Beyond Fine-Tuning: Procrustes Alignment of Multilingual Embeddings for Low-Resource Cross-Lingual Retrieval

Ali Faheem, Muhammad Hammad, Faizad Ullah, Ahmed Hassan, Fezan Rasool, Asim Karim

Department of Computer Science, Lahore University of Management Sciences, Lahore, Pakistan
{ali.faheem, hammad.muhammad, 20030057, 26100308, fezan.rasool, akarim}@lums.edu.pk
Correspondence: ali.faheem@lums.edu.pk

Abstract

Multilingual sentence-embedding models are widely used for cross-lingual retrieval; however, their performance drops significantly in low-resource languages. The Urdu language, which is considered a low-resource language by the NL community, poses this challenge, despite being spoken by over 246 million people worldwide. Its distribution in training corpora results in poor alignment with English within shared embedding spaces. To resolve this misalignment without model fine-tuning, we apply the Procrustes transformation, which is an orthogonal post-hoc alignment method with a closed-form solution. We utilize SQuAD and UQA datasets to learn a rotation matrix from a small set of sentence pairs and evaluate its effect across five multilingual embedding models (MiniLM, DistilUSE, E5-Base, LaBSE, and E5-Large) and perform geometric alignment, cross-lingual retrieval, and question-answering tasks. We find that cosine distances between parallel pairs decrease by up to 38.67%, and retrieval accuracy improves by 12.49% in Recall@1. We also analyze that models with better pre-trained cross-lingual representations exhibit a saturation effect, showing minimal retrieval change even as geometric tightening increases. Our error analysis reveals that morphologically complex queries and colloquial expressions remain challenging, indicating representational limitations beyond the scope of a linear transformation. These findings demonstrate that a computationally inexpensive alignment step can meaningfully improve cross-lingual retrieval for low-resource languages, with implications for retrieval-augmented generation (RAG) in resource-constrained settings.

Keywords: Cross-lingual information retrieval, Procrustes alignment, Urdu, Multilingual embeddings, Low-resource languages, Embedding space alignment

1. Introduction

Information on the web is easily available for English and a few other high-resource languages (Joshi et al., 2020). Low-resource languages, including Urdu, cannot take advantage of the high availability of English data. For instance, users who search in low-resource languages often cannot access relevant content. This issue mainly arises from the language gap between the input query and the available resources. Therefore, efficient solutions are required for deployment in resource-constrained settings. Cross-lingual information retrieval helps bridge this gap by enabling queries in one language to retrieve documents in another (Braschler et al., 1999).

Urdu is considered a low-resource language in the NLP research community. It is the 11th most widely spoken language in the world, with over 246 million speakers (Bhalloo and Molnar, 2025), and serves as the national language of Pakistan. However, due to low-resource constraints, Urdu remains significantly underrepresented in the training data of various multilingual pretrained language models (Conneau et al., 2020). Its Perso-Arabic script (Nastaliq), rich morphological structure, and the limited availability of large-scale digitized corpora further contribute to this gap. These factors pose significant challenges for NLP systems trained predominantly on high-resource languages with Latin-script data. Other low-resource languages such as Bengali, Sindhi, and Pashto face similar challenges, making progress on Urdu relevant to a broader class of under-resourced languages.

To address these challenges, recent research has increasingly focused on cross-lingual representation learning and retrieval methods. The most common approach to cross-lingual retrieval relies on multilingual sentence embedding models (Reimers and Gurevych, 2019). These models are trained on large multilingual corpora to map semantically similar text to nearby points in a shared vector space. However, the resulting space is often poorly aligned. Embeddings tend to cluster by language rather than by meaning, causing semantically equivalent Urdu and English queries to occupy distant regions of the vector space. This problem is particularly pronounced for Urdu, as it is underrepresented in the training data of most multilingual models. Alternative strategies such as query translation or model fine-tuning require either significant computational resources or large parallel datasets (Madankar et al., 2020; Conneau et al., 2020), neither of which is readily available for Urdu. In practical deployment scenarios, par-

Proceedings of SIGUL 2026 Joint Workshop with ELE, EURALI, and DCLRL @ LREC 2026, pages 233–241
11–12 May 2026.

ticularly in low-resource settings, even access to GPUs for model fine-tuning cannot be assumed. This motivates the development of a computationally efficient alignment method that can correct the misalignment in pre-trained embeddings without retraining or extensive bilingual supervision.

To address this gap, we apply the Procrustes transformation as a post-hoc method to align the Urdu and English subspaces within pre-trained multilingual embeddings. The Procrustes analysis learns an orthogonal rotation matrix from a small set of parallel sentence pairs, correcting the geometric offset between two language subspaces without modifying the underlying model or distorting its internal structure. Notably, the transformation has a closed-form solution via singular value decomposition, requiring no iterative optimization or GPU-based training, and adds only a single matrix multiplication at query time. While this technique has shown promise for word-level tasks such as bilingual lexicon induction (Lample et al., 2018), its effectiveness for sentence-level Urdu-English cross-lingual retrieval in low-resource settings has not been systematically studied. We address this gap by evaluating Procrustes alignment across five multilingual embedding models of varying size and architecture, using parallel data from the SQuAD and UQA datasets. Our evaluation covers geometric alignment of embedding spaces, cross-lingual retrieval accuracy, and downstream question answering quality in a retrieval-augmented pipeline. We make the following contributions:

- We quantify the alignment gaps in current multilingual embeddings for Urdu–English cross-lingual retrieval across five diverse architectures. Our results show that embedding spaces exhibit language-specific clustering, with cosine distance reductions of up to 38.67% after Procrustes alignment.
- We evaluate the practical impact of this alignment on cross-lingual retrieval and downstream question answering. Our results show that models with weaker baseline performance benefit the most: LaBSE gains 12.49 percentage points in Recall@1, and E5-Base more than doubles its Exact Match score, while larger models exhibit a saturation effect.

2. Related Work

Cross-lingual document retrieval has long been a central challenge in natural language processing. Early approaches relied on translating queries into the document language, or vice versa, before performing a monolingual search (Braschler et al., 1999). Such pipelines, however, introduced cascading translation errors and scaled poorly as the volume of queries or document collections increased. Neural representation learning offered a different approach: by projecting queries and documents into a shared vector space, retrieval reduces to a nearest-neighbour search, thereby eliminating the need for explicit translation.

Multilingual sentence encoders follow this approach. Multilingual BERT (Devlin et al., 2019) demonstrated that a single transformer trained across multiple languages could generalize to previously unseen ones. Subsequent models such as LaBSE (Feng et al., 2022) and multilingual E5 (Wang et al., 2024) advanced this line of work through translation-ranking and contrastive training objectives, with the goal of placing semantically equivalent sentences close together regardless of language. Empirically, however, the resulting embedding spaces remain poorly aligned. Representations tend to cluster by language rather than by semantic content, particularly for typologically distant language pairs and languages with limited pre-training data (Søgaard et al., 2018; Virtanen et al., 2019; Pires et al., 2019). This affects most severely the very languages for which cross-lingual retrieval is most needed. While fine-tuning on target-language pairs can mitigate this misalignment, it necessitates large parallel corpora and substantial computational resources, neither of which is typically available for low-resource languages.

Post-hoc alignment methods offer a lighter alternative. Procrustes analysis learns an orthogonal mapping between two embedding spaces from a small set of translation pairs, and has been applied effectively in bilingual lexicon induction (Lample et al., 2018; Grave et al., 2019; Aboagye et al., 2022). As the solution is obtained through a single Singular Value Decomposition, it requires neither iterative optimization nor GPU infrastructure, making it well-suited to resource-constrained settings. Canonical Correlation Analysis offers another projection-based approach (Faruqui et al., 2016). In parallel with this line of research, dense retrieval has made considerable progress: learned bi-encoder models now surpass traditional term-matching methods on monolingual benchmarks (Karpukhin et al., 2020), and Retrieval-Augmented Generation systems (Lewis et al., 2020) depend on accurate retrieval to ground their outputs. These two lines of work, however, have developed largely in isolation. Alignment research has concentrated on word-level tasks, while retrieval research has predominantly assumed monolingual settings. Whether lightweight alignment methods are effective at the sentence level for cross-lingual retrieval remains largely uninvestigated.

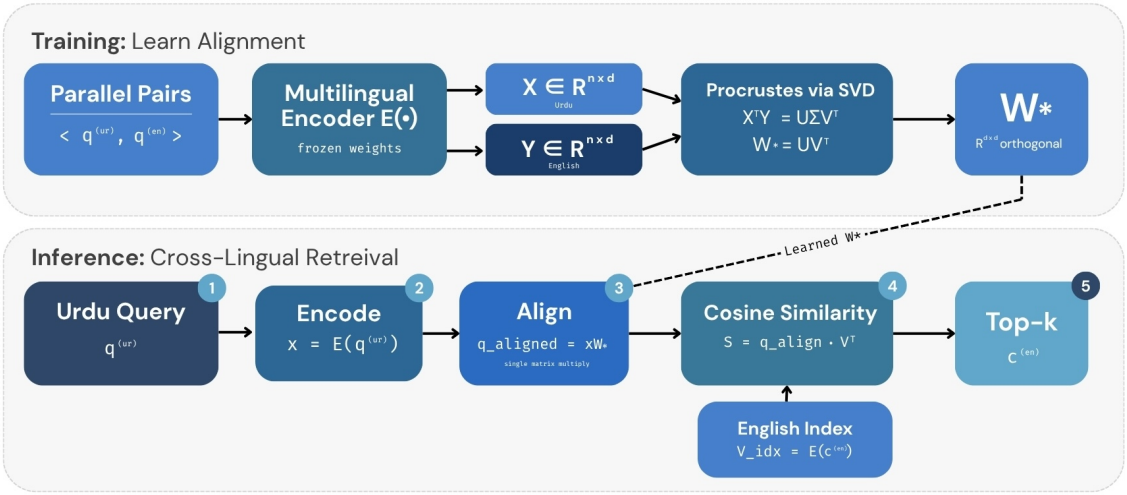

Figure 1: Overview of the proposed approach. (Top) The alignment matrix $\mathbf{W}^*$ is learned from parallel sentence pairs using SVD. (Bottom) In inference, the learned matrix aligns Urdu queries to the English embedding space for retrieval.

This gap is particularly significant for low-resource languages. Urdu remains significantly underrepresented in the training data of current multilingual models, despite being spoken by over 246 million people worldwide and serving as the national language of Pakistan and a scheduled language of India. Data scarcity, combined with the limited availability of large-scale digitized corpora in Urdu's Perso-Arabic script, makes conventional solutions impractical: fine-tuning large encoders or training bilingual models from scratch requires resources rarely available for such languages. Existing NLP work on Urdu has focused mainly on sentiment analysis (Mehmood et al., 2019), summarization (Faheem et al., 2025), and machine translation (Jayasakthi Velmurugan et al., 2025), leaving cross-lingual retrieval largely unaddressed. Other low-resource languages, including Bengali, Sindhi, and Pashto, face similar challenges: large speaker populations but limited support from multilingual systems. This work bridges the gap between these two research areas by applying Procrustes alignment, a method requiring minimal parallel data and negligible computational overhead, to sentence-level embeddings for Urdu to English retrieval.

3. Methodology

Multilingual sentence embedding models are designed to produce a shared semantic space across languages; however, the representations for low-resource languages, such as Urdu, remain underexplored. We apply a post-hoc Procrustes transformation to correct the misalignment between the Urdu and English embedding subspaces, as illustrated in Figure 1. This section describes (1) the dataset used for training and evaluation, (2) the alignment method, (3) the retrieval pipeline, and (4) the experimental setup.

3.1. Dataset

We use the Urdu Question Answering (UQA) dataset (Arif et al., 2024), a parallel bilingual corpus constructed by translating the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016). Each entry consists of an English question-context pair from SQuAD and its corresponding Urdu translation from UQA, providing aligned bilingual samples in which the English and Urdu queries are semantically equivalent. By aligning entries across the two datasets using shared SQuAD question IDs, we obtain 83,018 unique question-context pairs in both English and Urdu.

We partition this corpus into a training subset of 12,452 pairs (15%) used to learn the Procrustes alignment matrix and a held-out evaluation subset of 70,566 pairs (85%) used for all experiments reported in Section 4. This asymmetric split is motivated by the low-resource setting, as the closed-form Procrustes solution requires only a small number of parallel pairs to estimate a reliable alignment matrix. The two sets are strictly disjoint to ensure that the reported results generalize to unseen pairs.

3.2. Procrustes Alignment

Given two sets of embeddings $\mathbf{X} \in \mathbb{R}^{n \times d}$ (source language) and $\mathbf{Y} \in \mathbb{R}^{n \times d}$ (target language), where n is the number of parallel translation pairs and d is

the embedding dimension, Procrustes analysis determines an optimal orthogonal transformation matrix $\mathbf{W} \in \mathbb{R}^{d \times d}$ that minimizes the Frobenius norm $\|\cdot\|_F$, a measure of the total element-wise difference between two matrices:

$$\mathbf{W}^* = \arg\min_{\mathbf{W}} \|\mathbf{XW} - \mathbf{Y}\|_F^2 \quad \text{subject to} \quad \mathbf{W}^T\mathbf{W} = \mathbf{I}$$

The closed-form solution is obtained via Singular Value Decomposition (SVD):

$$\mathbf{W}^* = \mathbf{UV}^T \quad \text{where} \quad \mathbf{X}^T\mathbf{Y} = \mathbf{U\Sigma V}^T$$

The orthogonality constraint ensures that the transformation preserves distances and angles within the source embedding space, thereby maintaining semantic relations among Urdu representations after alignment. Any source-language embedding $\mathbf{x}$ can be projected into the target space as $\mathbf{x}_{\text{aligned}} = \mathbf{xW}^*$. In our setting, $\mathbf{X}$ contains the Urdu query embeddings and $\mathbf{Y}$ their English counterparts from the training split described in Section 3.1. The learned matrix $\mathbf{W}^*$ is subsequently applied to all Urdu embeddings at inference time.

This approach offers three methodological advantages for low-resource cross-lingual retrieval. First, it requires only a modest number of parallel pairs to learn the transformation. Second, the closed-form solution eliminates the need for iterative optimization. Third, alignment introduces negligible overhead at query time, as it reduces to a single matrix multiplication.

3.3. Experimental Setup

We evaluate the effectiveness of Procrustes alignment through two complementary experiments on the held-out evaluation set $\mathcal{D}_{\text{eval}}$. Both experiments are conducted across all five embedding models listed in Table 1, enabling an assessment of whether the alignment method generalizes across architectures of varying capacity and design. All five models include Urdu in their training data, though its representation remains limited compared to high-resource languages.

Experiment I: Geometric Alignment Analysis We quantify the topological distortion between the language subspaces by computing embeddings for all pairs in $\mathcal{D}_{eval}$. The Procrustes transformation matrix $\mathbf{W}^*$, learned from $\mathcal{D}_{train}$ via SVD, is applied to the Urdu embeddings such that $\hat{y}_i = y_i\mathbf{W}^*$, where $y_i \in \mathbb{R}^{1\times d}$ is the embedding vector of the i-th Urdu sentence. The effectiveness of this transformation is measured by calculating the mean Cosine Distance:

$$d_{cos} = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{x_i \cdot \hat{y}_i}{\|x_i\|\|\hat{y}_i\|}\right)$$

Table 1: Overview of the five multilingual embedding models evaluated in this study. Base Architecture, Number of Parameters and Embedding Dimensions.

Model	Base Architecture	Parameters	Dimensions
MiniLM[1]	BERT (Distilled)	118M	384
DistilUSE[2]	DistilBERT	135M	512
E5-Base[3]	XLM-RoBERTa	278M	768
LaBSE[4]	BERT	471M	768
E5-Large[5]	XLM-RoBERTa	560M	1024

[1] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2
[2] https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v2
[3] https://huggingface.co/intfloat/multilingual-e5-base
[4] https://huggingface.co/sentence-transformers/LaBSE
[5] https://huggingface.co/intfloat/multilingual-e5-large

A reduction in d_{cos} post-alignment indicates that the transformation has successfully mitigated the rotational misalignment between the English and Urdu subspaces.

Table 2: Average cosine distance between parallel Urdu–English sentence pairs before and after Procrustes alignment. Lower cosine distance values indicate closer semantic alignment.

Model	Before Alignment	After Alignment	Improvement
MiniLM	0.2048	0.2041	0.34%
DistilUSE	0.1947	0.1885	3.18%
E5-Base	0.1452	0.0912	37.19%
LaBSE	**0.1199**	0.0853	28.86%
E5-Large	0.1298	**0.0796**	**38.67%**

Experiment II: Cross-Lingual Retrieval To evaluate practical utility, we construct a cross-lingual dense retrieval task. The retrieval pool consists of all English contexts $C = \{c_j^{(\text{en})}\}$ from the evaluation set, each embedded into a dense index matrix $\mathbf{V}_{\text{idx}} \in \mathbb{R}^{M\times d}$. Given an Urdu query $q_i^{(\text{ur})}$, the system must retrieve the corresponding ground-truth English context $c_i^{(\text{en})}$ from this pool.

At query time, the Urdu query is embedded and transformed using the learned alignment matrix: $\mathbf{q}_{\text{align}} = \mathcal{E}(q^{(\text{ur})})\mathbf{W}^*$. Retrieval is performed by computing cosine similarity scores $S = \mathbf{q}_{\text{align}}\mathbf{V}_{\text{idx}}^T$ and ranking the English contexts accordingly. We report Recall@k for $k \in \{1, 3, 5\}$, measuring the proportion of queries for which the correct context appears in the top k results. This directly evaluates whether the geometric alignment achieved through Procrustes transformation yields improved retrieval accuracy.

Table 3: Cross-lingual retrieval performance (in %) before and after Procrustes alignment at multiple recall depths. Pre: Before Alignment; Post: After Alignment; Δ: Absolute change (Post $-$ Pre), where positive values indicate improvement. Bold values indicate the best performance per column.

	Recall@1			**Recall@3**			**Recall@5**		
Model	Pre	Post	Δ	Pre	Post	Δ	Pre	Post	Δ
LaBSE	30.24	42.73	**+12.49**	44.41	59.37	**+14.96**	51.11	65.69	**+14.58**
MiniLM	38.71	40.59	+1.88	53.57	55.43	+1.86	59.64	61.50	+1.86
DistilUSE	37.95	39.87	+1.92	53.76	55.57	+1.81	60.38	62.33	+1.95
E5-Base	56.52	56.45	-0.07	72.53	74.81	+2.28	77.89	80.33	+2.44
E5-Large	**70.42**	**70.38**	-0.04	**85.34**	**85.24**	-0.10	**89.48**	**89.34**	-0.14

Table 4: RAGAS evaluation metrics for generation quality before and after Procrustes alignment. Pre: Before Alignment; Post: After Alignment. Bold values indicate the best performance per column.

	Faithfulness		**Answer Relevance**		**Context Precision**		**Context Recall**	
Model	Pre	Post	Pre	Post	Pre	Post	Pre	Post
LaBSE	0.925	0.957	0.443	0.482	0.264	0.377	**0.320**	0.430
MiniLM	0.923	0.960	0.355	0.428	0.264	0.376	0.300	**0.450**
DistilUSE	**0.939**	**0.970**	0.359	0.426	0.264	0.376	0.300	**0.450**
E5-Base	0.935	0.960	**0.679**	**0.744**	**0.268**	0.377	0.310	0.440
E5-Large	0.923	0.963	0.663	0.727	0.264	**0.380**	0.310	**0.450**

Evaluation Metrics for Generation Quality To assess the quality of generated answers in the downstream QA pipeline, we use RAGAS (Es et al., 2024), an automated evaluation framework for RAG systems that evaluates both retrieval and generation quality without requiring human annotations. We report four metrics: (1) *Faithfulness*, (2) *Answer Relevance*, (3) *Context Precision*, and (4) *Context Recall*. *Faithfulness* measures whether each claim in the generated answer is entailed by the retrieved context, computed as the ratio of supported claims to total claims. *Answer Relevance* measures how well the answer addresses the question, computed by generating candidate questions from the answer and measuring their semantic similarity to the original question. *Context Precision* measures whether the retrieved context is relevant to the question, computed as a ranking-aware precision score over retrieved passages. *Context Recall* measures how much of the ground-truth answer is covered by the retrieved context, computed by checking entailment of ground-truth sentences against the retrieved passages. All metrics are scored on a scale of 0 to 1, with higher values indicating better performance.

4. Results and Discussion

In this section, we will discuss the results and findings. We evaluate Procrustes alignment across five embedding models across three dimensions: (1) geometric alignment, (2) cross-lingual retrieval accuracy, and (3) downstream QA quality.

Table 2 shows the mean cosine distance between parallel Urdu-English pairs before and after alignment. E5-Large and E5-Base show the reduction of 38.67% and 37.19%, respectively, followed by LaBSE at 28.86%. Other models like MiniLM and DistilUSE remain almost the same. This is consistent with the observation that larger models tend to place Urdu and English in similar but rotated subspaces, which is precisely what an orthogonal transformation corrects. The distilled models appear to embed the two languages in structurally different regions, where rotation alone cannot help much. As shown in Figure 2, the two languages form distinct clusters (before alignment) in both t-SNE and PCA projections. However, after alignment, the Urdu embeddings overlap substantially with the English space, confirming the expected effect of the orthogonal transformation.

In Table 3, these geometric shifts translate into retrieval performance. LaBSE gains Recall@1 by 12.49 points and Recall@5 by 14.58. MiniLM and DistilUSE each gain around 1.9 points, consistent with their smaller geometric improvements. The E5 models show different results: despite the most significant distance reductions, their retrieval scores barely changed. The E5-Base goes from 56.52% to 56.45%, and E5-Large stays near 70%. These models already retrieve reasonably well due to robust multilingual pretraining. Therefore, tightening the embedding space does not change which document ranks first. These results indicate that Procrustes alignment delivers the most value for weaker baseline models, which are also the

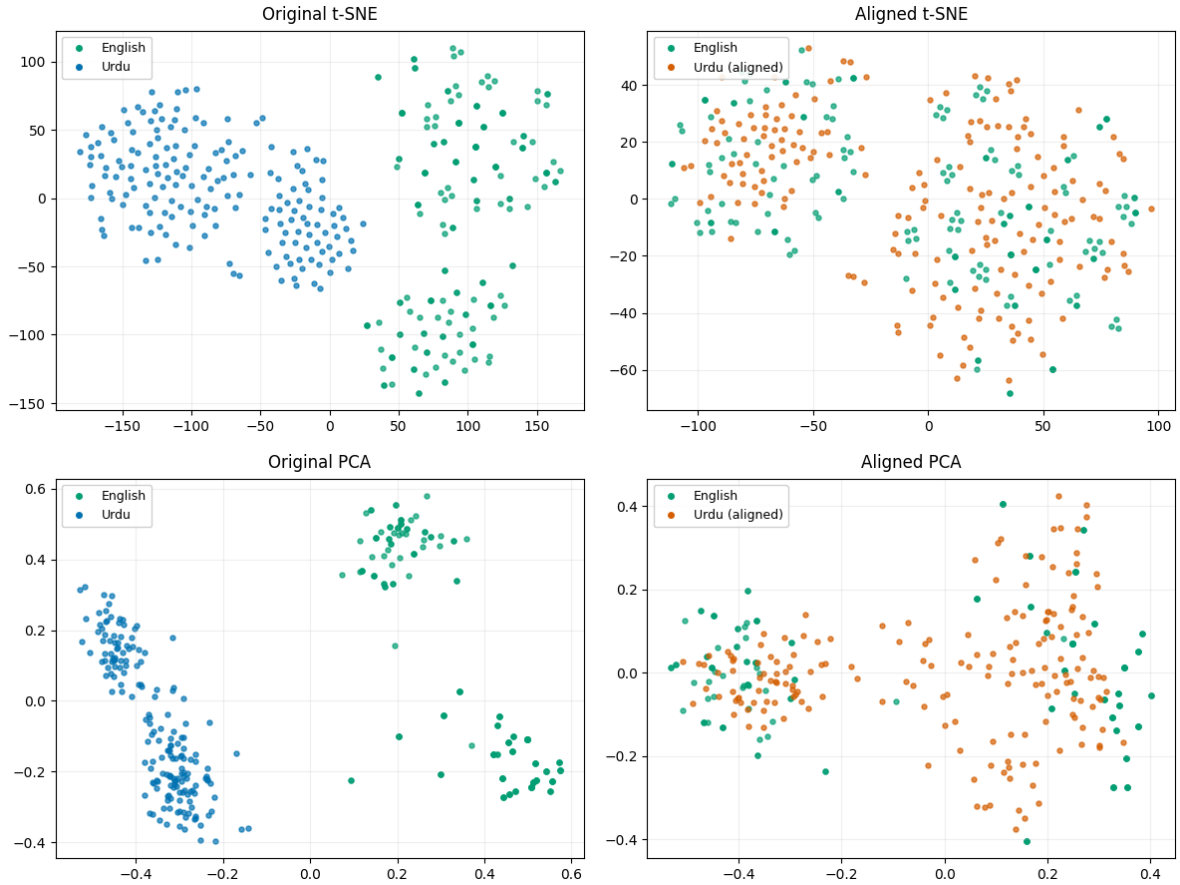

Figure 2: t-SNE (top) and PCA (bottom) projections of English (green) and Urdu (blue/orange) embeddings for LaBSE before (left) and after (right) Procrustes alignment.

ones most likely used in resource-constrained deployments.

Table 5: Exact Match and F1 scores before and after Procrustes alignment across all models.

	Exact Match			F1 Score		
Model	Pre	Post	Δ	Pre	Post	Δ
LaBSE	0.040	0.140	**+0.100**	0.040	0.168	**+0.128**
E5-Base	0.120	0.280	**+0.160**	0.162	0.308	**+0.146**
DistilUSE	0.080	0.160	+0.080	0.082	0.160	+0.078
MiniLM	0.180	0.180	0.000	0.221	0.208	-0.013
E5-Large	**0.320**	**0.300**	-0.020	**0.393**	**0.348**	-0.045

Tables 4 and 5 show whether retrieval gains carry through to answer quality. Our experiments show that retrieved contexts are relevant and that the answers they generate contain fewer unsupported claims, thereby raising context precision consistently from around 0.26 to 0.38. Faithfulness also improves consistently across all five models. The E5-Base almost doubles (from 0.12 to 0.28), and LaBSE jumps from 0.04 to 0.14. DistilUSE also improves from 0.08 to 0.16. MiniLM and E5-Large, whose retrieval was essentially unchanged, show no meaningful QA gains either. This confirms that geometric misalignment was the primary bottleneck for the mid-range models.

Despite these improvements, several limitations remain. The best post-alignment Recall@1 sits around 70%, leaving a meaningful gap. Manual inspection of failure cases reveals that errors often involve semantically related but different context passages. As shown in Table 6, queries regarding specific events in Beyoncé's career frequently retrieve topically related biography snippets instead of the exact context passage. Alignment improves the gold rank from 101 to 83 in one case and 59 to 33 in another, but does not fully resolve fine-grained factual discrimination between closely related events. These results suggest that alignment effectively repairs cross-lingual subspace rotation, but is bounded by the base model's inherent ability to disambiguate specific relations within high-density semantic regions.

5. Conclusion

This work demonstrates that a lightweight post-hoc alignment step can close the cross-lingual gap for Urdu, where conventional solutions such as fine-tuning or query translation are rarely prac-

Table 6: Qualitative comparison of retrieval results before and after Procrustes alignment (LaBSE).

Urdu Query (English translation)	**English Context Snippet**	**Pre-alignment Rank**	**Post-alignment Rank**	**Top-1 Retrieved Context Snippet**
نوٹری ڈیم کے کس صدر نے کالج آف سائنس کا قیام عمل میں لایا؟ (Which president of Notre Dame established the College of Science?)	...established in 1865 by president Father Patrick Dillon.	2	1	[Correct] ...established in 1865 by president Father Patrick Dillon...
بیونسی نے اپنا پہلا سولو البم کب جاری کیا؟ (When did Beyoncé release her first solo album?)	...first solo album Dangerously in Love was released on June 24, 2003...	13	1	[Correct] ...Her first solo album Dangerously in Love was released...
کس نے سنگل ڈیجا وو پر بیونسی کے ساتھ تعاون کیا؟ (Who collaborated with Beyoncé on the single "Déjà Vu"?)	...lead single "Déjà Vu", featuring Jay Z, reached the top five...	59	33	[Incorrect] ...revealed their marriage in a video montage at the listening party...
پہلے جوڑے کی افتتاحی گیند پر بیونسی نے کون سا گانا گایا؟ (Which song did Beyoncé sing at the first couple's inaugural ball?)	...starring as blues singer Etta James in the 2008 musical biopic...	101	83	[Incorrect] ...first solo recording was a feature on Jay Z's "'03 Bonnie & Clyde"...

tical. The results reveal a clear pattern: models with weaker baseline cross-lingual representations benefit most from Procrustes alignment, while stronger models exhibit diminishing returns, suggesting that alignment and pretraining quality interact in predictable ways that can guide model selection in resource-constrained deployments.

Several directions remain open. Combining Procrustes alignment with lightweight domain adaptation could help close the remaining gap from optimal retrieval. Testing in other low-resource languages, such as Bengali, Sindhi, and Pashto, would establish whether the findings generalize beyond Urdu. Integrating the approach into full retrieval-augmented generation pipelines would further assess its end-to-end utility. More broadly, this work demonstrates that a simple linear transformation, learned from limited parallel data, can meaningfully improve cross-lingual retrieval for low-resource languages where multilingual models fall short.

Ethics Statement

This work aims to improve information access for speakers of low-resource languages and raises no ethical concerns. All datasets used in this work are publicly available.

References

Prince O Aboagye, Yan Zheng, Michael Yeh, Junpeng Wang, Zhongfang Zhuang, Huiyuan Chen, Liang Wang, Wei Zhang, and Jeff Phillips. 2022. Quantized Wasserstein Procrustes Alignment of Word Embedding Spaces. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 200–214, Orlando, USA. Association for Machine Translation in the Americas.

Insiya Bhalloo and Monika Molnar. 2025. Urdu Phonological Tele-Assessment Tool (U-PASS) Tool Development: Supplementary Files (S1-S40). *Borealis*.

Martin Braschler, Jürgen Krause, Carol Peters, and Peter Schäuble. 1999. Cross-Language Information Retrieval (CLIR) Track Overview. In *TREC*.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of

deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th conference of the european chapter of the association for computational linguistics: system demonstrations*, pages 150–158.

Ali Faheem, Faizad Ullah, Muhammad Sohaib Ayub, and Asim Karim. 2025. Abstractive Summarization for Urdu Video Description Generation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 24(10).

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems With Evaluation of Word Embeddings Using Word Similarity Tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 30–35, Berlin, Germany. Association for Computational Linguistics.

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

Edouard Grave, Armand Joulin, and Quentin Berthet. 2019. Unsupervised alignment of embeddings with wasserstein procrustes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1880–1890. PMLR.

K. Jayasakthi Velmurugan, H. Faheem Nikhat, K. Suresh, S. Hemavathi, and V. Kavitha. 2025. Applying convolutional attention mechanisms and Human Memory Search for effective English-Urdu translation. *Engineering Applications of Artificial Intelligence*, 155:111043.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

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

Mangala Madankar, Manoj Chandak, and Nekita Chavhan. 2020. Information retrieval system based on query translation approach for cross-languages. In *International Conference on Automation, Signal Processing, Instrumentation and Control*, pages 1261–1269. Springer.

Khawar Mehmood, Daryl Essam, Kamran Shafi, and Muhammad Kamran Malik. 2019. Sentiment analysis for a resource poor language—Roman Urdu. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(1):1–15.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4996–5001.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Finnish. *arXiv preprint arXiv:1912.07076*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Samee Arif, Sualeha Farid, Awais Athar, and Agha Ali Raza. 2024. UQA: Corpus for Urdu question answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17237–17244.

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.