Language Bias in Multilingual Information Retrieval: The Nature of the Beast and Mitigation Methods

Jinrui Yang* **Fan Jiang*** **Timothy Baldwin*†**
*School of Computing & Information Systems, The University of Melbourne
†Mohamed bin Zayed University of Artificial Intelligence, UAE
{jinruiy,jifj}@student.unimelb.edu.au tbaldwin@unimelb.edu.au

Abstract

Language fairness in multilingual information retrieval (MLIR) systems is crucial for ensuring equitable access to information across diverse languages. This paper sheds light on the issue, based on the assumption that *queries in different languages, but with identical semantics, should yield equivalent ranking lists when retrieving on the same multilingual documents*. We evaluate the degree of fairness using both traditional retrieval methods, and a DPR neural ranker based on mBERT and XLM-R. Additionally, we introduce 'LaKDA', a novel loss designed to mitigate language biases in neural MLIR approaches. Our analysis exposes intrinsic language biases in current MLIR technologies, with notable disparities across the retrieval methods, and the effectiveness of LaKDA in enhancing language fairness.

1 Introduction

Information retrieval (IR) is the process of obtaining relevant information from a large collection of data according to a user's information needs. This information may exist in various formats, including text documents, images, or videos. Conventionally, the collection is a corpus of text documents, and user information needs are expressed in plain text queries. IR serves as a foundational technology in numerous NLP applications including question-answering systems (Abbasiyantaeb and Momtazi, 2020; Chen et al., 2017), and is also assuming an increasingly pivotal role in supporting the advancement of Large Language Models (LLMs) for text understanding and knowledge inference (Miao et al., 2024; Zhu et al., 2024).

Multilingual information retrieval (MLIR) entails queries being in different languages, with the results for a query in a given language being across multiple languages (including the source language of the query). MLIR has particular importance as it enables (multilingual) users to access information that may be unavailable or limited in their native language, thereby fostering cultural and linguistic diversity.

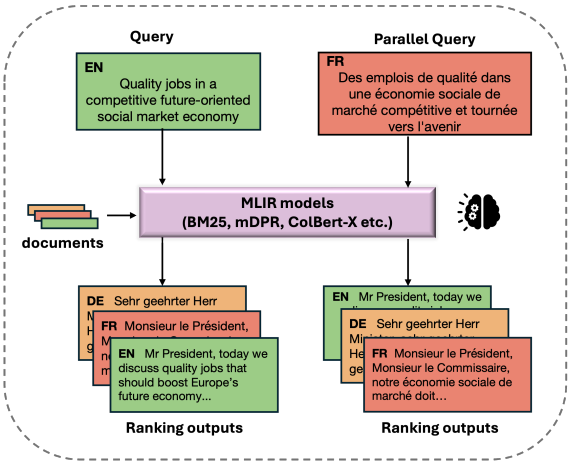

Figure 1: The case of language bias studied in this work. Semantically parallel queries retrieve the same documents, but the ranking outputs are inconsistent.

Research has shown that MLIR systems often exhibit biases towards certain languages due to factors like morphological complexity (Park et al., 2021) and resource availability (Lawrie et al., 2023; Huang et al., 2023). For instance, Lawrie et al. (2023) found that documents in higher-resource languages tend to be ranked higher in MLIR. This phenomenon is particular notable when the models are built upon multilingual pretrained models (Yang et al., 2024).

Another case of language bias in MLIR is shown in Figure 1. Given semantically equivalent queries in different languages and the same documents, we are interested in determining the consistency of the obtained ranking lists. This forms the basis for our investigation of language fairness in MLIR from a query-level perspective.

Our study compares MLIR methods using semantically equivalent queries in 24 European languages, which we use to search a fixed multilingual

document collection. These parallel query sets are from the original dataset, not machine-translated, and are based on human-annotated document tags. In repurposing them as queries and using the tags as relevance judgements, we fashion a multilingual IR collection with massively-multilingual parallel query sets.

Our work makes four main contributions:[1]

1. **Novel evaluation metric for fairness under ranking:** we propose the mean rank correlation (MRC) score to evaluate language fairness under MLIR, based on the premise that semantically-equivalent queries in different languages should yield consistent document rankings.
2. **Novel dataset:** we develop the MultiEup-v2 dataset, consisting of semantically parallel queries and multilingual documents, along with demographic attributes. This dataset serves as a benchmark for future fairness research in MLIR.
3. **Quantification of language (un)fairness:** we analyze language fairness in MLIR across different languages and language families, and find that BM25 exhibits larger language bias than neural retrieval frameworks like mDPR. Additionally, higher-resource languages tend to be associated with higher degrees of language fairness.
4. **Proposal of a new ranking bias mitigation method:** we propose the language KL-divergence alignment (LaKDA) loss to mitigate language bias in MLIR, demonstrating its effectiveness within the mDPR neural retrieval framework with multilingual text encoders mBERT and XLM-R.

2 Language Bias in MLIR

In this section, we examine language bias in MLIR. First, we introduce a novel metric for quantifying language fairness, our evaluation benchmark, and introduce a method for mitigating language bias.

2.1 MLIR Language Fairness Metric

We define fairness in MLIR as follows: queries in different languages but with identical semantics should yield equivalent ranking lists when executed against the same multilingual document collection.

[1]The dataset and code are available from https://github.com/jrnlp/MLIR_language_bias under an Apache 2.0 license.

Assume we have L languages and N queries for each language. For language pair $a, b \in \{\ell_1, \ell_2, \ldots, \ell_L\}$, let:

$$Q_a = \{q_{(1,a)}, q_{(2,a)}, \ldots, q_{(N,a)}\}$$

$$Q_b = \{q_{(1,b)}, q_{(2,b)}, \ldots, q_{(N,b)}\}$$

represent the sets of all queries in languages a and b, respectively, where $q_{(i,a)}$ is the i-th query in language a and $q_{(i,b)}$ is the i-th semantically parallel query in language b.

Assume a ranking method π produces a ranked result list $R(q_{(i,a)}, D)$ when given query $q_{(i,a)}$ against document collection D. Then for each query i and pair of languages (a, b), we compute the ranking correlation $\text{RC}^i_{(a,b)}$ between the ranking lists $R(q_{(i,a)}, D)$ and $R(q_{(i,b)}, D)$ using Spearman's rank correlation (Oakes, 2010; Spearman, 1904):

$$\text{RC}^i_{(a,b)} = \text{corr}(R(q_{(i,a)}, D), R(q_{(i,b)}, D))$$

Next, we compute the average correlation for language a with query i with the other $L-1$ language pairs, denoted as $\text{RC}^i_{(a)}$:

$$\text{RC}^i_{(a)} = \frac{1}{(L-1)} \sum_{1 \le a < b \le L} \text{RC}^i_{(a,b)}$$

The overall mean correlation score (MRC) for a specific language a among L languages with N queries is:

$$\text{MRC@k}_{(a)} = \frac{1}{N} \sum_{i=1}^{N} \text{RC}^i_{(a)}$$

The MRC@k represents the average degree of consistency between ranking lists for semantically identical queries across all language pairs in the top-k results. A higher MRC@k value indicates greater fairness, reflecting a higher degree of equivalence in the search results across different languages.

2.2 Mitigation Language Bias Methodology

Figure 2 demonstrates our co-training MLIR model framework with two losses. Section 2.2.1 introduces the unitized Dense Passage Retrieval (DPR) loss for IR, and in Section 2.2.2 we propose Language KL-Divergence Alignment (LaKDA) loss to improve language fairness.

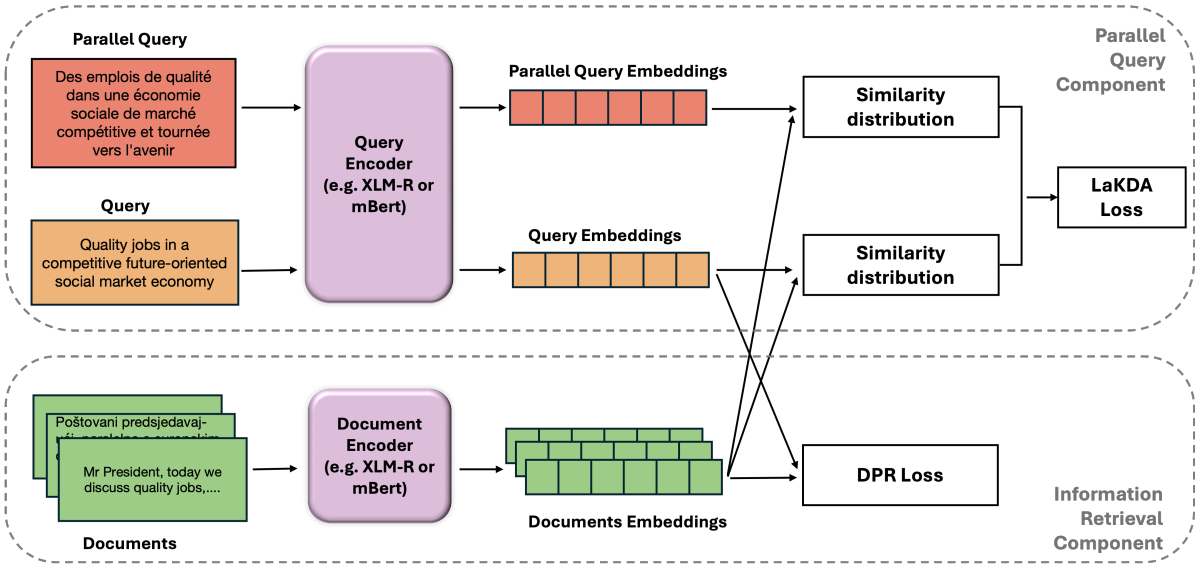

Figure 2: Our model framework contains two parts: the IR component, and the parallel query component. For the IR part, we adopt a DPR module for retrieval with DPR loss. For the parallel query part, we use the LaKDA loss to improve MLIR language fairness.

2.2.1 DPR Loss

Dense passage retrieval (Karpukhin et al., 2020) is a neural retrieval framework initially proposed for monolingual supervised fine-tuning. This architecture separately encodes queries and documents into dense vectors, optimizing their alignment through a contrastive loss. The goal is to maximize the similarity between queries and their relevant documents while minimizing it with irrelevant documents.

Assume we have a query q and a collection of documents $D = \{d_1^-, d_2^+, d_3^-, \ldots, d_M^-\}$, where d_i^+ indicates a relevant document and d_j^- an irrelevant document.

Let $\mathbf{q}$ be the dense vector representation of the query, and $\mathbf{d}_i^+$ and $\mathbf{d}_j^-$ be dense vector representations of the corresponding documents.

The similarity between the query and each document is computed using the dot product: $\text{sim}(q, d_i^+) = \mathbf{q} \cdot \mathbf{d}_i^{+\intercal}$ and $\text{sim}(q, d_j^-) = \mathbf{q} \cdot \mathbf{d}_j^{-\intercal}$.

We then define the loss to be the negative log-likelihood of the positive documents' similarity scores among all documents:

$$\mathcal{L}_{\text{DPR}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(q_i, d_i^+))}{Z_i}$$

$$Z_i = \exp(\text{sim}(q_i, d_i^+)) + \sum_{m=1}^{M} \exp(\text{sim}(q_i, d_{i,m}^-)).$$

This contrastive loss formulation ensures that the query embedding is closer to the positive document embedding than to any of the negative document embeddings, thereby enhancing the model's retrieval performance.

2.2.2 LaKDA Loss

To further mitigate language bias in MLIR, we add a Kullback-Leibler (KL) divergence term to measure the similarity of the distribution of retrieval scores between the original and parallel-language queries over a shared set of document embeddings.

For each query $q(i, \ell_a)$ and its parallel query $q(i, \ell_b)$, we compute their similarity distributions over the document embeddings as follows:

1. **Compute Similarity Scores:**

   For the original query $q(i, \ell_a)$ and the parallel query $q(i, \ell_b)$:

   $$\text{sim}(i, \ell) = \mathbf{q}(i, \ell) \cdot \mathbf{D}^{\intercal} \quad \text{where} \quad \ell \in \{\ell_a, \ell_b\}$$

2. **Transform to Probability Distributions:**

   The similarity scores are transformed into probability distributions using the softmax function:

   $$\mathbf{p}(i, \ell) = \frac{\exp(\text{sim}(i, \ell))}{\sum_{j=1}^{M} \exp(\text{sim}(i, \ell)_j)}$$

3. **KL Divergence Calculation:**

   The KL divergence between the similarity distributions of the original and parallel queries

is defined as:

$$D_{\text{KL}}(\mathbf{p}(i,\ell_b) \parallel \mathbf{p}(i,\ell_a)) = \sum_{j=1}^{M} \mathbf{p}(i,\ell_b)_j \log\left(\frac{\mathbf{p}(i,\ell_b)_j}{\mathbf{p}(i,\ell_a)_j + \epsilon}\right)$$

where ϵ is a small constant to avoid taking the log of zero.

4. **Overall LaKDA Loss:**

   The LaKDA Loss for all N queries is the mean of KL Divergence over all queries:

$$\mathcal{L}_{\text{LaKDA}} = \frac{1}{N}\sum_{i=1}^{N} D_{\text{KL}}(\mathbf{p}(i,\ell_b) \parallel \mathbf{p}(i,\ell_a))$$

Finally, to balance information retrieval performance and language fairness, we define a joint loss function as a weighted combination of the DPR loss $\mathcal{L}_{\text{DPR}}$ and the LaKDA loss $\mathcal{L}_{\text{LaKDA}}$:

$$\mathcal{L} = (1-\alpha)\mathcal{L}_{\text{DPR}} + \alpha\mathcal{L}_{\text{LaKDA}} \quad (1)$$

where α is a tunable hyperparameter.

2.3 MLIR Language Fairness Benchmark

Overview The European Parliament (EP) serves as a crucial forum for political debate and decision-making in the European Union. During debates, Members of the European Parliament (MEPs) discuss topics in their own languages, and debates are then transcribed in the original languages, and indexed with multilingual topics.

We constructed MultiEuP-v2 by expanding MultiEuP (Yang et al., 2023), taking the debate titles as queries, and individual MEP speeches in a given debate as documents. The documents are multilingual, encompassing 24 languages from 8 language families. Each query has parallel versions in all 24 languages, sourced from the original dataset. Additionally, we collected the basic demographic details of each of the MEPs, making it the perfect target for the study of fairness in an IR context, in terms of both language and other protected attributes.

Dataset Statistics We partition the dataset into mutually-exclusive train/dev/test sets to ensure that the queries and documents in the three sets are distinct. Table 1 details the statistics of the dataset. The number of unique queries is counted per language; i.e., for the dev and test sets, each language has 100 queries, with parallel versions across all 24 languages. The document collection is also made up of documents from all 24 languages. Table 6 in the Appendix shows the language distribution, with languages such as English (EN), German (DE), and French (FR) making up over 50% of the dataset in terms of document count.

	# Documents	# Unique Queries
Train	44,961	1,623
Dev	2,787	100
Test	2,589	100

Table 1: Data size and unique query IDs in train, dev, and test sets. The number of unique query IDs represents the counts for each language.

3 Experiments and Findings

Our language fairness experiment consists of two main parts: the detection and comparison of language bias among different ranking methods (Section 3.1) and the mitigation of fairness bias (Section 3.2).

3.1 Language Bias Detection

3.1.1 Detection Experiment Setting

We used the MultiEuP-v2 dataset in a *many-vs-many* setting for traning, where both queries and documents are multilingual to ensure language diversity. For evaluation, we adopted a parallel *one-vs-many* approach, with queries in one language and documents in multiple languages, enabling parallel comparison across different languages.

3.1.2 Detection Experiment Models

BM25 We implemented BM25, a commonly used traditional information retrieval baseline, using Pyserini (Lin et al., 2021). Pyserini is built upon Lucene (Yang et al., 2017). We used the default settings ($k_1 = 0.9$ and $b = 0.4$) and language-specific analyzers.

DPR Our neural IR approach is based on DPR and uses a bi-directional encoder to encode queries and documents separately. We compare DPR performance over two text encoders: mBERT with *bert-base-multilingual-uncased*, and XLM-R with *xlm-roberta-base*. In each case, the batch size was set to 96 and the learning rate was 5e-5, with each epoch taking approximately 40 minutes on a single Tesla V100 GPU.

MRR@100	Germanic					Romance					Slavic							Uralic			Baltic		Hellenic	Semitic	Celtic	
	EN	DE	NL	SV	DA	FR	ES	RO	IT	PT	PL	HR	BG	SK	SL	CS	HU	FI	ET	LT	LV	EL	MT	GA	Avg
BM25	87.6	59.8	29.6	25.5	21.4	59.6	58.2	33.9	51.7	49.7	39.9	33.4	22.6	30.9	32.5	28.2	22.9	20.3	19.6	22.8	21.6	18.1	12.6	16.5	34.1
mBERT																									
$\mathcal{L}_{\text{DPR}}$	42.8	39.4	37.1	36.3	33.9	38.3	41.0	38.3	39.2	40.1	39.9	37.9	36.9	39.0	39.7	37.9	32.5	32.2	30.7	35.8	33.7	30.6	27.0	14.5	35.6
$+\mathcal{L}_{\text{MSE}}$	29.1	28.7	25.8	22.5	26.5	26.2	27.0	26.2	26.9	26.8	26.9	25.9	22.8	25.7	26.0	26.3	22.7	23.7	23.3	20.0	22.9	16.3	15.6	12.2	24.0 (↓ 32.6%)
$+\mathcal{L}_{\text{LaKDA}}$	46.5	47.5	43.0	43.3	40.7	45.7	42.6	41.5	44.4	41.4	42.3	42.6	39.7	43.3	39.4	43.8	39.5	42.3	37.1	38.1	39.5	31.7	28.1	16.8	40.0 (↑ 12.4%)
XLM-R																									
$\mathcal{L}_{\text{DPR}}$	47.6	49.7	45.6	49.4	45.6	48.7	51.2	51.7	45.4	47.1	47.1	45.7	51.0	50.0	43.5	49.8	43.7	47.1	44.3	46.1	49.4	46.5	**40.6**	30.0	46.5
$+\mathcal{L}_{\text{MSE}}$	48.4	46.4	53.9	50.2	54.1	60.8	58.5	58.0	50.1	45.6	51.7	46.1	52.7	50.6	45.9	51.7	50.2	48.1	43.1	41.4	48.7	47.5	28.8	24.7	48.2 (↑ 3.7%)
$+\mathcal{L}_{\text{LaKDA}}$	**70.0**	**65.3**	**65.6**	**68.8**	**69.1**	**69.0**	**62.0**	**66.0**	**67.3**	**60.7**	**69.1**	**57.3**	**62.2**	**61.4**	**64.4**	**63.4**	**62.6**	**59.2**	**56.7**	**61.5**	**60.7**	**55.4**	34.9	**30.8**	**61.0 (↑ 31.2%)**

Table 2: The MLIR performance evaluation results on MultiEuP-v2. MRR@100 ($\times 100$) ranges from 0 to 100, where values closer to 100 indicate better performance. Underscore indicates the best performance for mBERT, and **bold** indicates the best performance for XLM-R. The symbol ↑/↓ indicates the percentage increase or decrease compared to the vanilla setting $\mathcal{L}_{DPR}$. All differences are significant at $p < 0.0005$. Note the broad similarities in results for a given language and also language family.

MRC@5	Germanic					Romance					Slavic							Uralic			Baltic		Hellenic	Semitic	Celtic	
	EN	DE	NL	SV	DA	FR	ES	RO	IT	PT	PL	HR	BG	SK	SL	CS	HU	FI	ET	LT	LV	EL	MT	GA	Avg
BM25	0.5	−1.0	1.4	−0.6	−0.5	−1.2	0.4	2.0	2.8	1.1	−0.3	1.8	3.2	−0.3	1.5	3.3	1.7	0.4	0.6	0.7	−2.4	1.0	−1.4	−0.5	0.6
mBERT																									
$\mathcal{L}_{\text{DPR}}$	12.9	15.0	15.9	15.2	12.8	12.9	15.4	15.1	14.4	13.7	14.6	15.2	15.6	15.5	13.8	16.0	7.3	13.0	10.3	10.6	11.3	12.9	10.2	5.3	13.1
$+\mathcal{L}_{\text{MSE}}$	14.3	15.6	12.2	14.5	14.6	15.5	13.5	18.3	17.1	15.5	15.5	14.4	10.7	15.5	14.3	17.2	13.2	14.2	12.1	12.1	12.7	6.5	13.7	7.7	13.8 (↑ 5.3%)
$+\mathcal{L}_{\text{LaKDA}}$	18.3	17.4	20.1	17.4	14.9	20.6	20.2	19.1	18.2	20.7	17.3	18.2	16.9	20.0	17.5	16.7	16.9	18.7	16.0	14.2	13.0	10.2	9.4	3.5	16.5 (↑ 25.6%)
XLM-R																									
$\mathcal{L}_{\text{DPR}}$	12.6	13.8	9.3	13.9	15.2	12.7	11.7	13.1	9.5	12.7	15.0	15.7	13.0	14.1	13.7	11.2	13.8	10.0	10.9	11.3	10.6	9.5	0.2	**7.4**	11.7
$+\mathcal{L}_{\text{MSE}}$	12.8	11.7	11.0	12.1	13.1	9.2	12.5	12.7	13.5	11.3	11.9	12.4	11.3	9.3	11.9	10.5	11.5	10.5	9.9	11.9	9.3	8.6	2.2	0.7	10.5 (↓ 10.3%)
$+\mathcal{L}_{\text{LaKDA}}$	**12.9**	**20.1**	**15.8**	**18.5**	**18.6**	**19.3**	**15.9**	**16.2**	**16.9**	**18.0**	**17.2**	**16.0**	**17.0**	**16.5**	**18.3**	**18.3**	**17.5**	**16.6**	**15.5**	**16.4**	**14.3**	**12.9**	**6.7**	6.2	**15.9 (↑ 35.9%)**

Table 3: The MLIR fairness evaluation results on MultiEuP-v2. MRC@5 ($\times 100$) ranges from -100 to 100, where values closer to 100 indicate better fairness. Underscore indicates the best fairness for mBERT, and **bold** indicates the best fairness for XLM-R.

3.1.3 Detection Evaluation and Findings

Performance Metrics To evaluate MLIR retrieval performance, we used the MRR@100 metric, which represents the Mean Reciprocal Rank for the top 100 documents (Radev et al., 2002; Voorhees and Tice, 2000). For a single query, the Reciprocal Rank (RR) is defined as $RR = \frac{1}{\text{rank}}$, where *rank* is the position of the highest-ranked relevant document. If no relevant document is returned, the RR is set to 0. For multiple queries N, the MRR is the mean of RRs (Yang et al., 2023).

$$MRR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\text{rank}_i}$$

Performance Findings Table 2 shows the MRR@100 results for semantically identical queries in different languages. The findings include: (1) for low-resource languages[2] like Maltese (MT) and Irish (GA), the MRR@100 is lower than high-resource languages; (2) interestingly, despite Maltese having more documents than Estonian (ET) in our dataset (Table 6), the MRR@100 disparity suggests that data augmentation alone does not eliminate the inherent bias in pre-trained IR models against low-resource languages; and (3) DPR with mBERT is slightly better overall than BM25, while DPR with XLM-R significantly outperforms both BM25 and DPR with mBERT.

Fairness Findings When we evaluate language fairness based on MRC@5 (see Section 2.1), we obtain the results shown in Table 7. The main findings are: (1) BM25 has lower language fairness than DPR; (2) similarly to MRR@100, low-resource languages (MT and GA) exhibit lower language fairness than high-resource languages; and (3) according to Figure 3, which shows the MRC@5 correlation between language pairs (noting that the results are symmetric), languages in the same language family (within the black squares) tend to have higher MRC scores, esp. for the Germanic, Romance, and Slavic language families (the dashed square).

3.2 Language Bias Mitigation

Next we turn to the question of language bias mitigation.

[2]Defined as those languages in Conneau et al. (2020) with less than 0.5 GiB in training data.

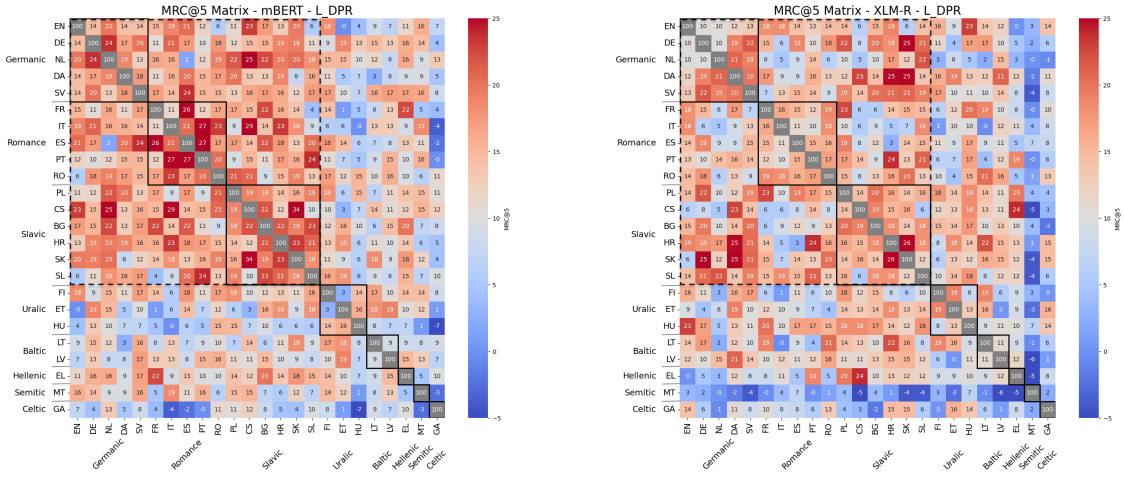

Figure 3: The MRC@5 matrix among parallel queries. The x-axis and y-axis both represent query languages.

3.2.1 Mitigation Experiment Setting

The training parameters and evaluation protocol and metrics used to measure language bias mitigation are consistent with those described in Section 3.1.

3.2.2 Mitigation Experiment Models

Vanilla Our vanilla setting is using only the DPR loss for MLIR (Karpukhin et al., 2020) and not incorporating any language fairness loss.

MSE Another baseline involves calculating the Mean Squared Error (MSE, Hastie et al. (2009)) between the embeddings of parallel queries to increase their similarity. We employ the same joint MSE loss with DPR loss.

LaKDA With our proposed LaKDA debiasing method (Section 2.2.2), for each query, we randomly sample a semantically identical query in a different language and compute the LaKDA loss, which is then jointly optimized with the DPR loss as shown in Equation (1). For both mBERT and XLM-R, we set $\alpha = 0.5$ for comparability (but return to investigate the hyperparameter sensitivity in Figure 4).

3.2.3 Mitigation Evaluation and Findings

Table 2 presents the IR performance (MRR@100), and Table 7 demonstrates language fairness (MRC@5) for the DPR framework with different pretrained multilingual models. Our observations are as follows:

mBERT Findings Compared to the vanilla setting (DPR only): (1) incorporating either MSE or LaKDA enhances language fairness (MRC@5) with mBERT, but LaKDA is more effective (25.6% vs. 5.1%); and (2) for MRR@100, LaKDA achieves an 11.3% improvement, whereas MSE loss reduces MRR@100 by 32.6%. Figure 4 also shows that during the hyperprameter α grid search, the DPR model with LaKDA loss is more robust in terms of MRR than MSE loss. This is because, unlike MSE loss, LaKDA loss considers not only the similarity between parallel queries but also their embedding similarity with documents, providing a better trade-off between fairness and performance.

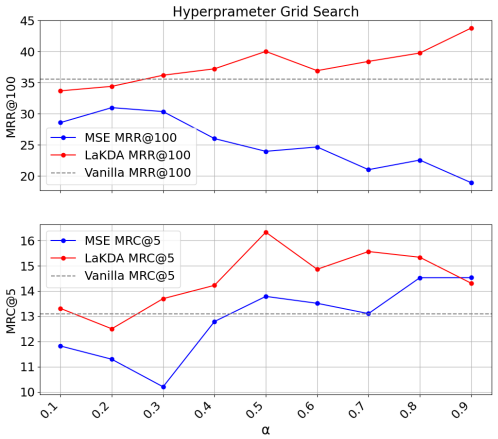

Figure 4: MSE and LaKDA sensitivity plot.

XLM-R Findings Compared to the vanilla setting (DPR only): (1) only LaKDA improves language fairness (MRC@5), by 35.9%, while MSE leads to a slight degradation; and (2) both MSE

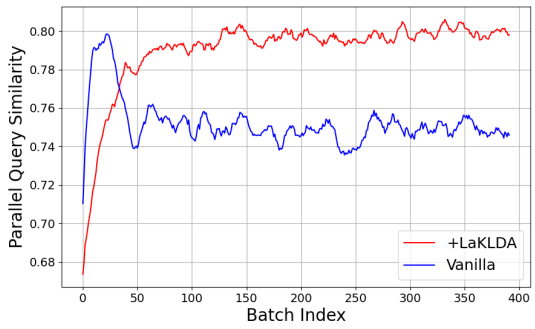

Figure 5: Parallel query similarity over training.

Parallel MT Query	MRR@100	MRC@5
Zero-shot	21.2	2.8
Translated	**36.2**	1.2
Original	34.9	**6.7**

Table 4: Maltese (MT) query MLIR results.

Parallel GA Query	MRR@100	MRC@5
Zero-shot	21.4	0.6
Translated	29.6	1.4
Original	**30.8**	**6.2**

Table 5: Irish (GA) query MLIR results.

and LaKDA improve IR performance (MRR@100), with increases of 3.7% and 16.6%, respectively. XLM-R not only achieves better IR performance but is also more robust. This observation aligns with other research, and is why XLM-R is more commonly used in MLIR (Hu et al., 2020; Conneau et al., 2020; Conneau and Lample, 2019).

4 Discussion

4.1 Improvement of Parallel Query Similarity

In our experimental setup, an important characteristic for enhancing language fairness is the increase in similarity of semantically parallel queries. We calculated the average parallel query similarity in each batch over training for mBERT, as depicted in Figure 5. We observe that with the addition of the LaKDA loss, the final stable value of parallel query similarity is higher compared to the vanilla setting. This result explains the enhancement in language fairness (MRC).

4.2 Effect of Size and Quality of Parallel Queries

To explore the impact of the number and quality of parallel queries on IR performance and language fairness, we selected queries in two languages, MT and GA, from the training dataset and conducted experiments under the following three settings:

Zero-shot: As low-resource languages, there is relatively little training data for MT and GA; we therefore excluded queries in MT and GA from the training dataset, keeping the other parallel queries unchanged, and then conducted the same training and evaluation settings.

Translation: Without the original MT and GA parallel queries, we translated English queries into MT and GA parallel queries using Google Translate.[3] The BLEU scores (Papineni et al., 2002) of the translation results compared to the original were 0.196 and 0.251, respectively.

Original: The original queries in MT and GA, as per the experiments in Section 3.2.

Findings: Tables 4 and 5 show the results for MT and GA, conducted on the XLM-R model with LaKDA loss. We found that:

1. The zeroshot setting had the worst MRR performance, indicating the importance of parallel queries.
2. The translated version serves as a silver-standard, with improvements in MRR compared to the zeroshot setting.
3. The original texts are the best choice, achieving the best MRR and MRC, demonstrating the value of our MultiEuP-v2 dataset in providing an original multilingual corpus.

4.3 Effect of Neural Retrieval Approaches

The MRC@5 results presented in Table 7 show more than a 20-fold disparity between BM25 and the neural retrieval ranker DPR, with scores of 0.6 and 11.7, respectively. To understand the underlying causes, we analyzed the top 100 ranking outputs from both methods. As shown in Figure 6, BM25's output document languages and query languages exhibit a strong correlation along the diagonal line, contributing to heightened language bias. Since BM25 is only able to retrieve documents containing keywords present within the query (Thakur et al., 2021) and suffers from lexical gap (Berger et al., 2000), resulting in high retrieval rates for documents in the same language as the query.

Meanwhile, DPR retrieves documents across different languages more effectively, with substantial

[3] https://translate.google.com/

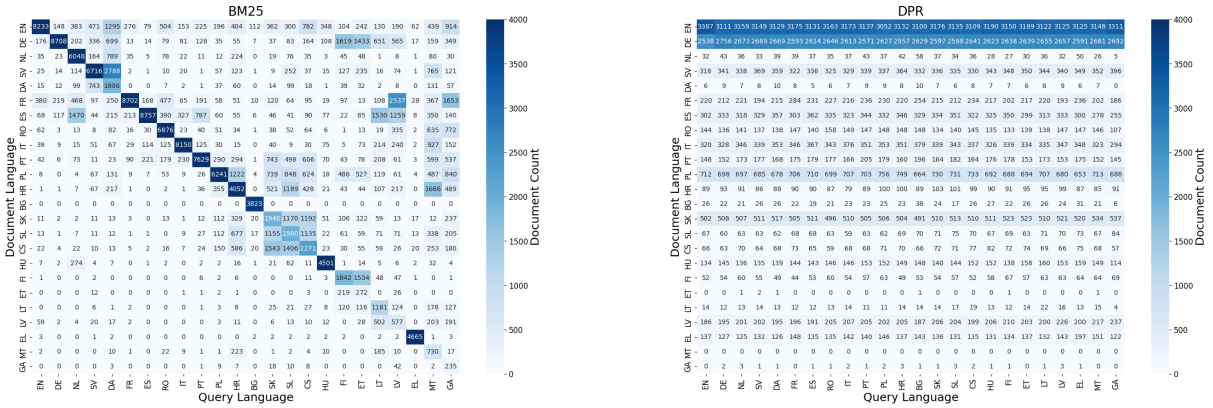

Figure 6: The correlation of query language with document language in top 100 ranking output.

off-diagonal values and reflecting the skewness of the dataset (see Table 6). This suggests that neural retrieval approaches can mitigate language bias to leveraging multilingual pre-trained models that understand semantic content regardless of the language.

5 Related Work

Fairness in Information Retrieval (IR) has been extensively studied through two primary dimensions: individual fairness and group fairness. These frameworks are crucial in ensuring equitable access to information, addressing concerns related to biases in ranking systems.

Individual fairness refers to the principle that similar items (in this case, documents) should be treated similarly (Biega et al., 2018; Dwork et al., 2011). In IR, this means that if two documents are equally relevant to a query, they should receive similar rankings. A violation of individual fairness occurs when two comparable documents are ranked differently due to irrelevant factors, such as their format or metadata. This concept is rooted in the idea of consistency and uniform treatment, ensuring that the system does not unfairly prioritize or penalize specific documents that are otherwise similar in content and relevance.

Group fairness, on the other hand, ensures that predefined groups (such as demographic groups or, in our case, languages) are treated equitably in the ranking process (Sapiezynski et al., 2019; Zehlike et al., 2022, 2017). The goal is to prevent bias against any group by ensuring that the system does not favor one group over another. In IR, this often translates to ensuring that documents associated with a protected group (e.g., underrepresented languages or communities) are not systematically ranked lower than those associated with unprotected groups. Group fairness frameworks attempt to mitigate historical and societal biases that might seep into the retrieval process, making sure that members of different groups have equitable access to information. In our work, we extend this concept to multilingual IR, treating each language as a group and ensuring that rankings are fair and consistent across languages.

Two key fairness metrics in group fairness that align with our work are Probability of Equal Expected Rank (PEER) and Attention Weighted Ranked Fairness (AWRF).

- **PEER** (Yang et al., 2024) is designed to ensure equity in ranking by guaranteeing that documents from different languages are treated equally when they are equally relevant. This metric is particularly valuable for multilingual retrieval, as it addresses the risk of language bias, ensuring that a document's rank does not depend on the language of the query if the content is of similar relevance across languages.

- **AWRF** (Sapiezynski et al., 2019) assesses group exposure by comparing how documents are distributed across ranked positions relative to a predefined target distribution. This metric focuses on ensuring that documents from all languages receive appropriate visibility within the top-ranked results, balancing relevance and fairness in exposure.

While these metrics primarily emphasize document-level fairness, our approach uniquely focuses on query-level fairness. In our context, we argue that the retrieval system should provide consistent performance across languages, ensuring

that the language of the query does not affect the user's ability to access relevant information. This promotes inclusivity, ensuring that users from different linguistic backgrounds experience similar outcomes when interacting with the system, ultimately fostering equal access to information.

6 Background Knowledge

MultiEuP The European Parliament (EP) serves as a crucial forum for political debate and decision-making in the European Union. During debates, Members of the European Parliament (MEPs) discuss topics in their own languages, and debates are then transcribed in the original languages, and indexed with multilingual topics. As such, the data is naturally occurring in the 24 official languages of the EU, and expertly transcribed and multilingually annotated. Additionally, we have access to basic demographic details of each of the MEPs, making it the perfect target for the study of fairness in an IR context, in terms of both language and protected attributes was crafted. The EU has published different language versions of all titles, providing semantically identical queries for investigating language fairness in MLIR.

An earlier version of the MultiEuP dataset was published in 2023 covering debates up to October 2022 (Yang et al., 2023). In this work, we have expanded the dataset using the same data collection and preprocessing procedures, to include debates up to 2024. This doubles the total data volume, and provides a sufficient sample size to research neural ranking methods. We additionally augment each document with comprehensive metadata of the author, including gender, nationality, political affiliation, and age, for use in exploring fairness with respect to protected attributes.

Unlike MLIR datasets such as mMARCO (Bonifacio et al., 2021), a multilingual version of the MS MARCO (Bajaj et al., 2016), that relies on machine translation, our benchmark queries and documents are original rather than translated versions. This reduces noise and ensures the linguistic authenticity of the corpus.

Another commonly used MLIR datasets Mr. TyDi (Zhang et al., 2021) and MIRACL (Zhang et al., 2023) , are actually mixed monolingual IR dataset, since they were structured such that queries in different languages are matched only with documents in the same language. This limits the comparability of results across different languages. Our benchmark addresses this limitation by introducing semantically parallel queries across multiple languages, enabling comprehensive analysis of language fairness in MLIR.

DPR Dense Passage Retrieval (DPR: Karpukhin et al. (2020)) is a neural retrieval framework initially proposed for monolingual supervised fine-tuning. DPR uses dual encoders: one for encoding queries and another for encoding passages (documents), both based on the BERT architecture (Devlin et al., 2019). The primary advantage of DPR over traditional retrieval models like BM25 is its ability to embed both queries and documents into a shared dense vector space, enabling efficient nearest-neighbor search for retrieval. The relevance of a document to a query is determined by the similarity between their embeddings, typically using the dot product as a similarity measure.

In our work, we employ mDPR using mBERT and XLM-R to handle multilingual queries and documents. These models are fine-tuned on parallel query-document pairs from multiple languages, allowing the system to generalize across different languages. The use of mDPR allows us to explore how multilingual language models handle language biases, which often favor high-resource languages over low-resource ones. Furthermore, we investigate the performance of these models on the MultiEuP dataset, assessing their ability to ensure fair and equitable retrieval across 24 languages, thus promoting fairness in multilingual IR.

7 Conclusion

We introduced a novel benchmark, MultiEup-v2, for investigating language fairness in multilingual information retrieval (MLIR) systems. Additionally, we proposed the mean rank correlation (MRC) score to assess language fairness in MLIR systems, which ensures that queries in different languages but with the same semantic meaning retrieve similar documents. Our findings indicate that the traditional IR method BM25 exhibits larger language biases than DPR with multilingual pretrained language models. Furthermore, we designed the language KL-divergence alignment (LaKDA) loss to mitigate language bias, and found that incorporating LaKDA loss into DPR improves language fairness substantially without sacrificing retrieval performance.

Ethics Statement

The dataset contains publicly-available EP data that does not include personal or sensitive information, with the exception of information relating to public officeholders, e.g., the names of the active members of the European Parliament, European Council, or other official administration bodies. The collected data is licensed under the Creative Commons Attribution 4.0 International licence.[4]

Limitations

Our investigation into language fairness in multilingual information retrieval (MLIR) is limited to European languages in this work. However, our approaches and evaluation methods are adaptable to other languages. Additionally, we focused exclusively on language fairness, leaving other dimensions of fairness in MLIR, such as demographic fairness, unexplored. We encourage the research community to conduct more comprehensive studies on fairness in MLIR, building upon the foundation of our benchmark.

Acknowledgement

We sincerely thank Trevor Cohn for his valuable suggestions and support in this work. This research was funded by the Melbourne Research Scholarship and was conducted using the LIEF HPC-GPGPU Facility hosted at the University of Melbourne. This facility was established with the assistance of LIEF Grant LE170100200.

References

Zahra Abbasiyantaeb and Saeedeh Momtazi. 2020. Text-based question answering from information retrieval and deep neural network perspectives: A survey. *CoRR*, abs/2002.06612.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Adam Berger, Rich Caruana, David Cohn, Dayne Freitag, and Vibhu Mittal. 2000. Bridging the lexical chasm: statistical approaches to answer-finding. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, page 192–199, New York, NY, USA. Association for Computing Machinery.

Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. 2018. Equity of attention: Amortizing individual fairness in rankings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18. ACM.

Luiz Henrique Bonifacio, Israel Campiotti, Roberto de Alencar Lotufo, and Rodrigo Frassetto Nogueira. 2021. mmarco: A multilingual version of MS MARCO passage ranking dataset. *CoRR*, abs/2108.13897.

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021. MultiEURLEX - a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6974–6996, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

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.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*.

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.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. 2011. Fairness through awareness. *Preprint*, arXiv:1104.3913.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

[4] https://eur-lex.europa.eu/content/legal-notice/legal-notice.html

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of the 37th International Conference on Machine Learning*, pages 4411–4421.

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*, pages 1208–1218.

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.

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

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. `https://github.com/castorini/pyserini`.

Jing Miao, Charat Thongprayoon, Supawadee Suppadungsuk, Oscar Garcia Valencia, and Wisit Cheungpasitporn. 2024. Integrating retrieval-augmented generation with large language models in nephrology: Advancing practical applications. *Medicina*, 60:445.

John G. Oakes. 2010. Commentary: Charles spearman and correlation: a commentary on 'the proof and measurement of association between two things'. *International Journal of Epidemiology*, 39(5):1151–1160.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Hyunji Hayley Park, Katherine J. Zhang, Coleman Haley, Kenneth Steimel, Han Liu, and Lane Schwartz. 2021. Morphology matters: A multilingual language modeling analysis. *Transactions of the Association for Computational Linguistics*, 9:261–276.

Dragomir R. Radev, Hong Qi, Harris Wu, and Weiguo Fan. 2002. Evaluating web-based question answering systems. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Piotr Sapiezynski, Wesley Zeng, Ronald E. Robertson, Alan Mislove, and Christo Wilson. 2019. Quantifying the impact of user attention on fair group representation in ranked lists. *Preprint*, arXiv:1901.10437.

Charles Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *CoRR*, abs/2104.08663.

Ellen M. Voorhees and Dawn M. Tice. 2000. The TREC-8 question answering track. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Eugene Yang, Thomas Jänich, James Mayfield, and Dawn Lawrie. 2024. Language fairness in multilingual information retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2024, page 2487–2491. ACM.

Jinrui Yang, Timothy Baldwin, and Trevor Cohn. 2023. Multi-EuP: The multilingual European parliament dataset for analysis of bias in information retrieval. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 282–291, Singapore. Association for Computational Linguistics.

Peilin Yang, Hui Fang, and Jimmy Lin. 2017. Anserini: Enabling the use of lucene for information retrieval research. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1253–1256.

Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. Fa*ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17. ACM.

Meike Zehlike, Tom Sühr, Ricardo Baeza-Yates, Francesco Bonchi, Carlos Castillo, and Sara Hajian. 2022. Fair top-k ranking with multiple protected groups. *Inf. Process. Manage.*, 59(1).

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

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

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zhicheng Dou, and Ji-Rong Wen. 2024. Large language models for information retrieval: A survey. *Preprint*, arXiv:2308.07107.

A Appendix

Language	ISO code	Countries where official lang.	Language Family	Total Usage	# Docs	Words per Doc
English	EN	United Kingdom, Ireland, Malta	Germanic	51%	14086	271/192
German	DE	Germany, Belgium, Luxembourg	Germanic	32%	5861	183/168
French	FR	France, Belgium, Luxembourg	Romance	26%	5313	267/210
Italian	IT	Italy	Romance	16%	3378	191/176
Spanish	ES	Spain	Romance	15%	4621	228/195
Polish	PL	Poland	Slavic	9%	2857	150/142
Romanian	RO	Romania	Romance	5%	1482	183/172
Dutch	NL	Netherlands, Belgium	Germanic	5%	1642	180/166
Greek	EL	Greece, Cyprus	Hellenic	4%	1104	180/171
Hungarian	HU	Hungary	Uralic	3%	979	131/131
Portuguese	PT	Portugal	Romance	3%	2185	183/169
Czech	CS	Czech Republic	Slavic	3%	913	155/143
Swedish	SV	Sweden	Germanic	3%	1038	168/154
Bulgarian	BG	Bulgaria	Slavic	2%	737	190/171
Danish	DA	Denmark	Germanic	1%	498	206/191
Finnish	FI	Finland	Uralic	1%	564	115/111
Slovak	SK	Slovakia	Slavic	1%	698	158/157
Lithuanian	LT	Lithuania	Baltic	1%	250	145/125
Croatian	HR	Croatia	Slavic	<1%	995	175/162
Slovene	SL	Slovenia	Slavic	<1%	549	188/158
Estonian	ET	Estonia	Uralic	<1%	88	167/162
Latvian	LV	Latvia	Baltic	<1%	176	128/113
Maltese	MT	Malta	Semitic	<1%	243	151/148
Irish	GA	Ireland	Celtic	<1%	80	179/163

Table 6: MultiEuP-v2 statistics, broken down by language: ISO language code; EU member states using the language officially; language family; proportion of the EU population speaking the language (Chalkidis et al., 2021); number of debate speech documents; and words per document (mean/median).

Recall@100	Germanic					Romance					Slavic						Uralic			Baltic		Hellenic	Semitic	Celtic	
	EN	DE	NL	SV	DA	FR	ES	RO	IT	PT	PL	HR	BG	SK	SL	CS	HU	FI	ET	LT	LV	EL	MT	GA	Avg
BM25	77.7	75.5	77.7	75.5	75.5	68.1	75.5	76.6	77.7	76.6	74.5	77.7	75.5	76.6	75.5	74.5	75.5	74.5	75.5	77.7	76.6	76.6	75.5	62.8	75.2
mBERT																									
$\mathcal{L}_{\text{DPR}}$	88.3	89.4	88.3	87.2	88.3	89.4	89.4	88.3	90.4	87.2	88.3	88.3	87.2	88.3	86.2	87.2	85.1	86.2	86.2	88.3	86.2	86.2	85.1	73.4	87.0
$+\mathcal{L}_{\text{MSE}}$	74.5	72.3	72.3	71.3	72.3	66.0	72.3	72.3	73.4	73.4	71.3	73.4	71.3	72.3	71.3	69.1	70.2	72.3	72.3	69.1	70.2	71.3	67.0	60.6	70.9
$+\mathcal{L}_{\text{LaKDA}}$	77.7	78.7	76.6	76.6	77.7	79.8	79.8	81.9	78.7	78.7	78.7	79.8	77.7	76.6	75.5	77.7	77.7	77.7	78.7	76.6	76.6	76.6	73.4	72.3	77.6
XLM-R																									
$\mathcal{L}_{\text{DPR}}$	86.2	89.4	86.2	84.0	86.2	85.1	89.4	89.4	86.2	87.2	86.2	84.0	87.2	88.3	90.4	86.2	89.4	90.4	84.0	80.9	86.2	86.2	88.3	81.9	86.6
$+\mathcal{L}_{\text{MSE}}$	91.5	92.6	90.4	86.2	88.3	69.1	91.5	91.5	90.4	91.5	90.4	90.4	90.4	91.5	88.3	92.6	91.5	88.3	87.2	88.3	91.5	90.4	87.2	78.7	88.7
$+\mathcal{L}_{\text{LaKDA}}$	93.6	96.8	93.6	93.6	93.6	67.0	94.7	94.7	92.6	96.8	95.7	96.8	95.7	94.7	92.6	92.6	94.7	92.6	95.7	93.6	93.6	92.6	89.4	75.5	92.2

Table 7: The MLIR additional evaluation results on MultiEuP-v2. Recall@100 ($\times$100) ranges from 0 to 100, where values closer to 100 indicate better performance.