SHIFT: Semantic Harmonization via Index-side Feature Transformation for Multilingual Information Retrieval

Youngjoon Jang[1], **Seongtae Hong**[1], **Hyeonseok Moon**[2*], **Heuiseok Lim**[1*]
[1]Department of Computer Science and Engineering, Korea University
[2]Sookmyung Women's University
{dew1701, ghdchlwls123, limhseok}@korea.ac.kr
hyns.moon@sookmyung.ac.kr

Abstract

With the rapid expansion of massive multilingual corpora, Multilingual Information Retrieval (MLIR) has emerged as a critical technology for global information access. MLIR enables users to retrieve semantically relevant documents from multilingual text collections using a single-language query. However, recent multilingual dense retrieval models often exhibit a strong preference for documents in the same language as the query. This leads to severe language bias, where top-ranked results are dominated by documents of specific languages, even when documents in other languages contain more semantically relevant information. To address this issue, we propose SHIFT, a training-free method applicable in the indexing stage. Specifically, SHIFT utilizes parallel translation pairs to estimate a relative language vector for each target language with respect to a source language. Subsequently, SHIFT corrects the language-specific offset by subtracting this relative language vector from document embeddings during indexing. Our comprehensive evaluation across four MLIR benchmarks and diverse dense retrieval models confirms that SHIFT can effectively mitigate language bias and enhance MLIR performance. The code for SHIFT is available at https://github.com/yjoonjang/SHIFT

1 Introduction

Information Retrieval (IR) constitutes the backbone of modern information systems, from traditional search engines to Retrieval-Augmented Generation (RAG) (Singhal et al., 2001; Kobayashi and Takeda, 2000; Lewis et al., 2020). With the rapid expansion of massive multilingual corpora, Multilingual Information Retrieval (MLIR) has become essential for equitable information access. MLIR enables users to retrieve relevant information from multilingual text collections using a single-language query (Peters and Braschler, 2002a). In real-world applications, such as global enterprise search and multilingual RAG, robust MLIR is essential to supply LLMs with diverse multilingual knowledge, preventing them from generating answers based solely on particular languages.

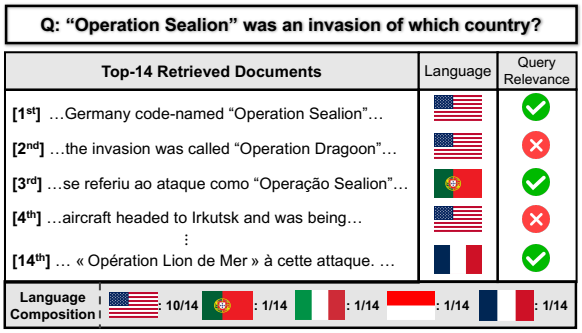

Figure 1: An illustration of language bias in MLIR using *multilingual-e5-large* on the Belebele dataset. The symbols ✓ and × denote semantic relevance. Despite the existence of semantically equivalent ground-truth documents across 14 languages, the top retrieved results are dominated by English documents (10/14).

Unlike Cross-Lingual Information Retrieval (CLIR), which assumes a one-to-one language mapping (Goswami et al., 2021), MLIR presents a more complex one-to-many challenge: simultaneously retrieving relevant documents across multiple target languages. Consequently, naive CLIR adaptations such as query translation, fail to perform effectively across massive multilingual corpora (Granell, 2014; Peters et al., 2012).

Ideally, an MLIR system should rank documents solely based on semantic similarity, invariant to linguistic surface forms. However, state-of-the-art multilingual dense retrievers suffer from severe language bias (Yang et al., 2024c; Hong et al., 2026). Rather than prioritizing semantic relevance, these models exhibit a disproportionate preference for documents written in the query language (Park and Lee, 2025; Sharma et al., 2025). As illustrated in

*Corresponding authors

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 2247–2269
October 24-29, 2026 ©2026 Association for Computational Linguistics

Figure 1, retrieving with *multilingual-e5-large* on the Belebele dataset reveals a pronounced language skew. Despite the existence of semantically equivalent ground-truth documents across 14 languages, the top ranks for an English query are saturated with semantically irrelevant English documents (10 out of 14). This indicates that language identity dominates semantic signals in the embedding space, distorting MLIR ranking behavior.

Conventional metrics often mask this bias by predominantly rewarding source-language retrievals. To explicitly quantify this, we introduce Target-Languages Recall@k (TLR@k), a novel metric measuring the retrieval effectiveness of non-query language documents. Furthermore, to mitigate this bias, we propose SHIFT, a training-free method applied during the indexing stage. By estimating a relative language vector from parallel translation pairs and subtracting it from document embeddings, SHIFT corrects the language-specific offset and forces the model to rely purely on underlying semantic signals. In our experiments, SHIFT yields consistent improvements, achieving up to a 16.4% relative gain in TLR@20. The main contributions of this work are as follows:

- We demonstrate that SHIFT is a universally effective solution, consistently improving MLIR performance and mitigating language bias across diverse dense retrieval models.
- We introduce Target-Languages Recall@k (TLR@k), a novel metric that explicitly quantifies language bias by measuring the retrieval of non-query language documents.
- Through detailed analysis and experiments, we provide empirical evidence that SHIFT successfully realigns multilingual documents based on semantics, correcting the skewed geometry of the embedding space.

2 Preliminaries

2.1 Related Works

The concept of "multilingual" in Information Retrieval (IR) can be interpreted in various contexts. For instance, multilingual IR is often referred to as monolingual retrieval across multiple languages (Hull and Grefenstette, 1996; Blloshmi et al., 2021) or as CLIR tasks involving multiple languages (Braschler, 2002a,b; Lawrie et al., 2022; Mitamura et al., 2008). In this work, however, we adopt the definition of MLIR established by the Cross-Language Evaluation Forum (CLEF): a task where a user retrieves information from a mixed-language document collection using a query in a single language (Peters and Braschler, 2002b).

Traditional MLIR approaches relied on Machine Translation (MT) to translate either the entire document collection or the query (Darwish and Oard, 2003; Kraaij et al., 2003; McNamee and Mayfield, 2002; Granell, 2014; Magdy and Jones, 2011; Peters et al., 2012; Si et al., 2008; Tsai et al., 2008). However, these methods incur significant computational costs or introduce systematic ranking bias due to disparate score distributions (Huang et al., 2023; Lawrie et al., 2023).

Recently, dense retrieval models based on multilingual Pre-trained Language Models (mPLMs) have become a standard, demonstrating robust performance in MLIR tasks (Conneau et al., 2020; Chen et al., 2024b; Yu et al., 2024). In MLIR scenarios, dense retrievers encode both queries and documents into a shared embedding space, enabling efficient retrieval via Approximate Nearest Neighbor (ANN) search using scoring functions such as dot product or cosine similarity (Arya et al., 1998; Liu et al., 2004; Dehak et al., 2010).

2.2 Language Bias in MLIR

However, the embedding spaces of dense retrievers fail to achieve perfect linguistic alignment. Consequently, documents representing identical meanings but written in different languages are often mapped to distant locations, leading to language bias that distorts the ranking (Yang et al., 2024c,a). An ideal MLIR system should rank documents solely based on semantic similarity, independent of the document's language. Yet, in practice, a recurring phenomenon is observed where documents in the same language as the query disproportionately occupy the top ranks (Park and Lee, 2025; Sharma et al., 2025). This bias increases the exposure of specific languages within multilingual collections, simultaneously undermining the fairness of information access and user experience.

To mitigate this issue, prior research has largely adopted two approaches. The first involves strengthening the model's linguistic alignment through additional training (Hu et al., 2023; Yang et al., 2024c; Huang et al., 2023; Yang et al., 2024b). While effective, this approach requires high computational costs due to large-scale training and necessitates retraining whenever the language combi-

nation of the model or data changes. The second approach addresses language bias at the pipeline level without modifying the model. A representative line of work translates each query into multiple languages and fuses language-specific ranked lists (Lawrie et al., 2025a; Yang et al., 2025). Although effective, such pipelines typically increase system complexity and inference-time latency. Another line mitigates bias by post-hoc normalization of multilingual embeddings (e.g., language-wise centering (Libovický et al., 2020) or projection-based methods (Yang et al., 2021)) without any additional training. While these strategies can improve retrieval performance, they may inadvertently suppress semantic signal when enforcing language-agnostic representations, and often require additional query-side processing, increasing inference-time latency.

3 SHIFT

In this work, we propose SHIFT, a training-free method designed to align multilingual representations at the indexing stage. Unlike approaches that necessitate computationally expensive training or impose additional inference latency, SHIFT employs a linear displacement strategy to correct spatial misalignment. This method preserves the high-dimensional structure of the original embeddings to the greatest extent possible, thereby improving MLIR performance with zero additional query-time overhead.

3.1 Estimating Relative Language Vectors

Let $\mathcal{L}_{\text{tgt}}$ denote the set of target languages, and let ℓ_{tgt} be an element of $\mathcal{L}_{\text{tgt}}$. The core mechanism of SHIFT is to estimate a relative language vector $V_{\ell_{\text{tgt}}}$ for each target language, with respect to a source language ℓ_{src}. To estimate these vectors, we utilize parallel documents consisting of source–target translation pairs from mMARCO (Bonifacio et al., 2022). For each target language ℓ_{tgt}, we collect a set of translation pairs:

$$\mathcal{P}_{\ell_{\text{tgt}}} = \left\{(D_i^{\ell_{\text{src}}},\, D_i^{\ell_{\text{tgt}}})\right\}_{i=1}^{N}, \quad \ell_{\text{tgt}} \in \mathcal{L}_{\text{tgt}}$$

Each pair consists of a source-language document $D_i^{\ell_{\text{src}}}$ and its translation $D_i^{\ell_{\text{tgt}}}$, sharing identical semantic meaning. We first encode each document into the embedding space using the dense retrieval model $f(\cdot)$:

$$z_i^{\ell_{\text{src}}} = f(D_i^{\ell_{\text{src}}}), \quad z_i^{\ell_{\text{tgt}}} = f(D_i^{\ell_{\text{tgt}}})$$

Algorithm 1 SHIFT Indexing

Require: Labeled Documents $\mathcal{D} = \{(d_j, \ell_j)\}_{j=1}^{M}$, Model f, Source Language ℓ_{src}, Target Language Set $\mathcal{L}_{\text{tgt}}$, Relative Language Vectors $\{V_\ell\}_{\ell \in \mathcal{L}_{\text{tgt}}}$, Scale factor $\alpha \in [0, 1]$

```
for each (d, ℓ) ∈ D do
    z ← f(d)
    if ℓ = ℓ_src then
        z̃ ← z
    else
        z̃ ← z − α V_ℓ
    end if
    Add z̃ to the index
end for
```

The difference between the two embeddings, $\Delta_i^{\ell_{\text{tgt}}} = z_i^{\ell_{\text{tgt}}} - z_i^{\ell_{\text{src}}}$, approximates the representational offset arising solely from the language shift from ℓ_{src} to ℓ_{tgt} within two semantically equivalent documents. We define the relative language vector of each target language ℓ_{tgt} with respect to the source language by averaging this offset over all N translation pairs:

$$V_{\ell_{\text{tgt}}} = \frac{1}{N} \sum_{i=1}^{N} \left(z_i^{\ell_{\text{tgt}}} - z_i^{\ell_{\text{src}}}\right), \quad \forall \ell_{\text{tgt}} \in \mathcal{L}_{\text{tgt}} \quad (1)$$

3.2 Index-side Language Shift

After estimating the relative language vectors, we calibrate document embeddings during indexing via linear subtraction. Specifically, for each document pre-labeled with its language ℓ, we subtract the corresponding vector if $\ell \neq \ell_{\text{src}}$. This adjustment is controlled by a hyperparameter α, which scales the transformation magnitude: larger α subtracts more of $V_{\ell_{\text{tgt}}}$, pulling target-language document embeddings closer to the source-language embedding space. We use a single global α shared by all documents and sweep $\alpha \in \{0.1, 0.2, \ldots, 1.0\}$ to report sensitivity to α. We do not tune α per dataset, language, or query set. The overall indexing procedure is summarized in Algorithm 1.

3.3 Retrieval

The retrieval process on the constructed index $\tilde{\mathcal{Z}} = \{\tilde{z}_1, \ldots, \tilde{z}_M\}$ is performed identically to standard dense retrieval procedures. Given a query q in the source language, we generate the query embedding $z_q = f(q)$ using the same embedding model f used for indexing. The final retrieval score is calculated

as the similarity between the query embedding and the calibrated document embedding $\tilde{z}$ in the index:

$$\text{score}(q, d) = \text{sim}(z_q, \tilde{z})$$

All models used in our experiments utilize cosine similarity for $\text{sim}(\cdot, \cdot)$. Since the documents processed with SHIFT are already aligned to the source language space, no additional transformation is required for the query. Consequently, our method introduces no additional latency at query-time compared to baseline models.

4 Experimental Setup

4.1 Models & Resources

Relative Language Vector Estimation To estimate relative language vectors, we use the mMARCO dataset (Bonifacio et al., 2022). mMARCO provides machine-translated versions of the original English content in 14 languages (via Google Translate), yielding aligned text pairs across languages. We leverage all 533k aligned passage pairs for language vector estimation.

Multilingual Retrievers To verify the generalizability of SHIFT, we employ a diverse set of multilingual dense retrievers featuring different architectures. We include encoder-based models such as *embeddinggemma-300m* (Vera et al., 2025), *multilingual-e5-large* (Wang et al., 2024), and *bge-m3* (Chen et al., 2024a), as well as decoder-based models including *Qwen3-Embedding-0.6B* (Zhang et al., 2025), *llama-nemotron-embed-1b-v2* (Babakhin et al., 2025), and *gte-Qwen2-1.5B-instruct* (Li et al., 2023). Implementation details are provided in Appendix B.

4.2 Evaluation Datasets

To rigorously evaluate the robustness of SHIFT, we employ four benchmarks adapted to the MLIR setting. In our main experiments (Section 5.1), we use English queries as the source language and retrieve from a multilingual corpus. We further evaluate Chinese, Vietnamese, and Hindi queries as source languages in Section 5.3. Below, we briefly describe each dataset and our MLIR construction procedure.

Belebele & XQuAD Belebele (Bandarkar et al., 2024) is a fully parallel 122-language MRC benchmark built on the FLORES-200 translation benchmark (Team et al., 2022), and XQuAD (Artetxe et al., 2020) is a fully parallel 10-language translation of a subset of SQuAD v1.1 (Rajpurkar et al., 2016). To adapt these datasets for MLIR, we construct a multilingual retrieval including all language versions of the context passages, inspired by the multilingual pool construction in LAReQA (Roy et al., 2020). Accordingly, for each source query, we treat the source-language gold context and all target-language versions of that context as positive documents. In our experiments, we utilize the subset of languages that overlap with mMARCO, resulting in 14 languages for Belebele and 8 languages for XQuAD.

MLQA MLQA (Lewis et al., 2019) is an extractive QA benchmark covering 7 languages. Unlike fully parallel datasets where every item exists in all languages, MLQA is constructed such that each instance is aligned across a subset of languages (typically four). For the MLIR setup, we aggregate contexts from all 7 languages in the document collection. For a given source query, all available contexts across different languages are treated as positive documents. This implies that the number of positive documents and their language composition may vary per query, but all semantically linked multilingual contexts are considered as positives.

MultiEuP-v2 MultiEuP-v2 (Yang et al., 2024c) is a dataset derived from European Parliament proceedings, spanning 24 languages, using debate titles as queries and individual speeches as documents. We configure the MLIR task such that for a single source query (debate title), all speeches belonging to that debate topic are considered positive documents. Thus, the positive document includes both the speech in the original language and speeches in other languages within the same topic. Consistent with our training setup, we evaluate on the 6 languages that overlap with mMARCO.

We summarize the dataset statistics in Table 1, with additional details in Appendix A.

4.3 Evaluation Metrics

We employ standard metrics, nDCG@20 and Recall@20, to measure retrieval accuracy. However, these conventional metrics do not reveal whether relevant documents beyond the query language are retrieved. In MLIR benchmarks such as MultiEuP-v2, for instance, a model can obtain high nDCG and Recall scores by disproportionately retrieving source-language documents, despite the presence of relevant documents in other languages.

Data Statistics	Belebele	MLQA	XQuAD	MultiEup-v2
# of Languages	14	7	8	6
Fully Parallel	O	X	O	X
# of Queries	900	11,582	1,190	1,781
# of Documents	6,832	36,799	1,920	35,444
# of d^+/q	14	3.65	8	19.9
# of d^+_{en}/q	1	1	1	7.91

Table 1: Statistics of the MLIR benchmark datasets. # of d^+/q denotes the number of positive documents per query, and # of d^+_{en}/q indicates the number of English positive documents per query.

To better capture this aspect, we introduce Target-Languages Recall@k (TLR@k). Essentially, this metric computes the recall score exclusively on the subset of relevant documents written in languages other than the query language. For a query q, let $\mathcal{D}_q^+$ denote the set of relevant documents to the query. We define the subset of relevant documents in target languages as $R_q^{\text{tgt}} = \{d \in \mathcal{D}_q^+ \mid L(d) \neq L(q)\}$. TLR@k is then calculated as the average recall of these documents:

$$\text{TLR@k} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \frac{\left|\text{top}_k(q) \cap R_q^{\text{tgt}}\right|}{\left|R_q^{\text{tgt}}\right|}$$

A higher TLR@k indicates that the model successfully retrieves target-language documents, complementing conventional accuracy metrics that may be dominated by source-language documents.

5 Results & Analysis

5.1 Main Results

Retrieval Performance Table 2 presents the retrieval performance before and after applying SHIFT across four MLIR benchmarks. The results demonstrate that SHIFT consistently achieves significant performance gains across all settings. For the encoder-based model, nDCG@20 for the *embeddinggemma-300m* model increased from 0.758 to 0.765 on average, while it significantly rose from 0.633 to 0.737 for *multilingual-e5-large* after applying SHIFT. This improvement extends to decoder-based models, suggesting that SHIFT effectively mitigates the representational discrepancies between source queries and multilingual documents across architectures and model scales.

Mitigation of Language Bias To verify that the performance improvement stems from increased exposure of relevant target-language documents rather than merely higher rankings of source-language documents, we analyze the TLR@20 metric in Table 3. Notably, standard metrics such as Recall can mask underlying language bias, as a model may achieve high scores solely by retrieving documents in the source language. For instance, on MultiEup-v2, *multilingual-e5-large* attains a reasonable Recall@20 of 0.302 in Table 2, yet its TLR@20 is only 0.095 in Table 3. This gap suggests that the model predominantly retrieves relevant documents in the source language, underexposing relevant target-language documents. Applying SHIFT alleviates this issue, yielding a substantial increase in TLR@20 across all models. This confirms that SHIFT bridges the representation gap between the query and document languages, effectively promoting relevant target-language documents to the top ranks. We provide additional comparisons to representative debiasing baselines in Appendix C.

5.2 Language Distribution Analysis

To complement the TLR analysis, Figure 2 visualizes the language distribution of the top-20 documents retrieved by *multilingual-e5-large* and *gte-Qwen2-1.5B-instruct*. The heatmap compares the density of retrieved documents by language before and after applying SHIFT, and thereby provides a qualitative view of the language composition and diversity of the retrieved results.

Our experimental results demonstrate that existing baseline models have a certain language preference. Given an English source query, the models assign disproportionately high similarity scores to English documents, filling the top ranks with English content. Conversely, documents in target languages are ranked lower, appearing as inactive or very faint regions on the heatmap. This indicates that the model's similarity space favors certain language regions even when relevant documents exist more broadly in target languages.

After applying SHIFT, this concentration becomes noticeably less pronounced. The density previously concentrated in English becomes more evenly dispersed across languages, and previously inactive regions become clearly visible. This suggests that SHIFT encourages stronger cross-lingual alignment between the source-language query and target-language documents, enabling the retriever to return a markedly more balanced language composition among the top-ranked results. Importantly, this qualitative trend is consistent with the

Archi	Model	Belebele		MLQA		XQuAD		MultiEup-v2		AVG	
		nDCG	Recall	nDCG	Recall	nDCG	Recall	nDCG	Recall	nDCG	Recall
encoder	embeddinggemma-300m	0.924	0.926	0.664	0.740	0.947	0.970	0.495	0.433	0.758	0.767
	+ SHIFT	**0.926**	**0.928**	**0.671**	**0.748**	**0.951**	**0.972**	**0.511**	**0.443**	**0.765**	**0.773**
	multilingual-e5-large	0.816	0.812	0.494	0.523	0.855	0.900	0.367	0.302	0.633	0.634
	+ SHIFT	**0.910**	**0.907**	**0.649**	**0.721**	**0.944**	**0.963**	**0.442**	**0.375**	**0.737**	**0.742**
	bge-m3	0.874	0.871	0.607	0.679	0.923	0.947	0.455	0.387	0.715	0.721
	+ SHIFT	**0.876**	**0.874**	**0.609**	**0.682**	**0.925**	**0.950**	**0.460**	**0.391**	**0.717**	**0.724**
decoder	Qwen3-Embedding-0.6B	0.890	0.887	0.611	0.688	0.914	0.940	0.468	0.403	0.721	0.729
	+ SHIFT	**0.895**	**0.892**	0.611	0.688	**0.916**	**0.942**	**0.476**	**0.411**	**0.725**	**0.733**
	llama-nemotron-embed-1b-v2	0.688	0.681	0.523	0.602	0.927	0.957	0.292	0.251	0.608	0.623
	+ SHIFT	**0.700**	**0.697**	**0.527**	**0.609**	**0.931**	**0.960**	**0.312**	**0.273**	**0.618**	**0.635**
	gte-Qwen2-1.5B-instruct	0.904	0.901	0.621	0.691	0.926	0.946	0.496	0.430	0.737	0.742
	+ SHIFT	**0.914**	**0.912**	**0.639**	**0.716**	**0.936**	**0.955**	**0.519**	**0.451**	**0.752**	**0.758**

Table 2: Top-20 retrieval performance (nDCG@20, Recall@20) of baseline models and SHIFT across four MLIR benchmarks, evaluated with English as the source query language. We conduct a global grid search over $\alpha \in 0.1, \ldots, 1.0$ and report the best α for each model in subsequent experiments, as described in Section 5.4.

Model	Belebele	MLQA	XQuAD	MultiEup-v2	AVG
embeddinggemma	0.922	0.706	0.966	0.318	0.728
+ SHIFT	**0.926**	**0.736**	**0.971**	**0.371**	**0.751**
mE5-large	0.798	0.380	0.886	0.095	0.540
+ SHIFT	**0.898**	**0.654**	**0.956**	**0.268**	**0.694**
bge-m3	0.864	0.637	0.942	0.293	0.684
+ SHIFT	**0.868**	**0.643**	**0.945**	**0.300**	**0.689**
Qwen3	0.880	0.650	0.933	0.302	0.691
+ SHIFT	**0.885**	**0.651**	**0.934**	**0.322**	**0.698**
llama-nemotron	0.667	**0.554**	0.951	0.123	0.574
+ SHIFT	**0.683**	0.553	**0.953**	**0.156**	**0.586**
gte-Qwen2	0.893	0.629	0.939	0.303	0.691
+ SHIFT	**0.908**	**0.683**	**0.951**	**0.378**	**0.730**

Table 3: Comparison of TLR@20 across different models and datasets.

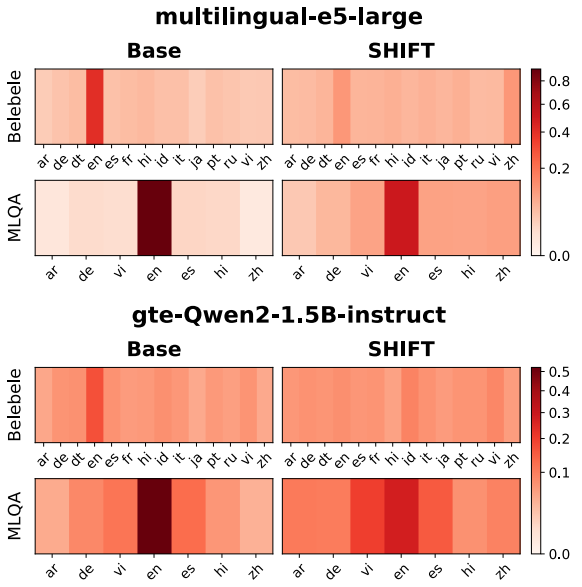

Figure 2: Visualization of language distribution in the top-20 retrieved documents for English queries. The heatmap intensity represents the density of retrieved documents for each language, where darker shades indicate a higher concentration of documents. Distributions for additional models and data are provided in Appendix D.

TLR@20 improvements in Table 3, suggesting that higher TLR@20 coincides with reduced query-language dominance and increased presence of other languages in the top ranks.

5.3 SHIFT in Multi-Source Languages

To demonstrate the universality of SHIFT and verify that it functions effectively regardless of the source language, we conduct additional experiments using three different languages as new source queries.

Experimental Setup We select Chinese (zh), Vietnamese (vi), and Hindi (hi) as non-English source (query) languages. These languages overlap with mMARCO and our evaluation benchmarks and cover distinct scripts and character distributions. For each run, we treat the selected query language as the source language ℓ_{src} and re-estimate the relative language vectors $V_{\ell_{\text{tgt}}}$ using Eq. 1 for all target languages $\ell_{\text{tgt}} \in \mathcal{L}_{\text{tgt}}$. Subsequently, during indexing (as in Section 3.2), all documents are shifted into the embedding space of the ℓ_{src}. For evaluation, we reprocess Belebele, MLQA, and XQuAD to construct MLIR test sets for each ℓ_{src}. We provide the constructed dataset statistics in Appendix E. We report nDCG@20 and TLR@20 to measure overall retrieval effectiveness and target-languages coverage.

Query	Dataset	multilingual-e5-large				bge-m3				llama-nemotron-embed-1b-v2				gte-Qwen2-1.5B-instruct			
		nDCG@20		TLR@20		nDCG@20		TLR@20		nDCG@20		TLR@20		nDCG@20		TLR@20	
		Base	SHIFT	Base	SHIFT	Base	SHIFT	Base	SHIFT	Base	SHIFT	Base	SHIFT	Base	SHIFT	Base	SHIFT
ZH	Belebele	0.321	**0.825**	0.195	**0.802**	0.824	**0.831**	0.813	**0.818**	0.597	**0.661**	0.572	**0.654**	0.875	**0.884**	0.867	**0.878**
	MLQA	0.317	**0.549**	0.059	**0.483**	0.599	**0.607**	0.637	**0.653**	0.483	**0.519**	0.487	**0.552**	0.598	**0.623**	0.633	**0.673**
	XQuAD	0.457	**0.880**	0.383	**0.907**	0.890	**0.898**	0.926	**0.931**	0.870	**0.889**	0.917	**0.930**	0.903	**0.916**	0.935	**0.942**
VI	Belebele	0.596	**0.866**	0.551	**0.853**	0.836	**0.839**	0.828	**0.831**	0.461	**0.605**	0.421	**0.581**	0.658	**0.851**	0.629	**0.842**
	MLQA	0.416	**0.630**	0.272	**0.674**	0.606	**0.610**	0.652	**0.663**	0.427	**0.494**	0.402	**0.535**	0.376	**0.543**	0.259	**0.572**
	XQuAD	0.689	**0.911**	0.717	**0.941**	0.890	**0.896**	0.920	**0.926**	0.794	**0.870**	0.832	**0.905**	0.673	**0.866**	0.702	**0.897**
HI	Belebele	0.534	**0.808**	0.482	**0.789**	0.775	**0.780**	0.764	**0.771**	0.629	**0.668**	0.611	**0.657**	0.512	**0.643**	0.466	**0.613**
	MLQA	0.414	**0.614**	0.275	**0.646**	0.591	**0.595**	0.641	**0.649**	0.493	**0.528**	0.522	**0.588**	0.352	**0.458**	0.262	**0.483**
	XQuAD	0.697	**0.902**	0.722	**0.927**	0.882	**0.887**	0.908	**0.913**	0.859	**0.877**	0.904	**0.921**	0.681	**0.799**	0.704	**0.838**

Table 4: Top-20 retrieval performance (nDCG@20, TLR@20) using non-English languages (Chinese, Vietnamese, Hindi) as the source query language.

Results Table 4 presents retrieval performance when non-English languages are set as the source query. The results show common performance improvements across all source languages and models upon applying SHIFT, proving that our method is not limited to a specific source language. The most notable change is observed in the *multilingual-e5-large* model; it records significant increases in both nDCG@20 and TLR@20 compared to the base model, demonstrating the method's effectiveness.

Another noteworthy observation is the *bge-m3* model. As shown in Table 2, when the source language is English, the baseline performance is already strong and the gains from SHIFT are marginal. However, when non-English languages serve as source languages, applying SHIFT substantially improves TLR@20 on MLQA. This suggests that while state-of-the-art models like *bge-m3* are strongly aligned around English, their alignment is relatively less consistent when the source language is non-English. Consequently, SHIFT effectively corrects these non-English-centric alignment discrepancies, demonstrating robustness across different source languages. To reflect more realistic scenarios where users query in various languages, we provide additional experiments on Multilingual-to-Multilingual Information Retrieval (M2MIR) in Appendix J.

5.4 Analysis of Scale Factor α

The scale factor α is a key hyperparameter in SHIFT that controls the strength of the language-vector subtraction. Varying α adjusts the degree of debiasing applied to the embedding space, and an overly aggressive subtraction may lead to over-correction for some models. To study this effect, we conduct a grid search over α from 0.1 to 1.0 and report the average nDCG@20 across all datasets.

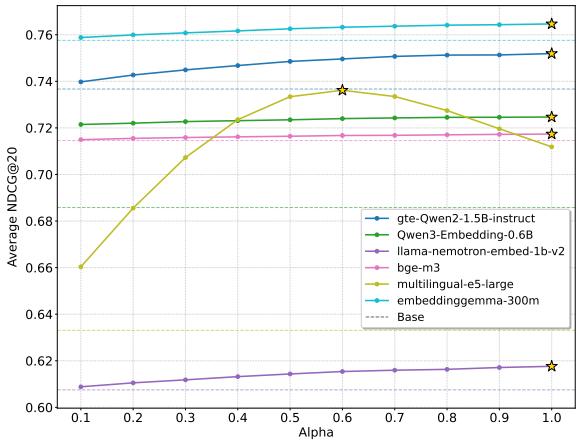

Figure 3: Impact of the scale factor α on the average nDCG@20 score.

Figure 3 illustrates the impact of α on retrieval performance. While the optimal α can vary across models, *multilingual-e5-large* exhibits an inverted-U-shaped curve with a peak at $\alpha = 0.6$. In contrast, all other models show a monotonic increase, reaching their maximum performance at $\alpha = 1.0$. Crucially, SHIFT consistently outperforms the baseline across the entire range of α. This demonstrates that our method effectively mitigates language bias without hinging on a narrowly tuned hyperparameter, although selecting a model-specific α can further maximize the performance gains.

5.5 Sensitivity Analysis on SHIFT Magnitude

While Section 5.4 explores the global impact of α across all languages, this section investigates its local impact on individual target languages and the resulting trade-off with overall system stability. We conduct a fine-grained sensitivity analysis by varying the α value for a specific target language, while fixing the α values for other languages to their global optimal settings identified previously.

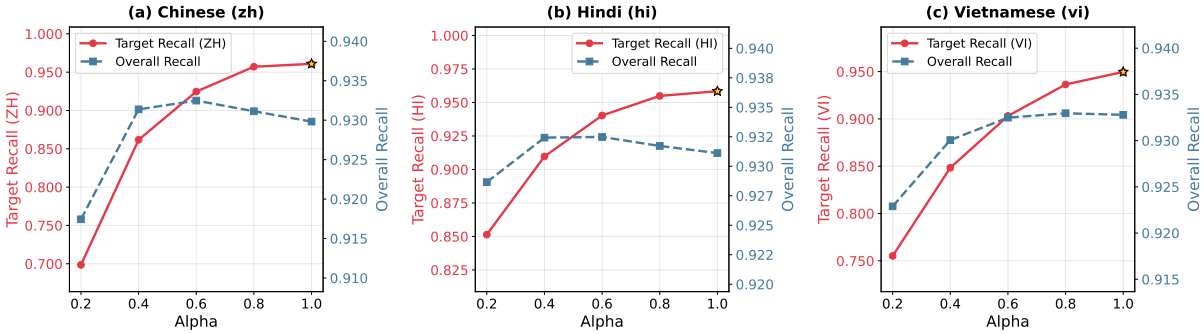

Figure 4: Sensitivity analysis of the scale factor α for specific target languages (Chinese, Hindi, Vietnamese). The solid red line represents performance on the specific target language (Target Recall@20), while the dashed blue line represents the overall system performance (Overall Recall@20).

Evaluation Setup To analyze target-language sensitivity to the scale factor α, we conduct experiments on two MLIR benchmarks, Belebele and XQuAD, using *multilingual-e5-large* with English as the fixed source (query) language. Based on Figure 3, we set the global optimal scale for *multilingual-e5-large* to $\alpha = 0.6$. We experiment on three target languages (zh, hi, vi) and perform a per-language ablation: we vary α for the target language while fixing $\alpha = 0.6$ for all other languages. We report Overall Recall@20, computed over all relevant documents regardless of language, together with Target Recall@20 for each target language. Specifically, unlike TLR@20 (which aggregates over all non-query languages), Target Recall@20 recomputes recall by restricting the ground-truth set to relevant documents in a single target language and treating relevant documents in other languages as non-relevant. All metrics are averaged over the two benchmarks to assess generalization.

Results Figure 4 illustrates the impact of α on retrieval performance. Regardless of the target language, we observe two consistent trends. First, Target Recall@20 monotonically improves as α increases. This confirms the high precision of the directionality of the relative language vectors estimated via mMARCO. Geometrically, if the relative language vector fails to capture the language's characteristics properly, increasing α would move document embeddings away from the query language space, causing a sharp performance drop. Therefore, the continuous rise in target language performance indicates that the relative language vectors extracted by our method form a correct trajectory from the target language space to the source language space.

Second, Overall Recall@20 exhibits an inverted-U trend, decreasing beyond a certain α. This suggests over-shifting: pushing a single target language too aggressively can expose documents of that language in high ranks, but it simultaneously pushes down relevant documents in other non-target languages from the top ranks. In other words, optimizing α for one language does not necessarily optimize overall retrieval across languages. Additional results for *multilingual-e5-large* on other datasets are reported in Appendix H, and results for additional retrievers are presented in Appendix I.

Controllability and Robustness Furthermore, these results suggest that α can serve as a controllable parameter in real-world environments. As observed, increasing α can explicitly boost the exposure of documents in a specific target language, even if it slightly compromises the overall ranking balance. This offers a practical advantage when dynamic adjustment of result composition is needed based on user language preferences or regional characteristics. In essence, SHIFT becomes an effective tool for managing the trade-off between semantic alignment and language preference.

6 Conclusion

In this work, we demonstrate that state-of-the-art multilingual dense retrieval models suffer from severe language bias, disproportionately favoring query-language documents. To address this, we propose SHIFT, a training-free, index-side transformation that aligns document embeddings by subtracting a relative language vector estimated from parallel data. Furthermore, we introduce Target-Languages Recall@k (TLR@k) to rigorously quantify this language bias. Extensive experiments across diverse benchmarks and multi-

ple model architectures confirm that SHIFT significantly improves MLIR performance, successfully redirecting the focus from linguistic surface forms to underlying semantics.

Limitations

Despite its effectiveness, our approach has certain limitations.

First, the estimation of language vectors relies on the mMARCO dataset, which contains machine-translated texts. Consequently, the quality of the shift vector may be influenced by the translation quality of the parallel corpus.

Second, SHIFT requires the target documents to be pre-labeled with their language to apply the correct vector subtraction during indexing. Future work will focus on relaxing this dependency on explicit language labels and exploring purely unsupervised methods for language vector estimation.

Finally, SHIFT requires re-estimating the relative language vectors whenever the source language or the embedding model changes. In practice, however, this cost is incurred only once per setting: once computed, the vectors can be cached and reused across indexing runs and datasets as long as the same model and source language are maintained.

Ethics Statement

This study utilizes publicly available datasets and open-source libraries, specifically sentence-transformers, for all experiments. We adhered to the licenses and terms of use associated with these resources. regarding the use of AI tools, we utilized AI assistants (e.g., ChatGPT, Gemini) exclusively for grammatical error correction and polishing the text to enhance readability. The scientific ideas and contributions presented in this paper are entirely our own.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425). This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (RS-2024-00398115, Research on the reliability and coherence of outcomes produced by Generative AI). This research was supported by Culture, Sports and Tourism R&D Program, funded by the Ministry of Culture, Sports and Tourism, through the Korea Culture Technology Planning and Evaluation Institute, an affilated institute of the Korea Creative Content Agency in 2026 (Project Name: Development of an AI Agent Integrating Korean Language Knowledge for Personalized Language Consultation Services, Project Number: RS-2026-25506607, Contribution Rate: 25%). This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the artificial intelligence star fellowship support program to nurture the best talents (IITP-2026-RS-2025-02304828) grant funded by the Korea government(MSIT).

References

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

Sunil Arya, David M Mount, Nathan S Netanyahu, Ruth Silverman, and Angela Y Wu. 1998. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *Journal of the ACM (JACM)*, 45(6):891–923.

Yauhen Babakhin, Radek Osmulski, Ronay Ak, Gabriel Moreira, Mengyao Xu, Benedikt Schifferer, Bo Liu, and Even Oldridge. 2025. Llama-embed-nemotron-8b: A universal text embedding model for multilingual and cross-lingual tasks. *Preprint*, arXiv:2511.07025.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Rexhina Blloshmi, Tommaso Pasini, Niccolò Campolungo, Somnath Banerjee, Roberto Navigli, Gabriella Pasi, and 1 others. 2021. Ir like a sir: Sense-enhanced information retrieval for multiple languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language*, pages 1030–1041.

Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2022. mmarco: A multilingual version of the ms marco passage ranking dataset. *Preprint*, arXiv:2108.13897.

Martin Braschler. 2002a. Clef 2001 — overview of results. In *Evaluation of Cross-Language Information Retrieval Systems*, pages 9–26, Berlin, Heidelberg. Springer Berlin Heidelberg.

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

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024b. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

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.

Kareem Darwish and Douglas W Oard. 2003. Probabilistic structured query methods. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 338–344.

Najim Dehak, Reda Dehak, James R Glass, Douglas A Reynolds, Patrick Kenny, and 1 others. 2010. Cosine similarity scoring without score normalization techniques. In *Odyssey*, volume 15.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, and 67 others. 2025. Mmteb: Massive multilingual text embedding benchmark. *Preprint*, arXiv:2502.13595.

Koustava Goswami, Sourav Dutta, Haytham Assem, Theodorus Fransen, and John Philip McCrae. 2021. Cross-lingual sentence embedding using multi-task learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9099–9113.

Ximo Granell. 2014. *Multilingual information management: Information, technology and translators*. Chandos Publishing.

Seongtae Hong, Youngjoon Jang, Hyeonseok Moon, Jungseob Lee, and Heuiseok Lim. 2026. Improving semantic proximity in english-centric information retrieval through cross-lingual alignment. In *The Fourteenth International Conference on Learning Representations*.

Xiyang Hu, Xinchi Chen, Peng Qi, Deguang Kong, Kunlun Liu, William Yang Wang, and Zhiheng Huang. 2023. Language agnostic multilingual information retrieval with contrastive learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9133–9146.

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.

David A Hull and Gregory Grefenstette. 1996. Querying across languages: A dictionary-based approach to multilingual information retrieval. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 49–57.

Mei Kobayashi and Koichi Takeda. 2000. Information retrieval on the web. *ACM computing surveys (CSUR)*, 32(2):144–173.

Wessel Kraaij, Jian-Yun Nie, and Michel Simard. 2003. Embedding web-based statistical translation models in cross-language information retrieval. *Computational Linguistics*, 29(3):381–419.

Dawn Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W. Oard, Luca Soldaini, and Eugene Yang. 2025a. Overview of the trec 2024 neuclir track. *Preprint*, arXiv:2509.14355.

Dawn Lawrie, James Mayfield, Douglas W Oard, and Eugene Yang. 2022. Hc4: A new suite of test collections for ad hoc clir. In *European Conference on Information Retrieval*, pages 351–366. Springer.

Dawn Lawrie, James Mayfield, Eugene Yang, Andrew Yates, Sean MacAvaney, Ronak Pradeep, Scott Miller, Paul McNamee, and Luca Soldani. 2025b. Neuclir-bench: A modern evaluation collection for monolingual, cross-language, and multilingual information retrieval. *Preprint*, arXiv:2511.14758.

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.

Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, Xiaoqi Ren, Shanfeng Zhang, Daniel Salz, Michael Boratko, Jay Han, Blair Chen, Shuo Huang, Vikram Rao, Paul Suganthan, and 28 others. 2025. Gemini embedding: Generalizable embeddings from gemini. *Preprint*, arXiv:2503.07891.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674, Online. Association for Computational Linguistics.

Ting Liu, Andrew Moore, Ke Yang, and Alexander Gray. 2004. An investigation of practical approximate nearest neighbor algorithms. *Advances in neural information processing systems*, 17.

Walid Magdy and Gareth JF Jones. 2011. Should mt systems be used as black boxes in clir? In *European Conference on Information Retrieval*, pages 683–686. Springer.

Paul McNamee and James Mayfield. 2002. Comparing cross-language query expansion techniques by degrading translation resources. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 159–166.

Teruko Mitamura, Eric Nyberg, Hideki Shima, Tsuneaki Kato, Tatsunori Mori, Chin-Yew Lin, Ruihua Song, Chuan-Jie Lin, Tetsuya Sakai, Donghong Ji, and 1 others. 2008. Overview of the ntcir-7 aclia tasks: Advanced cross-lingual information access. In *NTCIR*.

Jeonghyun Park and Hwanhee Lee. 2025. Investigating language preference of multilingual rag systems. *Preprint*, arXiv:2502.11175.

Carol Peters and Martin Braschler. 2002a. The importance of evaluation for cross-language system development: the clef experience. In *LREC*.

Carol Peters and Martin Braschler. 2002b. The importance of evaluation for cross-language system development: the clef experience. In *LREC*.

Carol Peters, Martin Braschler, and Paul Clough. 2012. *Multilingual information retrieval: From research to practice*. Springer.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. 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.

Nikhil Sharma, Kenton Murray, and Ziang Xiao. 2025. Faux polyglot: A study on information disparity in multilingual large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8090–8107, Albuquerque, New Mexico. Association for Computational Linguistics.

Luo Si, Jamie Callan, Suleyman Cetintas, and Hao Yuan. 2008. An effective and efficient results merging strategy for multilingual information retrieval in federated search environments. *Information Retrieval*, 11(1):1–24.

Amit Singhal and 1 others. 2001. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Ming-Feng Tsai, Yu-Ting Wang, and Hsin-Hsi Chen. 2008. A study of learning a merge model for multilingual information retrieval. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 195–202.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, and 1 others. 2025. Embeddinggemma: Powerful and lightweight text representations. *arXiv preprint arXiv:2509.20354*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *Preprint*, arXiv:2402.05672.

Eugene Yang, Thomas Jänich, James Mayfield, and Dawn Lawrie. 2024a. Language fairness in multilingual information retrieval. In *Proceedings of the*

47th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 2487–2491.

Eugene Yang, Dawn Lawrie, and James Mayfield. 2024b. Distillation for multilingual information retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2024, page 2368–2373. ACM.

Eugene Yang, Dawn Lawrie, Orion Weller, and James Mayfield. 2025. Hltcoe at trec 2024 neuclir track. *Preprint*, arXiv:2510.00143.

Jinrui Yang, Fan Jiang, and Timothy Baldwin. 2024c. Language bias in multilingual information retrieval: The nature of the beast and mitigation methods. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 280–292.

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.

Puxuan Yu, Luke Merrick, Gaurav Nuti, and Daniel Campos. 2024. Arctic-embed 2.0: Multilingual retrieval without compromise. *Preprint*, arXiv:2412.04506.

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.

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

A Evaluation Dataset Details

In this work, we leverage multilingual Question Answering (QA) datasets with parallel constructions, repurposed as retrieval tasks, to systematically assess MLIR performance. Since these datasets are originally designed for QA, the corresponding passages serve as exact positive documents within the retrieval framework. Consequently, datasets developed for QA are widely adopted for the evaluation of retrieval models in current literature (Enevoldsen et al., 2025; Lee et al., 2025; Zhang et al., 2025).

Belebele Belebele is a multilingual machine reading comprehension dataset spanning 122 language variants. Based on the FLORES-200 benchmark, each instance consists of a short passage and a corresponding multiple-choice question. For the retrieval task, we utilize the question stem as the query and the passage as the positive document. A defining feature of Belebele is its full parallelism: the entire set of questions and passages represents strictly parallel content across all 122 languages. For our evaluation, we select the 14 languages that overlap with the mMARCO dataset: Arabic (ar), German (de), Dutch (nl), English (en), Spanish (es), French (fr), Hindi (hi), Indonesian (id), Italian (it), Japanese (ja), Portuguese (pt), Russian (ru), Vietnamese (vi), and Chinese (zh).

XQuAD XQuAD is a cross-lingual QA benchmark consisting of a subset of the SQuAD v1.1 development set. The dataset was constructed by professionally translating both the questions and the paragraphs (contexts) from English into 10 target languages. This rigorous human translation process guarantees that XQuAD is fully parallel; every question (query) and paragraph (document) has an exact, semantically aligned counterpart in all other languages. We utilize the 8 languages intersecting with mMARCO for evaluation: Arabic (ar), German (de), English (en), Spanish (es), Hindi (hi), Russian (ru), Vietnamese (vi), and Chinese (zh).

MLQA MLQA is a multi-way aligned extractive QA benchmark covering seven languages. Unlike Belebele or XQuAD, MLQA is designed to be multi-way parallel; each QA instance is aligned across a subset of languages (typically four) rather than the entire set. This structure results from mining parallel sentences where overlaps exist across varying language combinations to maximize linguistic diversity. Consequently, queries are also not fully parallel across all seven languages for every instance. Instead, for a given document, queries exist only in the specific subset of languages aligned with that document. Despite this partial parallelism, the dataset ensures that within each 4-way subset, the questions and contexts are semantically equivalent. Since all 7 languages overlap with mMARCO, we utilize the complete language set for evaluation: Arabic (ar), German (de), English (en), Spanish (es), Hindi (hi), Vietnamese (vi), and Simplified Chinese (zh).

MultiEup-v2 Following Yang et al. (2024c), we utilize MultiEup-v2 as an MLIR benchmark. The task is defined as retrieving the relevant parliamentary speech segment (document) given a subject descriptor (query). The dataset is constructed from European Parliament proceedings, where professional archivists manually assign official subject descriptors (e.g., 'International Human Rights') to specific speech segments. Since the descriptors are professionally translated into 24 languages, the same information can be expressed as multilingual queries. In this framework, the ground truth is the specific speech text that was originally tagged with the descriptor by human experts. We use the 6 languages that overlap with mMARCO: German (de), English (en), Spanish (es), French (fr), Italian (it), and Portuguese (pt).

We note that Belebele, XQuAD, and MLQA are utilized in MMTEB (Enevoldsen et al., 2025), while MultiEup-v2 is adopted to evaluate MLIR tasks following Yang et al. (2024c). We do not include Mr. TyDi (Zhang et al., 2021) or MIRACL (Zhang et al., 2023) in our evaluation, since their relevance annotations are not explicitly labeled across languages and the datasets are neither parallel nor aligned at the document level, making them unsuitable for measuring retrieval performance under our MLIR setup.

B Implementation Details

Model Details We utilize the `sentence-transformers`[1] library for inference. For *bge-m3*, we exclusively employ the dense embedding component. The maximum sequence length is set to 512 tokens for all models. To ensure optimal performance, we strictly adhere to the specific prefixes and instruction prompts required by each retrieval model. Language vectors are computed from unnormalized embeddings, and

[1] https://github.com/huggingface/sentence-transformers

Model	query prefix	document prefix	# of trained_langs
embeddinggemma-300m	task: search result | query: {query}	title: none | text: {document}	100+
multilingual-e5-large	query: {query}	passage: {document}	94
bge-m3	-	-	100+
Qwen3-Embedding-0.6B	Instruct: Given a web search query, retrieve relevant passages that answer the query \nQuery: {query}	-	100+
llama-nemotron-embed-1b-v2	query: {query}	passage: {document}	26
gte-Qwen2-1.5B-instruct	Instruct: Given a web search query, retrieve relevant passages that answer the query \nQuery: {query}	-	33

Table 5: Comparison of Multilingual Embedding Models

during evaluation, we apply language shifts to the unnormalized embeddings before computing cosine similarity which internally normalizes the vectors. We provide the details in Table 5.

Hardware Details We conducted our experiments using an Intel Xeon Gold 6230R @2.10GHz CPU, 376GB RAM, and 8 NVIDIA RTX A6000 48GB GPUs. The software environment included nvidia driver, CUDA, and PyTorch, running on Ubuntu 20.04.6 LTS.

C Comparison with Baselines

To situate SHIFT among prior training-free approaches for mitigating language bias, we compare it with two representative post-hoc baselines on three retrievers: *embeddinggemma-300m*, *multilingual-e5-large*, and *bge-m3*. Table 6 reports nDCG@20 and TLR@20 on four MLIR benchmarks.

C.1 Baselines

Language-wise centering (Centering) Libovický et al. (2020) propose an unsupervised centering procedure motivated by the view that multilingual representations contain both language-specific and language-neutral components. Let $\mathbf{x} \in \mathbb{R}^d$ denote the d-dimensional embedding produced by the retriever for an input text (a query or a document). Given a language ℓ, the method estimates a language centroid as the mean embedding of texts in the language,

$$\boldsymbol{\mu}_\ell = \frac{1}{N_\ell} \sum_{i=1}^{N_\ell} \mathbf{x}_i^{(\ell)}$$

and applies language-wise re-centering by subtracting the centroid from each embedding:

$$\tilde{\mathbf{x}} = \mathbf{x} - \boldsymbol{\mu}_\ell$$

We use this centering variant as a training-free baseline and apply the same transformation to both queries and documents based on their language.

Language Information Removal (LIR) Yang et al. (2021) introduce Language Information Removal (LIR), a post-training method that suppresses language identification information by removing dominant language directions in the representation space. For each language ℓ, LIR constructs an embedding matrix $\mathbf{X}_\ell \in \mathbb{R}^{N_\ell \times d}$ by stacking embeddings of texts written in ℓ. It then computes the Singular Value Decomposition (SVD),

$$\mathbf{X}_\ell = \mathbf{U}_\ell \boldsymbol{\Sigma}_\ell \mathbf{V}_\ell^\top$$

and takes the top-r right-singular vectors $\mathbf{V}_{\ell,r} \in \mathbb{R}^{d \times r}$ as principal directions associated with language identity. Each embedding is updated by subtracting its projection onto the subspace spanned by these directions:

$$\tilde{\mathbf{x}} = \mathbf{x} - \mathbf{V}_{\ell,r} \mathbf{V}_{\ell,r}^\top \mathbf{x}$$

Equivalently, this removes components along the top directions, $\tilde{\mathbf{x}} = \mathbf{x} - \sum_{j=1}^{r} (\mathbf{v}_j^\top \mathbf{x}) \mathbf{v}_j$, where $\mathbf{v}_j$ is the j-th column of $\mathbf{V}_{\ell,r}$.

C.2 Implementation

For a controlled comparison, we compute all language-dependent statistics required by the baselines from mMARCO using the same retriever under evaluation.

Centering For each model and each language ℓ, we embed mMARCO texts written in ℓ and estimate the centroid $\boldsymbol{\mu}_\ell$ as the mean of their embeddings. At indexing time, each document embedding is transformed as $\mathbf{x} \mapsto \mathbf{x} - \boldsymbol{\mu}_\ell$. At retrieval time, we apply the same transformation to the query embedding before nearest-neighbor search.

Model	Belebele		MLQA		XQuAD		MultiEup-v2		AVG	
	nDCG	TLR	nDCG	TLR	nDCG	TLR	nDCG	TLR	nDCG	TLR
embeddinggemma-300m (base)	0.924	0.922	0.664	0.706	0.947	0.966	0.495	0.318	0.758	0.728
Centering (Libovický et al., 2020)	0.924	0.924	0.669	0.727	0.947	0.967	0.508	0.363	0.762	0.745
LIR (Yang et al., 2021)	**0.926**	0.925	**0.671**	0.728	**0.951**	0.970	0.509	0.357	0.764	0.745
SHIFT	**0.926**	**0.926**	**0.671**	**0.736**	**0.951**	**0.971**	**0.511**	**0.371**	**0.765**	**0.751**
multilingual-e5-large (base)	0.816	0.798	0.494	0.380	0.855	0.886	0.367	0.095	0.633	0.540
Centering (Libovický et al., 2020)	0.889	0.877	0.622	0.624	0.930	0.942	0.443	0.196	0.721	0.660
LIR (Yang et al., 2021)	0.891	0.880	0.623	0.625	0.931	0.943	**0.444**	0.261	0.722	0.677
SHIFT	**0.913**	**0.898**	**0.649**	**0.654**	**0.944**	**0.956**	0.442	**0.268**	**0.737**	**0.694**
bge-m3 (base)	0.874	0.864	0.606	0.637	0.923	0.942	0.455	0.293	0.715	0.684
Centering (Libovický et al., 2020)	0.849	0.838	0.598	0.626	0.912	0.931	0.454	0.295	0.703	0.673
LIR (Yang et al., 2021)	0.862	0.853	0.605	0.637	0.920	0.939	0.458	**0.300**	0.711	0.682
SHIFT	**0.876**	**0.868**	**0.609**	**0.643**	**0.925**	**0.945**	**0.460**	**0.300**	**0.717**	**0.689**

Table 6: Top-20 retrieval performance (nDCG@20, TLR@20) of base models, Libovický et al. (2020), Yang et al. (2021), and SHIFT across four MLIR benchmarks, evaluated with English as the source query language.

LIR For each model and each language ℓ, we construct the embedding matrix $\mathbf{X}_\ell$ by stacking embeddings of mMARCO texts in ℓ, and compute its SVD to obtain $\mathbf{V}_{\ell,r}$. Following Yang et al. (2021), which reports that removing only the top component performs best in their setting, we set $r = 1$ throughout. At indexing time, each document embedding is transformed as $\mathbf{x} \mapsto \mathbf{x} - \mathbf{V}_{\ell,r}\mathbf{V}_{\ell,r}^\top\mathbf{x}$, and we apply the same transformation to the query embedding at retrieval time.

C.3 Results & Analysis

Results Table 6 shows that SHIFT yields the strongest average performance across the three retrievers when considering both nDCG@20 and TLR@20. The difference is most pronounced for *multilingual-e5-large*, where the base model exhibits low TLR (AVG TLR@20 of 0.540). Both centering and LIR increase TLR@20 substantially, while SHIFT achieves the highest average TLR@20 (0.694) and also improves average nDCG@20. For *embeddinggemma-300m*, all methods yield modest but consistent improvements, with SHIFT attaining the best averages. For *bge-m3*, the improvements are smaller and both baselines are slightly worse than the base model on average, suggesting that removing global language components is not uniformly beneficial when the embedding space is already relatively well aligned. This behavior is consistent with prior observations that removing dominant components may also remove semantic information (Huang et al., 2023). Overall, the comparison indicates that SHIFT provides consistent gains across models, while maintaining strong target-language exposure as measured by TLR@20.

Inference-time practicality While all compared methods are training-free, the baselines require additional query-time transformations: centering subtracts a language-dependent centroid, and LIR removes language identification components via an additional projection step. In contrast, SHIFT can be applied to document embeddings offline during indexing, and the query-side operation is fixed given a source language. As a result, SHIFT reduces inference-time overhead and is simpler to deploy in latency-sensitive retrieval systems.

D Additional Language Distribution Analysis

Figure 5 provides a comprehensive visualization of language distribution heatmaps for all evaluated models across the four benchmarks. Consistent with the findings in Section 5.2, the baseline results (left panels) universally exhibit a strong English-centric bias, regardless of the model architecture. Whether using encoder-based models like *embeddinggemma-300m* or decoder-based models like *gte-Qwen2-1.5B-instruct*, the retrieval density is disproportionately concentrated in the 'en' column, leaving other target languages marginalized. However, applying SHIFT (right panels) alleviates this skew across all architectures. The density, previously locked on English, effectively disperses to the relevant target languages, resulting in a more semantically balanced distribution. This visual evidence further corroborates the quantitative improvements reported in Table 3, confirming that SHIFT

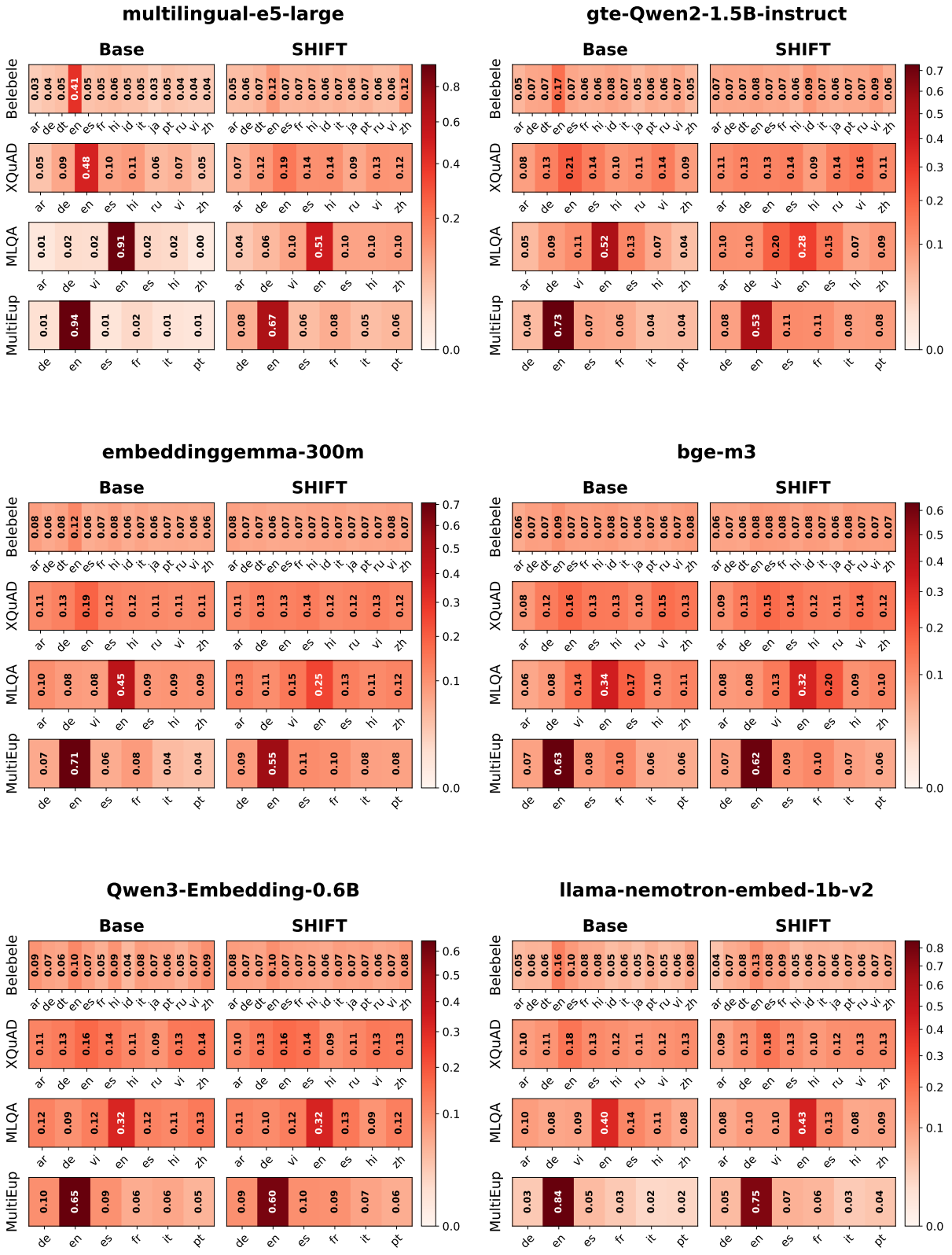

Figure 5: Extended visualization of language distribution heatmaps for all evaluated models.

serves as a robust and universal solution for rectifying language bias, independent of the underlying model type.

E Multi-Source Language Evaluation Dataset

Table 7 summarizes the detailed statistics of the datasets configured for the non-English source

Source Language	ZH			VI			HI		
Datasets	**Belebele**	**MLQA**	**XQuAD**	**Belebele**	**MLQA**	**XQuAD**	**Belebele**	**MLQA**	**XQuAD**
# of Languages	14	7	8	14	7	8	14	7	8
Fully Parallel	O	X	O	O	X	O	O	X	O
# of Queries	900	5,136	1,190	900	5,495	1,190	900	4,916	1,190
# of Documents	6,832	17,165	1,920	6,832	18,150	1,920	6,832	15,639	1,920
# Avg. d^+/q	14	3.79	8	14	3.75	8	14	3.77	8

Table 7: Detailed statistics of the datasets used for the multi-source language evaluation (Section 5.3), where Chinese (ZH), Vietnamese (VI), and Hindi (HI) serve as the source query languages.

query evaluation, where Chinese (ZH), Vietnamese (VI), and Hindi (HI) serve as the source query languages. As shown in the table, fully parallel datasets such as Belebele and XQuAD maintain consistent statistics across all source languages, including English in Table 1. In contrast, MLQA exhibits variations in the number of queries and documents depending on the source language. This is due to its partially parallel nature, where the availability of aligned query-passage pairs differs across language subsets.

F Additional Evaluation on NeuCLIR2023

To further evaluate SHIFT beyond parallel or translation-derived benchmarks, we conduct an additional experiment on a NeuCLIRBench dataset (Lawrie et al., 2025b), specifically NeuCLIR2023RetrievalHardNegatives [2]. Unlike our main experiments where relevance is constructed from semantically aligned multilingual passages, this benchmark represents a more open multilingual retrieval setting. The original task uses English queries and retrieves from a mixed multilingual document pool containing Chinese, Russian, and Persian documents. Since SHIFT requires relative language vectors estimated for each target language, we evaluate the subset of Chinese and Russian documents, which overlap with the languages supported by our vector-estimation corpus. We report nDCG@20 and apply the same model-specific scale factor α selected in the main experiments, without any dataset-specific tuning.

As shown in Table 8, SHIFT improves the average nDCG@20 from 0.4731 to 0.4839, yielding gains for 5 out of 6 evaluated retrievers. The most substantial improvement is observed for *multilingual-e5-large* (+0.039), which corresponds to the model exhibiting strong language bias in our previous analyses. Conversely, the performance of *bge-m3* changes only marginally (-0.002). This aligns with our observation that *bge-m3* already possesses relatively strong multilingual alignment, leaving less language-induced offset for SHIFT to correct.

Model	Base	SHIFT	Δ
embeddinggemma-300m	0.514	**0.520**	+0.006
multilingual-e5-large	0.409	**0.448**	+0.039
bge-m3	**0.472**	0.470	-0.002
Qwen3-Embedding-0.6B	0.461	**0.467**	+0.006
llama-nemotron-embed-1b-v2	**0.515**	**0.515**	+0.000
gte-Qwen2-1.5B-instruct	0.468	**0.483**	+0.015
Average	0.4731	**0.4839**	+0.0108

Table 8: Retrieval performance (nDCG@20) on the NeuCLIR2023 benchmark.

G Robustness to Imperfect Language Identification

Because SHIFT applies language-specific transformations during the indexing stage, its practical deployment depends on identifying the document languages. In our primary experiments, we utilized ground-truth document language labels to strictly isolate the effect of the proposed transformation. To evaluate a more realistic deployment scenario, we conduct a robustness experiment where document languages are automatically predicted using an off-the-shelf FastText language identification model[3], and SHIFT is applied based on these predicted labels.

We evaluate *multilingual-e5-large* using the same scale factor ($\alpha = 0.6$) as in the main experiments. We compare three settings: **Base**, **GT-SHIFT** (using ground-truth document language la-

[2] https://huggingface.co/datasets/mteb/NeuCLIR2023RetrievalHardNegatives

[3] https://huggingface.co/facebook/fasttext-language-identification

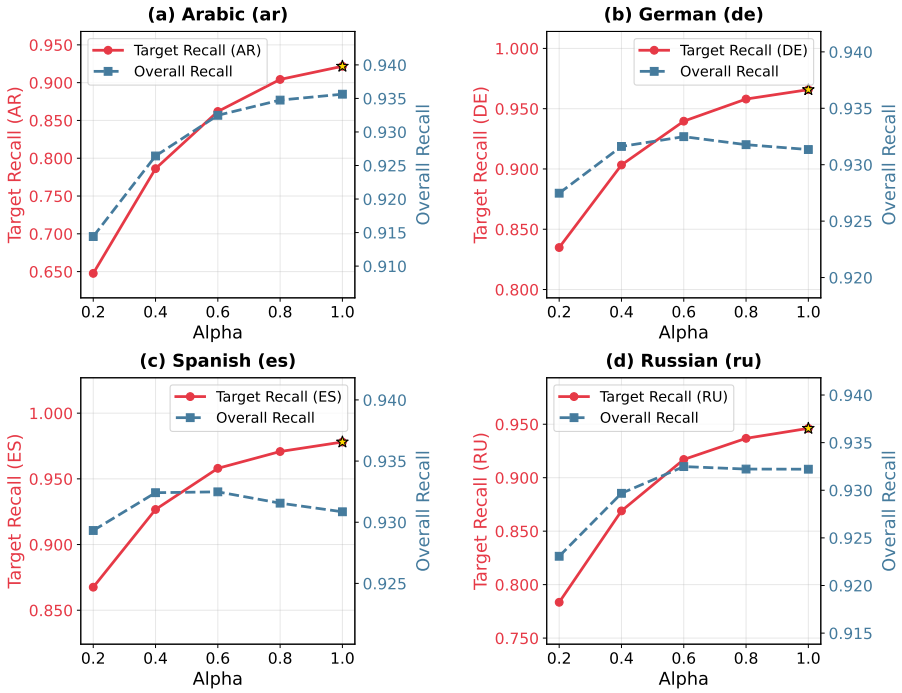

Figure 6: Extended sensitivity analysis of the scale factor α across remaining 4 target languages.

bels), and **FT-SHIFT** (using FastText-predicted document language labels). The evaluation spans the four MLIR benchmarks from the main experiments, as well as the NeuCLIR 2023 setting. We report Language Identification (LID) accuracy, defined as the percentage of corpus documents whose FastText-predicted language matches the ground-truth language.

Benchmark	Base	GT-SHIFT	FT-SHIFT	Δ (FT−GT)	LID Acc.
Belebele	0.623	0.813	0.807	-0.006	99.4%
MLQA	0.436	0.548	0.534	-0.016	93.4%
MultiEup	0.309	0.389	0.389	+0.000	99.1%
XQuAD	0.717	0.864	0.846	-0.018	96.3%
NeuCLIR 2023	0.409	0.448	0.421	-0.027	50.2%
Average	0.499	0.613	0.600	-0.013	87.7%

Table 9: Robustness of SHIFT to imperfect language identification using FastText on *multilingual-e5-large*. Performance is measured by nDCG@20.

The results are summarized in Table 9. **FT-SHIFT** consistently outperforms the **Base** model across all benchmarks, demonstrating that SHIFT remains effective even when relying on automatically predicted language labels. On average, **FT-SHIFT** improves nDCG@20 from 0.499 to 0.600, maintaining performance very close to **GT-SHIFT** with only a minimal average drop of 0.013.

The most noticeable degradation occurs on NeuCLIR 2023, where the FastText LID accuracy is only 50.2% (primarily due to a 7.4% accuracy drop on Chinese documents in this specific corpus). Nevertheless, even in this challenging setting, **FT-SHIFT** still surpasses the **Base** model (0.421 vs. 0.409). In contrast, on MultiEup, where all document languages are accurately identified, **FT-SHIFT** is virtually identical to **GT-SHIFT**. Ultimately, in deployment, SHIFT only requires predicted document languages at indexing time. Because both the language prediction and the corresponding vector transformation are applied offline, they introduce no additional overhead.

H Additional Sensitivity Analysis for *multilingual-e5-large*

We extend the sensitivity analysis presented in Section 5.5 to the remaining four languages available in the intersection of mMARCO, Belebele, and XQuAD: Arabic (ar), German (de), Spanish (es), and Russian (ru). The results are visualized in Figure 6.

As illustrated, the performance trajectories exhibit identical patterns to those reported in the main experiments. First, the Target Recall@20 (red line) monotonically increases as α grows, reaffirming the precise directionality of the estimated relative language vectors across diverse language families. Second, the Overall Recall@20 (blue dashed line) displays a characteristic inverted-U or saturation trend. This consistent behavior across all evaluated

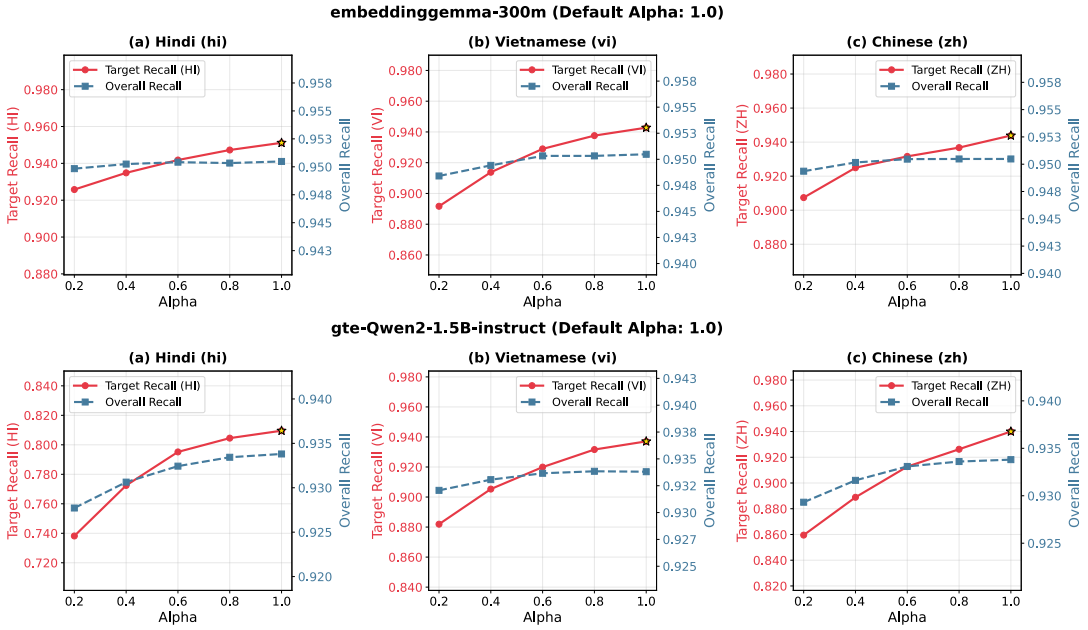

Figure 7: Sensitivity analysis with an Aggressive Default ($\alpha_{default} = 1.0$). We vary the α for the target language while fixing $\alpha = 1.0$ for all other languages.

languages further validates that while increasing α effectively boosts target-specific retrieval, an optimal threshold exists to maintain the global stability of the multilingual ranking system.

I Sensitivity Analysis on Diverse Architectures and Base Settings

To verify the universality of our findings beyond a single model architecture, we extend the sensitivity analysis to two distinct models: *embeddinggemma-300m* (encoder-based) and *gte-Qwen2-1.5B-instruct* (decoder-based). Furthermore, to investigate how the global shift intensity influences the local sensitivity of a specific target language, we conduct experiments under two extreme baseline settings:

- Aggressive Default ($\alpha_{default} = 1.0$): All non-target languages are shifted maximally. This simulates a fully transformed, source-aligned embedding space.
- Conservative Default ($\alpha_{default} = 0.1$): All non-target languages are shifted minimally. This simulates a scenario where the global embedding space remains close to the original, biased distribution.

In both settings, we sweep the α for the specific target language (Chinese, Hindi, Vietnamese) from 0.2 to 1.0 and observe the trade-off between Target Recall and Overall Recall.

Aggressive Default Figure 7 presents the results when the base system is fully aligned to the source language. In this setting, the continuous improvement in Target Recall and the high stability of Overall Recall are expected outcomes. As demonstrated in Section 5.5, these models achieve their global optimum at $\alpha = 1.0$. Therefore, increasing the specific target α towards 1.0 effectively moves the target language embeddings closer to the source language's semantic space, reinforcing the semantic signals that were previously diluted by language barriers. This confirms that the performance gains are a natural consequence of the embeddings converging to their optimal geometric alignment, where both the target-specific and global retrieval capabilities are maximized.

Conservative Default Figure 8 illustrates the results when the base shift is minimal. In this setting, the base system retains a strong bias towards the source language, typically suppressing target language documents. However, as we increase the α for the specific target language, we observe a dramatic surge in Target Recall (solid red line). This result is particularly significant from the perspective of controllability discussed in Section 5.5. It demonstrates that even if the global system remains conservative, we can effectively "boost" the exposure of a specific language by selectively increasing its shift magnitude. Although this targeted boosting involves a trade-off with Overall Recall (dashed

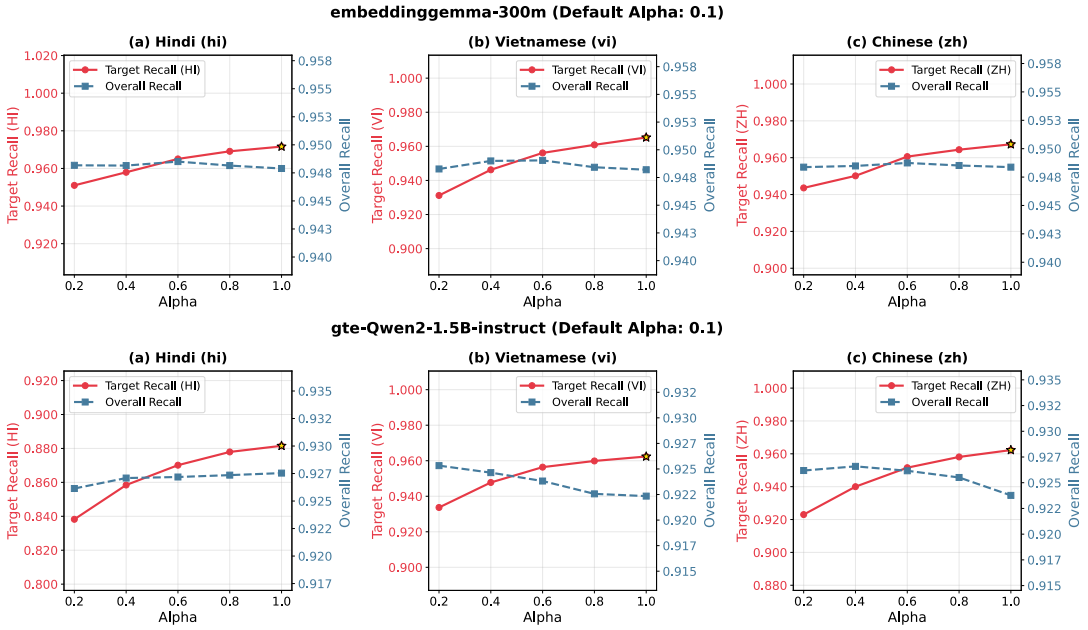

Figure 8: Sensitivity analysis with a Conservative Default ($\alpha_{default} = 0.1$). We vary the α for the target language while fixing $\alpha = 0.1$ for all other languages.

blue line), it empirically confirms that SHIFT provides a powerful mechanism to forcefully prioritize specific languages according to user requirements.

J M2MIR: Multilingual-to-Multilingual Information Retrieval

While standard MLIR tasks typically assume a monolingual query retrieving from a multilingual collection (one-to-many), real-world search scenarios often involve users issuing queries in diverse languages to find information across a globally mixed corpus. To evaluate SHIFT in this more practical and challenging environment, we introduce the M2MIR (Multilingual-to-Multilingual Information Retrieval) scenario, where both queries and documents originate from multiple languages.

J.1 Experimental Setup

Data Statistics	Belebele	MLQA	XQuAD	MultiEup-v2
# of Languages in queries	14	7	8	6
# of Languages in collection	14	7	8	6
Fully Parallel	O	X	O	X
# of Queries	12,600	42,245	9,520	10,686
# of Documents	6,832	36,799	1,920	35,444
# of d^+/q	14	3.65	8	19.9
# of d^+_{en}/q	1	1	1	7.91

Table 10: Statistics of the datasets configured for the M2MIR evaluation.

We reconfigure the test sets of Belebele, MLQA, XQuAD, and MultiEuP-v2 to include queries from all available languages overlapping with mMARCO, rather than restricting them to a single source language. The detailed statistics of this constructed dataset are summarized in Table 10. In this setting, the retrieval pool contains documents in varying languages, while the query stream simultaneously comprises a diverse mixture of languages. For all SHIFT operations in this scenario, we use English as the anchor (reference) language, i.e., we align embeddings toward the English embedding space. We set the scale factor α to 1.0 to apply the full shift toward the anchor space. Consistent with the main experiments, we employ nDCG@20 and Recall@20 as the primary evaluation metrics. We compare two strategies for applying SHIFT to align the embedding space toward English:

- *Doc Only*: We apply SHIFT exclusively to document embeddings. Specifically, we subtract the relative language vector corresponding to each document's language, aligning it with the English space. The query embeddings remain unmodified in their original language spaces.
- *Query and Doc*: We apply SHIFT to both queries and documents. Queries in non-English languages are also shifted towards the English space using their respective relative language vectors. This ensures that both the queries and the documents reside in the unified English-centric geometry.

Model	Belebele		MLQA		XQuAD		MultiEup-v2		AVG	
	nDCG	Recall	nDCG	Recall	nDCG	Recall	nDCG	Recall	nDCG	Recall
embeddinggemma-300m	0.875	0.878	0.622	0.696	0.913	0.948	0.457	0.402	0.717	0.731
+ SHIFT (*Doc Only*)	0.892	0.895	0.630	0.707	0.928	0.957	0.490	0.428	0.735	0.747
+ SHIFT (*Query and Doc*)	**0.893**	**0.896**	**0.630**	**0.707**	**0.928**	**0.957**	**0.494**	**0.431**	**0.736**	**0.748**
multilingual-e5-large	0.623	0.601	0.436	0.452	0.717	0.771	0.309	0.256	0.521	0.520
+ SHIFT (*Doc Only*)	0.860	0.856	0.594	0.664	0.914	0.947	**0.414**	**0.355**	0.695	**0.706**
+ SHIFT (*Query and Doc*)	**0.862**	**0.860**	**0.597**	**0.668**	**0.916**	**0.948**	0.409	0.347	**0.696**	**0.706**
llama-nemotron-embed-1b-v2	0.541	0.535	0.449	0.519	0.851	0.899	0.218	0.186	0.515	0.535
+ SHIFT (*Doc Only*)	0.604	0.606	0.465	0.541	0.874	**0.918**	0.258	0.226	0.550	0.573
+ SHIFT (*Query and Doc*)	**0.617**	**0.618**	**0.468**	**0.543**	**0.876**	**0.918**	**0.268**	**0.235**	**0.557**	**0.579**
gte-Qwen2-1.5B-instruct	0.793	0.790	0.518	0.577	0.828	0.875	0.442	0.388	0.645	0.657
+ SHIFT (*Doc Only*)	0.846	0.845	0.558	0.633	0.886	0.923	**0.478**	**0.419**	0.692	0.705
+ SHIFT (*Query and Doc*)	**0.853**	**0.852**	**0.566**	**0.642**	**0.892**	**0.927**	0.477	0.417	**0.697**	**0.709**

Table 11: Retrieval performance in the M2MIR scenario. We compare the effectiveness of shifting only the documents (*Doc Only*) versus shifting both queries and documents (*Query and Doc*) into the unified embedding space.

J.2 Results

Table 11 presents the retrieval performance in the M2MIR setting. We first observe that the *Doc Only* strategy significantly improves performance over the baseline across all models. For instance, embeddinggemma-300m increases from 0.717 to 0.735 on average nDCG@20, and llama-nemotron-embed-1b-v2 gains from 0.515 to 0.550. This indicates that simply shifting the multilingual document mixture into a single English-aligned space makes the documents more accessible, even if the queries are multilingual.

However, the most robust performance is achieved with the *Query and Doc* strategy. By shifting the queries into the same English subspace as the documents, we maximize the geometric alignment between queries and relevant documents. As shown in the results, applying SHIFT to both sides consistently yields the highest nDCG@20 and Recall@20 scores. Notably, for embeddinggemma-300m, the average nDCG@20 further improves to 0.736, and llama-nemotron-embed-1b-v2 reaches 0.557, outperforming the *Doc Only* approach. Even for strong baselines like multilingual-e5-large and gte-Qwen2-1.5B-instruct, the *Query and Doc* strategy maintains or slightly exceeds the high performance of *Doc Only*. This confirms that in M2MIR scenario, unifying both query and document representations into a common pivot space is the optimal strategy.

K Qualitative Analysis

To complement our quantitative results, we present qualitative retrieval examples that illustrate how SHIFT changes the ranking behavior in practice. Each table shows English queries and the top-ranked passages retrieved by the base model and by applying SHIFT. We highlight cases where the base retriever over-prefers documents written in the query language (English), while SHIFT promotes semantically relevant passages in other languages, improving target-language exposure without introducing obvious topic drift. We report examples for four retrievers to demonstrate that these patterns are consistent across architectures and model families.

Q: What do some animals not have?						
Rank	**Base**			**SHIFT**		
	Content	**Lang**	**Relevant**	**Content**	**Lang**	**Relevant**
1	Animals are made of many cells. They eat things and digest them inside. Most animals can move. Only animals have brains (though not even all animals do; jellyfish, for example, do not have brains). Animals are found all over the earth. They dig in the ground, swim in the oceans, and fly in the sky.	en	TRUE	物由多胞成。物吃了食物后在体其消化。大部分物都能移。只有物有（但也不是所有物都有，比如水母就有）。地球上的各角落都有物的身影。它有的在地下挖洞，有的在海洋中游，有的在天空中翱翔。 *[Animals are made of many cells. They eat things and digest them inside. Most animals can move. Only animals have brains (though not even all animals do; jellyfish, for example, do not have brains). Animals are found all over the earth. They dig in the ground, swim in the oceans, and fly in the sky.]*	zh	TRUE
2	Out on the savanna, it is hard for a primate with a digestive system like that of humans to satisfy its amino-acid requirements from available plant resources. Moreover, failure to do so has serious consequences: growth depression, malnutrition, and ultimately death. The most readily accessible plant resources would have been the proteins accessible in leaves and legumes, but these are hard for primates like us to digest unless they are cooked. In contrast, animal foods (ants, termites, eggs) not only are easily digestible, but they provide high-quantity proteins that contain all the essential amino acids. All things considered, we should not be surprised if our own ancestors solved their "protein problem" in somewhat the same way that chimps on the savanna do today.	en	FALSE	Animals are made of many cells. They eat things and digest them inside. Most animals can move. Only animals have brains (though not even all animals do; jellyfish, for example, do not have brains). Animals are found all over the earth. They dig in the ground, swim in the oceans, and fly in the sky.	en	TRUE
3	Insects were the first animals to take to the air. Their ability to fly helped them evade enemies more easily and find food and mates more efficiently. Most insects have the advantage of being able to fold their wings back along the body. This gives them a wider range of small places to hide from predators. Today, the only insects that cannot fold back their wings are dragon flies and mayflies.	en	FALSE	Les animaux sont constitués de nombreuses cellules. Ils mangent toutes sortes de choses et les digèrent naturellement. La plupart des animaux peuvent se déplacer. Seuls les animaux ont un cerveau (même si tous les animaux n'en ont pas; les méduses, par exemple, n'ont pas de cerveau). On trouve des animaux partout sur la Terre. Ils creusent le sol, nagent dans les océans et volent dans le ciel. *[Animals are made of many cells. They eat things and digest them inside. Most animals can move. Only animals have brains (though not even all animals do; jellyfish, for example, do not have brains). Animals are found all over the earth. They dig in the ground, swim in the oceans, and fly in the sky.]*	fr	TRUE
4	Wildlife is among the most challenging motifs for a photographer, and needs a combination of good luck, patience, experience and good equipment. Wildlife photography is often taken for granted, but like photography in general, a picture is worth a thousand words. Wildlife photography often requires a long telephoto lens, though things like a flock of birds or a tiny creature need other lenses. Many exotic animals are hard to find, and parks sometimes have rules about taking photographs for commercial purposes. Wild animals might either be shy or aggressive. The environment might be cold, hot, or otherwise hostile.	en	FALSE	Hewan tersusun dari banyak sel. Hewan memakan dan mencernanya. Sebagian besar hewan bisa bergerak. Hanya binatang memiliki otak (tetapi tidak semua binatang memilikinya; ubur-ubur, sebagai contoh, tidak memiliki otak). Hewan bisa ditemukan di berbagai tempat di Bumi. Hewan menggali tanah, berenang di laut, dan terbang di angkasa. *[Animals are made of many cells. They eat things and digest them inside. Most animals can move. Only animals have brains (though not even all animals do; jellyfish, for example, do not have brains). Animals are found all over the earth. They dig in the ground, swim in the oceans, and fly in the sky.]*	id	TRUE
5	Plants make oxygen which humans breathe, and they take in carbon-dioxide which humans exhale (that is, breathe out). Plants make their food from the sun by photosynthesis. They also provide shade. We make our houses from plants and make clothes from plants. Most foods that we eat are plants. Without plants, animals could not survive.	en	FALSE	Os animais são feitos de várias células. Eles comem coisas e as digerem em seu interior. A maior parte dos animais consegue se locomover. Apenas animais têm cérebros (embora nem todos os animais tenham; a água-viva, por exemplo, não tem cérebro). Os animais são encontrados por todo o planeta. Eles cavam o solo, nadam nos oceanos e voam pelo céu. *[Animals are made of many cells. They eat things and digest them inside. Most animals can move. Only animals have brains (though not even all animals do; jellyfish, for example, do not have brains). Animals are found all over the earth. They dig in the ground, swim in the oceans, and fly in the sky.]*	pt	TRUE

Table 12: Qualitative analysis comparison on Belebele dataset. The left block shows the top retrieved documents by *multilingual-e5-large*, and the right block shows the results after applying our SHIFT method.

Q: Defending democracy from foreign interference						
Rank	**Base**			**SHIFT**		
	Content	**Lang**	**Relevant**	**Content**	**Lang**	**Relevant**
1	Madam President, defence of democracy means enhancing transparency and it also means public participation, protecting journalists, empowering civil society, investing in education, in critical thinking and informed decision-making. It means keeping a strong framework for the rule of law and sanctioning those who ignore it. It means free voices. While we are fully aware of the importance of such a package, we need to be extremely careful on which particular instruments we choose to defend and not...	en	FALSE	Monsieur le Président, que dire aujourd'hui sur l'ingérence étrangère mettant en danger la démocratie? Nous nous concentrons désormais – Nous devons nous protéger et détecter les failles dans notre système qui permettent ces prises de pouvoir, que je qualifierais de souterraines. Notre ouverture sur le monde, nos libertés d'expression, de la presse et bien d'autres, ces valeurs qui nous sont si chères contiennent en elles les risques qui, parfois, nous empêchent de nous protéger. Or, nous devon... *[Mr. President, what can be said today about foreign interference endangering democracy? We are now focusing – We must protect ourselves and detect the flaws in our system that allow these takeovers, which I would call underground. Our openness to the world, our freedoms of expression, of the press and many others, these values which are so dear to us contain within them the risks which, sometimes, prevent us from protecting ourselves. However, we must...]*	fr	TRUE
2	Madam President, Vice-Presidents, colleagues, this proposal arrives at an important time as we approach the European elections in a few months and I'm very proud of our work in fighting foreign interference in the INGE and ING2 special committees in the European Parliament. Sadly, we are being proven right almost every day about how important this topic is. I believe that our special committees' commitment has been instrumental in shaping the discourse around foreign interference, and it is enco...	en	FALSE	Madam President, defence of democracy means enhancing transparency and it also means public participation, protecting journalists, empowering civil society, investing in education, in critical thinking and informed decision-making. It means keeping a strong framework for the rule of law and sanctioning those who ignore it. It means free voices. While we are fully aware of the importance of such a package, we need to be extremely careful on which particular instruments we choose to defend and not...	en	FALSE
3	The work of the INGE Committee and this House has been a real source of inspiration for the Commission. I want to congratulate the rapporteur, Sandra Kalniete, for bringing forward this work, which seems to command broad support across this House. That is testament to the importance of the report in analysing the phenomena of foreign interference and reflecting the need for a truly whole-of-society approach. The report explores many dimensions. It looks at interference via elite capture, nationa...	en	FALSE	Señor presidente, ¿cómo podemos defender nuestra democracia de injerencias extranjeras? Pues no permitiéndolo, no amparándolo, no legitimándolo. Porque durante demasiado tiempo hemos permitido, amparado y legitimado que Rusia e Irán, Venezuela y otros países financien e impulsen movimientos políticos extremistas en diferentes países europeos y también en esta Cámara. Porque hemos permitido, legitimado y amparado que países como Rusia apoyaran a movimientos separatistas en Europa para debilitarla... *[Mr. President, how can we defend our democracy from foreign interference? Well, by not allowing it, not sheltering it, not legitimizing it. Because for too long we have allowed, sheltered and legitimized Russia and Iran, Venezuela and other countries to finance and promote extremist political movements in different European countries and also in this Chamber. Because we have allowed, legitimized and sheltered countries like Russia to support separatist movements in Europe to weaken it...]*	es	TRUE
4	Madam President, dear Commissioners, colleagues, I have the tradition of reviewing voting results after this House votes on things the Chinese Communist Party really dislikes, and the Greek Social Democrats tend to consistently abstain or vote against China-critical texts. The underpinning problem is, of course, Greece's critical port of Piraeus now being China-owned. This is foreign interference at work. In my country, the Netherlands, we consistently supported Nord Stream 2, the pipeline, toge...	en	FALSE	Comme vous le savez, la défense des principes démocratiques et la promotion d'élections libres et équitables comptent parmi les priorités de la présidence belge, et le renforcement de nos démocraties va de pair avec une autre résistance: la résistance à la désinformation, à la propagande et à l'ingérence étrangère. C'est pourquoi nous avons pris l'initiative de préparer des conclusions du Conseil sur la résilience électorale et la sauvegarde des processus démocratiques contre l'ingérence étrangè... *[As you know, defending democratic principles and promoting free and fair elections are among the priorities of the Belgian presidency, and strengthening our democracies goes hand in hand with another resistance: resistance to disinformation, propaganda and foreign interference. This is why we took the initiative to prepare Council conclusions on electoral resilience and safeguarding democratic processes against foreign interference...]*	fr	FALSE
5	Señor presidente, ¿cómo podemos defender nuestra democracia de injerencias extranjeras? Pues no permitiéndolo, no amparándolo, no legitimándolo. Porque durante demasiado tiempo hemos permitido, amparado y legitimado que Rusia e Irán, Venezuela y otros países financien e impulsen movimientos políticos extremistas en diferentes países europeos y también en esta Cámara. Porque hemos permitido, legitimado y amparado que países como Rusia apoyaran a movimientos separatistas en Europa para debilitarla... *[Mr. President, how can we defend our democracy from foreign interference? Well, by not allowing it, not sheltering it, not legitimizing it. Because for too long we have allowed, sheltered and legitimized Russia and Iran, Venezuela and other countries to finance and promote extremist political movements in different European countries and also in this Chamber. Because we have allowed, legitimized and sheltered countries like Russia to support separatist movements in Europe to weaken it...]*	es	TRUE	Madam President, Vice-Presidents, colleagues, this proposal arrives at an important time as we approach the European elections in a few months and I'm very proud of our work in fighting foreign interference in the INGE and ING2 special committees in the European Parliament. Sadly, we are being proven right almost every day about how important this topic is. I believe that our special committees' commitment has been instrumental in shaping the discourse around foreign interference, and it is enco...	en	FALSE

Table 13: Qualitative analysis comparison on MultiEup-v2 dataset. The left block shows the top retrieved documents by *Qwen3-Embedding-0.6B*, and the right block shows the results after applying our SHIFT method.