

# 000 IMPROVING SEMANTIC PROXIMITY IN ENGLISH- 001 CENTRIC INFORMATION RETRIEVAL THROUGH CROSS- 002 LINGUAL ALIGNMENT 003

004 **Anonymous authors**  
005  
006

007 Paper under double-blind review  
008  
009  
010  
011

## ABSTRACT

012 With the increasing accessibility and utilization of multilingual documents, Cross-  
013 Lingual Information Retrieval (CLIR) has emerged as an important research area.  
014 Conventionally, CLIR tasks have been conducted under settings where the lan-  
015 guage of documents differs from that of queries, and typically, the documents are  
016 composed in a single coherent language. In this paper, we highlight that in such  
017 a setting, the cross-lingual alignment capability may not be evaluated adequately.  
018 Specifically, we observe that, in a document pool where English documents coexist  
019 with another language, most multilingual retrievers tend to prioritize unrelated  
020 English documents over the related document written in the same language as the  
021 query. To rigorously analyze and quantify this phenomenon, we introduce various  
022 scenarios and metrics designed to evaluate the cross-lingual alignment performance  
023 of multilingual retrieval models. Furthermore, to improve cross-lingual perfor-  
024 mance under these challenging conditions, we propose a novel training strategy  
025 aimed at enhancing cross-lingual alignment. Using only a small dataset consist-  
026 ing of 2.8k samples, our method significantly improves the cross-lingual retrieval  
027 performance while simultaneously mitigating the English inclination problem. Ex-  
028 tensive analyses demonstrate that the proposed method substantially enhances the  
029 cross-lingual alignment capabilities of most multilingual embedding models.  
030

## 1 INTRODUCTION

031 Information Retrieval (IR) is a core technology that accurately identifies and provides relevant  
032 information from large-scale document collections based on user queries. With the recent increase  
033 in multilingual data and the growing demand for accessing it, the importance of Cross-Lingual  
034 Information Retrieval (CLIR) has become more pronounced (Adeyemi et al., 2024; Guo et al.,  
035 2024; Zhang et al., 2023a; Mayfield et al., 2023). Commonly adopted CLIR evaluation settings  
036 primarily focus on scenarios where the query language differs from the language of the document  
037 pool, which is composed of a single language. The objective is to measure retrieval performance  
038 on a set of documents in a language different from that of query, thereby assessing the cross-  
039 lingual representation capability when query and document languages are not equal. Similarly, the  
040 Multilingual Information Retrieval (MLIR) task involves retrieving and ranking relevant documents  
041 from an integrated collection containing three or more languages (Lawrie et al., 2024).  
042

043 Despite these efforts, we observe that the practical effectiveness of multilingual embedding models  
044 in this settings has not been adequately assessed, and significant performance blind spots remain.  
045 Our research question is whether multilingual embedding models can consistently maintain retrieval  
046 performance for queries given in different languages, in realistic scenarios where documents in  
047 two languages coexist, without performance degradation stemming from cross-lingual semantic  
048 misalignment or language-specific biases. In fact, our exploratory analysis reveals that when retrieving  
049 from a document pool containing both English and another language that matches the query language,  
050 many multilingual embedding models exhibit significant cross-lingual semantic misalignment and  
051 language bias toward certain languages, severely degrading retrieval performance (Wu & Dredze,  
052 2020; Park & Lee, 2025; Yang et al., 2024; Elmahdy et al., 2024). For instance, when a query is  
053 written in language A, ideally, both the relevant document written in language A and its semantically  
equivalent counterpart in English should be ranked at the top of the retrieval results. However, due to

054 a prioritization bias toward English—a high-resource language—irrelevant English documents often  
 055 appear higher in the rankings, while the correct documents written in language A are relegated to  
 056 lower ranks. Furthermore, we observe a pronounced language-dependent performance disparity, in  
 057 which retrieval results vary significantly depending on the language of the query. This phenomenon  
 058 highlights a critical limitation that is difficult to accurately measure or analyze using conventional  
 059 evaluation settings. Consequently, our findings underscore the necessity of designing a more practical  
 060 evaluation environment and corresponding metrics, distinct from existing evaluation settings, to  
 061 rigorously validate balanced cross-lingual semantic alignment.

062 In this paper, we define a scenario where  
 063 documents in two languages coexist to  
 064 comprehensively evaluate cross-lingual  
 065 alignment capabilities in retrieval, and in-  
 066 troduce a new evaluation metric, Max@R.  
 067 By assessing embedding models within this  
 068 scenario and using the proposed metric, we  
 069 provide a detailed analysis of both cross-  
 070 lingual alignment aspects and retrieval per-  
 071 formance, as well as an in-depth exami-  
 072 nation of whether there exists an inherent  
 073 preference for a particular language.

074 Furthermore, based on our analysis of this scenario, we discuss methods to effectively mitigate  
 075 cross-lingual misalignment. In particular, we propose a unified training strategy that combines two  
 076 loss functions to jointly optimize cross-lingual alignment and retrieval performance. Specifically,  
 077 our strategy integrates Jensen-Shannon Divergence (JSD) (Lin, 1991), which aligns the semantic  
 078 embedding distributions across languages by adjusting the embedding space, with InfoNCE (van den  
 079 Oord et al., 2019), which directly enhances the retrieval ability between query and document. By  
 080 applying our proposed method, we demonstrate that even with a relatively small dataset of only 2.8K  
 081 samples, multilingual embedding models can achieve substantially enhanced cross-lingual alignment  
 082 and retrieval capabilities. Moreover, we find that performance of our method in monolingual settings,  
 083 including conventional CLIR, can be maintained or even improved relative to the baseline models.

## 084 2 PRELIMINARY

086 In this section, we rigorously define the cross-lingual alignment problem that arises in the evaluation  
 087 of multilingual embedding models. To accomplish this, we introduce the experimental environment  
 088 of both CLIR and MLIR, followed by a detailed explanation of the problem definition.

### 090 2.1 CONVENTIONAL SETTINGS

092 The objective of a typical IR task is to identify the most relevant documents from a collection in re-  
 093 sponse to a given query. This collection of documents is usually represented as  $D = \{d_1, d_2, \dots, d_n\}$ .  
 094 When a query ( $q$ ) is provided, similarity scores are calculated for the documents within the collection  
 095 ( $D$ ), and the most similar documents are retrieved based on these scores. CLIR is characterized  
 096 by the difference between the query language ( $L_q$ ) and the language of the document collection  
 097 ( $L_d$ ), such that  $L_q \neq L_d$ . For instance, when a query is provided in a specific language ( $L_1$ ), the  
 098 goal is to retrieve the most relevant documents written in a different language ( $L_2$ ). MLIR, on the  
 099 other hand, aims to effectively retrieve the documents from a collection composed of documents  
 100 in multiple languages ( $L_d \in \{L_2, L_3, L_4, \dots\}$ ) when the query is presented in a single specific  
 101 language ( $L_q = L_1$ ).

### 102 2.2 PROBLEM DEFINITION

104 Existing evaluation settings primarily address scenarios where the query and document collection  
 105 languages differ, or where multiple languages are included within documents (Lawrie et al., 2023).  
 106 However, these approaches are limited to assessing a model’s basic cross-lingual retrieval capabilities  
 107 or shallow multilingual retrieval performance, fail to reveal deeper issues such as inaccurate cross-  
 108 lingual alignment and language bias. Therefore, we consider a multi-reference cross-lingual setup that

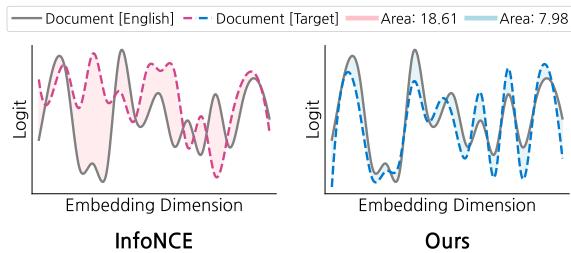


Figure 1: [Illustration of distribution-level alignment](#). Our method shows better alignment than InfoNCE, even when both share the same cosine similarity of 0.99.

can comprehensively evaluate both cross-lingual alignment and retrieval performance. For instance, in a scenario where the document pool contains a mixture of two languages,  $L_1$  and  $L_2$ , an effective model should rank all semantically relevant documents at the top, irrespective of the languages used in queries or documents. In other words, documents related to the query, even if expressed in different languages, should be retrieved with equal importance. Ideally, the model should demonstrate equal performance for queries in each language. To simulate an information retrieval scenario in such an environment, we construct and analyze an experimental setup using a dataset that is fully parallel across languages.

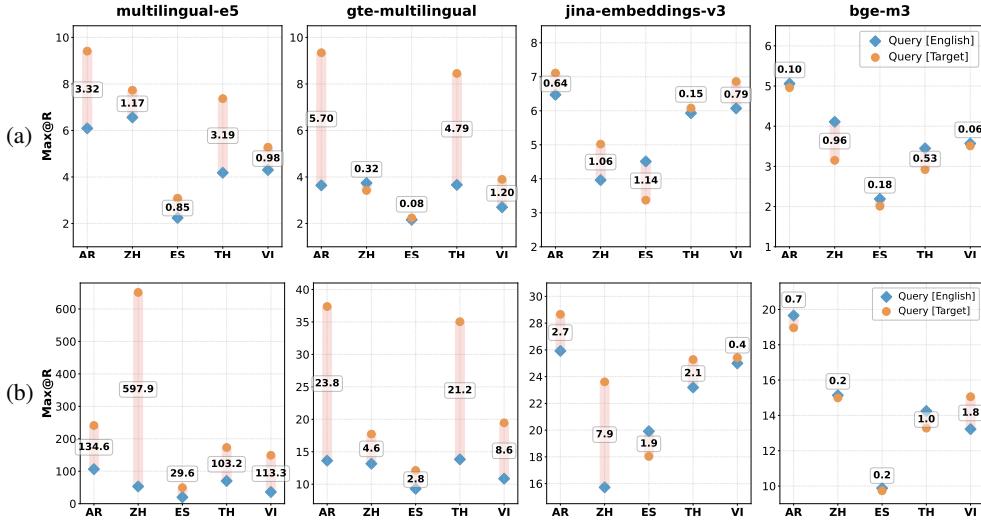


Figure 2: Performance comparison on the XQuAD dataset: (a) CLIR setting, and (b) Multi scenario. We evaluate four multilingual embedding models on five target languages: Arabic (AR), Chinese (ZH), Spanish (ES), Thai (TH), and Vietnamese (VI). Values indicate Max@R gaps.

### 3 MULTI-REFERENCE IN CROSS-LINGUAL INFORMATION RETRIEVAL

Insufficient cross-lingual semantic alignment results in the similarity between a query and documents in other languages not being properly reflected, causing two relevant documents to fail to be ranked at the top. To quantitatively measure this issue and to effectively evaluate performance in this scenario where two languages coexist, we propose a new evaluation metric.

**Max@R** We evaluate retrieval accuracy for a given query ( $q$ ) using a set of reference documents, denoted as  $R_q = \{r_1, r_2, \dots, r_m\}$ . In our multi-reference cross-lingual scenario, these documents are parallel ground-truths written in different languages. For the given query  $q$ , we denote an ordered list of retrieved documents as  $D'(q) = \{d'_1, d'_2, \dots, d'_n\}$ . The documents are sorted by retrieval priority based on semantic similarity to  $q$ . We define the position  $i$  of a document  $d'_i$  in the list  $D'(q)$  as its rank. Because existing metrics (MAP, MRR, or NDCG@k) are not designed to measure when all parallel ground-truths in this multi-reference scenario have been retrieved, we propose Max@R as an essential diagnostic metric. Then, we define **Max@R** as follows:

$$\text{Max@R} = \max(\{i \mid d'_i \in R_q\}) \quad (1)$$

By calculating Max@R in this manner, Max@R represents the highest (i.e., worst) rank at which all relevant documents in the reference set are first found in retrieval results. This is, in effect, the ‘actual cutoff point’ one must reach to retrieve all  $R_q$  documents. In other words, a lower Max@R value indicates that all relevant documents can be retrieved within a smaller portion of the top-ranked results, signifying a more efficient and higher-performing retrieval model.

162 3.1 LIMITATIONS OF CROSS-LINGUAL INFORMATION RETRIEVAL IN PRACTICAL SCENARIOS  
163

164 Figure 2 presents the evaluation results of cross-lingual retrieval performance using four different  
165 multilingual embedding models, focusing on English as the source language and five target languages.  
166 The experiments are conducted under two distinct settings: (a) conventional CLIR, and (b) our  
167 proposed [multi-reference cross-lingual setup](#), as defined in [Section 2.2](#). This setup (b) features a doc-  
168 ument pool with parallel documents in English and the target language, resulting in two ground-truth  
169 documents per query. [To thoroughly investigate potential biases, we report the performance achieved](#)  
170 [when querying in both English and the respective target language](#). Based on these experimental  
171 results, we identify and discuss three critical limitations of existing multilingual embedding models  
172 and CLIR environment that were not evident in conventional assessment settings.  
173

174 **Performance Disparities by Query Language** One of the most significant limitations observed  
175 is the disparity in retrieval performance depending on the query language. In Fig. 2 (a), the overall  
176 Max@R is notably low, and the performance differences by query language are relatively minor. In  
177 contrast, Fig. 2 (b) shows a substantial gap in performance based on the query language. For the  
178 multilingual-e5 model, the results for Chinese (ZH) queries display a marked difference of 597.9  
179 in Max@R compared to the results for English queries. This finding indicates that cross-lingual  
180 semantic alignment is inadequate, resulting in amplified performance discrepancies based on the  
181 query language in the proposed scenario.  
182

183 **Representation Instability among Target Languages** In CLIR, multilingual embedding models  
184 tend to exhibit relatively consistent performance across target languages. However, as shown in Fig. 2  
185 (b), there is a significant language-level instability even within the same model. For example, for  
186 the gte-multilingual model, the Max@R for Spanish (ES) remains relatively low and stable at 12.12,  
187 but rises sharply to 37.38 and 35.04 for Arabic and Thai, respectively—nearly a threefold increase.  
188 These results indicate that multilingual embedding models exhibit significant semantic representation  
189 inconsistency among target languages.  
190

191 **Excessively High Retrieval Ranks** In Fig. 2 (b), most models and languages exhibit excessively  
192 high Max@R values, making them impractical. Notably, for the multilingual-e5 with Chinese (ZH)  
193 queries, the Max@R reaches 650.95, meaning that hundreds of documents must be reviewed to find  
194 all relevant documents. This result implies that the existing CLIR setting (Fig. 2 (a)) fails to capture  
195 problems that emerge in practical scenarios, particularly the comprehensive issues of cross-lingual  
196 semantic alignment and retrieval capability.  
197

198 In summary, the proposed scenario and metric have revealed critical issues that were overlooked in  
199 conventional settings. This shows that there is still significant room for improvement in multilingual  
200 embedding models regarding cross-lingual alignment and retrieval. This underscores the necessity  
201 for more realistic and rigorous evaluations across various cross-lingual conditions, thereby providing  
202 a clear direction for the advancement of multilingual embedding models.  
203

204 4 METHODS  
205

206 In Section 3.1, we identified the issue of language misalignment within multilingual embedding  
207 models. Specifically, the identified problems and corresponding solutions are as follows: First, there  
208 is a lack of semantic alignment between text embeddings expressed in different languages, leading to  
209 an increased semantic distance between sentences and a decrease in overall retrieval performance.  
210 To resolve this cross-lingual semantic misalignment, we align the embedding distributions between  
211 English document ( $p_{en}$ ) and its semantically corresponding target language document ( $p_{tgt}$ ) using a  
212 loss function based on Jensen–Shannon Divergence (JSD), explicitly optimizing semantic alignment  
213 in the embedding space. Second, to mitigate the language bias where retrieval results show an  
214 inclination towards documents in a particular language (primarily English), we employ InfoNCE  
215 contrastive loss by using positive pairs between English query ( $q_{en}$ ) and target language document  
216 ( $p_{tgt}$ ). To this end, we utilize a dataset in the form of  $(q_{en}, p_{en}, p_{tgt})$  to train the model to minimize a  
217 combined loss of (1) JSD-based cross-lingual alignment loss  $L_{JSD}$  and (2) InfoNCE contrastive loss  
218  $L_{NCE}$ , defined as follows:  
219

$$L = \mathbb{E}_{(q_{en}, p_{en}, p_{tgt})} [L_{JSD} + L_{NCE}] \quad (2)$$

216 4.1 SEMANTIC EMBEDDING ALIGNMENT VIA  $L_{JSD}$   
217

218 While standard CLIR objectives are designed to increase the similarity score between a query in  
219 one language and a document in another, this approach may be insufficient for achieving robust  
220 semantic proximity. Achieving this requires more than just high similarity scores for retrieval; it  
221 also necessitates that their underlying embedding representations are fundamentally aligned. As  
222 illustrated in Figure 1, representations produced by different objectives can share an identical cosine  
223 similarity score yet remain significantly misaligned at the distribution level. To align semantically  
224 equivalent passages, our goal is to minimize the divergence between the embedding distributions of  
225 texts expressed in two different languages. A prominent method for measuring the disparities between  
226 two probability distributions is Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951). It  
227 quantifies how a distribution  $P$  diverges from another distribution  $Q$ , and it is defined as follows:  
228

$$229 D_{KL}(P\|Q) = \sum_i P_i \log \frac{P_i}{Q_i} \quad (3)$$

230 However, KL-divergence is asymmetric; that is,  $D_{KL}(P\|Q) \neq D_{KL}(Q\|P)$ . To address this asym-  
231 metry, Jensen-Shannon Divergence (JSD) is widely adopted. JSD first defines an intermediate  
232 averaged distribution  $M$  between two distributions and then measures the average KL-divergence of  
233 each distribution from this intermediate distribution. It is defined as follows:  
234

$$235 236 JSD(P\|Q) = \frac{1}{2}D_{KL}(P\|M) + \frac{1}{2}D_{KL}(Q\|M), \quad M = \frac{1}{2}(P + Q) \quad (4)$$

237 Divergence measures such as KL-divergence and JS-Divergence are widely used to quantify the  
238 differences between predicted probability distributions and reference distributions. Previous studies  
239 typically focused on modeling the similarities between queries and documents as probability distribu-  
240 tions and minimizing the divergence between these predicted distributions and reference distributions.  
241 For example, in information retrieval tasks, a model is trained to minimize the distributional diver-  
242 gence between a predicted distribution reflecting query-document similarities and a given reference  
243 distribution. In contrast, we propose directly aligning semantic embeddings at the distribution level.  
244 Specifically, semantic embedding vectors are interpreted as probability distributions. By explicitly  
245 aligning these embedding distributions using JSD, our approach aligns the embedding dimensions  
246 across languages, thereby effectively achieving enhanced cross-lingual semantic alignment. To illus-  
247 trate, let us denote the English document embedding vector as  $\mathbf{z}_{d_{en}} \in \mathbb{R}^{\text{dim}}$ , and the corresponding  
248 target-language embedding vector as  $\mathbf{z}_{d_{tgt}} \in \mathbb{R}^{\text{dim}}$ , where  $\mathbb{R}^{\text{dim}}$  represents the embedding dimension.  
249 To transform these embedding vectors into categorical probability distributions over dimensions, we  
250 apply the following softmax function:  
251

$$252 253 P(\mathbf{z})_i = \frac{\exp(z_i)}{\sum_{k=1}^{\text{dim}} \exp(z_k)}, \quad i = 1, 2, \dots, \text{dim} \quad (5)$$

254 After transforming the document embedding vectors into probability distributions  $P(\mathbf{z}_{d_{en}})$  and  
255  $P(\mathbf{z}_{d_{tgt}})$ , we perform distribution-level semantic embedding alignment by minimizing the JSD  
256 between these probability distributions. The proposed loss function is:  
257

$$258 259 \min L_{JSD} = \sqrt{JSD(P(\mathbf{z}_{d_{en}}) | P(\mathbf{z}_{d_{tgt}})) + \epsilon}. \quad (6)$$

260 Finally, this loss function employs square root of the Jensen-Shannon divergence, a rigorous distribu-  
261 tional distance measure between embedding distributions of two languages. Specifically, taking the  
262 square root of JSD satisfies the three distance axioms (identity, symmetry, and triangle inequality),  
263 thereby forming a valid metric space (Endres & Schindelin, 2003). Minimizing this distance-based  
264 loss during optimization encourages close alignment in the dimension-level probabilistic structures  
265 of embeddings, facilitating more effective cross-lingual semantic alignment.  
266

267 4.2 RETRIEVING OBJECTIVE VIA  $L_{NCE}$   
268

269 To improve similarity for cross-lingual query-document pairs, we train the model to maximize the  
semantic similarity between an English query  $q_{en}$  and its corresponding target language passage

270  $p_{\text{tgt}}$ . We employ the InfoNCE loss, a representative loss function of contrastive learning, as the  
 271 objective function. Specifically, the InfoNCE loss for positive pairs  $(p_{tgt_i}, q_{en_i}^+)^n_{i=1}$  is defined by  
 272 the following equation:  
 273

$$274 \min L_{\text{NCE}} = -\frac{1}{n} \sum_i \log \frac{\exp(s(p_{tgt_i}, q_{en_i}^+))}{\exp(s(p_{tgt_i}, q_{en_i}^+)) + \sum_j \exp(s(p_{tgt_i}, q_{en_{ij}}^-))} \quad (7)$$

$$275$$

$$276$$

277 where  $n$  and  $m$  denote the total number of data and the batch size, respectively. The negative examples  
 278  $\{q_{en_{ij}}^-\}_{j=1}^m$  are in-batch negatives (queries of other instances within the same batch).  $s(p, q)$  is the  
 279 relevance score of  $p$  and  $q$ , measured by the cosine similarity between their respective representations.  
 280 Through the optimization of this contrastive objective function, the semantic similarity between  
 281 related pairs is maximized, while the similarity with unrelated examples is minimized. This process  
 282 ultimately facilitates the semantic alignment within an embedding space for cross-lingual retrieval.  
 283

## 284 5 EXPERIMENTS

$$285$$

286 In this section, we conduct comprehensive experiments across a variety of scenarios to evaluate the  
 287 proposed method for cross-lingual alignment capability and retrieval performance.  
 288

### 289 5.1 EXPERIMENT SETUP

$$290$$

291 **Scenario** We design three experimental scenarios as follows: (1) **Multi** evaluates whether the model  
 292 can retrieve the two ground-truth documents per query in each language from a set of fully parallel  
 293 documents in two languages. (2) **Multi-1** excludes the ground-truth document in the same language  
 294 as the query, requiring the model to retrieve the correct document in the opposite language. This  
 295 setup evaluates the semantic-based retrieval capabilities under a strict setting. (3) **Mono** assesses  
 296 retrieval performance in a single-language environment, further divided into cases where the query  
 297 and document are in the same language (Mono-Same) and in different languages (Mono-Cross), with  
 298 the latter corresponding to conventional CLIR. Through these scenarios, we evaluate the reliability of  
 299 the model in both cross-lingual and monolingual settings.  
 300

301 **Datasets** To evaluate the three cross-lingual retrieval scenarios, it is essential to use datasets that are  
 302 fully parallel across languages. **This fully parallel structure is a necessary prerequisite for rigorously**  
 303 **validating our scenarios and the Max@R metric**. This means that the same question-document  
 304 pairs must exist in multiple languages to enable the evaluation of retrieval performance across  
 305 different languages. For this purpose, we utilize the multilingual benchmarks XQuAD (Artetxe et al.,  
 306 2020) and Belebele (Bandarkar et al., 2024). **These are high-quality, human-translated benchmarks,**  
 307 **widely adopted in standard multilingual retrieval evaluation benchmarks Enevoldsen et al. (2025)**. A  
 308 **comprehensive description of these datasets and our rationale is provided in Appendix B**. Both datasets  
 309 ensure complete parallelism, making them suitable for comprehensively validating performance across  
 310 the proposed scenarios.  
 311

312 **Metric** We employ evaluation metrics that align with the characteristics of each retrieval scenario  
 313 environment. For the Multi scenario, we utilize Complete@K (Qu et al., 2024), which considers an  
 314 answer correct only if all relevant documents are included within the top-k results. **This metric is**  
 315 **reported as a percentage on a 0–100 scale**. Additionally, we use Max@R, as proposed in Section 3,  
 316 and introduce an intuitively interpretable and generalized metric, Max@R<sub>norm</sub>, which normalizes  
 317 the varying Max@R values across different datasets on a logarithmic scale. This metric normalizes  
 318 the maximum rank for each query to a value between 0 and 100, represented as Max@R<sub>norm</sub> =  

$$319 \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} [100 \times \frac{\log_2(|D|) - \log_2(\text{Max@R})}{\log_2(|D|) - \log_2(|\mathcal{R}|)}],$$
 320 where  $|D|$  is the size of the document pool, and  $|\mathcal{R}|$  is the  
 321 number of ground-truth documents for each query. For the Multi-1 and Mono scenarios, where there  
 322 is only one correct document per query, we evaluate retrieval performance using metrics such as  
 323 NDCG@1 (Järvelin & Kekäläinen, 2002) and MRR (Nogueira & Cho, 2019; Khattab & Zaharia,  
 324 2020; Xiong et al., 2020; Karpukhin et al., 2020). By employing appropriate metrics for each scenario,  
 325 we systematically compare and analyze the proposed method across various scenarios, thereby  
 326 verifying its validity from multiple perspectives.  
 327

Table 1: Main performance results under the **Multi scenario**

Doc	Query	XQuAD						Belebte					
		Comp@10		Max@R (J)		Max@R <sub>norm</sub>		Comp@10		Max@R (J)		Max@R <sub>norm</sub>	
		Base	Ours	Base	Ours	Base	Ours	Base	Ours	Base	Ours	Base	Ours
<i>multilingual-e5-base</i>													
En+Ar	En	15.46	<b>60.34</b>	106.42	<b>18.56</b>	43.88	<b>68.54</b>	26.22	<b>87.22</b>	187.78	<b>17.46</b>	33.23	<b>68.14</b>
	Ar	8.91	<b>53.53</b>	241.06	<b>30.77</b>	32.33	<b>61.40</b>	15.44	<b>75.67</b>	231.79	<b>29.55</b>	30.13	<b>60.41</b>
En+Zh	En	29.24	<b>61.60</b>	53.04	<b>17.35</b>	53.71	<b>69.50</b>	63.44	<b>86.78</b>	44.74	<b>15.96</b>	54.31	<b>69.47</b>
	Zh	0.50	<b>65.88</b>	65.95	<b>13.00</b>	18.31	<b>65.45</b>	3.11	<b>82.44</b>	476.45	<b>18.55</b>	19.54	<b>67.26</b>
En+Es	En	58.99	<b>65.63</b>	19.83	<b>15.38</b>	67.61	<b>71.19</b>	68.67	<b>88.89</b>	49.14	<b>13.44</b>	52.94	<b>71.99</b>
	Es	36.30	<b>62.52</b>	49.46	<b>18.14</b>	54.70	<b>68.87</b>	55.44	<b>85.22</b>	69.81	<b>16.50</b>	47.77	<b>68.98</b>
En+Th	En	24.37	<b>57.23</b>	69.94	<b>18.81</b>	49.81	<b>68.35</b>	41.11	<b>87.33</b>	112.49	<b>15.44</b>	40.76	<b>69.95</b>
	Th	11.76	<b>46.81</b>	173.13	<b>33.66</b>	37.01	<b>60.13</b>	22.44	<b>75.89</b>	188.02	<b>29.60</b>	33.21	<b>60.39</b>
En+Vi	En	45.88	<b>63.53</b>	35.95	<b>17.31</b>	59.20	<b>69.53</b>	56.44	<b>88.89</b>	69.83	<b>12.92</b>	47.77	<b>72.58</b>
	Vi	18.15	<b>59.50</b>	149.21	<b>24.15</b>	39.11	<b>64.82</b>	33.44	<b>82.89</b>	153.05	<b>15.54</b>	36.23	<b>69.86</b>
<i>gte-multilingual-base</i>													
En+Ar	En	66.05	<b>67.23</b>	13.63	<b>12.64</b>	72.90	<b>73.97</b>	87.44	<b>88.56</b>	10.03	<b>9.17</b>	76.29	<b>77.62</b>
	Ar	49.58	<b>52.10</b>	37.38	<b>31.38</b>	58.65	<b>61.12</b>	79.67	<b>80.33</b>	26.23	<b>23.55</b>	62.16	<b>63.75</b>
En+Zh	En	68.15	<b>68.99</b>	13.16	<b>12.51</b>	73.40	<b>74.11</b>	91.00	<b>91.56</b>	9.23	<b>8.90</b>	77.52	<b>78.06</b>
	Zh	58.82	<b>63.11</b>	17.73	<b>15.24</b>	69.19	<b>71.32</b>	86.67	<b>88.56</b>	13.93	<b>12.80</b>	71.47	<b>72.71</b>
En+Es	En	77.14	<b>78.40</b>	9.30	<b>8.92</b>	78.29	<b>78.89</b>	90.44	<b>93.22</b>	8.77	<b>6.50</b>	78.27	<b>82.68</b>
	Es	71.01	<b>73.78</b>	12.12	<b>10.51</b>	74.56	<b>75.67</b>	88.11	<b>90.78</b>	14.18	<b>11.06</b>	71.21	<b>74.86</b>
En+Th	En	64.96	<b>67.48</b>	13.83	<b>13.02</b>	72.69	<b>73.55</b>	88.67	<b>89.33</b>	9.97	<b>9.03</b>	76.39	<b>77.84</b>
	Th	46.13	<b>51.34</b>	35.04	<b>30.50</b>	59.57	<b>61.53</b>	77.11	<b>78.67</b>	29.97	<b>25.30</b>	60.21	<b>62.69</b>
En+Vi	En	73.11	<b>73.19</b>	10.85	<b>10.78</b>	76.12	<b>76.22</b>	90.56	<b>91.33</b>	8.33	<b>7.71</b>	79.03	<b>80.16</b>
	Vi	62.10	<b>64.71</b>	19.46	<b>16.75</b>	67.87	<b>69.99</b>	86.33	<b>87.56</b>	11.63	<b>9.94</b>	74.12	<b>76.43</b>
<i>jina-embeddings-v3</i>													
En+Ar	En	55.13	<b>70.92</b>	25.92	<b>11.75</b>	63.82	<b>74.99</b>	85.11	<b>90.67</b>	19.83	<b>9.91</b>	66.28	<b>76.47</b>
	Ar	57.65	<b>65.04</b>	28.66	<b>19.29</b>	62.41	<b>68.00</b>	82.78	<b>86.22</b>	22.93	<b>16.47</b>	64.14	<b>69.10</b>
En+Zh	En	65.46	<b>72.44</b>	15.72	<b>10.96</b>	70.88	<b>75.98</b>	87.98	<b>91.56</b>	14.30	<b>10.00</b>	71.07	<b>76.34</b>
	Zh	58.57	<b>70.67</b>	23.61	<b>13.53</b>	65.14	<b>73.00</b>	85.33	<b>91.44</b>	18.52	<b>11.51</b>	67.28	<b>74.27</b>
En+Es	En	68.32	<b>75.63</b>	19.92	<b>10.10</b>	67.54	<b>77.14</b>	85.11	<b>92.22</b>	24.16	<b>9.88</b>	63.38	<b>76.52</b>
	Es	68.74	<b>73.53</b>	18.03	<b>11.44</b>	68.95	<b>75.37</b>	88.78	<b>90.22</b>	18.67	<b>12.59</b>	67.16	<b>72.96</b>
En+Th	En	56.39	<b>71.51</b>	23.19	<b>11.59</b>	65.39	<b>75.19</b>	86.44	<b>91.89</b>	17.63	<b>9.77</b>	68.03	<b>76.69</b>
	Th	58.07	<b>68.32</b>	25.28	<b>16.41</b>	64.18	<b>70.28</b>	84.33	<b>86.67</b>	22.17	<b>15.78</b>	64.65	<b>69.63</b>
En+Vi	En	59.75	<b>71.43</b>	24.99	<b>11.69</b>	64.34	<b>75.07</b>	83.00	<b>91.89</b>	22.91	<b>9.91</b>	64.16	<b>76.48</b>
	Vi	61.51	<b>67.06</b>	25.42	<b>18.64</b>	64.10	<b>68.48</b>	87.33	<b>90.67</b>	19.88	<b>12.27</b>	66.25	<b>73.33</b>
<i>bge-m3</i>													
En+Ar	En	59.50	<b>68.32</b>	19.66	<b>14.87</b>	67.73	<b>71.67</b>	86.44	<b>89.56</b>	13.78	<b>10.18</b>	71.62	<b>76.08</b>
	Ar	65.71	<b>67.98</b>	18.96	<b>17.39</b>	68.24	<b>69.46</b>	85.11	<b>85.44</b>	18.91	<b>15.82</b>	66.98	<b>69.60</b>
En+Zh	En	67.82	<b>71.09</b>	15.14	<b>13.08</b>	71.42	<b>73.48</b>	89.44	<b>90.56</b>	12.14	<b>10.90</b>	73.49	<b>75.07</b>
	Zh	63.19	<b>68.49</b>	14.99	<b>13.01</b>	71.56	<b>73.56</b>	88.67	<b>90.00</b>	12.65	<b>11.56</b>	72.88	<b>74.20</b>
En+Es	En	77.90	<b>78.82</b>	9.88	<b>9.36</b>	77.44	<b>78.21</b>	91.78	<b>93.00</b>	10.25	<b>9.29</b>	75.98	<b>77.43</b>
	Es	77.65	<b>77.39</b>	9.73	<b>9.72</b>	77.66	<b>77.68</b>	91.11	<b>92.56</b>	10.70	<b>9.92</b>	75.34	<b>76.46</b>
En+Th	En	67.90	<b>71.34</b>	14.25	<b>12.67</b>	72.27	<b>73.93</b>	90.89	<b>92.11</b>	10.24	<b>9.03</b>	75.99	<b>77.83</b>
	Th	67.23	<b>70.08</b>	13.28	<b>12.95</b>	73.27	<b>73.63</b>	87.00	<b>87.67</b>	16.73	<b>14.57</b>	68.78	<b>70.81</b>
En+Vi	En	71.01	<b>73.53</b>	13.22	<b>11.74</b>	73.33	<b>75.01</b>	90.56	<b>91.67</b>	10.53	<b>8.97</b>	75.59	<b>77.93</b>
	Vi	66.72	<b>70.92</b>	15.06	<b>13.40</b>	71.49	<b>73.14</b>	89.00	<b>90.78</b>	11.01	<b>9.80</b>	74.93	<b>76.63</b>

**Implementation Details** The experiments are conducted on a total of 10 languages. In the main results, we report for five languages: Arabic (AR), Chinese (ZH), Spanish (ES), Thai (TH), and Vietnamese (VI). Detailed experiments for the remaining five languages, German (DE), Greek (EL), Hindi (HI), Romanian (RO), and Turkish (TU), are provided in the Appendix D. We employ four multilingual embedding models for the experiments: multilingual-E5-base (Wang et al., 2024), gte-Multilingual-base (Zhang et al., 2024), jina-embeddings-v3 (Sturua et al., 2024), and bge-M3 (Chen et al., 2024). We utilize the MIRACL train dataset (Zhang et al., 2023b), which consists of 2.8k English query-document pairs. To obtain documents in each target language, the positive documents in English are translated into each target language using the GPT-4o (OpenAI, 2024). Additional details regarding the training procedure are provided in Appendix C.

## 5.2 MAIN RESULTS

In this section, we evaluate our method within a multi-scenario, where documents in English and the target language coexist, and each query has two gold documents. The results, presented in Table 1, are analyzed by querying in both English and the target language.

**Enhanced Cross-lingual Alignment** The proposed method demonstrates clear and consistent improvements over baseline models across all evaluated language pairs and metrics. Notably, based on Complete@10, while baseline models exhibit relatively limited performance when queried in non-English languages, our proposed method achieves significant performance enhancements for all languages and models considered. These improvements indicate that our method effectively facilitates

semantic alignment among different languages within an embedding space. Ultimately, these results confirm that our method successfully addresses the limitations of existing multilingual embedding models, which are not captured by the conventional CLIR setting.

**Reducing Variance in Language Bias** In addition to improving retrieval performance, the proposed method reduces the issue of performance disparity between languages. Baseline models typically exhibit a bias toward English data, resulting in significantly reduced performance for non-English queries. However, the proposed method results in a relative decrease in the quantitative performance disparity between English and target languages. For instance, the language performance gap for jina-embeddings-v3 (En+Zh) consistently decreases from  $6.89\%p \rightarrow 1.77\%p$  on XQuAD and from  $4.45\%p \rightarrow 0.12\%p$  on Belebele. This result demonstrates that the model has mitigated the previous misalignment between English and target languages, reducing language bias and contributing to improved language equity.

**Enhanced Full-Recall Ranking Performance** To intuitively assess the cross-lingual alignment, we focus on analyzing the Max@R metric in our experiments. The results show that our method consistently achieves significantly lower Max@R scores across all languages and datasets compared to the baseline. For instance, with Chinese queries on the multilingual-e5-base model, Max@R significantly improves from 650.95 to 23.10 on XQuAD. This indicates that our method is effective in positioning relevant passages near the top, demonstrating robust performance in cross-lingual multi-reference scenarios. These improvements are also observed by consistent gains on the Max@R<sub>norm</sub>.

### 5.3 CASE STUDY

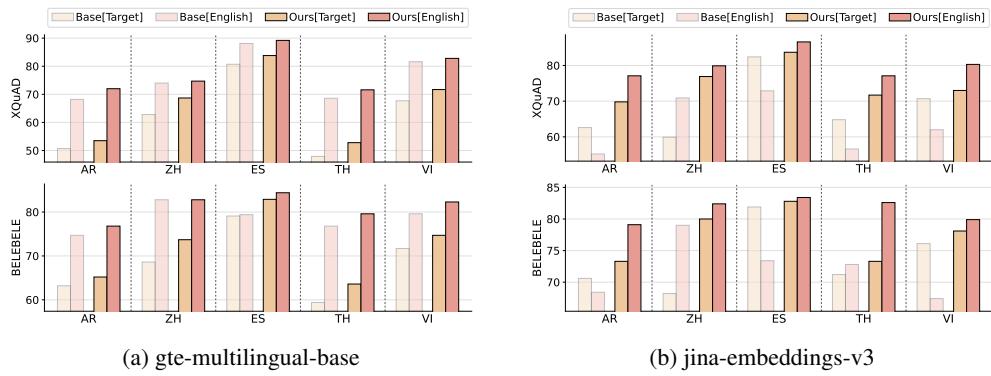


Figure 3: NDCG@1 comparison in the **Multi-1 scenario** with [lang] as the query language

**Cross-Lingual Information Retrieval in the Multi-1 Scenario** The Multi-1 scenario provides a more rigorous assessment of cross-lingual semantic alignment capabilities. As shown in Figure 3, existing baselines encounter substantial difficulties in retrieving a relevant document written in languages different from the query language. In contrast, our proposed method consistently improves the NDCG@1 across all language pairs, demonstrating enhanced cross-lingual semantic proximity within the semantic embedding space. Furthermore, the observed performance gains are consistent and substantial, irrespective of whether the queries are in English or in the respective target languages.

**Performance Validation in Monolingual Settings** To confirm whether the proposed method maintains robust retrieval performance in monolingual environments, we conduct evaluations under two settings: Mono-Same and Mono-Cross. The results, summarized in Table 2, demonstrate that our approach largely preserves or even modestly improves upon the baseline models in the Mono-Same scenario. Notably, slight performance gains are observed for target language queries, suggesting that the quality of single language semantic representations is indirectly enhanced via our alignment method. In the Mono-Cross scenario, our method further surpasses the baseline models, providing consistent improvements. Collectively, these findings indicate that our approach does not compromise and sometimes even improves, monolingual retrieval performance. This affirms that reducing language misalignment in embedding spaces improves representational quality and retrieval performance in both cross-lingual and monolingual settings.

Table 2: **Mono scenario** performance using jina-embeddings-v3: English/target language queries.

Lang	Model	XQuAD				Belebte			
		Mono-Same		Mono-Cross		Mono-Same		Mono-Cross	
		NDCG@1	MRR	NDCG@1	MRR	NDCG@1	MRR	NDCG@1	MRR
Ar	Base	89.2/81.9	0.933/0.879	79.5/80.6	0.863/0.874	88.4/83.0	0.919/0.873	80.1/77.0	0.856/0.833
	Ours	<b>90.8/83.9</b>	<b>0.945/0.898</b>	<b>87.1/81.9</b>	<b>0.919/0.884</b>	<b>88.4/87.0</b>	<b>0.922/0.871</b>	<b>82.6/79.8</b>	<b>0.880/0.856</b>
Zh	Base	89.2/88.2	0.933/0.926	83.3/83.8	0.895/0.895	<b>88.4/86.9</b>	0.919/0.912	83.4/81.9	0.882/0.874
	Ours	<b>90.7/89.5</b>	<b>0.943/0.937</b>	<b>88.0/86.5</b>	<b>0.927/0.916</b>	88.1/87.0	<b>0.920/0.912</b>	<b>84.9/84.4</b>	<b>0.895/0.896</b>
Es	Base	89.2/85.6	0.933/0.907	85.3/87.5	0.905/0.922	<b>88.4/84.9</b>	<b>0.919/0.895</b>	83.4/84.4	0.882/0.890
	Ours	<b>89.2/88.8</b>	<b>0.936/0.931</b>	<b>89.3/88.0</b>	<b>0.935/0.925</b>	87.7/85.4	0.918/0.898	<b>85.6/84.7</b>	<b>0.901/0.892</b>
Th	Base	89.2/82.4	0.933/0.885	80.4/81.7	0.873/0.879	88.4/82.2	0.919/0.875	80.7/77.6	0.865/0.844
	Ours	<b>91.2/86.4</b>	<b>0.947/0.916</b>	<b>87.6/83.1</b>	<b>0.925/0.894</b>	<b>88.7/82.2</b>	<b>0.923/0.876</b>	<b>85.2/78.6</b>	<b>0.899/0.848</b>
Vi	Base	<b>89.2/84.4</b>	<b>0.933/0.897</b>	79.4/81.1	0.862/0.878	<b>88.4/86.1</b>	<b>0.919/0.900</b>	80.7/82.1	0.859/0.873
	Ours	88.9/86.9	0.933/0.918	<b>87.6/81.8</b>	<b>0.922/0.884</b>	87.6/85.0	0.914/0.897	<b>83.4/83.6</b>	<b>0.889/0.887</b>

## 5.4 ABLATION STUDY

In this section, we perform an ablation study to clearly examine the roles of two loss components in our method: the Jensen–Shannon divergence-based embedding alignment loss  $L_{JSD}$  and the InfoNCE-based query-document relevance learning loss  $L_{NCE}$ . We also include a comparative experiment,  $L_{NCE_{psq}}$ , on enhancing the similarity between English documents ( $p_{en}$ ) and target language documents ( $p_{tgt}$ ) (Feng et al., 2020; Chi et al., 2020).

Results in Table 3 indicate clearly complementary roles for these components. Specifically, the absence of  $L_{JSD}$  negatively affects cross-lingual semantic embedding alignment and overall retrieval performance, whereas excluding  $L_{NCE}$  limits the retrieval effectiveness, despite embedding-level alignment. This demonstrates that the combination of  $L_{JSD}$  and  $L_{NCE}$  is essential for effectively achieving objectives of semantic alignment within an embedding space and securing retrieval performance. It also proves that relying solely on one of these components is insufficient to enhance overall cross-lingual retrieval.

Additionally, we analyze  $L_{NCE_{psg}}$ . The results show that this approach generally improves retrieval performance over the Baseline, confirming that enhancing document-level similarity is a valid strategy for improving cross-lingual representations. Crucially, regardless of the base model, Ours consistently and significantly outperforms the  $L_{NCE_{psg}}$  approach. This result highlights the source of our method's superiority: rather than simply increasing a similarity score between documents, it provides a more fundamental solution by directly aligning the distribution of the output representations themselves, making it more ef-

Table 3: Ablation results (Max@R<sub>norm</sub>) for loss components in multi scenario experiments on Belebele (English query / Target language query).

Methods	Ar	Zh	Es	Th	Vi
gte-multilingual-base					
Baseline	76.29 / 62.16	77.52 / 71.47	78.27 / 71.21	76.39 / 60.21	79.03 / 74.12
$LNCE_{pgg}$	75.45 / 63.28	76.07 / 71.72	79.33 / 74.13	76.34 / 62.53	77.79 / 74.77
Ours	<b>77.62 / 63.75</b>	<b>78.06 / 72.71</b>	<b>82.68 / 74.86</b>	<b>77.84 / 62.69</b>	80.16 / <b>76.43</b>
w/o $L_{JSD}$	75.63 / 63.19	75.18 / 71.42	80.43 / 74.14	74.75 / 61.93	78.99 / 75.39
w/o $L_{NCE}$	76.14 / 59.96	78.01 / 70.09	82.00 / 71.66	77.13 / 59.65	<b>80.20 / 73.84</b>
jina-embeddings-v3					
Baseline	66.28 / 64.14	71.07 / 67.28	63.38 / 67.16	68.03 / 64.65	64.16 / 66.25
$LNCE_{pgg}$	72.37 / 68.12	75.00 / 72.92	68.06 / 69.93	72.35 / 68.66	70.52 / 70.46
Ours	<b>76.47 / 69.10</b>	<b>76.34 / 74.27</b>	<b>76.52 / 72.96</b>	<b>76.69 / 69.63</b>	<b>76.48 / 73.33</b>
w/o $L_{JSD}$	71.58 / 67.73	74.64 / 72.96	68.31 / 69.88	71.90 / 68.19	70.42 / 70.19
w/o $L_{NCE}$	40.68 / 34.29	37.52 / 38.29	57.44 / 54.18	15.47 / 14.26	63.73 / 42.41

## 6 RELATED WORKS

Most existing studies in Cross-Lingual Information Retrieval (CLIR) focus on bridging the semantic gap by constructing cross-lingual embedding spaces or leveraging knowledge transfer approaches to minimize query-document distances across languages (Huang et al., 2023a; Yu et al., 2021; Litschko et al., 2021; Valentini et al., 2025; Lin et al., 2023). Unsupervised methods, in particular, have attempted to reduce dependencies on translation resources by training shared embedding spaces only from monolingual corpora (Litschko et al., 2018). Concurrently, several studies address low-resource languages, where parallel corpora are limited, by proposing optimal transport-based knowledge distillation or multi-stage knowledge distillation techniques to transfer ranking knowledge from high-resource languages (Huang et al., 2023a;b). Additionally, integrating knowledge graphs into query-document representations has shown promise in alleviating cross-lingual semantic gaps (Zhang et al., 2022; Litschko et al., 2022). Collectively, these studies underscore that aligning and refining cross-lingual embedding representations is critically important for CLIR. However, most prior studies

486 assume either purely monolingual or entirely multilingual document pool, thus failing to adequately  
 487 address biases and misalignments that arise when two languages coexist in a single document pool.  
 488

489 In parallel, studies have increasingly focused on explicit approaches for aligning multilingual em-  
 490 bedding spaces. Hu et al. (2020) leverages parallel corpora and explicit alignment objectives to  
 491 enhance sentence-level cross-lingual transferability. More granular studies addressing contextual  
 492 embedding alignment have introduced nuanced evaluation tasks, such as dependency parsing or  
 493 token-level semantic retrieval, highlighting the importance of fine-grained measures (Schuster et al.,  
 494 2019; Liu et al., 2019). While insightful, these works primarily emphasize alignment in general  
 495 representation tasks, offering limited consideration of practical retrieval challenges associated with  
 496 combined-language environments.

## 497 7 CONCLUSION

500 To investigate severe semantic misalignment and language disparities exhibited by existing multilin-  
 501 gual embedding models, we propose a new evaluation scenario and a metric, Max@R. Through our  
 502 experiments, we reveal previously unseen issues that were not observable using existing evaluation  
 503 scenarios. To address these issues, we present a training strategy that effectively achieves semantic  
 504 proximity in the cross-lingual embedding space by leveraging Jensen-Shannon Divergence for se-  
 505 mantic embedding alignment and InfoNCE for enhancing cross-lingual retrieval performance. Our  
 506 method mitigates linguistic misalignment and language bias, significantly improving cross-lingual  
 507 retrieval performance and effectively reducing performance disparities across languages. Additionally,  
 508 our method demonstrates stable performance even in monolingual settings.

## 509 ETHICS STATEMENT

511 This study proposes a cross-lingual embedding alignment methodology that can positively impact  
 512 applications requiring effective information retrieval across multiple languages. The proposed method  
 513 accurately aligns semantic representations across different languages to enhance retrieval performance,  
 514 while simultaneously strengthening the generalization and robustness of multilingual embedding  
 515 models. In particular, our approach is applicable even to low-resource languages, potentially miti-  
 516 gating global information disparity and contributing to the creation of a more equitable information  
 517 access environment. Moreover, by effectively utilizing existing large-scale embedding models and  
 518 data resources, our approach significantly reduces additional costs associated with data construction  
 519 and labeling. However, because the training process involves translations generated by large language  
 520 models, there are potential risks of subtle cultural nuances being distorted or the introduction of  
 521 data biases, which may lead to inaccurate or unintended outcomes for certain cultural or linguistic  
 522 groups. This study transparently acknowledges and carefully discusses these limitations and risks.  
 523 Nevertheless, we firmly believe that the expected benefits and positive impacts of our research  
 524 substantially outweigh the aforementioned concerns.

## 525 REPRODUCIBILITY STATEMENT

528 Our research is designed for full reproducibility. Details on the experimental setup, including datasets  
 529 and implementation specifics, are described in Section 5.1. Further training information, such as the  
 530 specific hyperparameters used, can be found in Appendix C. For the computational resources required  
 531 to run our experiments, please refer to Appendix C.

## 533 REFERENCES

535 Mofetoluwa Adeyemi, Akintunde Oladipo, Ronak Pradeep, and Jimmy Lin. Zero-shot cross-lingual  
 536 reranking with large language models for low-resource languages. In Lun-Wei Ku, Andre Martins,  
 537 and Vivek Srikanth (eds.), *Proceedings of the 62nd Annual Meeting of the Association for  
 538 Computational Linguistics (Volume 2: Short Papers)*, pp. 650–656, Bangkok, Thailand, August  
 539 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-short.59. URL  
<https://aclanthology.org/2024.acl-short.59/>.

540 Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of mono-  
 541 lingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Com-  
 542 putational Linguistics*. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.  
 543 acl-main.421. URL <http://dx.doi.org/10.18653/v1/2020.acl-main.421>.

544

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

552

553 Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding:  
 554 Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge  
 555 distillation, 2024. URL <https://arxiv.org/abs/2402.03216>.

556

557 Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-  
 558 Ling Mao, Heyan Huang, and Ming Zhou. Infoilm: An information-theoretic framework for  
 559 cross-lingual language model pre-training, 2020.

560

561 Adel Elmahdy, Sheng-Chieh Lin, and Amin Ahmad. Synergistic approach for simultaneous op-  
 562 timization of monolingual, cross-lingual, and multilingual information retrieval, 2024. URL  
 563 <https://arxiv.org/abs/2408.10536>.

564

565 D.M. Endres and J.E. Schindelin. A new metric for probability distributions. *IEEE Transactions on  
 566 Information Theory*, 49(7):1858–1860, 2003. doi: 10.1109/TIT.2003.813506.

567

568 Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap,  
 569 Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala,  
 570 Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan  
 571 Rystrøm, Roman Solomatin, Ömer Çağatan, Akash Kundu, Martin Bernstorff, Shitao Xiao, Akshita  
 572 Sukhlecha, Bhavish Pahwa, Rafał Poświata, Kranthi Kiran GV, Shawon Ashraf, Daniel Auras,  
 573 Björn Plüster, Jan Philipp Harries, Loïc Magne, Isabelle Mohr, Mariya Hendriksen, Dawei Zhu,  
 574 Hippolyte Gisserot-Boukhlef, Tom Aarsen, Jan Kostkan, Konrad Wojtasik, Taemin Lee, Marek  
 575 Šuppa, Cristina Zhang, Roberta Rocca, Mohammed Hamdy, Andrianos Michail, John Yang,  
 576 Manuel Fayolle, Aleksei Vatolin, Nandan Thakur, Manan Dey, Dipam Vasani, Pranjal Chitale,  
 577 Simone Tedeschi, Nguyen Tai, Artem Snegirev, Michael Günther, Mengzhou Xia, Weijia Shi,  
 578 Xing Han Lù, Jordan Clive, Gayatri Krishnakumar, Anna Maksimova, Silvan Wehrli, Maria  
 579 Tikhonova, Henil Panchal, Aleksandr Abramov, Malte Ostendorff, Zheng Liu, Simon Clematide,  
 580 Lester James Miranda, Alena Fenogenova, Guangyu Song, Ruqiya Bin Safi, Wen-Ding Li, Alessia  
 581 Borghini, Federico Cassano, Hongjin Su, Jimmy Lin, Howard Yen, Lasse Hansen, Sara Hooker,  
 582 Chenghao Xiao, Vaibhav Adlakha, Orion Weller, Siva Reddy, and Niklas Muennighoff. Mmteb:  
 583 Massive multilingual text embedding benchmark, 2025. URL <https://arxiv.org/abs/2502.13595>.

584

585 Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic  
 586 bert sentence embedding, 2020.

587

588 Ping Guo, Yue Hu, Yanan Cao, Yubing Ren, Yunpeng Li, and Heyan Huang. Query in your tongue:  
 589 Reinforce large language models with retrievers for cross-lingual search generative experience. In  
 590 *Proceedings of the ACM Web Conference 2024*, pp. 1529–1538, 2024.

591

592 Junjie Hu, Melvin Johnson, Orhan Firat, Aditya Siddhant, and Graham Neubig. Explicit alignment  
 593 objectives for multilingual bidirectional encoders. *arXiv preprint arXiv:2010.07972*, 2020.

594

595 Zhiqi Huang, Puxuan Yu, and James Allan. Improving cross-lingual information retrieval on low-  
 596 resource languages via optimal transport distillation. In *Proceedings of the Sixteenth ACM  
 597 International Conference on Web Search and Data Mining*, WSDM ’23, pp. 1048–1056. ACM,  
 598 February 2023a. doi: 10.1145/3539597.3570468. URL <http://dx.doi.org/10.1145/3539597.3570468>.

594 Zhiqi Huang, Puxuan Yu, and James Allan. Cross-lingual knowledge transfer via distillation for  
 595 multilingual information retrieval. *arXiv preprint arXiv:2302.13400*, 2023b.  
 596

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

599 Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi  
 600 Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *EMNLP*  
 601 (1), pp. 6769–6781, 2020.

602

603 Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized  
 604 late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on*  
 605 *research and development in Information Retrieval*, pp. 39–48, 2020.

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

608

609 Samuel Läubli, Rico Sennrich, and Martin Volk. Has machine translation achieved human parity? a  
 610 case for document-level evaluation. *arXiv preprint arXiv:1808.07048*, 2018.

611 Dawn Lawrie, Eugene Yang, Douglas W. Oard, and James Mayfield. Neural approaches to multilin-  
 612 gual information retrieval, 2023. URL <https://arxiv.org/abs/2209.01335>.  
 613

614 Dawn Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W. Oard, Luca Soldaini,  
 615 and Eugene Yang. Overview of the trec 2023 neuclir track, 2024. URL [https://arxiv.org/](https://arxiv.org/abs/2404.08071)  
 616 [abs/2404.08071](https://arxiv.org/abs/2404.08071).

617 Jungseob Lee, Hyeonseok Moon, Seungjun Lee, Chanjun Park, Sugyeong Eo, Hyunwoong Ko,  
 618 Jaehyung Seo, Seungyoon Lee, and Heui-Seok Lim. Length-aware byte pair encoding for mitigating  
 619 over-segmentation in korean machine translation. In *Findings of the Association for Computational*  
 620 *Linguistics ACL 2024*, pp. 2287–2303, 2024.

621

622 Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information*  
 623 *theory*, 37(1):145–151, 1991.

624 Jimmy Lin, David Alfonso-Hermelo, Vitor Jeronymo, Ehsan Kamaloo, Carlos Lassance, Rodrigo  
 625 Nogueira, Odunayo Ogundepo, Mehdi Rezagholizadeh, Nandan Thakur, Jheng-Hong Yang, and  
 626 Xinyu Zhang. Simple yet effective neural ranking and reranking baselines for cross-lingual  
 627 information retrieval, 2023. URL <https://arxiv.org/abs/2304.01019>.  
 628

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

632 Robert Litschko, Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. Evaluating multilingual text  
 633 encoders for unsupervised cross-lingual retrieval, 2021. URL [https://arxiv.org/abs/](https://arxiv.org/abs/2101.08370)  
 634 [2101.08370](https://arxiv.org/abs/2101.08370).

635

636 Robert Litschko, Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. On cross-lingual retrieval  
 637 with multilingual text encoders. *Information Retrieval Journal*, 25(2):149–183, 2022.

638 Qianchu Liu, Diana McCarthy, Ivan Vulić, and Anna Korhonen. Investigating cross-lingual alignment  
 639 methods for contextualized embeddings with token-level evaluation. 2019.

640

641 Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL <https://arxiv.org/abs/1711.05101>.  
 642

643 James Mayfield, Eugene Yang, Dawn Lawrie, Samuel Barham, Orion Weller, Marc Mason, Suraj  
 644 Nair, and Scott Miller. Synthetic cross-language information retrieval training data. *arXiv preprint*  
 645 *arXiv:2305.00331*, 2023.

646

647 Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with bert. *arXiv preprint*  
 648 *arXiv:1901.04085*, 2019.

648 OpenAI. Gpt-4o system card, 2024. URL <https://arxiv.org/abs/2410.21276>.  
649

650 Jeonghyun Park and Hwanhee Lee. Investigating language preference of multilingual rag systems,  
651 2025. URL <https://arxiv.org/abs/2502.11175>.  
652

653 Changle Qu, Sunhao Dai, Xiaochi Wei, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Jun Xu, and Ji-  
654 Rong Wen. Towards completeness-oriented tool retrieval for large language models. In *Proceedings*  
655 *of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM  
656 '24, pp. 1930–1940. ACM, October 2024. doi: 10.1145/3627673.3679847. URL <http://dx.doi.org/10.1145/3627673.3679847>.  
657

658 Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions  
659 for machine comprehension of text. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings*  
660 *of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392,  
661 Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL <https://aclanthology.org/D16-1264/>.  
662

663 Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. Cross-lingual alignment of con-  
664 textual word embeddings, with applications to zero-shot dependency parsing. *arXiv preprint*  
665 *arXiv:1902.09492*, 2019.  
666

667 Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel,  
668 Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, and Han Xiao. jina-embeddings-  
669 v3: Multilingual embeddings with task lora, 2024. URL <https://arxiv.org/abs/2409.10173>.  
670

671 Antonio Toral and Andy Way. What level of quality can neural machine translation attain on literary  
672 text? *Translation quality assessment: From principles to practice*, pp. 263–287, 2018.  
673

674 Francisco Valentini, Diego Kozlowski, and Vincent Larivière. Clirudit: Cross-lingual information  
675 retrieval of scientific documents, 2025. URL <https://arxiv.org/abs/2504.16264>.  
676

677 Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive  
678 coding, 2019. URL <https://arxiv.org/abs/1807.03748>.  
679

680 Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilin-  
681 gual e5 text embeddings: A technical report, 2024. URL <https://arxiv.org/abs/2402.05672>.  
682

683 Shijie Wu and Mark Dredze. Are all languages created equal in multilingual bert? *arXiv preprint*  
684 *arXiv:2005.09093*, 2020.  
685

686 Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and  
687 Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text  
688 retrieval. *arXiv preprint arXiv:2007.00808*, 2020.  
689

690 Eugene Yang, Thomas Jänich, James Mayfield, and Dawn Lawrie. Language fairness in multilingual  
691 information retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research*  
692 *and Development in Information Retrieval*, SIGIR 2024, pp. 2487–2491. ACM, July 2024. doi:  
693 10.1145/3626772.3657943. URL <http://dx.doi.org/10.1145/3626772.3657943>.  
694

695 Puxuan Yu, Hongliang Fei, and Ping Li. Cross-lingual language model pretraining for retrieval. In  
696 *Proceedings of the Web Conference 2021*, WWW '21, pp. 1029–1039, New York, NY, USA, 2021.  
697 Association for Computing Machinery. ISBN 9781450383127. doi: 10.1145/3442381.3449830.  
698 URL <https://doi.org/10.1145/3442381.3449830>.  
699

700 Fuwei Zhang, Zhao Zhang, Xiang Ao, Dehong Gao, Fuzhen Zhuang, Yi Wei, and Qing He. Mind the  
701 gap: Cross-lingual information retrieval with hierarchical knowledge enhancement. In *Proceedings*  
*of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 4345–4353, 2022.

702 Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong  
703 Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. mGTE: General-  
704 ized long-context text representation and reranking models for multilingual text retrieval. In  
705 Franck Dernoncourt, Daniel Preoziuc-Pietro, and Anastasia Shimorina (eds.), *Proceedings of the*  
706 *2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp.  
707 1393–1412, Miami, Florida, US, November 2024. Association for Computational Linguistics.  
708 doi: 10.18653/v1/2024.emnlp-industry.103. URL <https://aclanthology.org/2024.emnlp-industry.103/>.  
709

710 Xinyu Zhang, Kelechi Ogueji, Xueguang Ma, and Jimmy Lin. Toward best practices for training  
711 multilingual dense retrieval models. *ACM Transactions on Information Systems*, 42(2):1–33,  
712 2023a.

713 Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo,  
714 Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. MIRACL: A multilingual  
715 retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational*  
716 *Linguistics*, 11:1114–1131, 2023b. doi: 10.1162/tacl\_a\_00595. URL <https://aclanthology.org/2023.tacl-1.63/>.  
717

718

719

720

721

722

723

724

725

726

727

728

729

730

731

732

733

734

735

736

737

738

739

740

741

742

743

744

745

746

747

748

749

750

751

752

753

754

755

756 **A LIMITATIONS**

759 A major limitation of this study arises from the experimental setup and evaluation process for cross-  
 760 lingual alignment, which are primarily centered around English despite the extensive possibilities of  
 761 language combinations. In real-world scenarios, various language pairs that do not involve English  
 762 frequently exist, requiring consideration. Nonetheless, experiments and evaluations in this study were  
 763 intentionally centered around English, the most widely used and high-resource language, under the  
 764 assumption that multilingual models would exhibit significant bias toward alignment with English,  
 765 and to reflect the most practical real-world scenarios.

766 Additionally, although our evaluations considered various scenarios (Multi, Multi-1, Mono) to assess  
 767 cross-lingual retrieval performance, the current experimental design focused mainly on two languages  
 768 settings. In practice, multilingual contexts involving more than two languages frequently occur.  
 769 However, we restricted our evaluations primarily to cross-lingual settings because existing models  
 770 still struggle with performance even in these simpler setups.

771 Moreover, our approach utilized machine translation based on language models to build the training  
 772 dataset. Compared to human translation, automatically translated data may fail to capture subtle  
 773 linguistic nuances and cultural contexts sufficiently Toral & Way (2018); Läubli et al. (2018); Lee  
 774 et al. (2024). However, we adopted this methodology as it currently represents the most practical and  
 775 efficient solution available for constructing large-scale multilingual datasets.

776 **B EVALUATION BENCHMARK DETAILS**

777 In this study, we utilize multilingual Question-Answering (QA) datasets with parallel structures,  
 778 converted into retrieval tasks, to evaluate cross-lingual retrieval performance. Specifically, we use  
 779 XQuAD<sup>1</sup>, a multilingual question-answer dataset derived from SQuAD 1.1 Rajpurkar et al. (2016),  
 780 which consists of fully parallel question-answer pairs across 13 languages including English. XQuAD  
 781 was directly translated by professional translators, ensuring a precise one-to-one correspondence  
 782 between documents and queries across languages. This high-quality translation process naturally  
 783 preserves linguistic expressions and semantic meanings in each target language, making XQuAD  
 784 particularly suitable for assessing the robustness of embedding models against language variations  
 785 in cross-lingual scenarios. Additionally, we utilize Belebele<sup>2</sup>, another multilingual QA dataset  
 786 that includes diverse language pairs. Belebele was carefully translated from English into multiple  
 787 languages by native-speaking translators proficient in English, effectively capturing contextual  
 788 nuances and cultural subtleties Bandarkar et al. (2024). Due to these characteristics, Belebele provides  
 789 realistic and varied scenarios reflecting practical multilingual retrieval environments, enabling fine-  
 790 grained comparison and analysis of retrieval performance across languages. **Furthermore, our use**  
 791 **of these specific benchmarks aligns with standard practices for robust multilingual evaluation.** Both  
 792 **XQuAD and Belebele are core benchmarks used for assessing multilingual retrieval capabilities in the**  
 793 **prominent Massive Multilingual Text Embedding Benchmark (MMTEB) Enevoldsen et al. (2025)<sup>3</sup>.**

795 **C TRAINING DETAILS**

796 **Dataset Translation** We employed a Large Language Model to translate datasets for training  
 797 purposes. The format template for translation prompts used as inputs to the model is as follows:

801 System: #Instructions  
 802       Translate the following English passage fully and  
 803       accurately into {target language}  
 804  
 805 User: {English passage}  
 806 Assistant: {Translated passage}

808 <sup>1</sup><https://huggingface.co/datasets/google/xquad>

809 <sup>2</sup><https://huggingface.co/datasets/facebook/belebele>

810 <sup>3</sup><https://github.com/embeddings-benchmark/mteb>

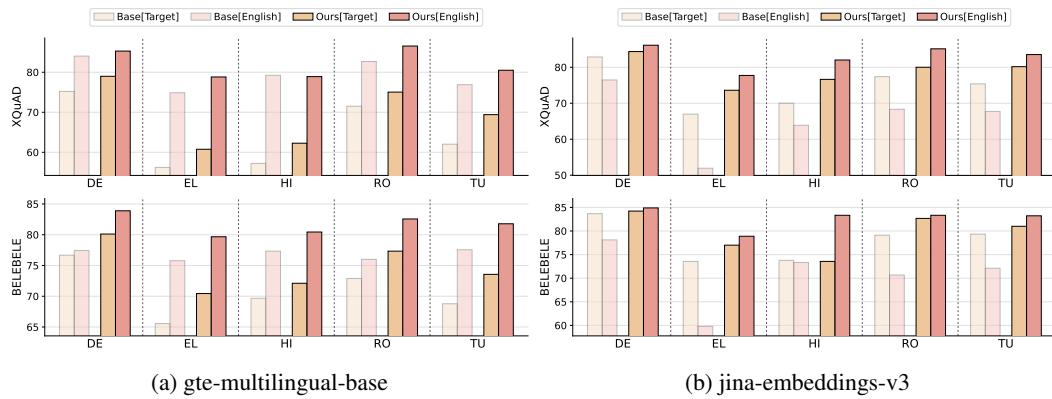
810  
 811 **Hyperparams** All models were trained for a total of one epoch with a batch size of 32, em-  
 812 ploying a linear learning rate scheduler with a warm-up ratio of 0.15. We used the AdamW opti-  
 813 mizer (Loshchilov & Hutter, 2019) (with parameters  $\beta_1 = 0.9$ ,  $\beta_2 = 0.99$ , and weight decay=0.01),  
 814 and adopted bfloat16 mixed precision for computational efficiency. Considering the characteris-  
 815 tics of each model, we set the initial learning rates as follows: 4e-6 for bge-M3, 2e-5 for multilingual-E5-  
 816 large, 1e-5 for gte-multilingual-base, and 3e-5 for jina-embeddings-v3.

817 **Hardware** We used 2 NVIDIA A100 GPUs, each with 80GB of memory capacity, along with  
 818 AMD EPYC 7513 processors featuring 32 cores, to train models. For evaluation, we employed a  
 819 single accelerator.

820 **Reproducibility** To ensure a fair and consistent comparison, all experimental results reported in  
 821 this paper are based on a single run using a fixed random seed (42). This fixed seed controls all  
 822 stochastic elements, including data shuffling and the batch composition and sampling order. This  
 823 approach allows us to strictly control the training environment for both baselines and our proposed  
 824 model, ensuring that the comparison fairly isolates the effect of our methodology.

## 826 D EXTENDED RESULTS FOR ADDITIONAL LANGUAGES

827 In this section, we provide experimental results for three scenarios involving five languages not  
 828 previously included in the main body of the paper: German (DE), Greek (EL), Hindi (HI), Romanian  
 829 (RO), and Turkish (TU).



830  
 831 Figure 4: NDCG@1 comparison on additional language pairs in the **Multi-1 scenario** with [lang] as  
 832 the query language.

833  
 834  
 835  
 836  
 837  
 838  
 839  
 840  
 841  
 842  
 843  
 844  
 845  
 846  
 847  
 848  
 849  
 850  
 851  
 852  
 853  
 854  
 855  
 856  
 857  
 858  
 859  
 860  
 861  
 862  
 863

Table 4: Performance comparison on remaining language pairs under the **Multi scenario**.

Doc	Query	XQuAD						Belebele					
		Comp@10		Max@R (↓)		Max@R-norm		Comp@10		Max@R (↓)		Max@R-norm	
Base	Ours	Base	Ours	Base	Ours	Base	Ours	Base	Ours	Base	Ours	Base	Ours
<i>multilingual-e5-base</i>													
En+De	En	54.20	<b>65.04</b>	24.16	<b>16.10</b>	64.82	<b>70.55</b>	66.89	<b>88.33</b>	52.44	<b>13.98</b>	51.98	<b>71.41</b>
	De	39.41	<b>62.35</b>	45.38	<b>21.00</b>	55.91	<b>66.80</b>	53.00	<b>86.44</b>	70.03	<b>16.05</b>	47.73	<b>69.38</b>
En+El	En	24.45	<b>62.10</b>	63.15	<b>17.59</b>	51.25	<b>69.30</b>	33.67	<b>86.56</b>	140.91	<b>15.18</b>	37.45	<b>70.20</b>
	El	7.73	<b>53.18</b>	248.40	<b>30.10</b>	31.91	<b>61.71</b>	11.78	<b>76.78</b>	287.99	<b>22.93</b>	26.94	<b>64.14</b>
En+Hi	En	46.97	<b>62.61</b>	26.87	<b>18.84</b>	63.32	<b>68.33</b>	65.33	<b>87.89</b>	47.43	<b>14.08</b>	53.46	<b>71.31</b>
	Hi	26.55	<b>53.70</b>	101.12	<b>30.74</b>	44.60	<b>61.41</b>	45.78	<b>76.00</b>	86.64	<b>27.37</b>	44.60	<b>61.54</b>
En+Ro	En	53.61	<b>65.45</b>	23.14	<b>15.57</b>	65.43	<b>71.03</b>	64.22	<b>89.11</b>	55.41	<b>14.91</b>	51.17	<b>70.47</b>
	Ro	23.28	<b>61.01</b>	89.76	<b>24.89</b>	46.28	<b>64.40</b>	39.56	<b>82.44</b>	145.65	<b>20.94</b>	39.96	<b>65.48</b>
En+Tu	En	42.44	<b>62.69</b>	35.54	<b>17.82</b>	59.37	<b>69.12</b>	59.11	<b>87.51</b>	10.04	<b>7.51</b>	76.28	<b>80.54</b>
	Tu	24.20	<b>55.21</b>	94.54	<b>30.75</b>	45.55	<b>61.41</b>	44.22	<b>80.00</b>	126.08	<b>24.93</b>	39.08	<b>62.91</b>
<i>gte-multilingual-base</i>													
En+De	En	75.13	<b>75.71</b>	10.45	<b>9.86</b>	76.65	<b>77.48</b>	88.56	<b>92.67</b>	8.79	<b>6.99</b>	78.23	<b>81.61</b>
	De	67.90	<b>71.09</b>	16.27	<b>14.08</b>	70.40	<b>72.45</b>	87.67	<b>89.22</b>	13.30	<b>9.79</b>	72.15	<b>76.66</b>
En+El	En	67.14	<b>70.84</b>	11.95	<b>10.98</b>	74.75	<b>75.95</b>	88.67	<b>90.22</b>	10.34	<b>8.44</b>	75.85	<b>78.84</b>
	El	51.68	<b>55.71</b>	25.08	<b>20.28</b>	64.29	<b>67.29</b>	81.00	<b>84.33</b>	24.88	<b>18.72</b>	62.94	<b>67.12</b>
En+Hi	En	71.93	<b>72.27</b>	12.07	<b>11.52</b>	74.61	<b>75.28</b>	89.00	<b>91.67</b>	9.35	<b>8.82</b>	77.33	<b>78.18</b>
	Hi	54.96	<b>58.66</b>	27.76	<b>23.51</b>	62.86	<b>65.20</b>	82.11	<b>84.00</b>	23.81	<b>19.66</b>	63.59	<b>66.40</b>
En+Ro	En	73.70	<b>76.47</b>	11.04	<b>9.60</b>	75.88	<b>77.85</b>	89.22	<b>92.89</b>	9.28	<b>7.19</b>	77.44	<b>81.20</b>
	Ro	64.54	<b>67.56</b>	16.94	<b>14.13</b>	69.83	<b>72.40</b>	84.89	<b>88.22</b>	17.44	<b>12.57</b>	68.17	<b>72.97</b>
En+Tu	En	69.24	<b>71.68</b>	12.82	<b>11.03</b>	73.76	<b>75.89</b>	90.22	<b>92.33</b>	10.04	<b>7.51</b>	76.28	<b>80.54</b>
	Tu	55.29	<b>61.51</b>	26.18	<b>20.31</b>	63.68	<b>67.27</b>	83.22	<b>86.89</b>	19.65	<b>14.53</b>	66.41	<b>70.85</b>
<i>jina-embeddings-v3</i>													
En+De	En	69.50	<b>75.80</b>	15.12	<b>10.08</b>	71.44	<b>77.16</b>	89.00	<b>92.89</b>	14.25	<b>9.00</b>	71.14	<b>77.88</b>
	De	72.86	<b>74.79</b>	13.94	<b>11.08</b>	72.58	<b>75.82</b>	91.33	<b>92.33</b>	11.16	<b>9.30</b>	74.75	<b>77.41</b>
En+El	En	53.03	<b>71.76</b>	31.84	<b>11.98</b>	60.93	<b>74.73</b>	80.56	<b>89.22</b>	36.69	<b>13.11</b>	57.21	<b>72.35</b>
	El	57.31	<b>67.31</b>	27.76	<b>16.49</b>	62.86	<b>70.21</b>	83.22	<b>87.33</b>	32.88	<b>22.58</b>	58.83	<b>64.37</b>
En+Hi	En	61.93	<b>74.54</b>	20.27	<b>10.88</b>	67.30	<b>76.00</b>	87.56	<b>91.67</b>	20.73	<b>11.15</b>	65.62	<b>74.74</b>
	Hi	62.86	<b>69.58</b>	24.82	<b>16.52</b>	64.43	<b>70.19</b>	85.00	<b>86.67</b>	28.77	<b>20.32</b>	60.83	<b>65.92</b>
En+Ro	En	64.54	<b>74.20</b>	24.53	<b>10.65</b>	64.63	<b>76.38</b>	83.22	<b>92.22</b>	31.87	<b>10.17</b>	59.30	<b>76.10</b>
	Ro	65.55	<b>70.76</b>	19.67	<b>12.90</b>	67.73	<b>73.68</b>	86.78	<b>91.00</b>	25.46	<b>14.12</b>	62.59	<b>71.27</b>
En+Tu	En	62.44	<b>73.87</b>	21.33	<b>10.54</b>	66.58	<b>76.53</b>	86.67	<b>93.00</b>	18.45	<b>9.39</b>	67.35	<b>77.27</b>
	Tu	65.71	<b>71.60</b>	18.89	<b>13.19</b>	68.28	<b>73.36</b>	87.56	<b>89.89</b>	18.65	<b>12.71</b>	67.19	<b>72.82</b>
<i>bge-m3</i>													
En+De	En	76.13	<b>77.48</b>	10.02	<b>9.96</b>	77.24	<b>77.33</b>	93.00	<b>93.78</b>	9.40	<b>8.54</b>	77.25	<b>78.66</b>
	De	<b>76.22</b>	75.97	<b>10.14</b>	10.44	<b>77.08</b>	76.67	92.22	<b>93.22</b>	8.44	<b>8.12</b>	78.83	<b>79.41</b>
En+El	En	68.24	<b>72.94</b>	13.62	<b>12.11</b>	72.91	<b>74.58</b>	89.33	<b>91.67</b>	11.11	<b>9.73</b>	74.79	<b>76.75</b>
	El	68.32	<b>70.84</b>	13.23	<b>13.23</b>	73.32	<b>73.32</b>	86.22	<b>87.89</b>	17.75	<b>14.01</b>	67.91	<b>71.38</b>
En+Hi	En	69.66	<b>72.86</b>	11.96	<b>11.55</b>	74.75	<b>75.24</b>	89.89	<b>91.44</b>	11.10	<b>10.04</b>	74.80	<b>76.29</b>
	Hi	68.07	<b>70.42</b>	13.69	<b>13.19</b>	72.84	<b>73.36</b>	83.11	<b>85.89</b>	21.98	<b>19.88</b>	64.76	<b>66.24</b>
En+Ro	En	78.07	<b>78.32</b>	9.97	<b>9.66</b>	77.31	<b>77.75</b>	92.67	<b>93.22</b>	9.66	<b>8.56</b>	76.84	<b>78.63</b>
	Ro	74.03	<b>75.13</b>	11.66	<b>11.17</b>	75.10	<b>75.71</b>	91.22	<b>92.67</b>	10.13	<b>9.41</b>	76.15	<b>77.23</b>
En+Tu	En	73.28	<b>74.79</b>	11.48	<b>11.25</b>	75.32	<b>75.61</b>	91.56	<b>92.11</b>	9.52	<b>8.46</b>	77.07	<b>78.79</b>
	Tu	71.34	<b>72.27</b>	13.38	<b>12.33</b>	73.16	<b>74.32</b>	89.56	<b>90.78</b>	13.82	<b>11.75</b>	71.58	<b>73.96</b>

Table 5: Performance comparison under the **Mono scenario** for five additional languages using the jina-embeddings-v3 model: English/target language queries.

Lang	Model	XQuAD				Belebele			
		Mono-Same		Mono-Cross		Mono-Same		Mono-Cross	
		NDCG@1	MRR	NDCG@1	MRR	NDCG@1	MRR	NDCG@1	MRR
De	Base	89.2/90.0	0.933/0.937	87.8/87.9	0.923/0.925	<b>88.4/89.3</b>	<b>0.919/0.926</b>	84.9/86.0	0.897/0.903
	Ours	<b>90.6/88.4</b>	<b>0.943/0.929</b>	<b>89.1/88.9</b>	<b>0.933/0.932</b>	87.6/88.1	0.916/0.916	<b>86.8/86.3</b>	<b>0.909/0.905</b>
El	Base	89.2/80.5	0.933/0.873	79.5/81.7	0.863/0.884	88.4/83.6	0.919/0.883	80.1/78.3	0.852/0.840
	Ours	<b>90.0/86.8</b>	<b>0.940/0.916</b>	<b>88.4/83.5</b>	<b>0.928/0.895</b>	<b>88.8/84.0</b>	<b>0.924/0.887</b>	<b>83.4/81.1</b>	<b>0.883/0.864</b>
Hi	Base	89.2/83.3	0.933/0.891	81.6/84.4	0.883/0.899	88.4/81.1	0.919/0.860	82.3/79.4	0.875/0.852
	Ours	<b>91.6/87.5</b>	<b>0.950/0.922</b>	<b>89.0/85.2</b>	<b>0.933/0.907</b>	<b>89.1/82.2</b>	<b>0.926/0.872</b>	<b>86.0/79.4</b>	<b>0.902/0.857</b>
Ro	Base	<b>89.2/84.7</b>	<b>0.933/0.899</b>	<b>82.4/85.3</b>	<b>0.885/0.909</b>	<b>88.4/85.1</b>	<b>0.919/0.892</b>	81.7/82.6	0.869/0.875
	Ours	88.7/86.5	0.932/0.917	88.1/85.2	0.927/0.908	87.2/87.0	0.914/0.907	<b>84.8/85.2</b>	<b>0.895/0.898</b>
Tr	Base	89.2/84.9	0.933/0.901	82.2/84.8	0.883/0.903	88.4/85.9	0.919/0.896	82.0/82.2	0.874/0.874
	Ours	<b>89.5/88.6</b>	<b>0.938/0.928</b>	<b>89.1/85.2</b>	<b>0.935/0.907</b>	<b>88.4/86.7</b>	<b>0.921/0.908</b>	<b>87.0/84.9</b>	<b>0.911/0.892</b>