Language Concept Erasure for Language-invariant Dense Retrieval

Zhiqi Huang[1]*, **Puxuan Yu**[2]*, **Shauli Ravfogel**[3] and **James Allan**[4]
[1]Capital One, [2]Snowflake Inc., [3]Bar-Ilan University,
[4]University of Massachusetts Amherst
zhiqi.huang@capitalone.com puxuan.yu@snowflake.com
shauli.ravfogel@gmail.com allan@cs.umass.edu

Abstract

Multilingual models aim for language-invariant representations but still prominently encode language identity. This, along with the scarcity of high-quality parallel retrieval data, limits their performance in retrieval. We introduce LANCER, a multi-task learning framework that improves language-invariant dense retrieval by reducing language-specific signals in the embedding space. Leveraging the notion of linear concept erasure, we design a loss function that penalizes cross-correlation between representations and their language labels. LANCER leverages only English retrieval data and general multilingual corpora, training models to focus on language-invariant retrieval by semantic similarity without necessitating a vast parallel corpus. Experimental results on various datasets show our method consistently improves over baselines, with extensive analyses demonstrating greater language agnosticism.

1 Introduction

Multilingual text encoders (Devlin et al., 2019; Conneau et al., 2020; Xue et al., 2021) aim to provide representations that capture the essential *meaning* of texts, irrespective of *language identity*. These models show promise in information retrieval (IR) tasks, outperforming traditional lexical-based methods like BM25 (Ni et al., 2022; Gao and Callan, 2022; Yates et al., 2021). Despite their potential, these models often underperform with unseen or underrepresented languages (Joshi et al., 2020; Nooralahzadeh et al., 2020).

This paper explores strategies to enhance language invariance in multilingual retrieval settings. We introduce **Language Concept Erasure for Language-invariant Dense Retrieval** (**LANCER**), a multi-task learning framework designed to induce language invariance in the representation space of dense retrieval models. We focus particularly on the challenges posed by *English-only fine-tuning*, where a multilingual model is fine-tuned using only English queries and documents. While the ideal is for document relevance to transcend language barriers, training regimes typically lack comprehensive multilingual coverage. This often leads to a performance gap between languages well-represented in the training data and those that are not, exacerbated by the scarcity of reliable relevance judgments in many languages.

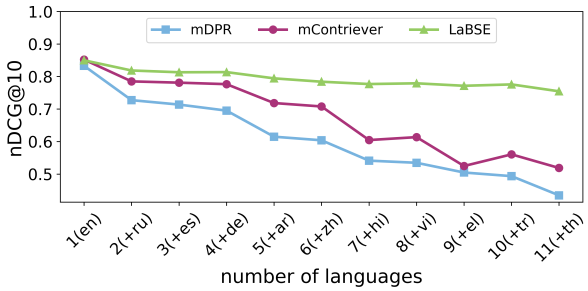

Figure 1: nDCG@10 decreases while the number of languages used in queries and documents increases. Results based on parallel data from LAReQA.

Language bias (Wu and Dredze, 2020) and data scarcity (Litschko et al., 2022) are known as limiting factors of the use of dense retrieval model performance on monolingual (non-English), cross-lingual, and multilingual[1] retrieval tasks. This degrading performance is illustrated in Figure 1, which presents retrieval results on the LAReQA dataset (Roy et al., 2020), a retrieval dataset with parallel queries and documents in 11 languages. When the number of languages on which models are evaluated is increased, performance degrades across all models.

Our approach, LANCER, regularizes the fine-

*Work completed while at UMass Amherst.

[1]To avoid confusion, we define "multilingual retrieval" as the task of retrieving information where multiple languages are involved in either the query, the document, or both. Conversely, we refer to tasks where both the query and the document are in the same language as "monolingual retrieval."

tuning process of the retrieval model by an auxiliary language invariance objective. Building upon the concept of *guardedness* (Ravfogel et al., 2023; Belrose et al., 2024), the inability to linearly predict a given concept (in our case, language identity) from a representation, our regularization scheme encourages guarded representations with respect to language identity. Specifically, we penalize the *correlation* between the representations and the language of the text that is represented. This auxiliary objective can be computed using *any* texts that are annotated by their language, i.e., it can make use of any multilingual non-parallel corpus, and does not require any specialized parallel data or human-annotation of semantic similarity. As such, it is particularly useful for languages that are not well-represented in commonly used retrieval datasets.

The dense retrieval models developed using our framework exhibit reduced language bias, improving all three backbone encoders on two multilingual retrieval datasets. Benefiting from language-invariant representations, these models also show substantial improvements in monolingual and cross-lingual retrieval tasks. Notably, on MIRACL (Zhang et al., 2023b), LANCER improves mDPR by 13.6% and mContriever by 32.5% at nDCG@10, also outperforming an LLM-based data augmentation method[2].

2 Related Works

2.1 Language-invariant Dense Retrieval

Multilingual versions of PLMs (mPLMs), such as mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020) and mT5 (Xue et al., 2021), allow for the joint learning of representations for many languages. Fine-tuning these models with retrieval-specific data enables them to learn the knowledge of query-document matching and perform retrieval tasks across diverse linguistic settings (e.g., monolingual, cross-lingual, and multilingual). Following the dense retrieval paradigm, mDPR (Zhang et al., 2021) extends English DPR (Karpukhin et al., 2020) to the multilingual setting by substituting the backbone encoder for mBERT. Later, mContriever (Izacard et al., 2022) adopts an unsupervised contrastive pre-training objective using data from mC4 (Xue et al., 2021). Cross-lingual pre-training tasks, such as translation ranking tasks, are applied prior to retrieval fine-tuning to improve cross-lingual alignment (Feng et al., 2022; Abulkhanov et al., 2023; Lin et al., 2023). However, compared to a large number of possible language configurations, the limited availability of multilingual retrieval training data significantly impacts the performance of dense retrieval models.

Some studies focused on building multilingual datasets for better training or evaluation. Datasets based on human annotations, such as NeuCLIR (Lawrie et al., 2024) and LAReQA (Roy et al., 2020), have been proposed for model evaluation. Large-scale synthetic training data generation involves two main simulation strategies: translating English retrieval datasets into target languages and building pseudo-labels using a corpus in target languages. For example, Sasaki et al. (2018) proposed a large cross-lingual retrieval collection, WikiCLIR. It uses the title of articles in target languages linked from Wikipedia pages as the query to simulate relevance. Bonifacio et al. (2021) built a multilingual passage ranking dataset, mMARCO, by translating the queries and passages in MS MARCO into the target language using NMT models. Thakur et al. (2023) leverage LLMs to generate queries in target languages with minimal supervision. Since language bias also exists in the tools used for synthetic data generation – e.g., neural machine translation (NMT) models for translating (Wang et al., 2021) or large language models (LLMs) for query generation (Yu et al., 2023) – it is hard to ensure uniform data quality across languages (Navigli et al., 2023). In this work, instead of resorting to creating new data, we explore a different approach by reducing the language-specific information when training dense retrieval models.

2.2 Concept Erasure

The problem of linear concept erasure falls under the scope of information removal, typically tackled using either adversarial methods (Edwards and Storkey, 2016; Xie et al., 2017; Zhang et al., 2018) during training or post-hoc linear methods (Haghighatkhah et al., 2022). Adversarial methods incorporate a gradient-reversal layer to encourage representations that do not encode the protected attribute. However, Elazar and Goldberg (2018) demonstrated that these methods do not completely eliminate all related information, allowing new adversaries to recover it. As a more tractable alternative, linear methods involve identifying and neutralizing a linear subspace associ-

[2]The code for model training available at https://github.com/zhiqihuang/lancer

ated with the target concept using algebraic techniques such as PCA (Kleindessner et al., 2023), orthogonal rotation (Dev et al., 2021), or spectral approaches (Shao et al., 2023b,a). We adopt the concept of guardedness as defined by Ravfogel et al. (2023) to develop a loss function specifically designed for the language concept erasure task in the training of dense retrieval models.

Towards language-invariant models, prior works primarily focus on supervised embedding disentanglement and unsupervised matrix decomposition. Embedding disentanglement mainly leverages the same semantics of parallel sentences to separate the original representation from multilingual encoders into language-specific and language-invariant embeddings (Tiyajamorn et al., 2021). The unsupervised approach employs singular value decomposition (SVD) to extract the most significant vectors from the representation matrix as the language-specific factor (Yang et al., 2021; Xie et al., 2022).

3 Methodology

Our objective is to diminish the language-dependent features within the embedding space of a dense retrieval model, enabling it to produce language-invariant representations across various linguistic contexts while preserving effective retrieval results. The guardedness objective (Ravfogel et al., 2023)—the inability to linearly predict the concept from the representation—has been shown to be equivalent to having a zero cross-correlation matrix between the representations and the concept labels (Belrose et al., 2024). We take inspiration from this condition, and propose a correlation-based loss function that penalizes any linear correlation between the language labels and the representations, while at the same time, we aim to preserve effective retrieval results in the training language (that is, English).

3.1 Preliminary: Dense Retrieval Models

A dense retrieval model consists of a query encoder E_Q and a document encoder E_D to map the query q and document d into k-dimensional dense vectors $\mathbf{h}_q$ and $\mathbf{h}_d$, respectively. The model computes the relevance score of q and d using the dot product as

$$s(q, d) = \mathbf{h}_q \cdot \mathbf{h}_d^\top$$

For a given query, documents from a collection $\{d\}_{i=1}^N$ are ranked based on the relevance scores. Recent works in dense retrieval models mostly adopt multilingual pre-trained language models (mPLMs) as backbone encoders for feature extraction. Typically, these models use Siamese architecture (Xiong et al., 2021) for the query and document encoders, sharing weights to ensure consistent processing. Based on output representations, some models, such as mDPR, utilize the representations of the leading `[CLS]` token as the dense vector; others, such as mContriever, take the pooling of all token representations as the dense vector.

1. Input features are taken from mPLMs.				
Models	mBERT	XLM-R	mT5	mE5
Accuracy	98.1	96.1	97.2	98.0
2.Input features are taken from dense retrievers.				
Models	mDPR (mBERT)	mDPR (mT5)	mContriever	LaBSE
Accuracy	97.9	96.7	98.0	81.6

Table 1: Language identification accuracy of logistic regression on mPLMs and retrieval models. Train test splits are sampled from mC4 dataset.

3.2 Language Identification

Texts can be similar due to various factors, such as language identity, syntactic structure, or semantics. In most search scenarios, we aim to retrieve documents based on their semantics, where the relevance between a query and a document is independent of the language in which they are expressed (Opitz and Frank, 2022). For example, if a document is labeled relevant to a query in English, translating both the query and the document into Arabic should not alter the judgment of their relevance. This principle highlights the importance of language-independent factors in determining the relevance of search results.

However, we find that mPLMs and the retrieval models built upon them still retain strong language-specific signals in their output representations. As shown in Table 1, we use the dense vectors from 16 languages as input features to train a logistic regression classifier for predicting language labels. The high accuracy achieved on a held-out test set indicates that the language factor remains strong on the dense vectors used for relevance scoring. This suggests that despite the intent to transcend language-identity, current models still embed significant language-specific information, which we hypothesize could impact their effectiveness in information retrieval tasks across diverse linguistic settings.

3.3 Language Concept Erasure

Based on this observation, we propose language concept erasure to reduce the influence of language in relevance scoring. Specifically, our goal is to prevent any linear classifier from detecting the language label given the dense vectors. Suppose $\mathrm{X} \in \mathbb{R}^k$ is the multilingual dense vector generated by the backbone encoder, and Z (the one-hot labels), taking values in $\mathcal{Z} = \{\mathbf{z} \in \{0,1\}^n \mid \|\mathbf{z}\|_1 = 1\}$, is the language labels tied to input instances. We adopt the idea of linear **guardedness** from Ravfogel et al. (2023) to formally define the language concept erasure.

Definition 3.1. *Let* $\mathcal{V} = \{\eta(\cdot;\boldsymbol{\theta}) : \mathbb{R}^k \to \mathbb{R}^n \mid \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ *be the class of all linear predictors, taking form* $\eta(\mathrm{X}) = \mathbf{W}\mathrm{X} + \mathbf{b}$ *for some weight matrix* $\mathbf{W} \in \mathbb{R}^{n\times k}$ *and bias* $\mathbf{b} \in \mathbb{R}^n$. X linearly guards Z *if no classifier in* $\mathcal{V}$ *can outperform a constant function at predicting* Z.

We say language concept is erased from dense vectors if linear guardedness is achieved between X and Z. Belrose et al. (2024) proved that the definition of linear guardedness is equivalent to zero covariance between every component of X and every component of Z.

Theorem 3.1. *Suppose* $\mathcal{L}$ *is convex loss functions defined on* $(\eta(\mathrm{X}), \mathrm{Z})$. *Then, if the cross-covariance matrix* Σ_{XZ}, *whose* $(i,j)^{th}$ *entry is* $\mathrm{Cov}(\mathrm{X}_i, \mathrm{Z}_j)$, *is a zero matrix, the constant predictor cannot be improved upon.*

Proof: See Belrose et al. (2024), theorem 3.2.

We can also use Theorem 3.1 to establish a concrete condition for language concept erasure. Therefore, if (X, Z) satisfies zero cross-covariance matrix, then the dense retrieval model prevents any linear classifier from detecting languages from its outputs.

3.4 Multi-task Learning

In Belrose et al. (2024), it is shown that there is a closed-form solution for a projection matrix that ensures the condition of Theorem 3.1 is satisfied while minimally modifying the pre-trained representations (in the L_2 distance). In this work, we use this solution as a baseline. In the method proposed here, we instead adopt a regularization scheme that encourages the zero-covariance condition *during training*. As we do not work on the pre-trained representations directly but rather adapt them simultaneously for the retrieval and language-invariance objectives, we can achieve a more expressive solution.

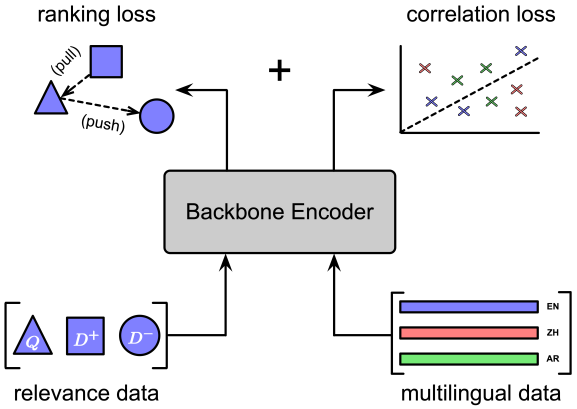

Figure 2: LANCER training objectives

During training, the model takes two types of data inputs in each batch:
(i) retrieval data, which includes triplets consisting of queries (Q), positive documents (D^+), and negative documents (D^-). We calculate the ranking loss, L_R, through contrastive learning (Chen et al., 2020):

$$\mathcal{L}_R = \sum_{q\in Q} \sum_{d^+\in D^+} -\log \frac{e^{s(q,d^+)}}{e^{s(q,d^+)} + \sum\limits_{d^-\in D^-} e^{s(q,d^-)}}$$

(ii) multilingual data, which is a group of passages (p) with language label (z), $\{(p_i, z_j)_{i=1}^m\}_{j=1}^n$, where n is the number of languages and m is the number of passages per language. We compute the cross-covariance matrix between dense vectors of input passages $\mathbf{X}$, and language labels $\mathbf{Z}$.

$$\Sigma_{\mathbf{XZ}} = \mathbb{E}[(\mathbf{X} - \bar{\mathbf{X}})(\mathbf{Z} - \bar{\mathbf{Z}})^\top]$$

The scale of the embedding values significantly influences the magnitude of covariance. Unnormalized outputs from some encoders result in covariance values that vary widely across different input instances. Therefore, we standardize the covariance matrix into the correlation matrix by dividing the standard deviations: $\rho_{\mathbf{XZ}} = \Sigma_{\mathbf{XZ}}/\sigma_{\mathbf{X}}\sigma_{\mathbf{Z}}$. The concept erasure loss is defined as the mean absolute value of the correlation matrix:

$$\mathcal{L}_C = \frac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n} |corr(\mathrm{X}_i, \mathrm{Z}_j)|$$

These two types of data and their corresponding losses complement each other effectively. The

concept erasure task effectively removes language-specific information from the dense vectors, enabling the retrieval task to concentrate on language-invariant knowledge. Simultaneously, the retrieval task, which focuses on semantic matching, ensures that the model maintains meaningful representations throughout the training. This balance prevents the concept erasure task from degenerating. Finally, as shown in Figure 2, we add the primary ranking loss $\mathcal{L}_R$ for retrieval and the correlation loss for concept erasure $\mathcal{L}_C$ to conduct the training of dense retrievers.

$$\mathcal{L} = \mathcal{L}_R + \mathcal{L}_C$$

Training data requirement. Because the language label is an intrinsic attribute separate from semantic meaning, the language concept erasure task of LANCER has minimal data requirements for multilingual input. A clean corpus from each language is sufficient to support the running of this task. On the other hand, as the retrieval task is less influenced by language-specific information, it can utilize query-document pairs in any language. In this work, we only use retrieval data in English for training.

4 Experimental Setup

4.1 Modeling Details

We apply LANCER to multilingual dense retrieval models with different degrees of multilingual pre-training: mBERT (mDPR) (Zhang et al., 2021), mContriever (Izacard et al., 2022), and LaBSE (Feng et al., 2022). The pre-training of mBERT is only an extension of masked language modeling (MLM) and next sentence prediction (NSP) to 104 languages. Based on mBERT, mContriever is further pre-trained on unsupervised contrastive learning over 29 languages using mC4 dataset. LaBSE, also built on mBERT, is further pre-trained on the translation ranking task, leveraging millions of parallel text.

To compare with existing baselines, we use the MS MARCO passage ranking dataset (Nguyen et al., 2016) as the retrieval training data. Note that there is no existing LaBSE-based dense retriever built on MS MARCO, so we created one by fine-tuning LaBSE on MS MARCO. For language concept erasure, we use a multilingual corpus containing 16 languages from different language families[3]. We sample 3M textual data per language from the mC4 (Xue et al., 2021) dataset. We set the batch size to 64 for the retrieval task and 256 (16 examples per language) for the concept erasure task. Based on the size of the MS MARCO train split, we train each model for 4 epochs with a learning rate of $2e^{-5}$. To monitor the language invariance during training (Section 5.3), we create a small held-out multilingual dataset for language label recovery using a logistic regression classifier. After every 1600 steps, we build a new classifier based on the current model outputs.

To evaluate language-invariant dense retrievers refined by LANCER, we conduct experiments on various benchmark retrieval datasets covering multilingual, cross-lingual, and monolingual (in many languages) tasks. Some of the evaluation datasets include training splits. To assess language agnosticism, we do not perform any additional fine-tuning using those splits to keep zero-shot evaluations of our approach and all compared methods.

4.2 Datasets and Metrics

4.2.1 Multilingual

CLEF. We evaluate searching a multilingual collection using English queries. The data is from the Cross-Language Evaluation Forum (CLEF) 2000-2003 campaign for bilingual ad-hoc retrieval tracks (Braschler, 2002). We include documents in French, German, and Italian to build a multilingual collection with 241K documents. Among 200 topic queries in English, we only consider a topic with relevant documents in all three languages as a valid query, leading to 133 queries in total.

LAReQA. We evaluate the retrieval performance when the query and collection are both multilingual. LAReQA (Roy et al., 2020) is a benchmark for language-invariant answer retrieval from a multilingual candidate pool. Each query appears in 11 languages[4] and has 11 parallel relevant documents.

4.2.2 Cross-lingual and Monolingual

XOR-Retrieve. We evaluate searching English collection using queries in another language. XOR-Retrieve (Asai et al., 2021) is a benchmark for evaluating cross-lingual retrieval systems. It includes 7 cross-lingual tasks between target language queries and English documents. The corpus contains 18.2M passages with a maximum of 100 word tokens from the English Wikipedia.

[3]List of training languages (ISO code): ar, bn, de, en, es, fa, fi, fr, hi, id, ja, ko, ru, te, th, zh

[4]Languages in LAReQA (ISO code): ar, de, el, en, es, hi, ru, th, tr, vi, zh

XTREME-UP. XTREME-UP (Ruder et al., 2023) focuses on extremely low-resource languages. Similar to XOR-Retrieve, it includes 20 cross-lingual tasks of queries in low-resource language and documents in English.

MIRACL. We evaluate monolingual retrieval across multiple languages. MIRACL (Zhang et al., 2023b) has a broad language coverage for evaluating monolingual retrieval. Developed on top of Mr. TyDi (Zhang et al., 2021), MIRACL comprises data in 18 languages, with both queries and documents presented in the same language. Our training covers 16 languages in the MIRACL dataset, except German and Yoruba.

4.2.3 Metrics

We report mAP and nDCG@10 for both multilingual evaluation datasets (CLEF and LAReQA). Following prior work (Zhang et al., 2023a; Li et al., 2022), we evaluate Recall@5kt and Recall@2kt on XOR-Retrieve, nDCG@10 and Recall@100 on MIRACL, and MRR@10 on XTREME-UP.

4.3 Compared Methods

Across all evaluations, we compared the performance of models incorporating LANCER to those without, e.g., mDPR+LANCER vs. mDPR. For multilingual evaluation, we compare LANCER with other language debiasing approaches, including post-hoc methods and knowledge distillation framework.

LSAR: As an unsupervised method, LSAR (Xie et al., 2022) is based on matrix decomposition to identify a language-invariant subspace and then directly projects the original multilingual embeddings onto that subspace to reduce the effects of language on downstream tasks.

LEACE: Also worked as an unsupervised method, LEACE (Belrose et al., 2024) derives a projection in closed-form to prevent linear classifiers from detecting a concept. We apply it upon baseline retrievers to reduce the effects of language concepts.

KD-SPD: Based on knowledge distillation, KD-SPD (Huang et al., 2023) designed a language-aware decomposition prompt for the encoder to transfer knowledge from an English retriever to multiple languages. Using parallel corpora to create supervision signals, this method is more resource intensive than both post-hoc baselines and LANCER.

Method	CLEF		LAReQA (Full)		LAReQA (Sampled)	
	mAP ↑	nDCG@10 ↑	mAP ↑	nDCG@10 ↑	mAP ↑	nDCG@10 ↑
KD-SPD	22.0	41.6	48.4	50.4	55.5	60.0
mDPR	20.2	34.6	25.5	31.7	41.0	41.6
+ LSAR	19.8	35.8	34.0	39.2	48.9	53.3
+ LEACE	18.9	34.6	33.4	38.7	48.9	53.2
+ LANCER	21.6	39.1	39.3	43.3	53.1	57.7
mContriever	27.2	46.1	31.1	37.3	48.8	52.5
+ LSAR	26.9	47.4	38.8	43.8	55.8	60.2
+ LEACE	28.3	48.8	39.1	44.2	56.3	60.7
+ LANCER	30.0	50.7	42.6	47.6	58.4	62.8
LaBSE	24.0	44.2	62.9	64.0	72.4	76.2
+ LSAR	22.8	42.5	61.4	62.1	70.9	74.9
+ LEACE	23.9	44.6	61.2	62.0	71.3	75.2
+ LANCER	25.8	47.0	64.5	65.2	74.5	78.1

Table 2: Results for multilingual retrieval on CLEF and LAReQA. LAReQA (Full) includes parallel queries and documents in 11 languages. LAReQA (Sampled) refers to randomly selecting a language from 11 languages for each query and document. Results are averaged over five folds. Our approaches are *highlighted* in light blue.

For cross-lingual and monolingual tasks, we include results from **SWIM-X** (Thakur et al., 2023), a synthetic query generation method using LLMs. It utilizes in-domain documents to generate synthetic queries and then performs fine-tuning to build multilingual dense retrieval models.

5 Experimental Results

5.1 Retrieval Performance

Multilingual. Table 2 lists the multilingual evaluation results. We observe that when LANCER is applied, all three baseline models show substantial improvements on two datasets in terms of both mAP and nDCG@10. Note that retrieval data used for training remained consistent across these experiments. Because of the language concept erasure, models built with LANCER have less language bias, leading to better performance on multilingual tasks. Moreover, LANCER outperforms post-hoc methods (LSAR and LEACE). Compared with the knowledge transfer method, LaBSE+LANCER uniformly improves KD-SPD, while mContriever+LANCER also performs better except on LAReQA (Full). Lastly, from a task perspective, LAReQA presents a greater challenge than CLEF due to the inclusion of more languages in its queries and documents. Because LaBSE is pre-trained on a wide range of languages using parallel sentences, mContriever is able to surpass LaBSE on CLEF but falls behind on LAReQA.

Cross-lingual. Table 3 and Table 4 list cross-lingual results on XOR-Retrieve and XTREME-UP

Method	Avg. (↑)	ar	bn	fi	ja	ko	ru	te
SWIM-X (500K)	59.0	54.0	67.4	59.2	52.7	55.1	54.4	70.2
mDPR	39.3	34.3	35.5	45.2	40.2	36.5	43.9	39.5
+ LANCER	41.4	36.2	37.8	47.1	37.8	45.3	42.2	43.3
mContriever	44.0	37.5	38.2	50.6	41.1	37.2	49.8	53.8
+ LANCER	45.7	43.0	35.9	56.4	39.4	46.0	43.5	55.5
LaBSE	56.8	56.0	63.5	57.6	50.2	50.2	48.1	71.8
+ LANCER	57.2	54.4	62.5	58.3	51.0	52.6	47.3	74.4

Table 3: Results showing Recall@5kt (%) for cross-lingual retrieval on XOR-Retrieve dev.

respectively. On XOR-Retrieve, LANCER demonstrates competitive performance compared to corresponding baseline models, improving 2.1 points on mDPR and 1.7 points on mContriever. When applied to LaBSE, LANCER aligns closely with the baseline. SWIM-X performs the best on XOR-Retrieve. However, SWIM-X utilizes in-domain data to generate cross-lingual training pairs, while our experiments are completely zero-shot evaluations. For collections with strong domain features like Wikipedia, synthetic data not only supports language-specific training but also acts as a form of domain adaptation, contributing to this strong performance. For results on Recall@2kt, refer to the Appendix B.

On XTREME-UP, LANCER consistently enhances performance over the baseline models. Both LaBSE and LaBSE+LANCER surpass SWIM-X. The performance of SWIM-X suggests that using LLMs for data augmentation does not always yield high-quality data, particularly for tasks involving low-resource languages.

Monolingual. Table 5 lists monolingual results on MIRACL, covering 18 languages. Compared to cross-lingual, LANCER improves the corresponding baseline models on monolingual tasks by a large margin. Specifically, in terms of nDCG@10, LANCER achieves an improvement of 5.7 points (13.6%) over mDPR, 12.3 points (32.5%) over mContriever, and 2.5 points (5.5%) over LaBSE. Surprisingly, when LANCER is applied, all three models outperform SWIM-X. This suggests that LANCER has robust zero-shot capability in monolingual tasks, highlighting its effectiveness without retrieval training for specific languages. From the data perspective, this also suggests that when language bias is reduced in embedding space, retrieval knowledge provided by MS MARCO data (English) is more comprehensive than language-specific synthetic data generated by

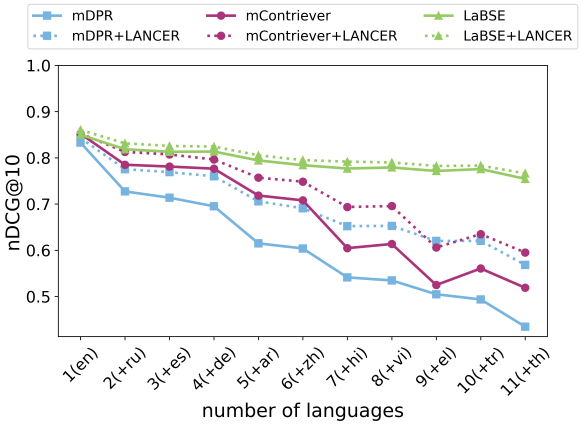

Figure 3: Compared to corresponding baselines, LANCER shows more robust nDCG@10 against the increase of languages. Results based on LAReQA.

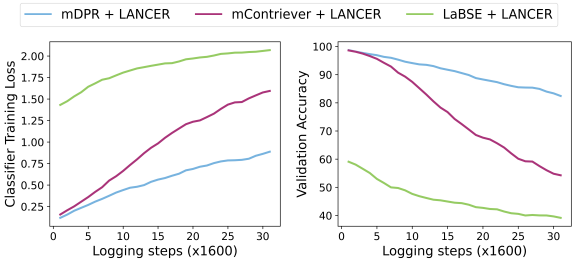

Figure 4: Training loss of logistic regression (Left) and prediction accuracy (Right) for language label recovery.

current LLMs. For results on Recall@100, refer to the Appendix A.

5.2 Effect of Multilingualism

In Figure 1, we show the language bias in multilingual retrieval by increasing the number of languages used in queries and documents. Here, we replicate the experiment with models trained using LANCER, observing how their performance shifts as more languages are incorporated into the queries and documents.

As shown in Figure 3, the models demonstrate improved resilience to language bias as the number of languages increases, maintaining higher levels of nDCG@10 compared to those without LANCER. This suggests that LANCER effectively mitigates the challenges posed by linguistic diversity, enhancing the model's ability to handle multilingual information retrieval more robustly.

5.3 Analysis of Training

To study the impact of language concept erasure on the dense vectors, we leverage held-out multilingual train and test splits to monitor language

Method	Avg. (↑)	as	bho	brx	gbm	gom	gu	hi	hne	kn	mai	ml	mni	mr	mwr	or	pa	ps	sa	ta	ur
SWIM-X (120K)	25.2	24.4	27.7	4.3	28.3	25.4	29.4	32.4	28.8	30.1	31.8	34.4	5.1	30.7	25.7	15.8	29.6	20.6	26.1	27.9	26.1
mDPR	5.9	2.6	6.5	0.6	7.0	2.2	5.4	13.9	5.7	6.3	6.9	8.7	0.3	8.7	6.1	0.7	9.5	2.6	4.1	7.7	13.3
+LANCER	9.8	5.0	9.5	0.8	11.2	6.3	11.9	19.2	10.0	10.5	11.5	14.1	0.8	15.9	9.9	0.3	15.5	3.7	9.5	13.8	16.4
mContriever	4.6	3.6	5.4	0.9	6.3	1.8	2.2	10.9	5.3	5.5	7.0	4.3	0.9	6.1	6.6	0.8	5.3	2.0	4.4	7.9	5.7
+LANCER	6.5	5.1	6.4	1.0	9.7	3.3	4.2	13.4	7.4	8.8	9.0	6.3	0.7	9.3	8.8	0.7	6.9	3.0	8.4	8.0	8.5
LaBSE	28.3	25.0	28.3	2.8	29.4	21.0	36.2	38.5	27.6	36.3	31.9	36.9	4.5	37.9	28.6	27.0	35.5	22.2	27.4	35.6	34.1
+LANCER	29.2	26.1	29.2	2.4	27.4	22.5	37.7	40.7	26.2	38.9	31.7	38.5	4.1	39.0	28.4	29.6	36.5	22.9	28.1	37.3	36.3

Table 4: Results showing MRR@10 (%) for cross-lingual retrieval on XTREME-UP test.

Method	Avg. (↑)	ar	bn	en	es	fa	fi	fr	hi	id	ja	ko	ru	sw	te	th	zh	de	yo
SWIM-X (180K)	46.4	60.2	57.1	34.7	33.4	36.3	40.6	64.3	33.0	39.5	40.8	43.3	49.7	40.0	55.9	56.3	63.3	60.2	57.1
mDPR	41.8	49.9	44.3	39.4	47.8	48.0	47.2	43.5	38.3	27.2	43.9	41.9	40.7	29.9	35.6	35.8	51.2	49.0	39.6
+ LANCER	47.5	55.6	50.5	43.4	46.3	49.1	56.6	46.1	36.7	34.3	49.0	47.4	46.7	39.4	52.7	46.1	50.0	46.8	58.5
mContriever	37.8	49.1	48.4	32.7	33.3	37.1	48.4	27.0	35.9	32.7	34.1	40.2	35.1	44.5	46.2	45.0	27.5	29.7	33.7
+ LANCER	50.1	61.4	56.9	40.7	46.1	38.0	65.4	41.2	35.7	43.6	48.1	54.5	46.2	58.0	67.9	58.2	45.1	43.2	51.7
LaBSE	45.6	50.2	53.7	35.6	37.7	42.4	57.2	40.6	41.4	37.8	34.6	46.2	40.5	57.4	53.9	50.1	34.9	39.7	67.7
+ LANCER	48.1	52.9	57.2	37.5	38.0	45.9	60.6	41.7	43.8	39.5	39.4	48.8	42.2	57.9	58.9	55.2	37.3	38.1	70.2

Table 5: Results showing nDCG@10 (%) for monolingual retrieval on MIRACL dev.

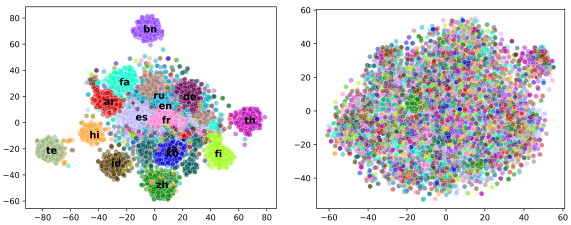

Figure 5: t-SNE visualization of multilingual representations from mDPR (Left) versus mDPR+LANCER (Right). Best viewed in color.

label recovery. At each logging step, we map the held-out splits into dense vectors and build a logistic regression classifier to predict language labels. Figure 4 records the loss on train split (Left) and prediction accuracy on test split (Right) on three models. As training continues, it is harder for the classifier to identify the language label according to rising loss and declining accuracy. This trend indicates that the language concept erasure task effectively reduces the language information in the dense vectors, making the model more language-invariant.

5.4 Analysis of Representation

To further demonstrate the impact of LANCER, we analyze the representations produced by dense retrieval models, both with and without the language concept erasure task. We sample 300 passages per language from 16 training languages and use t-SNE to visualize their representations. In Figure 5, the visualizations from mDPR show that the representations are predominantly clustered by language. However, after integrating LANCER, the representations from different languages are intermingled. This indicates that LANCER effectively diminishes language-specific clustering, resulting in a more language-invariant embedding space.

6 Conclusion

In this work, we introduce LANCER, a multi-task training framework designed to improve language-invariant dense retrieval. The core of LANCER is the language concept erasure task, which reduces the language-specific signals present in the multilingual dense vectors by preventing linear classifiers from detecting the language labels. Paired with the retrieval task, LANCER enables the model to prioritize learning language-invariant knowledge for query-document matching.

We conduct experiments across all possible linguistic settings of an IR task (e.g., monolingual, cross-lingual, and multilingual). The extensive results from these experiments demonstrate the effectiveness of LANCER in building language-invariant dense retrieval models. In multilingual contexts, LANCER outperforms knowledge transfer using parallel data. Furthermore, in monolingual tasks across 18 languages, LANCER, as a zero-shot approach, surpasses an in-domain data augmentation method based on LLMs. For future work, we are interested in extending the language concept to more general concept(s) to improve domain adaptation and convert LANCER to a general

framework for model debiasing for downstream applications.

7 Limitations

Our approach principally defines the idea of building language-invariant models by preventing linear classifiers from detecting the language label of input text. Nevertheless, language is inherently complex, intertwined with cultural nuances and semantic subtleties. The strategy of disabling linear classifiers to promote language-invariant retrieval within the dense retrieval paradigm may not be applicable to other tasks or model architectures, such as encoder-decoder models used for text generation. In fact, from our model checkpoints, the language labels can still be recovered by non-linear classifiers like multi-layer perceptron (MLP). Additionally, our experimental results (e.g., Figure 3 and 4) indicate the potential for further reducing language bias and enhancing retrieval performance in diverse linguistic settings. Therefore, our community still has a considerable path to tread in order to overcome language bias in AI models. Furthermore, we only conducted experiments in a zero-shot evaluation setting. Further investigation is needed for in-domain evaluation from both retrieval data and multilingual data perspectives.

Potential Risks. As most other research focused on multilingualism, our study covers only a small group of languages compared to more than 7,000 spoken languages on this planet. There exist risks that these findings may not be generalizable across all linguistic landscapes.

Acknowledgements

This work was supported in part by the Center for Intelligent Information Retrieval. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

References

Dmitry Abulkhanov, Nikita Sorokin, Sergey Nikolenko, and Valentin Malykh. 2023. Lapca: Language-agnostic pretraining with cross-lingual alignment. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, New York, NY, USA. Association for Computing Machinery.

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

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2024. Leace: perfect linear concept erasure in closed form. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

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

Martin Braschler. 2002. Clef 2002—overview of results. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 9–27. Springer.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Sunipa Dev, Tao Li, Jeff M Phillips, and Vivek Srikumar. 2021. OSCaR: Orthogonal subspace correction and rectification of biases in word embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5034–5050, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Harrison Edwards and Amos Storkey. 2016. Censoring representations with an adversary. In *International Conference in Learning Representations (ICLR2016)*, pages 1–14. 4th International Conference on Learning Representations, ICLR 2016 ; Conference date: 02-05-2016 Through 04-05-2016.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium. Association for Computational Linguistics.

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.

Luyu Gao and Jamie Callan. 2022. Unsupervised corpus aware language model pre-training for dense passage retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2843–2853, Dublin, Ireland. Association for Computational Linguistics.

Pantea Haghighatkhah, Antske Fokkens, Pia Sommerauer, Bettina Speckmann, and Kevin Verbeek. 2022. Better hit the nail on the head than beat around the bush: Removing protected attributes with a single projection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8395–8416, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhiqi Huang, Hansi Zeng, Hamed Zamani, and James Allan. 2023. Soft prompt decoding for multilingual dense retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 1208–1218, New York, NY, USA. Association for Computing Machinery.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

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)*, pages 6769–6781, Online. Association for Computational Linguistics.

Matthäus Kleindessner, Michele Donini, Chris Russell, and Muhammad Bilal Zafar. 2023. Efficient fair pca for fair representation learning. In *International Conference on Artificial Intelligence and Statistics*, pages 5250–5270. PMLR.

Dawn Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W Oard, Luca Soldaini, and Eugene Yang. 2024. Overview of the trec 2023 neuclir track. *arXiv preprint arXiv:2404.08071*.

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

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.

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.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: Origins, inventory, and discussion. *J. Data and Information Quality*, 15(2).

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. *choice*, 2640:660.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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)*, pages 4547–4562, Online. Association for Computational Linguistics.

Juri Opitz and Anette Frank. 2022. SBERT studies meaning representations: Decomposing sentence embeddings into explainable semantic features. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 625–638, Online only. Association for Computational Linguistics.

Shauli Ravfogel, Yoav Goldberg, and Ryan Cotterell. 2023. Log-linear guardedness and its implications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9413–9431, Toronto, Canada. Association for Computational Linguistics.

Uma Roy, Noah Constant, Rami Al-Rfou, Aditya Barua, Aaron Phillips, and Yinfei Yang. 2020. LAReQA: Language-agnostic answer retrieval from a multilingual pool. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5919–5930, Online. Association for Computational Linguistics.

Sebastian Ruder, Jonathan Clark, Alexander Gutkin, Mihir Kale, Min Ma, Massimo Nicosia, Shruti Rijhwani, Parker Riley, Jean-Michel Sarr, Xinyi Wang, John Wieting, Nitish Gupta, Anna Katanova, Christo Kirov, Dana Dickinson, Brian Roark, Bidisha Samanta, Connie Tao, David Adelani, Vera Axelrod, Isaac Caswell, Colin Cherry, Dan Garrette, Reeve Ingle, Melvin Johnson, Dmitry Panteleev, and Partha Talukdar. 2023. XTREME-UP: A user-centric scarce-data benchmark for under-represented languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1856–1884, Singapore. Association for Computational Linguistics.

Shota Sasaki, Shuo Sun, Shigehiko Schamoni, Kevin Duh, and Kentaro Inui. 2018. Cross-lingual learning-to-rank with shared representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 458–463.

Shun Shao, Yftah Ziser, and Shay B. Cohen. 2023a. Erasure of unaligned attributes from neural representations. *Transactions of the Association for Computational Linguistics*, 11:488–510.

Shun Shao, Yftah Ziser, and Shay B. Cohen. 2023b. Gold doesn't always glitter: Spectral removal of linear and nonlinear guarded attribute information. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1611–1622, Dubrovnik, Croatia. Association for Computational Linguistics.

Nandan Thakur, Jianmo Ni, Gustavo Hernández Ábrego, John Wieting, Jimmy Lin, and Daniel Cer. 2023. Leveraging llms for synthesizing training data across many languages in multilingual dense retrieval. *arXiv preprint arXiv:2311.05800*.

Nattapong Tiyajamorn, Tomoyuki Kajiwara, Yuki Arase, and Makoto Onizuka. 2021. Language-agnostic representation from multilingual sentence encoders for cross-lingual similarity estimation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7764–7774, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shuo Wang, Zhaopeng Tu, Zhixing Tan, Shuming Shi, Maosong Sun, and Yang Liu. 2021. On the language coverage bias for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4778–4790, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. 2017. Controllable invariance through adversarial feature learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 585–596, Red Hook, NY, USA. Curran Associates Inc.

Zhihui Xie, Handong Zhao, Tong Yu, and Shuai Li. 2022. Discovering low-rank subspaces for language-agnostic multilingual representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5617–5633, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

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.

Ziyi Yang, Yinfei Yang, Daniel Cer, and Eric Darve. 2021. A simple and effective method to eliminate the self language bias in multilingual representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5825–5832, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. 2021. Pretrained transformers for text ranking: BERT and beyond. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials*, pages 1–4, Online. Association for Computational Linguistics.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large language model as attributed training data generator: A tale of diversity

and bias. In *Advances in Neural Information Processing Systems*, volume 36, pages 55734–55784. Curran Associates, Inc.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 335–340, New York, NY, USA. Association for Computing Machinery.

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

Xinyu Zhang, Kelechi Ogueji, Xueguang Ma, and Jimmy Lin. 2023a. Toward best practices for training multilingual dense retrieval models. *ACM Trans. Inf. Syst.*, 42(2).

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023b. MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

Method	Avg. (↑)	ar	bn	en	es	fa	fi	fr	hi	id	ja	ko	ru	sw	te	th	zh	de	yo
SWIM-X (180K)	78.9	89.2	87.8	72.9	70	76.3	91.6	75.8	72.5	74.3	77.6	76.8	77.9	87.8	84.9	92.9	69.9	72.4	69.3
mDPR	79.4	84.1	81.9	76.8	86.4	89.8	78.8	91.5	77.6	57.3	82.5	73.7	79.7	61.6	76.2	67.8	94.4	89.8	79.5
+ LANCER	81.0	86.1	83.3	76.1	81.3	87.0	84.7	88.5	70.0	64.0	83.7	78.3	80.1	67.5	85.0	77.4	90.1	87.1	87.4
mContriever	60.6	73.5	80.8	52.1	49.5	61.7	66.0	51.8	50.3	63.5	65.6	56.3	58.9	73.5	85.9	76.6	58.2	36.3	30.2
+ LANCER	84.3	92.4	93.2	77.5	83.4	78.0	65.4	86.3	75.4	78.5	86.5	87.5	84.1	90.4	95.0	91.8	86.9	85.6	80.5
LaBSE	80.9	84.0	88.1	72.1	72.9	85.4	87.9	81.5	78.3	69.3	70.5	76.4	77.4	88.1	88.8	82.2	78.9	80.1	94.2
+ LANCER	82.4	85.8	90.8	73.5	74.4	86.6	88.6	82.3	81.0	70.6	75.5	76.7	77.8	88.6	92.3	86.1	80.1	78.9	93.7

Table 6: Results showing Recall@100 (%) for monolingual retrieval on MIRACL dev.

Method	Avg. (↑)	ar	bn	fi	ja	ko	ru	te
SWIM-X (500K)	49.2	46.3	57.2	49.0	42.7	45.6	44.7	58.8
mDPR	30.6	26.2	26.0	37.9	32.8	24.6	34.6	32.4
+ LANCER	31.8	28.8	25.3	40.4	28.6	35.4	31.6	32.4
mContriever	33.8	27.8	24.3	42.4	29.9	31.2	40.5	40.3
+ LANCER	38.4	37.8	27.0	50.3	32.4	36.5	36.7	48.3
LaBSE	47.1	44.6	52.6	49.7	36.1	44.9	39.6	62.2
+ LANCER	47.1	44.6	52.0	51.0	38.6	41.4	37.5	64.7

Table 7: Results showing Recall@2kt (%) for cross-lingual retrieval on XOR-Retrieve dev.

Appendix

A Additional Results on MIRACL

In Table 6, we report the Recall@100 scores on MIRACL for all denser retrievers. Models trained using LANCER outperform the models trained solely on retrieval tasks using English supervision data. Specifically, in terms of Recall@100, LANCER achieves an improvement of 1.6 points over mDPR, 23.7 points over mContriever, and 1.5 points over LaBSE. All three LANCER models surpass SWIM-X. This suggests that when language bias is reduced in embedding space, retrieval knowledge provided by MS MARCO data (English) is more comprehensive than language-specific synthetic data generated by LLMs.

B Additional Results on XOR-Retrieve

Table 7 lists the results of Recall@2kt on XOR-Retrieve. In terms of 2K tokens, LANCER improves by 1.2 points on mDPR and 4.6 points on mContriever. When applied to LaBSE, LANCER performs the same as the baseline. SWIM-X continues to outperform all other retrieval models. It leverages in-domain data to create cross-lingual training pairs, which is particularly effective for collections with distinct domain characteristics, such as Wikipedia. The synthetic data, in this case, not only caters to language-specific training needs but also functions as domain adaptation, which leads to this strong performance.