Investigating Language Preference of Multilingual RAG Systems

Anonymous ACL submission

Abstract

Multilingual Retrieval-Augmented Generation (mRAG) systems enhance language models by integrating external multilingual information to produce context-aware responses. However, mRAG systems struggle with retrieving relevant information due to linguistic variations between queries and documents, generating inconsistent responses when multilingual sources conflict. In this work, we systematically investigate language preferences in both retrieval and generation of mRAG through a series of experiments. Our analysis indicates that retrievers tend to prefer high-resource and query languages, yet this preference does not consistently improve generation performance. Moreover, we observe that generators prefer the query language or Latin scripts, leading to inconsistent outputs. To overcome these issues, we propose Dual Knowledge Multilingual RAG (DKM-RAG), a simple yet effective framework that fuses translated multilingual passages with complementary model knowledge. Empirical results demonstrate that DKM-RAG mitigates language preference in generation and enhances performance across diverse linguistic settings.

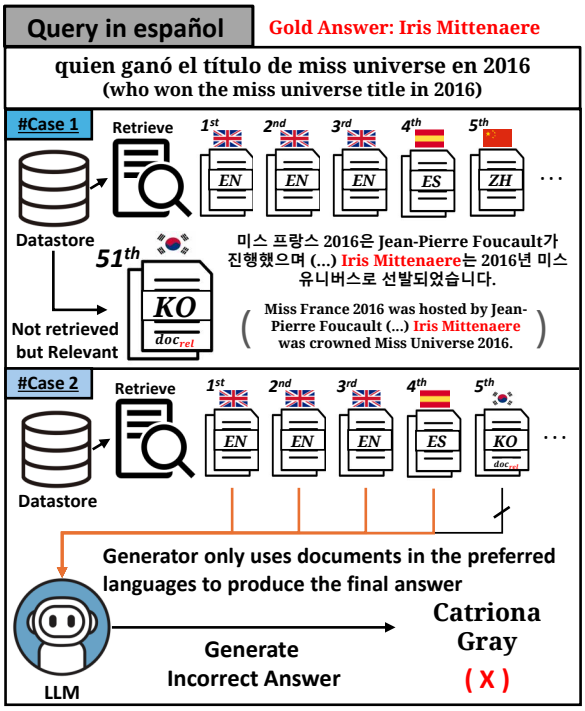

Figure 1: Failure cases of multilingual RAG system showing degraded generation ability because of language preference of retriever and generator in mRAG pipeline. doc_{rel} in the Korean (KO) document represents the relevant document to the given query that can be utilized to generate a final answer.

1 Introduction

Multilingual Retrieval-Augmented Generation (mRAG) extends traditional Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) by leveraging multilingual external sources to generate accurate, contextually and linguistically aware responses. However, mRAG systems face challenges in retrieving relevant information due to linguistic discrepancies between queries and documents (Wu et al., 2024a). Moreover, conflicts among multilingual sources can lead to inconsistencies in the generated responses (Chataigner et al., 2024).

Beyond retrieval challenges and source conflicts, language preference is another critical issue in mRAG systems, often leading to inaccurate outputs. As illustrated in **Case 1** of Figure 1, the retriever may prioritize particular languages—especially high-resource or query-language documents—at the expense of truly relevant information in low-resource language. Consequently, the Large Language Model (LLM) either produces an incorrect answer or deems the query unanswerable due to irrelevant content in the documents. Likewise, in **Case 2**, even if relevant documents are retrieved, the generator might favor passages in the query language or Latin scripts, ignoring essential evidence in lower-resource languages and resulting in inaccurate outputs. These preferences ultimately limit the effectiveness of mRAG, yielding biased rank-

ings, reduced answer quality, and inconsistencies across languages (Sharma et al., 2024).

Prior studies (Yang et al., 2024a; Telemala and Suleman, 2022; Sharma et al., 2024) have investigated this issue by introducing language fairness metrics to assess whether documents from different languages are ranked equitably via statistical equivalence testing, by proposing Language-Preference-Based Re-ranking for multilingual information retrieval, and investigating LLM's linguistic preference in across-language RAG-based information search setting. However, these approaches primarily focus on a limited set of languages and fail to reflect the true ranking dynamics of documents across languages.

In this work, we aim to understand language preference phenomena in mRAG systems comprehensively. We focus on the following three key research questions:

- **RQ1 (§4):** *Which languages does the retriever prefer?*
- **RQ2 (§5):** *Which languages does the generator prefer, and how do these preferences correlate with mRAG performance?*
- **RQ3 (§6):** *How can we mitigate language preference in mRAG?*

To address these questions, we present a comprehensive evaluation of language preferences throughout the entire mRAG pipeline across multiple languages. To systematically investigate the language preference problem of multilingual retrievers, we propose MultiLingualRank (MLR), a novel metric that quantifies language preference at the retriever level by measuring the shift in document rankings when non-query-language passages are translated into the query language. Our extensive experiments with diverse language combinations demonstrate that the retriever strongly prefers documents that are in high-resource languages and also share the same language as the query, confirming the presence of significant preference (§4).

At the generator level, we evaluate language preference by generating responses in multiple languages for the same query and the same retrieved document set, measuring their semantic similarity. Our results show that the generator favors both query languages and Latin script languages, with a relatively modest preference for query languages. This ultimately results in a decline in answer quality. Moreover, we uncover a nuanced relationship between language preference and overall mRAG performance. We observe that a strong preference for high-resource languages does not always lead to improved mRAG performance (§5). This occurs because the retriever may retrieve high-resource but irrelevant documents so that the generator cannot generate accurate answers from them. Therefore, language preference can degrade performance by overlooking lower-resource but relevant documents, thereby causing inconsistencies.

Finally, we propose Dual Knowledge Multilingual RAG (DKM-RAG), a simple yet effective framework that mitigates the language preference of mRAG. DKM-RAG enhances mRAG by combining externally retrieved, translated passages with internally rewritten passages enriched by the model's knowledge. Empirical results demonstrate that DKM-RAG significantly reduces language preference issues in the generation process, leading to improved performance across a range of linguistic settings (§6).

2 MultiLingualRank

To evaluate the language preference of a multilingual retriever in the mRAG system, we introduce *MultiLingualRank* (MLR), a novel metric that quantifies how much the ranking of retrieved documents improves when non-query language documents are translated into the query language. As shown in Figure 2, MLR is computed in three stages: (i) retrieving documents across multiple languages, (ii) translating documents that are not in the query language into query language, and (iii) re-ranking the translated documents to measure rank improvements.

2.1 Stage 1: Initial Document Retrieval

For each query $q \in Q$ (where Q is the set of all queries), we retrieve a ranked list of documents D_q from multilingual datastores. Each document $d \in D_q$ is assigned an initial rank r_d^{init} (with 1 being the highest rank). Let L_q denote the language of the query and L_d the language of document d.

2.2 Stage 2: Translation of Non-Query Language Documents

To ensure language consistency when assessing ranking improvements, we extract documents whose language differs from that of the query. Formally, we define:

$$D_q^{\text{diff}} = \{\, (d,\, r_d^{\text{init}}) \mid d \in D_q,\ L_d \neq L_q \,\}.$$

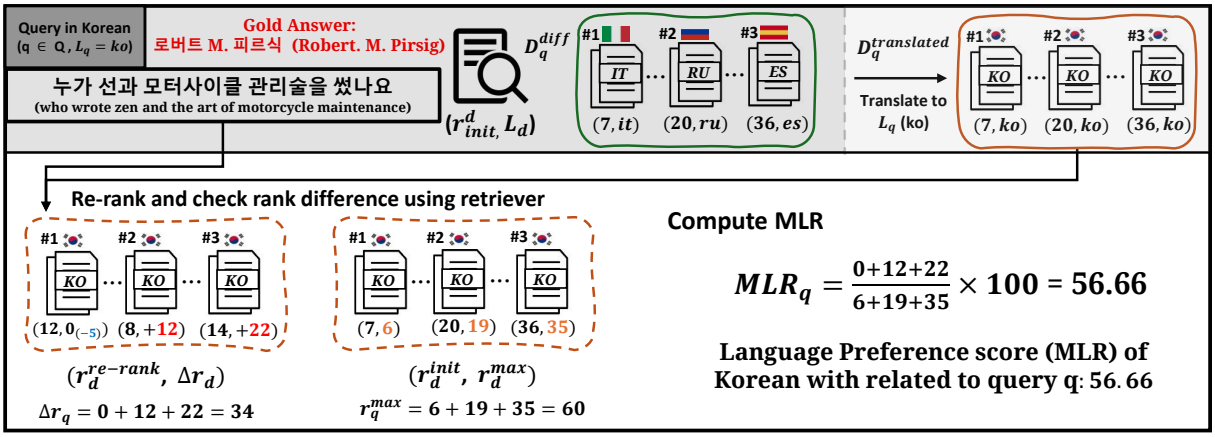

Figure 2: Overall framework of calculating MLR. For simplicity, we only consider three documents to calculate the MLR score.

Each document in D_q^{diff} is then translated into the query language L_q, resulting in the set:

$$D_q^{\text{Translated}} = \{\, d \mid d \text{ has been translated into } L_q \,\}.$$

2.3 Stage 3: Re-Ranking and MLR Score Computation

The translated documents in $D_q^{\text{Translated}}$ are re-ranked using retrievers in conjunction with the original query. Let $r_d^{\text{re-rank}}$ denote the new rank of document d after re-ranking. To capture ranking improvements, we compute the rank difference for each document d as:

$$\Delta r_d = \max\left(r_d^{\text{init}} - r_d^{\text{re-rank}}, 0\right).$$

A positive value of Δr_d indicates that the document has moved up in the ranking. For each query q, the total observed improvement is given by:

$$\Delta r_q = \sum_{d \in D_q^{\text{Translated}}} \Delta r_d.$$

To normalize this improvement, we first define the maximum possible improvement for each document as:

$$\Delta r_d^{\max} = r_d^{\text{init}} - 1,$$

and then compute the total maximum improvement for query q:

$$\Delta r_q^{\max} = \sum_{d \in D_q^{\text{Translated}}} \Delta r_d^{\max}.$$

The query-specific MLR score is then calculated as:

$$\text{MLR}_q = \begin{cases} \dfrac{\Delta r_q}{\Delta r_q^{\max}} \times 100, & \text{if } \Delta r_q^{\max} > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Finally, the overall MultiLingualRank is obtained by averaging the scores over all queries:

$$\text{MLR} = \frac{1}{|Q|} \sum_{q \in Q} \text{MLR}_q.$$

3 General Setup

3.1 Dataset

By following previous study (Chirkova et al., 2024), we use MKQA (Longpre et al., 2021) dataset, a multilingual open domain question answering evaluation set in our experiments. MKQA consists of 10k examples from the Natural Questions (NQ) dataset (Kwiatkowski et al., 2019), translated into 25 languages. This dataset is therefore parallel between languages and grounds knowledge primarily in English Wikipedia. In our experiments, we also select a subset of 2.7K samples, overlapping between MKQA and KILT NQ datasets [1], thus recovering relevant documents information from KILT NQ.

3.2 Models

Multilingual Retrievers Following previous work (Chirkova et al., 2024), we use a strong and publicly available BGE-m3 (Chen et al., 2024) as our multilingual retriever which can encode various languages we consider in our experiments. Consistent with the retriever, we use BGE-m3 (Chen et al., 2024) as a re-ranking encoder for computing MLR. In addition, we use two other Sentence-BERT series re-ranking encoders (Reimers and Gurevych, 2019), paraphrase-multilingual-MiniLM-L12-v2

[1] https://huggingface.co/datasets/facebook/kilt_tasks

and paraphrase-multilingual-mpnet-base-v2 to generalize the experimental results. We abbreviate them as p-mMiniLM and p-mMpNet for better visibility of the table.

Generators We use recently released various strong multilingual LLM, aya-expanse-8b (Dang et al., 2024) that can deal with various languages well. Also, we use strong LLMs, qwen 2.5-7B Instruct (Team, 2024) and Phi-4 14B (Abdin et al., 2024), Llama-3.1-8B-Instruct (Dubey et al., 2024) as our generators.

3.3 Other Implementation Details

Translation model We utilize a robust translation model for various languages, NLLB-200-distilled-600M (Costa-jussà et al., 2022) in our experiments.

Datastore Following previous study (Chirkova et al., 2024), we use Wikipedia as our datastore documents collection. In most of our experiments, we retrieve from two main Wikipedia sources: the KILT version of English Wikipedia[2] and the Wikipedia edition in the user's native language[3]. For detailed statistics, please refer to Appendix D.

Baseline We conduct several experiments to measure the language preference of mRAG based on Bergen (Chirkova et al., 2024). Bergen explores the components and adjustments necessary to develop an effective mRAG pipeline, serving as a robust baseline for future research.

4 Language Preference of Retrievers

In this section, we examine two factors that may affect the retriever's language preference: (i) the relationship between the query language (L_q) and the document language (L_d), and (ii) the resource availability of the languages involved.

4.1 Effect of the L_q and L_d Relationship

4.1.1 Experimental Setup

We evaluate eight language pairs under two scenarios: (1) **monolingual settings** where $L_q = L_d$, and (2) **cross-lingual settings** where $L_q \neq L_d$. In each case, our primary metric is *MLR* (MultiLingualRank), computed using three re-ranker encoders (bge-m3, p-mMiniLM, and p-mMpNet). For example, if the query is in English (en) but the target translation is in Korean (ko), we translate all non-English passages into Korean and then measure the rank changes with MLR.

4.1.2 Results for Monolingual Settings ($L_q = L_d$)

Strong Preference When the Query and Document Languages Match. As shown in the leftmost column of Table 1, when the query and document languages are identical, the retriever shows a high preference. This is expected, as direct linguistic alignment avoids the complexities of cross-lingual mapping and translation, thereby yielding stronger preference.

4.1.3 Results for Cross-Lingual Settings ($L_q \neq L_d$)

Lower Overall MLR in Cross-Lingual Matching. When $L_q \neq L_d$, the retriever performs cross-lingual matching, which typically results in lower MLR values than in monolingual cases. As indicated by the right-hand columns in Table 1 (highlighted in blue), cross-lingual setups are generally less preferred than their monolingual counterparts—except in cases involving English.

English as a Dominant Target Language. We observe that when the translated document language L_d is English, the retriever exhibits nearly the highest language preference as stated in the English column (en) in Table 1. In fact, English often outperforms even monolingual configurations, likely due to the abundance of English data in pretraining, which biases the model towards stronger English representations.

Influence of Language Family Similarity. Language family and geographic proximity also play a role. For example, Romance languages (fr, it, pt, es) share extensive lexical and structural similarities, which help maintain a relatively high cross-lingual preference and narrow the performance gap with monolingual setups, as illustrated by the joint L_q and L_d pairs in Table 1. Similarly, East Asian languages (ko, ja, zh) tend to show moderate declines in cross-lingual scenarios compared to the monolingual baseline, although they still lag behind the highest scores.

4.2 Impact of Language Resource Availability

4.2.1 Experimental Setup

We also investigate whether the volume of available language resources affects MLR. We categorize

[2] https://huggingface.co/datasets/facebook/kilt_wikipedia

[3] https://huggingface.co/datasets/wikimedia/wikipedia

Query Lang.	Encoder	$L_q = L_d$	$L_q \neq L_d$							
			en	ko	zh	fr	ja	it	pt	es
en	bge-m3	**56.03**	–	33.02 (-23.01)	33.10 (-22.93)	36.61 (-19.42)	33.36 (-22.67)	35.89 (-20.14)	35.86 (-20.17)	36.62 (-19.41)
	p-mMiniLM	**56.85**	–	34.34 (-22.51)	34.61 (-22.24)	38.17 (-18.68)	34.52 (-22.33)	37.15 (-19.70)	36.73 (-20.12)	37.96 (-18.89)
	p-mMpNet	**57.49**	–	34.45 (-23.04)	34.27 (-23.22)	37.94 (-19.55)	34.67 (-22.82)	37.34 (-20.15)	37.02 (-20.47)	37.90 (-19.59)
ko	bge-m3	41.15	**43.49** (+2.34)	–	34.42 (-6.73)	36.42 (-4.73)	37.18 (-3.97)	35.72 (-5.43)	35.30 (-5.85)	35.93 (-5.22)
	p-mMiniLM	42.95	**44.62** (+1.67)	–	36.04 (-6.91)	37.08 (-5.87)	38.47 (-4.48)	36.07 (-6.88)	36.18 (-6.77)	36.45 (-6.50)
	p-mMpNet	42.53	**44.98** (+2.45)	–	35.85 (-6.68)	37.20 (-5.33)	39.01 (-3.52)	36.21 (-6.32)	35.65 (-6.88)	36.34 (-6.19)
zh	bge-m3	44.98	**45.26** (+0.28)	34.52 (-10.46)	–	36.34 (-8.64)	36.05 (-8.93)	35.86 (-9.12)	35.73 (-9.25)	36.45 (-8.53)
	p-mMiniLM	**46.18**	45.39 (-0.79)	35.46 (-10.72)	–	36.98 (-9.20)	36.77 (-9.41)	36.38 (-9.80)	36.05 (-10.13)	36.85 (-9.33)
	p-mMpNet	**46.27**	45.41 (-0.86)	35.21 (-11.06)	–	36.87 (-9.40)	36.71 (-9.56)	36.28 (-9.99)	35.94 (-10.33)	36.78 (-9.49)
fr	bge-m3	43.18	**47.23** (+4.05)	33.29 (-9.89)	33.58 (-9.60)	–	34.07 (-9.11)	36.70 (-6.48)	36.30 (-6.88)	37.25 (-5.93)
	p-mMiniLM	44.09	**48.15** (+4.06)	34.54 (-9.55)	34.52 (-9.57)	–	34.83 (-9.26)	37.65 (-6.44)	37.05 (-7.04)	38.03 (-6.06)
	p-mMpNet	43.96	**48.14** (+4.18)	34.25 (-9.71)	34.37 (-9.59)	–	34.61 (-9.35)	37.59 (-6.37)	36.93 (-7.03)	38.01 (-5.95)
ja	bge-m3	45.03	**45.18** (+0.15)	35.45 (-9.58)	34.86 (-10.17)	36.71 (-8.32)	–	36.11 (-8.92)	35.88 (-9.15)	36.56 (-8.47)
	p-mMiniLM	**45.80**	45.54 (-0.26)	35.90 (-9.90)	35.57 (-10.23)	37.18 (-8.62)	–	36.53 (-9.27)	36.25 (-9.55)	36.91 (-8.89)
	p-mMpNet	**45.67**	45.39 (-0.28)	35.73 (-9.94)	35.30 (-10.37)	36.94 (-8.73)	–	36.24 (-9.43)	35.98 (-9.69)	36.62 (-9.05)
it	bge-m3	41.06	**46.63** (+5.57)	33.30 (-7.76)	33.47 (-7.59)	37.92 (-3.14)	33.86 (-7.20)	–	36.44 (-4.62)	37.68 (-3.38)
	p-mMiniLM	42.11	**47.69** (+5.58)	34.57 (-7.54)	34.59 (-7.52)	39.07 (-3.04)	34.80 (-7.31)	–	37.55 (-4.56)	38.83 (-3.28)
	p-mMpNet	41.98	**47.59** (+5.61)	34.48 (-7.50)	34.68 (-7.30)	38.94 (-3.04)	34.67 (-7.31)	–	37.27 (-4.71)	38.67 (-3.31)
pt	bge-m3	39.19	**46.64** (+7.45)	33.37 (-5.82)	33.46 (-5.73)	37.83 (-1.36)	34.02 (-5.17)	37.13 (-2.06)	–	38.61 (-0.58)
	p-mMiniLM	40.17	**47.75** (+7.58)	34.67 (-5.50)	34.91 (-5.26)	39.02 (-1.15)	35.03 (-5.14)	38.25 (-1.92)	–	39.68 (-0.49)
	p-mMpNet	39.91	**47.30** (+7.39)	34.68 (-5.23)	34.50 (-5.41)	38.70 (-1.21)	34.72 (-5.19)	38.01 (-1.90)	–	39.35 (-0.56)
es	bge-m3	40.76	**46.93** (+6.17)	33.36 (-7.40)	33.42 (-7.34)	37.73 (-3.03)	33.87 (-6.89)	37.22 (-3.54)	36.88 (-3.88)	–
	p-mMiniLM	41.81	**47.90** (+6.09)	34.63 (-7.18)	34.52 (-7.29)	38.86 (-2.95)	34.76 (-7.05)	38.33 (-3.48)	37.84 (-3.97)	–
	p-mMpNet	41.33	**47.34** (+6.01)	34.39 (-6.94)	34.19 (-7.14)	38.34 (-2.99)	34.39 (-6.94)	37.73 (-3.60)	37.25 (-4.08)	–

Table 1: Language preference measured by MLR with different re-ranking encoders for various query–document language pairs. The $L_q = L_d$ column shows scores for matching query and document languages, while the remaining columns represent cross-lingual scenarios. Parentheses indicate the change from the $L_q = L_d$ column (positive for improvement, negative for decline). The highest score per row is in bold, and the second highest is underlined.

languages into three groups based on their distribution in the pre-training corpus of recent LLMs: high-resource (e.g., English), mid-resource (e.g., Spanish), and low-resource (e.g., Korean). We use the same query set across all setups while systematically varying L_q and L_d.

4.2.2 Results

Limited Impact of Query Language Resources. The resource level of the query language (L_q) has a limited effect on cross-lingual preference. As shown in Table 1, when L_q is high-resource (e.g., English), strong preference is observed only if L_d also matches a high-resource language. Otherwise, the MLR scores remain similar regardless of whether L_q is high-, mid-, or low-resource.

Document Language Resources Are More Influential. In contrast, the language resource level of the document language (L_d) has a pronounced impact on MLR. As shown in Table 1, documents from high-resource languages consistently achieve the highest preference scores, followed by mid-resource and then low-resource languages. This trend (High > Mid > Low) suggests that extensive pre-training on high-resource languages enables stronger alignment, yielding higher MLR even across diverse query languages. Conversely, low-resource datastores typically produce lower MLR scores unless the query language also corresponds to that low-resource setting.

Overall, our results indicate that the resource availability of L_d critically influences the language preference of the retriever. These findings lay the groundwork for further investigation into the language dynamics within mRAG systems.

5 Language Preference of Generators

In this section, we explore LLM generators' language preferences in mRAG and their impact on overall performance.

5.1 Do LLMs Prefer Certain Languages for Contextual Knowledge?

5.1.1 Experimental Setup

To assess the generator's language preference, we measure answer consistency across eight languages: English (en), Korean (ko), Chinese (zh), French (fr), Japanese (ja), Italian (it), Portuguese (pt), and Spanish (es). For a given query, the generator produces responses in each of these languages using the same retrieved document set from multilingual datastores. We then compute the embedding

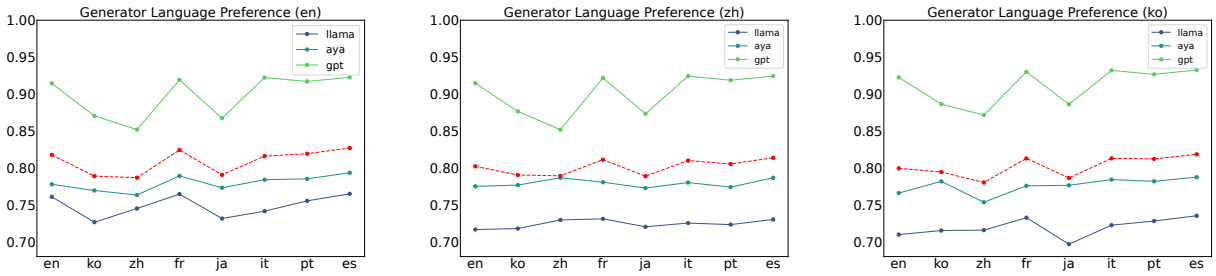

Figure 3: Language Preference of the Generators. In each figure, "aya" represents aya-expanse-8B, "llama" represents Llama-3.1-8B-Instruct, and "gpt" represents gpt-4o-mini. The red dotted line indicates the average generator preference.

similarity between each pair of generated answers, resulting in an 8×8 similarity matrix. We define the preference for a specific language as the average similarity score of the responses in that language.

We use LaBSE for measuring multilingual semantic similarity (Feng et al., 2022). And we use aya-expanse-8B, Llama-3.1-8B-Instruct, and GPT-4o-mini as our generators. We use language-specific prompts that incorporate the retrieved passages to induce responses in the target language, enabling us to capture the generator's inherent language preference.

5.1.2 Results

Strong Preference for Latin Script Languages. Figure 3 indicates that the generator produces more consistent responses in languages that use Latin scripts (e.g., en, fr, it, pt, es) compared to non-Latin languages (e.g., ko, zh, ja). This suggests that the model benefits from structural advantages in token alignment when processing Latin-based languages.

Modest Preference for the Query Languages. In addition, the generator shows a slight increase in consistency when the output language matches a query language. For instance, Korean (ko) queries yield somewhat more consistent responses when the generator replies in Korean rather than when the query is in English. However, this improvement is marginal, suggesting that the overall preference toward Latin scripts remains influential. Despite the modest gain, the model still demonstrates some capacity to handle multilingual queries effectively, indicating that matching the query language can provide a small but measurable benefit in non-Latin contexts.

5.2 Correlation between Language Preference and mRAG Performance

5.2.1 Experimental Setup

To examine how language preference impacts overall mRAG performance, we isolate language effects by providing generators with retrieved passages unified in a single target language—selected from the eight candidates (en, ko, zh, fr, ja, it, pt, es). We retrieve data from multilingual sources, enabling a direct comparison between language preference (measured by MLR, as shown in Table 1) and performance across three query languages (en, ko, zh).

We evaluate four generators (aya-expanse-8B, Phi-4, Qwen2.5-7B-Instruct, and Llama3.1-8B-Instruct) using character 3-gram recall (Chirkova et al., 2024) under three configurations. In the first configuration, passages are retrieved from multilingual resources, denoted as *all*. In the second, all retrieved passages are unified into a single target language (single-language document set). Finally, in the third configuration, we employ our proposed DKM-RAG framework (detailed in Section 6) to mitigate language preference. We also compute the average MLR score (across different query languages) for each language to indicate its overall preference.

5.2.2 Results and Analysis

Strong Correlation for English Queries. As stated in Table 2, for queries in English (L_q = en), RAG performance shows a strong correlation with language preference. English achieves the best results—likely due to its high-resource availability and the model's familiarity with it. In this setting, the *all* strategy is particularly effective, as it leverages cross-lingual knowledge fusion. We observe an exception for Japanese (ja), where performance is lower despite moderate preference, possibly due to challenges with non-Latin scripts and complex morphology.

	all	en	zh	ko	fr	ja	it	pt	es	DKM-RAG
L_q = **en**										
aya-expanse-8b	80.09	79.34	63.08	64.46	76.13	61.20	75.47	75.65	76.32	**82.60**
Phi-4	79.69	78.89	63.06	52.30	74.43	48.86	74.02	74.39	75.32	**82.59**
Qwen2.5-7B-Instruct	80.15	79.11	50.31	64.90	76.28	62.62	75.47	75.97	76.54	**82.60**
Llama3.1-8B-Instruct	80.25	79.28	61.99	65.81	76.40	62.58	75.89	76.09	76.47	**82.57**
L_q = **zh**										
aya-expanse-8b	32.55	25.62	38.31	26.64	24.00	25.27	23.63	23.63	23.79	**44.57**
Phi-4	16.75	17.57	36.76	17.50	18.15	17.56	18.19	17.89	18.44	**44.56**
Qwen2.5-7B-Instruct	34.28	27.33	38.31	27.91	25.15	27.78	25.90	25.37	25.30	**44.70**
Llama3.1-8B-Instruct	28.50	24.36	38.48	23.84	22.48	23.78	23.18	23.32	23.02	**44.51**
L_q = **ko**										
aya-expanse-8b	40.60	38.08	26.01	49.66	25.37	26.82	24.98	25.26	25.51	**55.01**
Phi-4	26.80	20.24	17.54	49.25	19.03	17.91	18.93	19.19	19.19	**54.82**
Qwen2.5-7B-Instruct	36.50	22.87	20.08	49.44	21.79	20.94	21.65	21.44	21.52	**54.85**
Llama3.1-8B-Instruct	37.18	26.48	22.88	49.87	24.46	24.86	25.23	24.87	25.22	**54.99**
MLR (Preference)	-	**47.70**	35.90	35.47	37.94	37.59	37.66	37.15	37.97	-

Table 2: Performance comparison between DKM-RAG and single/all language retrieval settings, showing character 3-gram recall scores for three query languages ($L_q \in \{\text{en}, \text{ko}, \text{zh}\}$) and eight passage languages. The bottom row shows average preference (MLR) scores. We highlight the cells corresponding to matching query and passage languages with a yellow background. The highest score per row is in bold, and the second highest is underlined.

Weaker Correlation for Non-English Queries. When $L_q \neq$ en, the relationship between language preference and performance becomes less pronounced. Although the generator generally prefers English passages overall, it achieves optimal performance when it receives retrieved passages that directly match the query language. In these cases, translating all passages into English does not enhance performance; instead, maintaining language consistency between the query and passages yields better results. This finding underscores the importance of linguistic compatibility in mRAG systems.

Optimal mRAG Strategy. Based on our experiments, different strategies depending on the query language prove more effective. As stated in Table 2, for English queries, employing the *all* strategy capitalizes on the high cross-lingual preference for English. In contrast, for non-English queries, translating retrieved passages into the query language L_q bridges the comprehension gap and ensures better alignment between query intent, passage semantics, and output language. This targeted approach ultimately leads to improved RAG performance by accommodating the specific language dynamics of the generator.

6 Dual Knowledge Multilingual RAG

Translating retrieved documents into the query language benefits mRAG, but it may also reflect retrieval outputs from high-resource languages including irrelevant content. Therefore, leveraging the LLM's internal knowledge can help filter inaccuracies and enrich the retrieved information with more reliable content. So we rewrite translated passages to refine the relevancy of documents by leveraging LLM's internal information.

Based on this insight, we propose Dual Knowledge Multilingual RAG (DKM-RAG), a framework that leverages both external translated passage and internal knowledge as shown in Figure 4. First **(#1)**, we retrieve documents for a given query from the *all* strategy and re-rank them. Next **(#2)**, we obtain external translated passages, $P_{\text{translated}}$ by translating into the query language. And **(#3)**, the rewriter LLM refines each translated passage in the context of the given query to produce refined passages, P_{refined}. This refining process utilizes a prompt to guide the model in integrating its internal knowledge, removing redundancy, and highlighting relevant information in a coherent and consistent style. For detailed prompts, please refer to Appendix B. Finally **(#4)**, We concatenate the two sets to form the final passage set as input to the generator LLM, ensuring that responses are both contextually enriched and linguistically aligned with the query:

$$P_{\text{final}} = \text{concat}\big(P_{\text{translated}}, P_{\text{refined}}\big).$$

Results. As shown in Table 2, DKM-RAG outperforms other document-based generator settings.

For non-English queries ($L_q \neq$ en), it leverages translated passages and enriched content to handle linguistic diversity. Even for English queries ($L_q =$ en), it surpasses the *all* baseline, highlighting the importance of integrating translated and refined knowledge.

7 Related Works

7.1 Multilingual RAG

Researchers explore challenges in mRAG such as problem of cross-lingual dense passage retrieval for low-resource languages (Wu et al., 2024a) and research various techniques to address key challenges in mRAG, such as enhancing the performance of language models in low-resource languages (Deshpande et al., 2024), resolving low-resource scenarios (Zhang et al., 2024), and adapting language models for multilingual reasoning tasks (Yoon et al., 2024). Benchmarks like MMTEB (Enevoldsen et al., 2025) enable systematic evaluation of multilingual retrieval.

Earlier mRAG systems frequently focus on high-resource languages (e.g., English), but a growing body of research aims to make advanced Natural Language Processing (NLP) technology accessible across a wide spectrum of linguistic contexts. Proposed solutions include code-mixed prompts for in-context learning (Shankar et al., 2024) and self-distillation from resource-rich to low-resource languages (Zhang et al., 2024).

7.2 Language Preference in mRAG

Despite significant progress, language preference—a systematic tendency to favor certain languages—remains a critical issue in mRAG systems. This preference arises from imbalances in training data, tokenization mismatches, script differences, and uneven resource availability (Sharma et al., 2024; Wu et al., 2024b). Studies show that high-resource languages (e.g., English) often overshadow relevant content in lower-resource languages during retrieval (Yang et al., 2024b; Chirkova et al., 2024), leading to suboptimal evidence retrieval (Yang et al., 2024a) and causing inconsistencies or hallucinations in outputs (Chataigner et al., 2024). These disparities also raise broader fairness concerns in multilingual NLP, as pre-trained models exhibit group fairness issues across languages (Cabello Piqueras and Søgaard, 2022; Ramesh et al., 2023).

Researchers propose several methods to counteract language preferences, including language-preference-based re-ranking (Telemala and Suleman, 2022), evaluate knowledge consistency across languages (Qi et al., 2023), and specialized datasets designed to detect such imbalances (Li et al., 2024). However, these approaches often focus on a single mRAG stage or overlook the actual ranking of retrieved documents (Sharma et al., 2024; Yang et al., 2024a). We introduce a metric that quantifies language preference in retrieval via ranking differences and propose a simple framework to mitigate these preferences across the entire mRAG pipeline.

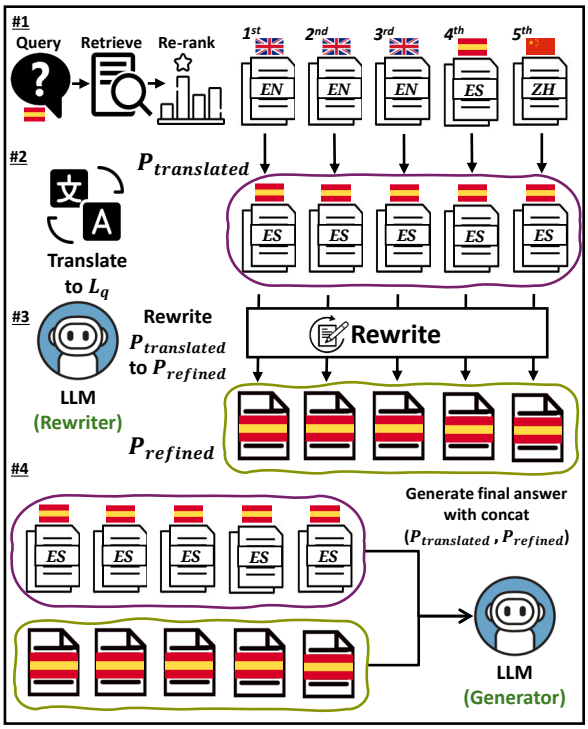

Figure 4: Overall flow of proposed DKM-RAG.

8 Conclusion

In this work, we investigate language preferences in mRAG systems. We propose a metric that measures the language preference of retrievers by checking the rank difference between the translated passage and the original one. Our experiments reveal that retrievers prefer high-resource and query language but do not always yield better generation performance. We also find that generators often favor the query language or Latin scripts, resulting in inconsistent outputs. To address this, we propose DKM-RAG which integrates translated passages with internal knowledge. Empirical results show that DKM-RAG consistently enhances mRAG performance across diverse languages.

Limitations

Our approach involves translating documents to measure rank shifts and unify linguistic representations. This process relies heavily on the quality of the translation model employed. Errors or inaccuracies in translation can distort the original meaning of passages and potentially introduce noise into both the retrieval and generation stages.

MLR entails translation and re-ranking steps. While this approach offers a principled way to quantify language preference, it also adds latency and computational cost, especially when dealing with large-scale multilingual corpora or real-time systems.

DKM-RAG framework which combines external translated passages and parametric (internal) knowledge, improves performance yet remains relatively straightforward. Future work could explore more sophisticated techniques for merging external and internal knowledge (e.g., trainable fusion mechanisms, dynamic weighting) to further reduce preferences and enhance overall system capabilities.

Lastly, our experiments focus on Wikipedia-based datasets in a specific set of languages, which may not generalize to all linguistic varieties or specialized domains. Future research should examine broader contexts, including low-resource languages not present in widely available corpora or domain-specific retrieval settings, to fully assess how language preferences manifest across diverse real-world scenarios.

Ethics Statement

We conduct our experiments using publicly available, multilingual dataset and models that follow recognized research and data-sharing guidelines. These resources are widely utilized in the academic community and are distributed with the intent to minimize harmful biases, inappropriate content, or stereotypes. However, they may still not fully represent the diversity of all languages and cultural contexts. We adhere strictly to the usage protocols and license agreements set forth by the original providers, who have taken steps to ensure compliance with established ethical standards.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. 2024. Phi-4 technical report.

Laura Cabello Piqueras and Anders Søgaard. 2022. Are pretrained multilingual models equally fair across languages? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3597–3605, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Cléa Chataigner, Afaf Taïk, and Golnoosh Farnadi. 2024. Multilingual hallucination gaps in large language models.

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

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

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier.

Tejas Deshpande, Nidhi Kowtal, and Raviraj Joshi. 2024. Chain-of-translation prompting (cotr): A novel prompting technique for low resource languages. *arXiv preprint arXiv:2409.04512*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Veysel Çağatan, Akash Kundu, Martin Bernstorff, Shitao Xiao, Akshita Sukhlecha, Bhavish Pahwa, Rafał Poświata, Kranthi Kiran GV, Shawon Ashraf, Daniel Auras, Björn Plüster, Jan Philipp Harries, Loïc Magne, Isabelle Mohr, Dawei Zhu, Hippolyte Gisserot-Boukhlef, Tom Aarsen, Jan Kostkan, Konrad Wojtasik, Taemin Lee, Marek Suppa, Crystina Zhang, Roberta Rocca, Mohammed Hamdy, Andrianos Michail, John Yang, Manuel Faysse, Aleksei Vatolin, Nandan Thakur, Manan Dey, Dipam Vasani, Pranjal A Chitale, Simone Tedeschi, Nguyen Tai, Artem Snegirev, Mariya Hendriksen, Michael Günther, Mengzhou Xia, Weijia Shi, Xing Han Lù, Jordan Clive, Gayatri K, Maksimova Anna, Silvan Wehrli, Maria Tikhonova, Henil Shalin Panchal, Aleksandr Abramov, Malte Ostendorff, Zheng Liu, Simon Clematide, Lester James Validad Miranda, Alena Fenogenova, Guangyu Song, Ruqiya Bin Safi, Wen-Ding Li, Alessia Borghini, Federico Cassano, Lasse Hansen, Sara Hooker, Chenghao Xiao, Vaibhav Adlakha, Orion Weller, Siva Reddy, and Niklas Muennighoff. 2025. MMTEB: Massive multilingual text embedding benchmark. In *The Thirteenth International Conference on Learning Representations*.

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.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Bryan Li, Samar Haider, Fiona Luo, Adwait Agashe, and Chris Callison-Burch. 2024. BordIRlines: A dataset for evaluating cross-lingual retrieval augmented generation. In *Proceedings of the First Workshop on Advancing Natural Language Processing for Wikipedia*, pages 1–13, Miami, Florida, USA. Association for Computational Linguistics.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10650–10666, Singapore. Association for Computational Linguistics.

Krithika Ramesh, Sunayana Sitaram, and Monojit Choudhury. 2023. Fairness in language models beyond English: Gaps and challenges. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2106–2119, Dubrovnik, Croatia. Association for Computational Linguistics.

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*. Association for Computational Linguistics.

Bhavani Shankar, Preethi Jyothi, and Pushpak Bhattacharyya. 2024. In-context mixing (ICM): Code-mixed prompts for multilingual LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4162–4176, Bangkok, Thailand. Association for Computational Linguistics.

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

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Joseph P. Telemala and Hussein Suleman. 2022. Language-preference-based re-ranking for multilingual swahili information retrieval. In *Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '22, page 144–152, New York, NY, USA. Association for Computing Machinery.

Jie Wu, Zhaochun Ren, and Suzan Verberne. 2024a. What are the limits of cross-lingual dense passage retrieval for low-resource languages?

Suhang Wu, Jialong Tang, Baosong Yang, Ante Wang, Kaidi Jia, Jiawei Yu, Junfeng Yao, and Jinsong Su. 2024b. Not all languages are equal: Insights into multilingual retrieval-augmented generation.

Eugene Yang, Thomas Jänich, James Mayfield, and Dawn Lawrie. 2024a. Language fairness in multilingual information retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2487–2491, New York, NY, USA. Association for Computing Machinery.

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

Dongkeun Yoon, Joel Jang, Sungdong Kim, Seungone Kim, Sheikh Shafayat, and Minjoon Seo. 2024. LangBridge: Multilingual reasoning without multilingual supervision. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7502–7522, Bangkok, Thailand. Association for Computational Linguistics.

Yuanchi Zhang, Yile Wang, Zijun Liu, Shuo Wang, Xiaolong Wang, Peng Li, Maosong Sun, and Yang Liu. 2024. Enhancing multilingual capabilities of large language models through self-distillation from resource-rich languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11189–11204, Bangkok, Thailand. Association for Computational Linguistics.

A Implementation Details

When retrieving from the datastore in all languages, we utilize the approach outlined in (Chirkova et al., 2024) as our baseline. Specifically, we employ the basic_translated_langspec prompt template, as detailed in Table 4 to generate our final mRAG answer from the generator. In our method, we retrieve and re-rank the top-50 documents for each query, and then use only the top-5 documents to generate the final answer. The document retrieval and re-ranking are carried out using bge-m3. We do not translate documents already in query language in the framework of DKM-RAG to reduce costs.

We conduct our experiments using an AMD EPYC 7313 CPU (3.0 GHz) paired with four NVIDIA RTX 4090 GPUs. We use Python 3.11.5 and PyTorch 2.3.1 for the software environment.

Language	Passage Count (M)	Percentage (%)
ja	27	20.53
en	25	19.00
de	14	10.64
fr	13	9.88
zh	11	8.36
es	10	7.60
ru	8.6	6.54
it	8.2	6.23
pt	4.7	3.57
th	3.7	2.81
ar	3.3	2.51
ko	1.6	1.22
fi	1.5	1.14

Table 3: Language distribution of wikipedia we use in our experiment.

B Prompts

As shown in Table 4, we provide the prompts used to generate our final answer with the retrieved documents in our mRAG baseline. *Docs* refers to retrieved documents and question refers to the current query. We also provide prompts during the passage rewriting phase in the DKM-RAG framework as stated in Table 5. We only provide english prompts for simplicity. And we provide prompts to measure the language preference of GPT-4o-mini, regarding answering in the specific languages as stated in Table 6.

C Language Notation

In this work, we use standard ISO 639-1 language codes to represent the various languages involved in our experiments. Specifically, en denotes English, ko represents Korean, ar corresponds to Arabic, zh refers to Chinese (Simplified), fi indicates Finnish, fr stands for French, de represents German, ja corresponds to Japanese, it refers to Italian, pt denotes Portuguese, ru stands for Russian, es represents Spanish, and th corresponds to Thai. These concise notations facilitate the identification and processing of language-specific data across datasets and models in multilingual NLP research.

D Dataset Statistics

We present the statistics of the datasets used in our experiments. MKQA serves as the primary dataset, and its details, including the number of examples and the median lengths of questions and answers, are summarized in Table 8. Additionally, we utilize Wikipedia as the external source for the retriever datastore, with its statistics (number of passages and median lengths) also provided in Table 8. And we provide the number of passages in each language and the ratio of them in Table 3. These details offer a clear overview of the data resources supporting our experiments.

Language Distribution of Pre-trained LLM We provide language distribution in the pre-training corpus of Llama-2. As stated in Table 7, we use English (EN) as a high-resource, Spanish (ES) as a mid-resource, and Korean (KO) as a low-resource language in our experiment based on their ratios.

E Language Preference of Other Languages

We also perform additional experiments to explore language preferences for languages not covered in Table 1, using the MLR score that we propose. As shown in Table 9, similar to the results in Table 1, the highest preferences are typically observed when $L_q = L_d$ across all query languages. English is also the most preferred language. For clarity, we omit results for other languages.

For most languages, such as Arabic, Finnish, German, and Russian, switching to a cross-lingual setup leads to a significant drop in MLR. For example, Arabic queries using the bge-m3 encoder achieve a monolingual score of 40.39, but cross-lingual retrieval (e.g., with Thai) results in a 6.80-point decrease.

Interestingly, for Thai queries, some cross-lingual pairs show a slight improvement over the monolingual baseline (as indicated by the positive

System	Prompt
With Documents	`You are a helpful assistant. Your task is to extract relevant information from provided documents and to answer to questions as short as possible. Please reply in English.` `user: f"Background:{docs}\n\nQuestion:{question}"`
Without Documents	`You are a helpful assistant. Answer the questions as short as possible. Please reply in English.` `user: f"Question:{question}"`

Table 4: System prompts with and without documents. The table outlines how instructions and prompts differ when documents are provided or omitted.

Prompt
Original Passage: {passage} Question: {question} Please create an independent document according to the following requirements: 1) Utilize known facts (parametric knowledge) related to the question. 2) Seamlessly combine with the original passage by removing redundant or unnecessary sentences. No additional explanations are allowed. 3) All content must be written smoothly and concisely in English.

Table 5: The prompt used for generating P_{refined} based on the passage and question. The instructions guide the generator to combine parametric knowledge with the original passage while ensuring clarity and conciseness.

differences in red), suggesting that for low-resource languages like Thai, cross-lingual signals might sometimes offer complementary benefits

F Similarity Matrices

We provide similarity matrix measured by LaBSE for each query language en, zh, ko and each generator in Figure 5, Figure 6, Figure 7, Figure 8, Figure 9, Figure 10, Figure 11, Figure 12 and Figure 13. Each entry represents the embedding similarity score between answers generated in different languages, with the diagonal values all equal to 1 (i.e., comparing an answer with itself). Moreover, the values shown in Figure 3 are computed by averaging over the rows or columns for each language.

G Case study

MLR We provide an example of a document that improved MLR score, where the rank of a relevant document significantly increases after translation. In Table 10, the user query *"영국 캐리비안에 언제 노예제가 폐지됐나요? (When was slavery abolished in the British Caribbean?)"* is in Korean, whereas the original passage is in English. Initially, the document's rank ($\mathbf{r}_{\mathbf{d}}^{\text{init}} = 34$) was relatively low, but after translating the passage into Korean and re-ranking ($\mathbf{r}_{\mathbf{d}}^{\text{re-rank}} = 2$), the document moved much closer to the top. This demonstrates how cross-lingual alignment can substantially improve retrieval performance in a multilingual setting. Notably, even if the passage content is semantically the same, language preference in the model can lead to poor alignment when the query and document are in different languages, adversely affecting retrieval. Translating the document into the query language effectively mitigates this issue.

Answer Generation in Language Preference of Generator We also provide an example of generated answers in different languages with a generator, GPT-4o-mini as shown in Table 11. The

	Content

System Message	`You are a highly capable multilingual assistant.`
	`Here are some reference documents: {top5_passages}`
	
	`The user wants answers in multiple languages.`
	`Please follow these rules strictly:`
	
	`1) Return your final answer as a valid JSON object.`
	
	`2) The JSON object must contain exactly these keys: {TARGET_LANGUAGES}.`
	
	`3) Each field's value must be the answer written in that respective language.`
	
	`4) Do not include any additional text outside the JSON (e.g., no Markdown or explanations).`
	
	`5) Ensure it is valid JSON with correct format.`

User Message	`Question: {question}`
	
	`Please provide the answers in JSON form for each of the following languages: {TARGET_LANGUAGES}.`

Table 6: Prompts used for measuring language preference of GPT-4o-mini in mRAG pipeline.

preference score in the rightmost column of Table 11 indicates that the generator prefers the query language and Latin-script languages over other languages.

Unified Document of DKM-RAG Additionally, we provide a sample of $P_{\text{translated}}$ and P_{refined} obtained via our proposed DKM-RAG framework in Table 12. This example illustrates how the crucial answer component, *"the executive branch"*, which is not apparent from the translated passage alone, emerges through the model's internal knowledge. Consequently, this shows that DKM-RAG can effectively leverage additional knowledge sources that is not included in the translated passage to achieve better performance.

G.1 Failure Case

MLR We present a failure case of the MLR metric in Table 13. Due to the difficulty of translating documents in low-resource languages, repetitive phrases such as *Changing the line-up* appear in the translated passage. This repetition causes the re-ranker to misinterpret the content, leading to an improvement in the rank even though the content is irrelevant.

DKM-RAG We also present a failure case of DKM-RAG in Table 14. The retriever retrieves an English document that is irrelevant to the query due to its language preference. Additionally, the LLM lacks relevant knowledge related to the query, resulting in a failed generation.

H Language Preference of Generators in average

We provide language preference of generators in terms of average as shown in Figure 14. Consistent with the result of each query language in Figure 3, the generator shows preferences for Latin-script languages. And GPT-4o-mini shows more consistent outputs than other generators. This is because it is a larger model than the others, providing more stable answers regardless of language preference. Between Llama and Aya, Aya produces slightly more consistent outputs, demonstrating its multilingual capability in handling diverse linguistic contexts.

I Ablation study of DKM-RAG

To prove the effectiveness of concatenating translated passages and refined passages in DKM-RAG framework, we provide an ablation study of each component in DKM-RAG. As stated in Table 15, removing any component from DKM-RAG results in decreased performance, highlighting that every part is crucial to its overall effectiveness.

Language	Percentage
EN	**89.70**%
Unknown	8.38%
DE	0.17%
FR	0.15%
SV	0.15%
ES	**0.13**%
ZH	0.15%
RU	0.12%
NL	0.11%
IT	0.11%
JP	0.11%
PL	0.09%
PT	0.09%
VI	0.08%
RO	0.03%
SR	0.04%
CA	0.04%
KO	**0.06**%
UK	0.07%
Other	0.21%

Table 7: Language distribution in the pre-training corpus of Llama-2. Unknown represents languages we cannot know because of closed-source access of model and other denotes other languages.

Dataset	**en**	**ar**	**es**	**fi**	**fr**	**de**	**ja**	**it**	**ko**	**pt**	**ru**	**zh**	**th**
MKQA													
# examples	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827
len question.	43	38	48	46	49	47	26	48	22	45	42	16	41
len answer.	11	10	11	11	11	11	8	11	6	11	12	6	12
Wikipedia													
# ex. (M)	25	3.3	10	1.5	13	14	27	8.2	1.6	4.7	8.6	11	3.7
len passage.	624	585	619	833	627	720	208	650	431	619	721	206	217

Table 8: Statistics of the datasets used in our experiments. MKQA Number of examples and median lengths of questions and answers (in Unicode characters). Wikipedia: Number of passages (in millions) and their median lengths.

J MLR Analysis

We prove the effectiveness of our proposed language preference metric, MLR by comparing language preference between MLR score and the average document language ratio of retrieved documents for each dataset. As stated in Table 16, the tendency of average language ratio of retrieved documents and MLR score is similar. To prove it, we also report Pearson and Spearman correlation coefficients and each p-value between them. Pearson value (0.98558) indicates a very strong positive linear correlation between the average MKQA language distribution values (mkqa_avg) and the MLR (Preference) scores. The p-value (7.75e-10) is extremely small, showing that the probability of observing such a strong correlation by chance is almost negligible. In short, there is a statistically significant, nearly perfect linear relationship between these two sets of values. Similarly, the Spearman value (0.86264) also indicates a strong association, and the corresponding p-value (1.47e-4) confirms that this correlation is statistically significant. By these results, we prove that MLR is efficient for measuring language preference of retriever.

Query Lang.	Encoder	$L_q = L_d$	$L_q \neq L_d$				
			ar	fi	de	ru	th
ar	**bge-m3**	**40.39**	–	34.10 (-6.29)	35.91 (-4.48)	36.22 (-4.17)	33.59 (-6.80)
	p-mMiniLM	**41.25**	–	34.90 (-6.35)	36.58 (-4.67)	37.13 (-4.12)	34.46 (-6.79)
	p-mMpNet	**41.34**	–	34.64 (-6.70)	36.34 (-5.00)	36.87 (-4.47)	34.36 (-6.98)
fi	**bge-m3**	**36.65**	33.47 (-3.18)	–	36.33 (-0.32)	35.42 (-1.23)	33.07 (-3.58)
	p-mMiniLM	**37.37**	34.60 (-2.77)	–	37.14 (-0.23)	36.48 (-0.89)	34.12 (-3.25)
	p-mMpNet	**37.27**	34.41 (-2.86)	–	36.92 (-0.35)	36.28 (-0.99)	34.12 (-3.15)
de	**bge-m3**	**39.81**	33.21 (-6.60)	34.16 (-5.65)	–	34.63 (-5.18)	32.95 (-6.86)
	p-mMiniLM	**40.80**	34.62 (-6.18)	35.25 (-5.55)	–	35.94 (-4.86)	34.18 (-6.62)
	p-mMpNet	**40.92**	34.81 (-6.11)	35.33 (-5.59)	–	36.13 (-4.79)	34.37 (-6.55)
ru	**bge-m3**	**45.05**	33.84 (-11.21)	34.20 (-10.85)	35.63 (-9.42)	–	33.24 (-11.81)
	p-mMiniLM	**46.08**	34.85 (-11.23)	35.18 (-10.90)	36.73 (-9.35)	–	34.23 (-11.85)
	p-mMpNet	**45.82**	34.63 (-11.19)	34.83 (-10.99)	36.28 (-9.54)	–	34.12 (-11.70)
th	**bge-m3**	34.52	33.68 (-0.84)	34.11 (-0.41)	**35.99** (+1.47)	35.60 (+1.08)	–
	p-mMiniLM	35.38	34.65 (-0.73)	34.77 (-0.61)	**36.63** (+1.25)	36.40 (+1.02)	–
	p-mMpNet	34.73	34.10 (-0.63)	34.14 (-0.59)	**36.08** (+1.35)	35.84 (+1.11)	–

Table 9: Language preference measured by MLR with various re-ranking encoders for various query and document language combinations in a multilingual retriever. The $L_q = L_d$ column reports the diagonal scores where the query language matches the translated document language, while the remaining columns represent cross-lingual scenarios (i.e., where the query language differs from the document language). Scores in parentheses indicate the difference from the diagonal value (positive for an improvement, negative for a decline). The highest score for each row is highlighted in bold, and the second highest is underlined.

Field	Value
query	영국 캐리비안에 언제 노예제가 폐지됐나요? (When was slavery abolished in the British Caribbean?)
gold answer	1834-08-01
doc id	kilt-100w_6947054 (English)
$\mathbf{r_d^{init}}$	34
$\mathbf{r_d^{re\text{-}rank}}$	2
d (content)	*History of the Caribbean. Empire remained slaves, however, until Britain passed the Slavery Abolition Act in 1833. When the Slavery Abolition Act came into force in 1834, roughly 700,000 slaves in the British West Indies immediately became free; other enslaved workers were freed several years later after a period of forced apprenticeship. Slavery was abolished in the Dutch Empire in 1814. Spain abolished slavery in its empire in 1811, with the exceptions of Cuba, Puerto Rico, and Santo Domingo; Spain ended the slave trade to these colonies in 1817, after being paid £400,000 by Britain. Slavery itself was not abolished in Cuba until 1886.*
d (translated)	*1834년 노예제 폐지법이 시행되자, 영국 서인도 제도에서 약 700,000명의 노예가 즉시 해방되었고, 다른 노예 노동자들은 강제 연습생 생활을 한 후 몇 년 후에 해방되었다. 1814년 네덜란드 제국에서 노예제는 폐지되었다. 1811년 스페인은 쿠바, 푸에르토리코, 산토 도밍고를 제외하고는 제국에서 노예제를 폐지했다. 1817년 영국이 400만원을 지불한 후 스페인은 이들 식민지에서의 노예 무역을 종식시켰다. 노예제는 1886년까지 쿠바에서 폐지되지 않았다.*

Table 10: An example of an improved MLR case. After translating the document into Korean, its rank improved from 34 to 2, illustrating language preference of retriever.

Question	which type of air pressure is associated with warm air rising

Language	**Answer**	**Preference Score**
en	Low pressure is associated with warm air rising.	0.9179
ko	따뜻한 공기가 상승하는 것과 관련된 공기 압력은 저압입니다.	0.9060
zh	暖空上升相的空力是低 。	0.8853
fr	La pression basse est associée à l'ascension de l'air chaud.	0.9231
ja	暖かい空が上昇することに連するは低です 。	0.9187
it	La bassa pressione è associata all'aria calda che sale.	0.9256
pt	A pressão baixa está associada ao ar quente que sobe.	0.9316
es	La presión baja está asociada con el aire caliente que asciende.	0.9317

Table 11: An example of generated answers in different languages with gpt-4o-mini. Also, we report the average similarity score between each pair of answers.

Field	**Value**
query	*in the united states the president is the head of which branch of government?*
gold answer	the executive branch
doc id	kilt-100w_5089743 (English)
$P_{\text{translated}}$	*President of the United States. President of the United States (POTUS) is the head of state and head of government of the United States of America. The president is the commander-in-chief of the United States Armed Forces. In contemporary times, the president is looked upon as one of the world's most powerful political figures as the leader of the only remaining global superpower. The role includes responsibility for the world's most expensive military, which has the second largest nuclear arsenal. The president also leads the nation with the largest economy.*
P_{refined}	*The president of the United States is the head of the executive branch of the federal government. The president directs the executive branch and is the commander-in-chief of the United States Armed Forces/ In contemporary times, the president is looked upon as one of the world's most powerful political figures as the leader of the only remaining global superpower. The role includes responsibility for the world's most expensive military, which has the second-largest nuclear arsenal. The president also leads the nation with the largest economy.*

Table 12: A DKM-RAG case study illustrating how $P_{\text{translated}}$ and P_{refined} correspond to the retrieved passage (translated into the query language) and the rewritten passage leveraging parametric knowledge, respectively. The overlap with the gold answer is highlighted in red.

Field	Value
query	*연속으로 가장 많은 자유투 기록 (who holds the record for most free throws made in a row)*
gold answer	톰 앰베리
doc id	wiki-100w-ja_8993041
$\mathbf{r}_{\mathbf{d}}^{\text{init}}$	31
$\mathbf{r}_{\mathbf{d}}^{\text{re-rank}}$	5
d (content)	*'林直明. を 更\n\n記 \n イニング最多連 四球：5（日本記） 1946年4月29日\n 同一年に2球 で勝利：1948年 ※史上3人目\n ゲム最多失点：14（セリグ記） 1950年6月7日\n\n背番'*
d (translated)	*Changing the line-up, Changing the line-up, Changing the line-up, Changing the line-up, Changing the line-up, Changing the line-up, Changing the line-up, Changing the line-up, Changing the line-up, Changing the line-up, Changing the line-up, Changing the line-up, Changing the line-up, Changing the line-up, Changing the line-up, Changing the line-up, (...)*

Table 13: A failure case of MLR because bad translation quality due to difficulty in translating low-resource language.

Field	Value
query	*who is the first british governor in india?*
gold answer	William Bentinck, Warren Hastings
doc id	kilt-100w_10189364 (English)
$P_{\text{translated}}$	*Sir John Child, 1st Baronet. Sir John Child, 1st Baronet (died 1690) was a governor of Bombay, and de facto (although not officially) the first governor-general of the British settlements in India. Born in London, Child was sent as a young boy to his uncle, the chief of the factory at Rajapur. On 27 October 1681, he was appointed chief of the East India Company's affairs at Surat and Bombay, while at the same time his namesake—stated to be unrelated by the Oxford Dictionary of National Biography—Sir Josiah Child, was governor of the company at home. The two*
P_{refined}	*Sir John Child, 1st Baronet was the first governor of Bombay, and de facto (although not officially) the first governor-general of the British settlements in India. He was born in London but later moved to India where he became involved in the East India Company's affairs. On 27th October 1681, he took up the position of Chief of the East India Company's affairs at both Surat and Bombay. This made him the first British Governor in . . .*

Table 14: A failure case of DKM-RAG because of preference of retriever so that high-resource but irrelevant document is retrieved.

	DKM-RAG	w/o P_{refined}	w/o $P_{\text{translated}}$
$L_q =$ **en**			
aya-expanse-8b	82.60	79.34	81.10
Phi-4	82.59	78.89	81.08
Qwen2.5-7B-Instruct	82.60	79.11	81.06
Llama3.1-8B-Instruct	82.57	79.28	81.19
$L_q =$ **zh**			
aya-expanse-8b	44.57	38.31	39.44
Phi-4	44.56	36.76	38.95
Qwen2.5-7B-Instruct	44.70	38.31	39.78
Llama3.1-8B-Instruct	44.51	38.48	39.35
$L_q =$ **ko**			
aya-expanse-8b	55.01	49.66	46.15
Phi-4	54.82	49.25	45.24
Qwen2.5-7B-Instruct	54.85	49.44	45.32
Llama3.1-8B-Instruct	54.99	49.87	45.55

Table 15: Ablation study on DKM-RAG. "DKM-RAG" denotes the DKM-RAG setting (i.e., the DKM-RAG column in Table 2), "w/o P_{refined}" indicates the performance corresponding to the highlighted cells, and "w/o $P_{\text{translated}}$" represents the results using only refined passages.

	en	ko	ar	zh	fi	fr	de	ja	it	pt	ru	es	th
mkqa_en	44.12	1.60	1.19	1.30	2.54	10.03	6.90	1.44	8.32	7.67	4.85	9.90	0.13
mkqa_ko	23.07	17.35	1.99	4.81	2.04	7.90	5.96	10.36	6.16	5.06	6.85	6.85	1.58
mkqa_ar	24.93	3.30	15.29	4.07	2.10	8.30	6.53	6.64	6.80	5.71	7.78	7.65	0.89
mkqa_zh	24.70	3.17	1.76	23.22	2.01	7.47	6.17	6.27	6.08	5.24	6.37	7.27	0.27
mkqa_fi	30.32	2.27	1.63	2.33	7.92	11.11	8.20	3.78	8.77	7.18	6.51	9.42	0.58
mkqa_fr	29.90	1.48	1.25	1.55	2.50	21.44	6.96	2.06	9.40	7.96	4.77	10.55	0.19
mkqa_de	32.54	1.46	1.17	1.44	2.96	11.40	15.12	1.89	9.09	7.69	4.83	10.17	0.24
mkqa_ja	24.56	4.80	1.69	3.99	2.19	7.97	5.99	22.55	6.38	5.66	6.49	7.45	0.28
mkqa_it	28.72	1.59	1.30	1.58	2.52	12.30	6.97	1.95	17.46	8.47	5.26	11.70	0.17
mkqa_pt	28.82	1.71	1.40	1.63	2.60	11.92	6.74	2.23	10.24	13.78	5.38	13.33	0.24
mkqa_ru	27.02	2.53	1.92	1.98	2.45	8.83	6.44	2.71	7.36	6.24	23.83	8.43	0.26
mkqa_es	29.45	1.73	1.27	1.60	2.66	11.85	6.93	1.83	10.55	9.33	5.27	17.36	0.16
mkqa_th	32.39	3.10	2.10	2.96	2.53	10.00	7.40	4.43	8.06	7.43	6.80	9.70	3.10
mkqa_avg	**29.27**	**3.55**	**2.61**	**4.04**	**2.85**	**10.81**	**7.41**	**5.24**	**8.82**	**7.49**	**7.31**	**9.98**	**0.62**
MLR (Preference)	**47.70**	**35.47**	**35.59**	**35.90**	**35.13**	**37.94**	**37.20**	**37.59**	**37.66**	**37.15**	**37.99**	**37.97**	**34.09**

Pearson correlation coefficient: 0.98558 (p-value: 7.75e-10)
Spearman correlation coefficient: 0.86264 (p-value: 1.47e-4)

Table 16: Language distribution ratios of documents retrieved from datasets composed of each query language. The table lists the raw MKQA language distribution values (without the percent sign) for each dataset. The row **mkqa_avg** shows the average distribution across all MKQA datasets for each language, while the row **MLR (Preference)** provides the corresponding MLR scores. Additionally, we report Pearson and Spearman correlation coefficients between MLR and mkqa_avg.

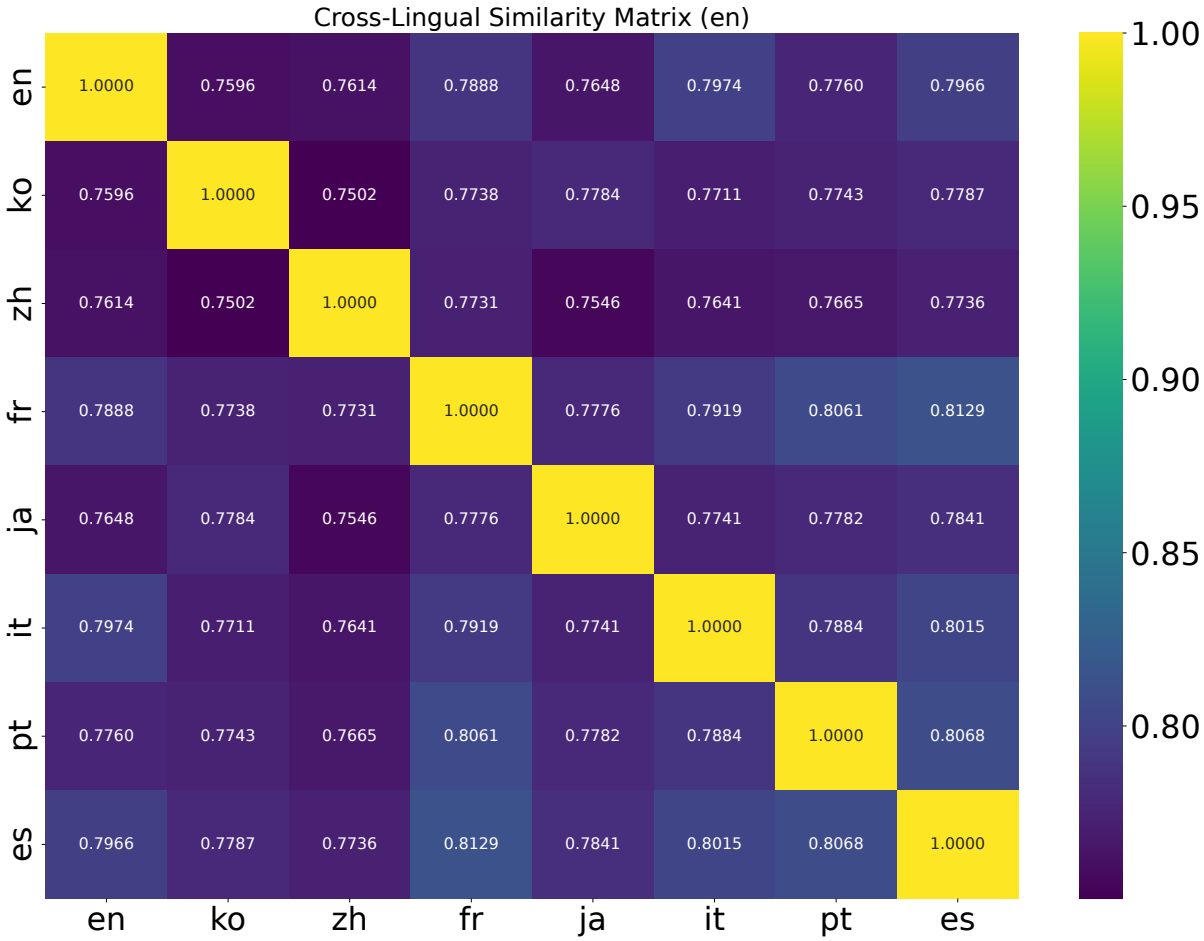

Figure 5: LaBSE Similarity Matrix of aya-expanse-8b (en).

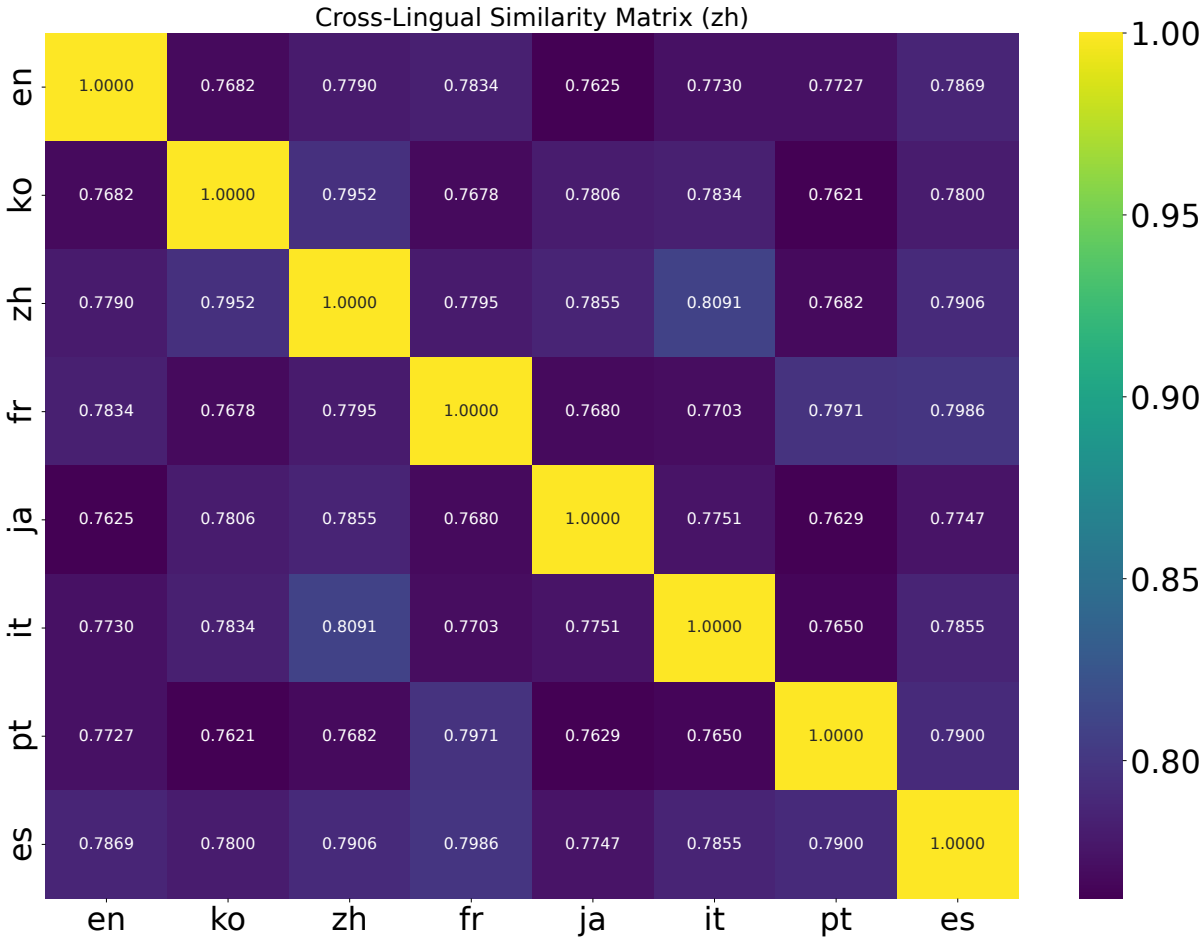

Figure 6: LaBSE Similarity Matrix (zh) of aya-expanse-8b.

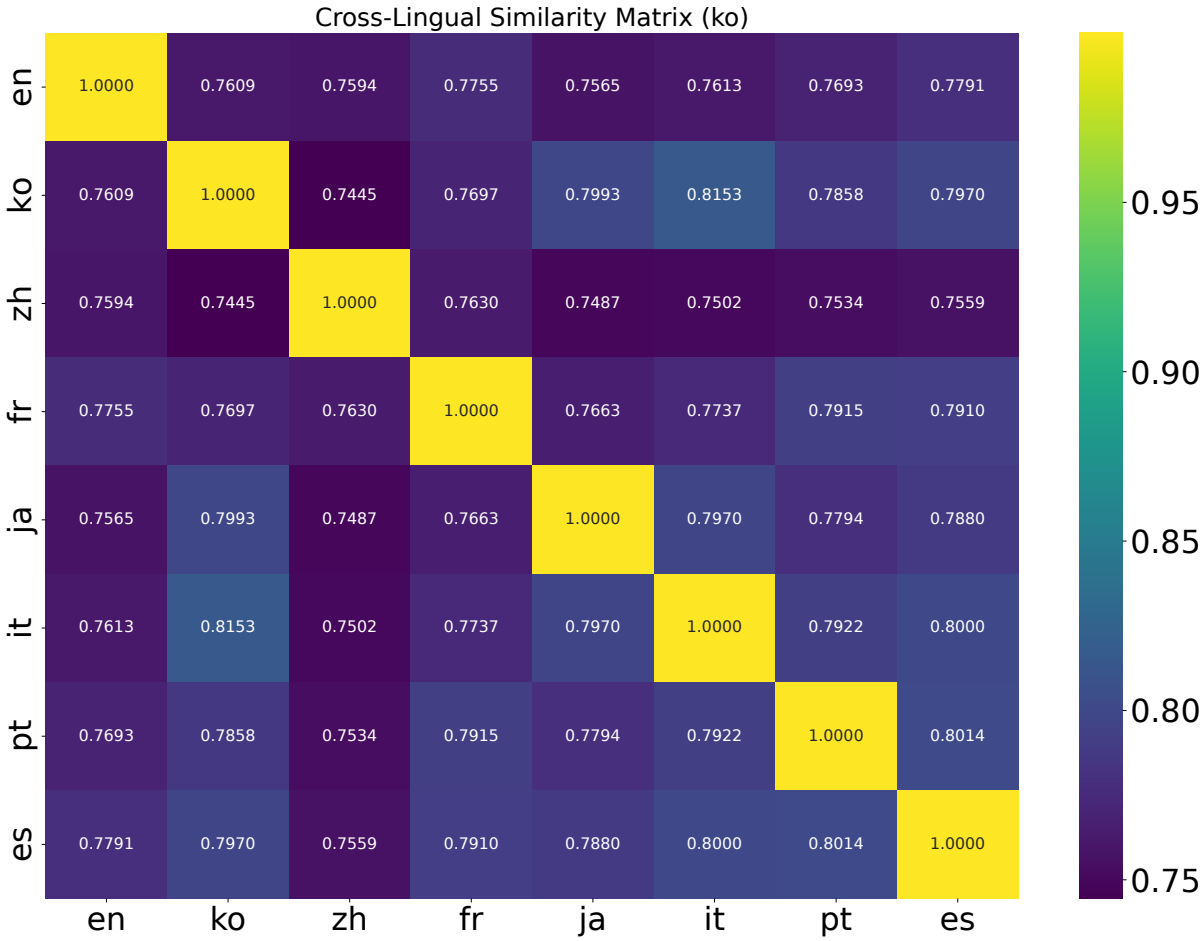

Figure 7: LaBSE Similarity Matrix (ko) of aya-expanse-8b.

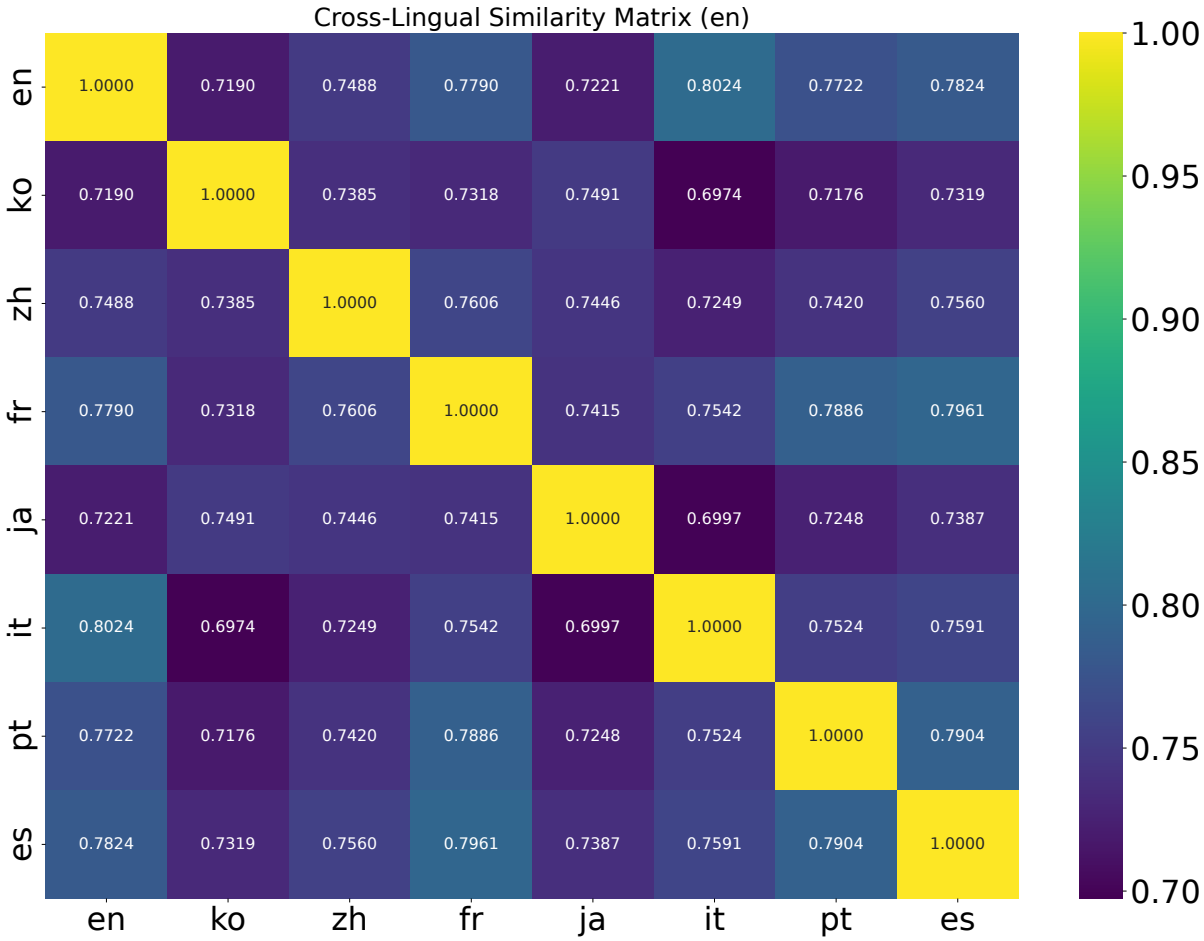

Figure 8: LaBSE Similarity Matrix (en) of Llama-3.1-8B-instruct.

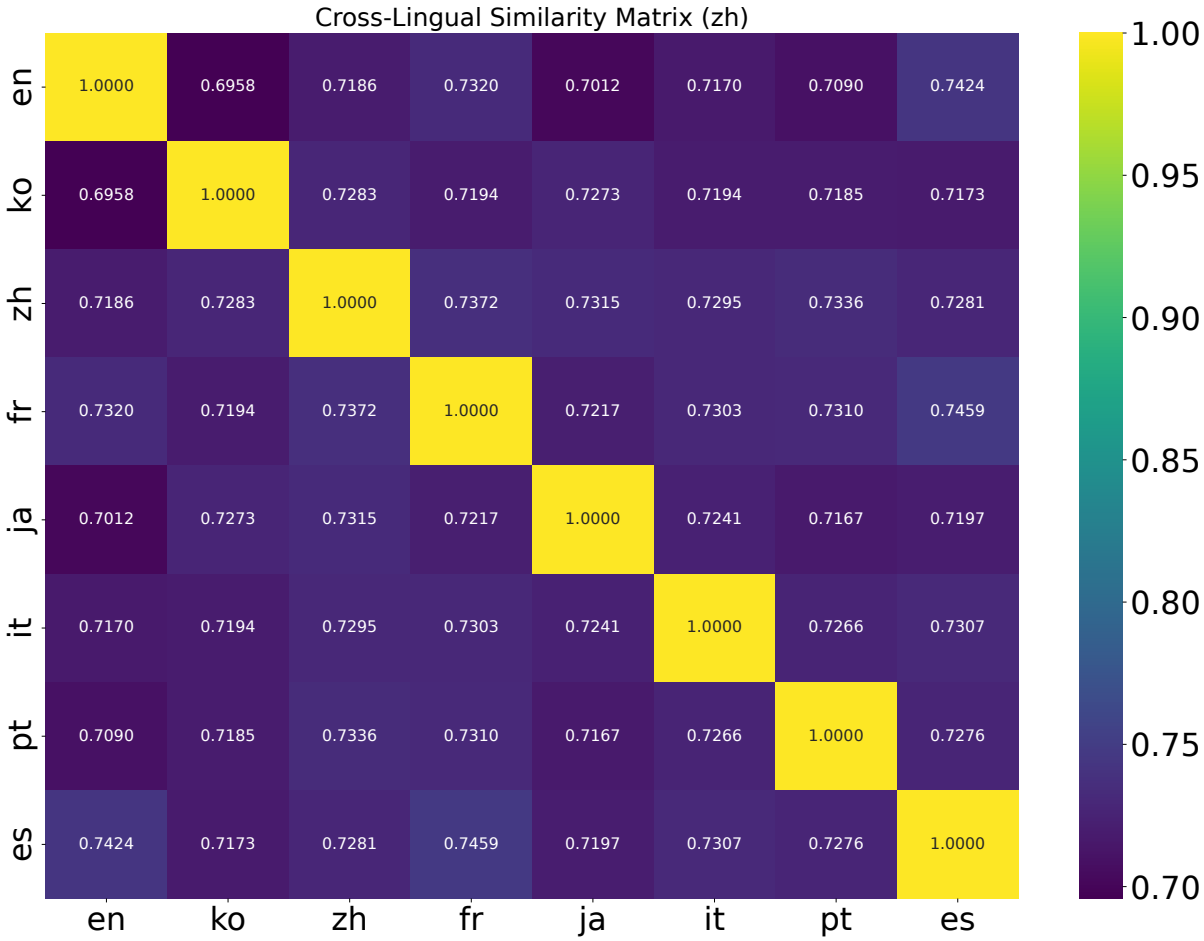

Figure 9: LaBSE Similarity Matrix (zh) of Llama-3.1-8B-instruct.

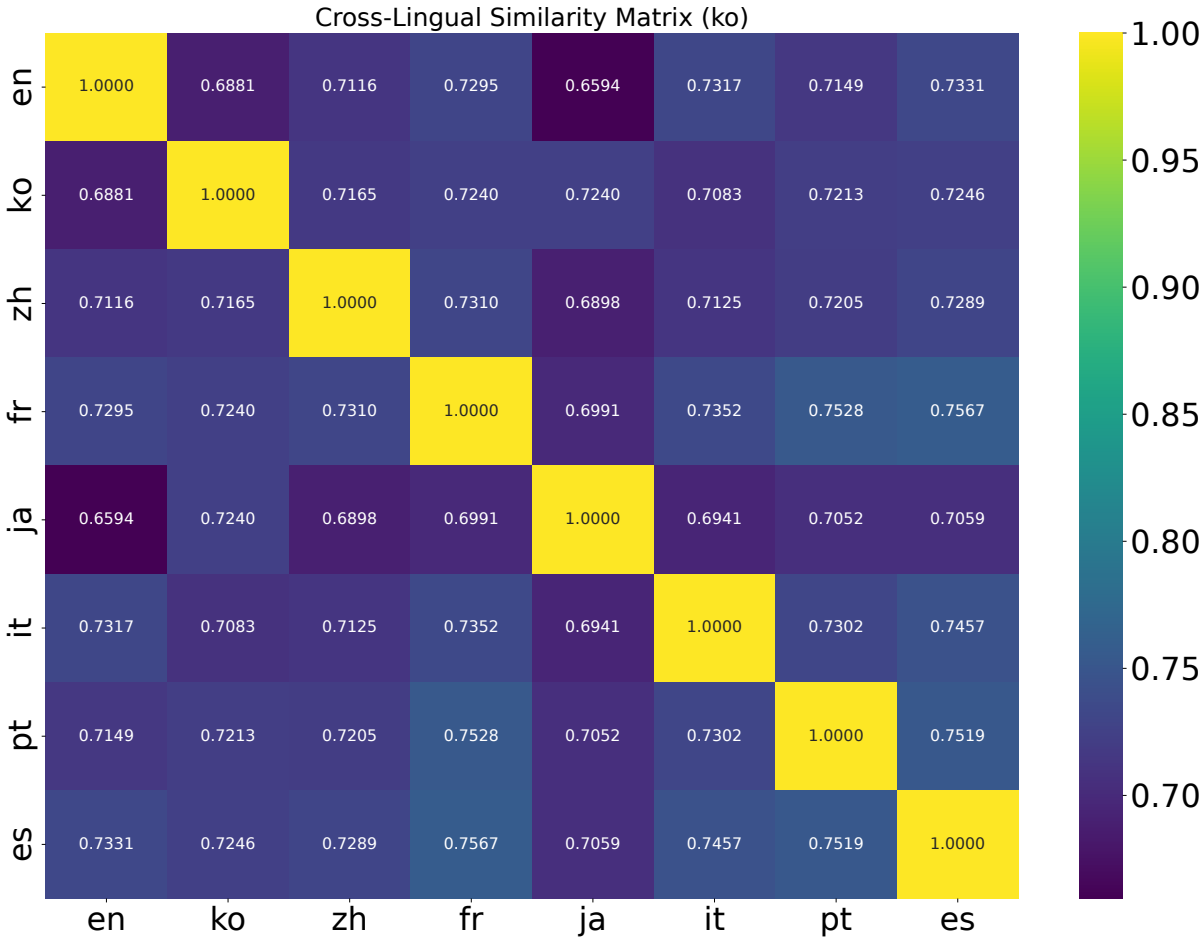

Figure 10: LaBSE Similarity Matrix (ko) of Llama-3.1-8B-instruct.

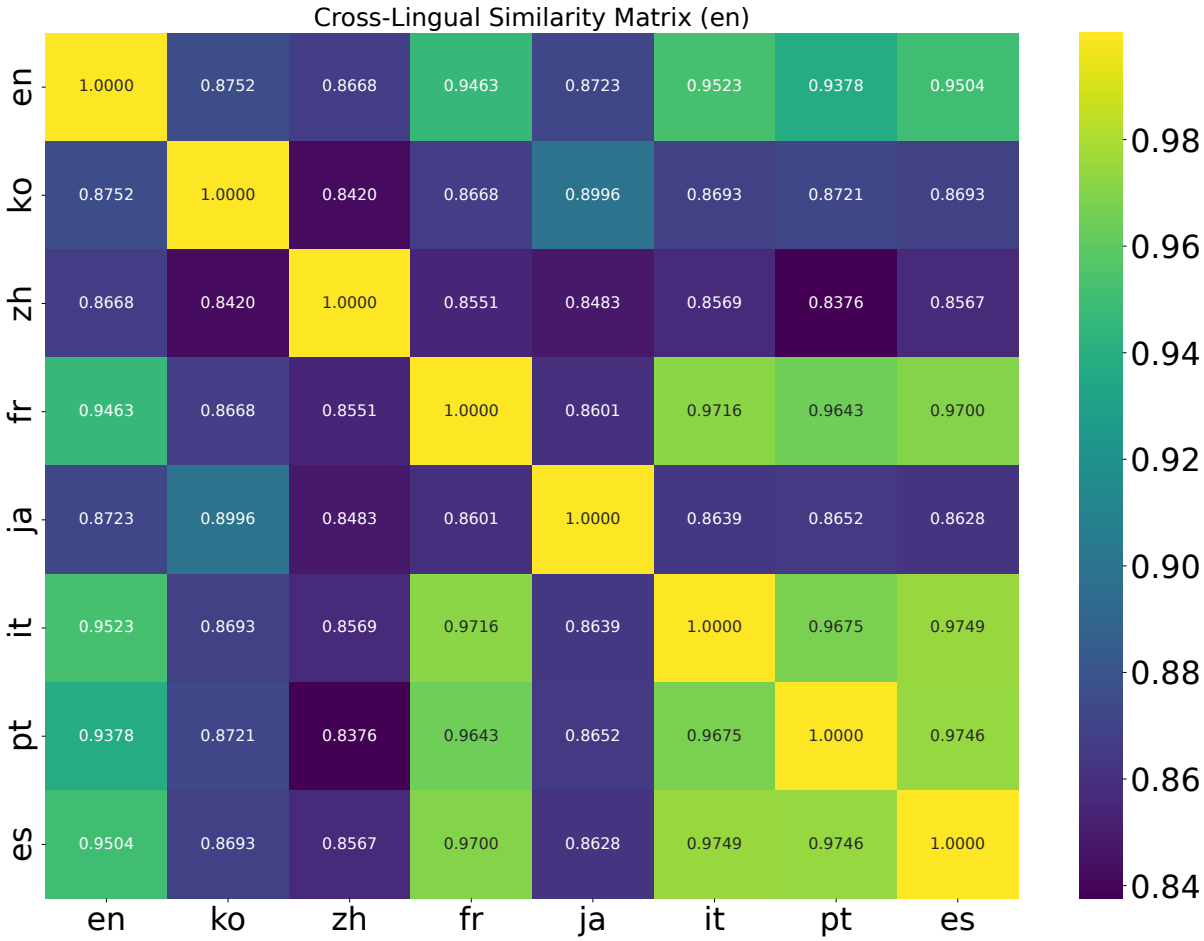

Figure 11: LaBSE Similarity Matrix (en) of gpt-4o-mini.

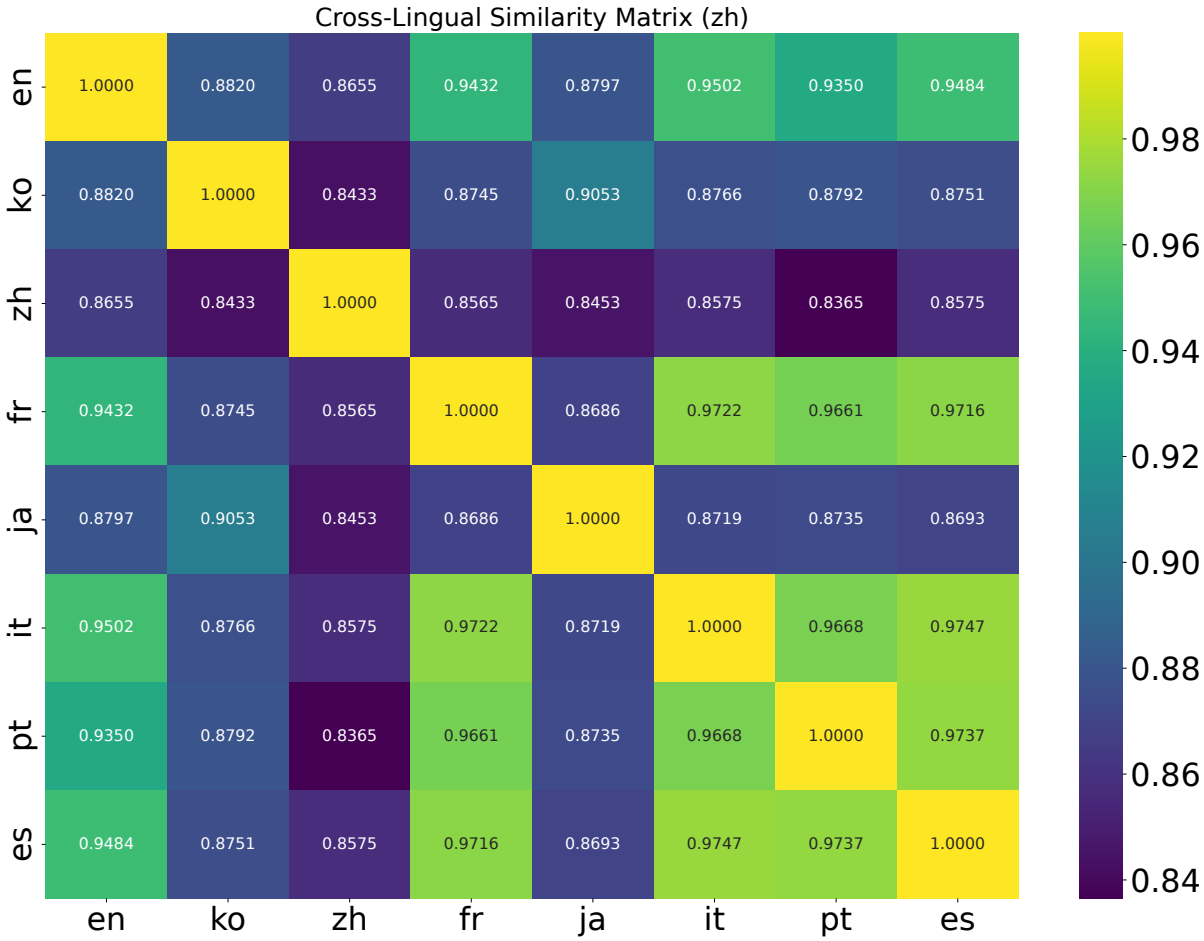

Figure 12: LaBSE Similarity Matrix (zh) of gpt-4o-mini.

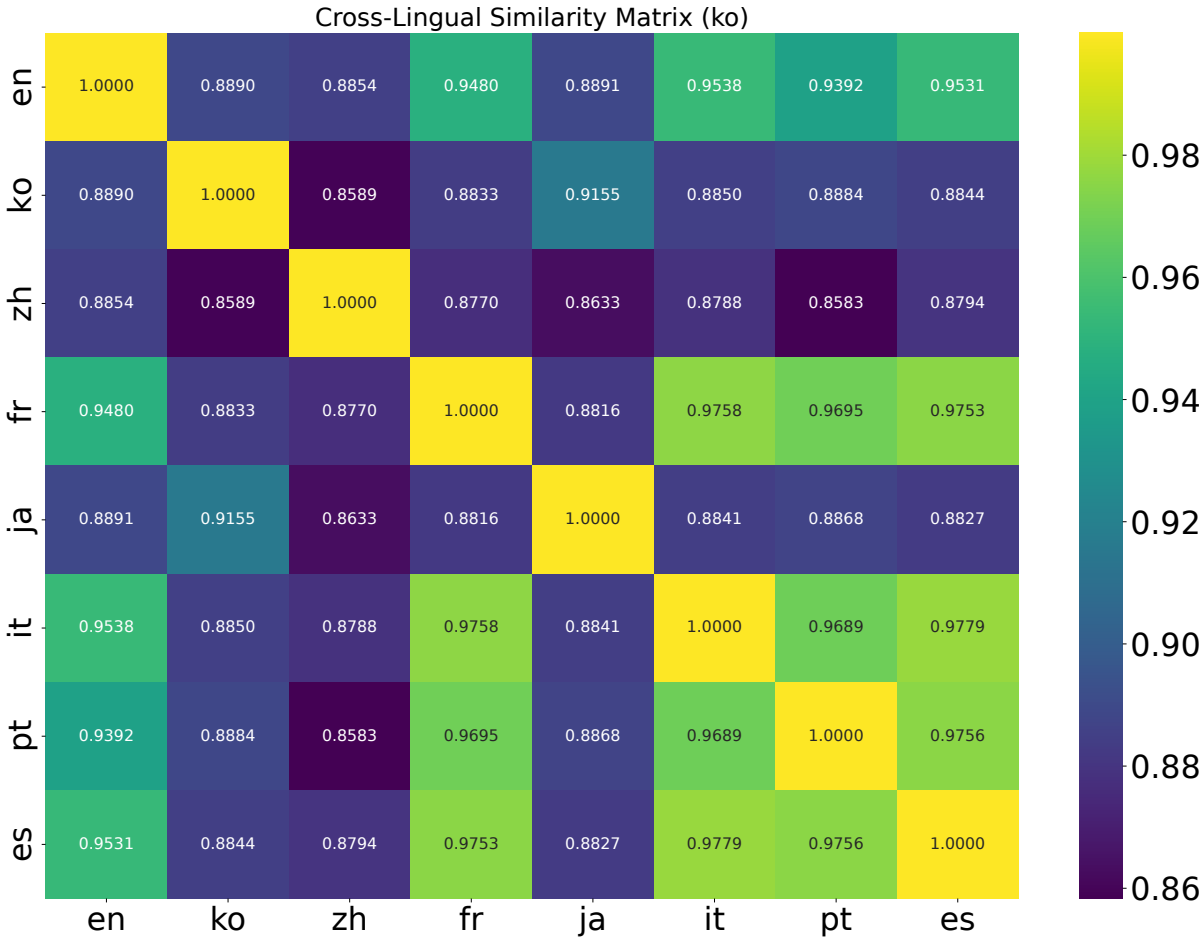

Figure 13: LaBSE Similarity Matrix (ko) of gpt-4o-mini.

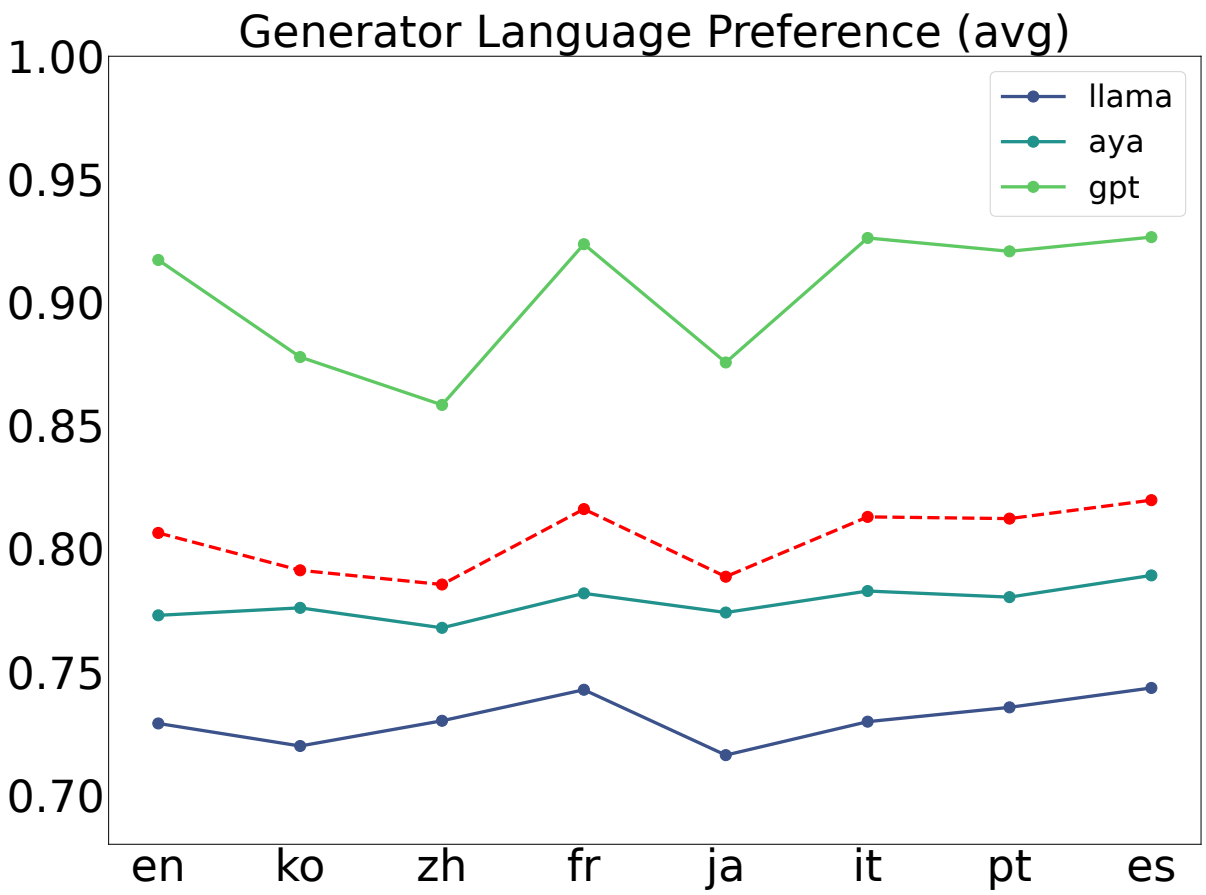

Figure 14: Average Generator Preference for three query languages: en, zh, ko.