

Improving Korean-English Cross-Lingual Retrieval: A Data-Centric Study of Language Composition and Model Merging

Youngjoon Jang¹, Junyoung Son¹, Taemin Lee¹, Seongtae Hong¹, Hyeonseok Moon¹
Seungyoon Lee¹, Andrew Matteson², Heuseok Lim^{1†}

¹Korea University ²Uracle
{dew1701, s0ny, taeminlee, ghdchlwlsl23,
glee889, dltmddb100, limhseok}@korea.ac.kr
amatteson@uracle.co.kr

Abstract

With the increasing utilization of multilingual text information, Cross-Lingual Information Retrieval (CLIR) has become a crucial research area. However, the impact of training data composition on CLIR and Mono-Lingual Information Retrieval (Mono-IR) performance remains underexplored. In this work, we conduct a data-centric study on Korean–English CLIR by systematically varying the language composition of the dataset for dense retriever fine-tuning. Using linguistically parallel training sets that support all cross-lingual permutations, we analyze how different compositions shape retrieval behavior across multiple embedding models and evaluation benchmarks. Our results reveal a consistent trade-off: configurations that substantially improve CLIR can simultaneously degrade Mono-IR, indicating that retrieval gains are coupled through inter-lingual correlations induced during training. We further identify which components of the training data are primarily responsible for the observed gains and regressions, providing actionable guidance for dataset construction. Finally, as a lightweight practical option, we show that a simple post-hoc weight averaging with the base model can partially recover Mono-IR performance while retaining most of the CLIR improvements.

1 Introduction

Information Retrieval (IR) aims to retrieve documents that are relevant to a user query, and it has become a core component of modern Retrieval-Augmented Generation (RAG) and agentic systems (Lewis et al., 2020; Gao et al., 2023; Singh et al., 2025). Although the significance of IR is widely recognized, the majority of research has centered on monolingual settings (Zhao et al., 2024). This focus often overlooks Cross-Lingual Information Retrieval (CLIR) scenarios, where the languages of the queries and documents differ.

Previous CLIR studies introduce various training data compositions to enhance cross-lingual performance. This includes pairing queries and passages in different languages and diversifying the languages of query–passage pairs (Goswami et al., 2021; Zhang et al., 2022; Gao et al., 2001; Bhat-tacharya et al., 2016; Saleh and Pecina, 2020). Despite these efforts, there has been scarce discussion on how different training data compositions affect the performances of Mono-IR and CLIR.

In this work, we present a data-centric study of Korean–English CLIR that systematically examines the role of language composition in training data. We first construct a synthetic Korean–English parallel training dataset that supports controlled variations of language assignments within training triples. Using this dataset, we fine-tune dense retrievers under different compositions and analyze their effects on both Mono-IR and CLIR across multiple models and evaluation benchmarks. Our experiments reveal a consistent trade-off: certain language pairings in training substantially improve CLIR in a given direction while degrading Mono-IR, indicating that training data composition can induce coupled changes in retrieval behavior.

Beyond documenting this trade-off, we identify which aspects of composition most strongly influence the retrieval performance, providing guidance for constructing CLIR training data. Finally, as a lightweight practical baseline, we report that simple weight-averaged model merging (Matena and Raffel, 2022) with the base model can partially reduce the Mono-IR drop while retaining much of the CLIR improvement.

2 Related Work

Information Retrieval (IR) aims to retrieve documents that are relevant to a given query (Singhal, 2001; Zhu et al., 2023; Li et al., 2025). Recent retrieval methods include dense retrieval (Karpukhin

[†]Corresponding author.

et al., 2020; Xiong et al., 2020; Wang et al., 2024; Zhang et al., 2024; Chen et al., 2024), sparse retrieval (Dai and Callan, 2020; Formal et al., 2021b,a), and multi-vector retrieval (Khattab and Zaharia, 2020; Liu and Mao, 2023). Most of these approaches have been developed and evaluated primarily in monolingual settings, where the query and document collection are written in the same language.

Cross-Lingual Information Retrieval (CLIR) extends this setting by retrieving relevant documents written in a language different from that of the query. Recent studies have improved CLIR through multilingual representation learning, cross-lingual pretraining, translation-based supervision, and knowledge transfer across languages (Goswami et al., 2021; Yu et al., 2021; Litschko et al., 2022; Huang et al., 2023; Nguyen et al., 2025). While these studies mainly focus on improving cross-lingual retrieval accuracy, the role of training data composition itself remains less explored.

In this work, we conduct a data-centric analysis of Korean–English CLIR by isolating the language composition of training triples. Specifically, we systematically vary the languages of queries, positives, and negatives while keeping the underlying semantics parallel across variants. This controlled setup allows us to analyze how different language compositions affect both CLIR and monolingual retrieval behavior, revealing a trade-off between cross-lingual gains and monolingual degradation.

3 Synthetic Data Generation

Our study aims to analyze the impact of language composition on IR performance. As recent embedding models utilize a diverse range of datasets, employing open-source data might risk overlapping with previously trained data. Furthermore, there is a scarcity of linguistically parallel query-positive (a passage providing answer to query) data. To address these issues and more robustly assess the effect of language composition, we build a high-quality training dataset that is linguistically parallel and covers a diverse range of domains.

KDC Sampling To ensure broad domain coverage and overcome potential LLM domain limitations, we incorporate the Korean Decimal Classification (KDC) (Oh, 2020). KDC is a Korean adaptation of the Dewey Decimal Classification (DDC) (Comaromi, 1984), a comprehensive library classification system. KDC has a hierarchical struc-

ture, where each digit represents a specific subject level: the first digit signifies a broad subject area, the second a subcategory, and the third a more specific category. For example, in the code 761, the 7 stands for the main class 'Language', the 6 for the subclass 'French language', and the 1 for the specific topic 'Phonology and writing systems'. We use all levels of KDC, which consists of 910 topics.

Task Brainstorming & Triple Generation Following the synthetic data construction procedure of E5-Mistral (Wang et al., 2023), we generate training triples in two steps. For each sampled KDC category, we first obtain a set of task descriptions, which are then used as prompts to produce {query, positive, hard_negative} triples. We deduplicate the resulting queries and passages, and mine six hard negatives per query to strengthen training signals. Finally, we construct six Korean–English cross-lingual variants of the triples to cover all cross-lingual permutations. Additional details on dataset generation, dataset statistics, and human quality assessment are provided in Appendix A.

4 Experimental Setup

Training Details We fine-tune four base models: bge-m3 (Chen et al., 2024), multilingual-e5-large, multilingual-e5-base (Wang et al., 2024), and gte-multilingual-base (Zhang et al., 2024) using each of the language combination datasets constructed in the previous section. The training employs GIS-TEmbedLoss (Solatorio, 2024), a contrastive loss function, reducing noise from random in-batch negatives. Comprehensive training parameters and environment details are provided in Appendix A.5.

Evaluation Datasets & Metrics Since there are few retrieval benchmarks covering both Korean and English, we employ two retrieval benchmarks: Belebele (Bandarkar et al., 2023) and StrategyQA (Geva et al., 2021). Both benchmarks are included in the official retrieval tasks of Massive Multilingual Text Embedding Benchmark (MMTEB) (Enevoldsen et al., 2025), which are widely adopted in contemporary retrieval works. Details about our evaluation benchmark are stated in Appendix B. We report nDCG@10 (Wang et al., 2013) as our primary evaluation metric, following the convention of MMTEB.

Data Notation We use specific notations to describe the language composition of our training data and evaluation task. The training data language

Data	Cross-Lingual							Mono-Lingual						
	Belebele		StrategyQA		AVG		OVR	Belebele		StrategyQA		AVG		OVR
	en-ko	ko-en	en-ko	ko-en	en-ko	ko-en		ko-ko	en-en	ko-ko	en-en	ko-ko	en-en	
<i>bge-m3</i>														
base	90.37	88.36	81.24	71.65	85.81	80.01	82.91	93.16	95.55	79.41	84.42	86.29	89.99	88.14
kokoen	90.32	87.95	81.64	71.39	85.98	79.67	82.83	92.80	95.16	79.68	84.29	86.24	89.73	87.98
koenko	89.32	89.08	79.44	72.92	84.38	81.00	82.69	91.77	94.64	78.06	83.92	84.92	89.28	87.10
koenen	91.46	89.65	81.69	72.86	86.58	81.26	83.92	92.65	95.03	79.13	84.06	85.89	89.55	87.72
enenko	89.09	88.58	80.76	71.68	84.93	80.13	82.53	93.19	95.38	79.34	84.43	86.27	89.91	88.09
enkoen	91.59	88.29	81.68	71.89	86.64	80.09	83.36	91.89	94.78	78.99	83.92	85.44	89.35	87.40
enkoko	91.24	89.00	81.81	72.73	86.53	80.87	<u>83.70</u>	92.29	94.83	79.31	84.24	85.80	89.54	87.67
<i>multilingual-e5-large (mE5-large)</i>														
base	92.03	86.45	82.04	67.30	87.04	76.88	81.96	94.50	96.50	80.35	84.20	87.43	90.35	88.89
kokoen	92.82	86.40	82.63	70.72	87.73	78.56	83.14	92.87	95.84	80.13	84.71	86.50	<u>90.28</u>	88.39
koenko	81.02	89.43	68.20	73.33	74.61	81.38	78.00	84.66	94.60	71.11	84.08	77.89	89.34	83.61
koenen	93.34	89.27	82.63	72.98	87.99	81.13	84.56	92.12	95.12	79.89	84.13	86.01	89.63	87.82
enenko	90.65	87.66	80.46	70.21	85.56	78.94	82.25	92.96	95.47	80.14	84.94	<u>86.55</u>	90.21	88.38
enkoen	93.05	74.77	82.11	58.88	87.58	66.83	77.20	91.22	89.94	79.33	79.98	85.28	84.96	85.12
enkoko	93.17	88.95	82.91	71.92	88.04	80.44	<u>84.24</u>	91.46	95.63	80.06	84.45	85.76	90.04	87.90
<i>multilingual-e5-base (mE5-base)</i>														
base	82.28	77.52	77.94	54.44	80.11	65.98	73.05	92.87	95.79	76.36	83.21	84.62	89.50	87.06
kokoen	85.26	79.34	79.19	57.89	82.23	68.62	75.42	90.55	95.81	75.80	84.01	83.18	89.91	86.54
koenko	75.30	83.35	60.05	62.84	67.68	<u>73.10</u>	70.39	83.31	93.75	62.48	82.68	72.90	88.22	80.56
koenen	89.02	83.29	80.06	63.23	84.54	73.26	78.90	89.61	94.17	74.71	83.65	82.16	88.91	85.54
enenko	85.17	80.31	76.93	58.59	81.05	69.45	75.25	90.92	95.22	75.48	84.20	<u>83.20</u>	89.71	86.46
enkoen	88.94	77.93	80.29	55.28	<u>84.62</u>	66.61	75.61	87.37	94.74	73.90	82.35	80.64	88.55	84.59
enkoko	89.43	82.50	80.60	61.83	85.02	72.17	<u>78.59</u>	88.10	94.79	73.76	83.90	80.93	89.35	85.14
<i>gte-multilingual-base (mgte-base)</i>														
base	88.06	85.66	80.54	64.38	84.30	75.02	79.66	87.96	94.63	75.12	84.26	81.54	89.45	85.49
kokoen	90.32	82.31	80.97	63.33	85.65	72.82	79.23	91.57	94.21	76.92	84.28	<u>84.25</u>	89.25	<u>86.75</u>
koenko	80.19	86.43	76.19	64.91	78.19	75.67	76.93	82.62	94.86	68.53	84.63	75.58	89.75	82.66
koenen	89.95	86.56	81.22	65.10	85.59	<u>75.83</u>	80.71	89.21	94.71	75.44	84.48	82.33	89.60	85.96
enenko	86.67	87.09	79.13	64.53	82.90	75.81	79.36	89.38	95.69	75.08	84.47	82.23	90.08	86.16
enkoen	90.31	77.49	81.26	64.16	<u>85.79</u>	70.83	78.31	89.81	89.91	75.49	84.00	82.65	86.96	84.80
enkoko	90.64	86.12	81.56	64.93	86.10	75.53	80.81	89.66	95.12	75.71	84.81	82.69	89.97	86.33

Table 1: CLIR (Left) and Mono-IR (Right) performance. Data notation: languages of query, positive, negatives (e.g., kokoen for Korean query, Korean positive and English negatives). Task notation: $Q_L - P_L$ (e.g., en-ko for English query, Korean documents). Highest scores in **bold**, second highest underlined. Cell colors indicate changes relative to the corresponding base model, with green/red denoting improvement/degradation. OVR denotes the micro-average over the two language directions in each block: *en-ko/ko-en* for CLIR and *ko-ko/en-en* for Mono-IR.

notation $\mathcal{D}(l_1, l_2, l_3)$ specifies the languages of the query, positive, and negatives sequentially. For example, $\mathcal{D}(ko, en, ko)$ indicates a Korean query, an English positive, and Korean negatives dataset. Task notations $\mathcal{T}(l_1 - l_2)$ indicate the languages of the query and the document pool respectively. For instance, $\mathcal{T}(en - ko)$ refers to a task for retrieving relevant documents from a Korean document pool given an English query. Additionally, to simplify exposition in the Analysis section, we denote the query language by Q_L , the positive language by P_L , and the language of the negatives by N_L .

5 Experiments and Analysis

5.1 Cross-Lingual Retrieval Results

On the left side of Table 1, we present a comprehensive summary of CLIR performance for the base models and all six training data combinations.

Enhancement by task-aligned Query-Positive

For a given CLIR language direction $\mathcal{T}(Q_L - P_L)$, training on query-positive pairs that match this direction improves performance on the specific task. For example, in the $\mathcal{T}(en - ko)$ task, the bge-m3 model trained with $\mathcal{D}(en, ko, en)$ achieves a higher AVG score (86.64) than when trained with $\mathcal{D}(en, en, ko)$ (84.93). Similarly, for the mE5-

base model on the $\mathcal{T}(ko - en)$ task, training with $\mathcal{D}(ko, en, en)$ (73.26) or $\mathcal{D}(ko, en, ko)$ (73.10) outperforms the base model (65.98) or training with $\mathcal{D}(ko, ko, en)$ (68.62). These results indicate that aligning (Q_L, P_L) with the evaluation direction reliably improves performance on the corresponding CLIR task.

Enhancement by Positive-Negative Language Match

Also, our results demonstrate that when $Q_L \neq P_L$, using negatives in the same language as the positive ($P_L = N_L$) shows the best performance overall. For instance, with mE5-large, $\mathcal{D}(ko, en, en)$ outperforms $\mathcal{D}(ko, en, ko)$ on OVR (84.56 vs. 78.00), and with mgte-base, $\mathcal{D}(en, ko, ko)$ exceeds $\mathcal{D}(en, ko, en)$ (80.81 vs. 78.31). These findings suggest that if $Q_L \neq P_L$, using $P_L = N_L$ makes the model learn to distinguish fine-grained semantics in the target language from contrastive learning, thereby boosting CLIR capability. By contrast, if $Q_L \neq P_L$ and $P_L \neq N_L$ (e.g., $\mathcal{D}(ko, en, ko)$, $\mathcal{D}(en, ko, en)$), it does acquire some semantic understanding, but it simultaneously learns to prioritize linguistic difference over semantic relevance. This is because the contrastive loss trains the model to pull the positive sample (P_L) (which is in a different language) closer to the query (Q_L), while pushing the negative sample (N_L) (which is in the same language as the query). This enhances the model’s sensitivity to language differences but reduces its discernment of subtle target language semantics, thus lowering CLIR performance.

5.2 Mono-IR Retrieval Results

The right side of Table 1 reports the Mono-IR performance for models trained on all data combinations. In contrast to the CLIR results, Mono-IR performance generally decreases after training. A notable observation is the large Mono-IR drop when training with $Q_L = N_L$ and $Q_L \neq P_L$ (e.g., $\mathcal{D}(ko, en, ko)$ and $\mathcal{D}(en, ko, en)$). For instance, considering the OVR score for mE5-large, training on $\mathcal{D}(ko, en, ko)$ (83.61) and $\mathcal{D}(en, ko, en)$ (85.12) leads to substantial degradation compared to the base model (88.89). In contrast, models trained on $\mathcal{D}(ko, en, en)$ or $\mathcal{D}(en, ko, ko)$, where $Q_L \neq P_L$ and $P_L = N_L$, experience only minor decreases, or even show improvements in the case of mgte-base. The larger degradation under $Q_L = N_L$ and $Q_L \neq P_L$ can be explained by the language configuration of the negatives. In these

settings, each query is paired with a cross-lingual positive, while its hard negatives are written in the same language as the query. As a result, the contrastive objective pulls the cross-lingual positive toward the query while repeatedly pushing away same-language negatives. This can disrupt the same-language semantic neighborhoods that are important for Mono-IR, leading to a larger performance drop. In other words, the model may learn a language-level separation signal that competes with semantic alignment.

5.3 Retrieval Results for Merged Models

Our preceding experiments reveal that optimizing for CLIR performance can lead to a degradation in Mono-IR capabilities, creating a performance trade-off. To address this issue, we incorporate a model merging strategy. Inspired by Model Soup (Wortsman et al., 2022), which is used to resolve task-specific conflicts (Lee et al., 2025; Zhang et al., 2025), we investigate whether simple weight-averaged model merging (Matena and Raffel, 2022) can mitigate such trade-offs. We merge the base and fine-tuned parameters with a fixed equal ratio, $\theta_{\text{merged}} = 0.5\theta_{\text{base}} + 0.5\theta_{\text{ft}}$. Since the data combinations that improved CLIR performance for the mgte-base model did not degrade its Mono-IR performance, we only use the remaining three models in this merging experiment.

MODEL	CROSS-AVG		MONO-AVG	
	en-ko	ko-en	ko-ko	en-en
<i>bge-m3</i>				
base	85.81	80.01	86.29	89.99
└ koenen FT	86.58	81.26	85.89	89.55
└ Merged (base+ koenen FT)	86.96	81.21	86.49	90.08
└ enkoko FT	86.53	80.87	85.80	89.54
└ Merged (base + enkoko FT)	86.85	80.99	86.51	90.15
<i>multilingual-e5-large (mE5-large)</i>				
base	87.04	76.88	87.43	90.35
└ koenen FT	87.99	81.13	86.01	89.63
└ Merged (base + koenen FT)	88.81	81.73	87.68	90.52
└ enkoko FT	88.04	80.44	85.76	90.04
└ Merged (base + enkoko FT)	89.20	81.23	87.17	90.84
<i>multilingual-e5-base (mE5-base)</i>				
Base	80.11	65.98	84.62	89.50
└ koenen FT	84.54	73.26	82.16	88.91
└ Merged (base + koenen FT)	84.93	72.71	84.20	89.82
└ enkoko FT	85.02	72.17	80.93	89.35
└ Merged (base + enkoko FT)	85.66	72.16	83.70	90.00

Table 2: Average performance on CLIR and Mono-IR tasks including merged models.

Merging Effects on Mono-IR Table 2 compares the IR performance of the base models, models fine-tuned on $\mathcal{D}(koenen)$ and $\mathcal{D}(enkoko)$ (strong CLIR performers), and the merged models. Crucially, the merged models consistently improve Mono-IR performance across both tasks compared to the individually fine-tuned CLIR-optimized models. For example, merging the mE5-large model trained on $\mathcal{D}(koenen)$ with the base model increases Mono-IR score by approximately 1 point on both tasks, in some cases even surpassing the original base model’s performance.

Merging Effects on CLIR Encouragingly, model merging does not lead to a substantial decline in CLIR performance. For mE5-base, the merged models maintain CLIR performance comparable to the individually fine-tuned models. Moreover, bge-m3 and mE5-large even show slight CLIR improvements after merging.

These findings indicate that weight-averaged model merging with the base model provides a simple and practical way to mitigate the trade-off between CLIR gains and Mono-IR degradation. Moreover, our results provide empirical evidence that model merging can, in some instances, further enhance CLIR performance beyond what is achieved by individually trained specialist models.

6 Conclusion

In this study, we conduct an in-depth analysis of how training data language composition impacts CLIR and Mono-IR in multilingual embedding models. We generated linguistically parallel training data and trained models with various language combinations. Through our experiments, we revealed a consistent trade-off: while aligning the languages of positives and negatives substantially enhances CLIR performance, this configuration can concurrently degrade Mono-IR capabilities. To address this trade-off, we demonstrate that simple weight-averaged model merging can partially mitigate Mono-IR degradation while preserving most CLIR gains. Our findings provide valuable insights for future data design and training for robust Cross-Lingual and Mono-Lingual Information Retrieval, and further suggest a methodology applicable to other language pairs.

Limitations

This study is subject to several limitations that open avenues for future research.

First, our investigation primarily focuses on Korean-English CLIR and Mono-IR. Consequently, the generalizability of our findings and the observed effectiveness of the model merging strategy to other language pairs, particularly those with typological or cultural differences, require further validation.

Second, the synthetic training data, despite efforts to ensure diversity using LLMs and the Korean Decimal Classification, may not fully capture the nuances and breadth of real-world user-generated query-document pairs, especially in highly specialized or niche domains.

Third, we employ a relatively straightforward Weight-Averaged Model Merging approach. Exploring more sophisticated or task-specific merging methodologies, such as those involving differential layer weighting or the use of Fisher information, might yield different outcomes or superior performance improvements.

Finally, the observed trade-offs and merging effects are contingent on the specific evaluation datasets (Belebele and StrategyQA) used. Different datasets or additional evaluation metrics could reveal different patterns or magnitudes of these effects.

Ethics Statement

Synthetic training data in this study was generated via Large Language Models (LLMs) strictly for research on Cross-Lingual and Mono-Lingual Information Retrieval, guided by the Korean Decimal Classification (KDC) to ensure diverse and neutral content. We utilized publicly available models and datasets, promoting transparency and reproducibility. Throughout the research process, including drafting and refining textual content for this paper, an AI assistant was utilized; however, all core research ideas, experimental design, analysis, and final content were directed and validated by the human authors. While acknowledging potential LLM biases, our focus was on the structural properties of synthetic data for training retrieval models, not direct user-facing content generation. Our goal is to improve multilingual information access. The findings aim to contribute to robust and balanced retrieval systems, and we advocate for responsible AI development considering societal impacts.

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 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 work was supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. RS-2022-II220369, (Part 4) Development of AI Technology to support Expert Decision-making that can Explain the Reasons/Grounds for Judgment Results based on Expert Knowledge)

References

- Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *arXiv preprint arXiv:2308.16884*.
- Paheli Bhattacharya, Pawan Goyal, and Sudeshna Sarkar. 2016. Query translation for cross-language information retrieval using multilingual word clusters. In *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016)*, pages 152–162.
- Andrei Z Broder. 1997. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, pages 21–29. IEEE.
- 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.
- John P Comaromi. 1984. Dewey decimal classification. *New Trends in International Librarianship: SR Ranganathan Festschrift to Mark the Platinum Jubilee of the Madras University Library*, page 163.
- Zhuyun Dai and Jamie Callan. 2020. Context-aware term weighting for first stage passage retrieval. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 1533–1536.
- Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Sibli, Dominik Krzemiński, and Genta Indra Winata. 2025. Mmteb: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*.
- Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021a. Splade v2: Sparse lexical and expansion model for information retrieval. *arXiv preprint arXiv:2109.10086*.
- Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021b. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2292.
- Jianfeng Gao, Jian-Yun Nie, Endong Xun, Jian Zhang, Ming Zhou, and Changning Huang. 2001. Improving query translation for cross-language information retrieval using statistical models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 96–104.
- Luyu Gao, Yunyi Zhang, Jiawei Han, and Jamie Callan. 2021. Scaling deep contrastive learning batch size under memory limited setup. *arXiv preprint arXiv:2101.06983*.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2:1.
- Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.
- Koustava Goswami, Sourav Dutta, Haytham Assem, Theodorus Franssen, and John Philip McCrae. 2021. Cross-lingual sentence embedding using multi-task learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9099–9113.
- Zhiqi Huang, Puxuan Yu, and James Allan. 2023. 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.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781.

- Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.
- Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, and Henrique Schechter Vera. 2025. Gemini embedding: Generalizable embeddings from gemini. *arXiv preprint arXiv:2503.07891*.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, and Tim Rocktäschel. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.
- Xiaoxi Li, Jiajie Jin, Yujia Zhou, Yuyao Zhang, Peitian Zhang, Yutao Zhu, and Zhicheng Dou. 2025. From matching to generation: A survey on generative information retrieval. *ACM Transactions on Information Systems*, 43(3):1–62.
- Robert Litschko, Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. 2022. On cross-lingual retrieval with multilingual text encoders. *Information Retrieval Journal*, 25(2):149–183.
- Qi Liu and Jiaxin Mao. 2023. Understanding the multi-vector dense retrieval models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 4110–4114.
- Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Michael S Matena and Colin A Raffel. 2022. Merging models with fisher-weighted averaging. *Advances in Neural Information Processing Systems*, 35:17703–17716.
- Gabriel de Souza P Moreira, 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*.
- Toan Ngoc Nguyen, Nam Le Hai, Nguyen Doan Hieu, Dai An Nguyen, Linh Ngo Van, Thien Huu Nguyen, and Sang Dinh. 2025. Improving vietnamese-english cross-lingual retrieval for legal and general domains. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 142–153.
- Dong-Geun Oh. 2020. Korean decimal classification (kdc). *ISKO Encyclopedia of Knowledge Organization*.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.
- Shadi Saleh and Pavel Pecina. 2020. Document translation vs. query translation for cross-lingual information retrieval in the medical domain. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6849–6860.
- Aditi Singh, Abul Ehtesham, Saket Kumar, and Tala Talaie Khoei. 2025. Agentic retrieval-augmented generation: A survey on agentic rag. *arXiv preprint arXiv:2501.09136*.
- Amit Singhal. 2001. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43.
- Aivin V Solatorio. 2024. Gistembed: Guided in-sample selection of training negatives for text embedding fine-tuning. *arXiv preprint arXiv:2402.16829*.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.
- Yining Wang, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. 2013. A theoretical analysis of ndcg type ranking measures. In *Conference on learning theory*, pages 25–54. PMLR.
- Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, and Simon Kornblith. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. PMLR.
- Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*.
- Puxuan Yu, Hongliang Fei, and Ping Li. 2021. Cross-lingual language model pretraining for retrieval. In *Proceedings of the Web Conference 2021*, pages 1029–1039.
- Fuwei Zhang, Zhao Zhang, Xiang Ao, Dehong Gao, Fuzhen Zhuang, Yi Wei, and Qing He. 2022. Mind the gap: Cross-lingual information retrieval with hierarchical knowledge enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4345–4353.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, and Fei Huang. 2024. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. *arXiv preprint arXiv:2407.19669*.

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

Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2024. Dense text retrieval based on pretrained language models: A survey. *ACM Transactions on Information Systems*, 42(4):1–60.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*.

A Details for Synthetic Data Generation

A.1 Prompt Examples

We provide representative prompts used in our synthetic data generation pipeline. The pipeline consists of two stages: (1) task brainstorming, which produces diverse retrieval task descriptions conditioned on a KDC topic, and (2) triple generation, which creates bilingual {query, positive, hard_negative} examples for each task. To support scalable filtering and automatic parsing, we enforce strict output formats: Prompt 1 outputs a Python list of strings, and Prompt 2 outputs a single JSON object.

Prompt Variables The prompts are parameterized with placeholders to control diversity and ensure coverage. In Table 3 (Prompt 1), {GIVEN_TOPIC} denotes the sampled KDC topic and {NUM_TASKS} controls the number of task descriptions. In Table 4 (Prompt 2), {TASK_DESCRIPTION} is a single task description sampled from Prompt 1 outputs. We further vary query and document attributes via the following placeholders:

- {query_type} ∈ {extremely long-tail, long-tail, common}
- {query_length} ∈ {less than 5 words, 5 to 15 words, at least 15 words}
- {difficulty} ∈ {high school, college, PhD}
- {clarity} ∈ {clear, understandable with some effort, ambiguous}
- {num_words} ∈ {50, 100, 200, 300, 400, 500}

Prompt 1: Task Brainstorming Prompt 1 (Table 3) generates a list of retrieval task descriptions in Korean. Each task specifies what the query represents and what constitutes a relevant document, while discouraging overly narrow tasks to improve coverage.

Prompt 2: Triple Generation Prompt 2 (Table 4) generates one training instance in a strict JSON format, including Korean and English versions of each component. A hard negative is designed to be topically plausible yet incorrect with respect to the task requirement, encouraging the retriever to learn fine-grained relevance beyond superficial overlap.

```
Brainstorm a list of potentially useful text retrieval tasks related to the given topic. Here are a few examples for your reference:  
- Provided a scientific claim as query, retrieve documents that help verify or refute the claim.  
- Search for documents that answers a FAQ-style query on children's nutrition.  
Please adhere to the following guidelines:  
- Specify what the query is, and what the desired documents are.  
- Each retrieval task should cover a wide range of queries, and should not be too specific.  
Given topic: {GIVEN_TOPIC}  
Your output should always be a python list of strings only, with about {NUM_TASKS} elements, and each element corresponds to a distinct retrieval task in one sentence written in Korean related to the given topic. Do not explain yourself or output anything else. Be creative!
```

Table 3: Prompt 1: Task Brainstorming

```
You have been assigned a retrieval task: {TASK_DESCRIPTION}  
Your mission is to write one text retrieval example for this task in JSON format.  
The JSON object must contain the following keys:  
- "user_query (kor)": a string, a random user search query specified by the retrieval task.  
- "user_query (eng)": a string, an English version of "user_query (kor)"  
- "positive_document (kor)": a string, a relevant document for the user query.  
- "positive_document (eng)": a string, an English version of "positive_document (kor)"  
- "hard_negative_document (kor)": a string, a hard negative document that only appears relevant to the query.  
- "hard_negative_document (eng)": a string, an English version of "hard_negative_document (kor)"  
Please adhere to the following guidelines:  
- The "user_query" should be {query_type}, {query_length}, {clarity}, and diverse in topic.  
- All documents should be at least {num_words} words long.  
Your output must always be a JSON object only, do not explain yourself or output anything else. Be creative!
```

Table 4: Prompt 2: Triple Generation

Post-processing After generation, we parse the model outputs and apply lightweight filtering. First, we perform format validation: any sample that cannot be parsed as a JSON object (or does not contain the required keys) is discarded. We then apply near-deduplication to reduce redundancy in the synthetic training set. Concretely, we deduplicate based on the English query field (user_query (eng)) using MinHash LSH (Broder, 1997) with a Jaccard-similarity threshold of 0.5 and 128 permutations. We tokenize each query using

the FacebookAI/xlm-roberta-large¹ tokenizer, build a MinHash signature over the resulting token set, and insert it into an LSH index. We retrieve candidate duplicates via LSH queries, and for each detected duplicate pair we remove the later-occurring instance. To avoid over-pruning across unrelated topics, we first run this MinHash deduplication within each KDC category and then run the same procedure once again globally over the merged set to eliminate remaining cross-category duplicates. The number of instances in the deduplicated dataset is 33k.

A.2 Details for Hard Negative Mining

In this section, we describe our hard negative mining procedure for constructing an effective training dataset based on the generated synthetic data.

Our synthetic triples are generated to be fully parallel across languages, i.e., each instance provides aligned query/positive/negative texts in both English and Korean that correspond to the same underlying semantics. This one-to-one parallel structure allows us to mine hard negatives in one language and reuse them for other language variants without changing the intended negative meaning. Accordingly, we perform hard negative mining on the English side and transfer the mined negatives to other variants via their parallel counterparts.

Concretely, we precompute embeddings for all English positive documents (`positive_document(eng)`) using NV-Embed-v2 (Moreira et al., 2024), and mine hard negatives for each English query based on these embeddings. We intentionally use NV-Embed-v2 (instead of a multilingual retriever) because hard negative mining benefits from a model that is strong at fine-grained semantic discrimination in the mining language; in our experience, a high-performing English retriever yields sharper similarity rankings and thus more challenging negatives. For each query, we rank documents by similarity, ignore the top-50 retrieved candidates to reduce potential false negatives, and then sample 5 English hard negatives per query from the remaining candidates. Finally, by adding the originally generated synthetic hard negative, we obtain a total of 6 hard negatives for each query. For variants involving Korean queries or Korean negatives, we directly use the Korean counterparts of these mined negatives provided by the parallel synthetic generation, so that language variants share the same

underlying negative semantics while differing only in language.

A.3 Final Training Dataset

Finally, we instantiate language-composed training data by selecting either the Korean or English translation for each of {query, positive, hard_negative}, resulting in six Korean–English cross-lingual variants that cover all language configurations.

# of Data	Variant	Query	Positive	Negatives(6)
33k	1	Kor	Kor	Eng
	2	Kor	Eng	Kor
	3	Kor	Eng	Eng
	4	Eng	Eng	Kor
	5	Eng	Kor	Eng
	6	Eng	Kor	Kor

Table 5: Six Korean–English cross-lingual variants used in training (excluding fully monolingual cases).

We provide an example of one final training data composition, $\mathcal{D}(ko, en, en)$, in Table 7. The example illustrates a cross-lingual training triple where an “Anchor” query in Korean is paired with a “positive” document in English that directly answers the query. The “negative” documents are topically related to the anchor (e.g., French learning, grammar, or dialects for a query about French pronunciation rules), yet they do not satisfy the information need, making them challenging distractors during training. This structure is designed to train a system for cross-lingual information retrieval, emphasizing semantic understanding.

A.4 Human Quality Assessment

To assess the quality of our synthetic training dataset, we conduct a manual human evaluation on a random sample of generated instances. We sample 100 training instances in total, consisting of 50 English-only instances ($\mathcal{D}(en, en, en)$) and 50 Korean-only instances ($\mathcal{D}(ko, ko, ko)$). Four expert annotators, two native English speakers and two native Korean speakers, evaluate each sample using a 3-point scale (0–2); for each language group, one annotator is an M.S. student and the other is a Ph.D. student in Computer Science. The criteria are as follows:

- Grammatical Fluency (0–2): whether the query is natural and clear (2 = native-level, 0 = unintelligible).
- Positive Relevance (0–2): whether the positive document contains an answer to the query (2 = explicit answer, 0 = irrelevant).

¹FacebookAI/xlm-roberta-large

- Hard Negative Validity (0–2): whether the hard negative is topically related but off-target (2 = valid hard negative, 0 = false negative / actually correct).

Table 6 summarizes the results (mean \pm std), where we additionally report the percentage of the maximum possible score in parentheses. Overall, the evaluation yields consistently high scores across all criteria. In particular, positive documents are almost always judged to directly answer the query (approximately 99% of the maximum score), and hard negatives are largely valid while remaining non-answering distractors (approximately 80–84% of the maximum score). Despite the inherent subjectivity of linguistic judgments, these results indicate strong annotator agreement on the quality of the generated corpus, supporting its suitability for retrieval model fine-tuning.

Metric	$\mathcal{D}(en, en, en)$	$\mathcal{D}(ko, ko, ko)$
Fluency	1.82 ± 0.39 (91%)	1.82 ± 0.39 (91%)
Pos. Rel.	1.98 ± 0.14 (99%)	1.97 ± 0.14 (99%)
HN Validity	1.68 ± 0.51 (84%)	1.57 ± 0.54 (80%)

Table 6: Human quality assessment of the synthetic training corpus (mean \pm std over a 0–2 scale). Percentages denote the ratio to the maximum possible score.

A.5 Experimental Setup Details

All models were trained for a total of one epoch with a batch size of 512, employing a linear learning rate scheduler with a warm-up ratio of 0.1. We used the AdamW optimizer (Loshchilov and Hutter, 2017) (with parameters $\beta_1 = 0.9$, $\beta_2 = 0.99$, and weight decay=0.01), and adopted bfloat16 mixed precision for computational efficiency. The learning rate for all models is fixed to $2e-5$.

We used 2 NVIDIA A100 GPUs, each with 80GB of memory capacity, along with AMD EPYC 7513 processors featuring 32 cores, to train models. Additionally, to increase the batch size within limited GPU resources, we adopt CachedGISTEmbedLoss from sentence-transformers (Reimers and Gurevych, 2019), which combines the GISTEmbedLoss with the gradient cache technique (Gao et al., 2021). We use bge-m3 as a guide model for the GISTEmbedLoss. For evaluation, we employed a single A100 GPU.

Type	Content
Anchor	프랑스어의 발음 규칙은 어떻게 되나요? (<i>What are the pronunciation rules in French?</i>)
Positive	The pronunciation rules of French are relatively clear and systematic. The French alphabet consists of 26 letters, each with its own unique pronunciation rules. For instance, 'e' is often silent when it appears at the end of a word or is unstressed. Additionally, 'h' is generally not pronounced, and 'c' is pronounced as 's' when followed by 'e', 'i', or 'y'. These rules help French learners to understand pronunciation more easily.
Negative 1	French is a Romance language primarily spoken in France, Belgium, Switzerland, and Canada. It is spoken by approximately 270 million people and is one of the six official languages of the United Nations. French originated from Latin and has evolved over time into modern French. The language is known for its distinct pronunciation, particularly the unique pronunciation of the letter 'r' compared to other languages. Learning French offers an opportunity to deepen understanding of the language's history and culture.
Negative 2	For practicing French pronunciation, it is important to focus on particularly difficult or commonly mispronounced sounds. Since pronunciation is very important in French, constant repetition and practice are necessary to get it right. The 'r' sound and nasal vowels are particularly challenging for many learners, and special exercises may be required for these.
Negative 3	French verb conjugation has a complex structure. Generally, verbs are divided into three main groups: regular -er verbs, -ir verbs, and irregular verbs. Each group follows different conjugation rules depending on various tenses such as present tense, past tense, and future tense. For instance, the present form of -er verbs changes according to the endings, and compound tenses are formed by using the appropriate form of 'avoir' or 'être' along with the past participle. In compound past tense, 'avoir' is mainly used, but for verbs of movement or state change, 'être' is used as an exception. These rules form the basis of French verb conjugation, and understanding the irregularities of each verb is crucial.
Negative 4	The most important aspect of starting to learn French is understanding the basic grammar rules. Nouns, adjectives, and verbs in French change form according to gender and number, and mastering these rules is essential. Additionally, French pronunciation is quite different from English, so practicing pronunciation is also necessary. To learn French, it is advisable to use a variety of learning materials. For example, reading French novels or news articles can be beneficial. Hiring an online tutor or participating in language exchange programs can also be a great way to practice French conversation. Finally, consistency is key in learning French, so it's important to continue with daily study, even if it's just a little each day.
Negative 5	French spelling often does not match its pronunciation, which can confuse learners. For instance, 'beaucoup' is pronounced as 'boku,' but its spelling does not suggest this. These discrepancies are related to the historical development of the French language. Additionally, there are various accents and pronunciations in French that can vary by region. Specifically, the French 'r' is different from the English 'r,' which can be challenging for learners to grasp initially. These mismatches between spelling and pronunciation are significant challenges to overcome when learning French.
Negative 6	Understanding the important rules of French grammar is fundamental to language learning. Key elements of French grammar include the agreement of gender and number in nouns and adjectives, verb tense and conjugation, pronoun placement, and the use of articles. In French, nouns and adjectives are divided into masculine and feminine forms, and their forms change depending on whether they are singular or plural. Verbs change according to the subject and tense, and it is particularly important to learn the conjugation of regular and irregular verbs. Furthermore, pronouns are used according to their position and role in a sentence. Lastly, articles change according to the gender and number of the noun, and understanding the difference between definite and indefinite articles is essential.

Table 7: Training Data Example of $\mathcal{D}(\textit{koenen})$

B Evaluation Benchmark

In this work, we leverage multilingual Question Answering (QA) datasets with parallel constructions, repurposed as retrieval tasks, to systematically assess CLIR 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 (query and its corresponding passage for answering) 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 Belebele² 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 both English and target language, 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.

StrategyQA StrategyQA is a crowdsourced dataset of diverse yes/no questions that often require implicit multi-step reasoning. Ko-StrategyQA³ was created by translating the original StrategyQA questions and evidence paragraphs using DeepL, and is included in the MMTEB benchmark. In our experiments, we additionally use the original English StrategyQA as a paired counterpart of Ko-StrategyQA to form an English–Korean parallel evaluation setting aligned by construction.

Beyond these, our evaluation was limited by the lack of other retrieval datasets that provide parallel Korean-English pairs.

C Robustness to Hard Negative Mining Strategies

We conduct additional experiments by adopting the hard negative mining strategies proposed in NV-Retriever (Moreira et al., 2024). This paper suggests that utilizing absolute 0.05 margin and relative 0.05 margin in the similarity score yields optimal results for hard negative mining. We follow

²https://github.com/embeddings-benchmark/mteb/blob/main/mteb/tasks/retrieval/multilingual/belebele_retrieval.py

³https://github.com/embeddings-benchmark/mteb/blob/main/mteb/tasks/retrieval/kor/ko_strategy_qa.py

this procedure and repeat the *bge-m3* experiments under both mining strategies. The results are provided in Table 8 and Table 9.

Results Although different mining strategies lead to different absolute performance levels, we observe that the qualitative trends reported in Section 5 remain consistent under the additional experiment. Specifically, the relative comparisons between training configurations are largely preserved, and similar trends are observed across both CLIR and Mono-IR settings.

<i>bge-m3 with Absolute 0.05 Margin Hard Negative Mining</i>														
Data	Cross-Lingual							Mono-Lingual						
	Belebele		StrategyQA		AVG		OVR	Belebele		StrategyQA		AVG		OVR
	en-ko	ko-en	en-ko	ko-en	en-ko	ko-en		ko-ko	en-en	ko-ko	en-en	ko-ko	en-en	
base	90.37	88.36	81.24	71.65	85.81	80.01	82.91	93.16	95.55	79.41	84.42	86.29	89.99	88.14
koenko	89.67	89.51	79.41	72.78	84.54	81.15	82.84	91.90	94.42	78.08	83.73	84.99	89.08	87.03
kokoen	90.68	87.18	82.18	70.65	86.43	78.92	82.67	93.16	95.44	79.66	84.40	86.41	89.92	88.16
koenen	91.49	89.10	81.71	72.73	86.60	80.92	83.76	92.42	94.78	79.36	84.09	85.89	89.44	87.66
enenko	89.65	88.60	80.41	71.64	85.03	80.12	82.58	93.07	95.18	79.32	84.53	86.20	89.86	88.03
enkoen	92.03	88.92	82.06	72.67	87.05	80.80	83.92	92.34	94.87	79.45	83.86	85.90	89.37	87.63
enkoko	91.59	89.38	81.95	72.73	<u>86.77</u>	<u>81.06</u>	<u>83.91</u>	92.56	95.13	79.06	84.22	85.81	89.68	87.74

Table 8: CLIR and Mono-Lingual IR performance for bge-m3 with Absolute 0.05 Margin Hard Negative Mining. Highest scores in **bold**, second highest underlined.

<i>bge-m3 with Relative 0.05 Margin Hard Negative Mining</i>														
Data	Cross-Lingual							Mono-Lingual						
	Belebele		StrategyQA		AVG		OVR	Belebele		StrategyQA		AVG		OVR
	en-ko	ko-en	en-ko	ko-en	en-ko	ko-en		ko-ko	en-en	ko-ko	en-en	ko-ko	en-en	
base	90.37	88.36	81.24	71.65	85.81	80.01	82.91	93.16	95.55	79.41	84.42	86.29	89.99	88.14
koenko	89.95	89.52	79.05	72.81	84.50	81.17	82.83	92.09	94.67	77.51	83.97	84.80	89.32	87.06
kokoen	90.78	87.15	82.09	70.75	86.44	78.95	82.69	93.01	95.41	79.68	84.53	86.35	89.97	88.16
koenen	91.45	89.21	81.63	72.76	86.54	80.99	<u>83.76</u>	92.57	94.74	79.59	84.18	86.08	89.46	87.77
enenko	89.82	88.80	80.93	71.28	85.38	80.04	82.71	93.40	95.29	79.43	84.80	<u>86.42</u>	90.05	88.23
enkoen	91.86	88.77	82.03	72.30	86.95	80.54	83.74	92.19	94.77	79.19	83.73	85.69	89.25	87.47
enkoko	91.64	89.61	81.77	72.77	86.71	81.19	83.95	92.74	94.95	79.28	84.12	86.01	89.54	87.77

Table 9: CLIR and Mono-Lingual IR performance for bge-m3 with Relative 0.05 Margin Hard Negative Mining. Highest scores in **bold**, second highest underlined.