: Cross-Lingual Enhancement in Retrieval via Reverse-training

Seungyoon Lee[1], **Minhyuk Kim**[1], **Seongtae Hong**[1], **Youngjoon Jang**[1],
Dongsuk Oh[2], **Heuiseok Lim**[1*]
[1]Department of Computer Science and Engineering, Korea University
[2]Department of English Language and Literature, Kyungpook National University
{dltmddbs100, mhkim0929, ghdchlwls123, dew1701, limhseok}@korea.ac.kr,
inow3555@knu.ac.kr

Abstract

Existing multilingual embedding models often encounter challenges in cross-lingual scenarios due to imbalanced linguistic resources and less consideration of cross-lingual alignment during training. Although standardized contrastive learning approaches for cross-lingual adaptation are widely adopted, they may struggle to capture fundamental alignment between languages and degrade performance in well-aligned languages such as English. To address these challenges, we propose **C**ross-**L**ingual **E**nhancement in Retriev**A**l via **R**everse-training (**CLEAR**), a novel loss function utilizing a reverse training scheme to improve retrieval performance across diverse cross-lingual retrieval scenarios. CLEAR leverages an English passage as a bridge to strengthen alignments between the target language and English, ensuring robust performance in the cross-lingual retrieval task. Our extensive experiments demonstrate that CLEAR achieves notable improvements in cross-lingual scenarios, with gains up to 15%, particularly in low-resource languages, while minimizing performance degradation in English. Furthermore, our findings highlight that CLEAR offers promising effectiveness even in multilingual training, suggesting its potential for broad application and scalability. We release the code at `https://github.com/dltmddbs100/CLEAR`.

1 Introduction

The recent progress in Large Language Models (LLMs) has enabled multilingual applications such as question answering via retrieval-augmented generation (RAG), substantially increasing the demand for robust cross-lingual information retrieval system (Lewis et al., 2020b; Siriwardhana et al., 2022; Li et al., 2022a; Gao et al., 2023; Zhang et al., 2023; Kamalloo et al., 2023; Chirkova et al., 2024; Wang et al., 2024a, 2025).

*Corresponding author

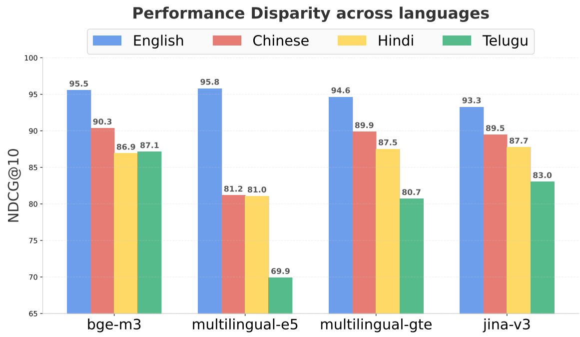


Figure 1: Performance disparity of various embedding models across languages in a cross-lingual setup where English passage with other language queries in Belebele benchmark.

Nevertheless, existing multilingual embedding models widely used for information retrieval often suffer from imbalanced linguistic distribution and insufficient attention to cross-lingual alignment during training (Izacard et al., 2021; Chen et al., 2024; Wang et al., 2024b; Zhang et al., 2024b; Sturua et al., 2024). This leads to biased representations and suboptimal retrieval performance, particularly in low-resource languages (Palta et al., 2022; Huang et al., 2023a; Yang et al., 2024b; Hong et al., 2026), exacerbating disparity in cross-lingual information retrieval scenarios.

This disparity leads to lower performance due to diminished expressiveness in particular languages. As illustrated in Figure 1, user queries in low-resource languages lead to a considerable drop in retrieval performance due to the constraints of the model's representative capability. This increases the risk of providing users with inaccurate information, thereby posing a considerable challenge to ensuring equitable access to multilingual information (Lawrie et al., 2023; Park and Lee, 2025). This limitation not only hinders the accuracy and fairness of cross-lingual retrieval systems but also highlights the need for training strategies that ex-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 347–362
July 2-7, 2026

plicitly enhance cross-lingual alignment, especially in resource-scarce contexts.

One of the prevalent approaches to address this issue involves enhancing cross-lingual alignment through an additional contrastive learning stage (Shuaibo et al., 2022; Wang et al., 2022; Zhang et al., 2024b; Chen et al., 2024). The commonly adopted InfoNCE (Oord et al., 2018) loss minimizes the distance between queries and gold passages while increasing the distance to negative samples. Although this loss function can be effective in encouraging query-passage similarity, it primarily focuses on distinguishing relevant passages based solely on the query. Consequently, it often captures only superficial representation ability and may fail to ensure fundamental alignment between different languages. Furthermore, these training strategies can degrade monolingual performance in the dominant language (e.g., English) during cross-lingual training.

In this paper, we propose Cross-Lingual Enhancement in Retrieval via Reverse-training (CLEAR), a novel training objective designed to improve retrieval performance across all cross-lingual scenarios where queries and passages are in different languages. CLEAR jointly trains on English and cross-lingual alignment by leveraging the English passage as a bridge to the target language, promoting diverse interactions among various components. In contrast to conventional methods that train models to retrieve related passages in response to a query, we introduce a reverse training scheme that fosters the model to capture multifaceted representation capabilities. This approach strengthens the foundational alignment between the languages, ensuring robustness across all cross-lingual scenarios.

Across extensive experiments spanning nine languages, our empirical findings demonstrate that CLEAR achieves substantial improvements in cross-lingual scenarios, especially exhibiting notable performance in low-resource languages, while concurrently mitigating the degradation of original proficiency in English. Furthermore, our approach substantiates its validity in the multilingual configuration where multiple languages are jointly trained. Our contributions are as follows:

- We propose a novel cross-lingual specialized loss, CLEAR, that leverages a reverse training scheme based on English passage bridge to enhance cross-lingual capability.
- We empirically verify the effectiveness of CLEAR for cross-lingual retrieval tasks through extensive experiments using various embedding models on a range of high- and low-resource languages while minimizing the degradation of English performance compared to the standard training approach.
- We show that CLEAR extends beyond cross-lingual scenarios, also proving highly effective in multilingual training when multiple target languages are concurrently addressed.

2 Related Work

In the field of cross-lingual retrieval, existing studies can largely be categorized into two primary directions. The first direction employs translated pairs for direct fine-tuning to adapt retrieval models to target languages (Litschko et al., 2018; Shi et al., 2021; Shuaibo et al., 2022; Zhang and Misra, 2022; Zhuang et al., 2023). For example, Shi et al. (2021) and Zhuang et al. (2023) utilize the query generation model based on the translation of English query-passage pairs to generate synthetic queries, followed by the training of retriever on this dataset.

On the other hand, the other line of research centers on knowledge distillation methods focusing on distilling insights from monolingual models into multilingual frameworks (Reimers and Gurevych, 2020; Limkonchotiwat et al., 2022; Li et al., 2022b; Huang et al., 2023b; Yang et al., 2024a; Zhang et al., 2024a). Huang et al. (2023b) introduces the Optimal Transport Distillation strategy to facilitate the transfer of knowledge from high to low resource languages by utilizing a well-trained monolingual retrieval model. Similarly, other studies distill representation or ranking knowledge of well-aligned language models into student models using parallel query-document pairs (Li et al., 2022b; Limkonchotiwat et al., 2022; Yang et al., 2024a).

More recently, the focus has shifted toward training retrieval models on large-scale multilingual query-document datasets to map language-specific representations into a shared embedding space (Chen et al., 2024; Zhang et al., 2024b; Sturua et al., 2024; Wang et al., 2024b; Lee et al., 2025).

While these approaches have succeeded in scaling coverage to a broader set of languages, they typically depend on vast amounts of parallel data and are often limited to capturing shallow cross-lingual interactions due to their reliance on the conventional InfoNCE loss. In this paper, we introduce a cross-lingual specialized loss function based on shared English passages to establish connections among languages. Our approach promotes diverse interactions among the components to enhance cross-lingual alignment in a resource-constrained environment.

3 CLEAR

We design CLEAR as a cross-lingual training loss function that leverages a reversal scheme to induce robust alignment between English and the target language.

3.1 Overview of InfoNCE

In general, retrievers learn meaningful representations based on similarity to identify the gold passage relevant to a given user query from a large passage pool. The prevalent approaches employ the InfoNCE loss (Oord et al., 2018) with multiple negatives. The formula for the retrieval task is as follows.

Given a text pairs (q_i, p_i^+), we assign a negative passage p_{ij}^- for the i-th example:

$$\mathcal{L}_{NCE} = -\frac{e^{\text{sim}(q_i,p_i^+)/\tau}}{e^{\text{sim}(q_i,p_i^+)/\tau} + \sum_j e^{\text{sim}(q_i,p_{ij}^-)/\tau}} \quad (1)$$

This formula promotes the model to distinguish between related pairs $(q_i,\ p_i^+)$ and unrelated passages (p_{ij}^-) within the embedding space, as quantified by a cosine similarity sim. Based on the InfoNCE loss, we modify the loss to promote cross-lingual alignment to suit our goal.

3.2 Proposed Strategy

CLEAR combines traditional contrastive learning with cross-lingual considerations to provide a solid foundation for enhancing retrieval capabilities. In Figure 2, a comparison of the interactions between queries and passages induced by conventional InfoNCE and CLEAR during training is shown. Compared to InfoNCE, CLEAR achieves sophisticated cross-lingual alignment by establishing diverse direct and indirect interactions centered around P_{en}^+.

CLEAR consists of a universal InfoNCE term designed for learning English representations alongside a cross-lingual term that encourages alignment between the target language (ℓ) and English (en). We define our overall loss function as follows:

$$\mathcal{L}_{\text{CLEAR}} = \lambda_1 \cdot \mathcal{L}_{\text{NCE}_{en}} + \lambda_2 \cdot \mathcal{L}_{\text{CL}} + \lambda_3 \cdot \mathcal{L}_{\text{KL}} \quad (2)$$

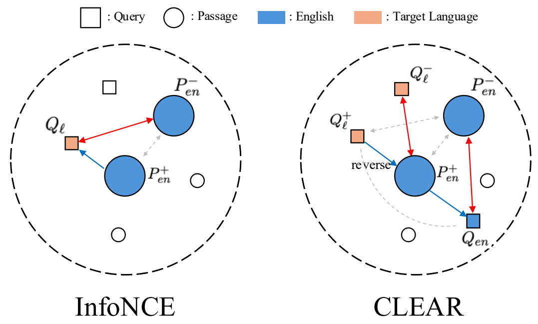


Figure 2: Comparison of the core idea of CLEAR with the standard InfoNCE loss. Solid arrows indicate the direct effects of training, while dashed arrows represent the indirect interactions, such as the resulting attraction or repulsion between representations.

$\mathcal{L}_{\text{NCE}_{en}}$ represents the standard NCE loss associated with English pairs (Equation 1), aiming at preserving the model's inherent performance in English while providing a language bridge via the passage. The direct training signal of cross-lingual alignment occurs through $\mathcal{L}_{\text{CL}}$, which is the reversed cross-lingual loss term to align English with the target language:

$$\mathcal{L}_{\text{CL}}(p_{en_i}, q_{\ell_i}^+, q_{\ell_{ij}}^-) = -\frac{e^{z_i^+/\tau}}{e^{z_i^+/\tau} + \sum_j e^{z_{ij}^-/\tau}} \quad (3)$$

where z_i^+ and z_{ij}^- is defined as:

$$z_i^+ = \text{sim}(p_{en_i}, q_{\ell_i}^+), \ z_{ij}^- = \text{sim}(p_{en_i}, q_{\ell_{ij}}^-) \quad (4)$$

where p_{en_i} denotes the English passage that serves as an anchor, while $q_{\ell_i}^+$ and $q_{\ell_{ij}}^-$ refer to the positive and negative target language queries corresponding to p_{en_i}. Thus, the training objective induces the model to pull the gold query closer to p_{en_i} and push unrelated queries further.

Next, $\mathcal{L}_{\text{KL}}$ denotes the KL-Divergence (Kullback and Leibler, 1951) between the similarity matrices S_{en} and S_{CL}:

$$\text{KL}(S_{en} \parallel S_{CL}) = \sum_{i,j} S_{en}[i,j] \log \frac{S_{en}[i,j]}{S_{CL}[i,j]} \quad (5)$$

where $S_{en}[i,j] = \text{sim}(q_{en_i}, p_{en_j})$ and $S_{CL}[i,j] = \text{sim}(p_{en_j}, q_{\ell_i})$ in the batch. In this manner, we harmonize similarity distributions between language pairs to support consistent representation spaces. The detailed strategies of CLEAR are as follows.

Reversal Scheme We introduce a reversal scheme in $\mathcal{L}_{\text{CL}}$ that swaps the roles of the query and passage to provide a new perspective on cross-lingual training signals. Unlike standard approaches, which consider $(q_{en_i}, p^+_{\ell_i}, p^-_{\ell_{ij}})$ as anchor, positive and negative respectively, this scheme encourages the model to use passage as an anchor and maximize the similarity with the corresponding gold target language query as shown in Equation 3.

This scheme enables the utilization of unrelated target language queries from in-batch samples as negative samples during training, thereby facilitating contrastive learning. By offering a new direction of training signal beyond the standard training signal, the reversal scheme supports the model to learn from passage-to-query perspectives, thereby promoting more robust cross-lingual adaptation.

Passage Bridge We share the same English passage p_{en_i} for both the target language query and corresponding English query. This approach creates a bridge between languages by jointly optimizing relevant query-passage pairs and achieving distributional alignment. As illustrated by the dotted arrows in Figure 2, Q^+_ℓ moves closer to Q_{en}, which conveys the same meaning, and further from the negative English passage, due to the interaction rise via the passage bridge. This strategy promotes the model to consider a broader range of mutual interactions during training, facilitating robust alignment between language representations.

Distribution Approximation Relying solely on instance-level contrastive signals may lead to less robust or more fragmented representations. To mitigate this issue, we employ a KL-Divergence loss to align the similarity distribution between P_{en} and Q_ℓ with that between P_{en} and Q_{en}. While other loss terms operate at the instance level by optimizing query-passage pairs, the KL term goes beyond individual point-to-point relationships; it shapes the overall semantic topology, ensuring that the global organization of meanings remains coherent across languages. This mechanism complements mutual interactions among the loss components and encourages the model to maintain shared representations across English and the target language.

Through the composite design, CLEAR benefits from a strong alignment signal between the target language and English while preserving the model's proficiency in English during the training.

4 Experimental Setup

4.1 Training

Dataset To enable cross-lingual training, target language queries paired with corresponding English passages are necessary. We employ the English portion of the MIRACL (Zhang et al., 2023) training set and MLQA (Lewis et al., 2020a), both of which provide queries mapped to gold passages. To gain parallel queries with English, we use the NLLB (Costa-Jussà et al., 2022) translation model[1]. We exclude queries sharing the same passage to prevent false negatives that may arise from duplicate passages within a single batch, resulting in a collection of unique query-passage pairs. This process yields 12,698 training samples for each language.

Regarding hard-negative selection, following the findings of Gabriel de Souza et al. (2024), which reports that selecting top-k candidates within the range of 30 to 100 effectively reduces false negatives during the negative mining, we sample 5 hard negatives for each training sample using each embedding model. Further details and exact training example can be found in Appendix A and D.

Models We adopt four widely used multilingual embedding models across various tasks: bge-m3 (Chen et al., 2024), multilingual-e5 (Wang et al., 2024b), gte-multilingual (Zhang et al., 2024b), and jina-v3 (Sturua et al., 2024). The models are trained under the identical hyper-parameters and evaluations are conducted under the same conditions. We heuristically set weights for each loss term in our experiments: $\lambda_1 = 0.4$, $\lambda_2 = 0.4$, and $\lambda_3 = 0.2$. Regarding this, the sensitivity analysis of loss weight parameters is provided in Appendix F.

4.2 Baseline

We employ the standard InfoNCE loss, which integrates in-batch negative sampling with external multiple negatives (Henderson et al., 2017) as a major baseline for our experiments. To be specific, we train the model to increase the similarity between a target language query (used as the anchor) and its relevant English passage, focusing exclusively on the cross-lingual alignment.

To ensure fair comparison, we utilize the same five hard negative samples for target language queries. We also conduct query negative mining for our cross-lingual term, which leverages queries

[1] https://huggingface.co/facebook/nllb-200-3.3B

Model	Language	English-English		English-Lang			Lang-English		
		InfoNCE	CLEAR	Base	InfoNCE	CLEAR	Base	InfoNCE	CLEAR
Belebele									
bge-m3	zh	94.65 (-0.90)	**95.47 (-0.08)**	90.35	91.69	**92.64**	89.81	91.63	**92.02**
	es	95.51 (-0.04)	**96.14 (+0.59)**	91.99	92.81	**93.39**	92.05	93.50	**93.82**
	de	95.61 (+0.06)	**96.11 (+0.56)**	92.81	93.98	**94.58**	92.90	94.30	**94.52**
	hi	95.15 (-0.40)	**95.75 (+0.20)**	86.90	89.14	**90.16**	90.39	92.86	**93.16**
	vi	95.14 (-0.41)	**95.79 (+0.24)**	92.58	93.04	**93.12**	93.04	92.58	**93.54**
	te	94.93 (-0.62)	**95.74 (+0.19)**	87.12	89.07	**90.12**	89.19	92.08	**92.74**
	bn	95.18 (-0.37)	**95.81 (+0.26)**	87.58	89.79	**90.85**	90.17	92.37	**93.08**
multilingual-e5	zh	94.18 (-1.61)	**95.06 (-0.73)**	81.16	87.00	**88.89**	85.49	90.02	**90.75**
	es	94.38 (-1.41)	**95.40 (-0.39)**	89.46	89.84	**91.61**	92.43	92.15	**93.13**
	de	94.49 (-1.30)	**95.06 (-0.73)**	90.77	90.97	**91.94**	92.16	92.68	**92.86**
	hi	94.11 (-1.68)	**95.28 (-0.51)**	81.05	83.11	**86.10**	88.44	91.08	**92.13**
	vi	94.22 (-1.57)	**95.07 (-0.72)**	86.03	87.11	**88.63**	88.58	91.81	**91.95**
	te	93.80 (-1.99)	**95.11 (-0.68)**	69.90	77.14	**80.97**	84.88	88.05	**88.99**
	bn	93.83 (-1.96)	**95.28 (-0.51)**	75.55	81.81	**85.44**	84.31	88.91	**89.92**
gte-multilingual	zh	94.38 (-0.25)	**95.15 (+0.52)**	89.86	92.30	**92.67**	91.51	92.05	**92.80**
	es	95.51 (+0.88)	**95.73 (+1.10)**	91.71	92.86	**93.28**	90.20	93.55	**93.87**
	de	95.32 (+0.69)	**95.67 (+1.04)**	91.21	92.92	**93.26**	89.42	93.45	**93.95**
	hi	94.93 (+0.30)	**95.40 (+0.77)**	87.51	89.34	**89.96**	89.55	92.45	**93.05**
	vi	95.02 (+0.39)	**95.71 (+1.08)**	89.37	91.59	**92.23**	90.48	93.14	**93.52**
	te	94.63 (0.00)	**95.32 (+0.69)**	80.70	84.77	**86.31**	88.46	89.96	**90.92**
	bn	94.54 (-0.09)	**95.40 (+0.77)**	82.13	85.18	**86.64**	87.81	91.03	**92.14**
jina-v3	zh	94.97 (+1.71)	**95.31 (+2.05)**	89.46	91.90	**92.67**	89.64	91.74	**92.18**
	es	95.21 (+1.95)	**95.66 (+2.40)**	91.40	93.57	**94.29**	92.64	**93.88**	93.86
	de	95.29 (+2.03)	**95.62 (+2.36)**	91.75	93.88	**94.46**	92.39	94.12	**94.46**
	hi	94.01 (+0.75)	**95.59 (+2.33)**	87.74	89.90	**90.70**	91.50	93.02	**93.34**
	vi	94.01 (+0.75)	**95.50 (+2.24)**	90.26	**92.68**	92.63	90.98	93.12	**93.40**
	te	91.33 (-1.93)	**95.69 (+2.43)**	83.02	85.57	**87.32**	88.99	91.87	**92.58**
	bn	94.01 (+0.75)	**95.75 (+2.49)**	86.56	89.22	**90.89**	91.14	**93.52**	93.43
Total	Average	94.58 (-0.22)	**95.52 (+0.71)**	87.00	89.36	**90.56**	89.95	92.18	**92.72**
XQuAD									
bge-m3	ar	96.29 (-0.88)	**96.72 (-0.45)**	92.24	93.08	**93.50**	92.38	94.41	**94.71**
	zh	96.13 (-1.04)	**96.70 (-0.47)**	94.04	94.03	**94.68**	93.18	94.21	**94.82**
	es	96.80 (-0.37)	**96.91 (-0.26)**	96.14	95.97	**96.14**	95.96	**96.42**	96.34
	ru	96.24 (-0.93)	**96.91 (-0.26)**	95.58	95.13	**95.83**	94.57	94.98	**95.16**
multilingual-e5	ar	94.70 (-3.32)	**96.23 (-1.79)**	87.29	87.58	**89.83**	91.22	91.69	**92.61**
	zh	95.22 (-2.80)	**96.01 (-2.01)**	89.60	90.64	**91.71**	91.02	92.44	**93.44**
	es	95.51 (-2.51)	**95.94 (-2.08)**	**96.15**	93.39	94.00	**96.41**	94.42	94.62
	ru	95.18 (-2.84)	**95.94 (-2.08)**	**93.06**	91.90	92.54	**93.22**	92.54	93.11
gte-multilingual	ar	94.70 (-3.32)	**96.23 (-1.79)**	87.29	87.58	**89.83**	91.22	91.69	**92.61**
	zh	96.84 (-1.09)	**97.43 (-0.50)**	93.98	94.27	**94.87**	92.07	93.67	**94.05**
	es	97.56 (-0.37)	**97.59 (-0.34)**	96.14	95.88	**96.45**	95.79	96.53	**96.66**
	ru	97.04 (-0.89)	**97.40 (-0.53)**	94.38	94.06	**94.57**	94.24	94.88	**95.21**
jina-v3	ar	**97.24 (+1.15)**	97.16 (+1.07)	90.58	93.25	**93.90**	93.21	95.02	**95.19**
	zh	97.31 (+1.22)	**97.33 (+1.24)**	92.65	94.72	**95.05**	92.79	95.36	**95.75**
	es	**97.36 (+1.27)**	97.17 (+1.08)	94.76	**96.00**	95.92	96.12	96.99	**97.02**
	ru	**97.25 (+1.16)**	97.19 (+1.10)	93.97	95.88	**96.06**	94.38	95.53	**95.82**
Total	Average	96.47 (-0.84)	**96.88 (-0.44)**	93.00	93.40	**94.05**	93.65	94.52	**94.89**

Table 1: Comprehensive cross-lingual evaluation results. In each cross-lingual setting, 'Lang' refers to the target language. The first word denotes the language of the passage, and the second one denotes the language of the query. The value in '()' indicates the difference in performance from the original model.

as negative samples by calculating similarity scores between the gold passage and queries.

4.3 Evaluation

Language Scope We perform downstream task evaluation on nine languages: Arabic (ar), German (de), Chinese (zh), Russian (ru), Spanish (es), Hindi (hi), Vietnamese (vi), Telugu (te) and Bengali (bn). They were chosen to provide a mix of high-, medium- and low-resource languages, typological and script diversity while satisfying the practical constraints of available evaluation datasets. We refer to German, Chinese, Russian and Spanish as high-resource, Arabic, Hindi and Vietnamese as medium-resource, and Telugu and Bengali as low-resource languages.

Cross-lingual Scenario We conduct a comprehensive evaluation across a wide range of cross-lingual scenarios. Since our aim covers cross-lingual evaluation across all directions (P_{en} - Q_{ℓ}

/ P_ℓ - Q_{en}), the same question-passage pairs must exist in multiple languages to enable the evaluation of retrieval performance across different languages, making essential to employ datasets that are fully parallel between English and target languages. To this end, we employ two cross-lingual retrieval benchmarks: Belebele (Bandarkar et al., 2024), which covers 122 language variants including English, and XQuAD (Artetxe et al., 2020), which includes 11 languages. Both benchmarks are included in the authorized evaluation framework MMTEB tasks (Enevoldsen et al., 2025), which are driven by the expansion of MTEB (Muennighoff et al., 2023). As a part of MTEB, these are widely adopted evaluation datasets in contemporary retrieval works (Chen et al., 2024; Sturua et al., 2024; Lee et al., 2025; Zhang et al., 2025). More details about benchmarks can be found in Appendix B.

For each language present in both benchmarks, we evaluate on both Belebele and XQuAD; languages exclusive to Belebele are evaluated only with that benchmark. Furthermore, we assess the preservation of the model's English retrieval capabilities by measuring the difference in English performance after the cross-lingual training. We use nDCG@10 (Järvelin and Kekäläinen, 2002) as the primary evaluation metric.

Multilingual Expansion We also evaluate CLEAR in a multilingual setup where multiple languages are concurrently learned. We construct a multilingual training set by combining 1,410 non-overlapping samples per language from a cross-lingual training dataset. Then we train the model on this combined dataset. By assessing performance in both cross-lingual and target-language-only scenarios across all languages considered in our experiments, we demonstrate the scalability of CLEAR to multilingual training scenarios.

5 Results

We first investigate the effects of CLEAR on cross-lingual scenarios under various languages and embedding models in Section 5.1. Then in Section 5.2, we examine the generalization of CLEAR within a multilingual training setup involving nine mixed languages. Finally, by conducting an ablation study of CLEAR's core strategies in Section 5.3, we demonstrate the validity of our approach. A further experiment report on all languages and metrics is available in Appendix G.

5.1 Comprehensive Cross-lingual Retrieval

Table 1 shows the effectiveness of CLEAR for cross-lingual adaptation, regardless of language or model. CLEAR consistently outperforms InfoNCE in most cases, regardless of whether the passage or query is in the target language.

Low-resource Languages Notably, this performance gap is more pronounced for low-resource languages. As shown in Table 1, on Bengali and Telugu, CLEAR achieves scores of 80.97 and 85.44 with multilingual-e5, which exceed the original model by more than 13% and surpass InfoNCE (77.14 and 81.81) by 4 points in the English-Lang setup. Similarly, for gte-multilingual, CLEAR reaches 86.31 and 86.64, showing a greater margin compared to other high-resource languages.

This is related to the performance of the original models concerning target languages. Both multilingual-e5 and gte-multilingual yield the lowest scores for low-resource languages among all models. This implies that an imbalance in training data causes a representational gap between English and the target languages. While InfoNCE seeks to address this gap solely with respect to the target language, CLEAR leverages English passages as a bridge, allowing the model to share interactions between English queries and underrepresented target language queries. As a result, CLEAR generalizes well in low-resource languages.

Generalization in Lang-English Setup We observe that CLEAR remains effective even when the passage is presented in the target language (Lang-English), a setting not accounted for during training. This stems from fundamental alignment. CLEAR does not merely learn target language expressiveness dependent on the query; rather, it enhances the fundamental representational ability between the target language and English by considering reverse directions, generalizing well to unseen target language passages. This suggests that CLEAR can be a robust training approach in real-world environments where target language passage corpora are limited for training the retriever and demonstrates its effectiveness in diverse cross-lingual scenarios.

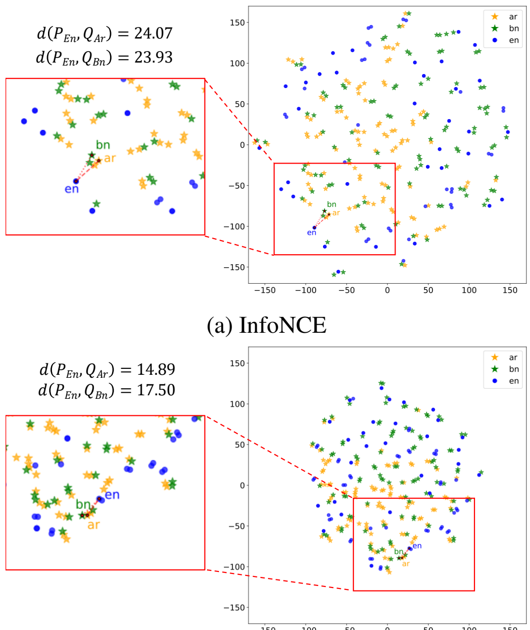
Preservation of English Ability In general, training methods that target cross-lingual alignment inevitably have a negative impact on monolingual performance. Table 1 also shows that cross-lingual training affects monolingual performance in English. However, CLEAR achieves superior perfor-

Model	Passage-Query	Method	ar	de	zh	ru	es	hi	vi	te	bn
multilingual-e5	English-Lang	InfoNCE	82.58	89.99	86.84	89.16	89.46	83.10	86.69	76.48	81.85
		CLEAR	**85.12**	**91.87**	**88.85**	**91.68**	**91.98**	**85.94**	**88.85**	**80.22**	**84.51**
	Lang-English	InfoNCE	88.08	92.29	90.02	91.39	91.58	90.97	90.44	88.63	88.93
		CLEAR	**88.89**	**93.47**	**90.78**	**92.25**	**93.14**	**92.39**	**91.47**	**89.50**	**90.68**
gte-multilingual	English-Lang	InfoNCE	86.38	92.12	91.92	91.80	92.44	89.13	91.26	83.53	84.42
		CLEAR	**87.77**	**92.92**	**92.70**	**92.56**	**93.29**	**90.13**	**91.82**	**85.19**	**85.70**
	Lang-English	InfoNCE	90.75	93.20	92.11	92.71	93.10	92.53	92.89	89.91	91.08
		CLEAR	**91.44**	**93.55**	**92.52**	**93.28**	**93.53**	**92.78**	**93.23**	**91.03**	**92.02**

Table 2: Cross-lingual evaluation results in multilingual training setup in Belebele benchmark.

Language	multilingual-e5		gte-multilingual	
	InfoNCE	CLEAR	InfoNCE	CLEAR
en	94.94	**95.61**	94.66	**95.46**
ar	86.82	**88.42**	88.37	**89.10**
de	91.15	**93.27**	92.89	**93.47**
zh	91.28	**93.05**	92.85	**93.65**
ru	92.15	**93.09**	93.07	**93.83**
es	90.91	**92.51**	92.55	**93.44**
hi	84.24	**86.07**	88.48	**89.34**
vi	90.77	**92.43**	92.18	**93.08**
te	80.67	**82.81**	83.71	**84.86**
bn	85.28	**87.56**	86.01	**87.63**

Table 3: Monolingual performances in multilingual training setup. Each row presents the result where the passage and query are in the same language.



(b) CLEAR

Figure 3: T-SNE visualization of the embeddings for multilingual-e5 with English passage and Arabic, Bengali queries after multilingual training. We randomly select 100 pairs from the Belebele and measure the distance between the embeddings of identical gold pairs.

mance in cross-lingual scenarios while reducing the leakage of the model's inherent capabilities in English or even achieving improvements. CLEAR yields a total average score of 96.88 in English results in XQuAD, reflecting a smaller decrease compared to InfoNCE's 96.47. Notably, CLEAR even improves performance for most models on the Belebele benchmark. In contrast, InfoNCE shows a greater decrease (multilingual-e5) or a marginal increase (jina-v3).

This indicates that the strategy of CLEAR helps preserve the original English alignment while also considering the target language. $\mathcal{L}_{\text{NCE}_{\text{en}}}$ encourages the model to maintain its existing English proficiency during the training, and the integration with $\mathcal{L}_{\text{CL}}$ enables the model to consider the capability in the target language. Through this joint training approach, CLEAR offers a practical solution that can address real-world needs where both cross-lingual and English performance must be considered.

5.2 Multilingual Training

We also find that CLEAR provides significant benefits in a multilingual configuration, where multiple languages are trained together. As shown in Table 2, CLEAR outperforms InfoNCE by a large margin across all languages.

Furthermore, despite targeting cross-lingual retrieval, CLEAR remains highly effective in a monolingual setup. Table 3 shows monolingual retrieval results after the multilingual training, where CLEAR consistently surpasses the baseline in all languages. This suggests that CLEAR can help enhance semantic representation within individual languages, in addition to improving alignment across languages.

These advantages can be attributed to the formation of a shared embedding space. In Figure 3, we observe that CLEAR demonstrates significantly better language-agnostic alignment for languages and achieves closer semantic proximity between the gold passage and query than InfoNCE. This

Method	ar	de	zh	ru	es	hi	vi	te	bn	Average
CLEAR$_{e5}$	83.78 / 88.23	**91.94 / 92.86**	**88.89 / 90.75**	**90.59 / 92.16**	**91.61 / 93.13**	**86.10 / 92.13**	**88.63 / 91.95**	**80.97 / 88.99**	**85.44 / 89.92**	**87.55 / 91.12**
w/o $\mathcal{L}_{\text{KL}}$	**83.82 / 88.49**	91.61 / 92.29	88.59 / 90.43	89.85 / 91.73	90.76 / 92.49	85.05 / 91.10	88.52 / 91.69	79.68 / 88.67	84.40 / 89.62	86.92 / 90.72
w/o Reverse	82.81 / 87.49	91.31 / 92.47	87.44 / 89.69	89.77 / 90.90	89.91 / 92.01	84.12 / 90.84	87.41 / 91.36	78.21 / 87.96	82.77 / 88.89	85.97 / 90.18
w/o $\mathcal{L}_{\text{NCE}_{\text{en}}}$	81.95 / 87.24	90.26 / 91.55	86.98 / 89.09	89.19 / 90.98	89.88 / 91.45	83.71 / 90.21	87.67 / 90.38	79.00 / 87.21	83.58 / 88.95	85.80 / 89.67
CLEAR$_{gte}$	**87.91 / 91.46**	**93.26 / 93.95**	**92.67 / 92.80**	**92.94 / 93.33**	**93.28 / 93.87**	**89.96 / 93.05**	**92.23** / 93.52	**86.31** / 90.92	**86.64 / 92.14**	**90.58 / 92.78**
w/o $\mathcal{L}_{\text{KL}}$	87.13 / 91.16	92.90 / 93.70	92.60 / 92.48	92.50 / 93.28	92.92 / 93.46	89.59 / 92.89	91.80 / 93.27	85.01 / **91.18**	85.79 / 91.70	90.03 / 92.57
w/o Reverse	87.19 / 90.93	93.01 / 93.74	92.46 / 92.35	92.54 / 93.32	93.16 / 93.63	89.73 / 92.68	91.67 / **93.62**	85.08 / 91.01	85.91 / 91.47	90.08 / 92.53
w/o $\mathcal{L}_{\text{NCE}_{\text{en}}}$	84.36 / 88.29	90.96 / 91.30	89.59 / 89.93	90.17 / 90.65	91.46 / 92.10	86.37 / 90.67	89.79 / 90.87	81.41 / 87.42	85.17 / 90.77	87.70 / 90.22

Table 4: Ablation results (English-Lang / Lang-English) for key components in CLEAR on multilingual-e5 and gte.

supports our claim that CLEAR constructs a robust language-agnostic space by narrowing the fundamental distance between language spaces and mapping them into a similar embedding space via passage bridge. CLEAR can be extended to encompass multilingualism, showing broader applicability.

5.3 Ablation Study

In this section, we analyze the influence of core strategies in CLEAR. In the cross-lingual scenario, we sequentially remove each proposed strategy from CLEAR and evaluate the impact on performance. Table 4 shows that each has a substantial impact on performance and also proves that the three strategies work in synergy for enhancing overall cross-lingual retrieval performance.

Validity of Passage Bridge We find that the passage bridge plays a vital role in cross-lingual alignment. In general, when the loss function jointly optimizes both the English training ($\mathcal{L}_{\text{NCE}_{\text{en}}}$) and the cross-lingual training objective ($\mathcal{L}_{\text{CL}}$), the model's capacity is distributed between these objectives, which may reduce the concentration on cross-lingual alignment during training.

However, the exclusion of the passage bridge via the removal of the English objective term leads to the most notable decrease in performance. We ascribe this phenomenon to the shared passages, which function as bridges that interlink the representation spaces of different languages and facilitate the sharing of semantic information. By positioning target languages closer to English within the embedding space via the passage bridge, CLEAR benefits from cross-lingual alignment.

Importance of Reversal Scheme Training with the conventional direction (where the query serves as the anchor in $\mathcal{L}_{\text{CL}}$) instead of the reversal scheme consistently degrades performance in the majority of cases. This is surprising given that reverse training does not directly align with the retrieval task's objective of finding relevant passages for a query. Also, this approach weakens the synergy with the passage bridge. These findings imply that aligning gold pairs through various perspectives beyond the conventional direction, in conjunction with a passage bridge, enhances the robustness of cross-lingual alignment.

Moreover, CLEAR's efficacy does not simply arise from increased computational demands. Substituting the reversal scheme with conventional direction also maintains all loss terms, leaving the computation amount identical to the case where the reversal scheme is applied. Given this, we can attribute CLEAR's benefit to the proposed reversal training scheme itself.

KL-Divergence The introduction of $\mathcal{L}_{\text{KL}}$ leads to a slight improvement in overall performance. This gain stems from harmonizing the patterns of similarity observed among English pairs with those between English passages and target language queries within each training batch. By ingraining a more fundamental understanding of cross-lingual semantic equivalence, $\mathcal{L}_{\text{KL}}$ complements the instance-level alignments fostered by other strategies.

6 Conclusion

We propose CLEAR, an innovative approach designed to enhance cross-lingual alignment within the realm of cross-lingual information retrieval. Our reversal training scheme, coupled with several strategies, promotes diverse interaction between English and the target language. Through the experiments on nine languages, CLEAR outperforms the standard approach, demonstrating notable robustness and adaptability in diverse linguistic environments. Furthermore, we highlight the efficacy of CLEAR in extending beyond cross-lingual to multilingual setups, showcasing its utility across broader scenarios. Our study suggests that CLEAR offers a promising avenue for future research and application in global information retrieval systems, which can be directly integrated into existing dense retrieval frameworks. For future work, we plan to explore the expansion of its application to other text embedding tasks beyond retrieval.

Limitation

Our study primarily focused on cross-lingual scenarios involving English and other target languages. Although broader coverage could be achieved by considering scenarios where both the passage and query languages are non-English, this was challenging for us due to the limited resources of the parallel dataset, especially at the passage level. Nevertheless, by achieving consistent improvements in cross-lingual retrieval adopted in most previous works, motivated by the practical demand for English resources in real-world applications, we were able to demonstrate strong generalization across a wide range of languages.

Ethics Statement

In this research, we utilized only publicly available datasets and models. All data used for training and evaluation were sourced from open-access repositories and applied in accordance with their respective licenses. We strictly adhered to the copyright, licensing terms, and guidelines of the original works and datasets, including those pertaining to language resources and translated data. We confirm that there were no distinct ethical concerns related to the collection, use, or processing of the datasets and resources used in this study.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2021R1A6A1A03045425). This work was supported by the Commercialization Promotion Agency for R&D Outcomes (COMPA) grant funded by the Korea government (Ministry of Science and ICT)(2710086166). This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (RS-2024-00398115, Research on the reliability and coherence of outcomes produced by Generative AI). This research was supported by Culture, Sports and Tourism R&D Program through the korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Toruism in 2024 (Project Name : Development of generative AI-based publishing content analysis and sharing platform technology to respond to changes in the publishing environment. Project Number : : RS-2024-00442061).

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. XOR QA: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 547–564, Online. Association for Computational Linguistics.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Nadezhda Chirkova, David Rau, Hervé Déjean, Thibault Formal, Stéphane Clinchant, and Vassilina Nikoulina. 2024. Retrieval-augmented generation in multilingual settings. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 177–188, Bangkok, Thailand. Association for Computational Linguistics.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzeminski, Genta Indra Winata, et al. 2025. Mmteb: Massive multilingual text embedding benchmark. In *International Conference on Learning Representations*. International Conference on Learning Representations.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

P Moreira Gabriel de Souza, Radek Osmulski, Mengyao Xu, Ronay Ak, Benedikt Schifferer, and Even Oldridge. 2024. Nv-retriever: Improving text embedding models with effective hard-negative mining. *arXiv preprint arXiv:2407.15831*, 1.

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

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*.

Seongtae Hong, Youngjoon Jang, Jungseob Lee, Hyeonseok Moon, and Heuiseok Lim. 2026. Improving semantic proximity in information retrieval through cross-lingual alignment. *arXiv preprint arXiv:2604.05684*.

Zhiqi Huang, Puxuan Yu, and James Allan. 2023a. Improving cross-lingual information retrieval on low-resource languages via optimal transport distillation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 1048–1056.

Zhiqi Huang, Puxuan Yu, and James Allan. 2023b. Improving cross-lingual information retrieval on low-resource languages via optimal transport distillation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23, page 1048–1056. ACM.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606, Toronto, Canada. Association for Computational Linguistics.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

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

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

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020a. Mlqa: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330.

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

Yulong Li, Martin Franz, Md Arafat Sultan, Bhavani Iyer, Young-Suk Lee, and Avirup Sil. 2022a. Learning cross-lingual IR from an English retriever. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4428–4436, Seattle, United States. Association for Computational Linguistics.

Yulong Li, Martin Franz, Md Arafat Sultan, Bhavani Iyer, Young-Suk Lee, and Avirup Sil. 2022b. Learning cross-lingual ir from an english retriever. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4428–4436.

Peerat Limkonchotiwat, Wuttikorn Ponwitayarat, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, and Sarana Nutanong. 2022. Cl-relkt: Cross-lingual language knowledge transfer for multilingual retrieval question answering. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2141–2155.

Robert Litschko, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. 2018. Unsupervised cross-lingual information retrieval using monolingual data only. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1253–1256.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037.

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

Shramay Palta, Haozhe An, Yifan Yang, Shuaiyi Huang, and Maharshi Gor. 2022. Investigating information inconsistency in multilingual open-domain question answering. *arXiv preprint arXiv:2205.12456*.

Jeonghyun Park and Hwanhee Lee. 2025. Investigating language preference of multilingual rag systems. *arXiv preprint arXiv:2502.11175*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. *Preprint*, arXiv:2004.09813.

Peng Shi, Rui Zhang, He Bai, and Jimmy Lin. 2021. Cross-lingual training of dense retrievers for document retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 251–253, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wang Shuaibo, Di Hui, Huang Hui, Lai Siyu, Ouchi Kazushige, Chen Yufeng, and Xu Jinan. 2022. Supervised contrastive learning for cross-lingual transfer learning. In *Proceedings of the 21st Chinese National Conference on Computational Linguistics*, pages 884–895, Nanchang, China. Chinese Information Processing Society of China.

Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, R. Rana, and Suranga Nanayakkara. 2022. Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 11:1–17.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, et al. 2024. jina-embeddings-v3: Multilingual embeddings with task lora. *arXiv preprint arXiv:2409.10173*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916.

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

Yaushian Wang, Ashley Wu, and Graham Neubig. 2022. English contrastive learning can learn universal cross-lingual sentence embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9122–9133, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zilong Wang, Zifeng Wang, Long Le, Steven Zheng, Swaroop Mishra, Vincent Perot, Yuwei Zhang, Anush Mattapalli, Ankur Taly, Jingbo Shang, Chen-Yu Lee, and Tomas Pfister. 2025. Speculative RAG: Enhancing retrieval augmented generation through drafting. In *The Thirteenth International Conference on Learning Representations*.

Eugene Yang, Dawn Lawrie, James Mayfield, Douglas W. Oard, and Scott Miller. 2024a. Translate-distill: Learning cross-language dense retrieval by translation and distillation. *Preprint*, arXiv:2401.04810.

Jinrui Yang, Fan Jiang, and Timothy Baldwin. 2024b. Language bias in multilingual information retrieval: The nature of the beast and mitigation methods. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 280–292, Miami, Florida, USA. Association for Computational Linguistics.

Bryan Zhang and Amita Misra. 2022. Machine translation impact in E-commerce multilingual search. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 99–109, Abu Dhabi, UAE. Association for Computational Linguistics.

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. 2024a. Jasper and stella: distillation of sota embedding models. *arXiv preprint arXiv:2412.19048*.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024b. 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*, pages 1393–1412, Miami, Florida, US. Association for Computational Linguistics.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. TyDi: A multi-lingual benchmark for dense retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 127–137, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. MIRACL: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Shengyao Zhuang, Linjun Shou, and Guido Zuccon. 2023. Augmenting passage representations with query generation for enhanced cross-lingual dense retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1827–1832.

A Training Details

We leveraged the Pytorch framework (Paszke et al., 2019) and the Sentence-Transformers library[2]. For the loss function, we use MultipleNegativesRankingLoss[3] as a baseline, which incorporates positive passages with negative samples (Henderson et al., 2017). We used the cached version of loss provided by sentence-transformers for memory efficiency (Gao et al., 2021).

We conducted all experiments under a uniform setup across all languages and models, employing four NVIDIA A6000 GPUs to perform fine-tuning. For hyper-parameters, we set a maximum sequence length of 512 and utilized a batch size of 64 with mini-batch size 32. The learning rate was established at 5e-5, using a cosine scheduler and a warmup ratio of 0.05 for stable training. We reported our experimental results by adopting the last checkpoint after training all models for only one epoch. The details of embedding models employed in our study are shown in Table 5.

[2]`https://github.com/UKPLab/sentence-transformers`

[3]`https://github.com/UKPLab/sentence-transformers/blob/master/sentence_transformers/losses/MultidpleNegativesRankingLoss.py`

Model	Details
bge-m3	`BAAI/bge-m3` (Chen et al., 2024)
multilingual-e5	`intfloat/multilingual-e5-base` (Wang et al., 2024b)
gte-multilingual	`Alibaba-NLP/gte-multilingual-base` (Zhang et al., 2024b)
jina-v3	`jinaai/jina-embeddings-v3` (Sturua et al., 2024)

Table 5: Embedding models details.

B Evaluation Benchmark

In this work, we leverage multilingual Question Answering (QA) datasets with parallel constructions, repurposed as retrieval tasks, to systematically assess cross-lingual retrieval performance in all directions. Since these datasets are originally designed for QA, the corresponding passages serve as exact gold labels within the retrieval framework. Consequently, datasets developed for QA are widely adopted for the evaluation of the retriever in the current literature (Enevoldsen et al., 2025; Lee et al., 2025; Zhang et al., 2025).

Belebele Belebele is a high-quality, professionally translated multilingual QA dataset featuring a broad range of language pairs. All translations were conducted by native speakers proficient in English, thereby capturing both contextual meaning and cultural nuances. Owing to these strengths, Belebele offers diverse and realistic multilingual retrieval scenarios, enabling detailed comparative analyses of retrieval models across different languages.

XQuAD XQuAD is a multilingual QA resource based on SQuAD 1.1 (Rajpurkar et al., 2016), comprising fully parallel question-answer pairs spanning 13 languages, including English. The dataset was translated by professional translators, ensuring strict one-to-one mapping between documents and queries across languages. This rigorous translation approach preserves both linguistic characteristics and semantic content in each target language, rendering XQuAD especially well-suited for evaluating the stability of embedding models with respect to linguistic variation in cross-lingual contexts.

We also considered a wide range of datasets for cross-lingual evaluation. However, most retrieval datasets either have quality issues that affect their reliability or do not align well with the objectives of our study. For example, since Mr.TyDi (Zhang et al., 2021) and MIRACL (Zhang et al., 2023)

Lang	bge-m3		multilingual-e5		gte-multilingual		jina-v3	
	InfoNCE	CLEAR	InfoNCE	CLEAR	InfoNCE	CLEAR	InfoNCE	CLEAR
Cross-lingual Scenario								
ar	72.44	**72.95**	64.11	**64.99**	66.22	**66.75**	71.77	**71.89**
ru	**76.93**	76.77	70.84	**71.83**	74.36	**74.76**	76.99	**77.59**
te	68.78	**69.66**	55.56	**59.42**	68.36	**69.26**	67.63	**67.83**
bn	69.96	**69.90**	60.30	**62.24**	63.67	**64.30**	68.59	**68.95**
Avg.	72.03	**72.32**	62.20	**64.12**	68.15	**68.77**	71.25	**71.57**
Multilingual Expansion								
ar	72.41	**72.96**	64.19	**65.46**	65.71	**66.43**	71.02	**71.76**
ru	**76.69**	76.62	70.01	**71.52**	74.29	**74.83**	**77.07**	77.05
te	69.61	**70.42**	54.96	**57.63**	68.33	**68.98**	67.45	**67.98**
bn	70.11	**70.43**	59.78	**60.70**	63.32	**64.14**	68.38	**68.71**
Avg.	72.21	**72.61**	62.23	**63.83**	67.41	**68.09**	70.98	**71.38**

Table 6: Cross-lingual evaluation results in XOR-TyDi. In the cross-lingual scenario, the results are from models fine-tuned individually for each language. In the multilingual expansion, the model is trained considering all languages equally, as in our main experiments.

are not fully parallel with English, they cannot be used for cross-lingual evaluation. In this regime, we carefully select Belebele and XQuAD as our main evaluation datasets.

C XOR-TyDi

Additionally, to further substantiate our evaluation, we report the cross-lingual evaluation results in the common retrieval task, XOR-TyDi (Asai et al., 2021). We utilize the XOR-Retrieve task from XOR-TyDi and conduct supplementary experiments on the four languages that overlap with those covered in our study, out of the seven languages included in the benchmark. The evaluation is performed using an English passage and target language queries for both the cross-lingual and multilingual training scenarios, as XOR-TyDi does not provide support for passages in target languages.

As can be seen in the Table 6, the result is in line with our main experimental result. This emphasizes again that CLEAR is indeed an effective training approach for cross-lingual retrieval tasks.

D Training Example

Compared to the standard InfoNCE, CLEAR requires translated queries that are parallel to the English query for training. CLEAR does not require target language passages. Thus, Q_{en} is used as the anchor to increase similarity with P^+_{en} and decrease similarity with P^-_{en} in $L_{NCE_{en}}$. In L_{CL}, P^+_{en} serves as the anchor, with Q_ℓ treated as a positive and Q^-_ℓ as a hard negative sample in contrastive learning. An example of inputs for the training is shown in Table 7.

Q_{en}	When was Winehouse quoted about the album release?
P^+_{en}	Winehouse and Ronson contributed a cover of Lesley Gore's "It's My Party" to the Quincy Jones tribute album Q Soul Bossa Nostra released 9 November 2010 ... In July 2010, Winehouse was quoted as saying her next album would be released no later than January 2011, saying "It's going to be very much the same as my second album, where there's a lot of jukebox stuff and songs that are... just jukebox, really."
P^-_{en}	On the other hand, Arabic is divided into over 27 dialects. Almost every Arab state has at least one local dialect of its own ...
Q_ℓ	zh: 懷恩豪斯何時提到這張專輯的發行? de: Wann hat Winehouse die Veröffentlichung dieses Albums erwähnt? ru: Когда Уайнхаус упомянула о выпуске этого альбома? es: ¿Cuándo mencionó Winehouse el lanzamiento de este álbum?
Q^-_ℓ	zh: 超級女孩電視劇什時候首次播出? de: Was tut die Polizei, um Sarah zu beschützen? ru: Когда родился Жак-Луи Давид? es: ¿Cuántos huevos pueden poner las aves en las colinas de Sri Lanka?

Table 7: Example of inputs in the training.

Model	Lang	English-Lang		Lang-English	
		InfoNCE	CLEAR	InfoNCE	CLEAR
bge-m3	ar	88.66	**89.23**	90.87	**91.17**
	zh	91.86	**92.40**	91.63	**92.10**
	es	92.66	**93.17**	93.07	**93.44**
	de	93.77	**94.30**	93.99	**94.66**
	ru	93.72	**94.17**	93.42	**93.54**
	hi	89.15	**89.91**	93.33	**93.47**
	vi	92.79	**93.34**	92.77	**93.35**
	bn	89.49	**90.43**	92.24	**93.04**
multilingual-gte	ar	86.49	**88.10**	90.40	**91.62**
	zh	92.34	**92.83**	91.99	**92.70**
	es	93.05	**93.52**	93.41	**93.93**
	de	92.78	**93.42**	93.53	**94.00**
	ru	92.34	**92.88**	93.22	**93.78**
	hi	89.17	**89.92**	92.66	**93.22**
	vi	91.46	**92.11**	92.99	**93.55**
	bn	85.29	**86.28**	90.98	**91.91**

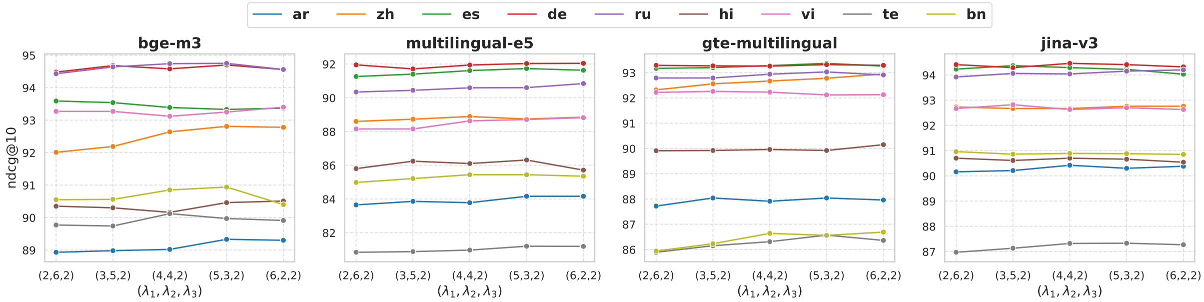
Table 8: Cross-lingual evaluation result in Belebele trained with m2m100 translated dataset.

E Robustness Analysis on Translation Quality

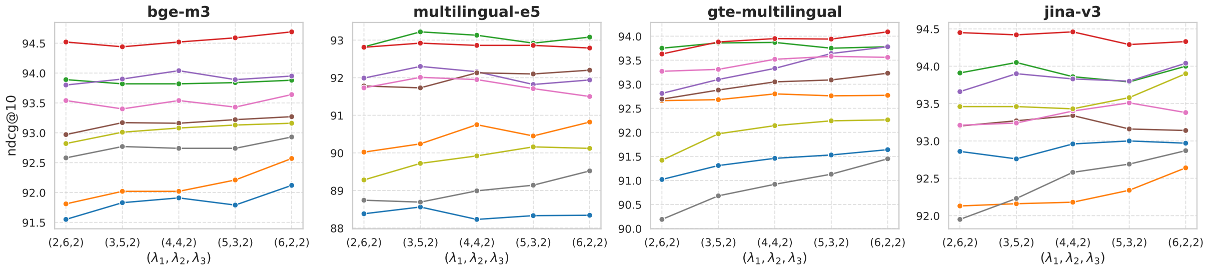
To evaluate the robustness of our approach against noise that may influence the results because of machine-translated training data, we conduct further experiments using another translation model, m2m100 (Fan et al., 2021)[4], within the same pipeline (except Telugu language, since m2m100 does not support). As shown in Table 8, CLEAR consistently yields performance improvements at the same level of data quality, aligning with our main result.

From the perspective of robustness, we observe no significant performance difference between models trained on m2m100-translated data and those trained on NLLB-translated data. Since CLEAR requires only translated queries rather

[4]https://huggingface.co/facebook/m2m100_1.2B



(a) English - Lang



(b) Lang - English

Figure 4: Performance variation depending on the loss component weights (λ_1, λ_2, λ_3) across multiple languages.

than complex passages, potential translation errors are likely minimized. Our result demonstrates that CLEAR can be an essentially robust approach with respect to variations in translation quality, thereby enhancing its practical applicability in real-world scenarios where human translation is costly.

F Loss Parameter Sensitivity Analysis

We report the nDCG@10 on Belebele for the settings of English passage-target language query and target language passage-English query in Figure 4. While exhaustively exploring every possible combination of values for all three components is computationally prohibitive, we focus our analysis on the components hypothesized to exert the most significant influence on CLEAR's core objectives: λ_1, which controls the English NCE term, and λ_2, which governs the cross-lingual reversal loss.

Overall, the results indicate that variations in the λ values do not cause substantial performance fluctuations, suggesting that the proposed method is relatively robust to the choice of loss-weight parameters. In the case of English passages with target-language queries (Figure 4 (a)), the combinations (4, 4, 2) and (5, 3, 2) yield the best average performance across all models. However, we observe that the Lang-English setup benefits from a stronger alignment of English representations (Figure 4 (b)). This pattern highlights that our bridging strategy, which adopts the English representation as a semantic anchor between languages, plays a crucial role in enhancing the passage representation capability of different languages. Moreover, these results suggest that there remains room for further improvement by refining the loss weight configuration, particularly in the Lang-English setup.

G Extended Evaluation Results

Among the selected nine languages, Belebele encompasses all languages, and XQuAD includes only five. We report the extended evaluation results in this section. Table 9, 10 presents the results for each target language under the cross-lingual setup, and Table 11, 12 illustrates the performance in the multilingual training setup.

Model	Language	English-Lang						Lang-English					
		nDCG@5			nDCG@10			nDCG@5			nDCG@10		
		Base	InfoNCE	CLEAR	Base	InfoNCE	CLEAR	Base	InfoNCE	CLEAR	Base	InfoNCE	CLEAR
bge-m3	ar	86.62	88.22	**88.29**	87.77	**89.06**	89.02	86.42	90.66	**90.93**	87.63	91.61	**91.91**
	zh	89.70	90.93	**92.14**	90.35	91.69	**92.64**	89.08	90.97	**91.51**	89.81	91.63	**92.02**
	es	91.48	92.61	**93.05**	91.99	92.81	**93.39**	91.47	92.99	**93.59**	92.05	93.50	**93.82**
	de	92.20	93.62	**94.31**	92.81	93.98	**94.58**	92.47	93.87	**94.30**	92.90	94.30	**94.52**
	ru	92.61	93.28	**94.30**	92.92	93.67	**94.74**	91.18	92.85	**93.61**	91.69	93.55	**94.04**
	hi	85.78	88.15	**89.72**	86.90	89.14	**90.16**	89.57	92.35	**92.62**	90.39	92.86	**93.16**
	vi	90.02	92.26	**92.79**	90.70	92.58	**93.12**	89.73	92.53	**93.10**	90.44	93.04	**93.54**
	te	85.98	88.33	**89.49**	87.12	89.07	**90.12**	88.27	91.41	**92.25**	89.19	92.08	**92.74**
	bn	86.49	88.95	**90.12**	87.58	89.79	**90.85**	89.18	91.65	**92.71**	90.17	92.37	**93.08**
multilingual-e5	ar	76.23	80.68	**82.79**	77.65	82.27	**83.78**	82.51	87.24	**87.39**	83.56	87.96	**88.23**
	zh	80.04	85.82	**87.83**	81.16	87.00	**88.89**	84.43	89.15	**89.87**	85.49	90.02	**90.75**
	es	88.53	89.01	**91.08**	89.46	89.84	**91.61**	91.92	91.79	**92.85**	92.43	92.15	**93.13**
	de	89.96	90.26	**91.44**	90.77	90.97	**91.94**	91.53	91.78	**92.17**	92.16	92.68	**92.86**
	ru	86.48	88.82	**89.79**	87.29	89.60	**90.59**	87.73	90.72	**91.48**	88.98	91.55	**92.16**
	hi	79.46	81.82	**84.88**	81.05	83.11	**86.10**	87.65	90.34	**91.52**	88.44	91.08	**92.13**
	vi	85.05	86.18	**87.76**	86.03	87.11	**88.63**	87.52	91.21	**91.44**	88.58	91.81	**91.95**
	te	67.87	75.08	**79.40**	69.90	77.14	**80.97**	83.57	86.90	**88.36**	84.88	88.05	**88.99**
	bn	73.83	80.34	**84.62**	75.55	81.81	**85.44**	82.95	88.22	**89.22**	84.31	88.91	**89.92**
gte-multilingual	ar	82.60	86.27	**87.32**	83.81	86.92	**87.91**	86.92	90.58	**91.06**	87.99	90.97	**91.46**
	zh	89.13	91.82	**92.11**	89.86	92.30	**92.67**	90.92	91.50	**92.14**	91.51	92.05	**92.80**
	es	90.78	92.09	**92.85**	91.71	92.86	**93.28**	89.73	93.08	**93.40**	90.20	93.55	**93.87**
	de	90.56	92.51	**93.04**	91.21	92.92	**93.26**	88.63	93.11	**93.63**	89.42	93.45	**93.95**
	ru	90.27	91.95	**92.50**	91.11	92.46	**92.94**	88.57	92.28	**93.04**	89.56	92.69	**93.33**
	hi	86.49	88.60	**89.39**	87.51	89.34	**89.96**	88.96	91.88	**92.60**	89.55	92.45	**93.05**
	vi	88.68	91.01	**91.68**	89.37	91.59	**92.23**	89.88	92.65	**93.05**	90.48	93.14	**93.52**
	te	79.13	83.53	**85.00**	80.70	84.77	**86.31**	87.49	89.18	**90.41**	88.46	89.96	**90.92**
	bn	80.76	83.89	**85.65**	82.13	85.18	**86.64**	87.09	90.47	**91.68**	87.81	91.03	**92.14**
jina-v3	ar	85.85	88.43	**89.52**	86.84	89.37	**90.42**	89.12	91.63	**92.31**	89.86	92.45	**92.96**
	zh	88.93	91.24	**92.08**	89.46	91.90	**92.67**	88.88	91.06	**91.65**	89.64	91.74	**92.18**
	es	90.88	93.14	**94.00**	91.40	93.57	**94.29**	91.95	93.46	**93.59**	92.64	**93.88**	93.86
	de	91.20	93.42	**93.91**	91.75	93.88	**94.46**	91.68	93.61	**94.05**	92.39	94.12	**94.46**
	ru	91.53	93.39	**93.56**	91.85	93.77	**94.04**	90.82	92.98	**93.46**	91.66	93.46	**93.83**
	hi	86.94	89.21	**90.34**	87.74	89.90	**90.70**	90.77	92.52	**93.01**	91.50	93.02	**93.34**
	vi	89.39	91.76	**92.01**	90.26	**92.68**	92.63	90.10	92.90	**93.05**	90.98	93.12	**93.40**
	te	81.49	84.35	**86.39**	83.02	85.57	**87.32**	88.15	91.33	**92.10**	88.99	91.87	**92.58**
	bn	85.45	88.27	**89.94**	86.56	89.22	**90.89**	90.48	93.05	**92.95**	91.14	93.52	**93.43**

Table 9: Results on all languages under cross-lingual scenario in Belebele.

Model	Language	English-Lang						Lang-English					
		nDCG@5			nDCG@10			nDCG@5			nDCG@10		
		Base	InfoNCE	CLEAR	Base	InfoNCE	CLEAR	Base	InfoNCE	CLEAR	Base	InfoNCE	CLEAR
bge-m3	ar	91.84	92.69	**93.24**	92.24	93.08	**93.50**	91.91	94.03	**94.44**	92.38	94.41	**94.71**
	zh	93.85	93.66	**94.34**	94.04	94.03	**94.68**	92.74	93.80	**94.50**	93.18	94.21	**94.82**
	es	**95.98**	95.86	95.97	96.14	95.97	**96.14**	95.80	**96.20**	96.09	95.96	**96.42**	96.34
	de	95.29	95.64	**95.91**	95.52	95.75	**96.04**	95.89	95.53	**95.91**	**96.17**	95.67	95.93
	ru	95.29	94.91	**95.67**	95.58	95.13	**95.83**	94.17	94.73	**94.83**	94.57	94.98	**95.16**
multilingual-e5	ar	86.36	86.88	**89.31**	87.29	87.58	**89.83**	90.82	91.27	**92.36**	91.22	91.69	**92.61**
	zh	88.65	90.22	**91.27**	89.60	90.64	**91.71**	90.47	92.18	**93.36**	91.02	92.44	**93.44**
	es	**96.01**	93.06	93.75	**96.15**	93.39	94.00	**96.25**	94.18	94.39	**96.41**	94.42	94.62
	de	**94.84**	92.01	92.49	**95.18**	92.39	92.93	**95.41**	93.45	93.49	**95.63**	93.76	93.79
	ru	**92.50**	91.42	92.09	**93.06**	91.90	92.54	**92.88**	92.13	92.62	**93.22**	92.54	93.11
gte-multilingual	ar	86.65	87.99	**89.27**	87.44	88.59	**89.78**	91.29	92.83	**93.34**	91.89	93.26	**93.79**
	zh	93.70	94.08	**94.60**	93.98	94.27	**94.87**	91.61	93.43	**93.74**	92.07	93.67	**94.05**
	es	95.85	95.70	**96.29**	96.14	95.88	**96.45**	95.48	96.36	**96.55**	95.70	96.53	**96.66**
	de	94.81	94.37	**94.97**	95.11	94.59	**95.37**	94.47	95.42	**95.97**	94.75	95.73	**96.11**
	ru	**94.23**	93.58	94.22	94.38	94.06	**94.57**	93.82	94.52	**94.85**	94.24	94.88	**95.21**
jina-v3	ar	90.01	92.92	**93.68**	90.58	93.25	**93.90**	92.83	94.82	**95.11**	93.21	95.02	**95.19**
	zh	92.32	94.49	**94.77**	92.65	94.72	**95.05**	92.37	95.11	**95.59**	92.79	95.36	**95.75**
	es	94.37	**95.76**	95.75	94.76	**96.00**	95.92	95.91	96.75	**96.88**	96.12	96.90	**97.02**
	de	94.52	96.02	**96.23**	94.69	96.16	**96.29**	95.15	96.35	**96.61**	95.30	96.44	**96.69**
	ru	93.77	95.76	**95.92**	93.97	95.88	**96.06**	94.06	95.34	**95.57**	94.38	95.53	**95.82**

Table 10: Results on all languages under cross-lingual scenario in XQuAD.

Model	Language	English-Lang						Lang-English					
		nDCG@5			nDCG@10			nDCG@5			nDCG@10		
		Base	InfoNCE	CLEAR	Base	InfoNCE	CLEAR	Base	InfoNCE	CLEAR	Base	InfoNCE	CLEAR
bge-m3	ar	86.62	88.85	**89.31**	87.77	89.56	**89.97**	86.42	90.35	**91.01**	87.63	91.30	**91.90**
	zh	89.70	92.31	**92.66**	90.35	92.92	**93.10**	89.08	91.55	**92.23**	89.81	92.10	**92.67**
	es	91.48	92.03	**93.22**	91.99	92.66	**93.44**	91.47	92.60	**93.29**	92.05	93.26	**93.81**
	de	92.20	93.44	**94.40**	92.81	93.77	**94.59**	92.47	93.72	**93.98**	92.90	94.11	**94.53**
	ru	92.61	93.43	**94.03**	92.92	93.94	**94.43**	91.18	92.72	**93.44**	91.69	93.14	**93.90**
	hi	85.78	88.64	**89.48**	86.90	89.39	**90.27**	89.57	92.52	**93.15**	90.39	93.06	**93.61**
	vi	90.02	92.39	**93.03**	90.70	92.58	**93.12**	89.73	92.54	**93.25**	90.44	93.04	**93.54**
	te	85.98	88.26	**89.09**	92.58	92.94	**93.50**	88.27	91.78	**92.63**	93.04	93.08	**93.62**
	bn	86.49	89.03	**89.81**	87.58	90.00	**90.68**	89.18	91.67	**92.82**	90.17	92.38	**93.16**
multilingual-e5	ar	76.23	81.37	**83.98**	77.65	82.58	**85.12**	82.51	87.19	**88.17**	83.56	88.08	**88.89**
	zh	80.04	85.70	**88.10**	81.16	86.84	**88.85**	84.43	89.18	**89.98**	85.49	90.02	**90.78**
	es	88.53	88.42	**91.50**	89.46	89.46	**91.98**	91.92	90.84	**92.74**	92.43	91.58	**93.14**
	de	89.96	89.23	**91.32**	90.77	89.99	**91.87**	91.53	91.67	**92.97**	92.16	92.29	**93.47**
	ru	86.48	88.68	**90.91**	87.29	89.16	**91.68**	87.73	90.64	**91.63**	88.98	91.39	**92.25**
	hi	79.46	82.08	**84.83**	81.05	83.10	**85.94**	87.65	90.31	**91.81**	88.44	90.97	**92.39**
	vi	85.05	85.46	**88.31**	86.03	86.69	**88.85**	87.52	89.79	**90.95**	84.88	90.44	**91.47**
	te	67.87	74.88	**78.65**	69.90	76.48	**80.22**	83.57	87.61	**88.96**	88.58	88.63	**89.50**
	bn	73.83	80.33	**83.62**	75.55	81.85	**84.51**	82.95	88.17	**90.17**	84.31	88.93	**90.68**
gte-multilingual	ar	82.60	85.71	**87.33**	83.81	86.38	**87.77**	86.92	90.06	**90.87**	87.99	90.75	**91.44**
	zh	89.13	91.20	**92.19**	89.86	91.92	**92.70**	90.92	91.46	**91.97**	91.51	92.11	**92.52**
	es	90.78	91.49	**92.64**	91.71	92.44	**93.29**	89.73	92.63	**93.21**	90.20	93.10	**93.53**
	de	90.56	91.65	**92.60**	91.21	92.12	**92.92**	88.63	92.87	**93.22**	89.42	93.20	**93.55**
	ru	90.27	91.11	**92.13**	91.11	91.80	**92.56**	88.57	92.15	**92.83**	89.56	92.71	**93.28**
	hi	86.49	88.46	**89.59**	87.51	89.13	**90.13**	88.96	91.99	**92.45**	89.55	92.53	**92.78**
	vi	88.68	90.58	**91.20**	89.37	91.26	**91.82**	89.88	92.47	**92.98**	90.48	92.89	**93.23**
	te	79.13	82.51	**84.03**	80.70	83.53	**85.19**	87.49	89.30	**90.32**	90.48	89.91	**91.03**
	bn	80.76	83.33	**84.78**	82.13	84.42	**85.70**	87.09	90.42	**91.50**	87.81	91.08	**92.02**
jina-v3	ar	85.85	88.22	**89.42**	86.84	88.88	**90.06**	89.12	91.84	**92.39**	89.86	92.37	**92.98**
	zh	88.93	91.77	**92.37**	89.46	92.54	**93.00**	88.88	91.48	**91.91**	89.64	92.12	**92.47**
	es	90.88	92.94	**93.71**	91.40	93.23	**94.01**	91.95	93.19	**93.62**	92.64	93.61	**93.98**
	de	91.20	93.43	**93.82**	91.75	93.86	**94.31**	91.68	93.57	**93.94**	92.39	94.04	**94.27**
	ru	91.53	93.59	**93.87**	91.85	93.90	**94.24**	90.82	**93.26**	93.19	91.66	**93.78**	93.48
	hi	86.94	88.97	**90.08**	87.74	89.59	**90.51**	90.77	92.53	**92.60**	91.50	92.96	**93.08**
	vi	89.39	91.97	**92.35**	90.26	92.84	**93.02**	90.10	92.77	**92.93**	90.98	93.09	**93.26**
	te	81.49	83.37	**85.19**	90.26	92.84	**93.02**	88.15	91.40	**91.88**	88.99	93.09	**93.26**
	bn	85.45	88.04	**89.79**	86.56	88.96	**90.50**	90.48	92.91	**93.24**	91.14	93.27	**93.76**

Table 11: Results on all languages under the multilingual training setup in Belebele.

Model	Language	English-Lang						Lang-English					
		nDCG@5			nDCG@10			nDCG@5			nDCG@10		
		Base	InfoNCE	CLEAR	Base	InfoNCE	CLEAR	Base	InfoNCE	CLEAR	Base	InfoNCE	CLEAR
bge-m3	ar	91.84	92.29	**93.19**	92.24	92.53	**93.46**	91.91	94.12	**94.26**	92.38	94.52	**94.62**
	zh	93.85	**94.22**	94.21	94.04	94.50	**94.58**	92.74	93.53	**93.98**	93.18	93.83	**94.26**
	es	95.98	95.12	**96.04**	96.14	95.37	**96.29**	95.80	95.68	**96.03**	95.96	95.95	**96.26**
	de	95.29	95.06	**95.65**	95.52	95.22	**95.81**	**95.89**	95.29	95.81	**96.17**	95.46	95.91
	ru	95.29	95.01	**95.40**	95.58	95.15	**95.73**	94.17	94.26	**94.50**	94.57	94.56	**94.88**
multilingual-e5	ar	86.36	87.23	**89.19**	87.29	87.70	**89.79**	90.82	91.64	**92.33**	91.22	92.14	**92.76**
	zh	88.65	90.39	**91.64**	89.60	90.78	**92.06**	90.47	92.09	**92.96**	91.02	92.39	**93.21**
	es	**96.01**	92.79	93.80	**96.15**	93.04	93.96	**96.25**	93.73	94.64	**96.41**	94.01	94.80
	de	**94.84**	91.97	93.50	**95.18**	92.22	93.81	**95.41**	93.50	94.26	**95.63**	93.75	94.39
	ru	**92.50**	91.09	91.89	**93.06**	91.51	92.28	92.88	92.46	**93.02**	93.22	92.87	**93.62**
gte-multilingual	ar	86.65	87.48	**88.76**	87.44	88.20	**89.35**	91.29	92.61	**93.23**	91.89	93.06	**93.63**
	zh	93.70	94.07	**94.67**	93.98	94.28	**94.86**	91.61	92.78	**93.47**	92.07	93.11	**93.83**
	es	95.85	95.05	**95.94**	96.14	95.29	**96.13**	95.48	95.65	**96.15**	95.70	95.84	**96.31**
	de	94.81	93.70	**94.82**	95.11	94.06	**95.16**	94.47	94.93	**95.84**	94.75	95.24	**96.01**
	ru	**94.23**	93.29	94.18	94.38	93.68	**94.47**	93.82	94.12	**94.63**	94.24	94.54	**95.07**
jina-v3	ar	90.01	92.72	**93.66**	90.58	93.12	**93.90**	92.83	94.76	**94.84**	93.21	95.02	**95.05**
	zh	92.32	94.46	**94.67**	92.65	94.65	**94.87**	92.37	94.84	**95.21**	92.79	95.06	**95.34**
	es	94.37	95.91	**96.02**	94.76	96.13	**96.13**	95.91	96.44	**96.64**	96.12	96.61	**96.78**
	de	94.52	95.87	**95.98**	94.69	96.05	**96.14**	95.15	96.26	**96.54**	95.30	96.34	**96.62**
	ru	93.77	95.81	**95.81**	93.97	95.98	**95.98**	94.06	95.50	**95.59**	94.38	95.77	**95.81**

Table 12: Results on all languages under the multilingual training setup in XQuAD.