Efficient and Versatile Model for Multilingual Information Retrieval of Islamic Text: Development and Deployment in Real-World Scenarios

Vera Pavlova[1,2], **Mohammed Makhlouf**[1]
[1]rttl labs [2]burevestnik.ai
[1]v@rttl.ai, mm@rttl.ai [2]v@burevestnik.ai

Abstract

Despite recent advancements in Multilingual Information Retrieval (MLIR), a significant gap remains between research and practical deployment. Many studies assess MLIR performance in isolated settings, limiting their applicability to real-world scenarios. In this work, we leverage the unique characteristics of the Quranic multilingual corpus to examine the optimal strategies to develop an ad-hoc IR system for the Islamic domain that is designed to satisfy users' information needs in multiple languages. We prepared eleven retrieval models employing four training approaches: monolingual, cross-lingual, translate-train-all, and a novel mixed method combining cross-lingual and monolingual techniques. Evaluation on an in-domain dataset demonstrates that the mixed approach achieves promising results across diverse retrieval scenarios. Furthermore, we provide a detailed analysis of how different training configurations affect the embedding space and their implications for multilingual retrieval effectiveness. Finally, we discuss deployment considerations, emphasizing the cost-efficiency of deploying a single versatile, lightweight model for real-world MLIR applications. The system is deployed online[1].

1 Introduction

MLIR is a challenging area of research that has seen significant advancements recently, mainly due to the use of large language models (LLMs) (Nair et al., 2022; Lawrie et al., 2022). However, there remains a considerable gap between research efforts and the actual deployment of MLIR systems in real-world scenarios. Many studies show impressive results in controlled environments or benchmark datasets, but typically focus on evaluating the IR model in a specific setting. However, many real-world applications often require a combination of various search scenarios within a single IR system —be it multilingual, cross-lingual, or monolingual. One example of such an application is a retrieval task for the Holy Quran. Retrieving passages from the Holy Quran is uniquely challenging. With translations in over 100 languages, it offers a rich parallel collection of high-quality human translations (Bashir et al., 2023). This unique feature provides an excellent opportunity to explore the multilingual potential of retrieval models and eliminates the bottleneck of applying machine translation (MT), simplifying and streamlining the evaluation process.

This study examines training approaches for deploying a single retrieval model across diverse MLIR settings, enabling modern search capabilities in the Islamic domain. The goal is to help users efficiently locate relevant Quranic passages in multiple languages and access the cultural and religious heritage preserved within Islamic texts, serving both scholars and the general public. We utilize the XLM-R$_{\text{Base}}$ model (Conneau et al., 2020), a multilingual model trained for a general domain, as a backbone for retrieval. It is known that the performance of retrieval models typically deteriorates due to domain shift (Thakur et al., 2021; Pavlova, 2023). As a preliminary step, we conduct a brief domain adaptation of the XLM-R$_{\text{Base}}$ model using a small multilingual domain-specific corpus (approximately 100M words). This short round of pre-training resulted in significant performance improvements in retrieval tasks. Moreover, to reduce the model's size, we conduct language reduction on the XLM-R$_{\text{Base}}$ model, which allows us to eliminate languages that are not needed for the current deployment, resulting in more than a 50% reduction in the model's size. We prepare eleven retrieval models using this lightweight domain-specific multilingual large language model (MLLM) by applying four different training approaches: monolingual, cross-lingual, translate-train-all, and a proposed mixed approach that com-

[1] https://rttl.ai/

bines cross-lingual and monolingual techniques. Evaluation across monolingual, cross-lingual, and multilingual retrieval scenarios demonstrates that our proposed mixed training approach produces promising results in all settings. We conduct an in-depth analysis of the potential effects of different training configurations on the embedding space and their impact on multilingual retrieval. Additionally, we discuss the advantages of deploying a single lightweight model for various potential deployment scenarios.

Our main contributions are: (1) We prepared eleven retrieval models trained using different approaches and conducted rigorous testing. This allowed us to evaluate how these models perform in various retrieval settings without limiting the evaluation to each model's specific training approach. (2) We propose a mixed training approach that achieves competitive performance across monolingual, cross-lingual, and multilingual retrieval scenarios. (3) We deploy our model as a part of a free online multilingual search tool designed to explore Quranic text in multiple languages.

2 Preliminaries

In this work, we use the term MLIR in its broadest sense. This includes monolingual IR in any language other than English, cross-lingual IR as a special case of MLIR, and MLIR itself, which enables the processing of queries in any language while retrieving relevant documents in multiple languages (Oard and Dorr, 1996, 1998). We experimented with four languages: English, Arabic, Urdu, and Russian. The choice of languages is motivated by the availability of the evaluation dataset and diversity. Before we proceed to the details of the training approach and experiment, we briefly discuss the preparation of a lightweight domain-specific MLLM that serves as a backbone of the retrieval models.

MLLMs provide cross-lingual functionality but are heavy to deploy in low-resource settings due to their large size (Devlin et al., 2019; Lample and Conneau, 2019; Conneau et al., 2020). Language reduction is a promising approach in the deployment environment (Abdaoui et al., 2020). It decreases the model size by pruning only the embedding matrix and removing the languages that are not needed in deployment while preserving all encoder weights. We use XLM-R$_{\text{Base}}$ to perform language reduction and trim its size from 1.1 GB down to 481MB by keeping the vocabulary only of languages of interest: English, Arabic, Urdu, and Russian (XLM-R-4 model). We follow the technique of Pavlova and Makhlouf (2024). The detailed steps and model comparison are listed in the Appendix A.

Models	EN	AR	UR	RU	avg.
XLM-R-EN	0.365	0.057	0.291	0.305	0.254
EN-monolingual	0.377	0.430	0.373	0.339	0.380
AR-monolingual	0.436	0.416	0.400	0.337	0.397
ENq-ARc	**0.441**	0.358	0.381	0.333	0.378
ARq-ENc	0.133	0.406	0.427	**0.401**	0.342
ENq-Bic	0.418	0.434	0.368	0.314	0.384
ARq-Bic	0.358	0.377	0.405	0.332	0.368
Biq-ENc	0.430	**0.452**	**0.454**	0.365	**0.426**
Biq-ARc	0.417	0.389	0.422	0.361	0.397
Biq-Bic	0.349	0.365	0.407	0.333	0.363
Bilingual-train-all	0.407	0.386	0.360	0.327	0.370
4lingual-train-all	0.421	0.366	0.273	0.341	0.350

Table 1: Monolingual evaluation MRR@10.

2.1 Domain Adaptation of MLLM

While it is relatively easy to find a small amount of data for unsupervised pre-training, the absence of domain-specific labeled data for downstream tasks is a common problem. By leveraging continued pre-training and the integration of new domain-specific vocabulary (Lee et al., 2019; Huang et al., 2019; Gu et al., 2020; Beltagy et al., 2019; Pavlova and Makhlouf, 2023; Pavlova, 2025), we perform domain adaptation using a small corpus of 100 million words only. We combine a random subset of 50 million words from The Open Islamicate Texts Initiative (domain-specific corpus in Arabic) (Romanov and Seydi, 2019) with texts in English, Russian, and Urdu, mainly consisting of Tafseer and Hadith, also totaling 50 million words. We train a new Islamic tokenizer based on this corpus, add new domain-specific tokens to the existing vocabulary of the XLM-R-4 model, and continue a short round of pre-training on this assembled multilingual Islamic corpus (XLM-R-4-ID model). For more details on the hyperparameters, refer to Appendix A. As we will show below, this short pre-training round significantly boosted model performance on retrieval tasks.

3 Training Approaches of Multilingual Retrieval Models

For retrieval, we employ a dense retrieval approach (Karpukhin et al., 2020) using the sentence transformer framework that adds a pooling layer on top

Models	**EN**	**AR**	**UR**	**RU**	**avg.**
XLM-R-EN	0.237	0.197	0.165	0.255	0.222
EN-monolingual	0.293	0.345	0.324	0.289	0.326
AR-monolingual	0.329	0.336	0.313	0.265	0.328
ENq-ARc	0.350	0.272	0.286	0.264	0.310
ARq-ENc	0.112	0.359	0.292	0.302	0.281
ENq-Bic	0.330	0.341	0.276	0.259	0.318
ARq-Bic	0.297	0.332	0.297	0.260	0.311
Biq-ENc	0.383	0.339	**0.349**	0.275	0.354
Biq-ARc	0.343	0.349	0.307	0.283	0.336
Biq-Bic	0.341	0.315	0.265	0.251	0.307
Bilingual-train-all	0.328	0.324	0.278	0.283	0.317
4lingual-train-all	**0.404**	**0.414**	0.296	**0.319**	**0.357**

Table 2: Multilingual evaluation MRR@10.

Models	**AR-UR**	**UR-AR**	**AR-EN**	**EN-AR**
XLMR-EN	0.175	0.04	0.097	0.019
EN-monolingual	0.323	0.284	0.247	0.284
AR-monolingual	0.366	0.352	0.352	0.36
ENq-ARc	0.225	0.342	0.2	0.342
ARq-ENc	**0.446**	0.34	0.357	0.277
ENq-Bic	0.34	0.344	0.272	0.369
ARq-Bic	0.376	0.368	**0.382**	0.337
Biq-ENc	0.424	0.385	0.37	**0.423**
Biq-ARc	0.409	0.368	0.341	0.343
Biq-Bic	0.369	**0.39**	0.36	0.383
Bilingual-train-all	0.349	0.329	0.328	0.316
4lingual-train-all	0.207	0.099	0.32	0.281

Table 3: Cross-lingual evaluation MRR@10.

of LLM embeddings and produces fixed-sized sentence embedding (Reimers and Gurevych, 2019). The loss function is designed within the framework of contrastive learning, which helps create an embedding space that brings related queries and their relevant passages closer together while pushing away queries and irrelevant passages (van den Oord et al., 2018), and formally defined as:

$$J_{\mathrm{CL}}(\theta) = -\frac{1}{M}\sum_{i=1}^{M}\log\frac{\exp\sigma(f_\theta(x^{(i)}), f_\theta(y^{(i)}))}{\sum_{j=1}^{M}\exp\sigma(f_\theta(x^{(i)}), f_\theta(y^{(j)}))}$$

where σ is a similarity function (a cosine similarity), f_θ is the sentence encoder, $\{x^{(i)}, y^{(i)}\}_{i=1}^{M}$ (where *M* is batch size) are positive labels and other in-batch examples treated as negative (Henderson et al., 2017; Gillick et al., 2019; Karpukhin et al., 2020).

We explore four different training approaches:

(1) **Monolingual training**, in this method, both the query and the passages are in the same language L_i.

(2) **Cross-lingual training** exploits a pair of languages during training in a traditional way; while queries are in the language L_i, passages are in the language L_j.

(3) **Mixed approach**. In this strategy, we construct the training data by combining monolingual and cross-lingual methods. There are three specific ways we develop the training samples:

- **Monolingual queries with bilingual collection**: Half of the passages in the collection are in the same language as the queries L_i, while the other half is in a different language L_j.

- **Bilingual queries with monolingual collection**: Here, the queries are presented in two different languages, L_i and L_j, while the collection consists of passages in only one language, L_i.

- **Bilingual queries with bilingual collection**: In this case, both the queries and the passages can be in the languages L_i and L_j.

One of the main differences between cross-lingual training and mixed training is that in cross-lingual training, the queries and passages are always in different languages. In contrast, mixed training allows for the query language to be either in the same language as the passage language (L_i) or a different language (L_j). We hypothesize that a mixed approach can enhance the diversity of training examples and improve cross-lingual interaction between languages.

(4) **Translate-train-all approach**: This approach involves training different translations of the training dataset simultaneously. In the previous mixed approach, queries or the collection are evenly divided between two languages. In this training mode, we expand the collection by adding another translation. Translate-train-all resembles monolingual training in structure (same-language pairs) but improves language coverage by training in multiple languages simultaneously.

4 Experimental Setup

The variations of the settings described above can create numerous combinations depending on how many languages are involved in the experiment. For monolingual, cross-lingual, and mixed training approaches, we focus on using English and Arabic. For two main reasons: English is the primary language of the XLM-R$_{\text{Base}}$ model, and the language of the MS MARCO dataset, and Arabic is the language that was mainly used for domain adaptation. For the translate-train-all approach, we experiment with four languages: English, Arabic, Urdu and Russian.

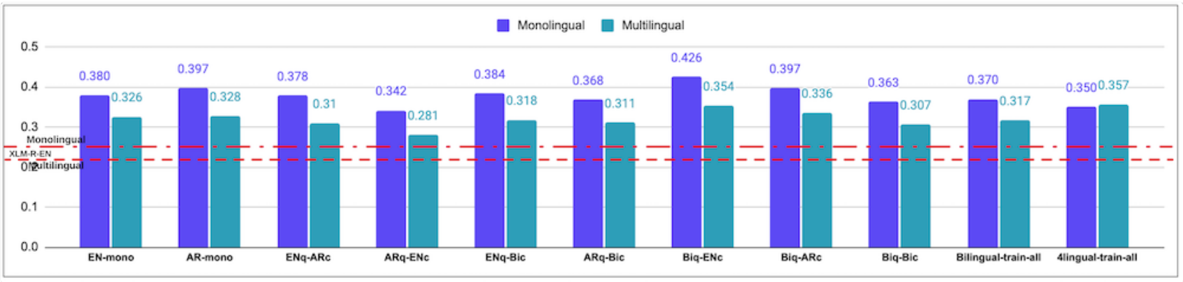

Figure 1: Comparison of the model performance averaged across languages between multilingual and monolingual evaluations. Red lines are the performance of a baseline model XLM-R-EN.

4.1 Datasets

For training, we use the MS MARCO (Bajaj et al., 2018), a large-scale English-language dataset widely adopted for training and evaluating dense retrieval models. It contains over 500,000 real-world queries paired with a collection of 8.8 million passages; Bonifacio et al. (2021) released machine-translated variants of MS MARCO for 13 languages, including Arabic and Russian (the collection was not translated into Urdu). Applying machine translation, we translated MS MARCO into Urdu to include this language in the experiment. For evaluation, we combined the train and development splits of the QRCD (Qur'anic Reading Comprehension Dataset) (Malhas and Elsayed, 2020) to increase the size of the test set, resulting in 169 queries used for evaluation. The answers provided are exhaustive, meaning all Qur'anic verses directly responding to the questions have been thoroughly extracted and annotated. The queries reflect contemporary, real user information needs, so they are both valid and salient today. This pairing of complete passage coverage with today's intents enables rigorous IR evaluation while preserving direct applicability to current user scenarios. The language of the QRCD dataset is Arabic; to evaluate in other languages, we use verified translations of this dataset to English, Russian, and Urdu. We use the Holy Quran text (Arabic), Sahih International translation (English), Elmir Kuliev (Russian), and Ahmed Raza Khan (Urdu) as retrieval collections.[2]

4.2 Evaluation Approach, Metrics and Baseline

We evaluate our methods in three different settings: monolingual, cross-lingual, and multilingual. In a monolingual setting, we evaluate using four languages. For the cross-lingual evaluation, we analyze two language pairs: Arabic and Urdu, which share similarities in writing systems and vocabulary, and Arabic and English, which represent linguistically distinct languages. In the multilingual evaluation, we combine collections of Quranic texts in four different languages. We then assess retrieval performance for each language by varying the query languages.

We use the MRR@10 (Mean Reciprocal Rate), the official evaluation metric of the MS MARCO dataset, as the main metric.

As a baseline, we train the XLM-R$_{\text{Base}}$ in a monolingual setting using English MS MARCO (XLM-R-EN model) that allows us to see the effects of domain adaptation. For the rest of the retrieval models described below, we use the XLMR-4-ID model for training.

4.3 Results

In all tables, the best score is in bold, and the second-best is underlined. In the monolingual evaluation (Table 1), the model with the highest average performance across languages (0.426) is the Biq-ENc (Bilingual Queries English Collection), which was trained with a mixed approach. It also demonstrates the best performance in both Arabic and Urdu. In the multilingual evaluation (Table 2), the best-performing model on average (0.357) is the 4lingual-train-all (translate-train-all approach). The Biq-ENc model achieves the highest score (0.349) in Urdu and also second best for average performance across four languages (0.354). As illustrated in Figure 1, all models outperform the baseline on average across languages. In the cross-lingual evaluation (Table 3), the Biq-ENc model is the top performer (0.423) for the EN-AR pair; for all other pairs, this model has the second-best results. Overall, the results indicate that the mixed training approach yields promising

[2] https://tanzil.net/trans/

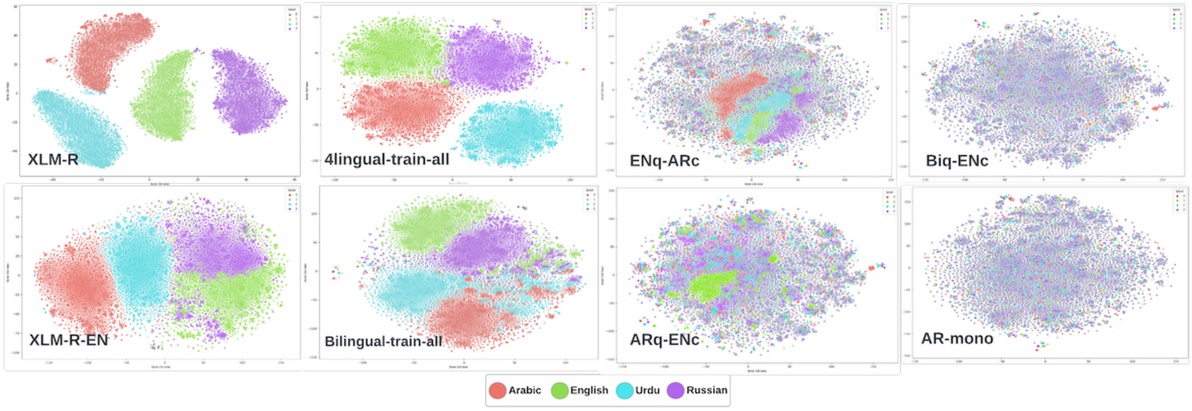

Figure 2: 2D t-SNE images of the representation of the Quranic verses embedding space in four languages.

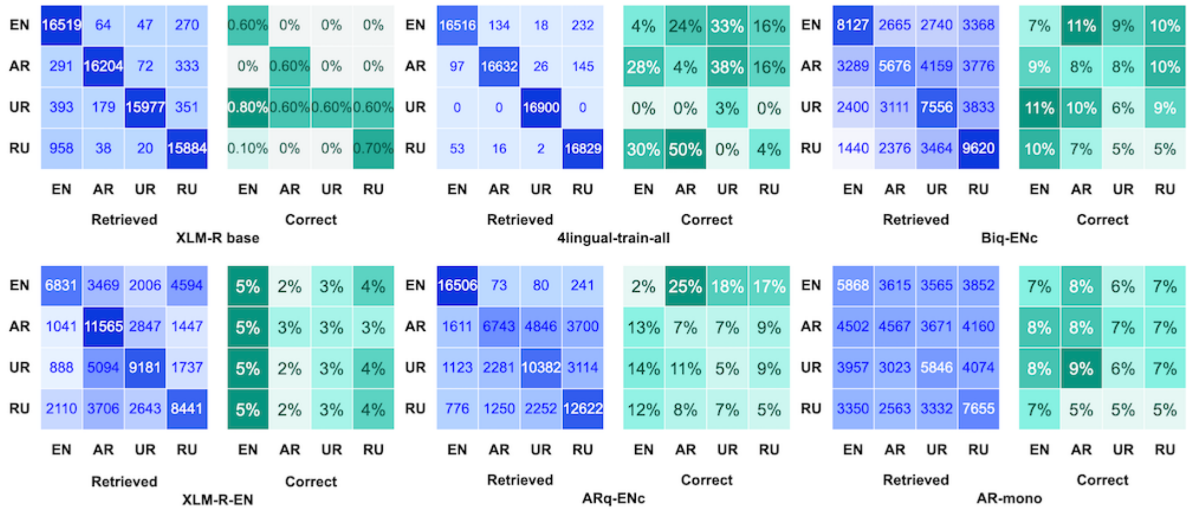

Figure 3: The heat map in blue shows how many passages from what language collection were retrieved. The heat map in green shows the percentage of correct passages out of retrieved passages. Language collection is the x-axis, and queries in a specific language are the y-axis.

outcomes, with the Biq-ENc model consistently demonstrating strong performance across all evaluation types: monolingual, multilingual, and cross-lingual.

5 Analysis

To assess how training strategies shape the multilingual embedding space, we apply the t-SNE algorithm (van der Maaten and Hinton, 2008) (Figure 2). To evaluate each model's cross-lingual capability, we examine heat maps for retrieval (blue) and correctness (green), which highlight monolingual (diagonal) and cross-lingual (off-diagonal) retrieval ability (Figure 3). To juxtapose models that demonstrate strong cross-lingual ability with those that remain monolingually biased, we include the XLM-R$_{\text{Base}}$ model, which is not fine-tuned for retrieval. The first t-SNE image in the upper row (Figure 2) shows that XLM-R$_{\text{Base}}$ produces four distinct language clusters and retrieves passages almost exclusively from the same language as the query. However, its accuracy remains extremely low, suggesting minimal ability for cross-lingual retrieval. The 4lingual-train-all (translate-train-all approach) model also clusters by language and is biased toward retrieving from the same language. However, it achieves a higher percentage of correct answers, including some off-diagonal results. The XLM-R-EN model (trained on English MS MARCO from XLM-R model without domain adaptation) produces less pronounced clusters, with partial overlap between English and Russian as shown in the t-SNE image. Remarkably, this model performs well on English and Russian datasets, but poorly on Urdu and Arabic. The green heat map aligns

Model(s)	Three Large Models	Single Versatile Lightweight Model		
Compute Substrate	**24GB A10 GPU**	**12 GB RTX 3080TI**	**CPU EM-A210R-HDD 16GB**	**Lambda Function**
Provider	Lambda Labs	Vast.ai	Scaleway	AWS
VRAM / RAM	24 GB	12 GB	16 GB	2 GB+
MRC	~ USD 540	~ USD 150	~ USD 48	~ USD 10 - USD 20

Table 4: MRC (monthly recurring cost) comparison from multiple compute substrate providers.

	Metric	MENA/EU Region	North America Region	APAC
Before	Avg. Latency	40.93ms	204.53ms	772.01ms
	Median P50 Latency	38.56ms	206.03ms	765.09ms
	P95 Latency	89.01ms	397.87ms	1380.37ms
After	Avg. Latency	25.33ms	160.03ms	408.99ms
	Median P50 Latency	23.69ms	150.73ms	402.66ms
	P95 Latency	48.99ms	220.27ms	980.11ms
%DELTA	P50 (lower better)	-38.6%	-26.8%	-47.4%

Table 5: RUM latency (ms) by region before/after model deployment; lower is better.

with language collections (vertical alignment), with the darkest column corresponding to English passages, which is expected given the model's English-centric training. The ARq-ENc model (Arabic queries, English collection) shows a more unified embedding space, though English cluster remains somewhat distinct. The blue heat map shows the darkest quadrant in the English-English retrieval, yet with a very low percentage of correct answers. Notably, this model performs poorly, particularly on English.

Conversely, t-SNE plots of Biq-ENc (bilingual queries, English collection) and AR-mono (domain-adapted XLM-R trained on Arabic MS MARCO) show a more homogeneous structure without clear language clusters. The heat maps are more uniformly colored for both retrieved and correct answers. Their off-diagonal results highlight stronger cross-lingual ability, supporting the idea that domain adaptation and mixed training encourage models to learn embedding space akin to language-agnostic representations, leading to improved performance in multilingual retrieval.

6 Deployment Considerations

Table 4 demonstrates possible deployment considerations and compares costs for deploying three separate retrieval models, each around 1GB (trained from XLM-R_{Base}), versus deploying one lightweight model of 400 MB in size (e.g. Biq-ENc model). The table shows that deploying one smaller model allows a reduction of costs by about 70% when deploying on a GPU-based server. Lower memory consumption and faster loading of a single lightweight model allow us to consider the deployment option on CPU-based servers, which further cut the cost by 70%. Python's runtime overhead, garbage collection, and large dependencies introduce inefficiencies in memory utilization, increasing overall deployment size. Leveraging Rust language capabilities to eliminate Python's inefficient memory management enables to reduce overall memory consumption to 30-50%, which can pave the way for deployment on compact serverless runtimes such as AWS Lambda functions, presenting the most cost-effective and scalable solution, potentially reducing monthly recurring costs to as low as USD 10 - USD 20.

7 Production Performance

We evaluate end-to-end latency with real-user monitoring (RUM) before and after deploying the new model (see Table 5). Measurements reflect browser-observed round-trip times from production traffic, exclude known bots, and are reported by region at three cut points: mean, median (P50), and tail (P95). As shown in Table 5, latency dropped across all regions after deployment. Median latency decreased by 38.6% in MENA/EU (38.6→23.7 ms), 26.8% in North America (206.0→150.7 ms), and 47.4% in APAC (765.1→402.7 ms). Means showed similar gains (−38.1%, −21.8%, −47.0%). Tail latency (P95) improved sharply: MENA/EU −45.0% (89.0→49.0 ms), North America −44.7% (397.9→220.3 ms), and APAC −29.0% (1380.4→980.1 ms). Lower tail latency materially improves perceived responsiveness under load and on slower networks. The APAC tail remains higher due to network distance; further reductions will require geographic routing and edge capacity in addition to model-side efficiency. The Appendix B provides additional details on user activity and other extrinsic evaluations, including the assessment of real-user queries.

8 Related work

Recent work in cross-lingual and multilingual information retrieval (CLIR and MLIR) explores extending monolingual dense retrievers such as ColBERT to the multilingual setting, often using XLM-R_{Base} as the backbone encoder (Nair et al., 2022; Lawrie et al., 2022). Our approach differs in that we use XLM-R_{Base} within a sentence embedding framework, which is more latency-efficient and scalable for real-world deployment. Unlike ColBERT-style late interaction models, we adopt full sentence representations with in-batch negatives—an approach shown to yield strong performance with lower computational cost (Qu et al., 2021; Ren et al., 2021; Karpukhin et al., 2020). Multilingual sentence embeddings have also been actively studied, with methods like LaBSE (Feng et al., 2022), mSimCSE (Hu et al., 2023), LASER (Artetxe and Schwenk, 2019), and multilingual variants of SBERT (Reimers and Gurevych, 2020) demonstrating strong performance across a variety of tasks. While many of these focus on general-purpose representation learning, our work specifically investigates how different training configurations affect multilingual retrieval quality and embedding space alignment in a domain-specific setting.

9 Conclusion

Our proposed mixed training approach has shown promising results across all evaluation settings, highlighting its beneficial properties for use in MLIR systems that need to handle various retrieval scenarios. Furthermore, the efficiency gained by deploying a single lightweight and versatile model proves to be a superior option for balancing performance, affordability, and scalability.

Limitations

Despite including different types of languages in our experiment and adding low-resource ones like Urdu, the results may vary with a significantly larger number of languages. Additionally, our mixed and cross-lingual training setups rely on parallel corpora, which may not generalize well to settings where such resources are unavailable or noisy. Finally, although we deployed a lightweight model, performance and efficiency trade-offs on truly resource-constrained devices (e.g., mobile or edge environments) remain to be fully explored.

Ethical Considerations

This work involves the retrieval of religious texts, specifically the Holy Quran, which holds deep cultural and spiritual significance for millions of people. We have taken care to use verified and official translations of Quranic text. However, variations in translation style and theological interpretation may still impact how passages are retrieved and understood across languages. We acknowledge the responsibility that comes with building search tools in sensitive domains. The system is designed to assist in information access, not to provide religious or legal rulings. Care should be taken in downstream use cases, particularly in educational, interfaith, or legal contexts. Additionally, as with any multilingual system, there remains a risk of uneven performance across languages, which could inadvertently prioritize or marginalize certain linguistic groups. We recommend future work consider input from domain experts, theologians, and community stakeholders to guide responsible deployment, especially when extending the system to broader religious or cultural corpora.

Acknowledgment

We thank the Program Chairs of the EMNLP Industry Track for their continuous support of research on underrepresented domains. We also gratefully acknowledge the Dubai AI Campus for providing the computing infrastructure that made this work possible. We extend our appreciation to our colleagues and collaborators for their support and inspiration throughout this project.

References

Amine Abdaoui, Camille Pradel, and Grégoire Sigel. 2020. Load what you need: Smaller versions of mutililingual BERT. In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 119–123, Online. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset. *Preprint*, arXiv:1611.09268.

Muhammad Huzaifa Bashir, Aqil M Azmi, Haq Nawaz, Wajdi Zaghouani, Mona Diab, Ala Al-Fuqaha, and Junaid Qadir. 2023. Arabic natural language processing for qur'anic research: A systematic review. *Artificial Intelligence Review*, 56(7):6801–6854.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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 3615–3620, Hong Kong, China. Association for Computational Linguistics.

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

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.

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

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.

Daniel Gillick, Sayali Kulkarni, Larry Lansing, Alessandro Presta, Jason Baldridge, Eugene Ie, and Diego Garcia-Olano. 2019. Learning dense representations for entity retrieval. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 528–537, Hong Kong, China. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing. *CoRR*, abs/2007.15779.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun hsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *Preprint*, arXiv:1705.00652.

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, Toronto, Canada. Association for Computational Linguistics.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *CoRR*, abs/1904.05342.

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 and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

Dawn J Lawrie, Eugene Yang, Douglas W. Oard, and James Mayfield. 2022. Neural approaches to multilingual information retrieval. In *European Conference on Information Retrieval*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Rana Malhas and Tamer Elsayed. 2020. Ayatec: building a reusable verse-based test collection for arabic question answering on the holy qur'an. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(6):1–21.

Suraj Nair, Eugene Yang, Dawn J Lawrie, Kevin Duh, Paul McNamee, Kenton Murray, James Mayfield, and Douglas W. Oard. 2022. Transfer learning approaches for building cross-language dense retrieval models. *ArXiv*, abs/2201.08471.

Douglas W. Oard and B. Dorr. 1996. A survey of multilingual text retrieval. In *Proceedings of the AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*.

Douglas W. Oard and B. Dorr. 1998. Evaluating cross-language text filtering effectiveness. In *Proceedings of the AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*.

Vera Pavlova. 2023. Leveraging domain adaptation and data augmentation to improve qur'anic IR in English and Arabic. In *Proceedings of ArabicNLP 2023*, pages 76–88, Singapore (Hybrid). Association for Computational Linguistics.

Vera Pavlova. 2025. Multi-stage training of bilingual islamic LLM for neural passage retrieval. In *Proceedings of the New Horizons in Computational Linguistics for Religious Texts*, pages 42–52, Abu Dhabi, UAE. Association for Computational Linguistics.

Vera Pavlova and Mohammed Makhlouf. 2023. BIOptimus: Pre-training an optimal biomedical language model with curriculum learning for named entity recognition. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 337–349, Toronto, Canada. Association for Computational Linguistics.

Vera Pavlova and Mohammed Makhlouf. 2024. Building an efficient multilingual non-profit IR system for the islamic domain leveraging multiprocessing design in rust. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 981–990, Miami, Florida, US. Association for Computational Linguistics.

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

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

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, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, QiaoQiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021. RocketQAv2: A joint training method for dense passage retrieval and passage re-ranking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2825–2835, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Maxim Romanov and Masoumeh Seydi. 2019. Openiti: a machine-readable corpus of islamicate texts. *Zenodo, URL: https://doi. org/10.5281/zenodo*, 3082464.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS 2021) - Datasets and Benchmarks Track (Round 2)*.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

Laurens van der Maaten and Geoffrey E. Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

A Appendix

Our reduction method consists of the following steps:

1. We select English, Arabic, Russian, and Urdu texts from a multilingual variant of the C4 corpus (Raffel et al., 2019) and train and SentencePiece BPE tokenizer.

2. Find the intersection between the new tokenizer and the XLM-R$_{\text{Base}}$ tokenizer[3], the tokens inside of intersection and the corresponding weights will be selected for the new embedding matrix of the XLM-R4 model (34k tokens).

3. The encoder weights from XLM-R$_{\text{Base}}$ get copied to the new XLM-R4 model as is.

Evaluation of XLMR-4 on the XNLI dataset demonstrates only a slight drop in performance (around 1.77% across all languages) compared to the XLM-R$_{\text{Base}}$ (see Table 7). At the same time, we manage to significantly reduce the number of parameters by trimming the embedding matrix (EM) (see Table 6).

The domain adaptation of XLM-R4 takes the following steps:

1. We train a new SentencePiece BPE tokenizer using a multilingual Islamic Corpus and find the intersection between the new Islamic tokenizer and the XLM-R4 tokenizer. All the tokens outside of the intersection (9k tokens) are added to the embedding matrix of the XLMR-4 model, and the weights for new tokens are assigned by averaging existing weights of subtokens from the XLM-R4 model.

2. We continue pre-training XLM-R4 using the domain-specific corpus which gives us the XLM-R4-ID (Islamic domain) model. For more details on the hyperparameters, refer to Appendix A.

Model	Size	#params	EM
mBERT	714 MB	178 M	92 M
XLM-R$_{\text{Base}}$	1.1 GB	278 M	192 M
XLM-R4	481 MB	119 M	33M

Table 6: Comparison of models' size

Model	**en**	**ru**	**ar**	**ur**
XLM-R$_{\text{Base}}$	**84.19**	**75.59**	**71.66**	**65.27**
XLM-R4	83.21	72.75	70.48	64.95
mBERT	82.1	68.4	64.5	57
mBERT 15lang	82.2	68.7	64.9	57.1
DistillmBERT	78.5	63.9	58.6	53.3

Table 7: Results on cross-lingual transfer for four languages of the XNLI dataset. XLM-R$_{\text{Base}}$ and XLM-R4 results are averaged over five different seeds.

B Real User Metrics

During August 2025, the deployed system processed 84,255 requests originating from 2,630 unique IP addresses (median $\approx$ 32 requests/IP), with a total data volume of approximately 1.9 GB served. The cache hit rate by bytes was approximately 11%, consistent with a predominantly dynamic, search-intensive workload. Request counts exclude known automated traffic through Cloudflare bot-score and hosting-ASN filters. Clear diurnal usage patterns, distribution across residential and mobile ASNs, and a balanced mix of HTML, asset, and API requests all indicate genuine end-user activity. The system operated in CPU-only serving mode with an uptime of $\geq$99.97%. (All metrics are aggregated and privacy-preserving.)

On the user-facing site over the same period (via real-user monitoring, RUM), Success@5 was approximately 58–62% (τ = 15 s; results page viewed $\geq$15 s with top-5 results rendered). Abandonment was approximately 18–22%, defined as either no top-5 impression, dwell time <15 s, or a query reformulation within 30 s.

Human evaluation. We constructed a parallel set of 25 real-user queries in AR/UR/RU/EN. For each language, two independent annotators rated all retrieved results on a 3-point relevance scale (0 = irrelevant, 1 = partially relevant, 2 = highly relevant). Inter-annotator agreement was substantial (weighted Cohen's $\kappa \approx 0.61$). Compared to an XLM-R Base baseline, our model achieved consistent gains in nDCG@10 of +0.06 - +0.10 across all four languages. Example queries are provided in Table 10.

[3] https://huggingface.co/FacebookAI/xlm-roberta-base

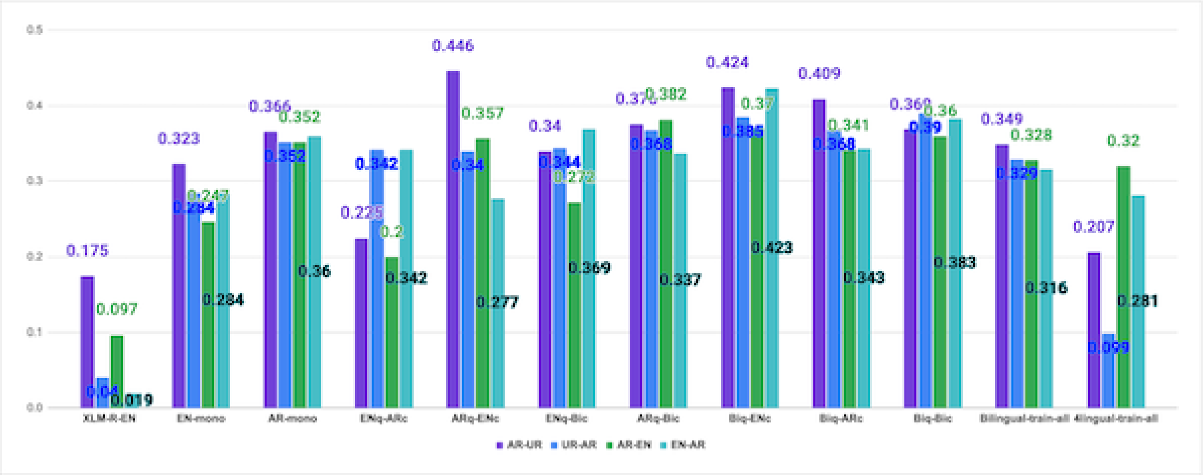

Figure 4: Comparison of model performance on cross-lingual evaluation.

Computing Infrastructure	1x H100 (80 GB)
Hyperparameter	**Assignment**
number of epochs	60
batch size	128
maximum learning rate	0.0005
learning rate optimizer	Adam
learning rate scheduler	None or Warmup linear
Weight decay	0.01
Warmup proportion	0.06
learning rate decay	linear

Table 8: Hyperparameters for pre-training of XLM-R4-ID model.

Computing Infrastructure	1x H100 (80 GB)
Hyperparameter	**Assignment**
number of epochs	10
batch size	256
learning rate	2e-5
pooling	mean

Table 9: Hyperparameters for training retrieval models.

Query ID	**Query**
1	How is the universe created?
2	What is the purpose of life of man on earth?
3	How is the fetus formed in the womb?
4	What function does the frontal lobe of the brain have?
5	How is the rain created?
6	What is the condition at the depth of the sea?
7	Why do mountains stand still on the surface of the earth?
8	Can animals communicate in their own languages?
9	Will we be held accountable for our deeds?
10	What is Hijab?
11	Will the world come to an end and how will it happen?
12	Who was Jesus?

Table 10: Examples of real user queries