McCrolin: Multi-consistency Cross-lingual Training for Retrieval Question Answering

Peerat Limkonchotiwat[♠*], **Wuttikorn Ponwitayarat**[♡*], **Lalita Lowphansirikul**[♡], **Potsawee Manakul**[♣], **Can Udomcharoenchaikit**[♡], **Ekapol Chuangsuwanich**[♢], **Sarana Nutanong**[♡]

[♠]AI Singapore, Singapore, [♣]SCB 10X, Thailand
[♡]School of Information Science and Technology, VISTEC, Thailand,
[♢]Department of Computer Engineering, Chulalongkorn University, Thailand
peerat@aisingapore.org, wuttikorn.p_s22@vistec.ac.th

Abstract

Automated question answering (QA) systems are increasingly relying on robust cross-lingual retrieval to identify and utilize information from multilingual sources, ensuring comprehensive and contextually accurate responses. Existing approaches often struggle with consistency across multiple languages and multi-size input scenarios. To address these challenges, we propose ***McCrolin***, a ***M***ulti-***c***onsistency ***Cro***ss-***lin***gual training framework, leveraging multi-task learning to enhance cross-lingual consistency, ranking stability, and input-size robustness. Experimental results demonstrate that McCrolin achieves state-of-the-art performance on standard cross-lingual retrieval QA datasets. Furthermore, McCrolin outperforms competitors when dealing with various input sizes on downstream tasks. In terms of generalizability, results from further analysis show that our method is effective for various encoder architectures and sizes. Codes and models are available at https://github.com/mrpeerat/McCrolin.

1 Introduction

Automated question answering is becoming more common thanks to the rapid advancement of large language models (LLMs). For reliability, these systems utilize a mechanism to integrate external knowledge into question answering, which, in turn, relies on robust retrieval capability. Jeong et al. (2024) show that the integration of external knowledge can substantially improve the performance of QA tasks. A modern QA system must contend with users with diverse language preferences while utilizing documents written in different languages. For example, a user might ask in English, *"What was the function of Surabaya pre-colonial?"* or in German, *"Was war die Funktion von Surabaya vor der Kolonialzeit?"* The system may retrieve relevant documents found in Indonesian historical records, as well as memoirs of missionaries and explorers written in Dutch and Portuguese from different perspectives. Robust cross-lingual retrieval is crucial to modern QA systems, as it provides the capability to identify and utilize information from diverse multilingual sources, forming the basis for comprehensive and contextually rich answers.

Cross-lingual representation learning is a common approach to obtaining the cross-lingual retrieval QA capability. In particular, one can employ bi-encoder learning, where a pre-trained language model (PLM) serves as the encoder for queries and documents. The training loss attempts to maximize the cosine similarity between the representation vectors of the query and its corresponding document (Karpukhin et al., 2020; Qu et al., 2021; Yang et al., 2021; Tasawong et al., 2023).

Despite the rapid advances in representation learning, previous works (Asai et al., 2021b; Limkonchotiwat et al., 2022b) have revealed that cross-lingual retrieval for QA requires further improvements in terms of consistency when dealing with multiple languages and input with different sizes. To address this limitation, existing methods utilize a larger PLM (330M parameters or over) and fine-tune it on extensive multilingual corpora (Asai et al., 2021b; Paranjape et al., 2022). While these approaches perform well in their training data distribution, they often lack robustness in handling unseen languages and out-of-domain scenarios.

In this paper, we propose a framework designed to address the challenges of language diversity and varying input sizes in cross-lingual retrieval QA, called ***M***ulti-***c***onsistency ***Cro***ss-***lin***gual training framework (***McCrolin***). Our approach leverages the reciprocal nature of multi-task learning, where the concurrent optimization of multiple interrelated objectives enhances the overall performance of the retrieval model. The crux of our proposed

[*]Equal contributions
[♠]Work was conducted while Peerat Limkonchotiwat was a PhD candidate at VISTEC

Findings of the Association for Computational Linguistics: EACL 2024, pages 2780–2793
November 12-16, 2024

method lies in multi-task learning objectives derived from the three properties vital to robust QA retrieval in a cross-lingual environment as follows:

- **Cross-lingual Consistency**: To handle multiple languages simultaneously, the method should produce an embedding space that is semantically consistent across languages.
- **Rank Stability**: For accurate retrieval, the method should minimize the distances between query-answer pairs and maximize the distances between queries and non-answers. In a cross-lingual environment, rank stability also extends to consistent ranking across multiple languages.
- **Input-Size Robustness**: To handle inputs of different lengths, the method should provide semantically consistent embeddings regardless of the input sizes across multiple languages. In other words, an original passage in English and a summary in another language should yield similar representations.

To achieve the three desired properties, we formulate three learning objectives; each corresponds to one property. First, for **cross-lingual language consistency** (§3.3.1), we introduce a training objective that enforces consistency across multilingual representations by transferring knowledge from English to a broad range of languages. Second, for **learning-to-rank** (§3.3.2), we propose a learning-to-rank loss function that transfers the English-to-English ranking capability to a broad range of language pairs, enhancing the consistency of multi-candidate relevance ranking in cross-lingual environments. Third, for **hierarchical alignment** (§3.3.3), we apply the mechanism presented in the first objective to a sub-unit, e.g., at the paragraph level, to extend the cross-lingual consistency concept. This objective enforces cross-lingual retrieval consistency at the paragraph level, improving the model's performance when retrieving paragraphs and sentences.

To evaluate the efficacy of our proposed framework, McCrolin, we conducted extensive experiments comparing it with 7 models across 5 experimental setups. The experimental results from cross-lingual retrieval QA demonstrate that McCrolin outperforms competitors in the average score case. When evaluating multi-candidate retrievals, our framework achieves state-of-the-art performance in Mean Reciprocal Rank at 10 (MRR@10). Moreover, our framework enables the base model to effectively handle multi-input levels, encompassing sentence and paragraph texts, improving retrieval performance at all levels.

The main contributions of this work include:

- We identify three desired properties for cross-lingual retrieval QA that are absent in existing approaches. To address this gap, we adopt a multi-task learning framework designed to integrate these properties.
- We propose a novel training pipeline that incorporates three loss components working reciprocally to enhance the cross-lingual retrieval capability, multi-candidate retrieval performance, and input-size robustness.
- We demonstrate that our framework achieves the SOTA performance in cross-lingual retrieval QA. Experimental results confirm the reciprocal nature of multi-task learning, i.e., the combination of three components provides the best results in all tasks. Further analysis demonstrates the generalizability of our framework regarding the encoder architectures and sizes.

2 Related Work

2.1 Sentence Embedding in Retrieval QA

QA representation learning aims to align query-answer pairs to each other. One widely adopted approach uses a similarity function as a learning objective. Karpukhin et al. (2020) proposed DPR, a query-answer representation alignment of two identical networks with N-pair loss, including the query and passage networks. Although DPR has demonstrated effectiveness and inference time efficiency in various datasets, their performance in multi- and cross-lingual retrievals needs further improvement. Asai et al. (2021b) proposed CORA, an end-to-end cross-lingual retrieval and reader QA framework using multilingual DPR (mDPR) and multilingual T5 (mT5) (Xue et al., 2021), respectively. However, CORA requires extensive multilingual training data, and their performance on unsupported languages (i.e., the language excluded from the training data) is inconsistent. Moreover, there are many possible answer levels in cross-lingual retrieval QA tasks (e.g., sentence, paragraph, and document levels). We found that existing works require one model for one input level and omit the multiple-level inputs from their works.

Recently, contrastive learning has become a popular approach for training sentence embedding models and achieving better ranking results. Given a tuple (anchor, positive, and negative), contrastive learning aims to maximize the similarity between

the representations of anchor and positive while minimizing the similarity between the representations of anchor and negative, which helps improve the ranking performance. Common techniques to generate positive and negative samples are data augmentation (Gao et al., 2021; Zhang et al., 2023; Xu et al., 2023; Zhao et al., 2024) and utilizing pairwise label datasets (i.e., natural language inference datasets) (Gao et al., 2021; Liu et al., 2022; Wang et al., 2022; Chen et al., 2024). However, recent research has demonstrated that when false negative or positive samples exist in the training data, contrastive learning fails to generate a meaningful representation (Zhou et al., 2022; Chanchani and Huang, 2023).

2.2 Language Knowledge Transfer (LKT)

Well-known sentence representation methods such as mUSE (Yang et al., 2020), LaBSE (Feng et al., 2022), and mE5 (Wang et al., 2024) have demonstrated the benefits of multilingual training. This technique is based on the fact that English accounts for approximately fifty percent of all multilingual training data, and other languages with less data can benefit from joint training with English. Thus, researchers have proposed a technique known as Language Knowledge Transfer (LKT), which entails transferring knowledge from the dominant language (English) to non-dominant languages using the same encoder (Lin et al., 2019; Nooralahzadeh et al., 2020; Wang et al., 2022; Limkonchotiwat et al., 2022b,a; Lin et al., 2023; Limkonchotiwat et al., 2023). The LKT principle leverages the superior performance of the dominant language to enhance the performance of other languages by allowing all languages to share the same encoder and updating the encoder weights only with English training data. These techniques achieve cross-lingual knowledge transfer without explicitly defining a transfer objective. Furthermore, previous works (Asai et al., 2021b; Limkonchotiwat et al., 2022b) discovered that, although the LKT is effective for retrieval, this concept did not improve the ranking score, particularly when retrieving multiple candidates (e.g., top-5 and top-10 retrievals).

3 McCrolin

3.1 Overview

In this work, we propose a multi-task learning framework to improve the accuracy of cross-lingual retrieval based on the following desired properties: cross-lingual consistency, rank stability, and input-size robustness. As shown in Figure 1, our proposed training process is built upon the teacher-student knowledge transfer concept. To achieve the desired properties, our loss function consists of three loss components with *cross-lingual consistency (Loss Component 1)* as the backbone, ensuring representation consistency across a broad range of languages. For *Loss Component 2*, the cross-lingual consistency concept is extended to contrastive learning to mimic the ranking capability from English (as the dominant language) to any language. Similarly, *Loss Component 3* extends the cross-lingual consistency concept to align paragraph representations, which is a document's subunit, with the query in multiple languages.

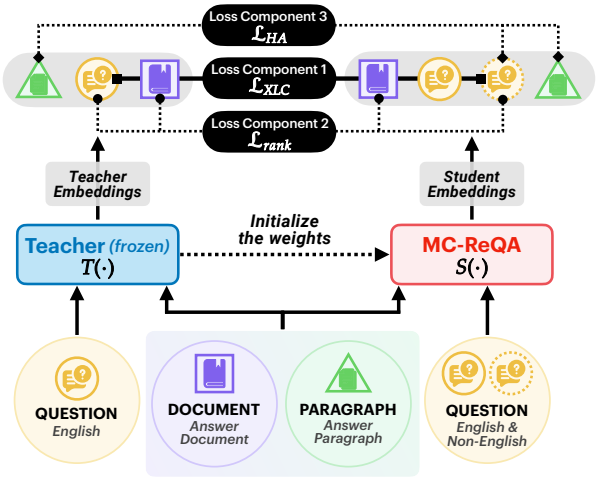

Figure 1: The teacher-student knowledge transfer process of ***M**ulti-**c**onsistency **Cro**ss-**ling**ual training framework (***McCrolin***) comprises three loss components: (i) Cross-lingual language consistency loss $\mathcal{L}_{\text{XLC}}$ (§3.3.1), (ii) Ranking loss $\mathcal{L}_{\text{rank}}$ (§3.3.2), and (iii) Hierarchical alignment loss $\mathcal{L}_{\text{HA}}$ (§3.3.3).

3.2 Teacher-Student Setup

As demonstrated in Figure 1, we adopt the teacher-student knowledge transfer approach to implement a multi-task learning pipeline with the desired cross-lingual transfer properties mentioned in the introduction. We describe the input-output setup of the teacher-student manner as follows:

- *Model initialization*. We initialize the student weights from the frozen teacher model. Note that we use the teacher model from Limkonchotiwat et al. (2022b) ($\text{mUSE}_{\text{teacher}}$).
- *Inputs*. We categorize the inputs into two groups: English (dominant language) and other languages (non-dominant languages) texts. We use English as the dominant language because it

accounts for most of the training and fine-tuning data used in the teacher model (Section 2.2).

- *Teacher Inputs*. During the fine-tuning process, we use the teacher to generate reference embedding vectors in the dominant language (English)[1].
- *Student Inputs*. In contrast, we aim to create an embedding space for the student model that produces identical embeddings for the same text regardless of whether it is posted in the dominant or non-dominant language. To achieve this, we input English and non-English texts into the student model and obtain student embeddings.

3.3 Learning Objectives

As shown in Figure 1, our framework distinguishes itself from other methods by its multi-task loss function:

$$\mathcal{L}_{\text{McCrolin}} = \gamma_1 \mathcal{L}_{\text{XLC}} + \gamma_2 \mathcal{L}_{\text{rank}} + \gamma_3 \mathcal{L}_{\text{HA}} \quad (1)$$

3.3.1 $\mathcal{L}_{\text{XLC}}$ – Cross-lingual Consistency Loss

As stated in Section 2.2, utilizing the same encoder for multiple languages at the same time, the cross-lingual transfer process ensures representation consistency across multiple languages by transferring the representation knowledge from the dominant language to others. While cross-lingual representation transfer objectives are useful, in this investigation, we discover that this process can be enhanced by ensuring the intra-lingual representation consistency between the teacher and student encoders. Consequently, we propose a new learning objective to minimize the teacher-student representation discrepancy for both intra- and cross-lingual settings. Given q^{en} is an English question, q^{ne} is a non-English question, $T()$ is a teacher model, $S()$ is a student model, and d is an answer document. The new cross-lingual language consistency loss $\mathcal{L}_{\text{XLC}}$ consists of four objectives:

$$\frac{1}{|M|}\sum_{i=1}^{M}[\beta_1||\underbrace{T(q_i^{\text{en}}) - S(q_i^{\text{ne}})}_{\text{XLC-Obj1}}||^2 + \beta_2||\underbrace{T(d_i) - S(d_i)}_{\text{XLC-Obj2}}||^2 + \beta_3||\underbrace{T(d_i) - S(q_i^{\text{ne}})}_{\text{XLC-Obj3}}||^2 + \beta_4||\underbrace{T(q_i^{\text{en}}) - S(q_i^{\text{en}})}_{\text{XLC-Obj4}}||^2], \quad (2)$$

where M is a mini-batch and β_1, β_2, β_3, and β_4 are the weight loss.

- *XLC-Obj 1: Query Representation Consistency.* The first objective is to transfer the knowledge from the same question expressed in English q^{en} to non-English q^{ne} using $T()$ and $S()$, respectively.
- *XLC-Obj 2: Document Representation Consistency.* This objective ensures that the document vectors produced by the student $S()$ are consistent with those of the teacher $T()$.
- *XLC-Obj 3: Retrieval Consistency.* For consistent cross-lingual QA pairing, we minimize the discrepancy between the student's question vector $S(q^{\text{ne}})$ and the teacher's document vector $T(d)$.
- *XLC-Obj 4: Intra-Lingual Consistency.* While we attempt to transfer knowledge from the teacher to the student model in a cross-lingual fashion, it is important to ensure the representation consistency for English, which is the dominant language. For this objective, we minimize the discrepancy between $T(q^{\text{en}})$ and $S(q^{\text{en}})$.

In contrast to cross-lingual retrieval works, we distill the knowledge from English to other languages while not omitting the English knowledge of the query samples. According to our ablation study (Table 1), using $\mathcal{L}_{\text{XLC}}$ to preserve the intra-lingual consistency is important and yields a substantial improvement.

3.3.2 $\mathcal{L}_{\text{rank}}$ – Ranking Loss

As discussed in the related work section, the cross-lingual transfer concept allows us to construct a multilingual embedding space by anchoring with the dominant language. However, previous works demonstrate poor downstream retrieval performance because they disregard the ranking aspect, especially in cross-lingual scenarios (Asai et al., 2021b; Wang et al., 2022).

We propose a mechanism to transfer knowledge from dominant to non-dominant languages while learning to rank simultaneously. As shown in Figure 1, $\mathcal{L}_{\text{rank}}$ uses contrastive learning CL$(\cdot)$ as the ranking objective to provide the distance contrast between positive and negative samples. For query-to-query cross-lingual transfer, we apply CL$(\cdot)$ to $T(q^{\text{en}})$ and $S(q^{\text{ne}})$ (the first term). For document-to-query cross-lingual transfer, we apply CL$(\cdot)$ to $T(d)$ and $S(q^{\text{ne}})$ (the second term). These two objectives work jointly to transfer the ranking capability from English to a broad range of target languages.

[1]Note that the questions in the training corpora are posted in non-English. As a result, they have to be translated. We use Google NMT for this purpose.

$$\mathcal{L}_{\text{rank}} = \lambda_1 \text{CL}(T(q^{\text{en}}), S(q^{\text{ne}})) + \lambda_2 \text{CL}(T(d), S(q^{\text{ne}})), \quad (3)$$

$$\text{CL}(a, p) = -\log \frac{e^{\text{sim}(\mathbf{a}_i, \mathbf{p}_i)/\tau}}{\sum_{j=1}^{M} e^{\text{sim}(\mathbf{a}_i, \mathbf{p}_j)/\tau}}, \quad (4)$$

where $\text{CL}(\cdot)$ is contrastive learning, $\text{sim}(\cdot)$ is a dot-product function, τ is a temperature scaling to extend the discrepancy of $T()$ and $S()$, and λ_1 and λ_2 are the loss weight. We use the anchor representations $\mathbf{a}$ from the teacher model, the positive representations $\mathbf{p}$ obtained from the student model, and the negative representations $\mathbf{n}$ are every sample in the mini-batch M except itself.

3.3.3 $\mathcal{L}_{\text{HA}}$ – Hierarchical Alignment Loss

In retrieval QA, a model should be able to handle input of different sizes, e.g., sentences, paragraphs, and documents. According to the discussion in Section 2.1, existing techniques perform well at document-level retrieval but perform poorly at the paragraph and sentence levels. However, retriever and machine reading QA models will require more computation when working on the document level since documents have more tokens to process. Thus, we aim to enable our model to better handle short text (sentence and paragraph levels) while maintaining the performance of long text (document level). Moreover, as demonstrated in previous works (Yang et al., 2020; Trijakwanich et al., 2021), adding the input-size robust property improves the generalization in retrieval tasks.

To achieve the input-size robustness, we change our model from being only document retrievers to three hierarchical retrieval units in a single model using the cross-lingual language consistency paradigm to force the representation of multi-level inputs to be the same. Given that PR is an answer paragraph, we simultaneously minimize the discrepancy between q and PR. This loss component leverages the backbone component, $\mathcal{L}_{\text{XLC}}$, by aligning PR to q^{ne} as shown in Figure 1. Therefore, our novel training objective $\mathcal{L}_{\text{HA}}$ is lying as follows:

$$\mathcal{L}_{\text{HA}} = \frac{1}{|M|} \sum_{i=1}^{M} [\omega_1 ||T(\text{PR}_i) - S(\text{PR}_i)||^2 + \omega_2 ||T(\text{PR}_i) - S(q_i^{\text{ne}})||^2] \quad (5)$$

where ω_1 and ω_2 are the loss weight.

3.4 Model Update

As discussed in Section 2.2, the cross-lingual transfer training objective enhances the performance of the student model, surpassing that of the teacher model. As a result, the student can function as the teacher in the next training round for iterative improvement. Experimental results indicate that omitting this teacher update step results in a significant performance penalty. Consequently, we adopt this iterative teacher update approach for our training pipeline consisting of three steps. First, we train the student model following the multi-task setup (Section 3.3). Then, at the end of the training process, we replace the teacher weights with the student weights. Finally, we repeat the process from the beginning until the performance of both models remains unchanged. For the result analysis, please refer to Appendix A.5.

4 Experimental Setup

4.1 Setting

Training Setup. We trained the student model with the Adam optimizer, a learning rate of 1e-3, and a batch size of 16 for 10 epochs. We used mUSE-small (#parameters: 68M) (Yang et al., 2020) as our primary encoder. We also explored our framework with other encoders such as mBERT, XLM-R, and E5 in Section 5.3. In addition, we set the number of teacher updates to three times. For hyperparameter settings, we used grid search to find the best parameter settings. The full hyperparameter configurations and sensitivity studies are given in Appendix A.1 and A.2, respectively.

Evaluation. We use the same evaluation setting as demonstrated in previous works (Yang et al., 2020; Karpukhin et al., 2020; Asai et al., 2021b; Limkonchotiwat et al., 2022b). We use recall@k as the main metric where we set k equal to 1 and 10. When applicable, we use the McNeMar test for hypothesis testing ($p < 0.05$). Note that we calculate the micro average score of all experiments from three random seeds.

4.2 Competitive Methods

- **mBERT- and XLM-R-mSimCSE**. We employed multilingual contrastive learning (Wang et al., 2022) to create a cross-lingual embedding space. We used questions as anchors and documents as positives and negatives.
- **DPR**. We employed a dense passage retrieval network (Karpukhin et al., 2020) where we used the learned weights from Asai et al. (2021b).
- **CORA**. We employed mT5 cross-lingual QA (Asai et al., 2021b). We used the same

weights without any fine-tuning for the XORQA dataset since this model was fine-tuned on this dataset. For other datasets, we fine-tuned CORA on them using the original weights.

- **LaBSE**. We employed a pre-trained sentence embedding based on mBERT (Feng et al., 2022) to cross-lingual retrieval QA. LaBSE fine-tuned on large-scale multilingual data and multiple sources of data and domains.
- **CL-ReLKT**. The mUSE-based encoder (Limkonchotiwat et al., 2022b) was fine-tined on the cross-lingual retrieval language knowledge transfer technique.
- **mE5-base**. We employed multilingual text embedding (Wang et al., 2024) as our baseline. The model was continually pre-trained from XLM-R and fine-tuned on multilingual retrieval datasets.

4.3 Benchmarks

We assess the effectiveness of our method and competitive methods on three main standard cross-lingual datasets for QA retrieval. For data statistics of each dataset, please refer to Appendix A.1.

XORQA (Asai et al., 2021a) is a cross-lingual open-retrieval question-answering dataset where questions are written in multiple languages and answers are written in English. The dataset contains 40k annotated samples across 7 languages.

XQuAD (Artetxe et al., 2020) is a cross-lingual question-answering dataset that evaluates cross-lingual robustness with questions and answers written in 11 languages. We employ the cross-lingual setting on this dataset. Questions are written in multiple languages, except for English, while the answers are written in English.

MLQA (Lewis et al., 2020) is a multi-way aligned extractive QA dataset containing multilingual questions and answers. The dataset consists of 7 languages. We used the same cross-lingual setting as XQuAD, with questions in multiple languages (excluding English) and answers in English.

5 Experimental Results

In this section, we present a series of experiments on cross-lingual retrieval question answering (Sections 5.1) and a downstream task, namely machine reading comprehension (Section 5.2). Additionally, we investigate the impact of our framework on different PLMs in Section 5.3.

5.1 Multi-task Training Objectives

In the first study, we perform a component-wise analysis of the McCrolin loss function consisting of the following components.

- **Cross-lingual Consistency Loss $\mathcal{L}_{\mathbf{XLC}}$**: The main loss component ensuring the consistency between question-answer pairs across multiple languages (Section 5.1.1).
- **Ranking Loss $\mathcal{L}_{\mathbf{rank}}$**: An auxiliary loss component improving the ability to rank question-answer pairs according to relevance (Section 5.1.2).
- **Hierarchical Alignment Loss $\mathcal{L}_{\mathbf{HA}}$**: An auxiliary loss component improving the hierarchical alignment between sentences and passages associated with the same document (Section 5.1.3).

5.1.1 Cross-lingual Consistency Loss $\mathcal{L}_{\mathbf{XLC}}$

This experiment assesses the cross-lingual consistency loss presented in Section 3.3. We compare our method to six competitors using three cross-lingual datasets for retrieval QA. For conciseness, the average scores across multiple languages are reported in Table 1 of this section, while the language-wise breakdowns are reported in Appendix A.4.

First, let us consider the component-wise results, i.e., the three rows started with "Only" in Table 1. We can see that just $\mathcal{L}_{\text{XLC}}$ alone outperforms CL-ReLKT in all cases. In addition, we can see that cross-lingual consistency $\mathcal{L}_{\text{XLC}}$ is the most crucial component, consistently outperforming the other two $\mathcal{L}_{\text{rank}}$ and $\mathcal{L}_{\text{HA}}$. This finding conforms with the design discussion presented in Section 3.1 showing $\mathcal{L}_{\text{XLC}}$ as the backbone component. The complete ablation analysis detailing all possible combinations of loss components and language-wise breakdowns is given in Appendix A.4.

Model	XORQA	XQuAD	MLQA	Avg.
mBERT-mSimCSE	44.8	71.3	25.7	53.2
XLM-R-mSimCSE	40.0	71.9	16.0	49.9
DPR	25.4	36.0	**48.2**	37.5
CORA	14.0	27.4	25.4	24.3
LaBSE	32.6	39.6	37.7	37.7
CL-ReLKT	50.9	75.1	47.0	62.4
Proposed model				
Only $\mathcal{L}_{\text{XLC}}$	51.1	75.9	47.3	63.0
Only $\mathcal{L}_{\text{rank}}$	48.1	72.6	45.7	60.2
Only $\mathcal{L}_{\text{HA}}$	46.3	58.0	44.0	51.8
McCrolin	**52.2**	**77.6**	47.8	**64.2**

Table 1: The average recall at 1 (R@1) on the cross-lingual retrieval QA task where the average score is micro averaging.

Second, the table also shows that McCrolin, which utilizes all three losses, achieves the highest average score, highlighting the reciprocal nature of our multi-task learning objectives. For example, McCrolin outperforms SOTA (CL-ReLKT) with statistical significance ($p < 0.05$) on XORQA, XQuAD, and MLQA by 1.3, 2.5, and 1.8 points, respectively. It is important to note that the performance of McCrolin on the MLQA dataset is lower than that of DPR. Our further analysis indicates irregularity of the mUSE tokenizer on HI texts, reporting a higher out-of-vocabulary (OOV) rate than in other languages. This is because HI was *not* present during the training process of mUSE-base, which is our base encoder. Nonetheless, on average, McCrolin achieves the best performance on 2 out of 3 datasets examined. Furthermore, we explore the robustness of our multi-task technique in out-of-domain scenarios in Appendix A.3.

5.1.2 Ranking Loss $\mathcal{L}_{\text{rank}}$

In this experiment, we evaluate the effectiveness of the ranking loss $\mathcal{L}_{\text{rank}}$ on the top-k retrieval task with $k = 10$. We use two measures: R@10 and MRR@10. The main results are reported in this section, while the additional results are provided in Table 12 in the appendix.

The results in Table 2 reveal three key findings: (i) The omission of the ranking loss ($\mathcal{L}_{\text{rank}}$) yields mixed results. We found that the retrieval performance of $\mathcal{L}_{\text{XLC}}$ and the CL-ReLKT method is similar for both metrics. This finding conforms with the discussion in Section 2.2 stating that only a cross-lingual consistency learning objective is inadequate to improve the ranking accuracy. (ii) The inclusion of $\mathcal{L}_{\text{rank}}$ (in addition to $\mathcal{L}_{\text{XLC}}$) improves the score in all cases, with a significant improvement on the average score ($p < 0.05$). (iii) Combining all losses yields SOTA results compared to competitive methods. The proposed method, which combines all losses, achieves the best performance on 5 out of 6 cases and obtains the highest average R@10 and MRR@10 scores. These findings underscore the importance of incorporating $\mathcal{L}_{\text{rank}}$ for improved ranking performance over single-task learning.

5.1.3 Hierarchical Alignment Loss $\mathcal{L}_{\text{HA}}$

This experiment focuses on assessing the hierarchical alignment loss $\mathcal{L}_{\text{HA}}$ by incorporating multiple-level inputs, specifically paragraph and sentence retrieval. In the paragraph-level setting, we utilize the provided paragraph (referred to as the "gold paragraph") available in the dataset. Meanwhile, for the sentence-level setting, we adopt the same methodology described in the work of Yang et al. (2020) for comparability. To make the experiment comparable, we fine-tune all competitive methods separately on sentence and paragraph texts.

Model	XORQA	XQuAD	MLQA	Avg.
R@10				
mBERT-mSimCSE	73.9	**92.7**	47.5	76.2
DPR	53.1	73.0	73.2	69.2
CL-ReLKT	76.6	92.0	67.8	82.1
$\mathcal{L}_{\text{XLC}}$	76.5	91.2	67.4	81.6
$\mathcal{L}_{\text{XLC}}$+$\mathcal{L}_{\text{rank}}$	77.5	92.4	68.1	82.6
McCrolin	**77.8**	92.5	**68.4**	**82.8**
MRR@10				
mBERT-mSimCSE	54.0	79.3	26.2	59.3
DPR	31.3	48.1	53.3	46.4
CL-ReLKT	57.3	79.8	53.9	68.1
$\mathcal{L}_{\text{XLC}}$	59.3	80.5	53.8	68.8
$\mathcal{L}_{\text{XLC}}$+$\mathcal{L}_{\text{rank}}$	59.5	81.9	54.1	69.7
McCrolin	**60.2**	**82.3**	**54.2**	**70.1**

Table 2: The average recall and MRR scores on the top-10 retrieval where the average score is micro averaging.

As shown in Table 3, the inclusion of $\mathcal{L}_{\text{HA}}$ on top of $\mathcal{L}_{\text{XLC}}$ results in a substantial performance increase for both sentence- and paragraph-level retrieval. In addition, the proposed method, McCrolin, which includes all three losses, demonstrates superior performance across both sentence and paragraph retrieval tasks, outperforming competitive methods in all cases. For instance, in sentence retrieval, McCrolin improves the performance of mBERT-mSimCSE from 22.0 to 35.3 on average. These results underscore the merits of the hierarchical alignment loss $\mathcal{L}_{\text{HA}}$ and the combination of the three learning objectives as a whole.

Model	XORQA	XQuAD	MLQA	Avg.
Sentence retrieval				
mBERT-mSimCSE	19.8	30.9	7.3	22.0
DPR	10.7	27.2	33.4	25.9
CL-ReLKT	22.2	18.4	33.4	23.4
$\mathcal{L}_{\text{XLC}}$	21.6	24.4	33.7	26.5
$\mathcal{L}_{\text{XLC}}$+$\mathcal{L}_{\text{HA}}$	22.3	36.4	34.2	33.1
McCrolin	**23.4**	**40.1**	**34.5**	**35.3**
Paragraph retrieval				
mBERT-mSimCSE	39.3	49.1	16.0	37.8
DPR	24.5	45.4	42.7	40.7
CL-ReLKT	47.9	40.1	42.7	42.3
$\mathcal{L}_{\text{XLC}}$	47.7	47.4	42.9	46.1
$\mathcal{L}_{\text{XLC}}$+$\mathcal{L}_{\text{HA}}$	48.1	59.7	43.5	52.9
McCrolin	**49.3**	**64.3**	**43.7**	**55.6**

Table 3: The average recall at 1 (R@1) on cross-lingual paragraph and sentence retrievals where the average score is micro averaging. For the full results, please refer to Table 13.

Model	XORQA				XQuAD				MLQA				Avg
	RU	KO	JA	FI	AR	DE	ZH	VI	AR	DE	ZH	VI	
mBERT (#parameters: 177M)													
mBERT-mSimCSE	52.1	**40.9**	**40.2**	46.0	77.3	83.9	80.7	71.9	11.0	31.1	38.5	25.6	49.9
+ $\mathcal{L}_{\text{McCrolin}}$	**53.0**	36.5	37.6	**48.5**	**82.8**	**88.2**	**84.9**	**74.5**	**32.7**	**56.2**	**41.7**	**46.6**	**56.2**
XLM-R (#parameters: 278M)													
XLM-R-mSimCSE	44.1	34.6	35.5	**46.0**	69.3	79.0	75.2	70.3	11.0	21.9	28.0	4.0	43.2
+ $\mathcal{L}_{\text{McCrolin}}$	**51.9**	**41.4**	**43.0**	45.0	**77.7**	**82.4**	**82.4**	**74.5**	**23.4**	**33.8**	**35.7**	**31.5**	**51.9**
mE5 (#parameters: 278M)													
mE5-base	73.6	53.6	59.1	71.2	70.2	78.2	67.2	72.3	49.9	70.9	**70.6**	59.9	66.4
+ $\mathcal{L}_{\text{McCrolin}}$	**75.6**	**67.7**	**69.2**	**76.3**	**85.7**	**89.9**	**86.1**	**85.3**	**51.5**	**73.2**	66.5	**66.5**	**74.5**

Table 4: Recall at 1 (R@1) on selected languages when changing base models from mUSE to PLMs.

5.2 Cross-lingual Machine Reading Comprehension

In the second study, we assess the effectiveness of McCrolin in cross-lingual machine reading comprehension with two different input levels: paragraph and document. We compare two retrieval models, CL-ReLKT and McCrolin, for the retrieval step and use GPT3.5-turbo as the reader model to extract the answer from the retrieved text. To keep the cost of invoking the GPT3.5-turbo API manageable, we conduct this study on a subset of languages presented in the XORQA, XQuAD, and MLQA benchmarks. Please refer to Table 14 for language-wise breakdowns.

As shown in Table 5, our method outperforms CL-ReLKT in all cases. In particular, McCrolin presents performance improvement over CL-ReLRT in paragraph and document retrievals from 27.1 and 39.2 to 38.0 and 39.9, respectively. In addition, we found an improvement in our method from 38.0 (paragraph-based) to 39.9 (document-based) F1 scores. However, using triple or quadruple amounts of tokens for the document-based, with only 1.9 points improvement compared to the paragraph-based, might not be worth the additional computational cost of using GPTs. This result emphasizes that achieving the input-size robustness property improves the downstream task's performance and saves money for the usage of LLMs.

Model	XORQA	XQuAD	MLQA	Avg.
Paragraph-based retrieval				
CL-ReLKT	31.3	24.1	26.7	27.1
McCrolin	**47.8**	**41.3**	**26.8**	**38.0**
Document-based retrieval				
CL-ReLKT	29.2	60.6	25.8	39.2
McCrolin	**30.1**	**61.4**	**26.4**	**39.9**

Table 5: The average F1 score on the cross-lingual MRC experiment using GPT3.5-Turbo as the reader.

5.3 McCrolin with Other Encoders

In this experiment, we examine the generalizability of McCrolin with different sentence encoders. Specifically, we continually fine-tuned the PLMs, namely mBERT-mSimCSE, XLM-R-mSimCSE, and mE5, using our proposed framework and evaluated their performance in selected languages.

As shown in Table 4, as expected, applying our framework to PLMs improved the cross-lingual capability for all models. For instance, the average performance of mBERT-mSimCSE increased from 49.9 to 56.2 points, while XLM-R-mSimCSE improved from 43.2 to 51.9 points. In addition, we observe an improvement in mE5 from 66.4 to 74.5. These findings also underscore the versatility of our framework, which can successfully improve the cross-lingual retrieval performance of various models. Moreover, we found that the performance of mE5 is higher than mUSE-based (Table 1). This is because this model was trained on massive training data, potentially leading to data contamination, e.g., mE5 used Mr.TyDi as the training data which is similar to XORQA. Therefore, we omit retrieval results from the main table and show them only in this experiment.

6 Conclusion

In this paper, we introduce ***McCrolin***, a ***M***ulti-***c***onsistency ***Cro***ss-***lin***gual training framework designed to address key challenges in cross-lingual QA retrieval, including cross-lingual consistency, ranking stability, and input-size robustness. We incorporate the desired properties into the base encoder using three loss components: cross-lingual language consistency $\mathcal{L}_{\text{XLC}}$, learning-to-rank $\mathcal{L}_{\text{rank}}$, and hierarchical alignment objectives $\mathcal{L}_{\text{HA}}$. Our component-wise analysis stresses the backbone status of $\mathcal{L}_{\text{XLC}}$, and the combination of all three loss components yields the best results. McCrolin's superior performance in handling various input sizes

on downstream tasks further underscores its robustness. Notably, analysis of various encoders demonstrates the generalizability of McCrolin across different encoder architectures and sizes, providing significant improvements when applied to various PLMs, such as mBERT, XLM-R, and mE5. This versatility enhances the cross-lingual consistency and robustness of QA systems, confirming the effectiveness of our approach.

Limitation

The experimental studies cover only general-domain standard benchmarks with pristine data quality. The out-of-domain study follows the standard practice of cross-benchmark evaluations, i.e., trained with one benchmark and evaluated by another. Although this study evaluates the models' capability to handle out-of-distribution inference, we did not experiment with in-the-wild data. Further studies should be conducted with application-oriented domain-specific data, such as those from medical chatbots and legal QA systems, to assess the models' behaviors in conditions closer to deployment.

Moreover, we found that one key limitation of our study is the out-of-vocabulary (OOV) issue, especially for the Hindi (HI) language in mUSE-based. The OOV rate for Hindi in XQuAD is 14.5%, while in MLQA, it is 34.4%. This is much higher than other languages, such as Greek (2.2% OOV rate) or German (1.3% OOV rate), which contributes to the observed performance gap between these languages.

References

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

Akari Asai, Jungo Kasai, Jonathan H. Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021a. XOR QA: Cross-lingual open-retrieval question answering. In *NAACL-HLT*.

Akari Asai, Xinyan Yu, Jungo Kasai, and Hannaneh Hajishirzi. 2021b. One question answering model for many languages with cross-lingual dense passage retrieval. *NeurIPS*.

Sachin Chanchani and Ruihong Huang. 2023. Composition-contrastive learning for sentence embeddings. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15836–15848, Toronto, Canada. Association for Computational Linguistics.

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

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. 2024. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. In *NAACL*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Association for Computational Linguistics*.

Peerat Limkonchotiwat, Weiwei Cheng, Christos Christodoulopoulos, Amir Saffari, and Jens Lehmann. 2023. mReFinED: An efficient end-to-end multilingual entity linking system. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15080–15089, Singapore. Association for Computational Linguistics.

Peerat Limkonchotiwat, Wuttikorn Ponwitayarat, Lalita Lowphansirikul, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, and Sarana Nutanong. 2022a. ConGen: Unsupervised control and generalization distillation for sentence representation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6467–6480, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Peerat Limkonchotiwat, Wuttikorn Ponwitayarat, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, and Sarana Nutanong. 2022b. CL-ReLKT: Cross-lingual language knowledge transfer for multilingual retrieval question answering. In *Findings of the Association for Computational Linguistics: NAACL 2022*,

pages 2141–2155. Association for Computational Linguistics.

Sheng-Chieh Lin, Amin Ahmad, and Jimmy Lin. 2023. mAggretriever: A simple yet effective approach to zero-shot multilingual dense retrieval. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11688–11696, Singapore. Association for Computational Linguistics.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Fangyu Liu, Yunlong Jiao, Jordan Massiah, Emine Yilmaz, and Serhii Havrylov. 2022. Trans-encoder: Unsupervised sentence-pair modelling through self- and mutual-distillations. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. 2020. Zero-shot cross-lingual transfer with meta learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Bhargavi Paranjape, Matthew Lamm, and Ian Tenney. 2022. Retrieval-guided counterfactual generation for QA. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1670–1686, Dublin, Ireland. Association for Computational Linguistics.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5835–5847, Online. Association for Computational Linguistics.

Panuthep Tasawong, Wuttikorn Ponwitayarat, Peerat Limkonchotiwat, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, and Sarana Nutanong. 2023. Typo-robust representation learning for dense retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1106–1115, Toronto, Canada. Association for Computational Linguistics.

Nattapol Trijakwanich, Peerat Limkonchotiwat, Raheem Sarwar, Wannaphong Phatthiyaphaibun, Ekapol Chuangsuwanich, and Sarana Nutanong. 2021. Robust fragment-based framework for cross-lingual sentence retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 935–944, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

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

Jiahao Xu, Wei Shao, Lihui Chen, and Lemao Liu. 2023. SimCSE++: Improving contrastive learning for sentence embeddings from two perspectives. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12028–12040, Singapore. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Nan Yang, Furu Wei, Binxing Jiao, Daxing Jiang, and Linjun Yang. 2021. xMoCo: Cross momentum contrastive learning for open-domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6120–6129, Online. Association for Computational Linguistics.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Junlei Zhang, Zhenzhong Lan, and Junxian He. 2023. Contrastive learning of sentence embeddings from scratch. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3916–3932, Singapore. Association for Computational Linguistics.

Yan Zhao, Huifang Ma, Jing Wang, Xiangchun He, and Liang Chang. 2024. Question-response representation with dual-level contrastive learning for improving knowledge tracing. *Information Sciences*, 658:120032.

Kun Zhou, Beichen Zhang, Xin Zhao, and Ji-Rong Wen. 2022. Debiased contrastive learning of unsupervised

sentence representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6120–6130, Dublin, Ireland. Association for Computational Linguistics.

A Appendix

A.1 Hyper-parameters and Datasets

We present our parameters and datasets' statistics in Tables 7 and 8, respectively. We use the grid search parameters starting from 1e0 to 1e-5. The full grid search step can observed from Figure 2. In addition, the fine-tuning time (mUSE-based) is only ∼20 minutes using a single V100.

A.2 Hyper-parameter Sensitivity

In this study, we observe the parameter sensitivity in each multi-task loss. In particular, we adjust the value according to Table 7 where we change only one value while keeping the other unchanged. In addition, we evaluate the performance only on the XORQA dataset. As shown in Figure 2, the most sensitive loss in our framework is the XLC loss $\mathcal{L}_{\text{XLC}}$. This is because $\mathcal{L}_{\text{XLC}}$ is the backbone of our multi-task training objective. The performance of $\mathcal{L}_{\text{XLC}}$ is changing when the parameter decreased to lower than 1e-2, while $\mathcal{L}_{\text{rank}}$ and $\mathcal{L}_{\text{HA}}$ almost remain unchanged.

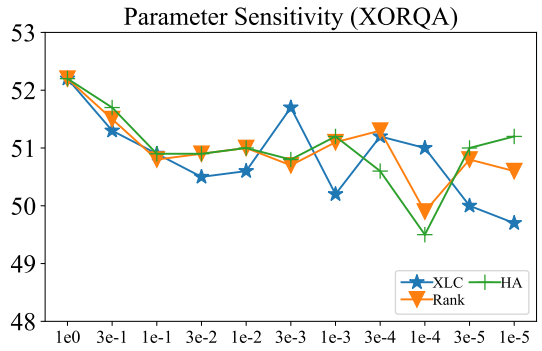

Figure 2: The sensitivity of each hyper-parameter in our multi-task framework, such as cross-lingual language consistency ($\mathcal{L}_{\text{XLC}}$), Ranking loss ($\mathcal{L}_{\text{rank}}$), and Hierarchical alignment loss ($\mathcal{L}_{\text{HA}}$).

A.3 Out-of-Domain Retrievals

In this experiment, we aim to assess the generalizability of our method in out-of-domain settings. Each model was trained on one dataset and tested on other datasets. Since there is an incompatibility in language overlap between the datasets, unsupported languages (the languages that were not in the original training data of mUSE-based) were excluded from this experiment.

As depicted in Table 6, our method demonstrates superior generalization compared to competitive methods in 22 out of 28 cases. In particular, when training our model on XQuAD and evaluating it on previously unseen datasets, our method consistently outperforms mSimCSE and CL-ReLKT with statistical significance ($p < 0.05$) across all cases, e.g., McCrolin exhibits an average improvement of 12.7 points over CL-ReLKT. This improvement in generalization can be attributed to the novel training objective we introduced, which enhances the robustness of our method.

A.4 Design Analysis

In this study, we conduct a comprehensive design analysis of our framework, McCrolin, by investigating the loss function configurations. For the loss functions, we explore three major configurations: (i) utilizing each training objective, $\mathcal{L}_{\text{XLC}}$, $\mathcal{L}_{\text{rank}}$, and $\mathcal{L}_{\text{HA}}$, individually; (ii) employing two training objectives in combination; and (iii) integrating all training objectives simultaneously. The results are shown in Table 9

Only one objective. The results indicate that the utilization of the cross-lingual transfer as a training objective, specifically $\mathcal{L}_{\text{XLC}}$, outperforms both contrastive learning ($\mathcal{L}_{\text{rank}}$) and hierarchical alignment ($\mathcal{L}_{\text{HA}}$). Furthermore, we observed significant overall improvements from our novel training objective, namely $\mathcal{L}_{\text{XLC}}$. On the other hand, relying solely on the ranking or hierarchical training objectives resulted in a degradation of the model's performance.

Two objectives. Integrating the cross-lingual consistency concept with the ranking training objective substantially improved the model's average performance. Conversely, when employing the $\mathcal{L}_{\text{rank}}$ and $\mathcal{L}_{\text{HA}}$ objectives in the base model, the performance drastically decreases compared to utilizing the $\mathcal{L}_{\text{XLC}}$ loss alone. These results serve to highlight the importance of the cross-lingual consistency concept in the context of cross-lingual retrieval QA.

Three objectives. Employing all three training objectives simultaneously within a single model yielded considerable performance enhancements in the average case, surpassing the improvements achieved by combining two training objectives alone. Moreover, the hierarchical alignment ($\mathcal{L}_{\text{HA}}$) loss predominantly influenced the performance of sentence retrieval, showcasing its crucial role in enhancing the model's ability to retrieve relevant

Model	XORQA			XQuAD							MLQA				Avg.
	RU	KO	JA	AR	DE	ES	RU	TH	ZH	TR	AR	DE	ES	ZH	
Train on XORQA. Test on XQuAD and MLQA datasets															
XLM-R-mSimCSE	-	-	-	34.0	45.8	45.8	45.4	42.0	48.7	46.6	23.0	42.6	40.8	35.1	40.9
CL-ReLKT	-	-	-	60.9	73.9	73.9	**67.2**	71.4	**67.6**	**73.1**	41.6	63.5	62.4	51.8	64.3
McCrolin	-	-	-	**63.4**	**74.4**	**75.6**	66.8	**74.4**	**67.6**	**73.1**	**44.7**	**66.2**	**63.0**	**53.8**	**65.7**
Train on XQuAD. Test on XORQA and MLQA datasets															
XLM-R-mSimCSE	22.1	17.5	15.5	-	-	-	-	-	-	-	17.8	31.6	30.2	24.4	22.7
CL-ReLKT	20.3	14.1	17.9	-	-	-	-	-	-	-	21.5	30.1	33.4	24.8	23.2
McCrolin	**37.2**	**28.5**	**29.0**	-	-	-	-	-	-	-	**28.4**	**45.3**	**46.8**	**35.9**	**35.9**
Train on MLQA. Test on XORQA and XQuAD datasets															
XLM-R-mSimCSE	14.0	16.5	19.9	13.0	31.1	22.3	32.8	35.7	41.2	35.7	-	-	-	-	26.2
CL-ReLKT	48.4	36.1	42.2	63.0	**79.4**	**76.1**	71.8	**74.8**	74.4	70.2	-	-	-	-	63.6
McCrolin	**51.0**	**37.3**	**42.7**	**64.7**	**79.4**	**76.1**	**73.5**	74.4	**75.6**	**70.6**	-	-	-	-	**64.5**

Table 6: Recall at 1 (R@1) on cross-lingual out-of-domain retrievals in *supported languages*. We trained a model on A dataset and tested on B and C datasets.

Parameters	Values
$\gamma_1, \gamma_2, \gamma_3$	1e0, 1e0, 1e0
$\beta_1, \beta_2, \beta_3, \beta_4$	1e-3, 1e0, 1e-3, 1e-1
$\lambda_1, \lambda_2, \tau$	3e-3, 1e-5, 0.05
ω_1, ω_2	1e-3, 1e-4

Table 7: Hyper-parameter settings.

Dataset	#Train	#Development	#Test	#Total	#Language
XORQA	6,264	894	1,781	8,949	4
XQuAD	97,006	10,779	13,090	120,875	11
MLQA	23,594	3,370	6,742	33,706	6

Table 8: Dataset statistics.

sentences accurately. These findings underscore the significance of incorporating all three training objectives in the multi-task setup, particularly in sentence retrieval.

A.5 Updating The Teacher Model

In this study, we conducted experiments to evaluate the effectiveness of the teacher update technique, as discussed in Section 3.4. Table 10 presents the performance improvements with and without the teacher update. The results of McCrolin with the teacher update are better than without in 15 out of 21 cases (71% improvement cases). The overall performance also increased from 63.6 to 64.2 points. These results emphasize the essential of the teacher update approach to increase the cross-lingual capability.

Model	XORQA				XQuAD											MLQA						Avg.
	RU	KO	JA	FI	AR	DE	ES	RU	TH	ZH	TR	RO	EL	HI	VI	AR	DE	ES	ZH	HI	VI	
Only one objective (Document retrieval)																						
Only $\mathcal{L}_{\text{XLC}}$	**59.3**	**46.0**	**51.0**	**48.2**	**81.5**	**85.3**	**86.6**	**83.6**	**86.1**	**83.2**	**81.1**	**78.5**	**61.9**	32.9	**73.9**	**49.5**	**64.6**	**64.2**	**57.7**	2.0	**46.0**	**63.0**
Only $\mathcal{L}_{\text{rank}}$	57.0	43.2	49.5	42.6	77.7	81.5	81.5	81.5	81.9	80.7	80.7	74.7	55.9	**35.7**	66.8	**49.5**	63.1	63.4	55.6	**3.2**	39.5	60.2
Only $\mathcal{L}_{\text{HA}}$	54.2	44.5	46.6	39.9	68.5	79.8	82.4	72.3	75.2	72.3	72.3	59.7	16.4	4.2	34.5	**49.5**	64.5	63.6	57.1	2.0	27.4	51.8
Two objectives (Document retrieval)																						
Previous best ($\mathcal{L}_{\text{XLC}}$)	**59.3**	46.0	51.0	48.2	81.5	85.3	**86.6**	83.6	86.1	83.2	81.1	78.5	61.9	32.9	73.9	49.5	64.6	**64.2**	**57.7**	2.0	46.0	63.0
$\mathcal{L}_{\text{XLC}}$ + $\mathcal{L}_{\text{rank}}$	58.7	46.0	**53.1**	**49.1**	81.5	**85.7**	**86.6**	**85.3**	**86.6**	**84.9**	**84.0**	79.0	**65.1**	**39.1**	**74.4**	**50.5**	64.8	63.6	**57.7**	**2.2**	**46.8**	**64.0**
$\mathcal{L}_{\text{XLC}}$ + $\mathcal{L}_{\text{HA}}$	57.3	45.1	49.2	46.4	**81.9**	84.9	**86.6**	**85.3**	**86.6**	83.6	83.2	**80.3**	59.7	29.8	69.7	49.1	64.5	63.8	57.3	2.0	43.2	62.4
$\mathcal{L}_{\text{rank}}$ + $\mathcal{L}_{\text{HA}}$	56.7	**46.2**	48.7	37.5	77.3	81.9	84.0	81.9	82.4	81.9	81.5	75.6	58.0	34.0	68.1	48.9	**65.4**	63.8	56.3	2.0	38.7	60.5
Three objectives (Document retrieval)																						
Previous best ($\mathcal{L}_{\text{XLC}}$ + $\mathcal{L}_{\text{rank}}$)	58.7	46.0	**53.1**	**49.1**	81.5	**85.7**	86.6	**85.3**	86.6	**84.9**	84.0	79.0	65.1	**39.1**	**74.4**	**50.5**	64.8	63.6	57.7	**2.2**	46.8	64.0
McCrolin ($\mathcal{L}_{\text{XLC}}$ + $\mathcal{L}_{\text{rank}}$ + $\mathcal{L}_{\text{HA}}$)	**59.9**	**48.1**	51.6	**49.1**	**82.8**	84.5	**87.4**	84.5	**86.7**	84.5	**84.9**	**81.1**	**65.6**	37.8	73.5	50.1	**65.8**	**63.8**	**57.9**	2.0	**47.4**	**64.2**
Three objectives (Sentence retrieval)																						
Previous best ($\mathcal{L}_{\text{XLC}}$ + $\mathcal{L}_{\text{rank}}$)	26.9	22.4	**20.2**	21.3	37.4	37.4	42.4	43.3	37.6	34.0	31.6	**36.3**	**23.2**	**11.8**	7.2	17.4	36.4	49.4	**51.2**	1.6	18.8	28.9
McCrolin ($\mathcal{L}_{\text{XLC}}$ + $\mathcal{L}_{\text{rank}}$ + $\mathcal{L}_{\text{HA}}$)	**28.4**	**22.8**	**20.2**	**22.1**	**49.6**	**57.2**	**54.6**	**53.2**	**47.9**	**50.6**	**51.9**	27.8	18.1	8.0	**22.0**	**36.6**	**51.6**	**52.0**	41.9	**1.8**	**23.1**	**35.3**

Table 9: Recall at 1 (R@1) on each loss component.

Model	XORQA				XQuAD											MLQA						Avg.
	RU	KO	JA	FI	AR	DE	ES	RU	TH	ZH	TR	RO	EL	HI	VI	AR	DE	ES	ZH	HI	VI	
McCrolin	**59.9**	**48.1**	**51.6**	**49.1**	**82.8**	**84.5**	**87.4**	**84.5**	85.7	**84.5**	**84.9**	**81.1**	65.6	37.8	**73.5**	**50.1**	**65.8**	**63.4**	**57.9**	2.0	**47.4**	**64.2**
- teacher update	57.0	46.6	**51.6**	47.7	79.8	83.6	87.0	84.0	**88.2**	84.0	83.0	78.6	**67.6**	**42.9**	71.4	49.7	65.4	**63.4**	57.3	**2.4**	44.5	63.6

Table 10: Recall at 1 (R@1) on removed and added teacher update setting.

Model	XORQA				XQuAD											MLQA						Avg.
	RU	KO	JA	FI	AR	DE	ES	RU	TH	ZH	TR	RO	EL	HI	VI	AR	DE	ES	ZH	HI	VI	
Competitive methods																						
mBERT-mSimCSE	52.1	40.9	40.2	46.0	77.3	83.9	86.1	83.5	15.5	80.7	82.8	73.2	65.0	**64.4**	71.9	11.0	31.1	26.2	38.5	21.7	25.6	53.2
XLM-R-mSimCSE	44.1	34.6	35.5	46.0	69.3	79.0	76.9	75.2	75.2	75.2	77.3	65.6	64.4	62.7	70.3	11.0	21.9	14.8	28.0	16.5	4.0	49.9
DPR	33.8	2.0	26.9	39.1	38.7	51.3	58.0	52.1	10.9	29.2	41.2	52.9	36.1	15.5	10.1	35.5	56.6	59.0	55.0	**50.9**	32.1	37.5
CORA	18.9	11.5	10.4	15.1	22.3	40.8	39.1	32.8	3.4	24.8	26.1	32.4	25.6	20.6	33.2	19.1	32.8	35.5	22.0	15.8	27.2	24.3
LaBSE	29.8	26.7	33.2	40.6	41.2	43.7	47.1	42.4	13.0	44.5	40.8	42.0	42.9	37.8	39.9	33.8	35.4	38.4	40.3	**50.9**	27.2	37.7
CL-ReLKT	58.2	47.7	49.5	48.2	79.4	83.2	84.0	83.6	86.1	82.4	80.3	76.9	64.3	34.0	71.8	48.5	64.8	62.8	**57.9**	3.6	44.2	62.4
Proposed model																						
McCrolin	**59.9**	**48.1**	**51.6**	**49.1**	**82.8**	**84.5**	**87.4**	**84.5**	**86.7**	**84.5**	**84.9**	**81.1**	**65.6**	37.8	**73.5**	**50.1**	**65.8**	**63.8**	**57.9**	2.0	**47.4**	**64.2**

Table 11: The full recall at 1 (R@1) score on the cross-lingual retrieval QA task.

Model	XORQA				XQuAD											MLQA						Avg.
	RU	KO	JA	FI	AR	DE	ES	RU	TH	ZH	TR	RO	EL	HI	VI	AR	DE	ES	ZH	HI	VI	
R@10																						
mBERT-mSimCSE	82.2	75.9	74.1	63.3	94.1	97.1	97.5	**97.5**	96.2	**97.1**	96.6	87.5	**85.0**	**85.4**	85.8	36.4	51.8	39.0	66.9	39.4	51.5	76.2
DPR	74.2	15.0	67.9	55.1	79.0	87.4	89.5	87.8	43.3	86.6	75.6	85.3	71.4	55.0	41.6	72.5	82.2	84.6	**85.7**	**59.8**	54.2	69.2
CL-ReLKT	85.4	78.3	78.8	63.9	95.8	96.6	97.9	96.6	**98.7**	**97.1**	**97.9**	95.0	83.6	64.3	88.2	76.4	**87.7**	**85.6**	84.7	10.5	61.7	82.1
$\mathcal{L}_{\text{XLC}}$	84.8	77.6	80.6	63.1	96.2	97.1	**98.3**	96.9	98.4	95.8	96.6	91.2	82.7	63.4	86.6	77.4	87.1	85.0	83.1	8.7	63.1	81.6
$\mathcal{L}_{\text{XLC}}$+$\mathcal{L}_{\text{rank}}$	**85.7**	78.3	81.9	64.2	**97.5**	**97.9**	**98.3**	97.1	98.6	96.6	97.1	95.0	83.3	67.2	88.2	77.4	**87.7**	85.2	83.1	9.5	65.7	82.6
McCrolin	85.4	**78.5**	**82.1**	**65.2**	95.4	97.5	97.5	97.1	98.3	**97.1**	97.1	**95.8**	83.6	68.1	**89.5**	**77.8**	87.5	85.0	83.5	7.7	**68.7**	**82.8**
MRR@10																						
mBERT-mSimCSE	60.8	51.7	51.4	52.1	77.9	84.9	84.3	83.5	82.6	82.8	84.4	74.0	71.7	**70.4**	75.4	17.9	30.8	21.9	39.4	16.6	30.7	59.3
DPR	45.6	5.0	39.5	35.2	49.8	63.5	67.3	63.3	20.2	61.4	50.6	62.6	46.7	26.4	17.6	46.8	65.0	67.2	65.2	**36.4**	39.2	46.4
CL-ReLKT	64.2	55.2	56.9	52.8	84.3	87.1	88.2	87.1	89.6	86.6	86.0	81.7	68.3	41.8	76.6	57.0	72.3	70.1	66.5	6.6	50.7	68.1
$\mathcal{L}_{\text{XLC}}$	67.6	56.8	60.0	53.0	86.2	87.8	90.2	87.8	90.3	87.6	87.0	81.7	68.3	41.8	76.6	57.9	72.2	71.2	66.0	3.4	52.4	68.8
$\mathcal{L}_{\text{XLC}}$+$\mathcal{L}_{\text{rank}}$	67.7	56.8	60.2	53.3	86.5	87.5	90.2	89.0	**90.7**	88.7	88.6	83.6	71.1	47.2	77.8	58.2	72.4	**71.4**	65.9	3.7	52.8	69.7
McCrolin	**68.2**	**57.6**	**61.0**	**54.0**	**86.8**	**88.9**	**90.3**	**89.2**	89.8	**88.8**	**89.0**	**85.3**	**72.0**	47.1	**78.1**	**58.4**	**72.9**	71.2	**66.7**	3.2	**53.1**	**70.1**

Table 12: The full recall and MRR scores on the top-10 retrieval.

Model	XORQA				XQuAD											MLQA						Avg.
	RU	KO	JA	FI	AR	DE	ES	RU	TH	ZH	TR	RO	EL	HI	VI	AR	DE	ES	ZH	HI	VI	
Sentence retrieval																						
mBERT-mSimCSE	26.1	19.0	17.1	17.1	30.7	42.4	39.5	33.8	29.4	36.7	34.6	22.2	**20.7**	**27.8**	**22.0**	2.9	12.5	2.8	14.5	3.3	7.8	22.0
DPR	14.0	0.0	9.6	19.1	32.2	41.0	49.4	42.7	7.5	41.9	32.4	22.5	17.2	7.5	5.3	29.5	45.5	50.4	37.7	**17.5**	20.0	25.9
CL-ReLKT	26.4	**23.0**	18.4	21.0	21.0	25.4	26.1	20.7	25.6	23.2	27.8	12.2	7.6	3.0	9.7	35.5	49.8	51.0	42.5	1.8	20.0	23.4
$\mathcal{L}_{\text{XLC}}$	26.4	20.5	19.4	20.2	25.7	33.6	34.1	29.7	27.4	35.0	28.8	21.5	11.3	5.5	15.7	35.8	50.8	51.6	41.3	1.8	20.9	26.5
$\mathcal{L}_{\text{XLC}}$+$\mathcal{L}_{\text{HA}}$	26.6	21.5	**20.2**	21.0	46.6	52.1	49.6	47.7	44.1	46.4	50.2	27.4	12.2	5.9	18.2	36.4	51.2	**52.0**	**42.6**	2.0	21.3	33.1
McCrolin	**28.4**	22.8	**20.2**	**22.1**	**49.6**	**57.2**	**54.6**	**53.2**	**47.9**	**50.6**	**51.9**	**27.8**	18.1	8.0	**22.0**	**36.6**	**51.6**	**52.0**	41.9	1.8	**23.1**	**35.3**
Paragraph retrieval																						
mBERT-mSimCSE	44.7	34.8	34.5	43.3	45.0	55.9	53.8	47.5	44.1	49.6	54.6	49.6	**43.3**	**45.4**	51.3	9.5	22.3	12.6	24.0	10.5	16.8	37.8
DPR	34.4	2.0	25.1	36.4	46.2	68.1	70.7	67.6	15.5	62.3	47.5	59.3	38.7	15.5	8.4	38.1	53.1	59.0	50.6	**27.4**	27.8	40.7
CL-ReLKT	56.4	44.1	45.3	45.8	48.3	51.7	51.3	48.7	51.7	48.3	53.8	31.1	17.6	15.1	23.5	44.3	61.9	**62.4**	52.0	3.2	32.2	42.3
$\mathcal{L}_{\text{XLC}}$	56.4	42.2	46.1	46.0	52.2	58.5	61.8	57.6	58.5	53.9	59.7	46.2	23.9	15.1	33.7	46.6	61.3	61.6	53.0	1.8	32.8	46.1
$\mathcal{L}_{\text{XLC}}$+$\mathcal{L}_{\text{HA}}$	56.4	42.2	47.4	46.4	66.4	70.6	74.4	73.1	73.9	67.6	70.2	60.8	37.3	16.4	46.1	46.8	**62.0**	62.3	**54.0**	2.2	33.8	52.9
McCrolin	**57.9**	**44.7**	**47.7**	**46.9**	**70.6**	**77.3**	**77.7**	**76.5**	**76.5**	**73.9**	**75.6**	**64.7**	**43.3**	19.3	**52.1**	**47.8**	61.7	62.2	53.4	2.2	**34.8**	**55.6**

Table 13: The full recall at 1 (R@1) score on cross-lingual sentence and paragraph retrievals.

Method	XORQA				XQuAD					MLQA					Avg
	RU	KO	JA	FI	AR	DE	ES	ZH	VI	AR	DE	ES	ZH	VI	
Paragraph-based retrieval															
CL-ReLKT	28.4	35.1	37.4	24.1	24.4	27.5	25.9	26.6	15.9	**17.7**	**40.7**	33.4	28.5	**13.1**	27.1
McCrolin	**33.7**	**55.2**	**57.9**	**44.3**	**40.7**	**47.4**	**46.4**	**44.8**	**27.3**	17.2	39.7	**33.6**	**30.6**	13.0	**38.0**
Document-based retrieval															
CL-ReLKT	20.1	34.6	37.0	25.3	56.6	65.9	63.8	60.1	56.5	20.2	**37.7**	30.6	**24.8**	15.9	39.2
McCrolin	**20.3**	**36.2**	**38.2**	**25.5**	**57.4**	**66.5**	**64.9**	**60.9**	**57.1**	**21.2**	**37.7**	**32.5**	24.5	**16.3**	**39.9**

Table 14: The full F1 score on the cross-lingual MRC experiment using GPT3.5 Turbo.