Multilingual RAG: Mitigating Cross-Language Noise for Contextually Aligned Retrieval

Anonymous ACL submission

Abstract

Large Language Models (LLMs) excel in cross-lingual question answering (QA) but often hallucinate due to mismatches between the query and the model's internal knowledge representation. Retrieval-augmented generation (RAG) mitigates this issue but struggles with cross-lingual retrieval inconsistencies. We propose a retrieval method that enhances recall and re-ranking by improving semantic alignment across languages. Our approach integrates a language-aware retrieval mechanism with a fine-tuned encoder model, ParaXLM-SR, to refine query-context matching and prioritize relevant information. By leveraging bias-adjusted similarity re-ranking, our method further mitigates cross-lingual retrieval noise and improves context relevance.

1 Introduction

RAG has emerged as a valuable method for enhancing LLMs by incorporating an external retrieval mechanisms to improve response quality and factual grounding (Lewis et al., 2020; Gao et al., 2023). By retrieving semantically relevant context before generation, RAG mitigates knowledge gaps and reduces the hallucination problem in LLMs, thereby increasing the factual reliability of generated responses. RAG has gained significant attention, particularly in tasks requiring knowledge-intensive reasoning, such as open-domain question answering and specialized domain-specific text generation (Chirkova et al., 2024a; Hu and Lu, 2024b). Despite these advancements, most RAG-based retrieval systems are still predominantly monolingual, with English being the primary language for both retrieval and generation.

To overcome these challenges, we introduce a cross-lingual retrieval question-answering RAG designed to improve multilingual retrieval effectiveness and reduce hallucinations in knowledge-intensive tasks on figure 1. Our framework employs a multi-step process that combines cross-lingual retrieval and re-ranking mechanisms to identify semantically aligned information from diverse languages. This approach ensures that LLM-generated responses are firmly grounded in reliable multilingual sources. Instead of restricting retrieval to monolingual knowledge bases, our approach dynamically queries multiple language-specific repositories, retrieving the most relevant content irrespective of the query's original language (Artetxe and Schwenk, 2019a; Ji et al., 2023). Through comprehensive evaluations on cross-lingual benchmarks, we demonstrate that our method significantly enhances cross-lingual recall while mitigating the hallucination rate. By integrating multilingual evidence retrieval into the RAG pipeline, our approach strengthens the factual consistency of LLM-generated outputs, thereby making retrieval-augmented architectures more reliable and linguistically inclusive.

ParaXLM-SR (Paraphrase-XLM-RoBERTa Semantic Retrieval) is a fine-tuned multilingual sentence encoder designed to enhance cross-lingual semantic alignment for retrieval tasks. Built upon `paraphrase-xlm-r-multilingual-v1`, this model refines sentence embeddings by optimizing for semantic relatedness, improving the alignment of multilingual representations while preserving retrieval robustness. Unlike its predecessor, ParaXLM-SR incorporates domain-adaptive fine-tuning on a filtered, high-quality multilingual dataset, ensuring that semantically similar question-context pairs exhibit greater coherence in vector space. By reducing retrieval noise and enhancing cross-lingual consistency, ParaXLM-SR enables more precise retrieval of multilingual knowledge, making it particularly effective for retrieval-augmented generation (RAG) pipelines and cross-lingual QA scenarios.

Our Key Findings. Through extensive evaluation, we observe the following improvements:

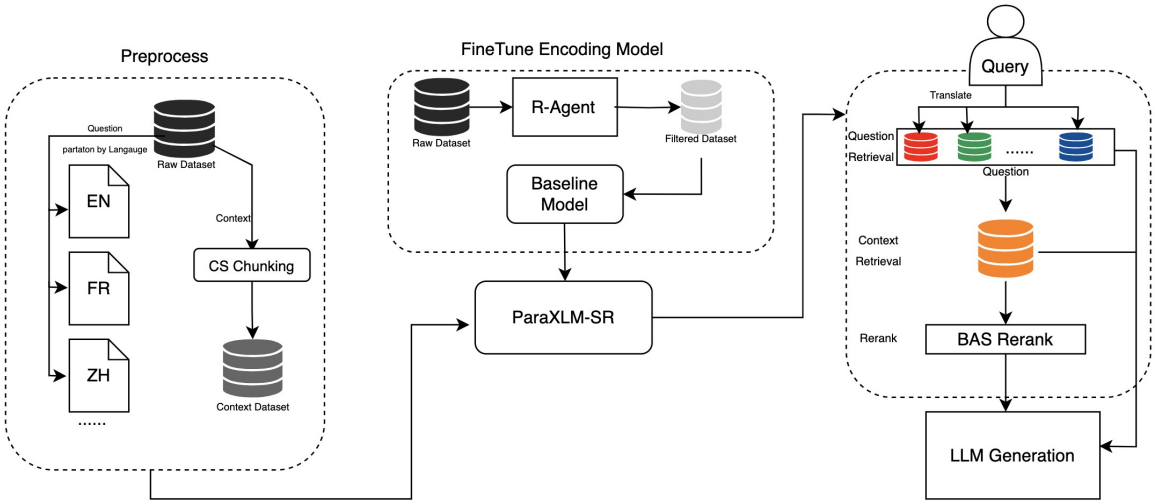

Figure 1: Illustration of the cross-lingual RAG architecture pipeline. Preprocessing partitions data by language with context-specific (CS) chunking. The R-Agent filters training data, enabling optimized retrieval. Queries are translated across languages, and retrieved contexts are re-ranked using the Bias-Adjusted Similarity (BAS) mechanism to enhance cross-lingual precision.

- **Embedding Alignment:** ParaXLM-SR reduces vector space distortion across languages, leading to more precise semantic similarity computations.
- **Retrieval Efficiency:** Our RAG framework enhances retrieval accuracy across multiple language pairs, demonstrating greater adaptability to cross-lingual QA tasks.
- **Re-ranking Stability:** The bias-adjusted reranking mechanism improves ranking consistency, particularly in typologically diverse languages, while preserving monolingual ranking quality.

Despite these advancements, challenges persist in retrieving high-quality results in low-resource languages, adapting retrieval mechanisms for complex linguistic structures, and refining ranking functions for domain-specific knowledge.

2 Related Work

2.1 Multilingual Sentence Embeddings and Retrieval Models

Multilingual sentence embeddings are fundamental to cross-lingual retrieval-augmented QA, where semantic misalignment often leads to retrieval discrepancies. Early distributed representation models, such as Word2Vec (Mikolov et al., 2013a,b), provided word-level embeddings but lacked contextual depth and cross-lingual adaptability. Subsequent models, including LASER (Artetxe and Schwenk, 2019b; Heffernan et al., 2022), Universal Sentence Encoder (Yang et al., 2020), and LaBSE (Feng et al., 2020), leveraged large-scale transformer architectures to enhance cross-lingual embedding alignment.

Advancements in transformer-based architectures, such as mBERT and XLM-RoBERTa, have further refined multilingual representation learning (Muller et al., 2021). RoBERTa introduced improved contextual embeddings through optimized pretraining objectives (Dadas et al., 2020a). To mitigate retrieval inconsistencies, `stsb-xlm-r-multilingual` (Huertas, 2023), a fine-tuned variant of XLM-RoBERTa, has been employed to filter weakly related question-context pairs, thereby improving retrieval precision. Several multilingual retrieval models have emerged, including `paraphrase-multilingual-mpnet-base-v2`, which is optimized for paraphrase identification, and `LaBSE`, which employs a dual-encoder framework for cross-lingual similarity. Lawrie et al. (Lawrie et al., 2022) introduced Multilingual ColBERT-X, a late interaction retrieval model that enhances multilingual dense retrieval by leveraging token-level representations. `paraphrase-xlm-r-multilingual-v1` (Reimers and Gurevych, 2019a), an extension of XLM-R, has been widely adopted for multilingual retrieval, while knowledge distillation techniques, such as

`distiluse-base-multilingual-cased-v2` and `distilbert-multilingual-nli-stsb-quora`, have enhanced model efficiency. Additionally, `xlm-r-bert-base-nli-stsb-mean-tokens` fine-tunes XLM-R for sentence similarity tasks, `LASER` enables zero-shot cross-lingual transfer, and `GloVe` (Pennington et al., 2014) serves as a non-contextual baseline.

2.2 Multilingual Retrieval Strategies

Cross-lingual retrieval presents unique challenges, particularly for low-resource languages, where naive corpus-wide retrieval amplifies retrieval mismatches and semantic drift. Chirkova et al. (Chirkova et al., 2024b) systematically analyzed multilingual retrieval pipelines, underscoring the significance of fine-tuned retrievers, re-ranking strategies, and retrieval optimization in mitigating hallucinations. Feng et al. (Feng et al., 2022) investigated the constraints of code-switching in cross-lingual transfer. Wang et al. (Wang et al., 2024) introduced retrieval partitioning, demonstrating that structured retrieval units improve both efficiency and recall. While their M-RAG framework applies structured retrieval to generative QA, our approach extends this paradigm to multilingual question answering by aligning queries with semantically relevant cross-lingual contexts.

Cosine similarity has been widely adopted as a fundamental metric for ranking retrieved candidates in multilingual retrieval models (Conneau and Lample, 2019; Reimers and Gurevych, 2019a). However, inherent biases in multilingual representations necessitate further refinement to ensure equitable ranking across languages. While prior approaches have explored metadata-based retrieval constraints (Pires et al., 2019), these methods often fail to generalize across typologically diverse languages. Our framework introduces a bias-adjusted re-ranking mechanism that refines cross-lingual retrieval precision while maintaining ranking stability for monolingual retrieval candidates.

2.3 Retrieval-Augmented Generation and Cross-Lingual Adaptations

RAG has become a key framework for integrating external knowledge retrieval into language models, improving response factuality in knowledge-intensive tasks (Finardi et al., 2024a; Hu and Lu, 2024a; Fan et al., 2024). Despite these advancements, most existing RAG systems operate in monolingual environments, limiting their effectiveness in multilingual retrieval (Sharma et al., 2024). Prior studies have explored domain adaptation techniques for RAG in open-domain QA (Siriwardhana et al., 2023; Veturi et al., 2024) and contextual response generation (Jin et al., 2024), offering insights into retrieval generalization.

Cross-lingual retrieval introduces additional challenges stemming from embedding space discrepancies and linguistic divergence. Recent efforts have proposed retrieval error correction techniques (Muller et al., 2021; Finardi et al., 2024b), but effective re-ranking strategies tailored to multilingual QA remain underexplored. Our approach introduces a bias-adjusted similarity re-ranking mechanism inspired by retrieval error compensation, optimizing ranking precision by addressing systematic cross-lingual retrieval biases. Moreover, retrieval caching mechanisms, such as RAGCache (Jin et al., 2024), emphasize the importance of improving retrieval efficiency, further aligning with our structured retrieval design.

Prior work has also investigated translation-based augmentation, such as using Google Translate to enhance retrieval (Balk et al., 2012). Our approach extends this by constructing a bias matrix that systematically refines cross-lingual ranking adjustments while mitigating retrieval distortions caused by linguistic variation.

3 Methodology

A cross-lingual QA task aims to retrieve the most semantically relevant context c_j for a given query q_i from a multilingual QA dataset $\mathcal{D} = (q_i, c_j, l_{ij})i = 1^{|\mathcal{D}|}$, where lij represents the synthetic language label indicating the language of the question and context. (e.g., "ende" signifies a question in English with a context in German). Unlike monolingual retrieval, cross-lingual retrieval introduces challenges such as semantic misalignment, retrieval mismatches, and linguistic variation across typologically diverse languages.

In the multi-step retrieval process, we use cosine similarity to measure the relevance between a query question and a context. Given a question q_i from the question set $\mathcal{Q}_l$ and a context c_j from the multilingual context set $\mathcal{C}$, the cosine similarity is calculated as:

$$S(q_i, c_j) = \frac{E(q_i) \cdot E(c_j)}{\|E(q_i)\|\|E(c_j)\|} \quad (1)$$

Cosine similarity is utilized in our retrieval pipeline to measure the angular distance between query

and context embeddings, prioritizing semantically aligned content across multiple languages.

Language	Question / Context	Answer
France	**Dans la citation biblique chrétienne, il est plus ...?**	Needle
German	... durch ein Nadelöhr zu gehen, ...	
Spanish	**¿Llevó a cabo una misión de predicación y ...?**	Jesus
German	Jesus führte etwa im Jahr 28–30 n. Chr. eine Predigt-...	
English	**The company Titleist manufacture which ... ?**	Golf
French	Retrouvez les fabricants de golf qui fabriquent le ...	
Spanish	**¿Un cóctel Molotov (bomba molotov casera) ... ?**	Molotov
English	petrol bomb [C20: named after V. M. Molotov] ...	
Swahili	**Sauti ya vokali isiyosisitizwa 'uh' kwa kawaida ... ?**	schwa
Swahili	inayoitwa schwa, inawakilishwa na kichwa chini ...	
Chinese	第一个被发现的抗生素是什么	Penicillin
German	Das erste Antibiotikum Penicillin wurde 1928 von ...	

Table 1: Examples of multilingual question-answering data. Each row consists of a question, corresponding context, and answer across different language pairs.

3.1 Fine-Tuning the Encoder Model

After R-Agent filtering, the dataset $\mathcal{D}_{\text{filtered}}$ is used to fine-tune the baseline retrieval model, `paraphrase-xlm-r-multilingual-v1` (Dadas, 2020), a variant of Sentence-BERT (SBERT) (Reimers and Gurevych, 2019b). The fine-tuning process updates the pooling layer and fully connected layer to optimize sentence similarity representation, enabling the encoder model to generate embeddings that improve the retrieval of semantically relevant contexts.

Given a batch of training instances $(q_i, c_i) \in \mathcal{D}_{\text{filtered}}$, the model computes the similarity logits $f(q_i, c_i)$ through a fully connected projection layer:

$$f(q_i, c_i) = W^\top \operatorname{pool}(E(q_i), E(c_i)) + b. \quad (2)$$

Here, the pooling operation refers to **average pooling**, where the final sentence representation is obtained by computing the mean over all token embeddings. This method ensures a stable and smooth aggregation of contextual information while reducing noise, making it effective for multilingual retrieval tasks. W is the learnable weight matrix, and b is the bias term. The objective function is the cross-entropy loss:

$$\mathcal{L} = -\sum_{i=1}^{|\mathcal{D}_{\text{filtered}}|} y_i \log f(q_i, c_i) + (1 - y_i) \log(1 - f(q_i, c_i)). \quad (3)$$

The binary label y_i indicates whether the pair (q_i, c_i) is semantically related, with $y_i = 1$ representing a semantically relevant pair and $y_i = 0$ indicating an irrelevant pair.

3.2 Partitioning and Sharding of Embedded Data

To improve retrieval efficiency and mitigate semantic misalignment in cross-lingual question answering, we introduce a structured retrieval approach that consists of two key components: partitioning and sharding. These mechanisms ensure that retrieval occurs within semantically and linguistically aligned spaces, reducing noise and improving cross-lingual recall.

Partitioning for Language-Specific Question Retrieval We first partition the embedded question representations based on their respective language labels. Formally, the partitioned question sets are defined as:

$$\mathcal{Q} = \{\mathcal{Q}_{l_{q_i}}\}, \quad \mathcal{Q}_{l_{q_i}} = \{q_i \mid (l_{q_i}, l_{c_j}) = \text{split}(l_{ij})\}. \quad (4)$$

Each partition $\mathcal{Q}lq_j$ contains only questions in a single language l_{q_j}. Since each question-context pair (q_i, c_j) in the dataset is assigned a synthetic language label l_{ij}, we first decompose it into its respective question and context languages, l_{q_i} and l_{c_j}. During retrieval, a given query q' in language $l_{q'}$ is first translated into its corresponding language $\hat{q}'$ before being embedded into the same vector space as $\mathcal{Q}lq'$. This ensures that similarity computations occur within a shared semantic representation space, minimizing retrieval mismatches.

Sharding for Question-Context Alignment To enhance retrieval accuracy, we introduce a sharding mechanism that decouples question embeddings and context embeddings while preserving their alignment within a multilingual retrieval space. Formally, the sharded dataset is structured as:

$$\mathcal{E} = \{\mathcal{E}_{(l_{q_j}, l_{c_k})}\},$$
$$\text{where } \mathcal{E}_{(l_{q_j}, l_{c_k})} = \{(E(q_j), E(c_k)) \mid (l_{q_j}, l_{c_k})\}. \quad (5)$$

Retrieval is conducted in two sequential stages. In the first stage, a query q' is compared against the question set $\mathcal{Q}_{l_{q'}}$ within its respective language partition. If the similarity score $S(q', q_j)$ exceeds a predefined threshold k_q, the retrieved question q_j is deemed semantically relevant and is subsequently used to retrieve the corresponding context. In the second stage, the most relevant context c^* is selected by maximizing the similarity score between the identified question and its candidate contexts:

$$c^* = \arg\max_{c_k} \mathcal{E}(E(q_j), E(c_k)), \quad c_k \in \mathcal{E}_{l_{q_j}}. \quad (6)$$

This two-step retrieval mechanism ensures that the initial retrieval step prioritizes semantically aligned questions before searching for the most contextually relevant passage. By enforcing this hierarchical retrieval structure, we mitigate retrieval noise, particularly in cross-lingual settings where direct query-to-context alignment may introduce inconsistencies.

R-Agent Effective retrieval in cross-lingual question answering requires high-quality training data that aligns semantically between the question and its associated context. However, large-scale QA datasets such as TriviaQA(Joshi et al., 2017) contain noise, where retrieved contexts may be weakly related or irrelevant to the corresponding questions. To mitigate this issue, we apply a filtering mechanism using a sentence similarity model before fine-tuning the retrieval encoder.

To refine the training data, we filter question-context pairs $(q_i, c_i) \in \mathcal{D}$ based on their semantic similarity score $S(q_i, c_i)$. Pairs with a similarity score below a predefined threshold τ are excluded, ensuring that only strongly correlated instances remain:

$$\mathcal{D}_{\text{filtered}} = \{(q_i, c_i) \mid S(q_i, c_i) \geq \tau\}. \quad (7)$$

By leveraging the R-Agent, we enhance the dataset quality, removing weakly related question-context pairs and ensuring that the retrieval model is trained on semantically robust data.

3.3 Cross-Language Reranking Algorithm

In cross-lingual retrieval tasks, semantic misalignment between languages can introduce errors, leading to suboptimal ranking of retrieved contexts. Standard similarity metrics, such as cosine similarity, do not account for the inherent variation in embeddings caused by differences in language structure. To address this issue, we introduce a cross-language re-ranking algorithm that applies weighted bias adjustment to the similarity scores. This method enhances retrieval accuracy by incorporating language-specific error adjustments.

Construction of the Bias Matrix To quantify language-induced variations in embedding similarity, we generate a cross-language error matrix $\mathbf{B}$. First, we select 200,000 context passages from TriviaQA and translate each passage into seven different target languages using Google Translate. Let $\mathcal{C} = \{c_i^l\}_{i=1}^N$ represent the translated contexts, where c_i^l denotes the i-th context in language l. Using our fine-tuned encoder model, each translated passage is embedded into a vector space:

$$v_i^l = E(c_i^l), \quad v_i^{l'} = E(c_i^{l'}). \quad (8)$$

For each context passage, we compute the pairwise embedding difference across languages:

$$\delta(l, l') = \frac{1}{N} \sum_{i=1}^{N} \|v_i^l - v_i^{l'}\|_2. \quad (9)$$

The resulting error matrix $\mathbf{B}$, where $B_{l,l'} = \delta(l, l')$, represents the average embedding discrepancy between languages l and l'. Higher values in $\mathbf{B}$ indicate greater semantic drift in embeddings due to language translation.

Bias-Adjusted Similarity Reranking (BAS Rerank) Given a query q in language l_q and a retrieved context c in language l_c, we apply a bias correction factor to the similarity score. Let $S(q, c)$ denote the initial cosine similarity between the query and context embeddings. The adjusted similarity $\tilde{S}(q, c)$ is computed as:

$$\tilde{S}(q, c) = \frac{S(q, c)}{1 - (B_{l_q, l_c} \cdot \alpha)}, \quad (10)$$

where α is a tunable weight that determines the influence of cross-language error correction. If no bias value exists for the given language pair, the similarity remains unchanged. This adjustment accounts for systematic language-specific variations and improves ranking fairness by reducing the impact of embedding distortions.

Interpretability of Weight Adjustment The parameter α allows control over the degree of bias correction. Setting $\alpha = 0$ recovers the standard cosine similarity, while higher values increase the correction strength. This flexibility enables optimization based on empirical performance across different language pairs. The re-ranking process ensures that retrieved contexts are ranked more reliably, particularly in low-resource languages where translation discrepancies are more pronounced.

To evaluate retrieval effectiveness, we employ the **Error Rate (ER)**, which quantifies the proportion of queries where the ground-truth context is absent from the top-K retrieved results. Formally, ER is defined as:

$$\text{ER}_K = \frac{1}{|\mathcal{Q}|} \sum_{i=1}^{|\mathcal{Q}|} I\left(c_i^* \notin \{c_1, ..., c_K\}\right), \quad (11)$$

A lower ER_K value indicates a higher retrieval success rate, making it a key metric for evaluating the effectiveness of cross-lingual retrieval strategies.

4 Data Preprocessing

Given that TriviaQA is originally an English dataset, we begin with an initial preprocessing step before multilingual translation to ensure data integrity and relevance. The first step involves filtering out instances where the ground-truth answer is not sufficiently represented within the associated context. Specifically, for each (q_i, c_j, a_k) triplet, a filtering mechanism is applied based on answer presence criteria. A context c_j is retained if the answer a_k satisfies at least one of the following conditions: (1) it appears as an exact substring within the context, (2) at least 75% of its tokens are present in the context, or (3) all its constituent words occur in any order within the context. Only instances meeting at least one of these criteria are preserved, ensuring that the dataset remains semantically.

To enhance adaptability to model constraints, we apply **context-specific (CS) chunking** as a preprocessing strategy for handling extended contexts while preserving semantic integrity. Unlike fixed-length truncation, CS chunking segments text at natural discourse boundaries, such as sentences or paragraphs, ensuring structural coherence and retention of key informational elements. Additionally, a lexical search mechanism is integrated to guarantee that the segmented context retains the complete answer span. This adaptive chunking framework balances the trade-off between maintaining linguistic continuity and adhering to model input limitations.

To ensure balanced multilingual representation and improve retrieval efficiency, the preprocessed TriviaQA dataset is partitioned by language into subsets. Initially, the dataset is divided into seven equal parts, and each of these subsets is further subdivided into seven smaller partitions. This stratified approach ensures a uniform distribution of question-context pairs across the dataset and mitigates any bias caused by imbalanced language combinations. After partitioning, both the questions and contexts in these subsets are translated into seven languages: English, French, German, Russian, Swahili, Spanish, and Chinese. Given that the answers remain in English, this translation strategy ensures comprehensive coverage of cross-lingual retrieval scenarios while preserving answer consistency across all languages. Notably, Swahili, a typologically distinct Bantu language, presents unique structural challenges due to its divergence from Indo-European languages, providing an opportunity to evaluate the model's performance on languages with significant typological variation.

In addition to TriviaQA, we incorporate TydiQA (Clark et al., 2020), a dataset containing 200K question-answer pairs spanning 11 typologically diverse languages, to assess cross-lingual generalization in retrieval tasks. Both datasets undergo partitioning and sharding to optimize retrieval efficiency and support multilingual evaluation. To accommodate the input size constraints of the encoder model, CS chunking is applied to contexts that exceed the maximum input length. This ensures that longer contexts are divided into semantically coherent units, preserving crucial information for question-context alignment while adhering to model constraints. The integration of structured dataset partitioning, language-aware translation, and adaptive chunking enables the construction of a balanced, high-fidelity multilingual dataset suitable for cross-lingual retrieval and evaluation.

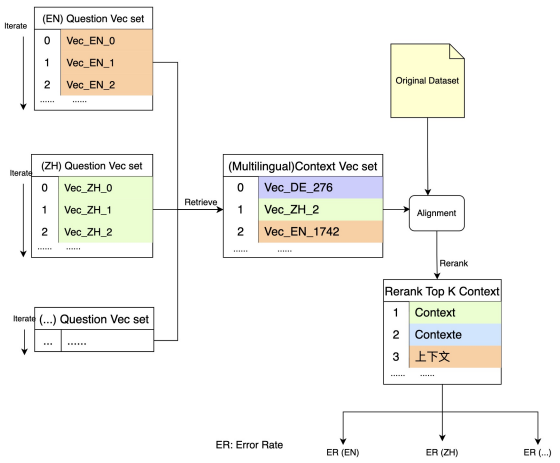

Figure 2: Pipeline of the experimental framework, including multi-step retrieval and re-ranking. Retrieved contexts are ranked, and the top-k results are checked for the true context and language identification to ensure cross-lingual alignment. We report the textual language of one example vector for English and Chinese vector setsdseq o90.

5 Experiments

As illustrated in Figure 2, the experimental pipeline consists of multi-step retrieval followed by a re-

ranking mechanism. Retrieved contexts are ranked based on their semantic similarity to the query, and the top-k results are analyzed to determine whether the correct context is present. Additionally, language identification is performed to assess cross-lingual retrieval accuracy, ensuring alignment between queries and retrieved content across diverse language pairs.

5.1 Finetuning RAG Encoder

After dataset refinement, the fine-tuning process is carried out on the `paraphrase-xlm-r-multilingual-v1` model. In this step, each question-context pair is labeled to indicate whether it is semantically related or not. If the pair is semantically related, it is assigned a label of 1, and if it is deemed irrelevant, it is assigned a label of 0. The model's pooling and fully connected layers are fine-tuned for 10 epochs to optimize the model's ability to compare and align semantically similar question-context pairs in a cross-lingual context. The fine-tuned model, referred to as the **ParaXLM-SR**, improves the retrieval performance by ensuring that semantically related pairs are better aligned, while irrelevant pairs are more dispersed in the embedding space.

The fine-tuning process utilizes the AdamW optimizer, with the learning rate controlling the step size during the gradient-based optimization. The training proceeds iteratively over T epochs. Once fine-tuning is completed, the optimized model is saved and used for retrieval tasks in subsequent applications.

To comprehensively assess the effectiveness of our proposed framework, we conduct a two-stage evaluation: (1) an intrinsic evaluation of the fine-tuned encoder model to measure its performance in semantic representation and cross-lingual alignment, and (2) an extrinsic evaluation of the full RAG architecture to analyze its retrieval efficacy and robustness in multilingual question-answering tasks.

5.2 Evaluation

5.2.1 Evaluation of Encoder Model

ParaXLM-SR was evaluated across multiple linguistic benchmarks by SentEval, a standard benchmarking framework for sentence embeddings (Conneau and Kiela, 2018) to assess its effectiveness in semantic representation (Table 2), paraphrase identification, natural language inference, and topic classification. In terms of **semantic relatedness**, and outperformed the baseline in the **CDSC-R** and **SICK-R** benchmarks(Wróblewska and Krasnowska-Kieraś, 2017; Zhang et al., 2019), CDSC-R evaluates sentence representations based on compositional distributional semantics in Polish, measuring the degree of semantic similarity between sentence pairs, while SICK-R assesses semantic relatedness through graded similarity scores across English sentence pairs, focusing on fine-grained distinctions in meaning and paraphrase variability. Additionally, the model exhibited superior performance in **paraphrase identification**, surpassing the baseline in the cross-lingual **PPC Benchmark**, indicating improved alignment in cross-lingual paraphrase detection. For **natural language inference**, the fine-tuned model achieved higher score in the **CDSC-E** benchmark (Wróblewska and Krasnowska-Kieraś, 2017), confirming its robustness in logical entailment recognition. Furthermore, in **topic classification**, the model attained competitive results on the **8-Tags Benchmark** (Dadas et al., 2020b), showcasing its capability to generalize across diverse classification tasks.

5.2.2 Evaluation of the RAG Architecture

In evaluating the proposed RAG architecture of ParaXLM-SR, Table 3 analyze Top-10 and Top-3 retrieval error rates on TriviaQA and TydiQA to report cross-lingual retrieval performance and compare with existing. It shows that Swahili exhibits higher error rates due to its typological divergence from Indo-European languages. ParaXLM-SR achieves notable improvements in cross-lingual retrieval, particularly in typologically closer language pairs. The BAS re-ranking algorithm further refines retrieval by reducing semantic misalignment, as evidenced by the Top-3 error rate converging with Top-10, demonstrating its efficacy in prioritizing semantically relevant contexts while mitigating retrieval noise.

To further evaluate the robustness of the proposed RAG architecture, we conduct experiments on the TydiQA dataset with additional retrieval challenges on longer contextual passages. To ensure compatibility with model constraints, we employ CS chunking, segmenting extended passages into 512-token units while preserving semantic integrity. Furthermore, the proposed RAG architecture was also evaluated on five languages in TydiQA which are absent from the training data, pre-

	Semantic Relatedness		Paraphrase Identification	Semantic Analysis		Natural Language Inference		Topic Classification
	CDSC-R	SICK-R	PPC	WCCRS HOTELS	WCCRS MEDICINE	CDSC-E	SICK-E	8TAGS
mbert	0.823	0.650	65.7	73.68	71.51	83.2	70.53	64.96
xlmr_base_all	0.795	0.712	67.2	83.72	80.89	87.7	75.19	69.60
roberta_base_all	0.819	0.759	72.4	84.62	82.62	86.4	75.68	70.04
paraphrase-multilingual-mpnet-base-v2	0.901	0.831	80.3	84.49	**82.67**	86.7	79.47	**72.58**
LaBSE	0.883	0.826	77.7	**86.16**	80.64	87.2	81.63	71.41
paraphrase-xlm-r-multilingual-v1	0.898	0.838	80.6	81.66	80.69	86.9	81.29	70.61
distiluse-base-multilingual-cased-v2	0.875	0.807	76.4	79.86	75.95	**87.9**	78.68	70.86
distilbert-multilingual-nli-stsb-quora-ranking	0.862	0.816	81.1	79.92	75.11	86.9	80.27	68.96
xlm-r-bert-base-nli-stsb-mean-tokens	0.853	0.820	**81.5**	81.72	79.26	85.2	80.31	69.37
LASER	0.880	0.816	81.1	82.50	77.78	87.7	**81.96**	64.98
GloVe	0.834	0.732	67.7	77.41	69.78	87.4	73.32	67.89
ParaXLM-SR	**0.909**	**0.842**	81.4	81.15	77.09	87.6	81.10	71.71

Table 2: Performance of ParaXLM-SR vs. Baseline on Semantic and Language Tasks. We report accuracy classification and Pearson correlation between true and predicted relatedness scores for semantic relatedness.

	Baseline		ParaXLM-SR		Re-Rank
	Top 10	Top 3	Top 10	Top 3	Top 3
TriviaQA					
en	0.123	0.213	0.058 (-0.065)	0.137 (-0.076)	0.068 (-0.069)
fr	0.177	0.286	0.126 (-0.077)	0.209 (-0.103)	0.111 (-0.098)
de	0.215	0.345	0.138 (-0.023)	0.242 (-0.033)	0.154 (-0.088)
ru	0.351	0.492	0.199 (-0.055)	0.326 (-0.039)	0.097 (-0.229)
sw	0.793	0.856	0.770 (-0.793)	0.822 (-0.856)	0.771 (-0.056)
es	0.081	0.173	0.060 (-0.081)	0.130 (-0.173)	0.074 (-0.099)
zh	0.172	0.250	0.117 (-0.172)	0.211 (-0.250)	0.153 (-0.058)
avg w. sw	0.273	0.373	0.210 (-0.063)	0.297 (-0.169)	0.204 (-0.093)
avg w.o. sw	0.187	0.293	0.117 (-0.070)	0.210 (-0.184)	0.109 (-0.101)
TydiQA					
ar	0.691	0.765	0.473 (-0.218)	0.615 (-0.150)	0.627 (0.012)
bn	0.702	0.828	0.557 (-0.145)	0.719 (-0.109)	0.733 (0.014)
en	0.326	0.404	0.255 (-0.071)	0.319 (-0.085)	0.326 (0.007)
fi	0.339	0.474	0.305 (-0.034)	0.451 (-0.023)	0.465 (0.014)
id	0.405	0.582	0.441 (0.036)	0.623 (0.041)	0.642 (0.019)
ko	0.580	0.713	0.437 (-0.144)	0.599 (-0.114)	0.616 (0.017)
ru	0.486	0.603	0.322 (-0.164)	0.514 (-0.089)	0.523 (0.009)
sw	0.730	0.806	0.443 (-0.287)	0.617 (-0.190)	0.629 (0.012)
avg	0.532	0.647	0.404 (-0.128)	0.557 (-0.090)	0.570 (0.013)

Table 3: Top 10 and 3 WER Evaluation of ParaXLM-SR, Baseline, and Re-Ranking across TriviaQA and TydiQA datasets.

senting a rigorous zero-shot evaluation scenario for cross-lingual retrieval.

BAS re-ranking adjusts similarity scores only for cross-lingual question-context pairs, while monolingual pairs remain unchanged. Therefore, the Top-3 error rate post re-ranking cannot fall below retrieval, since only cross-lingual candidates are modified. Empirical results show that the post-re-ranking Top-3 error rate is merely 0.013 higher than retrieval, indicating minimal disruption to monolingual ranking while effectively refining cross-lingual retrieval accuracy.

6 Conclusion and Limitations

We present a cross-lingual retrieval-augmented question-answering framework that improves retrieval accuracy across multiple languages while addressing retrieval biases. ParaXLM-SR fine-tuned on a semantically filtered multilingual dataset, enhances cross-lingual semantic alignment by reducing retrieval discrepancies. Experimental results on TriviaQA and TydiQA demonstrate its effectiveness in minimizing retrieval errors and improving ranking consistency. Additionally, BAS re-ranking compensates for cross-lingual discrepancies while maintaining monolingual retrieval stability, further refining retrieval fidelity in multilingual settings.

Despite the demonstrated improvements in retrieval accuracy, certain limitations remain. First, due to the absence of real-world user queries in our dataset, we did not evaluate the retrieval performance when user-generated queries are used instead of pre-defined questions. Future research will focus on refining the retrieval mechanism to better handle diverse user queries in practical applications. Second, given the rapid advancements in LLMs, our study did not evaluate the generation performance of our RAG architecture with contemporary LLM iterations. Since LLM architectures are continuously evolving, subsequent studies will explore how our framework performs when integrated with modern LLM-based text generation, ensuring robust end-to-end retrieval-augmented generation performance.

References

Mikel Artetxe and Holger Schwenk. 2019a. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. Transactions of the Association for Computational Linguistics, 7:597–610.

Mikel Artetxe and Holger Schwenk. 2019b. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. Transactions of the Association for Computational Linguistics, 7:597–610.

E. M. Balk, M. Chung, and N. Hadar. 2012. Accuracy of data extraction of non-english language trials with google translate.

Nadezhda Chirkova, David Rau, Hervé Déjean, Thibault Formal, Stéphane Clinchant, and Vassilina Nikoulina. 2024a. Retrieval-augmented generation in multilingual settings. arXiv preprint arXiv:2407.01463.

Nadezhda Chirkova, David Rau, Hervé Déjean, Thibault Formal, Stéphane Clinchant, and Vassilina Nikoulina. 2024b. Retrieval-augmented generation in multilingual settings. In Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024), pages 177–188. Association for Computational Linguistics.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. Preprint, arXiv:2003.05002.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. Preprint, arXiv:1803.05449.

Alexis Conneau and Guillaume Lample. 2019. Unsupervised cross-lingual representation learning at scale. In Proceedings of ACL.

S. Dadas, M. Perełkiewicz, and R. Poswiata. 2020a. Pretraining polish transformer-based language models at scale. In International Conference on Artificial Intelligence and Soft Computing, pages 301–314. Springer.

Szymon Dadas, Mateusz Perełkiewicz, and Rafał Poswiata. 2020b. Evaluation of sentence representations in polish. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 1674–1680, Marseille, France. European Language Resources Association.

Sławomir Dadas. 2020. Training effective neural sentence encoders from automatically mined paraphrases.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. arXiv preprint.

Fanchao Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. arXiv preprint.

Yukun Feng, Feng Li, and Philipp Koehn. 2022. Toward the limitation of code-switching in cross-lingual transfer. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 5966–5971, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Paulo Finardi, Leonardo Avila, Rodrigo Castaldoni, Pedro Gengo, Celio Larcher, Marcos Piau, Pablo Costa, and Vinicius Caridá. 2024a. The chronicles of rag: The retriever, the chunk and the generator. arXiv preprint.

Paulo Finardi, Leonardo Avila, Rodrigo Castaldoni, Pedro Gengo, Celio Larcher, Marcos Piau, Pablo Costa, and Vinicius Caridá. 2024b. The chronicles of rag: The retriever, the chunk and the generator.

Yufei Gao, Yuan Xiong, Xiaoyu Gao, and Jie Tang. 2023. Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. arXiv preprint.

Yucheng Hu and Yuxing Lu. 2024a. Rag and rau: A survey on retrieval-augmented language model in natural language processing. arXiv preprint.

Yucheng Hu and Yuxing Lu. 2024b. Rag and rau: A survey on retrieval-augmented language models in natural language processing. arXiv preprint arXiv:2406.14567.

Javier Huertas. 2023. Multilingual semantic textual similarity benchmark (stsb). GitHub Repository.

Zhijing Ji, Nayeon Lee, Ruitong Frieske, Tong Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. A survey on hallucination in natural language generation. arXiv preprint arXiv:2303.01117.

C. Jin, Z. Zhang, and X. Jiang. 2024. Ragcache: Efficient knowledge caching for retrieval-augmented generation. arXiv preprint.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Dawn Lawrie, Eugene Yang, Douglas W Oard, and James Mayfield. 2022. Multilingual colbert-x. arXiv preprint arXiv:2209.01335.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and Sebastian Riedel. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In Proceedings of NeurIPS.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In International Conference on Learning Representations.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In Neural Information Processing Systems, pages 3111–3119.

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021. First align, then predict: Understanding the cross-lingual ability of multilingual bert. In Proceedings of ACL.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In Proceedings of ACL.

Nils Reimers and Iryna Gurevych. 2019a. 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. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019b. Sentence-bert: Sentence embeddings using siamese bert-networks. Preprint, arXiv:1908.10084.

Nikhil Sharma, Kenton Murray, and Ziang Xiao. 2024. Faux polyglot: A study on information disparity in multilingual large language models. arXiv preprint.

S. Siriwardhana, R. Weerasekera, and E. Wen. 2023. Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering. Transactions of the Association for Computational Linguistics, 11:1–17.

S. Veturi, S. Vaichal, and R. L. Jagadheesh. 2024. Rag-based question-answering for contextual response prediction system. arXiv preprint.

Zheng Wang, Shu Xian Teo, Jieer Ouyang, Yongjun Xu, and Wei Shi. 2024. M-rag: Reinforcing large language model performance through retrieval-augmented generation with multiple partitions. arXiv preprint, arXiv:2405.16420.

Alina Wróblewska and Katarzyna Krasnowska-Kieraś. 2017. Polish evaluation dataset for compositional distributional semantics models. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 784–792.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 87–94. Association for Computational Linguistics.

Li Zhang, Steven Wilson, and Rada Mihalcea. 2019. Multi-label transfer learning for multi-relational semantic similarity. In Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019), page 44–50. Association for Computational Linguistics.

A Bias Matrix

	en	fr	es	sw	ru	zh	de
en	0.000	0.079	0.372	0.588	0.403	0.444	0.399
fr	0.079	0.000	0.353	0.574	0.381	0.421	0.384
es	0.372	0.353	0.000	0.436	0.106	0.166	0.092
sw	0.588	0.574	0.436	0.000	0.475	0.507	0.456
ru	0.403	0.381	0.106	0.475	0.000	0.154	0.130
zh	0.444	0.421	0.166	0.507	0.154	0.000	0.138
de	0.399	0.384	0.092	0.456	0.130	0.138	0.000

Table 4: Bias table in reranking for TriviaQA